curso y
ejercicios de

BIOESTADISTICA

j.r. vizmanos - r.asensio

curso y ejercicios de

BIOESTADISTICA

J. R. VIZMANOS

Profesor de la asignatura de BIOESTADISTICA
de la Facultad de Medicina y del
Colegio Universitario Arcos de Jalón de la
Universidad Complutense de Madrid

R. ASENSIO

Profesor de la asignatura de BIOESTADISTICA
de la Facultad de Medicina
de la Universidad Complutense de Madrid

MADRID
1976

PEDIDOS a:

J.R.Vizmanos
c/ Melilla, 12
M A D R I D-5

R.Asensio
c/ Abtao, 3
M A D R I D-7

Telfs: 2 66 18 60 - 2 52 57 92

Impreso en España - Printed in Spain
Depósito Legal:M - 31458 - 1976
I.S.B.N.: 84 - 400 - 1804 - 5

CENTRO DE PROMOCION REPROGRAFICA
Paseo de las Delicias, 52.MADRID-7

prólogo

El propósito de este libro es ofrecer algunas nociones de Estadística aplicada a la Medicina y a la Biología, tratando en todo momento de presentarlas de la manera más simple y didáctica posible sin perder por ello el debido rigor científico.

Cada tema va seguido de una serie de ejercicios totalmente resueltos con el fin de que el alumno ponga en práctica los conocimientos adquiridos.

Probablemente haya en esta obra algunos errores de cálculo o de exposición por lo que rogamos nos disculpen, aceptando gustosamente cuantas indicaciones se nos hagan en este sentido.

Finalmente, nos resta expresar nuestro agradecimiento al Dr. Sixto Rios, así como a tantos compañeros que, conocedores de nuestro esfuerzo por sacar adelante el presente libro, nos han estimulado y prestado su ayuda.

Septiembre 1976 **LOS AUTORES**

índice de materias

1. ESTADISTICA DESCRIPTIVA

6. ESTIMACION DE PARAMETROS

3. FORMULAS PARA LOS CONTRASTES

8. APLICACIONES DE LA X^2

9. ANALISIS DE LA VARIANZA

10. SERIES CRONOLOGICAS

* * * * * * * * *

1- estadística descriptiva

1. GENERALIDADES

1.1 Población
1.2 Unidades estadísticas o individuos
1.3 Muestra
1.4 Caracteres y modalidades
1.5 Caracteres estadísticos cualitativos
1.6 Caracteres estadísticos cuantitativos
1.7 Variables estadísticas
1.8 Variables estadísticas discretas
1.9 Variables estadísticas continuas

2. DISTRIBUCIONES ESTADISTICAS DE UN CARACTER

2.1 Tablas estadísticas
2.2 Tratamiento de la información
2.3 Representaciones gráficas
  2.3.1 Diagrama de barras
  2.3.2 Diagrama de sectores
  2.3.3 Pictogramas
  2.3.4 Histogramas
  2.3.5 Polígono de frecuencias
  2.3.6 Diagramas lineales
  2.3.7 Otras representaciones

3. MEDIDAS DE CENTRALIZACION

3.1 Media aritmética
3.2 Media geométrica
3.3 Media cuadrática
3.4 Media armónica
3.5 Relación entre estas medias
3.6 Mediana
3.7 Cuartiles
3.8 Deciles y percentiles
3.9 Moda
3.10 Relación empírica entre media, mediana y moda

4. MEDIDAS DE DISPERSION

4.1 Recorrido
4.2 Recorrido intercuartílico
4.3 Desviación media
4.4 Desviación típica
4.5 Varianza
4.6 Coeficiente de variación
4.7 Momentos respecto al origen
4.8 Momentos respecto a la media
4.9 Relación entre ambos momentos

5. MEDIDAS DE FORMA

5.1 Sesgo
5.2 Coeficiente de asimetría
5.3 Coeficiente de curtosis

1- ESTADISTICA DESCRIPTIVA

1. GENERALIDADES

1.1 POBLACION.- Se llama *población* al conjunto de todos los elementos que cumplen una o varias características.

Las poblaciones pueden ser *infinitas* (extracciones de una bolsa) o *finitas* (pacientes de un Centro Médico).

Para que una población quede bien definida, debemos saber si un elemento particular pertenece o no a la población.

A veces, a la población también se le llama *universo o colectivo*.

1.2 UNIDADES ESTADISTICAS O INDIVIDUOS.- Se llaman así a los elementos que componen la población.

El hecho de que se utilice el nombre de *individuos* se debe al origen demográfico de la Estadística Descriptiva.

1.3 MUESTRA.- Se llama *muestra* a un subconjunto de la población.

Si queremos extraer conclusiones sobre la población deberemos elegir una muestra *representativa* de ella. El proceso mediante el cual se extrae una muestra representativa de la poblacion se conoce con el nombre de *muestreo al azar* de manera que cada elemento de la población tiene la misma posibilidad de ser incluido en la muestra. Las muestras asi obtenidas se denominan *muestras aleatorias*.

Al número de elementos de la muestra, se le llama *tamaño*.

1.4 CARACTERES Y MODALIDADES.- Una vez que tengamos los individuos de una población o de una muestra nos interesará estudiar uno o varios *caracteres* de esos elementos.

Ahora bien, cada uno de esos caracteres de los individuos de la población pueden presentar varias *modalidades* de tal forma que cada individuo de la población presente *una y solamente una* de las modalidades del carácter.

Ej: Examinada una muestra de mujeres ingresadas en una Maternidad, respecto al carácter, estado civil, puede presentar las siguientes modalidades

Solteras - Casadas - Viudas - No consta

Ya se comprenderá que cuantas más modalidades se tengan de un carácter, más exhaustiva será la información que tendremos de los individuos de la población.

1.5 CARACTERES ESTADISTICOS CUALITATIVOS.- Se dice que un carácter estadístico es *cualitativo* si sus modalidades *no son medibles*.

Ej: Consideremos la población formada por los pacientes de un Centro Médico, son caracteres cualitativos, la raza, el sexo, el estado civil, la profesión. etc.

1.6 CARACTERES ESTADISTICOS CUANTITATIVOS.- Se dice que un carácter estadístico es *cuantitativo* si sus modalidades *son medibles*.

Ej: Consideremos la población del ejemplo anterior, serán caracteres estadísticos cuantitativos, la talla, el peso, la edad, el pulso, el perímetro torácico, la temperatura, etc.

1.7 VARIABLES ESTADISTICAS.- Ya hemos visto en el caso de caracteres estadísticos cuantitativos que sus modalidades son medibles es decir que a cada una de sus modalidades, se le puede asignar un número. Este número que es variable respecto a la modalidad, le llamaremos *variable estadística*.

1.8 VARIABLES ESTADISTICAS DISCRETAS.- Se dice que una variable estadística es *discreta*, cuando solo puede tomar determinados valores dentro de un intervalo. O también, cuando no puede tomar cualquier valor entre dos valores dados.

Es decir, que la variable estadística discreta está definida para *valores aislados*.

Generalmente las variables estadísticas discretas toman valores sobre el conjunto de los números enteros.

Ejs: Número de flores obtenidas de un clarkia elegans.
Número de hijos de un paciente.
Número de pétalos de una flor.

1.9 VARIABLES ESTADISTICAS CONTINUAS.- Se dice que una variable estadística es *contínua* cuando puede tomar cualquier valor dentro de los de un intervalo.

Es decir, que el conjunto de valores posibles es *infinito*.

Ejs: Temperatura de un paciente.
Presión sanguinea de una cobaya.
Espesor de la concha de una tortuga.

2. DISTRIBUCIONES ESTADISTICAS DE UN CARACTER

2.1 TABLAS ESTADISTICAS.- Supongamos una población formada por N unidades estadísticas tal que respecto de un determinado caracter C, presenta j modalidades distintas.

Designemos por f_i la frecuencia absoluta de la modalidad C_i y por h_i la frecuencia relativa de la clase C_i.

Llamaremos tabla estadística correspondiente al carácter C con sus distintas modalidades, a una tabla de la forma que a continuación se expone:

modalidades	frecuencias absolutas	frecuencias relativas
C_1	f_1	h_1
C_2	f_2	h_2
.	.	.
.	.	.
.	.	.
C_j	f_j	h_j
	$\sum_{k=1}^{j} f_k = N$	$\sum_{k=1}^{j} h_k = 1$

Ejs:

a) Sea el carácter C *cualitativo*.

Distribución según formas de masas de superficie forestal en España, en miles de hectáreas y referido al año 1.972. *Fuente: Anuario Estadístico 1974.*

forma de masas	miles de hectáreas
Monte alto	7.351
Monte mediano	608
Monte bajo	2.006
Matorral, pastos y otros	15.927

b) Sea el carácter C *cuantitativo*.

b-1) Caso de variable estadística *discreta*.

Distribución de los beneficiarios de título de familia numerosa de primera categoría según número de hijos referido al año 1971.

Fuente: Anuario estadístico 1974.

número de hijos	beneficiarios
1	128
2	1.192
3	4.528
4	201.621
5	84.303
6	38.198
7	17.878

b-2) Caso de variable estadística *contínua*

Distribución de porcentajes del perímetro torácico (en cms.) de españoles varones, referido al año 1973. *Fuente: Estadística de reclutamiento y reemplazo de los Ejercitos.*

Perímetro torácico en cms.	Porcentajes
Menos de 75	0,5
De 75 a 79	4,1
De 80 a 84	19,8
De 85 a 89	35,4
De 90 a 94	26,0
De 95 a 100	11,2
Más de 100	3,0

2.2 TRATAMIENTO DE LA INFORMACION.- En presencia de una distribución estadística existe un deseo espontaneo de simplificación.

Para ello se busca reemplazar la complejidad y multitud de datos por una característica única.

Ahora bien, conviene tener en cuenta que al reemplazar datos numéricos por una sola notación se traduce en una pérdida de la cantidad de información que poseemos de la distribución estadística.

A continuación, veamos como procederemos de manera ordenada para estudiar una muestra.

1.- Recogida de datos.- Consiste en recoger una colección de datos numéricos proporcionados por la muestra.

Ej: Tomemos las tallas de 100 soldados de un Regimiento escogidos de manera aleatoria.

2.- Ordenación de los datos.- Una vez recogidos los datos numéricos, los colocaremos en orden creciente o decreciente.

3.- Tablas de frecuencia.- Con el fin de poder trabajar de forma más sistemática, construiremos tablas de frecuencia como hemos visto en 2.1.

En nuestro ejemplo, la tabla de frecuencia sería de la siguiente forma:

Altura en cms.	Número de soldados f_i	h_i
160-164	14	0,14
165-169	28	0,28
170-174	24	0,24
175-179	18	0,18
180-184	12	0,12
185-189	4	0,04
	100 = N	1,-

A cada intervalo de la tabla de frecuencia se llama *clase*.

4.- Límites de clase.- La forma de determinar una clase es dando sus extremos, al extremo por la izquierda, le llamaremos *límite inferior* y al extremo por la derecha, le llamaremos *límite superior*.

5.- Límites reales de clase.- A veces, como en el ejemplo que estamos estudiando las clases no son consecutivas, es decir el extremo superior de la segunda clase es 169 y no coincide con el extremo inferior de la tercera clase que es 170, en este caso haremos una aproximación con el fin de que las clases sean consecutivas, de manera que el límite superior de una clase coincida con el límite inferior de la siguiente.

De esta forma construiremos unos nuevos límites que llamaremos, *límites reales de clase.*

En nuestro ejemplo, los límites reales de clase son los de la tabla adjunta.

límites reales
159,5 - 164,5
164,5 - 169,5
169,5 - 174,5
174,5 - 179,5
179,5 - 184,5
184,5 - 189,5

A los límites inferiores reales los representaremos por L_i y a los superiores por L_s.

6.- Tamaño de clase.- Se llama *tamaño de clase*, a la diferencia entre el límite superior real y el límite inferior real de dicha clase.

En general, es conveniente agrupar la distribución en clases de igual tamaño. Al tamaño de la clase lo representaremos por c.

7.- Marca de clase.- Se llama *marca de clase*, al punto medio de la clase.

Para calcular la marca de clase, se suman el límite superior real y el límite inferior real y se divide por dos.

Por ejemplo, la marca de la tercera clase, será:

$$c = \frac{169,5 + 174,5}{2} = 172$$

2.3 REPRESENTACIONES GRAFICAS.- Aún cuando la tabla estadística de una distribución encierra toda la información, a veces es necesario traducirla a un gráfico con el fín de hacerla más patente.

Veamos a continuación distintos tipos de representaciones gráficas.

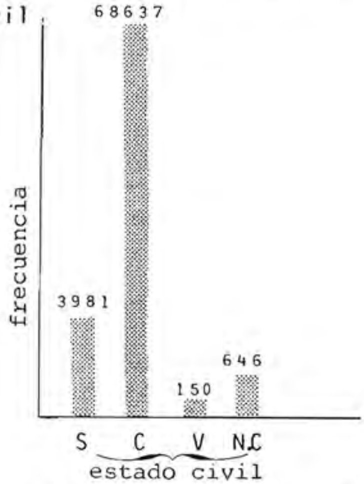
2.3.1. DIAGRAMA DE BARRAS.- Es especialmente útil si se desea comparar datos cualitativos o cuantitativos de tipo discreto.

a) *Cualitativo discreto.*

Ej: Distribución de mujeres ingresadas en Maternidades según su estado civil.

Estado civil	número de mujeres
Solteras	3.981
Casadas	68.637
Viudas	150
No consta	646

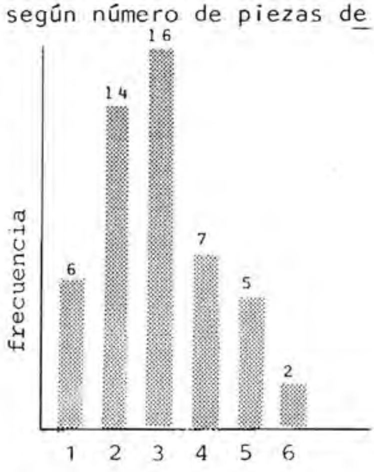
Fuente: Anuario Estadístico 1974



b) *Cuantitativo discreto.*

Ej: Distribución de lotes según número de piezas defectuosas

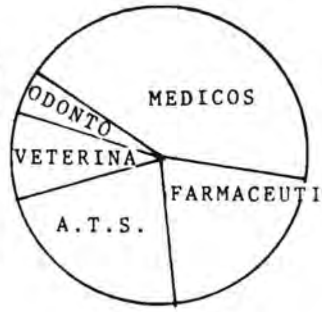
número de defectuosas por lote	frecuencia
1	6
2	14
3	16
4	7
5	5
6	2



2.3.2. DIAGRAMA DE SECTORES.- Consiste en representar mediante sectores circulares las distintas modalidades de un carácter. Teniendo en cuenta que los sectores circulares han de tener un ángulo central proporcional a la frecuencia absoluta correspondiente. El área del sector circular será proporcional a la frecuencia absolta.

Ej: Distribución de profesionales sanitarios en el año:

Profesiones	frecuencias
Médicos	51594
Odontólogos	3613
Farmaceuticos	17498
Veterinarios	7462
A.T.S.	25723



2.3.3. <u>PICTOGRAMAS</u>.- Se utilizan generalmente para representar la información mediante un dibujo alusivo a la distribución que se estudia.

Ej: Consideremos el número de plazas de albergues juveniles en las cuatro provincias siguientes:

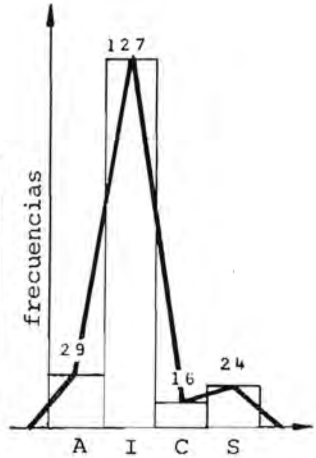
Palencia	100
Gerona	300
Oviedo	580
Madrid	850

Fuente: Anuario Estadístico 1974.

2.3.4. <u>HISTOGRAMAS</u>.- Ese tipo de representación se utiliza generalmente para variables contínuas y consiste en representar mediante un rectángulo, cada una de las modalidades, de tal manera que las alturas de dichos rectángulos son iguales a las frecuencias de clase, en el supuesto de que todas las clases tengan igual tamaño. En el caso contrario, las alturas han de ser calculadas teniendo en cuenta que las superficies de los rectángulos han de ser proporcionales a las frecuencias de cada clase.

Ej: Distribución de parados en miles según ramas de acti vidad.

Rama de actividad	frecuencia en miles
AGRICULTURA	29
INDUSTRIA	127
COMUNICACIONES Y TRANSPORTES	16
OTROS SERVICIOS	24



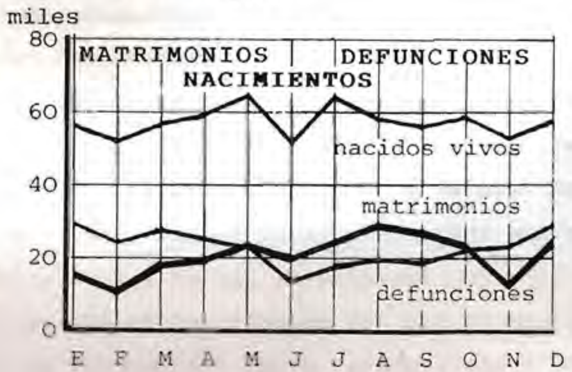
2.3.5. POLIGONO DE FRECUENCIAS.- Se llama *polígono de frecuencias* a la línea que une los puntos medios de las bases superiores de los rectángulos de un histograma de frecuencias.

En el caso anterior el polígono de frecuencias será el marcado en línea gruesa.

2.3.6. DIAGRAMAS LINEALES.- Este tipo de diagramas son muy utilizados para mostrar el cambio de valores con el paso del tiempo o edad.

Lo que interesa en este tipo de diagramas es solamente la altura de la línea referida a la base del diagrama.

Ej: Distribución de nacimientos, matrimonios y defunciones en el año 1974.



2.3.7. OTRAS REPRESENTACIONES.- Si en las representaciones gráficas anteriores representamos sobre el eje de ordenadas las frecuencias relativas en lugar de las absolutas, obtenemos lo que se llama, *polígono de frecuencias relativas* e *histograma de frecuencias relativas.*

La suma de las frecuencias absolutas de las clases inferiores a una dada, más la de la dada se llama *frecuencia absoluta acumulada.*

Las frecuencias absolutas acumuladas divididas por el número de datos nos dan las frecuencias relativas acumuladas. Utilizando estas frecuencias obtendremos *polígonos e histogramas de frecuencias absolutas acumuladas* o *polígonos e histogramas de frecuencias relativas acumuladas.*

3. MEDIDAS DE CENTRALIZACION

Se llaman *medidas de centralización*, a los valores que tienden a situarse en el centro del conjunto de datos ordenados respecto a su magnitud.

Veamos a continuación, mediante un organigrama las distintas medidas de centralización.

MEDIDAS DE CENTRALIZACION	DE TAMAÑO	media aritmética media geométrica media cuadrática media armónica
	DE POSICION	mediana cuartiles deciles y percentiles
	DE FRECUENCIA	moda

3.1 MEDIA ARITMETICA .- La *media aritmética* $\bar{x}$ de un conjunto de n números : $x_1, x_2, \ldots, x_n$ que se presentan con frecuencias absolutas $f_1, f_2, \ldots, f_n$, respectivamente es:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum f_i x_i}{N}$$

Ahora bien, cuando se trata de datos agrupados tomaremos las x_i como las marcas de clase, pero a pesar de ello resulta bastante incómodo su cálculo, por lo que conviene realizar un cambio de variable.

Para ello elegiremos un origen de trabajo que llamaremos x_o (que puede ser la marca de la clase con mayor frecuencia). Sea c la amplitud de los intervalos de clase que supondremos constante. En estas condiciones realizamos el siguiente cambio de variable: $x_i = x_o + cu_i$, siendo u_i la nueva variable.

Por tanto, tendremos:

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{\sum f_i (x_o + cu_i)}{N} = x_o \frac{\sum f_i}{N} + c \frac{\sum f_i u_i}{N}$$

Luego: $$\boxed{\bar{x} = x_o + c\,\bar{u}}$$

3.2 MEDIA GEOMETRICA.- La *media geométrica* G de un conjunto de n números $x_1, x_2, \ldots, x_n$ que se presentan con frecuencias $f_1, f_2, \ldots, f_n$, respectivamente, será:

$$G = \sqrt[N]{x_1^{f_1} . x_2^{f_2} \ldots x_n^{f_n}} \quad \text{siendo} \quad N = \sum_{i=1}^{n} f_i$$

3.3 MEDIA CUADRATICA.- La *media cuadrática* Q de un conjunto de n números $x_1, x_2, \ldots, x_n$ que se presentan con frecuencias $f_1, f_2, \ldots, \ldots, f_n$, respectivamente, será:

$$Q = \sqrt{\frac{f_1 x_1^2 + f_2 x_2^2 + \ldots + f_n x_n^2}{f_1 + f_2 + \ldots + f_n}}$$

3.4 MEDIA ARMONICA.- La *media armónica* H de un conjunto de números x_1, x_2, ... , x_n, que se presentan con frecuencias absolutas f_1, f_2,.. ..., f_n, respectivamente, será:

$$H = \frac{f_1 + f_2 + \dots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_n}{x_n}}$$

3.5 RELACION ENTRE ESTAS MEDIDAS.- A partir de las definiciones de las medias que acabamos de ver, se obtiene, la siguiente relación que liga las diferentes medias

$$H \leq G \leq \overline{x} \leq Q$$

3.6 MEDIANA.- La mediana de un conjunto de datos ordenados es un número que ocupa el valor medio de la distribución estadística.

a) Datos *no agrupados*

Supondremos los datos ordenados en orden creciente o decreciente.

Si hay un número impar de datos, la mediana viene dada por el valor medio.

Si hay un número par de datos, la mediana viene dada por la media aritmética de los valores medios.

b) Datos *agrupados*

En este caso el valor de la mediana se obtiene por interpolación lineal y viene dado por la siguiente expresión:

$$M = L_i + \left(\frac{\frac{N}{2} - (\Sigma f)_1}{f_{mediana}} \right) c$$

L_i = límite inferior real de la clase mediana

N = número total de datos, es decir, suma de las frecuencias absolutas

$(\Sigma f)_1$ = suma de las frecuencias absolutas de todas las clases anteriores a la clase mediana

$f_{mediana}$ = frecuencia absoluta de la clase mediana

c = tamaño del intervalo de la clase mediana.

Geometricamente.- La mediana es el valor de la abcisa x, que corresponde a la vertical que divide al histograma de frecuencias, en dos partes de igual área.

3.7 CUARTILES.- Acabamos de ver que la mediana divide al histograma de frecuencias en dos partes de igual área.

Del mismo modo se definen otras dos medidas de centralización que son los cuartiles.

Cuartil de primer orden Q_1.- Es el valor de la abcisa que corresponde a la vertical que divide al histograma de frecuencias en dos partes, dejando a la izquierda de Q_1 la cuarta parte de la distribución y a la derecha de Q_1 las tres cuartas partes de la distribución.

El valor de Q_1 viene dado por la siguiente expresión:

$$Q_1 = L_i + \left(\frac{\frac{N}{4} - (\Sigma f)_1}{f_{cuartil\ 1°}} \right) c$$

Cuartil de tercer orden Q_3 .- Es el valor de la abcisa que corresponde a la vertical que divide al histograma de frecuencias en dos partes, dejando a la izquierda de Q_3 las tres cuartas partes de la distribución y a la derecha de Q_3 la cuarta parte de la distribución.

El valor de Q_3 viene dado por la siguiente expresión:

$$Q_3 = L_i + \left(\frac{\frac{3N}{4} - (\Sigma f)_3}{f_{cuartil\ 3°}} \right) c$$

L_i = límite inferior real de la clase que contiene al $\left\{\begin{array}{l}\text{cuartil } 1^\circ\\ \text{cuartil } 3^\circ\end{array}\right.$

$(\Sigma f)_{1,3}$ = suma de frecuencias absolutas de todas las clases anteriores a la que contiene al $\left\{\begin{array}{l}\text{cuartil } 1^\circ\\ \text{cuartil } 3^\circ\end{array}\right.$

N = número total de datos

c = tamaño del intervalo de la clase $\left\{\begin{array}{l}\text{cuartil } 1^\circ\\ \text{cuartil } 3^\circ\end{array}\right.$

3.8 DECILES Y PERCENTILES.- Analogamente al concepto de cuartil, llamaremos *deciles* a los valores que dividen al conjunto de datos en diez partes iguales y los representaremos por:

$$D_1,\ D_2,\ D_3,\ \ldots\ ,\ D_9$$

Del mismo modo, llamaremos *percentiles*, a los valores que dividen el conjunto de datos en cien partes iguales y los representaremos por:

$$P_1,\ P_2,\ P_3,\ \ldots.\ ,\ P_{99}$$

Es evidente que $M = Q_2 = D_5 = P_{50}$ y que $Q_1 = P_{25}$; $Q_3 = P_{75}$

En general, a valores como cuartiles, deciles, percentiles y todos aquellos que obtengamos por subdivisiones análogas del conjunto de datos los llamaremos *cuantiles*.

3.9 MODA.- Llamaremos *moda* de un conjunto de datos y representaremos por M_o, al valor de la variable estadística que presenta mayor frecuencia. También se llama *valor dominante*.

La moda puede no existir, así como no ser única, en el caso de que haya dos modas la distribución se llamará *bimodal*, y en general cuando la moda sea múltiple la denominaremos, *multimodal*.

Para el cálculo de la moda, hemos de distinguir dos casos:

a) Variable estadística discreta.- En este caso la moda queda perfectamente definida, pues basta ver cual es el valor de la variable que presenta mayor frecuencia.

b) Variable estadística contínua.- En este caso, solo es posible conocer la clase modal que corresponderá a la clase que tenga mayor frecuencia.

Una primera forma de definir la moda sería la marca de la clase modal, pero veremos una expresión más exacta para el cálculo de la moda. Esta expresión es la siguiente:

$$M_o = Li + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c$$

Donde:

L_i = límite inferior real de la clase modal

Δ_1 = diferencia entre la frecuencia absoluta de la clase modal y la clase anterior

Δ_2 = diferencia entre la frecuencia absoluta de la clase modal y la clase siguiente

c = tamaño del intervalo de la clase modal.

3.10 RELACION EMPIRICA ENTRE MEDIA, MEDIANA Y MODA.- En el caso de que la distribución solo tenga una moda y no sea muy asimétrica se comprueba empiricamente la siguiente relación:

$$\bar{x} - M_o = 3\ (\bar{x} - M)$$

Por supuesto que en las curvas simétricas, los valores de la media, mediana y moda coinciden.

4. CARACTERISTICAS DE DISPERSION

A veces no basta con conocer los valores centrales, ya que es más importante conocer en que medida los datos numéricos estan agrupados o no alrededor de la media. A esto es a lo que se llama *dispersión*.

4.1 RECORRIDO.- Se llama *recorrido*, a la diferencia entre el límite superior real de la última clase y el límite inferior real de la primera clase. También se llama *rango*.

4.2 RECORRIDO INTERCUARTILICO.- Se llama así a la diferencia entre los cuartiles de tercero y primer orden. Es decir: $Q_3 - Q_1$.

El recorrido intercuartílico es una medida de dispersión que elimina la influencia de los valores extremos.

4.3 DESVIACION MEDIA.- Se llama *desviación media*, a la media aritmética de las desviaciones de los datos respecto a la media, consideradas estas desviaciones en valor absoluto.

Es decir:

$$D_{\bar{x}} = \frac{\sum f_i \, |x_i - \bar{x}|}{\sum f_i} = \overline{|x_i - \bar{x}|}$$

En realidad se debería llamar *desviación absoluta media, respecto a la media.*

4.4 DESVIACION TIPICA.- Se define la desviación típica como la raiz cuadrada de la media aritmética de los cuadrados de las desviaciones respecto a la media.

Es decir:

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}}$$

Haciendo operaciones en la expresión de la desviación típica, podemos obtener una expresión que con frecuencia resulta más cómoda

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum f_i (x_i^2 - 2\bar{x}x_i + \bar{x}^2)}{N}} =$$

$$= \sqrt{\frac{\sum f_i x_i^2}{N} - 2\bar{x}\frac{\sum f_i x_i}{N} + \bar{x}^2 \frac{\sum f_i}{N}} = \sqrt{\frac{\sum f_i x_i^2}{N} - \bar{x}^2}$$

ya que como se recordará $\frac{\sum f_i x_i}{N} = \bar{x}$ y $\frac{\sum f_i}{N} = 1$

Ahora bien, en los casos de datos agrupados, conviene realizar un cambio de variable, análogo a como vimos para la media aritmética, con el fín de hacer los cálculos más sencillos. Sea c la amplitud de los intervalos, elegimos x_o como el nuevo origen de trabajo y sea u_i la nueva variable, con lo que

$$\left.\begin{array}{l} x_i = x_o + c\,u_i \\ \overline{x} = x_o + c\,\overline{u} \end{array}\right\} \Longrightarrow \quad x_i - \overline{x} = c(u_i - \overline{u})$$

y sustituyendo resulta:

$$s = \sqrt{\frac{\sum f_i (x_i - \overline{x})^2}{N}} = \sqrt{\frac{\sum f_i c^2 (u_i - \overline{u})^2}{N}} = c\sqrt{\frac{\sum f_i u_i^2}{N} - \overline{u}^2}$$

A la desviación típica, también se la llama *desviación standar* o *desviación cuadrática media.*

4.5 VARIANZA.- Se define como *varianza*, el cuadrado de la desviación típica.

Es decir: $$s^2 = \frac{\sum f_i (x_i - \overline{x})^2}{N}$$

4.6 COEFICIENTE DE VARIACION.- Es el cociente entre la desviación típica y la media aritmética.

Es decir: $$d = \frac{s}{\overline{x}}$$

Obsérvese que el coeficiente de variación no tiene gran utilidad cuando la media $\overline{x}$, se aproxima a cero, ya que entonces d toma valores infinitamente grandes.

4.7 MOMENTOS RESPECTO AL ORIGEN.- Llamaremos *momento de orden r respecto al origen* y representaremos por α_r, a la siguiente expresión:

$$\alpha_r = \frac{\sum f_i x_i^r}{N}$$

Obsérvese que el momento de primer orden respecto al origen es la media aritmética.

Es decir: $$\alpha_1 = \frac{\sum f_i x_i}{N} = \overline{x}$$

4.8 MOMENTOS RESPECTO A LA MEDIA.- Llamaremos *momento de orden r respecto a la media* y representaremos por m_r a la siguiente expresión:

$$m_r = \frac{\sum f_i (x_i - \overline{x})^r}{N}$$

A los momentos respecto a la media, también se llaman *momentos centrales*.

Obsérvese que el momento central de primer orden, es nulo. En efecto, ya que:

$$m_1 = \frac{\sum f_i (x_i - \bar{x})}{N} = \frac{\sum f_i x_i}{N} - \bar{x}\frac{\sum f_i}{N} = \bar{x} - \bar{x} = 0$$

El momento de segundo orden coincide con la varianza. En efecto:

$$m_2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = s^2$$

4.9 RELACION ENTRE LOS MOMENTOS CENTRALES Y RESPECTO AL ORIGEN.-

$$m_1 = \frac{\sum f_i (x_i - \bar{x})}{N} = 0$$

$$m_2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = \alpha_2 - \alpha_1^2 = s^2$$

$$m_3 = \frac{\sum f_i (x_i - \bar{x})^3}{N} = \frac{\sum f_i x_i^3}{N} - 3\bar{x}\frac{\sum f_i x_i^2}{N} + 3\bar{x}^2\frac{\sum f_i x_i}{N} - \bar{x}^3\frac{\sum f_i}{N} =$$

$$= \alpha_3 - 3\,\alpha_1\,\alpha_2 + 2\,\alpha_1^3$$

$$m_4 = \frac{\sum f_i (x_i - \bar{x})^4}{N} = \alpha_4 - 4\,\alpha_3\,\alpha_1 + 6\,\alpha_1^2\,\alpha_2 - 3\,\alpha_1^4$$

Y así sucesivamente.

Para el cálculo de los momentos conviene realizar el cambio de variable utilizado en el cálculo de la media aritmética y desviación típica.

Es decir: $\left.\begin{array}{l} x_i = x_o + c\,u_i \\ \bar{x} = x_o + c\,\bar{u} \end{array}\right\} \Longrightarrow x_i - \bar{x} = c(u_i - \bar{u})$

Por tanto resulta:

$$m_r = \frac{\sum f_i (x_i - \bar{x})^r}{N} = \frac{\sum f_i \left[c(u_i - \bar{u})\right]^r}{N} = c^r\,\frac{\sum f_i (u_i - \bar{u})^r}{N}$$

5. MEDIDAS DE FORMA

Además de las medidas de centralización y de dispersión conviene al estudiar una distribución conocer su *forma* mediante un índice lo más simplificado posible.

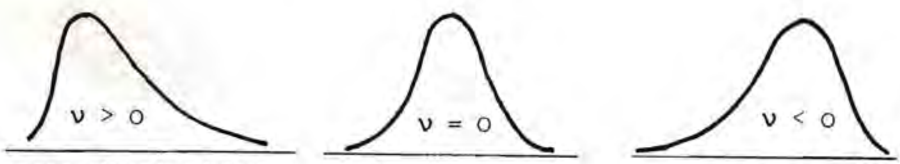
5.1 SESGO.- Se denomina *sesgo* a la mayor o menor simetría o asimetría de una distribución.

Si la curva de frecuencias de la distribución tiene una "cola" más larga hacia la izquierda que hacia la derecha respecto al máximo central, se dice que la distribución está *sesgada a la izquierda* o que tiene sesgo negativo. En el caso contrario se dice que la distribución está *sesgada a la derecha* o que tiene sesgo positivo.

A las distribuciones con sesgo nulo, se las llama *insesgadas*.

El coeficiente de sesgo viene dado por la expresión:

$$\nu = \frac{\bar{x} - M_o}{s}$$



sesgada a la dcha. insesgada sesgada a la izda.

A este coeficiente de sesgo, se le conoce como primer coeficiente de PEARSON.

5.2 COEFICIENTE DE ASIMETRIA.- Se denomina así a otra importante medida del sesgo, definida del siguiente modo:

$$\gamma_1 = \frac{m_3}{s^3} = \frac{1}{s^3} \frac{\sum f_i (x_i - \bar{x})^3}{N}$$

Como en el caso anterior, este coeficiente es nulo para las distribuciones insesgadas, negativo para las distribuciones con sesgo a la izquierda y positivo para las distribuciones con sesgo a la derecha.

A este coeficiente, también se le llama primer coeficiente de FISHER.

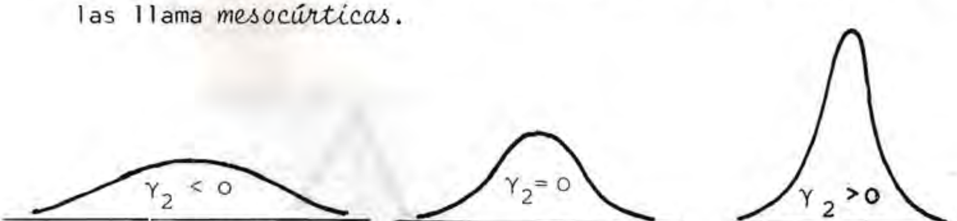
5.3 CURTOSIS.- Se denomina *curtosis* al grado de "apuntamiento" de una distribución.

Generalmente, se compara el mayor o menor apuntamiento de una distribución con la curtosis de la distribución normal. Por lo que viene expresado en la forma:

$$\gamma_2 = \frac{m_4}{s^4} - 3$$

De esta forma el coeficiente de curtosis γ_2, será nulo para la distribución normal. Si γ_2 es negativo, se trata de una distribución menos apuntada que la normal y recibe el nombre de *planicúrtica*. Si γ_2 es positivo, se trata de una distribución más apuntada que la normal y recibe el nombre de *leptocúrtica*.

A las distribuciones con coeficiente de curtosis nulo se las llama *mesocúrticas*.



platicúrtica mesocúrtica leptocúrtica

****

PROBLEMAS RESUELTOS

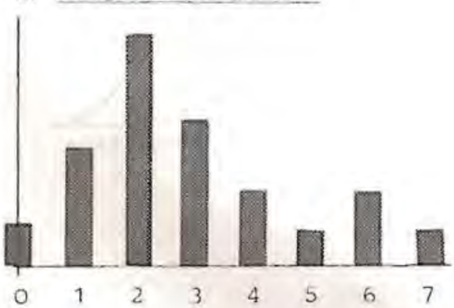
1.1 *El número de hijos de 20 familias seleccionadas al azar, es el siguiente: 3,1,2,2,1,5,2,2,0,6,3,2,4,3,4,2,3,1,7,6*
a) Formar la tabla de frecuencias.
b) Construir el correspondiente diagrama de barras.
c) Construir el polígono de frecuencias.

S O L U C I O N:

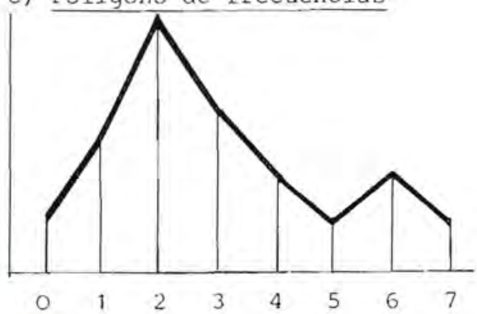
a) Tabla de frecuencias

nºde hijos	recuento	frecuencia absoluta f_i	frecuencia relativa h_i
0	/	1	0,05
1	///	3	0,15
2	///// /	6	0,30
3	////	4	0,20
4	//	2	0,10
5	/	1	0,05
6	//	2	0,10
7	/	1	0,05
		$N=20=\sum f_i$	$1=\sum h_i$

b) Diagrama de barras



c) Polígono de frecuencias



****

1.2 *Los valores del pH sanguineo en 80 individuos son los siguientes:*

7,33	7,32	7,34	7,40	7,28	7,29	7,35	7,33	7,34	7,28
7,31	7,35	7,32	7,33	7,33	7,36	7,32	7,31	7,35	7,36
7,26	7,39	7,29	7,32	7,34	7,30	7,34	7,32	7,39	7,30
7,33	7,33	7,35	7,34	7,33	7,36	7,33	7,35	7,31	7,33
7,37	7,38	7,38	7,33	7,35	7,30	7,31	7,33	7,35	7,33
7,27	7,33	7,32	7,31	7,34	7,32	7,34	7,32	7,31	7,36
7,30	7,37	7,33	7,32	7,31	7,33	7,32	7,30	7,29	7,38
7,33	7,35	7,32	7,33	7,32	7,34	7,32	7,34	7,32	7,33

a) *Formar la tabla de frecuencias.*
b) *Construir el histograma de frecuencias absolutas.*
c) *Polígono de frecuencias absolutas.*
d) *Construir el histograma de frecuencias absolutas acumuladas.*
e) *Polígono de frecuencias absolutas acumuladas.*

S O L U C I O N :

a) Tabla de frecuencias:

El rango es la diferencia entre el mayor y el menor valor del pH.

Es decir: $7,40 - 7,26 = 0,14$

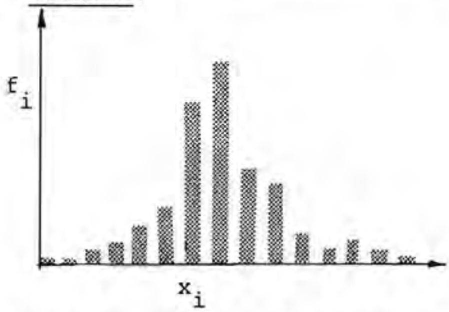
Construiremos 15 clases. Tomando 0,005 por la izquierda y por la derecha de las primera y última clase respectivamente, resulta:

$7,405 - 7,255 = 0,15$

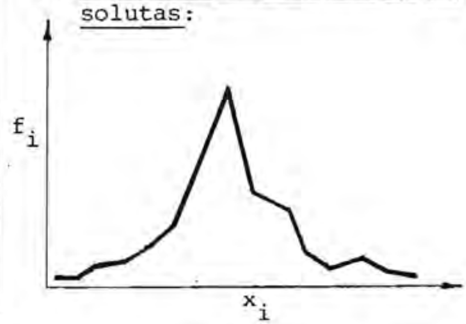
Luego el tamaño del intervalo de clase será: $\frac{0,15}{15} = 0,01$

clases	marcas de clase	recuento	f_i	h_i	F_i	H_i
7,255-7,265	7,26	/	1	0,0125	1	0,0125
7,265-7,275	7,27	/	1	0,0125	2	0,0250
7,275-7,285	7,28	//	2	0,0250	4	0,0500
7,285-7,295	7,29	///	3	0,0375	7	0,0875
7,295-7,305	7,30	/////	5	0,0625	12	0,1500
7,305-7,315	7,31	///// //	7	0,0875	19	0,2375
7,315-7,325	7,32	///// ///// ////	14	0,1750	33	0,4125
7,325-7,335	7,33	///// ///// ///// ///	18	0,2250	51	0,6375
7,335-7,345	7,34	///// ////	9	0,1125	60	0,7500
7,345-7,355	7,35	///// ///	8	0,1000	68	0,8500
7,355-7,365	7,36	////	4	0,0500	72	0,9000
7,365-7,375	7,37	//	2	0,0250	74	0,9250
7,375-7,385	7,38	///	3	0,0375	77	0,9625
7,385-7,395	7,39	//	2	0,0250	79	0,9875
7,395-7,405	7,40	/	1	0,0125	80	1,0000

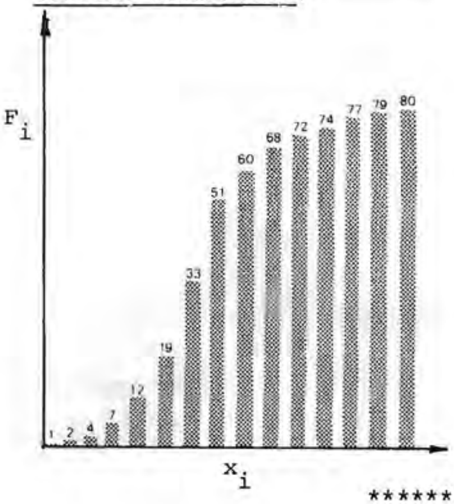
b) Histograma de frecuencias absolutas:



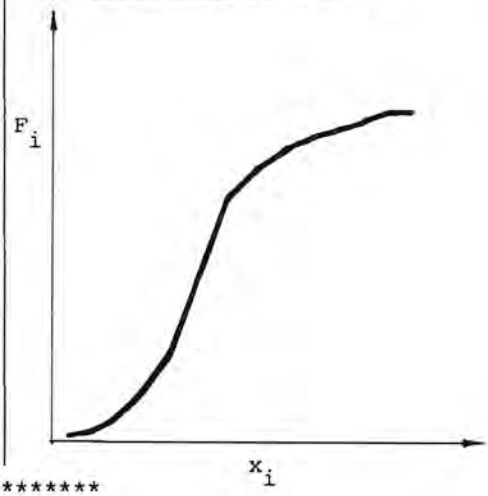
c) Polígono de frecuencias absolutas:



d) Histograma de frecuencias absolutas acumuladas:



e) Polígono de frecuencias absolutas acumuladas:



****

1.3 *Se ha medido la longitud de 50 individuos adultos de una determinada especie de rana, obteniendose los siguientes resultados:*

32,1	31,0	32,6	30,0	32,8	31,4	32,0	30,0	30,1	31,8
34,0	31,7	33,0	31,0	32,3	32,6	32,0	31,4	30,2	32,0
33,0	31,4	32,4	31,6	32,7	34,0	33,2	33,1	33,7	31,0
31,8	33,0	32,3	31,4	32,4	31,4	34,0	33,4	32,7	32,3
32,2	33,1	34,2	31,3	29,6	32,7	33,0	31,4	32,6	33,0

a) *Formar la tabla de frecuencias.*
b) *Construir el histograma de frecuencias relativas.*
c) *Polígono de frecuencias relativas.*
d) *Construir el histograma de frecuencias relativas acumuladas.*
e) *Polígono de frecuencias relativas acumuladas.*

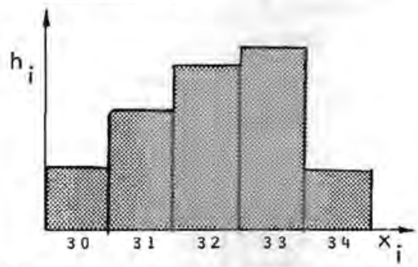
S O L U C I O N:

a) Tabla de frecuencias:

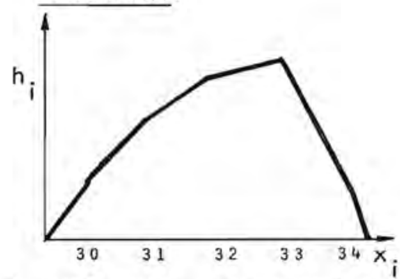
El rango es 34,2 - 29,6 = 4,6. Agruparemos los datos en cinco clases de 1 cm de amplitud cada una, siendo el límite inferior real de la primera 29,5 y el superior real de la última 34,5

limites reales	recuento	f_i	h_i	H_i
29,5 - 30,5	/////	5	0,1	0,1
30,5 - 31,5	///// /////	10	0,2	0,3
31,5 - 32,5	///// ///// ////	14	0,28	0,58
32,5 - 33,5	///// ///// ///// /	16	0,32	0,90
33,5 - 34,5	/////	5	0,1	1,00

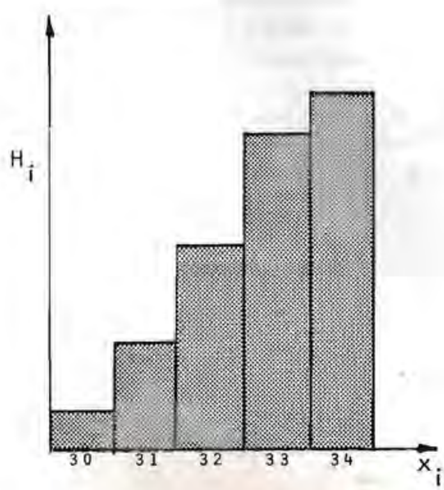
b) Histograma de frecuencias relativas:



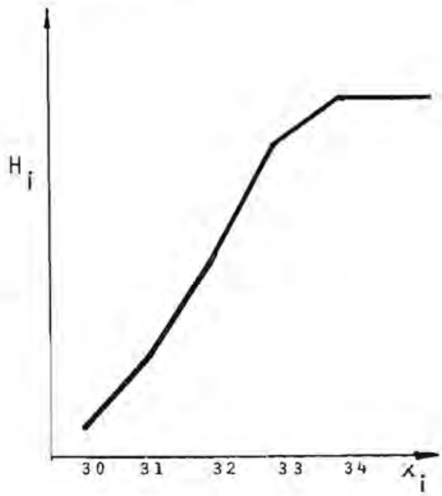
c) Polígono de frecuencias relativas:



d) Histograma de frecuencias acumuladas:



e) Polígono de frecuencias acumuladas:



****

1.4 *El número de accidentes mortales diarios en una gran ciudad en nueve días han sido: 6, 4, 8, 1, 5, 3, 3, 7, 2.*

a) Hallar la media aritmética.
b) Hallar la media geométrica.
c) Hallar la media cuadrática.
d) Hallar la media armónica.
e) Relación entre estas medias.

SOLUCION:

a) Media aritmética: $\bar{x} = \dfrac{\sum x_i}{N} = \dfrac{39}{5} = 4,333;$ $\boxed{\bar{x} \simeq 4,33}$

b) Media geométrica: $G = \sqrt[N]{x_1 . x_2 . \ldots . x_n} = (x_1 . x_2 . \ldots . x_n)^{1/N}$

tomando logaritmos, resulta: $\log G = \dfrac{1}{N} (\log x_1 + \log x_2 + \; . \; . \; . \; +$

$+ \log x_n) = \dfrac{1}{N} \sum \log x_i$

$\log G = \dfrac{5,08263}{9} = 0,56473$

tomando antilogaritmos resulta:

$G = \text{antilg } 0,56473 = 3,67054$

$\boxed{G \simeq 3,67}$

x_i	$\log x_i$	x_i^2	$\frac{1}{x_i}$
1	0,00000	1	1,000
2	0,30103	4	0,500
3	0,47712	9	0,333
3	0,47712	9	0,333
4	0,60206	16	0,250
5	0,69897	25	0,200
6	0,77815	36	0,166
7	0,84509	49	0,142
8	0,90309	64	0,125
39	5,08263	213	3,049

c) Media cuadrática:

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + \; . \; . \; . + x_n^2}{N}} = \sqrt{\frac{\sum x_i^2}{N}} = \sqrt{\frac{213}{9}} = 4,86$$

$\boxed{Q \simeq 4,86}$

d) Media armónica:

$$H = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \; . \; . \; . + \frac{1}{x_n}} = \frac{N}{\sum \frac{1}{x_i}} = \frac{9}{3,049} \simeq 2,95$$

$\boxed{H \simeq 2,95}$

e) Relación entre estas medias:

Como $\overline{x} \simeq 4,33$; $G \simeq 3,67$; $Q \simeq 4,86$; $H \simeq 2,95$.

Se tiene que : $\boxed{H \leq G \leq \overline{x} \leq Q}$

****

1.5 *El número de individuos muertos por cólera en un determinado país por año, en el transcurso de 11 años, han sido:*

2, 17, 5, 8, 12, 3, 2, 8, 12, 12, 3.

a) Hallar la media aritmética.
b) Hallar la media geométrica.
c) Hallar la media cuadrática.
d) Hallar la media armónica.
e) Relación entre estas medias.

S O L U C I O N:

En este caso podemos considerar los datos como una serie de datos agrupados, en la que el número de veces que se presenta cada dato sería la frecuencia absoluta correspondiente.

Con lo que construiremos la siguiente tabla:

x_i	f_i	$f_i x_i$	$\log x_i$	$f_i . \log x_i$	$f_i x_i^2$	$\frac{f_i}{x_i}$
2	2	4	0,30103	0,60203	8	1,0000
3	2	6	0,47712	0,95424	18	0,6666
5	1	5	0,69897	0,69897	25	0,2000
8	2	16	0,90309	1,80618	128	0,2500
12	3	36	1,07918	3,23754	432	0,2500
17	1	17	1,23045	1,23045	289	0,0588
	11	84	4,68984	8,52941	900	2,4254

a) Media aritmética: $\overline{x} = \dfrac{\sum f_i x_i}{N} = \dfrac{84}{11} = 7,63.$ $\boxed{x \simeq 7,63}$

b) Media geométrica: $G = \sqrt[N]{x_1^{f_1} . x_2^{f_2} . \ldots . x_n^{f_n}}$

tomando logaritmos resulta: $\log G = \dfrac{1}{N} \sum f_i . \log x_i = \dfrac{8,52941}{11} =$

$= 0,775403$ y tomando antilogaritmos resulta: $\boxed{G \simeq 5,96}$

c) Media cuadrática: $Q = \sqrt{\dfrac{\sum f_i x_i^2}{N}} = \sqrt{\dfrac{900}{11}} \simeq 9{,}04$; $\boxed{Q \simeq 9{,}04}$

d) Media armónica: $H = \dfrac{N}{\sum \dfrac{f_i}{x_i}} = \dfrac{11}{2{,}4254} \simeq 4{,}53$; $\boxed{H \simeq 4{,}53}$

e) Relación entre estas medias:

Como $\bar{x} \simeq 7{,}63$; $G \simeq 5{,}96$; $Q \simeq 9{,}04$; $H \simeq 4{,}53$.

Se tiene que: $\boxed{H \leq G \leq \bar{x} \leq Q}$

****

1.6 *La distribución de 500 estudiantes de primer curso de la Facultad de Medicina de la Universidad Complutense de Madrid, según el pulso (latidos por minuto), tras una hora de observación, es la siguiente:*

40-49	3
50-59	8
60-69	70
70-79	195
80-89	125
90-99	86
100-109	12
110-119	1

a) Calcular la media aritmética.

S O L U C I O N:

Se trata cono vemos, de una distribución de variable estadística contínua, en la que se han agrupado los datos en ocho clases de igual tamaño. Construiremos los límites reales de clase y haremos un cambio de variable para facilitar el cálculo de la media $\bar{x}$.

límites de clase	limites reales de clase	f_i	u_i	$f_i u_i$
40 - 49	39,5 - 49,5	3	-3	-9
50 - 59	49,5 - 59,5	8	-2	-16
60 - 69	59,5 - 69,5	70	-1	-70
70 - 79	69,5 - 79,5	195	0	0
80 - 89	79,5 - 89,5	125	1	125
90 - 99	89,5 - 99,5	86	2	172
100 - 109	99,5 -109,5	12	3	36
110 - 119	109,5 -119,5	1	4	4
		500		242

Media aritmética: $\bar{X} = x_o + c\,\bar{u} = \bar{x}_o + c\,\dfrac{\sum f_i u_i}{N}$

x_o es el nuevo origen de trabajo, que en nuestro caso es la marca de la cuarta clase, es decir: $x_o = 74,5$.

Por tanto: $\bar{x} = 74,5 + 10\,\dfrac{242}{500} = 79,34$ $\qquad \bar{x} = 79,34$

****

1.7 *El número de crías de 10 familias de una determinada especie, es el siguiente: 5, 4, 0, 8, 3, 6, 2, 9, 3, 7.*
a) Calcular la mediana.
b) Calcular la moda.

S O L U C I O N:

a) Mediana:

Ordenamos en orden creciente los datos:

0, 2, 3, 3, 4, 5, 6, 7, 8, 9.

Como se trata de un número "par" de terminos, la mediana vendrá dada por la media aritmética de los dos términos centrales .

Es decir: $M = \dfrac{4 + 5}{2} = 4,5$

b) Moda:

Obsérvese que se trata de una distribución de variable estadística discreta, por tanto, la moda será el valor que se presenta con mayor frecuencia.

Es decir: $M_o = 3$

Luego la distribución es unimodal.

****

1.8 *El número de pétalos de 13 flores de una determinada especie es el siguiente: 8, 10, 6, 5, 8, 11, 8, 10, 7, 10, 7, 10, 9.*
a) Calcular la mediana.
b) Calcular la moda.
c) Calcular los cuartiles de primero y tercer orden.
d) Calcular el recorrido intercuartílico.

S O L U C I O N:

a) Mediana:

Ordenamos en orden creciente los datos:

5, 6, 7, 7, 8, 8, 8, 9, 10, 10, 10, 10, 11

Como se trata de un número "impar" de términos, la mediana vendrá dada por el término central. Es decir: $M = 8$

b) Moda:

Será el valor que presente mayor frecuencia. Es decir: $M_o = 10$

c) Cuartiles de primero y tercer orden:

- Cuartil primero: Será la mediana de la semiserie de la izquierda en que la mediana divide a la serie de datos dada.
  Es decir: $Q_1 = \frac{7 + 7}{2} = 7$

- Cuartil tercero: Analogamente será la mediana de la otra semiserie. Es decir: $Q_3 = \frac{10 + 10}{2} = 10$

d) Recorrido intercuartílico:

$$r_Q = Q_3 - Q_1 = 10 - 7 = 3.$$

****

1.9 *Considerando el valor teórico del metabolismo basal igual a 100 los valores observados en 50 individuos han dado los siguientes resultados:*

102	98	93	100	98	105	115	110	99	120
115	130	100	86	95	103	105	92	99	134
116	118	88	102	128	99	119	128	110	130
112	114	106	114	100	116	108	113	106	105
120	106	110	100	106	117	109	108	105	106

Calcular agrupando los datos en 10 clases, los siguientes valores:

a) Media aritmética.
b) Mediana.
c) Moda.
d) Cuartiles de primero y tercer orden.
e) Recorrido intercuartílico.

S O L U C I O N:

Como el rango o recorrido es $134 - 86 = 48$, y hemos de construir 10 clases, el tamaño de cada clase será $\frac{48}{10} \simeq 5$

a) Media aritmética:

$$\bar{x} = x_o + c\,\bar{u} = x_o + c\,\frac{\sum f_i u_i}{N} = 107 + 5\frac{17}{50} = 108{,}7$$

b) Mediana:

Según observamos en la columna de frecuencias acumuladas la clase mediana es la quinta, de límites 104,5 - 109,5.

Por tanto:

$$M = L_i + c\left(\frac{\frac{N}{2} - \sum f_i}{f_{mediana}}\right) = 104,5 + 5\frac{25-17}{12} = 107,83$$

clases	marcas	f_i	F_i	u_i	$f_i u_i$
84,5 - 89,5	87	2	2	-4	- 8
89,5 - 94,5	92	2	4	-3	- 6
94,5 - 99,5	97	6	10	-2	-12
99,5 -104,5	102	7	17	-1	- 7
104,5 -109,5	107	12	29	0	0
109,5 -114,5	112	7	36	1	7
114,5 -119,5	117	7	43	2	14
119,5 -124,5	122	2	45	3	6
124,5 -129,5	127	2	47	4	8
129,5 -134,5	132	3	50	5	15
					17

c) <u>Moda:</u>

La clase modal es la que tiene mayor frecuencia. Por tanto, la clase modal también será la quinta de límites 104,5 - 109,5

$$M_o = L_i + c\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) = 104,5 + 5\,\frac{12 - 7}{(12-7)+(12-7)} = 107$$

d) <u>Cuartiles de primero y tercer orden:</u>

- <u>Cuartil de primer orden:</u>

La clase que contiene al cuartil primero será aquella en la que se alcanza la cuarta parte de los individuos de la distribución. Observando la columna de las frecuencias acumuladas, vemos que la clase que contiene al primer cuartil es la cuarta, de límites: 99,5-104,5

Por tanto:

$$Q_1 = L_i + c\left(\frac{\frac{N}{4} - \sum f_1}{f_{cuartil\ 1°}}\right) = 99,5 + 5\,\frac{12,5-10}{7} = 101,28$$

- <u>Cuartil de tercer orden:</u>

Del mismo modo a como hemos hecho para el cálculo del cuartil 1°, tendremos que la clase del cuartil de tercer orden será la séptima, de límites 114,5 - 119,5

Por tanto:

$$Q_3 = L_i + c\left(\frac{\frac{N}{4} - \sum f_3}{f_{cuartil\ 3°}}\right) = 114,5 + 5\frac{37,5-36}{7} = 115,57$$

e) Recorrido intercuartílico:

$$r_Q = Q_3 - Q_1 = 115,57 - 101,28 = 14,29$$

****

1.10 *La distribución por pesos de 70 empleados de un hospital, se expresa en la tabla adjunta:*
Calcular:
a) Media aritmética.
b) Mediana.
c) Moda.

Kgrs.	nº de empleados
50 - 60	8
60 - 70	15
70 - 80	21
80 - 90	18
90 -100	7
100 -110	1

S O L U C I O N:

Formemos en primer lugar la tabla de frecuencias.

clases	marcas	f_i	F_i	u_i	$f_i u_i$
50 - 60	55	8	8	-2	-16
60 - 70	65	15	23	-1	-15
70 - 80	75	21	44	0	0
80 - 90	85	18	62	1	18
90 -100	95	7	69	2	14
100 -110	105	1	70	3	3
					4

a) Media aritmética:

$$\bar{x} = x_o + c\,\bar{u} = x_o + c\,\frac{\sum f_i u_i}{N} = 75 + 10\frac{4}{70} = 75,57.$$

b) Mediana:

La clase mediana será la tercera, de límites: 70 - 80.

$$M = L_i + \left(\frac{\frac{N}{2} - \sum f_i}{f_{mediana}}\right) c = 70 + \left(\frac{35-23}{21}\right)10 = 75,71.$$

c) Moda:

La clase modal será la de mayor frecuencia, es decir la tercera.

$$M_o = L_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) c = 70 + \left(\frac{6}{6+3}\right)10 = 76,66.$$

****

1.11 *Examinadas 150 camadas de 8 ratones, respecto del número de ellos con pelo liso, se obtuvo la siguiente tabla:*

Calcular:

a) Desviación típica.

b) Varianza.

c) Momento de tercer orden respecto al origen.

d) Momento de tercer orden respecto a la media.

e) Coeficiente de asimetría.

n° de ratones con pelo liso en cada camada	n° de camadas
0	0
1	1
2	6
3	13
4	50
5	33
6	30
7	17
8	0
	150

S O L U C I O N:

Construyamos en primer lugar la tabla de cálculo.

x_i	f_i	$f_i x_i$ $\bar{x}$	$f_i x_i^2$	$f_i x_i^3$
0	0	0	0	0
1	1	1	1	11
2	6	12	24	48
3	13	39	117	351
4	50	200	800	3200
5	33	165	825	4125
6	30	180	1080	6480
7	17	119	833	5831
8	0	0	0	0
	150	716 $=\bar{x}$	3680	20036

a) Desviación típica:

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum f_i x_i^2}{N} - \bar{x}^2} = \sqrt{\frac{\sum f_i x_i^2}{N} - \left(\frac{\sum f_i x_i}{N}\right)^2} =$$

$$= \sqrt{\frac{3680}{150} - \left(\frac{716}{150}\right)^2} = \sqrt{1,74} = 1,32.$$

b) Varianza:

$$s^2 = (1,32)^2 = 1,74.$$

c) Momento de tercer orden respecto al origen:

$$\alpha_3 = \frac{\sum f_i x_i^3}{N} = \frac{20036}{150} = 133,57.$$

d) Momento de tercer orden respecto a la media:

$$m_3 = \frac{\sum f_i (x_i - \bar{x})^3}{N} = \frac{\sum f_i x_i^3}{N} - 3\,\bar{x}\,\frac{\sum f_i x_i^2}{N} + 2\,\bar{x}^3 =$$

$$= \frac{20036}{150} - 3\,\frac{716}{150}\cdot\frac{3680}{150} + 2\left(\frac{716}{150}\right)^3 = -0{,}22.$$

e) Coeficiente de asimetría:

$$\gamma_1 = \frac{m_3}{s^3} = \frac{-0{,}22}{(1{,}32)^3} = -0{,}09.$$

Obsérvese, que al ser γ_1 muy pequeño, se trata de una distribución prácticamente insesgada, si bien, al ser negativo, posee un ligero sesgo hacia la izquierda.

****

1.12 *Se ha realizado la determinación del contenido de Ca en la sangre de 25 pacientes. Los resultados han sido los siguientes:*

8,7	9,3	10,1	9,2	9,1	9,3	9,4	8,7	8,8	8,7
9,2	8,3	10,2	9,5	9,6	9,7	9,2	9,3	8,8	9,5
		9,8	9,1	9,2	9,6	8,4			

Calcular:
a) Media aritmética.
b) Moda.
c) Desviación típica.
d) Coeficiente de variación.
e) Sesgo.

S O L U C I O N:

En primer lugar formemos las clases. El rango es 10,2 - 8,3 = 1,9

Construiremos cinco clases de igual tamaño, del siguiente modo:

clases	marcas	f_i	F_i	u_i	$f_i u_i$	$f_i u_i^2$
8,25 - 8,65	8,45	2	2	-2	-4	8
8,65 - 9,05	8,85	5	7	-1	-5	5
9,05 - 9,45	9,25	10	17	0	0	0
9,45 - 9,85	9,65	6	23	1	6	6
9,85 -10,25	10,05	2	25	2	4	8
		25			1	27

a) Media aritmética:

$$\bar{x} = x_o + c\,\bar{u} = x_o + c\,\frac{\sum f_i u_i}{N} = 9{,}25 + 0{,}40\,\frac{1}{25} = 9{,}266.$$

b) Moda:

$$M_o = L_i + c\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) = 9{,}05 + 0{,}40\left(\frac{5}{5+4}\right) = 9{,}272.$$

c) Desviación típica:

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}} = c\sqrt{\frac{\sum f_i (u_i - \bar{u})^2}{N}} = c\sqrt{\frac{\sum f_i u_i^2}{N} - \bar{u}^2} =$$

$$= 0{,}40\sqrt{\frac{27}{25} - \left(\frac{1}{25}\right)^2} = 0{,}3984.$$

d) Coeficiente de variación:

$$d = \frac{s}{\bar{x}} = \frac{0{,}3984}{9{,}2666} = 0{,}04.$$

e) Sesgo:

$$\nu = \frac{\bar{x} - M_o}{s} = \frac{9{,}266 - 9{,}272}{0{,}398} = -0{,}015.$$

Al ser el coeficiente de sesgo muy pequeño y negativo, expresa que la distribución tiene un ligerisimo sesgo hacia la izquierda.

****

1.13 *Se ha determinado el peso de 20 niños normales, al nacer. Obteniéndose los siguientes resultados:*

2,3	3,0	3,1	3,2	3,3	2,6	2,7	3,5	3,5	3,7
4,1	4,4	3,0	3,2	3,3	3,3	3,1	2,8	3,6	3,4

Calcular:

a) Media aritmética.
b) Mediana.
c) Varianza.
d) Momentos de tercero y cuarto orden respecto a la media.
e) Coeficiente de curtosis.

S O L U C I O N:

En primer lugar formemos las clases. El rango es 4,4 - 2,3 = 2,1.

Construiremos seis clases de igual tamaño, del siguiente modo:

clases	f_i	u_i	f_iu_i	$f_iu_i^2$	$f_iu_i^3$	$f_iu_i^4$
2,15 - 2,55	1	-2	-2	4	-8	16
2,55 - 2,95	3	-1	-3	3	-3	3
2,95 - 3,35	9	0	0	0	0	0
3,35 - 3,75	5	1	5	5	5	5
3,75 - 4,15	1	2	2	4	8	16
4,15 - 4,55	1	3	3	9	27	81
	20		5	25	29	121

a) Media aritmética:

$$\bar{x} = x_o + c\,\bar{u} = x_o + c\,\frac{\sum f_iu_i}{N} = 3,15 + 0,40\,\frac{5}{20} = 3,25.$$

b) Mediana:

La clase mediana será la tercera, de límites: 2,95 - 3,35.

$$M = L_i + c\left(\frac{\frac{N}{2} - \sum f_i}{f_{\text{mediana}}}\right) = 2,95 + 0,4\left(\frac{10-4}{9}\right) = 3,21.$$

c) Varianza:

$$s^2 = \frac{\sum f_i(x_i - \bar{x})^2}{N} = c^2\left(\frac{\sum f_i(u_i - \bar{u})^2}{N}\right) = c^2\left(\frac{\sum f_iu_i^2}{N} - \bar{u}^2\right) =$$

$$= 0,4^2\left(\frac{25}{20} - \left(\frac{5}{20}\right)^2\right) = 0,19.$$

d) Momentos de tercero y cuarto orden respecto a la media:

$$m_3 = \frac{\sum f_i(x_i - \bar{x})^3}{N} = c^3\left(\frac{\sum f_i(u_i - \bar{u})^3}{N}\right) =$$

$$= c^3\left(\frac{\sum f_iu_i^3}{N} - 3\,\bar{u}\,\frac{\sum f_iu_i^2}{N} + 2\,\bar{u}^3\right) = 0,4^3\left(\frac{29}{20} - 3\,\frac{5}{20}\,\frac{25}{20} + 2\left(\frac{5}{20}\right)^3\right) =$$

$$= 0,034.$$

- Momento de cuarto orden:

$$m_4 = \frac{\sum f_i(x_i - \bar{x})^4}{N} = c^4\left(\frac{\sum f_i(u_i - \bar{u})^4}{N}\right) =$$

$$= c^4\left(\frac{\sum f_iu_i^4}{N} - 4\,\frac{\sum f_iu_i^3}{N}\,\bar{u} + 6\,\bar{u}^2\,\frac{\sum f_iu_i^2}{N} - 3\,\bar{u}^4\right) =$$

$$= 0{,}4^4\left(\frac{121}{20} - 4\,\frac{29}{20}\,\frac{5}{20} + 6\left(\frac{5}{20}\right)^2\cdot\frac{25}{20} - 3\left(\frac{5}{20}\right)^4\right) = 0{,}13.$$

e) Coeficiente de curtosis:

$$\gamma_2 = \frac{m_4}{s^4} - 3 = \frac{0{,}13}{0{,}19^2} - 3 = 0{,}60.$$

Al ser $\gamma_2 > 0$, se trata de una distribución leptocúrtica.

****

1.14 *Se ha estudiado la distribución de albúmina total circulante (en gramos) de 50 hombres normales comprendidos entre los 20 y los 30 años, obteniéndose los siguientes resultados:*

albúmina circulante (grs.)	*nº de hombres*
99,5 - 109,5	5
109,5 - 119,5	10
119,5 - 129,5	12
129,5 - 139,5	11
139,5 - 149,5	8
149,5 - 159,5	4

Calcular:
a) La media aritmética.
b) La mediana.
c) La moda.
d) Desviación típica.
e) El sesgo.
f) La curtosis.

S O L U C I O N:

Formemos la correspondiente tabla de cálculo:

clases	f_i	u_i	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
99,5 - 109,5	5	-2	-10	20	-40	80
109,5 - 119,5	10	-1	-10	10	-10	10
119,5 - 129,5	12	0	0	0	0	0
129,5 - 139,5	11	1	11	11	11	11
139,5 - 149,5	8	2	16	32	64	128
149,5 - 159,5	4	3	12	36	108	324
	50		19	109	133	553

a) Media aritmética:

$$\bar{x} = x_o + c\,\bar{u} = x_o + c\,\frac{\sum f_i u_i}{N} = 124{,}5 + 10\,\frac{19}{50} = 128{,}3.$$

b) Mediana:

$$M = L_i + c\left(\frac{\frac{N}{2} - \sum f_i}{f_{mediana}}\right) = 119{,}5 + 10\left(\frac{25 - 15}{12}\right) = 127{,}83$$

c) Moda:

$$M_o = L_i + c\left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) = 119,5 + 10\ \frac{2}{2+1} = 126,16.$$

d) Desviación típica:

$$s = c\sqrt{\frac{\sum f_i u_i^2}{N} - \bar{u}^2} = 10\sqrt{\frac{109}{50} - \left(\frac{19}{50}\right)^2} = 14,2$$

e) Sesgo:

$$\nu = \frac{\bar{x} - M_o}{s} = \frac{128,3 - 126,16}{14,2} = 0,15$$

Al ser el coeficiente de sesgo pequeño y positivo, se trata de una distribución ligeramente sesgada hacia la derecha.

f) Curtosis:

Calculemos en primer lugar el momento de cuarto orden respecto a la media.

$$m_4 = \frac{\sum f_i (x_i - \bar{x})^4}{N} = c^4\left(\frac{\sum f_i (u_i - \bar{u})^4}{N}\right) =$$

$$= c^4\left(\frac{\sum f_i u_i^4}{N} - 4\,\bar{u}\,\frac{\sum f_i u_i^3}{N} + 6\,\bar{u}^2\,\frac{\sum f_i u_i^2}{N} - 3\,\bar{u}^4\right) =$$

$$= 10^4\left(\frac{553}{50} - 4\cdot\frac{19}{50}\cdot\frac{133}{50} + 6\left(\frac{19}{50}\right)^2\cdot\frac{109}{50} - 3\left(\frac{19}{50}\right)^4\right) = 88400.$$

Por tanto el coeficiente de curtosis será:

$$\gamma_2 = \frac{m_4}{s^4} - 3 = \frac{88400}{(14,2)^4} - 3 = -0,83$$

Luego se trata de una distribución platicúrtica.

****

1.15 *Para establecer la importancia del pinzamiento precoz o tardío del cordón umbilical sobre el volumen sanguíneo del recien nacido, se ha estudiado los volumenes sanguíneos (expresados en porcentajes del peso corporal) de 20 niños a los que se pinzó el cordón precozmente.*

Se obtuvieron los siguientes resultados:

12,5	9,6	8,3	8,6	9,4	10,2	11,3	10,7	8,2	9,1
9,7	10,3	11,2	9,8	9,3	10,7	11,8	9,7	9,6	9,4

Se desea calcular:

a) *Media aritmética.*
b) *Mediana.*
c) *Moda.*
d) *Varianza.*
e) *Coeficiente de variación.*
f) *Sesgo.*

S O L U C I O N:

En primer lugar, formemos las clases. El rango es 12,5-8,2 = 4,3 Construimos nueve clases de igual **tamaño** del siguiente modo:

clases	f_i	u_i	$f_i u_i$	$f_i u_i^2$
8 - 8,5	2	-3	-6	18
8,5 - 9	1	-2	-2	4
9 - 9,5	4	-1	-4	4
9,5 -10	5	0	0	0
10 -10,5	2	1	2	2
10,5 -11	2	2	4	8
11 -11,5	2	3	6	18
11,5 -12	1	4	4	16
12 -12,5	1	5	5	25
	20		9	95

a) Media aritmética:

$$\bar{x} = x_o + c\,\bar{u} = x_o + c\,\frac{\sum f_i u_i}{N} = 9,75 + 0,5\,\frac{9}{20} = 9,97$$

b) Mediana:

$$M = L_i + c\left(\frac{\frac{N}{2} - \sum f_i}{f_{mediana}}\right) = 9,5 + 0,5\left(\frac{10-7}{5}\right) = 9,8.$$

c) Moda:

$$M_o = L_i + c\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) = 9,5 + 0,5\left(\frac{1}{1+3}\right) = 9,62.$$

d) Varianza:

$$s^2 = c^2\left(\frac{\sum f_i (u_i - \bar{u})^2}{N}\right) = c^2\left(\frac{\sum f_i u_i^2}{N} - \bar{u}^2\right) =$$

$$= 0,5^2\left(\frac{95}{20} - \left(\frac{9}{20}\right)^2\right) = 1,13$$

e) Coeficiente de variación:

$$d = \frac{s}{\overline{x}} = \frac{1,06}{9,97} = 0,1$$

d) Sesgo:

$$\nu = \frac{\overline{x} - M_o}{s} = \frac{9,97 - 9,62}{1,06} = 0,33$$

Luego se trata de una distribución sesgada hacia la derecha.

****

2 - probabilidades

1. ALGEBRA DE SUCESOS

- 1.1 Experimento aleatorio
- 1.2 Espacio muestral
- 1.3 Suceso
  - 1.3.1 Sucesos elementales
  - 1.3.2 Sucesos compuestos
  - 1.3.3 Suceso seguro
  - 1.3.4 Suceso contrario
  - 1.3.5 Suceso imposible
- 1.4 Inclusión de sucesos
- 1.5 Igualdad de sucesos
- 1.6 Operaciones con sucesos
  - 1.6.1 Unión de sucesos
  - 1.6.2 Intersección de sucesos
  - 1.6.3 Diferencia de sucesos
- 1.7 Sistema completo de sucesos
- 1.8 Algebra de Boole de los sucesos aleatorios
  - 1.8.1 Consecuencias del Algebra de Boole

2. FRECUENCIA Y PROBABILIDAD

- 2.1 Frecuencia absoluta de un suceso
- 2.2 Frecuencia relativa de un suceso
- 2.3 Propiedades de las frecuencias
- 2.4 Definición clásica de probabilidad
- 2.5 Definición axiomática de probabilidad
- 2.6 Consecuencias de los axiomas
- 2.7 Espacio probabilístico

3. PROBABILIDAD CONDICIONADA

- 3.1 Probabilidad condicionada
- 3.2 Sucesos dependientes e independientes
- 3.3 Teorema de la probabilidad compuesta
- 3.4 Aplicación a los esquemas de contagio y campañas de seguridad
- 3.5 Teorema de las probabilidades totales
- 3.6 Teorema de Bayes

2- PROBABILIDADES

1. ALGEBRA DE SUCESOS

1.1 EXPERIMENTO ALEATORIO.- Se dice que un experimento es *aleatorio* si se cumplen las siguientes condiciones:

1.) El experimento se puede repetir indefinidamente bajo análogas condiciones.

2.) En cada prueba del experimento, se obtiene un resultado que pertenece al conjunto de resultados posibles del experimento.

3.) Antes de realizar una nueva prueba del experimento, no se puede predecir el resultado que se obtendrá.

Esta última condición se conoce como *condición de aleatoriedad* o *condición de azar*.

1.2 ESPACIO MUESTRAL.- Llamaremos *espacio muestral* al conjunto de todos los resultados posibles del experimento aleatorio. A este conjunto lo representaremos en lo que sigue por E.

Ej: El espacio muestral del experimento aleatorio que consiste en considerar el sexo del próximo hijo de una familia, será: $E = \{ V, H \}$

Representando por V el sexo varón y por H el sexo hembra.

1.3 SUCESO.- Llamaremos *suceso aleatorio* o *estocástico*, a cada uno de los elementos del conjunto de partes de E, es decir, serán sucesos los elementos de $\mathcal{P}(E)$.

Ej: Para el ejemplo anterior, tendremos que:

$$E = \{V, H\} \Rightarrow \mathcal{P}(E) = \{ \phi , \{V\} , \{H\} , \{V, H\} \}$$

A cada uno de los elementos de $\mathcal{P}(E)$, lo llamaremos suceso aleatorio.

1.3.1 SUCESOS ELEMENTALES.- Se llaman así a cada uno de los elementos del espacio muestral E.
Es decir, en el ejemplo que nos ocupa son sucesos elementales $\{V\}$ y $\{H\}$.

1.3.2 SUCESOS COMPUESTOS.- Se dice que un suceso es compuesto cuando está formado por dos o más sucesos elementales.

1.3.3 SUCESO SEGURO.- Se llama así al suceso formado por todos los resultados posibles. Es decir, que suceso seguro equivale a espacio muestral.
Obsérvese que E es un elemento de $\mathcal{P}(E)$.
En nuestro caso el suceso seguro será: $E = \{ V, H \}$
Al suceso seguro también se le llama *suceso cierto* o *suceso universal*.

1.3.4 SUCESO CONTRARIO.- Llamaremos *suceso contrario* del suceso A y representaremos por $\overline{A}$, a un suceso que se presenta si y sólo si no se presenta A y reciprocamente.

1.3.5 SUCESO IMPOSIBLE.- Llamaremos *suceso imposible*, al suceso contrario del suceso seguro. Generalmente lo representaremos por la letra ϕ.
Por tanto: $\phi = \overline{E}$
Obsérvese que como el suceso seguro se presenta siempre, el suceso imposible no se presenta nunca.

1.4 INCLUSION DE SUCESOS.- Se dice que un suceso A está *contenido* o *incluido* en otro B y escribiremos $A \subset B$, si todos los sucesos elementales de A lo son de B.

1.5 IGUALDAD DE SUCESOS.- Decimos que dos sucesos A y B son *iguales* y escribiremos $A = B$ si se verifica que $A \subset B$ y $B \subset A$.

1.6 OPERACIONES CON SUCESOS

1.6.1 UNION DE SUCESOS.- Dados dos sucesos A y B llamaremos suceso unión de A y B y escribiremos $A \cup B$, al suceso que está formado por todos los sucesos elementales que pertenecen a A o a B o a ambos. S. disjuntos: $A \cup B = A + B$
no " $A \cup B = A + B - (A \cap B)$

1.6.2 INTERSECCION DE SUCESOS.- Dados dos sucesos A y B, llamaremos suceso intersección de A y B y escribiremos $A \cap B$, al suceso que está formado por los sucesos elementales que pertenecen simultaneamente a los dos sucesos dados.

Ej: Consideremos el experimento que consiste en lanzar un dado y anotar la cara que queda al posar sobre el suelo.

El espacio muestral o suceso seguro será:

$$E = \{1, 2, 3, 4, 5, 6\}$$

Consideremos los sucesos $A = \{1, 2\}$ y $B = \{2, 4, 6\}$

Los sucesos unión e intersección serán:

$A \cup B = \{1,2,4,6\}$ $\qquad$ $A \cap B = \{2\}$

Consideremos ahora, los sucesos $M = \{1,3\}$ y $N = \{2,5,6\}$

Los sucesos unión e intersección serán:

$M \cup N = \{1,2,3,5,6\}$ $\qquad$ $M \cap N = \phi$

Es decir, que los suceos M y N no se pueden presentar al mismo tiempo.

Definición: Se dice que dos sucesos M y N son *incompatibles* o *mutuamente excluyentes*, cuando su intersección es el suceso imposible.

Es decir: M y N incompatibles $\Longleftrightarrow$ $M \cap N = \phi$

1.6.3 DIFERENCIA DE SUCESOS.- Dados dos sucesos A y B, llamaremos suceso diferencia de A y B y escribiremos $A - B$, al suceso que está formado por todos los sucesos elementales de A que no son de B.

Obsérvese, que como consecuencia de la definición de diferencia se tiene: $A - B = A \cap \overline{B}$

1.7 SISTEMA COMPLETO DE SUCESOS.- Dado un conjunto de sucesos A_1, A_2, A_3, ... , A_n, se dice que constituyen un sistema completo de sucesos, si se verifican las siguientes condiciones:

1ª) $A_1 \cup A_2 \cup A_3 \cup \ . \ . \ . \ \cup A_n = E = \bigcup_{i=1}^{i=n} A_i$

2ª) $A_i \cap A_j = \phi$; siendo $i \neq j$

1.8 ALGEBRA DE BOOLE DE LOS SUCESOS ALEATORIOS.- Consideremos un determinado experimento aleatorio y llamemos E al espacio muestral o conjunto de resultados posibles. Formemos el conjunto de partes de E que representaremos por $\mathcal{A} = \mathcal{P}(E)$.

Dentro de este conjunto $\mathcal{A}$, definimos las operaciones internas de unión, intersección y complementación (que en este caso llamaremos contrario), respecto de las cuales se cumplen las siguientes propiedades:

1.- Conmutativa: $A \cup B = B \cup A$ $\qquad A \cap B = B \cap A$

2.- Asociativa: $(A \cup B) \cup C = A \cup (B \cup C)$ $\qquad (A \cap B) \cap C = A \cap (B \cap C)$

3.- Idempotente: $A \cup A = A$ $\qquad A \cap A = A$

4.- Simplificativa: $A \cup (B \cap A) = A$ $\qquad A \cap (B \cup A) = A$

Hasta aqui vemos que el conjunto $\mathcal{A}$, con las operaciones de unión e intersección constituye un *retículo*, veamos que además es distributivo y complementario.

5.- Distributiva: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

6.- Complementación: $\forall A \in \mathbf{A}$, $\exists \overline{A} \in \mathbf{A}$, $A \cup \overline{A} = E$ y $A \cap \overline{A} = \phi$

Por tanto, el conjunto $\mathbf{A}$, con las operaciones de unión, intersección y complementación, es un retículo distributivo y complementario que llamaremos *Algebra de Boole de los sucesos estocásticos o aleatorios*.

1.8.1 CONSECUENCIAS DEL ALGEBRA DE BOOLE.- Las consecuencias más importantes son:

1ª) Existe un elemento ínfimo: El suceso imposible ϕ, tal que: $A \cup \phi = A$ y $A \cap \phi = \phi$

2ª) Existe un elemento universal: El suceso seguro o cierto E, tal que: $A \cup E = E$ y $A \cap E = A$

3ª) Leyes de Morgan: Para dos sucesos cualesquiera del algebra, se verifica: $\overline{A \cap B} = \overline{A} \cup \overline{B}$ y $\overline{A \cup B} = \overline{A} \cap \overline{B}$.

2. FRECUENCIA Y PROBABILIDAD

2.1 FRECUENCIA ABSOLUTA DE UN SUCESO.- Supongamos que realizamos un determinado experimento aleatorio n veces y sea A un suceso de ese experimento, tal que de las n pruebas realizadas, se haya presentado el suceso A, n' veces.

Llamaremos *frecuencia absoluta* del suceso A y representaremos por f(A) al número n'.

Es decir: $f(A) = n'$

Es inmediato observar que $n' \leq n$

2.2 FRECUENCIA RELATIVA DE UN SUCESO.- Llamaremos *frecuencia relativa* del suceso A y representaremos por h(A), a la razón entre la frecuencia absoluta del suceso A y el número total de pruebas realizadas.

Es decir: $h(A) = \dfrac{f(A)}{n} = \dfrac{n'}{n}$

2.3 PROPIEDADES DE LAS FRECUENCIAS.-

1ª) "Cualquiera que sea el suceso A, se tiene que la frecuencia relativa del suceso A, es un número racional acotado entre 0 y 1"

Es decir: $0 \leq h(A) \leq 1$

- En efecto, si realizamos n pruebas y en ellas aparece n' veces el suceso A, se verificará:

$$0 \leq n' \leq n$$

y dividiendo por n, tendremos: $0 \leq \dfrac{n'}{n} \leq 1$

Es decir : $0 \leqslant h(A) \leqslant 1$

2ª) "Si A y B son dos sucesos incompatibles, (es decir, $A \cap B = \emptyset$), la frecuencia relativa de la unión de los dos sucesos es igual a la suma de las frecuencias relativas de cada uno de ellos"

Es decir: $h(A \cup B) = h(A) + h(B)$

- En efecto, si realizamos n pruebas y en ellas obtenemos respectivamente, n' veces el suceso $A \cup B$, n'' veces el suceso A y n''' veces el suceso B, sabemos que por definición de unión de sucesos, tendremos que siempre que se realice el suceso $A \cup B$, es que se ha realizado el suceso A o el B, pero solo uno de los dos, por ser ambos incompatibles.

Por tanto: $n' = n'' + n'''$

y dividiendo los dos miembros de la igualdad anterior por n resulta: $\frac{n'}{n} = \frac{n''}{n} + \frac{n'''}{n}$

Es decir: $h(A \cup B) = h(A) + h(B)$

3ª) "La frecuencia relativa del suceso cierto o seguro E es igual a la unidad"

Es decir: $h(E) = 1$

- En efecto, ya que si efectuamos n pruebas, está claro que se habrá presentado n veces el suceso E y por tanto

$$h(E) = 1$$

4ª) "Si un cierto experimento lo repetimos bajo análogas condiciones un número n de veces, se observa que la frecuencia relativa de un suceso determinado tiende experimentalmente a aproximarse a un valor fijo, es decir a estabilizarse.

2.4 DEFINICION CLASICA DE PROBABILIDAD.- La probabilidad de un suceso A, viene definida por el cociente entre el número de casos favorables y el número de casos posibles.

Es decir: $$p(A) = \frac{\text{número de casos favorables al suceso A}}{\text{número total de casos posibles}}$$

Esta definición fué enunciada por Laplace en el siglo XVII y supone que todos los sucesos elementales de un mismo experimento aleatorio han de tener la misma probabilidad.

2.5 DEFINICION AXIOMATICA DE PROBABILIDAD.- Llamamos probabilidad, a una aplicación del Algebra de Boole de los sucesos aleatorios $\mathcal{A}$, en el cuerpo de los números reales $\mathbb{R}$. De manera que a cada suceso A de $\mathcal{A}$, se le asocia un número real que llamamos probabilidad de A y representamos por p(A) y tal que verifica los siguientes axiomas:

1.- $\forall A \in \mathcal{A} \implies 0 \leq p(A) \leq 1$

2.- Si $A_1, A_2, \ldots, A_n$ son incompatibles, se verifica:

$$p(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} p(A_i) = p(A_1)+p(A_2)+\ldots+p(A_n)$$

3.- $p(E) = 1$

Este grupo de axiomas, se conocen con el nombre de CRAMER-KOLMOGOROV.

2.6 CONSECUENCIAS DE LOS AXIOMAS.-

1ª).- La probabilidad del suceso contrario $\overline{A}$, es igual a 1 menos la probabilidad del suceso A.

Es decir: $p(\overline{A}) = 1 - p(A)$

- En efecto, pues por definición de suceso contrario tenemos:

$$A \cup \overline{A} = E \quad \text{y} \quad A \cap \overline{A} = \phi$$

aplicando el axioma 3.- de probabilidad a los dos miembros de la primera igualdad, se tiene: $P(E) = p(A \cup \overline{A})$

Es decir: $1 = p(A) + p(\overline{A})$ de donde $p(\overline{A}) = 1 - p(A)$

2ª).- La probabilidad del suceso imposible es nula. $p(\phi) = 0$

Como el suceso imposible ϕ, es el contrario del suceso cierto E, por la consecuencia anterior resulta:

$$p(\phi) = 1 - p(E) = 1 - 1 = 0$$

3ª).- Sean A y B dos sucesos no incompatibles, es decir tales que $A \cap B \neq \phi$. Entonces, se tiene:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

- En efecto, ya que $A \cup B = A \cup (B \cap \overline{A})$
de tal forma que A y $(B \cap \overline{A})$ son incompatibles y por tanto podemos aplicar el axioma 3.-

$$p(A \cup B) = p\left[A \cup (B \cap \overline{A})\right] = p(A) + p(B \cap \overline{A}) \qquad [1]$$

Ahora bien, como: $B = (B \cap A) \cup (B \cap \overline{A})$, siendo $B \cap A$ y $B \cap \overline{A}$ incompatibles, también podemos aplicar el axioma 3.-

$$p(B) = p\left[(B \cap A) \cup (B \cap \overline{A})\right] = p(B \cap A) + p(B \cap \overline{A})$$

de donde: $\quad p(B \cap \overline{A}) = p(B) - p(B \cap A) \qquad [2]$

Sustituyendo [2] en [1], resulta:

$$p(A \cup B) = p(A) + p(B \cap \overline{A}) = p(A) + p(B) - p(A \cap B)$$

como queriamos demostrar.

4ª).- Sean A y B dos sucesos tales que $A \subset B$, entonces $p(A) \leq p(B)$

- En efecto, pues:

$$\text{Si } A \subset B \implies \exists C, \text{ tal que } \begin{cases} A \cup C = B \\ A \cap C = \phi \end{cases}$$

Como A y C son incompatibles podemos aplicar el axioma 2.-

$$p(A \cup C) = p(B), \qquad p(A) + p(C) = p(B)$$

Ahora bien, por el axioma 1.-, sabemos que para cualquier suceso C, se verifica que $p(C) \geq 0$,

Por tanto, se tiene: $\quad p(A) \leq p(B)$

2.7 ESPACIO PROBABILISTICO.- Dado un experimento aleatorio cualquiera, llamaremos espacio probabilístico asociado a dicho experimento a la terna (E, A, p), donde E es el espacio muestral del experimento, A es el algebra de Boole de los sucesos aleatorios y p es la probabilidad que acabamos de definir sobre A.

3. PROBABILIDAD CONDICIONADA

3.1 PROBABILIDAD CONDICIONADA.- Supongamos el siguiente ejemplo:
De 5000 individuos examinados en un determinado país se ha obser-

vado que 4.800 son de ojos claros y 200 de ojos oscuros, por otra parte hay 1.275 rubios y 3.725 castaños, lo que podemos expresar mediante la siguiente tabla:

		pelo		
		RUBIOS A	CASTAÑOS $\bar{A}$	
ojos	CLAROS B	1.225	3.575	4.800
	OSCUROS $\bar{B}$	50	150	200
		1.275	3.725	5.000

De esta tabla se obtiene fácilmente que la frecuencia relativa de rubios, será: $h(A) = 1275/5000$, que será distinta de la frecuencia relativa de rubios entre los individuos de ojos claros, que llamaremos frecuencia relativa de A condicionada a B, y escribiremos $h(A/B) = \frac{1225}{4800}$. Análogamente $h(B/A) = \frac{1225}{1275}$, $h(A \cap B) = \frac{1225}{5000}$

A partir de los valores obtenidos observamos que se verifica la siguiente relación: $\frac{1225}{5000} = \frac{1275}{5000} \cdot \frac{1225}{1275}$

Es decir: $h(A \cap B) = h(A) . h(B/A)$

Generalizando esta expresión obtenemos el concepto de probabilidad condicionada.

Definición: Consideremos un espacio probabilístico $(E, \mathcal{A}, p)$ y sea A un suceso del algebra $\mathcal{A}$, tal que $p(A) \neq 0$. Llamaremos *probabilidad condicionada* del suceso B respecto del suceso A y lo denotaremos por $p(B/A)$, al cociente:

$$\boxed{p(B/A) = \frac{p(A \cap B)}{p(A)}} \qquad \text{supuesto} \quad p(A) \neq 0$$

De aqui resulta: $P(A \cap B) = p(A).P(B/A)$

Del mismo modo, se define:

$$p(A/B) = \frac{p(A \cap B)}{p(B)} \qquad \text{supuesto} \quad p(B) \neq 0$$

Para tres sucesos, será: $p(A\cap B\cap C) = p(A).p(B/A).p(C/A\cap B)$

Y en general, para n sucesos será:

$$p(A_1\cap A_2\cap \ldots \cap A_n)=p(A_1).p(A_2/A_1)\ldots.p(A_n/A_1\cap A_2\cap\ldots\cap A_{n-1})$$

3.2 SUCESOS DEPENDIENTES E INDEPENDIENTES.- En general, tenemos que $p(B/A) \neq p(B)$, si así se verifica diremos que el suceso B, depende del suceso A.

Ahora bien, si $p(B/A) = p(B)$, diremos que el suceso B es independiente del suceso A.

En el caso de que B sea independiente de A quedará:

$$p(B/A) = p(B) = \frac{p(A\cap B)}{p(A)} \Longrightarrow \qquad p(A\cap B) = p(A).p(B)$$

Sustituyendo en p(A/B) resulta:

$$p(A/B) = \frac{p(A\cap B)}{p(B)} = \frac{p(A).P(B)}{p(B)} = p(A)$$

Por tanto, si B es independiente de A, también se verifica que A es independiente de B.

Definición: Se dice que los sucesos A y B son independientes, si se verifica: $p(A \cap B) = p(A).p(B)$.

3.3 TEOREMA DE LA PROBABILIDAD COMPUESTA.- Sean $A_1, A_2, \ldots, A_n$ n sucesos cualesquiera de un mismo experimento y tales que la probabilidad de la realización simultanea de los n sucesos es no nula, entonces se tiene:

$$p(A_1\cap A_2\cap \ldots \cap A_n) = p(A_1).p(A_2/A_1)\ldots.p(A_n/A_1\cap A_2\cap \ldots\cap A_{n-1})$$

Demostración: En efecto, ya que a partir de la definición de probabilidad condicionada, tenemos que:

$$p(A_1 \cap A_2) = p(A_1).P(A_2/A_1) \qquad \text{supuesto } p(A_1) \neq 0$$

Aplicando el principio de inducción, supondremos cierta la expresión anterior para n-1 y a partir de ella demostraremos para n

$$p(A_1\cap A_2\cap \ldots \cap A_n) = p(A_1\cap A_2\cap \ldots\cap A_{n-1}).p(A_n/A_1\cap A_2 \ldots\cap A_{n-1})$$

$$= p(A_1).p(A_2/A_1).p(A_3/A_1\cap A_2)\ldots.p(A_n/A_1\cap A_2\cap \ldots\cap A_{n-1}).$$

3.4 APLICACION A LAS CAMPAÑAS DE SEGURIDAD Y ESQUEMAS DE CONTAGIO.

Veamos un ejemplo de aplicación del teorema de la probabilidad compuesta a una situación frecuente en la vida humana, como es una epidemia.

Consideremos a la población humana compuesta por individuos sanos y por individuos afectados por la epidemia, como bolas de una gran urna que para distinguir, representaremos por bolas blancas a los individuos sanos y por bolas negras a los individuos enfermos. Sean b bolas blancas y n bolas negras.

Supongamos que cada cierto tiempo se extrae una bola de la urna y se devuelve a la urna con otras c del mismo color y d del color opuesto. Entonces la probabilidad de que aparezca una bola negra en cada una de las tres primeras extracciones la obtendremos aplicando el teorema de la probabilidad compuesta. Es decir:

$$p(N_1 \cap N_2 \cap N_3) = p(N_1).p(N_2/N_1).p(N_3/N_1 \cap N_2) =$$

$$= \frac{n}{b+n} \cdot \frac{n+c}{b+n+c+d} \cdot \frac{n+2c}{b+n+2c+2d}$$

Caso 1°: Si $c > 0$ y $d = 0$, observamos que después de cada extracción se aumenta la probabilidad de que la bola extraida en la próxima prueba sea del mismo color.

Esta situación es análoga a cuando se produce el contagio de una epidemia, ya que a medida que se registran nuevos casos de contagio va aumentando la probabilidad de quedar afectado por la epidemia. Esta situación fué estudiada por Polya bajo el nombre genérico de *esquema de contagio*.

Caso 2°: Si $c = 0$ y $d > 0$, tenemos el caso opuesto, en el que después de cada extracción disminuye la probabilidad de que la bola extraida sea del mismo color. Esto en la práctica es debido a que se toman medidas para la disminución de esa probabilidad, razon por la que Friedman denominó *campaña de seguridad*.

3.5 TEOREMA DE LAS PROBABILIDADES TOTALES.- Sean A_1, A_2, . . . ,A_n un sistema completo de sucesos, tales que $p(A_i) > 0$, $(1 \leq i \leq n)$ y sea B un suceso cualquiera para el que se conocen las probabilidades $p(B/A_i)$, entonces la probabilidad del suceso B viene dada por la siguiente expresión:

$$p(B) = p(A_1).p(B/A_1)+p(A_2).p(B/A_2)+...+p(A_n).p(B/A_n)$$

Es decir:

$$\boxed{p(B) = \sum p(A_i).p(B/A_i)}$$

Demostracción: Como $A_1,A_2,...,A_n$, constituyen un sistema completo de sucesos se tiene:

$$\left.\begin{array}{l} \bigcup A_i = E \\ A_i \cap A_j = \phi \quad (i \neq j) \end{array}\right\}$$

Entonces $p(B) = p(B \cap E) = p\left[B \cap (A_1 \cup A_2 \cup \ ... \cup A_n)\right] =$
$= p(B \cap A_1) + p(B \cap A_2) + \ . \ . \ . + p(B \cap A_n) = \sum p(B \cap A_i) =$
$= \sum p(A_i).p(B/A_i)$. c.q.d.

3.6 TEOREMA DE BAYES.- Sean A_1,A_2,A_3, ... ,A_n un sistema completo de sucesos, tales que $p(A_i) > 0$ y sea B un suceso cualquiera para el que se conocen las probabilidades condicionadas $p(B/A_i)$ que llamaremos *verosimilitudes*.

En estas condiciones, el teorema de Bayes establece que las probabilidades $p(A_i/B)$, vienen dadas por la siguiente expresión:

$$\boxed{p(A_i/B) = \frac{p(A_i).p(B/A_i)}{\sum_{i=1}^{n} p(A_i).p(B/A_i)}}$$

A las probabilidades $p(A_i)$ se las llama *probabilidades a priori*, y a las probabilidades $p(A_i/B)$, se las llama *probabilidades a posteriori*.

Demostracción: Por definición de probabilidad condicionada, se verifica:

$$p(A_i \cap B) = p(A_i).p(B/A_i) = p(B).p(A_i/B), \text{ de donde:}$$

$$p(A_i/B) = \frac{p(A_i).p(B/A_i)}{P(B)} \qquad [1]$$

Ahora bien, como p(B) por el teorema anterior, es igual a la siguiente expresión:

$$p(B) = \sum p(A_i).p(B/A_i) \qquad [2]$$

Sustituyendo [2] en [1], resulta:

$$p(A_i/B) = \frac{p(A_i).p(B/A_i)}{\sum p(A_i).p(B/A_i)} \qquad \text{c.q.d.}$$

****

PROBLEMAS RESUELTOS

2.1 *Probar que si los sucesos A y B son independientes, también lo son los sucesos siguientes:*
1°) A y $\overline{B}$ 2°) $\overline{A}$ y $\overline{B}$ 3°) $\overline{A}$ y B

SOLUCION:

Sea E el suceso cierto, por tanto A, $\overline{A}$, B, $\overline{B}$ están contenidos en E.

1°) Si A y B son independientes, se verifica: $p(A\cap B) = p(A).p(B)$

Para probar que A y $\overline{B}$ son independientes, hemos de ver que:

$$p(A \cap \overline{B}) = p(A).p(\overline{B})$$

Ahora bien, como: $A \cap \overline{B} = A \cap (E - B) = (A \cap E) - (A \cap B) = A - (A \cap B)$

Tomando probabilidades en los dos miembros, resulta:

$$p(A \cap \overline{B}) = p\left[A - (A \cap B)\right] = p(A) - p(A \cap B) = p(A) - p(A).p(B) =$$
$$= p(A)\left[1 - p(B)\right] = p(A).p(\overline{B}).$$

2°) Veamos que $\overline{A}$ y $\overline{B}$ son independientes, es decir, hemos de ver:

$$p(\overline{A} \cap \overline{B}) = p(\overline{A}).p(\overline{B})$$

Haciendo uso de las leyes de Morgan, tenemos que: $\overline{A} \cap \overline{B} = \overline{(A \cup B)}$

$$p(\overline{A} \cap \overline{B}) = p(\overline{A \cup B}) = 1 - p(A \cup B) = 1 - p(A) - p(B) + p(A \cap B)$$
$$= 1 - p(A) - p(B) + p(A).p(B) = 1 - p(A)\left[1 - p(B)\right] - p(B) =$$
$$= 1 - p(B) - p(A).p(\overline{B}) = p(\overline{B}) - p(A).p(\overline{B}) =$$
$$= p(\overline{B})\left[1 - p(A)\right] = p(\overline{B}).p(\overline{A}) = p(\overline{A}).p(\overline{B})$$

3°) Veamos que $\overline{A}$ y B son independientes, es decir

$$p(\overline{A} \cap B) = p(\overline{A}).P(B)$$

$$\overline{A} \cap B = (E - A) \cap B = (E \cap B) - (A \cap B) = B - (A \cap B).$$

$$p(\overline{A} \cap B) = p\left[B - (A \cap B)\right] = p(B) - p(A \cap B) = p(B) - p(A).p(B) =$$
$$= p(B)\left[1 - p(A)\right] = p(B).p(\overline{A}) = p(\overline{A}).p(B)$$

****

2.2 *Si la probabilidad de que una persona sea rubia es 0,4 y la probabilidad de que tenga los ojos negros es 0,3, calcular las siguientes probabilidades:*

1°) Que sea rubia y tenga los ojos negros.
2°) Que sea rubia o tenga los ojos negros.
3°) Que sean tres personas rubias.
4°) Que dos personas sean rubias o tengan los ojos negros.

S O L U C I O N:

1°) Que sea rubia y tenga los ojos negros.

Si representamos por R el suceso ser rubio y por N el suceso tener los ojos negros, será el suceso $R \cap N$.

$p(R \cap N) = p(R).p(N) = 0,4.0,3 = 0,12.$

2°) Que sea rubia o tenga los ojos negros.

En este caso se trata del suceso unión $R \cup N$. Por tanto:

$p(R \cup N) = P(R) + p(N) - p(R \cap N) = 0,4 + 0,3 - 0,12 = 0,58.$

3°) Que sean las tres personas rubias.

Se trata del suceso $R \cap R \cap R$. Por tanto:

$p(R\cap R\cap R) = p(R).p(R).p(R) = 0{,}4.0{,}4.0{,}4 = (0{,}4)^3 = 0{,}064.$

4°) Que dos personas sean rubias o tengan los ojos negros.

El suceso que dos personas sean rubias será $R\cap R$.

El suceso que dos personas tengan los ojos negros será $N\cap N$

Por tanto se trata del suceso $(R\cap R)\cup(N\cap N)$.

$$p\left[(R \cap R) \cup (N \cap N)\right] = p(R \cap R) + p(N \cap N) - p(R \cap R \cap N \cap N) =$$
$$= (0{,}4)^2 + (0{,}3)^2 - (0{,}4)^2.(0{,}3)^2 = 0{,}2356$$

****

2.3 *La compañía farmaceutica A suministró 300 unidades de un medicamento de las cuales 10 eran defectuosas; la compañía B entregó 100 unidades de las que había 20 defectuosas y la compañía C entregó 200 unidades de las que 25 eran defectuosas.*

Se almacenaron todas las unidades de forma que se mezclaron aleatoriamente. Calcular:

1°) Probabilidad de que una unidad tomada al azar sea de la compañía A.
2°) Probabilidad de que sea de la compañía B.
3°) Probabilidad que sea de C y defectuosa.
4°) Probabilidad que sea de A y buena.
5°) Probabilidad de que sea buena.
6°) Probabilidad de que sea defectuosa.
7°) Si resultó ser defectuoso, ¿cuál es la probabilidad de que sea de la compañía C?.
8°) Si es buena ¿cuál es la probabilidad de que sea de la compañía B?
9°) Si es de la compañía A ¿cuál es la probabilidad de que sea defectuosa?.
10°) Si es de la compañía B ¿cuál es la probabilidad de que sea buena?.

S O L U C I O N:

Representemos por A al suceso que la unidad elegida al azar sea de la compañía A. Análogamente representaremos por B y C que la unidad elegida al azar pertenezca a las compañias B y C respectivamente. Por D representaremos al suceso elegir unidad defectuosa y por $\overline{D}$ al suceso elegir unidad buena.

1°) Probabilidad que la unidad elegida sea de A.

$$p(A) = \frac{300}{600} = 0,5.$$

2°) Probabilidad que la unidad elegida sea de B.

$$p(B) = \frac{100}{600} = 0,166.$$

3°) Probabilidad de que la unidad elegida sea de C y defectuosa.

$$p(C \cap D) = p(D/C).p(C) = \frac{25}{200} \cdot \frac{200}{600} = 0,04166.$$

4°) Probabilidad de que la unidad elegida sea de A y buena.

$$p(A \cap \overline{D}) = p(\overline{D}/A).p(A) = \frac{290}{300} \cdot \frac{1}{2} = 0,4833$$

5°) Probabilidad de que la unidad elegida sea buena.

$$p(\overline{D}) = \frac{290 + 80 + 175}{600} = 0,90833.$$

6°) Probabilidad de que la unidad elegida sea defectuosa.

$$p(D) = 1 - p(\overline{D}) = 1 - 0,90833 = 0,09166.$$

7°) Siendo defectuosa, probabilidad de que perteneciera a la compañía C.

$$p(C/D) = \frac{p(C \cap D)}{p(D)} = \frac{p(C).p(D/C)}{p(D)} = \frac{0,04166}{0,09166} = 0,4545.$$

8°) Siendo buena, probabilidad de que perteneciera a la compañía B. Aplicando el teorema de Bayes, resulta:

$$p(B/\overline{D}) = \frac{p(\overline{D}/B).p(B)}{p(\overline{D}/A).p(A) + p(\overline{D}/B).p(B) + p(\overline{D}/C).p(C)} =$$

$$= \frac{\frac{80}{100} \cdot \frac{1}{6}}{\frac{290}{300} \cdot \frac{1}{2} + \frac{80}{100} \cdot \frac{1}{6} + \frac{175}{200} \cdot \frac{1}{3}} = 0,1467$$

9°) Siendo de la compañía A, probabilidad de que sea defectuosa.

$$p(D/A) = \frac{10}{300} = 0,0333$$

10°) Siendo de la compañía B, probabilidad de que sea buena.

$$p(\overline{D}/B) = \frac{80}{100} = 0,8$$

****

hecho

2.4 *Se ha comprobado que en una ciudad están enfermos con diarrea el 60% de los niños, con sarampión el 50% y el 20% con ambas enfermedades. Calcular:*

1°) Probabilidad de que elegido un niño al azar esté enfermo con diarrea o sarampión o ambas enfermedades.

2°) En un colegio con 450 niños ¿cuántos cabe esperar que estén enfermos con diarrea o sarampión o las dos enfermedades a la vez?.

S O L U C I O N:

Si designamos por A el suceso "estar enfermo con diarrea" y por B al suceso "estar enfermo con sarampión", tendremos:

1°) Probabilidad de que elegido un niño al azar tenga diarrea, sarampión o ambas. Se trata del suceso unión. Es decir $A \cup B$.

Por tanto: $p(A \cup B) = p(A) + p(B) - p(A \cap B) =$

$$= 0,6 + 0,5 - 0,2 = 0,9.$$

2°) En un colegio con 450 niños, ¿cuántos cabe esperar que estén enfermos con diarrea o sarampión o ambas?

$$n = 450.p(A \cup B) = 450.0,9 = 405.$$

****

2.5 *Se clasifican 155 varones que padecen una cierta enfermedad según su edad y su peso.*

hecho

PESO

		55-65	65-75	75-85	>85
EDAD	40-49	5	8	20	30
	50-59	2	8	25	30
	60-69	2	5	8	12

Calcular:

1°) Probabilidad de que un individuo esté en el intervalo de edad 40-49.

2°) Probabilidad de que un individuo esté en el intervalo de edad 50-59 y pese 75-85 Kgrs.

3°) Probabilidad de que un individuo pese menos de 75 Kgrs.

4°) Probabilidad de que un individuo pese menos de 65 Kgrs. y sea menor de 50 años.

5°) Probabilidad de que un individuo esté en el intervalo de edad

40-49 y pese de 65 a 75 Kgrs.

SOLUCION:

Si representamos al peso como la variable x, y a la edad como la variable y, tendremos:

1°) $p(40 \leqslant y \leqslant 49) = \dfrac{5+8+20+30}{155} = 0,406.$

2°) $p\left[(50 \leqslant y \leqslant 59) \cap (75 \leqslant x \leqslant 85)\right] = \dfrac{25}{155} = 0,1612$

3°) $p(x < 75) = \dfrac{5+8+2+8+2+5}{155} = \dfrac{30}{155} = 0,193.$

4°) $p\left[(x < 65) \cap (y < 50)\right] = \dfrac{5}{155} = 0,032.$

5°) $p\left[(40 \leqslant y \leqslant 49) \cap (65 \leqslant x \leqslant 75)\right] = \dfrac{8}{155} = 0,051.$

****

2.6 *Para la señalización de emergencia de un hospital se han instalado dos indicadores que funcionan independientemente. La probabilidad de que el indicador accione durante la avería es igual a 0,99 para el primero de ellos y 0,95 para el segundo. Hallar la probabilidad de que durante la avería accione solo un indicador.*

SOLUCION:

Consideremos los siguientes sucesos:

$$\left.\begin{array}{l} A = \text{"funcione el indicador 1°"} \\ B = \text{"funcione el indicador 2°"} \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} p(A) = 0,99 \\ p(B) = 0,95 \end{array}\right.$$

Como queremos que durante la avería funcione solo un indicador, será que funcione A y no B o que funcione B y no A, lo que expresaremos del siguiente modo: $(A \cap \overline{B}) \cup (\overline{A} \cap B)$

$$p\left[(A \cap \overline{B}) \cup (\overline{A} \cap B)\right] = p(A \cap \overline{B}) + p(\overline{A} \cap B) = p(A)p(\overline{B})+p(\overline{A})p(B) =$$
$$= 0,99.0,05 + 0,01.0,95 = 0,059$$

****

2.7 *La probabilidad de que una mujer viva dentro de 30 años es 0,25 y la probabilidad de que viva su hijo dentro de 30 años es 0,90. Calcular:*

a) Probabilidad de que ambos vivan.
b) Probabilidad de que únicamente viva la madre.
c) Probabilidad de que únicamente viva el hijo.
d) Probabilidad de que al menos viva uno de los dos.

SOLUCION:

Representemos por M y H respectivamente a los siguientes sucesos:

M = "que viva la madre dentro de 30 años"

H = "que viva el hijo dentro de 30 años"

cuyas probabilidades son concocidas, $p(M) = 0{,}25$ y $p(H) = 0{,}90$ por tanto las probabilidades de los sucesos contrarios serán:

$$p(\overline{M}) = 0{,}75 \quad \text{y} \quad p(\overline{H}) = 0{,}1$$

a) Probabilidad de que ambos vivan dentro de 30 años.

$$p(M \cap H) = p(M).p(H) = 0{,}25.0{,}90 = 0{,}225.$$

b) Probabilidad de que únicamente viva la madre dentro de 30 años.

$$p(M \cap \overline{H}) = p(M).p(\overline{H}) = 0{,}25.0{,}10 = 0{,}025.$$

c) Probabilidad de que únicamente viva el hijo dentro de 30 años.

$$p(\overline{M} \cap H) = p(\overline{M}).p(H) = 0{,}75.0{,}90 = 0{,}675.$$

d) Probabilidad de que al menos viva uno de los dos.

$$p(M \cup H) = 1 - p(\overline{M \cup H}) = 1 - p(\overline{M} \cap \overline{H}) = 1 - p(\overline{M}).p(\overline{H}) =$$
$$= 1 - 0{,}1.0{,}75 = 0{,}925.$$

****

2.8 *Un estudiante de Ciencias Biológicas busca una fórmula que necesita en tres tratados de Bioquímica. Las probabilidades de que la citada fórmula se encuentre en el primero, segundo o tercer tratado respectivamente son, 0,5; 0,6; 0,7. Hallar la probabilidad de que la fórmula se encuentre:*

a) Solamente en un tratado.

b) Unicamente en dos tratados.

c) En los tres tratados.

SOLUCION:

Representemos por A, B y C los siguientes sucesos:

A = "la fórmula se encuentra en el primer tratado"

B = "la fórmula se encuentra en el segundo tratado"

C = "la fórmula se encuentra en el tercer tratado"

siendo sus probabilidades: $p(A) = 0{,}5$; $p(B) = 0{,}6$; $p(C) = 0{,}7$.

a) Que la fórmula se encuentre en un solo tratado.

$$p\left[(A\cap\overline{B}\cap\overline{C})\cup(\overline{A}\cap B\cap\overline{C})\cup(\overline{A}\cap\overline{B}\cap C)\right] = 0{,}5.0{,}4.0{,}3 + 0{,}5.0{,}6.0{,}3 + + 0{,}5.0{,}4.0{,}7 = 0{,}29$$

b) Que la fórmula se encuentre únicamente en dos tratados.

$$p\left[(A\cap B\cap\overline{C})\cup(A\cap\overline{B}\cap C)\cup(\overline{A}\cap B\cap C)\right] = 0{,}5.0{,}6.0{,}3 + 0{,}5.0{,}4.0{,}7 + + 0{,}5.0{,}4.0{,}3 = 0{,}29$$

c) Que la fórmula se encuentre en los tres tratados.

$p(A\cap B\cap C) = p(A).p(B).P(C) = 0{,}5.0{,}6.0{,}7 = 0{,}21.$

****

2.9 *El 10% de las personas que contraen tuberculosis del aparato respiratorio no se curan ¿Cuál es la probabilidad de que de 4 personas que contraen la enfermedad no se cure ninguna?.*

Si aplicamos un nuevo tratamiento sobre un grupo de control compuesto por 30 personas que padecen la enfermedad y obtenemos como resultado la curación de todas ellas, ¿podemos decir que el tratamiento ha sido eficaz?. Para ello calcúlese la probabilidad de que hubiese ocurrido el resultado observado, sin haber aplicado el nuevo tratamiento.

SOLUCION:

Sea A el suceso curar a un paciente que padece tuberculosis del aparato respiratorio, por tanto el suceso contrario será $\overline{A}$.

Las probabilidades de ambos sucesos son: $p(\overline{A}) = 0{,}1$; $p(A) = 0{,}9$

En el primer caso, se desea saber:

$$p(\overline{A}\cap\overline{A}\cap\overline{A}\cap\overline{A}) = p(\overline{A})^4 = 0{,}1^4 = 0{,}0001$$

En el segundo caso, calculemos la probabilidad de que de 30 individuos curen todos:

$$p(\overbrace{A\cap A\cap A\cap\ldots\cap A}^{30\ \text{veces}}) = p(A)^{30} = 0{,}9^{30} = 0{,}0424.$$

Obsérvese que se trata de una probabilidad pequeñisima y si con el nuevo tratamiento hemos conseguido la curación de los 30 individuos podemos concluir que el tratamiento ha sido eficaz para el grupo de control.

****

2.10 *Elegido un individuo al azar y observado por Rayos X, se diagnosticó que estaba tuberculoso. La probabilidad de que en la po-*

blación de la que se eligió el individuo, uno de ellos sea tuberculoso es de 0,01. La probabilidad de que un aparato de Rayos X detecte que un individuo es tuberculoso siendolo es 0,97 y no siendolo es 0,001. ¿Que podemos decir acerca del diagnóstico?.

S O L U C I O N:

Representemos por T, $\overline{T}$ y A a los sucesos siguientes:

$$\begin{cases} T = \text{"que un individuo sea tuberculoso"} \\ \overline{T} = \text{"que un individuo no sea tuberculoso"} \\ A = \text{"detectar que un individuo es tuberculoso"} \end{cases}$$

Por tanto, sus probabilidades son:

$$\begin{cases} p(T) = 0{,}01 \\ p(A/T) = 0{,}97 \end{cases} \qquad \begin{cases} p(\overline{T}) = 1-p(T) = 0{,}99 \\ p(A/\overline{T}) = 0{,}001 \end{cases}$$

El problema nos pide encontrar la probabilidad siguiente: p(T/A).

Aplicando el teorema de Bayes, resulta:

$$p(T/A) = \frac{p(T).p(A/T)}{p(T).p(A/T) + p(\overline{T}).p(A/\overline{T})} = \frac{0{,}01.0{,}97}{0{,}01.0{,}97+0{,}99.0{,}001} =$$

$$= 0{,}907.$$

****

hecho

2.11 *Se cree que un enfermo ha podido contraer las enfermedades A o B, con probabilidades 0,25 y 0,4 respectivamente.*

Para establecer un diagnóstico diferencial, se somete al paciente a un análisis clínico, del que solo se pueden obtener dos resultados positivo o negativo. Por la experiencia acumulada en otros casos análogos, se sabe que la probabilidad de que el análisis resulte positivo para los que padecen la enfermedad A es 0,04 y 0,07 para los que padecen la enfermedad B.

Al enfermo se le hace el análisis, y da resultado positivo. ¿Cuál es la probabilidad de que el paciente tuviera la enfermedad A y cuál es la probabilidad de que tuviera la enfermedad B.

S O L U C I O N:

El enunciado nos proporciona las siguientes probabilidades:

$$\begin{cases} p(A) = 0{,}25 \\ p(B) = 0{,}4 \end{cases} \qquad \begin{cases} p(+/A) = 0{,}04 \\ p(+/B) = 0{,}07 \end{cases}$$

1°) Probabilidad de que padezca la enfermedad A, habiendo dado positivo el análisis.

$$p(A/+) = \frac{p(A).p(+/A)}{p(A).p(+/A) + p(B).p(+/B)} = \frac{0{,}25.0{,}04}{0{,}25.0{,}04+0{,}4.0{,}07} =$$

$$= 0{,}263.$$

2°) Probabilidad de que padezca la enfermedad B, habiendo dado positivo el análisis.

$$p(B/+) = \frac{p(B).p(+/B)}{p(A).p(+/A) + p(B).p(+/B)} = \frac{0{,}4.0{,}07}{0{,}25.0{,}04+0{,}4.0{,}07} =$$

$$= 0{,}73.$$

****

2.12 *En un hospital especializado en enfermedades del torax, ingresan un promedio de 50% de enfermos con bronquitis, 30% con neumonía y 20% con gripe. La probabilidad de curación completa de cada una de estas enfermedades es respectivamente 0,7; 0,8;y 0,9.*

Un enfermo internado en el hospital ha sido dado de alta completamente sano. Hallar la probabilidad de que el enfermo dado de alta, ingresara con bronquitis.

S O L U C I O N:

Representemos por A_1, A_2, A_3 a los siguientes sucesos:

$$\begin{cases} A_1 = \text{"el enfermo ingresa padeciendo bronquitis"} \\ A_2 = \text{"el enfermo ingresa padeciendo neumonía"} \\ A_3 = \text{"el enfermo ingresa padeciendo gripe"} \end{cases}$$

siendo sus probabilidades los siguientes valores:

$p(A_1) = 0{,}5.$ $\quad p(A_2) = 0{,}3.$ $\quad p(A_3) = 0{,}2.$

Sea C el suceso "el paciente cura".

También conocemos las siguientes probabilidades:

$p(C/A_1) = 0{,}7$ $\quad p(C/A_2) = 0{,}8$ $\quad p(C/A_3) = 0{,}9$

Deseamos saber la probabilidad de que el enfermo ingresara con bronquitis condicionado a que ha sido dado de alta totalmente curado.

Es decir: $p(A_1/C)$.

Aplicando el teorema de Bayes resulta:

$$p(A_1/C) = \frac{p(A_1)p(C/A_1)}{p(A_1)p(C/A_1) + p(A_2)p(C/A_2) + p(A_3)p(C/A_3)} =$$

$$= \frac{0{,}5.0{,}7}{0{,}5.0{,}7 + 0{,}3.0{,}8 + 0{,}2.0{,}9} = 0{,}4545.$$

****

2.13 *Una enfermedad puede ser producida por tres virus A, B y C. En un laboratorio se tienen tres tubos con virus A, dos con el virus B y cinco con el virus C. La probabilidad de que el virus A produzca la enfermedad es 1/3, que la produzca el virus B es 2/3 y que la produzca el virus C es 1/7.*

Se inocula un virus a un animal y contrae la enfermedad. ¿Cuál es la probabilidad de que el virus que se inoculó fuera el C?

S O L U C I O N:

Si representamos por E al suceso "el animal contrae la enfermedad".

Tendremos: $p(A) = 3/10$ $p(B) = 2/10$ $p(C) = 5/10$

$p(E/A) = 1/3$ $p(E/B) = 2/3$ $p(E/C) = 1/7$

Queremos calcular $p(C/E)$. Aplicando el teorema de Bayes, resulta:

$$p(C/E) = \frac{p(C)p(E/C)}{p(A)p(E/A) + p(B)p(E/B) + p(C)p(E/C)} =$$

$$= \frac{\frac{5}{10}\,\frac{1}{7}}{\frac{3}{10}\,\frac{1}{3} + \frac{2}{10}\,\frac{2}{3} + \frac{5}{10}\,\frac{1}{7}} = 0{,}304.$$

****

2.14 *En una población y en un determinado momento, se conocen las probabilidades de que un individuo tenga las enfermedades A, B y C y de que esté sano, siendo estas las siguientes:*

$p(A) = 0{,}02$ $p(B) = 0{,}05$ $p(C) = 0{,}03$ $p(S) = 0{,}9$

Consideremos un síntoma, representado por el símbolo ε. Las probabilidades de que se presente dicho síntoma en cada caso anterior son: $p(\varepsilon/A) = 0{,}99$ $p(\varepsilon/B) = 0{,}4$

$p(\varepsilon/C) = 0{,}1$ $p(\varepsilon/S) = 0{,}01$

Examinando a un individuo, se observa que presenta el cita-

do síntoma $\mathcal{E}$. Se desea calcular:
1°) $p(A/\mathcal{E})$; 2°) $p(B/\mathcal{E})$; 3°) $p(C/\mathcal{E})$; 4°) $p(S/\mathcal{E})$

SOLUCION:

Como tenemos todos los datos directamente en el enunciado, aplicaremos el teorema de Bayes.

1°) $$p(A/\mathcal{E}) = \frac{0{,}02.0{,}99}{0{,}02.0{,}99 + 0{,}05.0{,}4 + 0{,}03.0{,}1 + 0{,}9.0{,}01} = 0{,}382.$$

2°) $$p(B/\mathcal{E}) = \frac{0{,}05.0{,}4}{0{,}02.0{,}99 + 0{,}05.0{,}4 + 0{,}03.0{,}1 + 0{,}9.0{,}01} = 0{,}386.$$

3°) $$p(C/\mathcal{E}) = \frac{0{,}03.0{,}1}{0{,}02.0{,}99 + 0{,}05.0{,}4 + 0{,}03.0{,}1 + 0{,}9.0{,}01} = 0{,}057.$$

4°) $$p(S/\mathcal{E}) = \frac{0{,}9.0{,}01}{0{,}02.0{,}99 + 0{,}05.0{,}4 + 0{,}03.0{,}1 + 0{,}9.0{,}01} = 0{,}173$$

****

2.15 *En una determinada especie se ha observado que de padres con ojos oscuros nacen hijos de ojos oscuros en el 5% de los casos, e hijos con ojos claros en el 8% de los casos. Del mismo modo se ha observado que de padres con ojos claros nacen hijos con ojos oscuros en el 9% de los casos e hijos con ojos claros el 78% de los casos. Establecer las relaciones que existen entre el color de los ojos del padre y del hijo de la citada especie.*

SOLUCION:

		padres oscuros A	claros $\bar{A}$
hijos	oscuros B	0,05	0,09
	claros $\bar{B}$	0,08	0,78

Calculemos las siguientes probabilidades condicionadas:

1°) Probabilidad de que el hijo tenga ojos oscuros condicionada a que el padre también los tenga.

$$p(B/A) = \frac{p(A\cap B)}{p(A)} = \frac{0{,}05}{p(A\cap B) + p(A\cap \bar{B})} = \frac{0{,}05}{0{,}05+0{,}08} = 0{,}39.$$

2°) Probabilidad de que el hijo tenga ojos claros, condicionada a que el padre los tenga oscuros.

$$p(\overline{B}/A) = 1 - p(B/A) = 1 - 0,39 = 0,61$$

3°) Probabilidad de que el hijo tenga ojos oscuros, condicionada a que el padre tenga los ojos claros.

$$p(B/\overline{A}) = \frac{p(\overline{A} \cap B)}{p(\overline{A})} = \frac{0,09}{p(\overline{A} \cap B)+p(\overline{A} \cap \overline{B})} = \frac{0,09}{0,09 + 0,78} = 0,103.$$

4°) Probabilidad de que el hijo tenga ojos claros si el padre los tiene también claros.

$$p(\overline{B}/\overline{A}) = 1 - p(B/\overline{A}) = 1 - 0,103 = 0,897$$

****

3- distribuciones discretas

3 - DISTRIBUCIONES DISCRETAS

1. GENERALIDADES

1.1 VARIABLE ALEATORIA.- En un espacio probabilístico (E,A,p), llamaremos *variable aleatoria*, y representaremos por X, a una función del conjunto de los sucesos elementales de un experimento aleatorio en el cuerpo de los números reales.

De tal forma que dado un número real cualquiera x, si consideramos el conjunto de todos los sucesos elementales, para los que la función X toma valores menores o iguales a x, ese conjunto ha de ser un suceso del algebra **A**.

1.2 VARIABLE ALEATORIA DISCRETA.- Se dice que una variable aleatoria es *discreta*, cuando el conjunto imagen de la función X, es un subconjunto de $\mathbb{R}$ finito o infinito numerable.

[NOTA: *Se dice que un conjunto es infinito numerable, si se puede establecer una aplicación biyectiva entre dicho conjunto y el conjunto de los números naturales.*]

Es decir, las variables aleatorias discretas solo pueden tomar valores "aislados".

1.3 LEY DE PROBABILIDAD.- Consideremos un espacio probabilístico (E,A,p) y sea X una variable aleatoria discreta que toma los valores $x_1, x_2, \ldots, x_n$, para cada uno de los cuales se conoce su probabilidad. Es decir: $p(X = x_i) = p_i \geqslant 0$, $(\Sigma p_i = 1)$

Llamaremos *ley de probabilidad*, de la variable aleatoria X, a la correspondencia entre los valores x_i y sus probabilidades p_i.

1.4 FUNCION DE DISTRIBUCION.- Consideremos un espacio probabilístico (E,A,p), y sea X una variable aleatoria discreta, llamaremos *función de distribución* de la variable aleatoria X, a la función:

$$F(x) = p(X \leq x)$$

1.5 PROPIEDADES DE LA FUNCION DE DISTRIBUCION.-

1ª) $F(+\infty) = 1$.

- En efecto, ya que $F(+\infty) = \lim_{x\to+\infty} F(x) = \lim_{x\to+\infty} p(X \leq x) = 1$

2ª) $F(-\infty) = 0$

- En efecto, ya que $F(-\infty) = \lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} p(X \leq x) = 0$

3ª) Si $x_1 \leq x_2 \implies F(x_1) \leq F(x_2)$

- En efecto, pues:

$F(x_2) = p(X \leq x_2) = p(X \leq x_1) + p(x_1 < X \leq x_2) =$

$= F(x_1) + p(x_1 < X \leq x_2)$. Como $p(x_1 < X \leq x_2) \geq 0$

se deduce que $F(x_1) \leq F(x_2)$.

Como consecuencia de esta propiedad se tiene que:

$$p(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

4ª) $F(x)$ es una función contínua a la derecha cualquiera que sea el valor de x.

- En efecto, $F(x_1+h) - F(x_1) = p(x_1 < X \leq x_1+h)$

Cuando $h \to 0$, se sigue que $F(x_1+h) - F(x_1) \to 0$.

Resumiendo estas cuatro propiedades diremos:

Función de distribución es una función monótona no decreciente contínua a la derecha en cada punto y tal que

$$F(+\infty) = 1 \qquad y \qquad F(-\infty) = 0$$

1.6 ESPERANZA MATEMATICA DE UNA VARIABLE ALEATORIA DISCRETA.- Sea X una variable aleatoria discreta, de tal forma que se conocen las probabilidades $p(X = x_i) = p_i$

Llamamos *esperanza matemática, valor esperado, o media* de la variable aleatoria discreta X, a la suma de la serie:

$$E[X] = p_1x_1 + p_2x_2 + \ldots + p_nx_n = \sum_{i=1}^{n} p_ix_i$$

1.7 PROPIEDADES DE LA ESPERANZA MATEMATICA.-

1ª) La esperanza matemática de una constante es igual a ella misma. Es decir: $E[c] = c$

- En efecto, ya que la constante c, se considera como una variable aleatoria discreta que tiene un solo valor posible c y lo toma con probabilidad 1.

Por tanto: $E[c] = c.1 = c.$

2ª) La esperanza matemática de la suma de dos variables aleatorias es igual a la suma de las esperanzas matemáticas de los sumandos.

Es decir: $E[X + Y] = E[X] + E[Y]$

- En efecto, pues sean X e Y las variables aleatorias discretas , dadas por las siguientes leyes de probabilidad.

X	x_1	x_2
p	p_1	p_2

Y	y_1	y_2
p'	p'_1	p'_2

Al construir la nueva variable X + Y, tendrá la siguiente ley de probabilidad:

X + Y	x_1+y_1	x_1+y_2	x_2+y_1	x_2+y_2
p*	p_{11}	p_{12}	p_{21}	p_{22}

Entonces la esperanza de X + Y, será:

$$E[X + Y] = (x_1+y_1)p_{11} + (x_1+y_2)p_{12} + (x_2+y_1)p_{21}+(x_2+y_2)p_{22} =$$

$$= x_1(p_{11}+p_{12})+x_2(p_{21}+p_{22})+y_1(p_{11}+p_{21})+y_2(p_{12}+p_{22}) =$$

por el teorema de adición, tenemos:

$$\left.\begin{array}{l} p_{11}+p_{12} = p_1 \\ p_{11}+p_{21} = p'_1 \end{array}\right\} \qquad \left.\begin{array}{l} p_{21}+p_{22} = p_2 \\ p_{12}+p_{22} = p'_2 \end{array}\right\}$$

$$E[X + Y] = x_1p_1+x_2p_2+y_1p'_1+y_2p'_2 = E[X] + E[Y].$$

Observación.- Para la demostracción de esta propiedad 2ª) hemos supuesto únicamente dos valores posibles para cada una de las variables. Análogamente se demostraría para más valores.

Generalizando la propiedad anterior se tiene:

$$E[X_1+X_2+\ldots+X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]$$

3ª) La esperanza matemática del producto de una constante por una variable aleatoria es igual al producto de la constante por la esperanza de la variable.

Es decir: $$E[cX] = c\,E[X]$$

- En efecto, ya que:

$$E[cX] = \sum c\, x_i\, p_i = c \sum x_i\, p_i = c\,E[X]$$

4ª) La esperanza matemática del producto de dos variables aleatorias independientes es igual al producto de sus esperanzas matemáticas

Es decir: $$E[X.Y] = E[X].E[Y]$$

- Supongamos dos variables X e Y con leyes de probabilidad como las vistas para la demostracción de la propiedad 2ª).

X	x_1	x_2
p	p_1	p_2

Y	y_1	y_2
p'	p'_1	p'_2

Al construir la nueva variable X.Y, tendrá la siguiente ley de probabilidad:

X . Y	$x_1 . y_1$	$x_1 y_2$	$x_2 y_1$	$x_2 y_2$
p*	$p_1 . p'_1$	$p_1 p'_2$	$p_2 p'_1$	$p_2 p'_2$

Entonces la esperanza de X.Y, será:

$$E[X.Y] = x_1y_1p_1p'_1+x_1y_2p_1p'_2+x_2y_1p_2p'_1+x_2y_2p_2p'_2 =$$
$$= y_1p'_1(x_1p_1+x_2p_2) + y_2p'_2(x_1p_1+x_2p_2) =$$
$$=(x_1p_1+x_2p_2)(y_1p'_1+y_2p'_2) = E[X].E[Y].$$

Generalizando para un número arbitrario de variables alea torias independientes, tendremos:

$$E[X_1 X_2 \cdot \ldots \cdot X_n] = E[X_1]\, E[X_2] \cdot \ldots \cdot E[X_n]$$

1.8 MOMENTOS RESPETO AL ORIGEN.- Llamaremos *momento de orden r respecto al origen* de la variable aleatoria X, y representaremos por α_r, a la esperanza matemática de X^r.

Es decir: $\alpha_r = E[X^r]$

En el caso de variable aleatoria discreta, tendremos:

$$\alpha_r = E[X^r] = \sum x_i^r\, p_i$$

De lo que se deduce:

$$\alpha_1 = E[X] = \sum x_i\, p_i$$

Es decir, la esperanza matemática coincide con el momento de primer orden respecto al origen, que de ahora en adelante llamaremos media poblacional y representaremos por μ.

1.9 MOMENTOS RESPECTO A LA MEDIA.- Llamaremos *momento respecto a la media o momento central de orden r* de la variable aleatoria X y representaremos por μ_r, a la esperanza matemática de $(X - \alpha_1)^r$.

Es decir: $\mu_r = E\left[(X - \alpha_1)^r\right]$

Para el caso de variable aleatoria discreta, tendremos:

$$\mu_r = E\left[(X - \alpha_1)^r\right] = \sum (x_i - \alpha_1)^r\, p_i$$

Del mismo modo a como vimos en Estadística Descriptiva el momento de primer orden es nulo, ya que:

$$\mu_1 = E[X - \alpha_1] = E[X] - E[\alpha_1] = \alpha_1 - \alpha_1 = 0$$

El momento de segundo orden es igual a la varianza.

Es decir: $\mu_2 = E\left[(X - \alpha_1)^2\right] = \sum (x_i - \alpha_1)^2\, p_i = \sigma^2$

También se suele expresar por:

$$\mu_2 = \sigma^2 = V[X] = D^2[X].$$

1.10 RELACION ENTRE AMBOS MOMENTOS.-

$$\mu_r = E[(X - \alpha_1)^r] = E\left[X^r - \binom{r}{1} X^{r-1}\alpha_1 + \binom{r}{2} X^{r-2}\alpha_1^2 - \ldots \pm \alpha_1^r\right] =$$

$$= \alpha_r - \binom{r}{1}\alpha_1\,\alpha_{r-1} + \binom{r}{2}\alpha_1^2\,\alpha_{r-2} - \ldots \pm \alpha_1^r$$

Esta expresión nos permite pasar de los momentos respecto al origen a los momentos centrales.

Como caso particular, para la varianza tendremos:

$$\mu_2 = \alpha_2 - \alpha_1^2$$

Del mismo modo, desarrollando la expresión siguiente:

$$\mu_r = E\left[X^r\right] = E\left[\left[(X - \alpha_1) + \alpha_1\right]^r\right] =$$
$$= \alpha_r + \binom{r}{1}\alpha_1 \ \alpha_{r-1} + \binom{r}{2}\alpha_1^2 \ \alpha_{r-2} + \ldots + \alpha_1^r .$$

Obtenemos la expresión que nos da los momentos respecto al origen conocidos los momentos respecto a la media y la media.

2. DISTRIBUCION BINOMIAL

2.1 INTRODUCCION. En gran parte de los problemas clínicos, biológicos, etc. debemos considerar ciertas alternativas cuya probabilidad es constante.

Por ejemplo, en una determinada especie se sabe que la probabilidad de que un individuo vacunado contra cierta enfermedad, contraiga dicha enfermedad es 0,2, por tanto, la probabilidad de que no la contraiga será 0,8.

Entonces, se puede preguntar, ¿cuál es la probabilidad para que de una muestra de n individuos tomados al azar de una población, en la que todos han sido vacunados, se encuentren r individuos que contraen la enfermedad?

2.2 DEFINICION.- Consideremos un experimento aleatorio cualquiera, para el que únicamente se pueden presentar los sucesos A y contrario $\overline{A}$, con probabilidades p y q respectivamente.

Representemos al suceso A como *éxito* y como consecuencia el suceso $\overline{A}$, será *fracaso*. Realicemos n pruebas sucesivas del mismo experimento aleatorio.

Deseamos saber la probabilidad de obtener r éxitos en n pruebas.

Para ello consideremos uno de los casos en que se verifica, obtener r éxitos en n pruebas. Sea el suceso:

$$\overbrace{A \cap A \cap A \cap \ldots \cap A}^{r \text{ veces}} \cap \overbrace{\bar{A} \cap \bar{A} \cap \bar{A} \cap \ldots \cap \bar{A}}^{n-r \text{ veces}}$$

Esta claro que la probabilidad de este suceso, teniendo en cuenta la independencia de los resultados en sucesivas pruebas, será:

$$p(\overbrace{A \cap A \cap \ldots \cap A}^{r} \cap \overbrace{\bar{A} \cap \bar{A} \cap \ldots \cap \bar{A}}^{n-r}) = p^r . q^{n-r}$$

Ahora bien, conviene considerar todas las maneras posibles de obtener r éxitos y n-r fracasos, estas serán las permutaciones con repetición de n elementos entre los que r están repetidos y n-r también.

Es decir: $$P_n^{r,\, n-r} = \frac{n!}{(n-r)!\; r!}$$

Por tanto la probabilidad pedida será el producto de este número por la probabilidad $p^r . q^{n-r}$

Es decir: $$p(X = r) = \frac{n!}{(n-r)!\; r!}\, p^r\, q^{n-r}$$

2.3 LEY DE PROBABILIDAD.- Como acabamos de ver, sea X una variable discreta que representa el número de éxitos.

La función que nos da la probabilidad de que en n pruebas se produzcan r éxitos, viene dada por la expresión:

$$p(X = r) = \binom{n}{r}\, p^r\, q^{n-r}$$

Obsérvese que esta es una función puntual, que nos da la probabilidad para los distintos valores desde X = 0 hasta X = n. El cálculo de esta función sería laborioso, por esta razón se han construido tablas (Ver APENDICE tabla I), que nos proporcionan para los distintos valores de n y de p, la probabilidad de que la variable X tome los distintos valores de 0 a n.

2.4 FUNCION DE DISTRIBUCION.- Por lo visto en 1.4 del presente capítulo. Llamaremos *función de distribución binomial* y representaremos por F(x) a la siguiente expresión:

$$F(x) = p(X \leq x) = \sum_{r=0}^{x} \binom{n}{r} p^r q^{n-r}$$

Obsérvese que a partir de la definición de F(x), es inmediato comprobar que se verifican las propiedades vistas en 1.5.

2.5 CARACTERISTICAS ESTADISTICAS.- Veamos un caso particular de la distribución binomial, aquel en el que únicamente se realiza una prueba en lugar de n.

A esta variable se le llama variable aleatoria de Bernouilli, y es una función que toma los valores:

$$\begin{cases} x_1 = 1 & \text{con probabilidad } p \\ x_2 = 0 & \text{con probabilidad } q = 1 - p \end{cases}$$

Por tanto, la media será:

$$\mu = E[X] = \sum p_i x_i = 1.p + 0.q = p$$

La varianza será:

$$\sigma^2 = E\left[(X - \mu)^2\right] = \sum (x_i - \mu)^2 p_i = (1-p)^2 p + (0-p)^2 q =$$
$$= (1-2p+p^2)p + p^2(1-p) = p(1-p) = pq.$$

Consideremos ahora el caso de una distribución binomial con n pruebas repetidas.

Se tiene que:

1°) Media $\mu = E[X] = n\,p$

2°) Varianza $\sigma^2 = n\,p\,q$

3°) Desviación típica $\sigma = \sqrt{npq}$

4°) Coeficiente de asimetría $\gamma_1 = \dfrac{q - p}{\sqrt{n\,p\,q}}$

5°) Coeficiente de curtosis $\gamma_2 = \dfrac{1 - 6\,p\,q}{n\,p\,q}$

2.6 APLICACION DE LA DISTRIBUCION BINOMIAL A LA HERENCIA BIOLOGICA.- Consideremos un animal cualquiera cuyos padres pertenecen a dos grupos distintos A y B. Sea p_A la probabilidad de que el hijo pertenezca al grupo A y $p_B = 1 - p_A$, la probabilidad de que el hijo pertenezca al grupo B.

Si se cruzan n parejas de animales, siendo uno del grupo A y otro del grupo B, el número de descendientes que pertenecerán al grupo A sigue una distribución binomial de parámetros n y p_A. Del mismo modo, el número de descendientes que pertenecerán al grupo B, sigue una distribución binomial de parámetros n y p_B.

2.7 AJUSTE DE UN DISTRIBUCION EMPIRICA POR UNA DISTRIBUCION TEORICA

Si se observa que una variable estadística obtenida experimentalmente a partir de una muestra satisface las condiciones que conducen a una distribución binomial, tendrá una distribución empírica que se aproximará a una distribución binomial teórica.

El problema está en seleccionar entre todas las distribuciones binomiales, la que mejor se aproxima a la distribución empírica. Este es el fundamento del *ajuste estadístico*.

Por otra parte se demuestra que la distribución binomial que mejor se aproxima a una distribución empírica es aquella que tiene la misma media.

Para ello calcularemos la media muestral observada $\bar{x}$ y la haremos coincidir con la media poblacional

$$\mu = n\,p = \bar{x} \quad \Longrightarrow \quad p = \frac{\bar{x}}{n}$$

Por tanto, consideraremos la distribución binomial de parámetros n y $\frac{\bar{x}}{n}$.

En próximos capítulos, trataremos de dar una medida objetiva de la bondad del ajuste efectuado.

3. DISTRIBUCION HIPERGEOMETRICA

3.1 INTRODUCCION.- Ya hemos visto como la distribución binomial correspondía a sondeos de una población donde los individuos se clasificaban en dos categorias, teniendo en cuenta que estos sondeos se realizaban con reemplazamiento. Pues bien, la distribución hipergeométrica corresponde a sondeos sin reemplazamiento de una población finita, en la que los individuos se clasifican también en dos categorias.

3.2 DEFINICION.- Se dice que una variable aleatoria X que puede tomar los valores comprendidos entre 0 y n , tiene una distribución hipergeométrica, cuando:

$$p(X = r) = \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}}$$

siendo Np y Nq números enteros.

A continuación vamos a dar una idea física de esta distribución mediante un modelo de urna.

Consideremos una gran urna que contiene N bolas de las que:

$$\begin{cases} Np \text{ son blancas} & (p = \text{proporción de blancas}) \\ Nq \text{ son negras} & (q = \text{proporción de negras} = 1 - p) \end{cases}$$

Se extrae sin reemplazamiento una muestra de n bolas,Al ser sin reemplazamiento, la composición de la urna varía en cada extracción.

El número de muestras distintas de tamaño n que podemos realizar será: $\binom{N}{n}$.

El número de muestras diferentes con r bolas blancas será: $\binom{Np}{r}\binom{Nq}{n-r}$.

Si representamos por X el número de bolas blancas contenidas en la muestra de tamaño n, tendremos que:

$$p(X = r) = \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}}$$

donde $r \in I \subset \mathbb{N}$, $I = \{0,1,2,\ldots,n\}$

Esta función que acabamos de ver, es la ley de probabilidad de la distribución hipergeométrica.

3.3 FUNCION DE DISTRIBUCION.- Llamaremos *función de distribución hipergeométrica* y representaremos por F(x), a la siguiente expresión:

$$F(x) = p(X \leq x) = \sum_{r=0}^{x} \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}}$$

Obsérvese que a partir de la definición de F(x), se puede comprobar que cumple las propiedades de toda función de distribución vistas en 1.5 del presente capítulo.

3.4 CARACTERISTICAS ESTADISTICAS.-

1°) Media: $\mu = E[X] = np$

La media aritmética de esta distribución es independiente de N y es idéntica a la media de la distribución binomial.

2°) Varianza: $\sigma^2 = \mu_2 = E\left[(X - np)^2\right] = npq\,\frac{N-n}{N-1}$

Obsérvese que la varianza de la distribución hipergeométrica es menor que la de la distribución binomial, ya que el valor de $\frac{N-n}{N-1} \leq 1$. Al factor $\frac{N-n}{N-1}$, se le conoce con el nombre de *factor de exhaustividad.*

El factor de exhaustividad únicamente será igual a la unidad en los siguientes casos:

1°) n = 1, en cuyo caso solo se extrae una bola y por tanto existe identidad entre las extracciones con y sin reemplazamiento.

2°) $N \to \infty$, En este caso $\lim_{N\to\infty} \frac{N-n}{N-1} = 1$ y también aquí, supuesta la urna de extensión infinitamente grande existe identidad entre las dos formas de extracción.

3.5 CONVERGENCIA DE LA DISTRIBUCION HIPERGEOMETRICA CON LA DISTRIBUCION BINOMIAL.- Ya hemos visto, que la distribución hipergeométrica depende de tres parámetros, que son: la extensión de la urna N, la extensión de la muestra n, y la composición de la urna p, en cambio, la distribución binomial solo depende de estos dos últimos parámetros n y p, parece lógico que dada la mayor sencillez de esta última, convenga estudiar la posible aproximación de la distribución hipergeòmétrica a la distribución binomial.

Para ello, estudiemos a que tiende la probabilidad de la distribución hipergeométrica cuando $N \to \infty$.

$$\lim_{N \to \infty} \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}} = \lim_{N \to \infty} \frac{\frac{(Np)!}{r!\,(Np-r)!}\;\frac{(Nq)!}{(n-r)!\,(Nq-n+r)!}}{\frac{N!}{n!(N-n)!}} =$$

$$= \lim_{N\to\infty} \frac{n!}{r!(n-r)!}\;\frac{[Np(Np-1)\ldots(Np-r+1)][Nq(Nq-1)\ldots(Nq-n+r+1)]}{N(N-1)\ldots(N-n+1)} =$$

$$= \frac{n!}{r!(n-r)!}\;\frac{(Np)^r\,(Nq)^{n-r}}{N^n} = \binom{n}{r}\,p^r\,q^{n-r}$$

4. DISTRIBUCION DE POISSON

4.1 INTRODUCCION.- La distribución de Poisson se encuentra en el caso de probabilidades pequeñas, de ahí que se denomine *ley de los sucesos raros*.

Así por ejemplo, las variables siguientes obedecen a la ley de Poisson.

- Número de partos triples por año, en un determinado país.
- Número de átomos desintegrados por segundo, en una cierta cantidad de material radiactivo.
- Número de individuos albinos nacidos durante un año, en un país.
- Número de electrones emitidos por un cátodo excitado, en un cierto intervalo de tiempo.
- Número de individuos centenarios.

- Número de bacterias de una determinada especie contenidas en un centímetro cúbico de cultivo.
- Número de personas que fallecen diariamente por infarto de miocardio en una gran ciudad.

4.2 DEFINICION.- Se dice que la variable aleatoria X sigue una *ley de Poisson*, si puede tomar los valores enteros 0,1,2,...,n siendo sus probabilidades:

$$p(X = r) = e^{-\lambda} \frac{\lambda^r}{r!}$$

Siendo λ una constante positiva que se denomina parámetro de la distribución de Poisson.

4.3 LA DISTRIBUCION DE POISSON COMO LIMITE DE LA DISTRIBUCION BINOMIAL.- Supongamos que realizamos n pruebas independientes siendo p la probabilidad de que ocurra el suceso A, en cada una de las n pruebas.

Para hallar la probabilidad de que ocurra r veces el suceso A, utilizaremos la distribución binomial.

Es decir: $p(X = r) = \binom{n}{r} p^r q^{n-r}$

Ahora bien, si la probabilidad del suceso A es muy pequeña y n es muy grande, vamos a ver que obtenemos una aproximación muy buena, mediante otra distribución que resulta ser la distribución de Poisson.

Para ello hemos de admitir que el producto np conserva un valor constante que llamaremos λ. Esto significa que el promedio de apariciones del suceso, para distintos valores de n permanece invariable.

Calculemos el límite de la distribución binomial para p pequeño, n grande y $np = \lambda$.

$$p(X = r) = \binom{n}{r} p^r q^{n-r} = \frac{n(n-1)\ldots(n-r+1)}{r!} p^r q^{n-r} =$$

$$= \frac{n(n-1)\ldots(n-r+1)}{r!} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r} =$$

$$= \frac{\lambda^r}{r!} \frac{n(n-1)\ldots(n-r+1)}{n^r} \left(1 - \frac{\lambda}{n}\right)^{n-r} =$$

$$= \frac{\lambda^r}{r!} \frac{n(n-1)\ldots(n-r+1)}{n^r} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-r} =$$

$$= \frac{\lambda^r}{r!} \left[\left(1 - \frac{\lambda}{n}\right)^n \frac{\frac{n(n-1)\ldots(n-r+1)}{n^r}}{\left(1 - \frac{\lambda}{n}\right)^r} \right] =$$

$$= \frac{\lambda^r}{r!} \left(1 - \frac{\lambda}{n}\right)^n \frac{\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\ldots\left(1 - \frac{r-1}{n}\right)}{\left(1 - \frac{\lambda}{n}\right)^r} =$$

Ahora bien, $\lim\limits_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{\lim\limits_{n\to\infty} n\left(\frac{-\lambda}{n}\right)} = e^{-\lambda}$ y

$$\lim_{n\to\infty} \frac{\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\ldots\left(1 - \frac{r-1}{n}\right)}{\left(1 - \frac{\lambda}{n}\right)^r} = 1$$

por tener el mismo número de factores en el numerador que en el denominador.

Por tanto: $$p(X = r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

Que por la definición 4.2 sabemos que se trata de la distribución de Poisson.

En la práctica, se reemplaza la distribución binomial por la de Poisson, cuando simultaneamente n sobrepasa 50 y p es menor que 0,1.

4.4 LEY DE PROBABILIDAD.-Si X es la variable aleatoria de la distribución de Poisson, la función que nos da la probabilidad de que en n pruebas ocurra r veces el suceso A, viene dada por la expresión:

$$p(X = r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

donde $r = 0,1,2,\ldots,n$

El cálculo de esta función sería muy laborioso y dada la importancia práctica de la distribución de Poisson se han construido tablas, que nos dan la probabilidad para cada valor de r. (Ver APENDICE tabla II)

4.5 FUNCION DE DISTRIBUCION.- Llamaremos *función de distribución de Poisson* y representaremos por F(x) a la siguiente expresión:

$$F(x) = p(X \leq x) = \sum_{r=0}^{x} \frac{\lambda^r}{r!} e^{-\lambda}$$

Compruebe el alumno, que la función F(x) verifica las propiedades expuestas en 1.5.

4.6 CARACTERISTICAS ESTADISTICAS.-

1°) Media:

$$\mu = E[X] = \sum_{x_i=0}^{\infty} x_i p_i = \sum_{r=0}^{\infty} r\, e^{-\lambda} \frac{\lambda^r}{r!} = \lambda e^{-\lambda} \underbrace{\sum_{r=0}^{\infty} \frac{\lambda^{r-1}}{(r-1)!}}_{e^{\lambda}} =$$

$$= \lambda\, e^{-\lambda} e^{\lambda} = \lambda$$

Por tanto, $\mu = E[X] = \lambda$

2°) Varianza:

$$\sigma^2 = E\left[(X-\lambda)^2\right] = E\left[(X^2 - 2\lambda X + \lambda^2)\right] =$$

$$= E\left[X^2\right] - 2\lambda E\left[X\right] + \lambda^2 = E\left[X^2\right] - 2\lambda^2 + \lambda^2 =$$

$$= E\left[X^2\right] - \lambda^2$$

Calculemos $E\left[X^2\right]$

$$E\left[X^2\right] = \sum r^2 e^{-\lambda} \frac{\lambda^r}{r!} = \lambda e^{-\lambda} \sum \frac{r}{(r-1)!} \lambda^{r-1} =$$

Hagamos el cambio $z = r - 1$, entonces resulta:

$$E\left[X^2\right] = \lambda\, e^{-\lambda} \sum \frac{(z+1)}{z!} \lambda^z = \lambda\, e^{-\lambda} \sum \frac{z}{z!} \lambda^r +$$

$$+ \lambda\, e^{-\lambda} \sum \frac{\lambda^z}{z!} = \lambda\, e^{-\lambda} e^{\lambda} + e^{-\lambda} e^{\lambda} = \lambda^2 + \lambda$$

Por tanto, $\sigma^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

3°) Desviación típica: $\sigma = \sqrt{\lambda}$

4°) Coeficiente de sesgo: $\gamma_1 = \frac{1}{\sqrt{\lambda}}$

Como $\gamma_1 > 0$, se trata de una distribución sesgada a la derecha.

5°) Coeficiente de curtosis: $\gamma_2 = \frac{1}{\lambda}$

4.7 AJUSTE DE UNA DISTRIBUCION EMPIRICA POR UNA DISTRIBUCION DE POISSON.- Por lo visto en 2.7 del presente capítulo, tendremos que la distribución de Poisson que mejor se aproxima a una distribución empírica, es la que tiene la misma media.

Por tanto, calcularemos la media experimental de las observaciones $\overline{x}$ y la distribución de Poisson ajustada será la que tenga por parámetro $\lambda = \overline{x}$

****

PROBLEMAS RESUELTOS

3.1 *Hallar la esperanza matemática y la varianza de la variable discreta X, dada por la siguiente ley de probabilidad:*

X	2	3	7
p	0,2	0,3	0,5

SOLUCION:

a) Esperanza matemática:

$$E[X] = \sum x_i p_i = 2.0,2 + 3.0,3 + 7.0,5 = 4,8$$

b) Varianza:

$$\sigma^2 = E\left[(X - 4,8)^2\right] = \sum (x_i - 4,8)^2 p_i = (2-4,8)^2 0,2 + (3-4,8)^2 0,3 +$$

$$+ (7-4,8)^2 0,5 = 1,568 + 0,972 + 2,42 = 4,96.$$

****

3.2 *Una variable aleatoria discreta viene dada por la siguiente ley de probabilidad:*

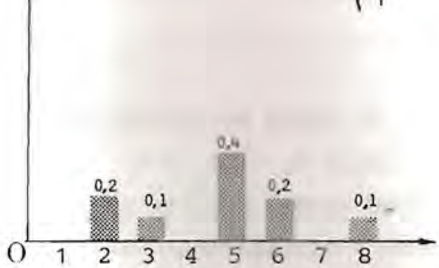
X	2	3	5	6	8
p	0,2	0,1	0,4	0,2	0,1

Hallar la función de distribución y representar la función de distribución y la ley de probabilidad.

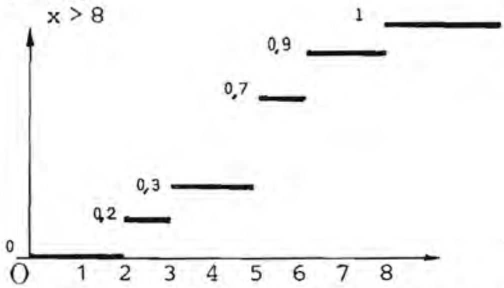
S O L U C I O N:

Esta claro que al ser la función de distribución una función monótona no decreciente contínua a la derecha y tal que $F(+\infty) = 1$, $F(-\infty) = 0$, para este caso será:

$$F(x) = \begin{cases} 0 & x \leq 2 \\ 0,2 & 2 < x \leq 3 \\ 0,3 & 3 < x \leq 5 \\ 0,7 & 5 < x \leq 6 \\ 0,9 & 6 < x \leq 8 \\ 1 & x > 8 \end{cases}$$



Gráfica de la ley de probabilidad



Gráfica de la función de distribución

****

3.3 *Una variable X viene definida por la siguiente ley de probabilidad:*

X	1	3	7
p	0,1	0,4	0,5

Calcular los momentos de 1°, 2°, y 3° orden respecto al origen.

S O L U C I O N:

Recordemos la expresión de momento de orden r respecto al origen

$$\alpha_r = E\left[X^r\right] = \sum x_i^r \, p_i$$

Por tanto tendremos:

a) Momento de primer orden respecto al origen:

$$\alpha_1 = E[X] = \Sigma\, x_i p_i = 1.0,1 + 3.0,4 + 7.0,5 = 4,8$$

b) Momento de segundo orden respecto al origen:

$$\alpha_2 = E[X^2] = \Sigma\, x_i^2 p_i = 1^2.0,1 + 3^2.0,4 + 7^2.0,5 = 28,2$$

c) Momento de tercer orden respecto al origen:

$$\alpha_3 = E[X^3] = \Sigma\, x_i^3 p_i = 1^3.0,1 + 3^3.0,4 + 7^3.0,5 = 182,4$$

****

3.4 *Una variable aleatoria X viene dada por la siguiente ley de probabilidad:*

X	3	4	5
p	0,1	0,3	0,6

Calcular los momentos de 1º, 2º, 3º y 4º orden respecto a la media o momentos centrales.

SOLUCION:

a) El momento de primer orden respecto a la media es siempre cero, ya que: $\mu_1 = E[X - \alpha_1] = E[X] - E[\alpha_1] = \alpha_1 - \alpha_1 = 0$

Para calcular los momentos centrales, calcularemos previamente los momentos respecto al origen y posteriormente utilizaremos las fórmulas vistas en 1.10.

$$\alpha_1 = E[X] = 3.0,1 + 4.0,3 + 5.0,6 = 4,5$$
$$\alpha_2 = E[X^2] = 3^2.0,1 + 4^2.0,3 + 5^2.0,6 = 20,7$$
$$\alpha_3 = E[X^3] = 3^3.0,1 + 4^3.0,3 + 5^3.0,6 = 96,7$$
$$\alpha_4 = E[X^4] = 3^4.0,1 + 4^4.0,3 + 5^4.0,6 = 459,9$$

b) Momento central de 2°orden:

$$\mu_2 = \alpha_2 - \alpha_1^2 = 20,7 - 4,5^2 = 0,45$$

c) Momento central de 3°orden:

$$\mu_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3 = 96,7 - 3.4,5.20,7 + 2(4,5)^3 = -0,5$$

d) Momento central de 4°orden:

$$\mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4 = 4,17$$

****

3.5 *Sea X una variable discreta cuya ley de probabilidad es:*

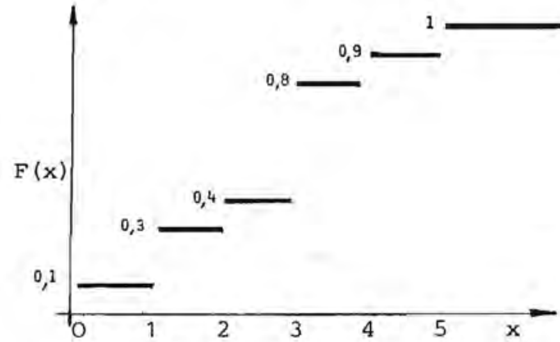
X	0	1	2	3	4	5
p	0,1	0,2	0,1	0,4	0,1	0,1

1°) Calcular y representar gráficamente la función de distribución.
2°) Calcular: a) $p(X < 4{,}5)$; b) $p(X \geq 3)$; c) $p(3 \leq X < 4{,}5)$.

SOLUCION:

1°)

$$F(x) = \begin{cases} 0 & x \geq 0 \\ 0{,}1 & 0 < x \leq 1 \\ 0{,}3 & 1 < x \leq 2 \\ 0{,}4 & 2 < x \leq 3 \\ 0{,}8 & 3 < x \leq 4 \\ 0{,}9 & 4 < x \leq 5 \\ 1 & x > 5 \end{cases}$$



2°) a) $p(X < 4{,}5) = p(X = 0) + p(X = 1) + \ldots + p(X = 4) = 0{,}9$

b) $p(X \geq 3) = 1 - p(X < 2) = 1 - \Big[p(X = 0) + p(X = 1) \Big] =$
$= 1 - (0{,}1 + 0{,}2) = 0{,}7.$

c) $p(3 \leq X < 4{,}5) = p(X = 3) + p(X = 4) = 0{,}4 + 0{,}1 = 0{,}5$

****

3.6 *Sea X una variable discreta que tiene por ley de probabilidad la siguiente: $p(X = r) = 1/8$, $r = 2,3, \ldots ,9$.*
Calcular:
a) $p(X > 6)$ b) $p(4 < X \leq 7)$ c) $p(X \geq 8)$
d) Función de distribución.
e) Representación de la ley de probabilidad y de la función de distribución.

SOLUCION:

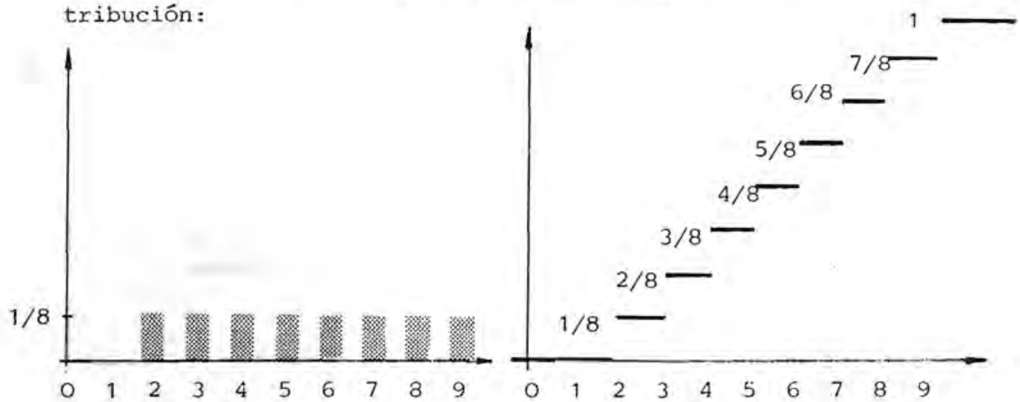
a) $p(X > 6) = p(X = 7) + p(X = 8) + p(X = 9) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$

b) $p(4 < X \leq 7) = p(X = 5) + p(X = 6) + p(X = 7) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$

c) $p(X \geq 8) = p(X = 8) + p(X = 9) = \frac{1}{8} + \frac{1}{8} = \frac{2}{8}$

d) $F(x) = p(X \leq x) = \sum_{r=2}^{x} \frac{1}{8} = \frac{x - 1}{8}$, para $x = 2,3, \ldots ,9$.

e) Representación de la ley de probabilidad y de la función de distribución:



Ley de probabilidad $p(X = r)$ — Función de distribución $F(x) = p(X \leq x)$

****

3.7 *Encontrar, (utilizando la tabla I del APENDICE), las probabilidades, $p(X = r)$; $p(X \geq r)$; $p(X < r)$, en las siguientes distribuciones binomiales:*

a) $n = 7$, $p = 0,3$, $r = 5$.
b) $n = 2$, $p = 0,1$, $r = 0$.
c) $n = 3$, $q = 0,9$, $r = 2$.

SOLUCION:

a) $n = 7$, $p = 0,3$, $r = 5$.

$p(X = 5) = 0,025$.

$p(X \geq 5) = p(X = 5) + p(X = 6) + p(X = 7) = 0,025+0,0036+0,0002 =$
$= 0,0288$.

$p(X < 5) = 1 - p(X \geq 5) = 1 - 0,0288 = 0,9712$.

b) $n = 2$, $p = 0,1$, $r = 0$.

$p(X = 0) = 0,81$.

$p(X \geq 0) = p(X = 0) + p(X = 1) + p(X = 2) = 1$

$p(X < 0) = 1 - p(X \geq 0) = 1 - 1 = 0$.

c) $n = 3$, $q = 0,9$, $r = 2$. Si $q = 0,9$ se sigue que $p = 0,1$.

$p(X = 2) = 0,027$.

$p(X \geq 2) = p(X = 2) + p(X = 3) = 0,027 + 0,001 = 0,028$.

$p(X < 2) = 1 - p(X \geq 2) = 1 - 0,028 = 0,9720$.

3.8 *La probabilidad de curación en un determinado tratamiento quirúrgico es de 0,65. Calcúlese la probabilidad de que en un grupo de 10 enfermos se curen la mitad.*

SOLUCION:

Representemos por A el suceso que un individuo cure despues del tratamiento quirúrgico y por $\bar{A}$ al suceso contrario.

Entonces: $p = p(A) = 0,65$ y $q = 1 - p = p(\bar{A}) = 0,35$

Sea X la variable que representa el número de individuos que curan tras el tratamiento. Es evidente que dicha variable sigue una distribución binomial de parámetros: $p = 0,65$, $n = 10$

Por tanto, $p(X = 5) = \binom{10}{5} 0,65^5 . 0,35^5 = 0,1536$

****

3.9 *Se supone que la probabilidad de nacer niño es 0,45. Calcular la probabilidad de que en una familia con ocho hijos, sean:*

a) Todos varones.
b) Al menos dos varones.
c) Ninguno sea varón.

SOLUCION:

Sea A el suceso nacer varón y $\bar{A}$ el suceso contrario. Por tanto,

$$p = p(A) = 0,45 \quad y \quad q = 1 - p = p(\bar{A}) = 0,55$$

Si representamos por X la variable que representa el número de varones que nacen en una familia de ocho hijos, esta claro que dicha variable obedece a una distribución binomial de parámetros:

$$n = 8, \qquad p = 0,45$$

a) Que todos sean varones:

$$p(X = 8) = \binom{8}{8} \; 0,45^8 . \; 0,55^0 = 0,0017.$$

b) Que al menos nazcan dos varones:

$$p(X \geq 2) = 1 - p(X < 2) = 1 - \left[p(X = 0) + p(X = 1) \right] =$$
$$= 1 - (0,0084 + 0,0548) = 0,9368.$$

c) Que ninguno sea varón:

$$P(X = 0) = 0,0084.$$

****

3.10 *Se sabe que la tercera parte de los enfermos que padecen hepatitis se curan al cabo de dos meses.*

Calcular la probabilidad de que de 5 pacientes con hepatitis se curen dos.

SOLUCION:

Esta claro que se trata de una distribución binomial de parámetros:

$$n = 5, \qquad p = 1/3$$

Por tanto, tendremos:

$$p(X = 2) = \binom{5}{2} (1/3)^2 . (2/3)^3 = 0{,}3292.$$

****

3.11 *La aplicación de un determinado tratamiento a enfermos de cirrosis, produce una cierta mejoría en el 70% de los casos.*

Si se aplica el tratamiento a 10 enfermos cirróticos, calcular:

a) ¿Cuál es la probabilidad de que mejoren 4?
b) ¿Cuál es la probabilidad de que al menos mejoren 3?

SOLUCION:

Se trata de una distribución binomial de parámetros: $p = 0{,}7$; $n=10$

a) Probabilidad de que mejoren 4:

$$p(X = 4) = \binom{10}{4} 0{,}7^4 . 0{,}3^6 = 0{,}0368$$

b) Probabilidad de que al menos mejoren 3:

$$p(X \geq 3) = 1 - p(X < 3) = 1 - \left[p(X = 0) + p(X = 1) + p(X = 2) \right] =$$
$$= 1 - (0 + 0{,}0001 + 0{,}0014) = 0{,}9985.$$

****

3.12 *La probabilidad de que un individuo vacunado contra una determinada enfermedad la contraiga es 0,2.*

De un grupo de 8 individuos vacunados, ¿cuál es la probabilidad de que:

a) Unicamente contraiga la enfermedad un solo inidividuo.
b) Al menos dos contraigan la enfermedad.
c) Todos contraigan la enfermedad.

SOLUCION:

Representemos por A al suceso "contraer la enfermedad", y por $\overline{A}$ al suceso contrario. Por tanto: $p = p(A) = 0{,}2$ y $q = p(\overline{A}) = 0{,}8$

Si representamos por X a la variable que representa el número de individuos que estando vacunados contraen la enfermedad, dicha variable obedece a una ley de probabilidad binomial de parámetros:

$$n = 8, \qquad p = 0,2.$$

a) Que contraiga la enfermedad solamente uno:

$$p(X = 1) = \binom{8}{1} 0,2^1 \cdot 0,8^{8-1} = 0,3355.$$

b) Que al menos dos contraigan la enfermedad:

$$p(X \geq 2) = 1 - p(X < 2) = 1 - \left[p(X = 0) + p(X = 1) \right] =$$
$$= 1 - (0,1678 + 0,3355) = 0,4967.$$

c) Que todos contraigan la enfermedad.

$$p(X = 8) = \binom{8}{8} 0,2^8 = 0,00000256.$$

****

3.13 *La probabilidad de que un estudiante obtenga el título de Licenciado en Ciencias Biológicas es 0,3.*

Calcular la probabilidad de que de un grupo de 7 estudiantes:

a) Ninguno de los 7 finalice los estudios.
b) Finalicen los 7.
c) Al menos 2 acaben la Licenciatura.
d) Solo finalice uno la Licenciatura.

SOLUCION:

Si representamos por A el suceso "obtener la Licenciatura", el suceso contrario será $\bar{A}$. Por tanto, $p = p(A) = 0,3$ y $q = p(\bar{A}) = 0,7$

Si representamos por X la variable que expresa el número de estudiantes que obtienen la Licenciatura, esta claro que se trata de una variable que sigue una distribución binomial de parámetros:

$$n = 7, \qquad p = 0,3.$$

a) Que ninguno de los 7 finalicen los estudios:

$$p(X = 0) = \binom{7}{0} 0,3^0 \cdot 0,7^7 = 0,082.$$

b) Que finalicen los 7:

$$p(X = 7) = \binom{7}{7} 0,3^7 \cdot 0,7^0 = 0,00021$$

c) Que al menos dos acaben:

$$p(X \geq 2) = 1 - p(X < 2) = 1 - \left[p(X = 0) + p(X = 1) \right] =$$

$$= 1 - (0{,}082 + 0{,}2471) = 0{,}671$$

d) Que solo obtenga la Licenciatura un estudiante:

$$p(X = 1) = \binom{7}{1} 0{,}3^1 . 0{,}7^6 = 0{,}2471$$

****

3.14 *Una máquina automática fabrica tabletas antitérmicas siendo la probabilidad de encontrar una defectuosa 0,02.*

Si las tabletas se colocan en tubos de 30 pastillas, calcular la probabilidad de que al adquirir un tubo no contenga ninguna defectuosa.

SOLUCION:

Si representamos por X la variable que expresa el número de tabletas defectuosas contenidas en un tubo de 30 pastillas, dicha variable obedece a una distribución binomial de parámetros:

$$n = 30, \qquad p = 0{,}02.$$

Por tanto la probabilidad pedida será:

$$p(X = 0) = \binom{30}{0} 0{,}02^0 . 0{,}98^{30} = 0{,}545.$$

****

3.15 *De 500 unidades de un producto farmaceútico, tal que cada unidad lleva dos precintos, se han observado las siguientes frecuencias de precintos rotos:*

nºde precintos rotos	*0*	*1*	*2*
nºde piezas	*220*	*160*	*120*

Ajustar una distribución binomial y calcular las frecuencias absolutas teóricas.

SOLUCION:

Calculemos la media observada:

$$\overline{x} = \frac{0.220 + 1.160 + 2.120}{500} = \frac{400}{500} = 0{,}8$$

Como la media de la distribución binomial es np, es decir 2p, resulta: $2p = 0{,}8 \qquad p = 0{,}4$. Por tanto, aproximaremos mediante un distribución binomial de parámetros: $n = 2, \quad p = 0{,}4$

Si representamos por X la variable que expresa el número de precintos rotos, podemos calcular las frecuencias absolutas teóricas, sin mas que multiplicar 500 por las probabilidades de o, 1, 2 roturas.

$$p(X = 0) = \binom{2}{0} 0{,}4^0 . 0{,}6^2 = 0{,}36 \quad \Longrightarrow \quad 500.\ 0{,}36 \simeq 180$$

$$p(X = 1) = \binom{2}{1} 0{,}4^1 . 0{,}6^1 = 0{,}48 \quad \Longrightarrow \quad 500.\ 0{,}48 \simeq 240$$

$$p(X = 2) = \binom{2}{2} 0{,}4^2 . 0{,}6^0 = 0{,}16 \quad \Longrightarrow \quad 500.\ 0{,}16 \simeq 80$$

Por tanto, las frecuencias absolutas teóricas serán:

n°de precintos rotos	0	1	2
n°de piezas	180	240	80

****

3.16 *Una urna contiene 20 bolas, de las que 16 son blancas y 4 negras. Se efectuan tres extracciones consecutivas sin reemplazamiento de la urna. Si representamos por X el número de bolas blancas extraidas, encontrar la ley de probabilidad de X.*

SOLUCION:

La variable X obedece a una distribución hipergeométrica de parámetros: $N = 20$, $n = 3$, $p = \frac{16}{20} = 0{,}8$.

Por tanto, tendremos:

$$p(X = r) = \frac{\binom{16}{r}\binom{4}{3-r}}{\binom{20}{3}}$$

Particularizando para valores de $r = 0,1,2,3$, se tiene:

$$p(X = 0) = \frac{\binom{16}{0}\binom{4}{3}}{\binom{20}{3}} = 0{,}0035.$$

$$p(X = 1) = \frac{\binom{16}{1}\binom{4}{2}}{\binom{20}{3}} = 0{,}0842$$

$$p(X = 2) = \frac{\binom{16}{2}\binom{4}{1}}{\binom{20}{3}} = 0{,}4211$$

$$p(X = 3) = \frac{\binom{16}{3}\binom{4}{0}}{\binom{20}{3}} = 0{,}4912$$

Observese que $p(X = 0) + p(X = 1) + p(X = 2) + p(X = 3) = 1$

****

3.17 *Una jaula de laboratorio contiene 15 cobayas, de las que 6 son blancas y 9 son pardas. Un ayudante de laboratorio con los ojos vendados extrae sin reemplazamiento 4 cobayas de la jaula.*

Calcular la probabilidad de que solamente una de las cobayas sea blanca.

S O L U C I O N:

Se trata de una distribución hipergeométrica de parámetros:

$$N = 15, \qquad n = 4, \qquad p = \frac{6}{15} = 0,4.$$

Por tanto, la probabilidad de que solamente una de las cobayas sea blanca será:

$$p(X = 1) = \frac{\binom{6}{1}\binom{9}{3}}{\binom{15}{4}} = 0,37.$$

****

3.18 *Por prescripción facultativa un enfermo debe hacer una toma de tres píldoras de un determinado medicamento. De las 25 píldoras que contiene el envase 5 están en malas condiciones. ¿Cuál es la probabilidad de que de las tres píldoras de la toma al menos una esté en malas condiciones?.*

S O L U C I O N:

Se trata de una distribución hipergeométrica de parámetros:

$$N = 25, \qquad n = 3 \qquad p = \frac{20}{25}$$

Por tanto: $N.p = 25\,\frac{20}{25} = 20;$ $\qquad N.q = 25\,\frac{5}{25} = 5$

Luego, la probabilidad de que las tres píldoras de la toma sean buenas será:

$$p(X = 3) = \frac{\binom{20}{3}\binom{5}{0}}{\binom{25}{3}} = 0,4956.$$

Y la probabilidad pedida será: $1 - 0,4956 = 0,5044.$

****

3.19 *La probabilidad de que se produzca un choque anafiláctico al suministro de un suero en un individuo es de 0,002. Sea X la variable aleatoria que representa el número de choques anafilácticos en un grupo de 1200 enfermos a los que se les ha suministrado el suero.*
Calcular: a) $p(X \leq 5)$; b) $p(X = 7)$.

SOLUCION:

Un enfermo al que se le suministra el suero puede:

- producirle un choque, con probabilidad $p = 0{,}002$
- no producirle choque, con probabilidad $q = 1 - p = 0{,}998$

La variable X sigue de modo trivial una ley binomial. Ahora bien, al ser $n = 1200$ muy grande y $p = 0{,}002$ muy pequeño, obtenemos una buena aproximación de la distribución binomial mediante la distribución de Poisson, que tiene por parámetro:

$$\lambda = n\,p = 1200.0{,}002 = 2{,}4.$$

a) $p(X \leq 5) = p(X = 0) + p(X = 1) + p(X = 2) + p(X = 3) + p(X = 4) + p(X = 5) = 0.0907 + 0{,}2177 + 0{,}2613 + 0{,}2090 + 0{,}1254 + 0{,}0602 = 0{,}9643.$

b) $p(X = 7) = 0{,}0083.$

****

3.20 *Un almacén farmaceútico recibió 1000 botellines de suero fisiológico. La probabilidad de que al transportar las botellas se rompa una es igual a 0,003. Hallar la probabilidad de que al desembalar los botellines, resulten:*
a) Menos de dos rotas.
b) Por lo menos una rota.
c) Exactamente dos rotas.

SOLUCION:

Observese que al ser $n = 1000$ muy grande y $p = 0{,}003$ muy pequeño y el producto $\lambda = n.p = 3$, es decir constante, se trata de una distribución de Poisson de parámetro $\lambda = 3$.

Si representamos por X el número de botellines rotos en el transporte, tendremos:

a) $p(X < 2) = p(X = 0) + p(X = 1) = 0{,}1992.$

b) $p(X \geq 1) = 1 - p(X < 1) = 1 - p(X = 0) = 1 - 0{,}04979 = 0{,}95.$

c) $p(X = 2) = 0{,}2245.$

****

3.21 *Se supone que en un determinado país la proporción de individuos albinos es de 0,005. Calcular la probabilidad de que elegida una muestra de la citada población de tamaño 1000, se presenten los siguientes casos:*
a) Ningun individuo sea albino.
b) Haya menos de 2 individuos albinos.
c) Al menos se encuentren 3 individuos albinos.

SOLUCION:

Al ser $p = 0{,}005$ y $n = 1000$ y $\lambda = n\,p = 5$ constante, se trata de una distribución de Poisson de parámetro $\lambda = 5$.

Por tanto, tendremos:

a) Que ningun individuo sea albino: $p(X = 0) = 0{,}0067.$

b) Que haya menos de dos individuos albinos:

$$p(X < 2) = p(X = 0) + p(X = 1) = 0{,}0067 + 0{,}0337 = 0{,}0404.$$

c) Que al menos se encuentren 3 individuos albinos:

$$p(X \geq 3) = 1 - p(X < 3) = 1 - \left[p(X = 0) + p(X = 1) + p(X = 2) \right] =$$
$$= 1 - (0{,}0067 + 0{,}0337 + 0{,}0842) = 0{,}8754.$$

****

3.22 *Sea X una variable aleatoria que sigue una distribución de Poisson y que representa el número de llamadas por minuto del teléfono de urgencia de un hospital. Calculado el número medio de llamadas por minuto, entre las 10 y las 12, se obtuvo 1,2.*

Calcular la probabilidad de que en un minuto determinado se produzcan:
a) dos llamadas
b) al menos una llamada
c) ninguna llamada
d) dos o tres llamadas
e) mas de dos llamadas

SOLUCION:

El número X de llamadas recibidas por minuto es una variable aleatoria que sigue la ley de Poisson. El valor del parámetro λ es el número medio de llamadas recibidas, es decir: $\lambda = 1,2$.

a) Probabilidad de recibir dos llamadas: $p(X = 2) = 0,2169$.

b) Probabilidad de recibir al menos una llamada:

$$p(X \geq 1) = 1 - p(X = 0) = 1 - 0,3012 = 0,6988.$$

c) Probabilidad de no recibir ninguna llamada: $p(X = 0) = 0,3012$.

d) Probabilidad de recibir dos o tres llamadas:

$$p(X = 2) + p(X = 3) = 0,2169 + 0,0867 = 0,3036.$$

e) Probabilidad de recibir más de dos llamadas:

$$p(X > 2) = 1 - \left[p(X = 0) + p(X = 1) + p(X = 2)\right] =$$
$$= 1 - (0,3012 + 0,3614 + 0,2169) = 0,1205.$$

****

3.23 *Tras una larga serie de estudios experimentales, se ha llegado a determinar que el número medio de una determinada especie de bacterias por centímetro cúbico contenida en el agua de la superficie de un lago es igual a 4.*

¿Calcular la probabilidad de no encontrar una sola bacteria en una gota de agua?. Se supone que la gota de agua es igual a 1/8 de centímetro cúbico.

SOLUCION:

Se trata de una distribución de Poisson de parámetro $\lambda = 4 \, \frac{1}{8} = 0,5$. Por tanto, si representamos por X la variable que expresa el número de bacterias contenidas en la gota de agua, tendremos que:

$$p(X = 0) = 0,6065.$$

****

3.24 *Un aparato de Rayos X algo antiguo, tiene una probabilidad de 0,01 de no detectar una lesión de pulmon existente durante la revisión de un paciente.*

Calcular la probabilidad de que al revisar 100 pacientes, el aparato no detecte el paso de 2 enfermos con lesión de pulmón.

SOLUCION:

Designemos por X la variable que representa el número de individuos que padeciendo lesión de pulmón, no es detectado por el aparato de Rayos X. Esta claro, que al ser $p = 0{,}01$ pequeño y $n = 100$ grande y el producto $n.p = 1$ constante, entonces, X será una variable aleatoria que sigue una distribución de Poisson de parámetro $\lambda = 1$.

La probabilidad pedida será: $p(X = 2) = 0{,}1839$.

****

3.25 *Se sabe que la probabilidad de que un individuo reaccione desfavorablemente tras la inyección de una vacuna es de 0,002.*

Determinar la probabilidad de que en un grupo de 2000 personas vacunadas haya como mucho tres que reaccionen desfavorablemente.

SOLUCION:

Representemos por A al suceso que un individuo reaccione desfavorablemente tras la vacuna y por $\overline{A}$ el suceso contrario.

Entonces: $p = p(A) = 0{,}002$ y $q = 1 - p = p(\overline{A}) = 0{,}998$

Si representamos por X la variable que expresa el número de personas que reaccionan desfavorablemente, observamos que se trata de una variable aleatoria que obedece a una distribución binomial de parámetros: $n = 2000$ y $p = 0{,}002$.

Ahora bien, como p es pequeño y n es grande y se verifica que $\lambda = np$ es constante e igual a 4, entonces la distribución anterior puede aproximarse mediante una distribución de Poisson de parámetro $\lambda = 4$.

La probabilidad de que 3 personas a lo sumo reaccionen desfavorablemente será:

$$p(X \leq 3) = p(X = 0) + p(X = 1) + p(X = 2) + p(X = 3) =$$
$$= 0{,}0183 + 0{,}0733 + 0{,}1465 + 0{,}1954 = 0{,}433$$

****

3.26 *La tabla adjunta muestra una distribución estadística en la que se expresa el número de días en los que ocurre 0,1,2,3,4,5 accidentes en una gran ciudad .*

1°) Calcular el número medio de accidentes por día, en el periodo de los 60 días considerados.

2°) *Ajustar la distribución anterior mediante una distribución de Poisson.*

3°) *Calcular con arreglo a la distribución ajustada, el número de días teórico en los que habrá 0,1,2,3,4,5 accidentes.*

número de accidentes	*número de días*
0	24
1	17
2	8
3	7
4	2
5	2

SOLUCION:

1°) El número medio de accidentes por día será:

$$\bar{x} = \frac{0.24 + 1.17 + 2.8 + 3.7 + 4.2 + 5.2}{60} = 1,2$$

2°) Para ajustar esta distribución empírica a una distribución de Poisson, hemos de identificar $\bar{x} = \lambda$. Es decir ajustaremos mediante una distribución de Poisson de $\lambda = 1,2$.

Calculemos las probabilidades para $X = 0,1,2,3,4,5$.

$p(X = 0) = 0,3012$	$p(X = 3) = 0,0867$
$p(X = 1) = 0,3614$	$p(X = 4) = 0,0260$
$p(X = 2) = 0,2169$	$p(X = 5) = 0,0062$

3°) Para calcular el número teórico de dias en los que habrá r accidentes, multiplicaremos estas probabilidades por $n = 60$.

r	frec. empírica	frec. teórica
0	24	18
1	17	22
2	8	13
3	7	5
4	2	2
5	2	0

Obsérvese que la aproximación es bastante mala, como veremos en próximos capítulos al estudiar la bondad del ajuste.

****

4- distribuciones continuas

4 - DISTRIBUCIONES CONTINUAS

1. GENERALIDADES

1.1 VARIABLE ALEATORIA CONTINUA.- En 1.1 del capítulo 3, vimos el concepto general de variable aleatoria, que decía así:
En un espacio probabilístico (E,**A**,p), llamaremos *variable aleatoria* y representaremos por X a una función del conjunto de los sucesos elementales de un experimento aleatorio, en el cuerpo de los números reales. De tal forma que dado un número real cualquiera x, si consideramos el conjunto de todos los sucesos elementales para los que la función X toma valores menores o iguales a x, ese conjunto ha de ser un suceso del algebra A.

Ahora bien, diremos que la variable aleatoria X es *contínua*, cuando el conjunto imagen de la función X es un intervalo de la recta real y por tanto infinito.

1.2 FUNCION DE DENSIDAD.- En un espacio probabilístico (E,A,p), llamaremos *función de densidad* de la variable X y representaremos por f(x), a una función que nos da la probabilidad para cada punto.

Es decir: $f(x) = p(X = x)$

Obsérvese, que la función de densidad para una variable aleatoria continua, equivale a la ley de probabilidad para una variable aleatoria discreta.

La función de densidad para el caso de variable aleatoria contínua es la curva límite del polígono de frecuencias.

1.3 FUNCION DE DISTRIBUCION.- En un espacio probabilístico (E,**A**,p) llamaremos *función de distribución* de la variable X y representaremos por $F(x)$ a la función:

$$F(x) = p(X \leq x)$$

Si X es una variable aleatoria contínua, la función de distribución será:

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

1.4 PROPIEDADES DE LA FUNCION DE DISTRIBUCION.- Como vimos en el apartado 1.5 del capítulo anterior, las propiedades de la función de distribución son:

1^a) $F(+\infty) = 1$. En efecto, ya que $F(+\infty) = \int_{-\infty}^{\infty} f(t)dt = 1$

2^a) $F(-\infty) = 0$. En efecto, ya que $F(-\infty) = \int_{-\infty}^{-\infty} f(t)dt = 0$

3^a) Si $x_1 \leq x_2 \implies F(x_1) \leq F(x_2)$

En efecto, pues $F(x_2) = \int_{-\infty}^{x_2} f(t)dt = \int_{-\infty}^{x_1} f(t)dt + \int_{x_1}^{x_2} f(t)dt$; $F(x_2) = F(x_1) + \int_{x_1}^{x_2} f(t)\,dt$.

Por tanto, $F(x_1) \leq F(x_2)$

Como consecuencia de esta propiedad, se tiene:

$$p(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

4^a) $F(x)$ es una función continua a la derecha, cualquiera que sea el valor de x.

En efecto, pues: $F(x_1+h) - F(x_1) = p(x_1 < X \leq x_1+h)$ y cuando $h \to 0$ se sigue que: $F(x_1+h) - F(x_1) \to 0$

Veamos ahora que si la función de densidad $f(x)$ es contínua para todo x, se tiene que $F'(x) = f(x)$

En efecto,

$$F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \to 0} \frac{1}{h} \int_{x}^{x+h} f(t)dt = f(x)$$

Luego la función de densidad es la derivada de la función de distribución.

Es decir: $$F'(x) = f(x)$$

Obsérvese que la función de distribución es el límite del polígono de frecuencias relativas acumulativo, cuando el tamaño del intervalo de clase tiende a cero.

1.5 ESPERANZA MATEMATICA DE UNA VARIABLE ALEATORIA CONTINUA.- Sea X una variable aleatoria contínua que tiene por imagen el intervalo $[a,b]$ y para la que se conocen las probabilidades puntuales mediante la función de densidad $f(x)$.

Llamaremos *esperanza matemática, valor esperado o media* de la variable aleatoria contínua X al valor de la siguiente integral

$$\mu = E\left[X\right] = \int_a^b x\ f(x)dx$$

Si la variable X puede recorrer toda la recta real, la esperanza matemática de X será:

$$\mu = E\left[X\right] = \int_{-\infty}^{\infty} x\ f(x)dx$$

1.6 PROPIEDADES DE LA ESPERANZA MATEMATICA.- Análogamente a como vimos para el caso de variable aleatoria discreta en el apartado 1.7 del capítulo anterior, para el caso de variable aleatoria contínua se comprueban las siguientes propiedades:

1ª) $E\left[c\right] = c$

2ª) $E\left[X + Y\right] = E\left[X\right] + E\left[Y\right]$

3ª) $E\left[c\ X\right] = c\ E\left[X\right]$

4ª) $E\left[X\ .\ Y\right] = E\left[X\right]\ .\ E\left[Y\right]$ supuesto que X e Y son variables aleatorias independientes.

1.7 MOMENTOS RESPECTO AL ORIGEN.- Llamaremos *momento de orden r respecto al origen* de la variable aleatoria X y representaremos -

por α_r, a la esperanza matemática de X^r.

Es decir: $\alpha_r = E\left[X^r\right]$

Si X es una variable aleatoria contínua, tendremos:

$$\alpha_r = E\left[X^r\right] = \int_{-\infty}^{\infty} x^r f(x)\,dx$$

De lo que se deduce que: $\alpha_1 = E\left[X\right] = \int_{-\infty}^{\infty} x f(x)\,dx$

Es decir, la esperanza matemática coincide con el momento de primer orden respecto al origen, que llamaremos media poblacional y representaremos por μ.

1.8 MOMENTOS RESPECTO A LA MEDIA.- Llamaremos *momento respecto a la media o momento central de orden r* de la variable aleatoria X y representaremos por μ_r, a la esperanza matemática de $(X-\alpha_1)^r$.

Es decir: $\mu_r = E\left[(X-\alpha_1)^r\right]$

Para el caso de variable aleatoria contínua, tendremos:

$$\mu_r = E\left[(X-\alpha_1)^r\right] = \int_{-\infty}^{\infty} (x-\alpha_1)^r f(x)\,dx$$

El momento de segundo orden es igual a la varianza. Es decir:

$$\sigma^2 = \mu_2 = E\left[(X-\alpha_1)^2\right] = \int_{-\infty}^{\infty} (x-\alpha_1)^2 f(x)\,dx$$

1.9 RELACION ENTRE AMBOS MOMENTOS.- En 1.10 del capítulo anterior, vimos las relaciones que existen entre los momentos centrales y los momentos respecto al origen, que son igualmente válidas para el caso de variable aleatoria contínua.

Estas relaciones son:

$$\begin{cases} \mu_r = \alpha_r - \binom{r}{1}\alpha_1\,\alpha_{r-1} + \binom{r}{2}\alpha_1^2\,\alpha_{r-2} - \cdots \pm \alpha_1^r \\ \alpha_r = \mu_r + \binom{r}{1}\mu_1\,\mu_{r-1} + \binom{r}{2}\mu_1^2\,\mu_{r-2} + \cdots + \mu_1^r \end{cases}$$

1.10 TEOREMA DE TCHEBYCHEFF.- Consideremos una variable aleatoria X que tenga de media y desviación típica μ y σ respectivamente y sea k un número arbitrario y positivo. Entonces se verifica que:

$$p\left[\,|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$$

Para $k < 1$, el teorema es evidente, pues cualquier probabilidad es menor o igual a la unidad.

Para $k \geq 1$, el teorema proporciona una importante información acerca de como disminuye la masa de la distribución a grandes distancias del valor medio μ, ya que la cantidad de masa situada fuera del intervalo $(\mu - k\sigma\ ,\ \mu + k\sigma)$ es menor que $1/k^2$.

El teorema es totalmente válido para cualquier distribución, pero solo haremos la demostración para una distribución de variable aleatoria contínua.

Demostración:

Queremos demostrar que $p\left[\,|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$

Sabemos que: $\sigma^2 = E\left[(X - \mu)^2\right] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx =$

$$= \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2 f(x)dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x - \mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)dx$$

Quitando la segunda integral, tendremos:

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2\, f(x)\, dx + \int_{\mu+k\sigma}^{+\infty} (x - \mu)^2 f(x)\, dx.$$

Ahora bien, como estamos bajo la hipotesis de que $|x - \mu| \geq k\sigma$ entonces $(x - \mu)^2 \geq k^2\sigma^2$, con lo que podemos minorar la expresión anterior, resultando:

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)dx \geq$$

$$\geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)dx + \int_{\mu+k\sigma}^{+\infty} k^2\sigma^2 f(x)dx =$$

$$= k^2\sigma^2 \left(\int_{-\infty}^{\mu-k\sigma} f(x)dx + \int_{\mu+k\sigma}^{\infty} f(x)dx \right) = k^2\sigma^2 \; p\left[|X-\mu| \geq k\sigma\right] .$$

Luego tenemos: $\sigma^2 \geq k^2\sigma^2 \; p\left[|X - \mu| \geq k\sigma\right]$

entonces dividiendo ambos miembros por $k^2\sigma^2$, que sabemos es distinto de cero, reulta:

$$\frac{1}{k^2} \geq p\left[|X - \mu| \geq k\sigma\right] \quad \text{o bien} \quad p\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$$

2. DISTRIBUCION NORMAL

2.1 DISTRIBUCION NORMAL.- DEFINICION.- Diremos que una variable aleatoria contínua X sigue una distribución normal de media μ y desviación típica σ y representaremos por $N(\mu,\sigma)$, si se cumplen las siguientes condiciones:

1ª) La variable X recorre el intervalo $(-\infty , +\infty)$

2ª) La probabilidad para cada punto viene definida por la siguiente expresión:

$$p(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad [1]$$

2.2 FUNCION DE DENSIDAD.- Acabamos de ver que la probabilidad para cada punto viene dada por la expresión [1]. Esta función es precisamente la función de densidad de la distribución normal de parámetros μ, α, que representaremos por $f(x)$.

Si representamos esta función, observamos lo siguiente:

1°) La función es simétrica respecto de la recta $x = \mu$.

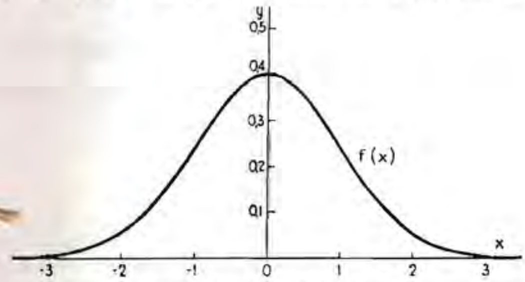
2°) La función f(x) tiene una asintota horizontal que es el eje x.

3°) Admite un máximo absoluto en el punto $\left(\mu , \frac{1}{\sigma\sqrt{2\pi}}\right)$

4°) Presenta dos puntos de inflexión para los valores de las abcisas: $x = \mu + \sigma$ y $x = \mu - \sigma$

5°) Como toda función de densidad, el valor del area encerrada bajo ella y el eje x es igual a la unidad.

Por tanto, la gráfica de la función f(x) será de la siguiente forma:



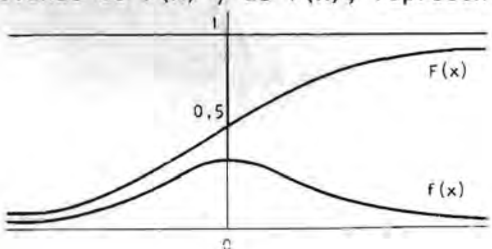
Función de densidad N(0 , 1)

2.3 FUNCION DE DISTRIBUCION.- Dada una variable aleatoria contínua X que sigue una distribución $N(\mu , \sigma)$, llamaremos función de distribución de la variable X y representaremos por F(x), a la siguiente expresión:

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt =$$

Es decir: $$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt .$$

Las gráficas cartesianas de F(x) y de f(x), representadas sobre los mismos ejes, son:



2.4 TIPIFICACION DE LA VARIABLE.- Llamaremos *tipificación de la variable* al paso de una variable aleatoria $X \in N(\mu, \sigma)$ a una variable $Y \in N(0, 1)$. Este paso consiste en:

1°) *Centrar*: es decir hacer la media nula.

2°) *Reducir*: hacer la desviación típica igual a la unidad.

Estos dos pasos se consiguen simultaneamente, sin mas que hacer el siguiente cambio de variable:

$$Y = \frac{X - \mu}{\sigma}$$

Para el caso de variable tipificada las funciones de densidad y de distribución se simplifican, quedando de la siguiente forma:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \Longrightarrow F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Existen tablas para la función de distribución de una variable aleatoria $N(0, 1)$ (Ver APENDICE, tabla III).

Facilita mucho el manejo de las tablas, el hecho de que la distribución normal es simétrica.

2.5 UN PROBLEMA RELACIONADO CON LA NORMAL.- Con frecuencia se presenta el problema de determinar la probabilidad de que una variable aleatoria X que obedece a una distribución $N(\mu, \sigma)$, se desvie de su media μ en menos de un cierto valor, que generalmente expresaremos como múltiplo de la desviación típica.

Es decir, se trata de calcular: $p\left[\,|X - \mu| > k\sigma\right]$

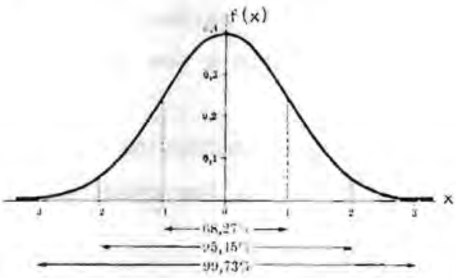
Si tipificamos, tendremos que $Y = \frac{X - \mu}{\sigma} \in N(0, 1)$, entonces

$$p\left[\,|X - \mu| > k\sigma\right] = p\left[\left|\frac{X - \mu}{\sigma}\right| > k\right] = p\left[\frac{X-\mu}{\sigma} < -k\right] + p\left[\frac{X-\mu}{\sigma} > k\right] =$$

$$= 2p\left[\frac{X - \mu}{\sigma} > k\right] = \frac{2}{\sqrt{2\pi}} \int_{k}^{\infty} e^{-\frac{y^2}{2}} dy$$

Buscando en las tablas se pueden calcular los valores de k, que para los casos más importantes, serán los siguientes:

El 50%	de las	observaciones	están	en	el	intervalo	$(\mu-0{,}68\sigma\ ,\ \mu+0{,}68\sigma)$
El 68,2%	"	"	"	"	"	"	$(\mu-\sigma\ ,\ \mu+\sigma)$
El 95%	"	"	"	"	"	"	$(\mu-1{,}96\sigma\ ,\ \mu+1{,}96\sigma)$
El 95,4%	"	"	"	"	"	"	$(\mu-2\sigma\ ,\ \mu+2\sigma)$
El 99%	"	"	"	"	"	"	$(\mu-2{,}58\sigma\ ,\ \mu+2{,}58\sigma)$
El 99,7%	"	"	"	"	"	"	$(\mu-3\sigma\ ,\ \mu+3\sigma)$
El 99,9%	"	"	"	"	"	"	$(\mu-3{,}29\sigma\ ,\ \mu+3{,}29\sigma)$



2.6 CARACTERISTICAS ESTADISTICAS.- Las características estadísticas de una distribución normal de parámetros μ y σ, son:

1°) Media aritmética: μ

2°) Varianza: σ^2

3°) Desviación típica: σ

4°) Sesgo: $\gamma_1 = 0$

5°) Curtosis: $\gamma_2 = 3$

2.7 AJUSTE DE UNA DISTRIBUCION EMPIRICA MEDIANTE UNA DISTRIBUCION NORMAL.- La idea del ajuste es análoga a lo expuesto para las distribuciones binomial y Poisson. De tal forma que la distribución normal que mejor se aproxima a una distribución empírica es la que tiene la misma media y la misma desviación típica.

Así pues, calcularemos la media de las observaciones $\overline{x}$ y la desviación típica s de la distribución empírica observada y la ley normal ajustada será, $N(\overline{x}\ ,\ s)$ y tendrá por función de densidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{x-\overline{x}}{s}\right)^2}$$

2.8 LA DISTRIBUCION NORMAL COMO LIMITE DE LA DISTRIBUCION BINOMIAL.- Cuando estudiamos la distribución binomial, vimos que el cálculo de las probabilidades se hace muy laborioso para valores grandes de n.

Para poder conseguir una buena aproximación mediante la ley normal, aplicaremos el siguiente teorema:

TEOREMA DE MOIVRE.- Si X es una variable aleatoria que obedece a una distribución binomial de parámetros n, p, entonces la variable $y = \dfrac{X - np}{\sqrt{npq}}$ obedece a una distribución N(0, 1) cuando $n \to \infty$.

En la práctica, la aproximación es bastante buena cuando se verifican simultaneamente las condiciones siguientes:

1°) $n > 30$

2°) $0,1 < p < 0,9$

Esta claro que cuanto mayor es n y p más se aproxima a 0,5, tanto mas óptima es la aproximación.

3. DISTRIBUCION χ^2 DE PEARSON

3.1 FUNCION GAMMA.- La integral definida de la forma

$$\int_0^\infty x^{n-1} \, e^{-x} \, dx$$

se conoce con el nombre de *función gamma* y se representa por $\Gamma(n)$.

Integrando n-1 veces por partes o utilizando las fórmulas de recurrencia obtendriamos:

$$\Gamma(n) = \int_0^\infty x^{n-1} \, e^{-x} \, dx = (n-1)!$$

3.2 PROPIEDADES DE LA FUNCION GAMMA.-

1ª) $\Gamma(1) = \Gamma(2)$

En efecto, ya que: $\Gamma(1) = (1 - 1)! = 0! = 1$
$\Gamma(2) = (2 - 1)! = 1! = 1$

2^a) $\Gamma(n+1) = n\,\Gamma(n)$

En efecto, pues:

$\Gamma(n+1) = (n+1-1)! = n! = n(n-1)! = n\,\Gamma(n)$.

3.3 DISTRIBUCION χ^2 DE PEARSON.- DEFINICION.- Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias $N(0\ ,\ 1)$ e independientes, entonces la expresión:

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

es una variable aleatoria que recibe el nombre de χ^2 de Pearson, (se lee chi-cuadrado o ji-dos).

3.4 FUNCION DE DENSIDAD.- La función de densidad de la variable aleatoria de la distribucion χ^2 es la siguiente:

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\,\Gamma(\frac{n}{2})}\, e^{-\frac{x}{2}}\, x^{(\frac{n}{2}-1)} & x \geq 0 \\[2ex] 0 & x < 0 \end{cases}$$

Obsérvese que la distribución "ji-cuadrado" solo depende del parámetro n.

Veamos a continuación algunas propiedades de esta función.

1^a) La variable aleatoria de una "chi-cuadrado", está definida en el intervalo $(0\ ,\ +\infty)$.

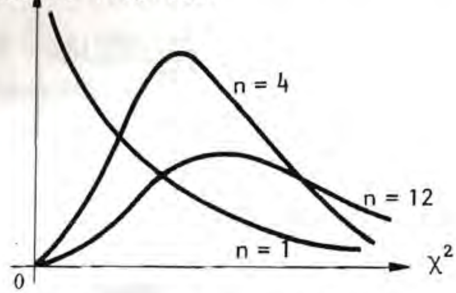
2^a) El numero de variables normales $N(0\ ,\ 1)$ e independientes, recibe el nombre de *grados de libertad* que representaremos por n.

3^a) La función f(x) es contínua para todo el dominio de definición.

4^a) La función f(x) presenta un único máximo para el valor n-2, supuesto $n \geq 2$

5ª) La función f(x) no es simétrica.

6ª) La función f(x) depende del número de grados de libertad n. En la figura se muestran algunas funciones de densidad de la distribución "chi-cuadrado" para distintos grados de libertad.



3.5 FUNCION DE DISTRIBUCION.- Llamaremos *función de distribución de la variable* χ^2 y representaremos por F(x) al valor de la siguiente integral:

$$F(x) = p(\chi^2 \leq x) = \int_0^x \frac{1}{2^{\frac{n}{2}}\,\Gamma(\frac{n}{2})}\, e^{-\frac{t}{2}}\, t^{(\frac{n}{2}-1)}\, dt$$

Como ejercicio compruebe el alumno, que la función F(x) cumple las propiedades de toda función de distribución.

Esta función esta tabulada, (Ver APENDICE, tabla IV) en la que aparece para distintas areas α de las colas de la derecha, las abcisas de los extremos izquierdos de las mismas según los distintos grados de libertad.

Para $n > 30$, se tiene que $T = \sqrt{2\,\chi^2} - \sqrt{2n-1}$ se distribuye segun la N(0 , 1).

3.6 CARACTERISTICAS ESTADISTICAS.-

1ª) Media aritmética: $\mu = n$

2ª) Varianza: $\sigma^2 = 2n$

3ª) Desviación típica: $\sigma = \sqrt{2n}$

3.7 TEOREMA DE LA ADICION.- Si X_1^2 es una "chi-cuadrado" con n_1 grados de libertad y X_2^2 es otra "chi-cuadrado" con n_2 grados de libertad e independientes entre si, entonces se construye a partir de estas otra "chi-cuadrado"

$$X^2 = X_1^2 + X_2^2$$

que tiene $n_1 + n_2$ grados de libertad.

3.8 APLICACION DE LA X^2 PARA ESTIMACIONES Y CONTRASTES ACERCA DE LA VARIANZA.- Una de las aplicaciones más importantes de la "chi-cuadrado", aparte del estudio de la bondad de un ajuste, es la utilización de esta distribución para la estimación de la varianza o para el constraste de una hipótesis acerca de la varianza, como veremos en próximos capítulos.

Sea σ^2 la varianza de una población normal y s^2 la varianza de una muestra aleatoria de tamaño n extraida de la citada población. Entonces $\frac{(n-1)\,s^2}{\sigma^2}$ es una variable aleatoria cuya distribución en el muestreo es una X^2 con (n-1) grados de libertad.

4. DISTRIBUCION t DE STUDENT

4.1 DEFINICION.- Sean $X^2, X_1^2, X_2^2, \ldots, X_n^2$, n+1 variables aleatorias independientes y N(0,1).

Llamaremos *variable t de la distribución de Student*, con n grados de libertad, a la siguiente expresión:

$$t = \frac{X}{\sqrt{\frac{1}{n}\,(X_1^2 + X_2^2 + \ldots + X_n^2)}}$$

Esta distribución fué obtenida por el matemático inglés Gosset, en el año 1907, el cual firmaba sus trabajos con el seudonimo de "student"

4.2 FUNCION DE DENSIDAD.- La función de densidad o de probabilidad de una variable t de Student con n grados de libertad, viene dada por la siguiente expresión:

$$t_n(x) = p(t = x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}$$

Estudiemos algunas propiedades de esta función:

1ª) El dominio de definición de $f(x)$ es $(-\infty, +\infty)$.

2ª) Para los distintos valores del parámetro n, la función $t_n(x)$ es campaniforme y simétrica.
Cuanto mayor es el número de grados de libertad más apuntada es la curva, tendiendo a la normal de parámetros $N(0, 1)$ para valores de n grandes.

3ª) Como el cálculo de las probabilidades es laborioso, la función t_n, está tabulada, (Ver APENDICE, tabla V).
Para valores de $n > 30$, se debe aproximar mediante una $N(0, 1)$.

4.3 FUNCION DE DISTRIBUCION.- Llamaremos *función de distribución* de una variable t de Student, con n grados de libertad y representaremos por $T_n(x)$, al valor de la siguiente integral:

$$T_n(x) = p(t \leq x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^{x} \left(1 + \frac{z^2}{n}\right)^{-\left(\frac{z+1}{2}\right)} dz$$

4.4 APLICACION DE LA t DE STUDENT PARA ESTIMACIONES Y CONTRASTES ACERCA DE LA MEDIA.- Sean μ y σ^2 la media y varianza respectivamente de una población normal y $\overline{x}$, s^2 la media y varianza respectivamente de una muestra aleatoria de tamaño n, extraida de la población anterior.

Entonces, $t = \dfrac{\overline{x} - \mu}{\frac{n\, s}{n-1}} \sqrt{n}$, es una variable aleatoria cuya distribución en el muestreo es una t de Student con n-1 grados de libertad.

5. DISTRIBUCION F DE FISHER-SNEDECOR

5.1 DEFINICION.- Sean χ_1^2 y χ_2^2 dos variables "chi cuadrado independientes y con n_1 y n_2 grados de libertad respectivamente. Llamaremos *variable* F *de la distribución de Fisher-Snedecor*, con n_1 y n_2 grados de libertad a la expresión:

$$F = \frac{\chi_1^2/n_1}{\chi_2^2/n_2}$$

5.2 FUNCION DE DENSIDAD.- La función de densidad de la variable F de Fisher-Snedecor con n_1, n_2 grados de libertad es:

$$f_{n_1,n_2}(x) = n_1^{n_1/2}\, n_2^{n_2/2}\, \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{x^{\frac{n_1}{2}-1}}{(n_2+n_1x)^{\frac{n_1+n_2}{2}}}$$

1°) El dominio de F es de 0 a ∞.

2°) Las tablas VI del APENDICE nos da los valores de $F_{\alpha,\, n_1,n_2}$ que representa el valor de la abcisa que deja a su derecha bajo la curva de densidad un area igual a α para los valores de $\alpha = 0{,}01$ y $\alpha = 0{,}05$.

5.3 FUNCION DE DISTRIBUCION.- Llamaremos *función de distribución* de una variable F de Fisher-Snedecor con n_1, n_2 grados de libertad y representaremos por $F_{n_1,n_2}(x)$, al valor de la siguiente integral:

$$F_{n_1,n_2}(x) = \int_0^x f_{n_1,n_2}(z)\, dz$$

5.4 APLICACION DE LA F DE FISHER-SNEDECOR AL CONTRASTE DE COMPARACION DE DOS VARIANZAS.- Sean σ_1^2 y σ_2^2 las varianzas de dos poblaciones normales $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ e independientes y queremos saber si existe diferencia significativa entre ellas.

Tomaremos una muestra aleatoria de tamaño n_1 de una población y otra muestra n_2 de la otra población. Las varianzas muestrales serán s_1^2 y s_2^2. Veamos que el cociente s_1^2/s_2^2 sigue una F de Fisher-Snedecor con n_1-1 y n_2-1 grados de libertad.

$$\left.\begin{array}{l} \text{Como } (n_1-1)\frac{s_1^2}{\sigma_1^2} \text{ sigue una } \chi^2_{n_1-1} \\ \text{y } \quad (n_2-1)\frac{s_2^2}{\sigma_2^2} \text{ sigue una } \chi^2_{n_2-1} \end{array}\right\} \Longrightarrow F = \frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

Si $\sigma_1^2 = \sigma_2^2$, resulta $F = s_1^2 / s_2^2$. c.q.d.

PROBLEMAS RESUELTOS

4.1 *Dada la función de distribución de una variable aleatoria contínua X, definida del siguiente modo:*

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \operatorname{sen} x & 0 < x \leq \pi/2 \\ 1 & x > \pi/2 \end{cases}$$

Calcular la función de densidad.

SOLUCION:

Como vimos en 1.4, la función de densidad de una variable aleatoria contínua es la derivada de la función de distribución.

Por tanto,

$$f(x) = F'(x) = \begin{cases} 0 & x \leq 0 \\ \cos x & 0 < x \leq \pi/2 \\ 0 & x > \pi/2 \end{cases}$$

****

4.2 *Dada la función de densidad de una variable aleatoria continua, definida del siguiente modo:*

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \cos x & 0 < x \leq \pi/2 \\ 0 & x > \pi/2 \end{cases}$$

Calcular la función de distribución de la citada variable.

SOLUCION:

La función de distribución viene dada por la siguiente expresión:

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(x)\, dx.$$ Por tanto,

1°) - Si $x \leq 0$, como $f(x) = 0$, se sigue $F(x) = \int_{-\infty}^{x} 0.dx = 0$

2°) - Si $0 < x \leq \pi/2$, como $f(x) = \cos x$, se deduce:

$$F(x) = \int_{-\infty}^{0} 0.dx + \int_{0}^{x} \cos x \, dx = \text{sen } x.$$

3°) - Si $x > \pi/2$, como $f(x) = 0$, se sigue:

$$F(x) = \int_{-\infty}^{0} 0 \, dx + \int_{0}^{\pi/2} \cos x \, dx + \int_{\pi/2}^{x} 0 \, dx = \Big[\text{sen } x\Big]_{0}^{\pi/2} = 1$$

****

4.3 *La función de densidad de una variable aleatoria contínua viene definida por la siguiente expresión:*

$$f(x) = \frac{C}{e^{x} + e^{-x}}$$

Hallar el valor de la constante C, para que $f(x)$ sea una auténtica función de densidad.

SOLUCION:

Se tiene que cumplir $\displaystyle\int_{-\infty}^{+\infty} \frac{C}{e^{x} + e^{-x}} dx = 1$

Es decir: $C = \dfrac{1}{\displaystyle\int_{-\infty}^{+\infty} \frac{dx}{e^{x} + e^{-x}}}$, Calculemos primeramente el valor de la integral del denominador.

$\displaystyle\int \frac{dx}{e^{x} + e^{-x}} = \int \frac{e^{x}}{1 + e^{2x}} dx = \text{arc tg } e^{x}$. Por tanto, la integral definida, será $\displaystyle\int_{-\infty}^{+\infty} \frac{dx}{e^{x} + e^{-x}} = \Big[\text{arc tag } e^{x}\Big]_{-\infty}^{+\infty} = \frac{\pi}{2}$

$$\text{Luego} \quad C = \frac{1}{\frac{\pi}{2}} = \frac{2}{\pi}$$

****

4.4 *La función de densidad de una variable aleatoria contínua X, viene definida por la función:*

$$f(x) = \begin{cases} x & \text{para} \quad 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

Calcular:

a) *Esperanza matemática de la variable X.*
b) *Varianza.*

SOLUCION:

a) Esperanza matemática:

Como $\alpha_1 = E[X] = \int_a^b x\, f(x)\, dx.$

Entonces $E[X] = \int_0^1 x\, x\, dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}$

b) Varianza:

Como $\sigma^2 = E\left[(X - \alpha_1)^2\right] = E\left[X^2\right] - \alpha_1^2$

Entonces: $\sigma^2 = E\left[X^2\right] - (\frac{1}{3})^2 = \int_0^1 x^2\, x\, dx - \frac{1}{9} =$

$$= \left[\frac{x^4}{4}\right]_0^1 - \frac{1}{9} = \frac{5}{36}$$

****

4.5 *Sea X una variable aleatoria cuya función de distribución es la siguiente:*

$$F(x) = \begin{cases} 0 & \text{para} \quad x \leq -2 \\ \frac{x}{4} + \frac{1}{2} & \text{para} \quad -2 < x \leq 2 \\ 1 & \text{para} \quad x > 2 \end{cases}$$

a) *Calcular la esperanza matemática de X.*
b) *Hallar la varianza.*
c) $p(X \leq 1)$
d) $p(1 < X \leq 2)$
e) $p(X > 3)$

SOLUCION:

En primer lugar calculemos la función de densidad.

$$f(x) = \begin{cases} 0 & \text{para} \quad x \leq -2 \\ 1/4 & \text{para} \quad -2 < x \leq 2 \\ 0 & \text{para} \quad x > 2 \end{cases}$$

a) Esperanza matemática:

$$E[X] = \int_{-2}^{2} x\, f(x)\, dx = \int_{-2}^{2} x \,\frac{1}{4}\, dx = 0$$

b) Varianza:

$$\sigma^2 = E\left[(X-0)^2\right] = \int_{-2}^{2} x^2 \,\frac{1}{4}\, dx = \frac{4}{3}$$

c) $p(X \leq 1) = F(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$

d) $p(1 < X \leq 2) = F(2) - F(1) = 1 - \frac{3}{4} = \frac{1}{4}$

e) $p(X > 3) = 1 - p(X \leq 3) = 1 - F(3) = 1 - 1 = 0$

****

4.6 *Dada la función*

$$f(x) = \begin{cases} \dfrac{C}{1+x^2} & 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

a) *Calcular el valor de C para que $f(x)$ sea una función de densidad.*
b) *Hallar la función de distribución.*
c) *Esperanza matemática de la variable X que tiene por función de densidad $f(x)$.*

S O L U C I O N:

a) Para que f(x) sea una función de densidad ha de ocurrir que $\int_0^1 f(x)dx = 1$. Por tanto $\int_0^1 \frac{C}{1+x^2}\, dx = C\left[\text{arc tg } x\right]_0^1 =$

$$= C\left(\frac{\pi}{4} - 0\right) = \frac{C\pi}{4} = 1. \quad \text{Luego} \quad C = \frac{4}{\pi}$$

b) Función de distribución:

- Para $x < 0$, $F(x) = 0$
- Para $0 \leq x \leq 1$ $\quad F(x) = \int_0^x f(x)\, dx = \int_0^x \frac{4/\pi}{1+x^2}\, dx =$

$$= \frac{4}{\pi}\left[\text{arc tg } x\right]_0^x = \frac{4}{\pi} \text{ arc tg } x$$

- Para $x > 1$, se deduce que $F(x) = 1$

Es decir:

$$F(x) = \begin{cases} 0 & \text{para} \quad x < 0 \\ \frac{4}{\pi} \text{ arc tg } x & \text{para} \quad 0 \leq x \leq 1 \\ 1 & \text{para} \quad x > 1 \end{cases}$$

c) Esperanza matemática:

$$E[X] = \int_0^1 x \frac{4}{\pi(1+x^2)} dx = \frac{4}{\pi}\left[\frac{1}{2} L(1+x^2)\right]_0^1 = \frac{2}{\pi} L\, 2$$

****

4.7 *Se llama ley de probabilidad exponencial negativa a la que tiene por función de densidad:*

$$f(x) = \begin{cases} 0 & \text{para} \quad x < 0 \\ k\, e^{-kx} & \text{para} \quad 0 \leq x \leq \infty. \qquad (k > 0) \end{cases}$$

a) Compruebese que en efecto, $f(x)$ es una autentica función de densidad.
b) Obtener la función de distribución.
c) Calcular la esperanza matemática.
d) Hallar la varianza.
e) Calcular las siguientes probabilidades:
$p(X \leq 1)$; $p(1 < X \leq 3)$; $p(X > 5)$

SOLUCION:

a) Veamos que $\int_0^\infty f(x)\, dx = 1.$

$$\int_0^\infty k\, e^{-kx} dx = k\left[\frac{e^{-kx}}{k}\right]_0^\infty = k \frac{1}{k} = 1$$

b) Función de distribución:

$$F(x) = \int_0^x k\, e^{-kx} dx = k\left[\frac{e^{-kx}}{-k}\right]_0^x = -e^{-kx} + 1.$$

c) Esperanza matemática:

$$E[X] = \int_0^x x\, f(x)\, dx = \int_0^x x\, k\, e^{-kx} dx$$

e integrando por partes, resulta:

$$E[X] = \left[\frac{e^{-kx}}{-k} \right]_0^{\infty} = \frac{1}{k}$$

d) Varianza:

$$\sigma^2 = \int_0^{\infty} x^2 \, k \, e^{-kx} \, dx - \left(\frac{1}{k}\right)^2$$

e integrando dos veces por parte, resulta:

$$\sigma^2 = \frac{2}{k^2} - \frac{1}{k^2} = \frac{1}{k^2}$$

e) $p(X \leq 1) = F(1) = -e^{-k} + 1.$

$p(1 < X \leq 3) = F(3) - F(1) = -e^{-3k} + e^{-k}$

$p(X > 5) = 1 - p(X \leq 5) = 1 - F(5) = 1 + e^{-5k} - 1 = e^{-5k}$

****

4.8 *La variable aleatoria X tiene la siguiente función de densidad:*

$$f(x) = \begin{cases} x & \text{para } 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

Calcular:

a) Momento de primer orden respecto al origen.
b) Momento de segundo orden respecto al origen.
c) Momento de tercer orden respecto al origen.
d) Momento de cuarto orden respecto al origen.

SOLUCION:

a) Momento de primer orden respecto al origen:

$$\alpha_1 = E[X] = \int_0^1 x \, x \, dx = \frac{1}{3}$$

b) Momento de segundo orden respecto al origen:

$$\alpha_2 = E\left[X^2\right] = \int_0^1 x^2 x \, dx = \frac{1}{4}$$

c) Momento de tercer orden respecto al origen:

$$\alpha_3 = E\left[X^3\right] = \int_0^1 x^3 x \, dx = \frac{1}{5}$$

d) Momento de cuarto orden respecto al origen:

$$\alpha_4 = E\left[X^4\right] = \int_0^1 x^4 x \, dx = \frac{1}{6}$$

****

4.9 *Para la variable aleatoria del problema anterior, calcúlese:*
a) Momento de primer orden respecto a la media.
b) Momento de segundo orden respecto a la media.
c) Momento de tercer orden respecto a la media.
d) Momento de cuarto orden respecto a la media.

S O L U C I O N:

a) Momento de primer orden respecto a la media:

$$\mu_1 = 0$$

Ya que sabemos que el moemento central de primer orden de cualquier variable es siempre cero.

b) Momento de segundo orden respecto a la media:

$$\mu_2 = \alpha_2 - \alpha_1^2 = \frac{1}{4} - \left(\frac{1}{3}\right)^2 = \frac{5}{36}$$

c) Momento de tercer orden respecto a la media:

$$\mu_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3 = \frac{1}{5} - 3\,\frac{1}{3}\,\frac{1}{4} + 2\left(\frac{1}{3}\right)^2 = 0{,}17$$

d) Momento de cuarto orden respecto a la media:

$$\mu_4 = \alpha_4 - 4\alpha_1\alpha_3 + 6\alpha_1^2\alpha_2 - 3\alpha_1^4 = \frac{1}{6} - 4\,\frac{1}{3}\,\frac{1}{5} + 6\left(\frac{1}{3}\right)^2\frac{1}{4} - 3\left(\frac{1}{3}\right)^4 = 0{,}05$$

****

4.10 *La vida de un virus es una variable aleatoria que tiene la siguiente función de densidad:*

$$f(x) = \begin{cases} 0 & \text{para} \quad x < 1 \\ \dfrac{k}{x^4} & \text{para} \quad x \geq 1 \end{cases}$$

a) Calcular el valor de k para que f(x) sea una función de densidad.
b) ¿Cuál es la vida media de los virus?
c) Hallar la función de distribución.
d) Calcular la mediana.
e) Hallar los cuartiles.

f) Hallar la probabilidad de que un virus tomado al azar, viva mas de 5 horas.

S O L U C I O N:

a) Valor de la constante k:

$$\int_1^{\infty} \frac{k}{x^4}\,dx = k\left[-\frac{1}{3\,x^3}\right]_1^{\infty} = \frac{k}{3} = 1 \implies k = 3$$

b) Vida media de los virus:

$$\alpha_1 = \int_1^{\infty} x\ \frac{3}{x^4}\,dx = \left[-\frac{3}{2x^2}\right]_1^{\infty} = \frac{3}{2}$$

c) Función de distribución:

- Para $x < 1$ $\qquad F(x) = 0$
- Para $x \geq 1$ $\qquad F(x) = \displaystyle\int_1^{x} \frac{3}{x^4}\,dx = -\frac{1}{x^3} + 1$

d) Mediana:

La mediana de X, es el valor de m para el que $F(m) = \frac{1}{2}$

Por tanto, $F(m) = -\dfrac{1}{m^3} + 1 = \dfrac{1}{2} \implies m = \sqrt[3]{2} = 1{,}26$

e) Cuartiles:

De primer orden: $F(Q_1) = \dfrac{1}{4} \implies -\dfrac{1}{Q_1^3} + 1 = \dfrac{1}{4} \implies Q_1 = 1{,}10$

De tercer orden: $F(Q_3) = \dfrac{3}{4} \implies -\dfrac{1}{Q_3^3} + 1 = \dfrac{3}{4} \implies Q_3 = 1{,}58$

f) $p(X > 5) = 1 - p(X \leq 5) = 1 - F(5) = 1 + \dfrac{1}{5^3} - 1 = 0{,}008$

****

4.11 *Utilizando la tabla III del APENDICE, calcúlense las siguientes areas:*

a) Area entre 0 y 0,25

b) Area desde $-\infty$ hasta 1,32.

c) Area entre -2,23 y 1,15.

S O L U C I O N:

a) $p(0 < X \leq 0{,}25) = 0{,}5987 - 0{,}5 = 0{,}0987.$

b) $p(X < 1{,}32) = 0{,}9066.$

c) $p(-2{,}23 < X \leq 1{,}15) = p(X \leq 1{,}15) - 1 + p(X < 2{,}23) =$

$= 0{,}8749 - 1 + 0{,}9871 = 0{,}862.$

****

4.12 *Se sabe que la distribución de los coeficientes intelectuales de los alumnos de una Colegio, sigue la ley normal, conocemos que:* $p(X \geq 1{,}4) = 0{,}1056$ *y* $p(X > 1) = 0{,}4013$.

Calcular los parámetros de la citada distribución.

SOLUCION:

Como $p(X \geq 1{,}4) = 1 - p(X < 1{,}4) = 0{,}1056$, entonces $p(X < 1{,}4) = 0{,}8944$

Luego, $p(Y < \frac{1{,}4 - \mu}{\sigma}) = 0{,}8944$. Buscando en las tablas de la distribución normal, obtenemos $Y = 1{,}25$, es decir, $\frac{1{,}4 - \mu}{\sigma} = 1{,}25$

Del mismo modo, de la otra igualdad se obtiene:

$p(X \geq 1) = 0{,}4013 = 1 - p(X < 1) \implies p(X < 1) = 0{,}5987$

$p(Y < \frac{1 - \mu}{\sigma}) = 0{,}5987$. Buscando en las tablas resulta:

$Y = 0{,}25.$

Por tanto, tenemos:

$$\left.\begin{aligned} \frac{1{,}4 - \mu}{\sigma} &= 1{,}25 \\ \frac{1 - \mu}{\sigma} &= 0{,}25 \end{aligned}\right\} \implies \mu = 0{,}9 \quad y \quad \sigma = 0{,}4$$

Luego se trata de una distribución $N(0{,}9\ ,\ 0{,}4)$

****

4.13 *Se supone que la estancia de los enfermos en un hospital sigue una distribución normal de media 8 días y desviación típica, 3.*

Calcular la probabilidad de que la estancia de un enfermo :

a) Sea inferior a 7 días.
b) Sea superior a 3 días.
c) Esté comprendida entre 10 y 12 días.
d) Esté comprendida entre 1 y 2 días.

SOLUCION:

Sea X la variable aleatoria que expresa el número de dias de estancia de los enfermos en un Hospital, que sigue una distribución N(8,3).

Para calcular las probabilidades pedidas, tendremos que tipificar la variable mediante el cambio $Y = \dfrac{X - 8}{3}$. Esta nueva variable sigue una distribución N(0 , 1).

a) $p(X < 7) = p(Y < \dfrac{7 - 8}{3}) = p(Y < -\dfrac{1}{3}) = 0{,}3707.$

b) $p(X > 3) = 1 - p(X \leq 3) = 1 - p(Y \leq \dfrac{3 - 8}{3}) = 1 - p(Y \leq -\dfrac{5}{3}) = 0{,}9515$

c) $p(10 < X \leq 12) = p(\dfrac{10 - 8}{3} < Y \leq \dfrac{12 - 8}{3}) = 0{,}9082 - 0{,}7454 = 0{,}1628$

d) $p(1 < X \leq 2) = p(\dfrac{1 - 8}{3} < Y \leq \dfrac{2 - 8}{3}) = 0{,}9901 - 0{,}9772 = 0{,}0129$

****

4.14 *El peso teórico de una tableta de aspirina es 324 mg. Si suponemos que los pesos de las tabletas de aspirina siguen una distribución normal de desviación típica 10 mg por tableta,*

a) ¿cuál será el tanto por ciento de tabletas con peso menor o igual a 310 mg?.

b) ¿cuál será el tanto por ciento de tabletas con peso superior a 330 mg?.

S O L U C I O N:

a) $p(X \leq 310) = p(Y \leq \dfrac{310 - 324}{10}) = p(Y \leq -1{,}4) = 1 - p(Y \leq 1{,}4) =$

$= 1 - 0{,}9192 = 0{,}0808$. Por tanto, el 8,1% de las tabletas tendrá peso menor o igual a 310 mg.

b) $p(X > 330) = 1 - p(X \leq 330) = 1 - p(Y \leq \dfrac{330 - 324}{10}) = 1 - 0{,}7257 =$

$= 0{,}2743$. Por tanto, el 27% de las tabletas tienen peso superior a 330 mg.

****

4.15 *Se llama cociente intelectual al cociente entre la edad mental y la edad real. Se sabe que la distribución de los cocientes intelectuales de 2000 reclutas sigue una distribución normal de media 0,80 y desviación típica 0,50.*

Calcular:

a) Número de reclutas con cociente intelectual comprendido entre

0,7 y 1,2.

b) *Número de reclutas con cociente intelectual inferior a 0,3.*

c) *Número de reclutas con cociente intelectual inferior a 0,9.*

d) *Número de reclutas con cociente intelectual superior a 1,4.*

SOLUCION:

La variable X, sigue una distribución N(0,8 , 0,5). Tipificando la variable, tendremos $Y = \dfrac{X - 0,8}{0,5}$

a) $p(0,7 < X \leq 1,2) = p(\dfrac{0,7 - 0,8}{0,5} < Y \leq \dfrac{1,2 - 0,8}{0,5}) = 0,7881 - 1 +$

$+ 0,5793 = 0,3674.$

Luego el número pedido será : $2000.0,3674 = 734,8 \simeq 735.$

b) $p(X < 0,3) = p(Y < \dfrac{0,3-0,8}{0,5}) = p(Y < -1) = 1 - 0,8413 = 0,1587.$

El número pedido será: $2000.0,1587 = 317,4 \simeq 317.$

c) $p(X < 0,9) = p(Y < \dfrac{0,9-0,8}{0,5}) = p(Y < 0,2) = 0,5793$

El número pedido será: $2000.0,5793 = 1159.$

d) $p(X > 1,4) = p(Y > \dfrac{1,4-0,8}{0,5}) = p(Y > 1,2) = 1 - p(Y \leq 1,2) =$

$= 1 - 0,8849 = 0,1151.$

El número pedido será: $2000.0,1151 = 230,2 \simeq 230.$

****

4.16 *La duración media sin ninguna rotura, de las sábanas de un Sanatorio Psiquiatrico de 300 camas, es de 50 dias con una desviación típica de 8 dias. Si suponemos que la duración de las sábanas sigue una distribución normal, calcular:*

a) *¿Cuantas sábanas se habrán tenido que reponer antes de los 35 dias?*

b) *¿Cuantas sábanas habrá que reponer pasados los 60 dias?*

SOLUCION:

Si representamos por X la variable que expresa el número de dias que dura una sábana sin ningun tipo de rotura, sabemos que X obedece a una distribución normal de parámetros: N(50 , 8)

a) $p(X \leq 35) = p(Y \leq \dfrac{35-50}{8}) = p(Y \leq -1,875) = 1 - 0,9693 = 0,0307.$

Como hay 300 camas, son 600 las sábanas que están en uso, por tanto el número de sábanas que habrá que reponer antes de los 35 dias, será: $600.0,0307 = 18,42 \simeq 18$ sábanas

b) $p(X > 60) = 1 - p(X \leq 60) = 1 - p(Y \leq \frac{60-50}{8}) = 1 - p(Y \leq 1,25) =$

$= 1 - 0,8944 = 0,1056.$

Luego después de los primeros 60 dias habrá que reponer:

$$600.0,1056 = 63,36 \simeq 63 \text{ sábanas}$$

****

4.17 *Una tableta de vitamina A, producida por una máquina automática se considera apta si la desviación de su peso observado respecto del teórico no es mayor que 10 mg. Las desviaciones aleatorias del peso observado respecto del teórico obedecen a una ley normal de media 0 y desviación típica 5 mg.*

Calcular que tanto por ciento de tabletas, producidas por la citada máquina automática, se considerarán aptas.

SOLUCION:

Consideraremos una tableta apta, si se desvia de la media menos de 10 mg.

Por tanto, $p(|X| < 10) = p(-10 < X < 10) = p(\frac{-10-0}{5} < Y < \frac{10-0}{5}) =$

$= p(-2 < Y < 2) = 2\,p(0 < Y < 2) = 2(0,9772-0,5) =$

$= 2.0,4772 = 0,9544.$

Luego el 95% de las tabletas producidas serán consideradas válidas.

****

4.18 *Se ha comprobado que la probabilidad de tener un determinado individuo los ojos pardos es 0,6. Sea X la variable aleatoria que representa el número de individuos que tienen los ojos pardos de un grupo de 1100.*

a) Obtener la ley de probabilidad de la varibale X.

b) Calcular $p(670 < X \leq 675)$.

c) Calcular $p(X \geq 680)$.

SOLUCION:

a) La variable X sigue una ley de probabilidad binomial, siendo:

A = "tener ojos pardos" y $\overline{A}$ = "no tener los ojos pardos"

$p = p(A) = 0,6$ $q = 1 - p = p(\overline{A}) = 0,4$

Como los n individuos que componen la muestra son independientes la ley de probabilidad de la variable X será:

$$p(X = r) = \binom{1100}{r} 0,6^r . 0,4^{1100-r}$$

b) Como vimos en 2.9 del presente capítulo, por cumplirse:

1°) $n > 30.$
2°) $0,1 < p < 0,9$

Podemos aproximar esta distribución binomial, mediante una distribución normal $N(0, 1)$, siendo la nueva variable:

$$Y = \frac{X - np}{\sqrt{npq}} \; ; \quad \begin{cases} np = 1100.0,6 = 660. \\ \sqrt{npq} = \sqrt{1100.0,6.0,4} = 16,2 \end{cases}$$

Es decir: $Y = \dfrac{X - 660}{16,2}$. Por tanto, tendremos:

$$p(670 < X \leq 675) = p\left(\frac{670 - 660}{16,2} < Y \leq \frac{675 - 660}{16,2}\right) =$$

$$= 0,8238 - 0,7324 = 0,0914.$$

c) $p(X \geq 680) = 1 - p(X < 680) = 1 - p\left(Y < \dfrac{680 - 660}{16,2}\right) = 1 - p(Y < 1,23) =$

$$= 1 - 0,8907 = 0,1093.$$

****

4.19 *Se ha estudiado la distribución de albúmina total circulante (en gramos) de 50 hombres normales, comprendidos entre los 20 y los 30 años, obteniéndose los siguientes resultados:*

Ajústese esta distribución empírica mediante una distribución normal y calcular:

a) Probabilidad de que un individuo elegido al azar, tenga albúmina circulante menor o igual a 100 grs.

b) Probabilidad de que tenga más de 150 gramos de albúmina.

Albúmina circulante grs.	*n°de hombres*
99,5 - 109,5	*5*
109,5 - 119,5	*10*
119,5 - 129,5	*12*
129,5 - 139,5	*11*
139,5 - 149,5	*8*
149,5 - 159,5	*4*

S O L U C I O N:

El cálculo de la media y desviación típica de esta distribución fué efectuado en el problema 1.14 del capítulo de Estadística Descriptiva. Por tanto, $\bar{x} = 128,3$ y $s = 14,2$

Luego la distribución normal que más se aproxima a esta distribución empírica será la que tenga por parámetros

$$\mu = 128,3 \quad \text{y} \quad \sigma = 14,2$$

a) $p(X \leq 100) = p(Y \leq \frac{100 - 128,3}{14,2}) = p(Y \leq -1,99) = 1 - p(Y < 1,99) =$

$= 1 - 0,9767 = 0,0233.$

b) $p(X > 150) = 1 - p(X \leq 150) = 1 - p(Y \leq \frac{150 - 128,3}{14,2}) = 1 - 0,9370 =$

$= 0,063.$

****

4.20 *Calcular los siguientes valores de la variable que sigue una F de Fisher-Snedecor.*

a) $F_{0,05;5;21}$ *b)* $F_{0,01;24;12}$ *c)* $F_{0,05;8;4}$

SOLUCION:

Mediante lectura directa de la tabla VI del APENDICE, se tiene:

a) $F_{0,05;5;21} = 2,68.$

b) $F_{0,01;24;12} = 3,78.$

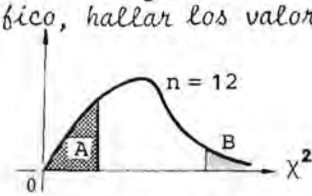
c) $F_{0,05;8;4} = 6,04.$

****

4.21 *Dada una distribución* χ^2 *con* $n = 12$ *grados de libertad y siendo A y B las areas rayadas del gráfico, hallar los valores* χ^2_p *tales que:*

a) El area A sea igual a 0,025.

b) El area B sea igual a 0,01.



SOLUCION:

La tabla IV del APENDICE, nos proporciona para un valor determinado del número de grados de libertad n, el valor de χ^2_p, tal que:

$$p(\chi^2 > \chi^2_p) = p$$

a) $p = 1 - A = 1 - 0,025 = 0,975, \quad p(\chi^2 > \chi^2_p) = 0,975.$

y mirando en las tablas, para n = 12, obtenemos $\chi_p^2 = 4{,}40$.

b) $p = B = 0{,}01$. $p(\chi^2 > \chi_p^2) = 0{,}01 \Longrightarrow \chi_p^2 = 26{,}22$.

****

4.22 *Calcular las siguientes probabilidades*

a) $p(\chi^2 \leq 20{,}4)$ *para* $n = 10$.
b) $p(\chi^2 > 23{,}2)$ *para* $n = 14$.
c) $p(11{,}9 \leq \chi^2 \leq 32{,}4)$ *para* $n = 18$.

SOLUCION:

a) $p(\chi_{10}^2 \leq 20{,}4) = 1 - p(\chi_{10}^2 > 20{,}4)$

para obtener $p(\chi_n^2 > 20{,}4)$, siendo n = 10, se busca en la tabla IV del APENDICE. En la fila correspondiente a n = 10, se busca el valor más próximo por defecto a 20,4, es decir 18,31, al cual corresponde p = 0,05. Para obtener con más aproximación la probabilidad pedida efectuaremos una interpolación lineal. Si a un incremento de χ_{10}^2 de 20,5 - 18,31 = 2,19 coresponde una disminución de p de 0,05 - 0,025 = 0,025, Por tanto a un incremento de χ^2 de 20,4 - 18,31 = 2,09 corresponderá una disminución de p igual a x.

Es decir: $$\frac{2{,}19}{2{,}09} = \frac{0{,}025}{x} \Longrightarrow x = 0{,}0238.$$

Restando 0,0238 del valor 0,050 de p, se obtiene 0,0262

Por tanto, $p(\chi_{10}^2 > 20{,}4) = 0{,}0262$ y $p(\chi_{10}^2 \leq 20{,}4) = 0{,}9738$

b) $p(\chi^2 > 23{,}2)$ para n = 14.

Del mismo modo, tendremos: $$\frac{23{,}68 - 21{,}06}{23{,}20 - 21{,}06} = \frac{0{,}1 - 0{,}05}{x}$$

x = 0,0408. Por tanto 0,1 - 0,0408 = 0,0592

c) $p(11{,}9 \leq \chi_{18}^2 \leq 32{,}4) = p(\chi_{18}^2 \geq 11{,}9) - p(\chi_{18}^2 \geq 32{,}4)$

Para el cálculo de ambas probabilidades procederemos como en los casos anteriores.

$p(\chi_{18}^2 \geq 11{,}9)$ $$\frac{25{,}99 - 10{,}86}{11{,}9 - 10{,}86} = \frac{0{,}9 - 0{,}1}{x} \Longrightarrow x = 0{,}055$$

Por tanto, $p(\chi_{18}^2 \geq 11{,}9) = 0{,}9 - 0{,}055 = 0{,}845$

$$p(\chi^2_{18} \geq 32,4) \Longrightarrow \quad \frac{34,80 - 31,3}{32,40 - 31,3} = \frac{0,025 - 0,010}{x}$$

$x = 0,0047$. Por tanto, $p(\chi^2_{18} \geq 32,4) = 0,025 - 0,0047 = 0,0203$

****

4.23 *Dada una distribución χ^2 con 45 grados de libertad, calcular el valor de χ^2_p de χ^2 correspondiente a $p = 0,025$.*

S O L U C I O N:

Si miramos en la tabla IV del APENDICE correspondiente a la distribución χ^2 observamos que solo nos proporciona valores de χ^2_p hasta n=30 grados de libertad.

Para valores de $n > 30$, vimos en 3.5 que $T = \sqrt{2\chi^2} - \sqrt{2n-1}$ se distribuye segun una $N(0\ ,\ 1)$. Por tanto, como en la tabla de la distribución normal tenemos, $p(T > 1,96) = 0,025$

$T = \sqrt{2\chi^2} - \sqrt{2.45 - 1} > 1,96$; $\quad \sqrt{2\chi^2} > 1,96 + 9,45$,

$2\chi^2 > 129,73$. $\quad \chi^2 = \dfrac{129,73}{2} = 64,87$. $\quad$ Luego $\chi^2_p = 64,87$.

****

4.24 *Dada una población que sigue una distribución normal con media y desviación típica desconocida, se extrae una muestra de tamaño 16 que tiene una varianza muestral conocida s^2.*

Calcular la probabilidad siguiente: $p(0,5 < \frac{s^2}{\sigma^2} < 1,8)$, siendo σ^2 la varianza poblacional.

S O L U C I O N:

En 3.8 del presente capítulo, vimos que $\dfrac{(n-1)\ s^2}{\sigma^2}$ es una variable aleatoria cuya distribución en el muestreo es una χ^2 con n-1 grados de libertad.

Por tanto, $\dfrac{15\ s^2}{\sigma^2}$ es una χ^2 con 15 grados de libertad.

Luego $\quad p(0,5 < \dfrac{s^2}{\sigma^2} < 1,8) = p(0,5.15 < \dfrac{15\ s^2}{\sigma^2} < 15.1,8) =$

$= p(7,5 < \dfrac{15\ s^2}{\sigma^2} < 27) = p(7,5 < \chi^2_{15} < 27) = p(\chi^2_{15} > 7,5) - p(\chi^2_{15} > 27)$

$p(\chi^2_{15} > 7,5)$; $\quad \dfrac{8,55 - 7,26}{7,50 - 7,26} = \dfrac{0,95 - 0,9}{x} \Longrightarrow x = 0,009$

Por tanto, $p(\chi^2_{15} > 7,5) = 0,95 - 0,009 = 0,941$.

$p(\chi^2_{15} > 27)$; $\dfrac{27,5 - 25}{27 - 25} = \dfrac{0,05 - 0,025}{x} \Longrightarrow x = 0,02$

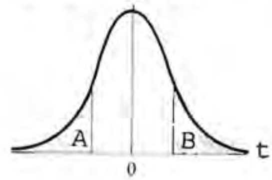
Por tanto, $p(\chi^2_{15} > 27) = 0,05 - 0,02 = 0,03$.

Luego $p(0,5 < \dfrac{s^2}{\sigma^2} < 1,8) = 0,941 - 0,03 = 0,911$.

****

4.25 *En una distribución t de Student con 7 grados de libertad, consideremos las areas A y B rayadas en la figura. Hallar los valores de t_p tales que:*

a) El area A sea igual a 0,025.
b) El area B sea igual a 0,01.
c) El area A + B sea igual a 0,05.



S O L U C I O N:

La tabla V del APENDICE nos proporciona para un valor determinado del número de grados de libertad n, el valor t_p, tal que:

$$p(|t| > t_p) = p.$$

Por tanto:

a) $A = 0,025 = \frac{p}{2}$, luego $p = 0,05$. $p(|t| > t_p) = 0,05$ mirando en las tablas para $n = 7$, tenemos $t_p = -2,365$.

b) $B = 0,01 = \frac{p}{2}$, $p = 0,02$. $p(|t| > t_p) = 0,02$ de donde se deduce que $t_p = 2,998$.

c) $A + B = 0,05 = p$, $p(|t| > t_p) = 0,05 \Longrightarrow t_p = \pm 2,365$.

****

4.26 *Dadas las distribuciones de la t de Student para $n = 15$ para $n = 25$ y para $n = 80$.*

Hallar los valores críticos t_p de t, para los cuales el area situada a la derecha de t es 0,1.

S O L U C I O N:

Los valores críticos t_p de t correspondientes a $p = 2.0,1 = 0,2$, son:

Para $n = 15$: $p(|t| > 1,34) = 0,2$ $t_p = 1,341$.

Para $n = 25$: $p(|t| > 1,316) = 0,2$ $t_p = 1,316$.

Para $n = 80$: $p(|t| > 1,296) = 0,2$ $t_p = 1,296$.

Para este último caso, como $n > 30$, se podría aproximar mediante la distribución normal, del siguiente modo: $p(t > 0{,}1)$, buscamos en la normal para una probabilidad igual a $1 - 0{,}1 = 0{,}9$ y resulta: $t \simeq 1{,}29$, que como se ve es análogo a lo obtenido directamente en la tabla de la t de Student.

****

4.27 *Calcular las siguientes probabilidades de una variable que sigue una distribución t de Student, siendo n el número de grados de libertad:*

a) $p(|t| < 1{,}83)$, *para* $n = 8$.

b) $p(-1{,}64 \leq t \leq 1{,}97)$, *para* $n = 12$.

SOLUCION:

a) $p(|t_8| \leq 1{,}83) = 1 - 2\,p(t_8 > 1{,}83)$;

$$\frac{1{,}86 - 1{,}397}{1{,}83 - 1{,}397} = \frac{0{,}1 - 0{,}05}{x} \implies x = 0{,}046.$$

$$p(|t_8| \leq 1{,}83) = 1 - 2(0{,}10 - 0{,}046) = 0{,}892.$$

b) $p(-1{,}64 \leq t_{12} \leq 1{,}97) = p(t_{12} \geq -1{,}64) - p(t_{12} \geq 1{,}97) =$

$$= 1 - p(t_{12} \geq 1{,}64) - p(t_{12} \geq 1{,}97).$$

- Cálculo de $p(t_{12} \geq 1{,}64)$:

$$\frac{1{,}782 - 1{,}356}{1{,}64 - 1{,}356} = \frac{0{,}10 - 0{,}05}{x} \implies x = 0{,}033.$$

$$p(t_{12} \geq 1{,}64) = 0{,}1 - 0{,}033 = 0{,}067.$$

- Cálculo de $p(t_{12} \geq 1{,}97)$:

$$\frac{2{,}179 - 1{,}782}{1{,}97 - 1{,}782} = \frac{0{,}025 - 0{,}01}{x} \implies x = 0{,}0071.$$

$$p(t_{12} \geq 1{,}97) = 0{,}025 - 0{,}0071 = 0{,}0178.$$

Por tanto: $p(-1{,}64 \leq t_{12} \leq 1{,}97) = 1 - 0{,}067 - 0{,}0178 = 0{,}9152.$

****

4.28 *Calcular las siguientes probabilidades de una variable que sigue una distribución t de Student, siendo n el número de grados de libertad:*

a) $p(t \geq 1,35)$, para $n = 16$.

b) $p(|t| > 2,34)$, para $n = 25$.

SOLUCION:

Utilizaremos la tabla V del APENDICE, realizando una interpolación lineal en los casos que la tabla no nos proporcione los valores directamente.

a) $p(t_{16} \geq 1,35)$: $\dfrac{1,746 - 1,337}{1,35 - 1,337} = \dfrac{0,10 - 0,05}{x}$

$x = 0,0015$. Por tanto $p(t_{16} \geq 1,35) = 0,1 - 0,0015 = 0,0085$.

b) $p(|t_{25}| > 2,34) = 2\, p(t_{25} > 2,34)$:

$$\frac{2,485 - 2,060}{2,34 - 2,060} = \frac{0,025 - 0,01}{x} \implies x = 0,098.$$

$p(|t_{25}| > 2,34) = 2\,(0,025 - 0,0098) = 0,0304$.

****

5- regresión y correlación

1. GENERALIDADES

1.1 Variables estadísticas bidimensionales
1.2 Diagrama de dispersión
1.3 Tipos de tablas de frecuencias bidimensionales
1.4 Distribuciones marginales
1.5 Distribuciones condicionadas
1.6 Momentos bidimensionales respecto al origen
1.7 Momentos bidimensionales respecto a la media
1.8 Relación entre ambos momentos

2. REGRESION

2.1 Concepto general de regresión
2.2 Ajuste de una linea de regresión a un diagrama de dispersión
2.3 Distintos tipos de curvas para realizar el ajuste
2.4 Método de mínimos cuadrados
2.5 Regresión lineal mínimo-cuadrática
2.6 Regresión parabólica mínimo-cuadrática

3. CORRELACION

3.1 Concepto general de correlación
3.2 Dependencia aleatoria
3.3 Dependencia funcional
3.4 Coeficiente de correlación lineal
3.5 Invariancia del coeficiente de correlación lineal ante un cambio de variable
3.6 Estudio del valor del coeficiente de correlación lineal a partir de la varianza residual

5- REGRESION Y CORRELACION

1. GENERALIDADES

1.1 VARIABLES ESTADISTICAS BIDIMENSIONALES.- Hasta ahora habiamos estudiado fenómenos en los que para cada observación se obtenía una medida.

Supongamos ahora un nuevo fenómeno, en el que para cada observación se obtienen un par de medidas.

Por ejemplo:

1.- Area superficial y proteinas circulantes en sangre de los pacientes de un Centro Médico.

2.- Volumen plasmático y albúmina total circulante.

3.- Pulso y temperatura.

4.- Talla y peso de recien nacidos.

5.- Tantos por ciento de pérdida de actividad de un preparado hormonal y meses transcurridos desde su envase.

6.- Altura y mayor diámetro de la cáscara de una determinada especie de caracol.

A estas variables estadísticas resultantes de la observación de un fenómeno respecto de dos modalidades se las llama *variables estadisticas bidimensionales*.

1.2 DIAGRAMA DE DISPERSION.- Dada una variable estadística bidimensional, representaremos por x, a los valores observados respecto de una de las modalidades y por y, a los valores observados respecto de la otra modalidad.

De esta forma obtenemos pares ordenados de la forma

(x,y), que representados sobre unos ejes cartesianos, da lugar a la *nube de puntos* o *diagrama de dispersión*.

1.3 TIPOS DE TABLAS DE FRECUENCIA BIDIMENSIONAL.- Las distribuciones de dos caracteres se pueden clasificar según la naturaleza de estos en: cualitativos, cuantitativos discretos y cuantitativos contínuos.

De esta forma se obtienen 6 tipos generales de distribución con dos caracteres.

TIPO 1.- Los dos caracteres cualitativos.

Ej: Clasificación de pacientes según sexo y profesión.

TIPO 2.- Un carácter cualitativo y otro cuantitativo discreto.

Ej: Clasificación de 100 cabezas de familia, según su categoría socio-profesional y el número de hijos.

TIPO 3.- Un carácter cualitativo y otro cuantitativo contínuo.

Ej: Clasificación de pacientes según raza y perímetro torácico.

TIPO 4.- Los dos caracteres cuantitativos discretos.

Ej: Clasificación de familias, según el número de hijos y el de hijas.

TIPO 5.- Los dos caracteres cuantitativos continuos.

Ej: Clasificación de varones adultos, según pulso y temperatura.

TIPO 6.- Un carácter cuantitativo contínuo y el otro discreto.

Ej: Clasificación de familias según edad de la madre y número de hijos.

En general, si designamos los valores o las modalidades de la variable x por $x_1, x_2, x_3, \ldots, x_n$ y los de la variable y por $y_1, y_2, y_3, \ldots, y_m$ y por f_{ij} la frecuencia absoluta del par (x_i, y_j), tendremos una tabla del siguiente tipo:

Donde $f_{.j} = \sum_{k=1}^{n} f_{kj}$ $\quad f_{i.} = \sum_{l=1}^{m} f_{il}$ y

$f_{..} = \sum_{j=1}^{m} f_{.j} = \sum_{i=1}^{n} f_{i.} = N =$ n°total de individuos que intervienen en la distribución.

$y \backslash x$	x_1	x_2	...	x_i	...	x_n	
y_1	f_{11}	f_{21}	...	f_{i1}	...	f_{n1}	$f_{.1}$
y_2	f_{12}	f_{22}	...	f_{i2}	...	f_{n2}	$f_{.2}$
⋮	⋮	⋮		⋮		⋮	⋮
y_j	f_{1j}	f_{2j}	...	f_{ij}	...	f_{nj}	$f_{.j}$
⋮	⋮	⋮		⋮		⋮	⋮
y_m	f_{1m}	f_{2m}	...	f_{im}	...	f_{nm}	$f_{.m}$
	$f_{1.}$	$f_{2.}$	...	$f_{i.}$	...	$f_{n.}$	$f_{..}$

Para formar las frecuencias relativas, bastará dividir la frecuencia absoluta por el número total de individuos que intervienen en la distribución.

De esta forma, la frecuencia relativa de la casilla (x_i , y_j), será: $h_{ij} = f_{ij}/N$

del mismo modo: $h_{.j} = f_{.j}/ N$ y $h_{i.} = f_{i.}/N$

Si una o las dos de las variables son cuantitativas continuas, agruparemos en clases siendo x_i e y_j las marcas de clase de los intervalos.

1.4 DISTRIBUCIONES MARGINALES.- Veamos la siguiente tabla estadística, donde x representa el pulso e y la temperatura de 26 pacientes

	70	75	80	85	
36	4	3	2	-	9
36,5	3	4	3	1	11
37	-	1	2	3	6
	7	8	7	4	26

Observamos que para la variable y se obtiene "al margen" la siguiente distribución:

y	frecuencia relativa
36	9/26
36,5	11/26
37	6/26

Y del mismo modo, para la variable x tendremos:

x	70	75	80	85
frecuencia relativa	7/26	8/26	7/26	4/26

A estas distribuciones, obtenidas "al margen", las llamamos *distribuciones marginales*.

De una manera general, llamaremos *distribución marginal* de la variable y, a la que tiene por frecuencias relativas:

$$h_{\cdot j} = \frac{f_{\cdot j}}{N} \qquad j = 1,2,...,m.$$

Estos valores los obtendremos en el margen de la derecha de la tabla estadística bidimensional.

Del mismo modo:

Llamaremos *distribución marginal* de la variable x, a la que tiene por frecuencias relativas:

$$h_{i\cdot} = \frac{f_{i\cdot}}{N} \qquad i = 1,2,...,n$$

1.5 <u>DISTRIBUCIONES CONDICIONADAS</u>.- Supongamos ahora que para el ejemplo anterior, fijamos el valor de y = 36°, obtenemos entonces la siguiente distribución de frecuencias relativas de la x:

	70	75	80	85
h(x/y = 36°)	4/9	3/9	2/9	0

A esta distribución la llamaremos *distribución de la variable* x *condicionada* al valor y = 36°.

Según que fijásemos otros valores para la y obtendriamos:

	70	75	80	85
h(x/y = 36,5°)	3/11	4/11	3/11	1/11
h(x/y = 37°)	0	1/6	2/6	3/6

Del mismo modo, fijado un valor correspondiente a la x, obtendriamos la correspondiente distribución condicionada de la variable y.

	f(y/x = 70)	f(y/x = 75)	f(y/x = 80)	f(y/x = 85)
36	4/7	3/8	2/7	0
36,5	3/7	4/8	3/7	1/4
37	0	1/8	2/7	3/4

De una manera general, llamaremos *distribución condicionada* de la variable x, para el valor de $y = y_j$, a la que tiene por frecuencias relativas:

$$h(x_i/y_j) = \frac{f_{ij}}{f_{.j}} \qquad \begin{cases} i = 1,\ldots,n. \\ j = 1,\ldots,m. \end{cases}$$

Análogamente, llamaremos *distribución condicionada* de la variable y, para el valor de $x = x_i$, a la que tiene por frecuencias relativas:

$$h(y_j/x_i) = \frac{f_{ij}}{f_{i.}} \qquad \begin{cases} i = 1,\ldots,n. \\ j = 1,\ldots,m. \end{cases}$$

1.6 MOMENTOS BIDIMENSIONALES RESPECTO AL ORIGEN.- Llamamos *momento de orden r y s respecto al origen* y representaremos por a_{rs} a la expresión: $$a_{rs} = \sum_i \sum_j h_{ij}\, x_i^r\, y_j^s$$

donde h_{ij} es la frecuencia relativa del par (x_i , y_j)

Veamos algunos casos particulares:

- Momento de primer orden de la variable x, respecto al origen.

$$\overline{x} = a_{10} = \sum \sum h_{ij}\, x_i = \sum h_{i.}\, x_i$$

- Momento de primer orden de la variable y, respecto al origen.

$$\overline{y} = a_{01} = \sum \sum h_{ij}\, y_j = \sum h_{.j}\, y_j$$

Al punto (a_{10}, a_{01}) se le llama centro de gravedad de la nube de puntos.

- Momentos de segundo orden:

$$a_{20} = \sum \sum h_{ij}\, x_i^2 = \sum h_{i.} x_i^2$$

$$a_{02} = \sum \sum h_{ij}\, y_j^2 = \sum h_{.j} y_j^2$$

- Momento de primer orden respecto de cada una de las dos variables, será:

$$a_{11} = \sum \sum h_{ij}\, x_i\, y_j$$

1.7 MOMENTOS BIDIMENSIONALES RESPECTO A LA MEDIA.- Llamamos *momento de orden r y s respecto a la media* y representaremos por m_{rs} a la siguiente expresión:

$$m_{rs} = \sum_i \sum_j h_{ij} (x_i - \overline{x})^r (y_j - \overline{y})^s$$

También se les llama *momentos centrales*.

Veamos algunos casos particulares.

- Momento de segundo orden de la variable x, respecto a la media.

$$s_x^2 = m_{20} = \sum \sum h_{ij} (x_i - \overline{x})^2 = \sum h_{i.} (x_i - \overline{x})^2$$

- Momento de segundo orden de la variable y, respecto a la media.

$$s_y^2 = m_{02} = \sum \sum h_{ij} (y_j - \overline{y})^2 = \sum h_{.j} (y_j - \overline{y})^2$$

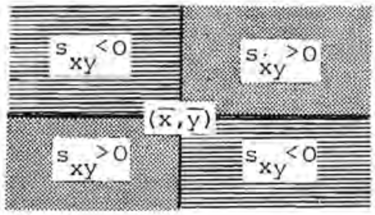
- Momento mixto o *covarianza*

$$s_{xy} = m_{11} = \sum \sum h_{ij} (x_i - \overline{x}) (y_j - \overline{y})$$

Obsérvese que la covarianza será positiva cuando simultaneamente las desviaciones $x_i - \overline{x}$ e $y_j - \overline{y}$ sean positivas o negativas, (es decir, para puntos situados en los cuadrantes 1° y 3°, siendo el origen el punto $(\overline{x}, \overline{y})$).

Del mismo modo, la covarianza será negativa cuando las desviaciones $x_i - \overline{x}$ e $y_j - \overline{y}$ sean de signos contrarios, (es decir, para los puntos situados en el cuadrante 2° y 4°).

Esta situación se aclara con el diagrama adjunto.



A veces, en los casos de datos agrupados, el cálculo de los momentos respecto a la media resulta muy laborioso, por lo que se suele realizar un cambio de variable, de manera análoga a como se hizo para las distribuciones unidimensionales.

Es decir:

$$\left.\begin{array}{l} x_i = x_o + c\,u_i \\ y_j = y_o + k\,v_j \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} \overline{x} = x_o + c\,\overline{u} \\ \overline{y} = y_o + k\,\overline{v} \end{array}\right.$$

y restando miembro a miembro, resulta:

$$\left.\begin{array}{l} x_i - \overline{x} = c(u_i - \overline{u}) \\ y_j - \overline{y} = k(v_j - \overline{v}) \end{array}\right\}$$

Entonces, el momento de órdenes r y s respecto a la media, será:

$$m_{rs} = \sum\sum h_{ij}(x_i - \overline{x})^r (y_j - \overline{y})^s = c^r k^s \sum\sum h_{ij}(u_i - \overline{u})^r (v_j - \overline{v})^s$$

Para los casos concretos de las varianzas y la covarianza, será:

$$\left\{\begin{array}{l} s_x^2 = c^2\, s_u^2 \\ \\ s_y^2 = k^2\, s_v^2 \\ \\ s_{xy} = c\,k\,s_{uv} \end{array}\right.$$

Las expresiones anteriores desarrolladas para los casos de variable continua con datos agrupados, son las siguientes:

Varianza de x:

$$s_x^2 = c_x^2\left(\frac{\sum f_x u_x^2}{N} - \left(\frac{\sum f_x u_x}{N}\right)^2\right)$$

Varianza de y : $$s_y^2 = c_y^2\left(\frac{\sum f_y u_y^2}{N} - \left(\frac{\sum f_y u_y}{N}\right)^2\right)$$

Covarianza: $$s_{xy} = c_x c_y\left(\frac{\sum f u_x u_y}{N} - \frac{\sum f_x u_x}{N}\,\frac{\sum f_y u_y}{N}\right)$$

1.8 RELACIONES ENTRE AMBOS MOMENTOS.- En forma análoga a como hemos venido haciendo para las distribuciones unidimensionales podemos pasar de los momentos centrales a los momentos respecto al origen. Veamos los casos mas sencillos:

- Varianza de la x:

$$s_x^2 = m_{20} = \sum\sum h_{ij}(x_i - \bar{x})^2 = \sum h_{i.}(x_i - \bar{x})^2 =$$
$$= \sum h_{i.}(x_i^2 - 2\,\bar{x}\,x_i + \bar{x}^2) = \sum h_{i.}x_i^2 - 2\bar{x}\sum h_{i.}x_i + \bar{x}^2 =$$
$$= \sum h_{i.}\,x_i^2 - \bar{x}^2 = a_{20} - \bar{x}^2$$

- Varianza de la y:

$$s_y^2 = m_{02} = \sum\sum h_{ij}(y_j - \bar{y})^2 = \sum h_{.j}(y_j - \bar{y})^2 =$$
$$= \sum h_{.j}(y_j^2 - 2\,\bar{y}\,y_j + \bar{y}^2) = \sum h_{.j}y_j^2 - 2\bar{y}\sum h_{.j}y_j + \bar{y}^2 =$$
$$= \sum h_{.j}\,y_j^2 - \bar{y}^2 = a_{02} - \bar{y}^2$$

- Covarianza:

$$s_{xy} = m_{11} = \sum\sum h_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \sum\sum h_{ij}(x_i y_j - x_i\bar{y} - \bar{x}y_j + \bar{x}\bar{y}) =$$
$$= \sum\sum h_{ij}x_i y_j - \bar{y}\sum\sum h_{ij}x_i - \bar{x}\sum\sum h_{ij}y_j + \bar{x}\,\bar{y} =$$
$$= \sum\sum h_{ij}x_i y_j - \bar{x}\,\bar{y} = a_{11} - \bar{x}\,\bar{y}$$

2. REGRESION

2.1 CONCEPTO GENERAL DE REGRESION.- Hemos visto como al estudiar un determinado fenómeno surgen un par de medidas que daban lugar a la idea de variable estadística bidimensional. Por ejemplo, las tallas y los pesos de hombres adultos.

En esta situación llamaremos *regresión o ajuste* a la teoría que trata de expresar mediante una ecuación matemática la relación que existe entre las dos variables, con lo que podemos obtener con cierta aproximación el valor de una de las variables conocida el de la otra.

2.2 AJUSTE DE UNA LINEA DE REGRESION A UN DIAGRAMA DE DISPERSION.-

Si representamos el diagrama de dispersión, observamos que los puntos del diagrama se condensan conforme a una linea que llamaremos *línea de regresión.*

El conocimiento de esta linea permite *estimar* el valor de una de las variables conocido el valor de la otra.

A la hora de tener que realizar el ajuste de una linea de regresión a un diagrama de dispersión, conviene tener en cuenta los siguientes puntos:

1°) Eleccion de la linea de regresión:
Se realizará de forma que la linea elegida sea la que mejor se ajuste al diagrama de dispersión.

2°) Estimación de los parámetros de la linea de regresión:
Supongamos que al representar la nube de puntos, observamos que se condensan en torno a una recta, de ecuación $y = ax + b$
Entonces, a y b son parámetros desconocidos y por tanto tendremos que estimarlos en función de la muestra bidimensional que tengamos.

3°) Estudio de la bondad del ajuste:
Como siempre que realizamos un ajuste, se trata de ver si la linea de regresión elegida, ha sido la idonea o no a la distribución. Para este estudio, en general, utilizaremos la prueba de χ^2 como veremos en próximos capítulos.

2.3 DIVERSOS TIPOS DE CURVAS PARA REALIZAR EL AJUSTE.-

Veamos, a continuación, algunos tipos sencillos de curvas para aproximar y sus ecuaciones:

1.- Recta: $y = a\,x + b$

2.- Parábola: $y = a\,x^2 + b\,x + c$

3.- Cúbica: $y = a\,x^3 + b\,x^2 + c\,x + d$

4.- Cuártica: $y = a\,x^4 + b\,x^3 + c\,x^2 + d\,x + e$

5.- Exponencial: $y = c\,a^x$

6.- Hipérbola: $y = \dfrac{1}{a + b\,x}$

2.4 <u>METODO DE MINIMOS CUADRADOS</u>.- Sean $x_1, x_2, x_3, \ldots, x_n$ los valores de la variable x e $y_1, y_2, y_3, \ldots, y_m$ los valores de la variable y, y f_{ij} la frecuencia absoluta del par $(x_i\,,\,y_j)$

Si los puntos de la forma $(x_i\,,\,y_j)$ fueran muchos y muy próximos, nos permitirian obtener la linea de regresión de forma empírica. Ahora bien, esto en general, no es posible, entonces procederemos del siguiente modo.

Elegiremos una familia de curvas dependientes de los parámetros a,b,c,...,k, que escribiremos mediante la ecuación:

$$y = \Psi\,(x;\ a,b,c,\ldots,k)$$

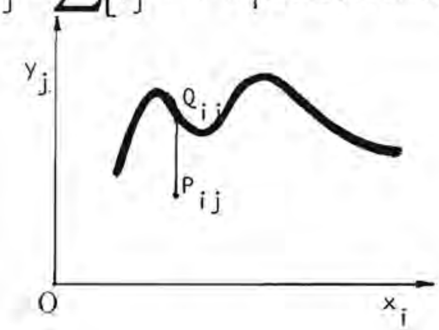
Ahora tratamos de elegir una curva de la familia, tal que haga mínima la expresión:

$$E = \sum (\overline{P_{ij}\,Q_{ij}})^2\, f_{ij}$$

donde Q_{ij} es un punto de la curva de la familia, correspondiente a $x = x_i$ y P_{ij} es un punto observado que tiene de coordenadas $(x_i\,,\,y_j)$.

Es decir: $$E = \sum (\overline{P_{ij}Q_{ij}})^2 f_{ij} = \sum \left[y_j - \Psi\,(x_i;\ a,b,\ldots,k)^2\right] f_{ij}$$

Para hacer mínima la expresión E, sabemos por el Análisis Matemático, que tenemos que anular las derivadas parciales respecto de cada uno de los parámetros.



A continuación veremos la aplicación de este método general de ajuste, a los casos lineal y parabólico.

2.5 REGRESION LINEAL MINIMO CUADRATICA.- Supongamos que el diagrama de dispersión se condensa conforme a una linea recta.

Llamaremos *recta de regresión de y sobre x*, a la recta que permite obtener aproximadamente los valores de y conocidos los valores de x.

Sea pues, la recta de ecuación: $y = a\,x + b$

Tratemos de determinar los parámetros a y b, con arreglo al método de mínimos cuadrados.

Por tanto la expresión siguiente ha de ser mínima.

$$E = \sum f_{ij}(y_j - a\,x_i - b)^2$$

Calculamos las derivadas parciales respecto de cada uno de los dos parámetros a y b.

$$\left.\begin{aligned} \frac{\partial E}{\partial a} &= 2\sum f_{ij}(y_j - a\,x_i - b)(-x_i) = 0 \\ \frac{\partial E}{\partial b} &= 2\sum f_{ij}(y_j - a\,x_i - b)(-1) = 0 \end{aligned}\right\}$$

$$\left.\begin{aligned} \sum f_{ij}\, x_i y_j &= a\sum f_{ij}\, x_i^2 + b\sum f_{ij}\, x_i \\ \sum f_{ij}\, y_j &= a\sum f_{ij}\, x_i + b\sum f_{ij} \end{aligned}\right\}$$

Y dividiendo por N resulta:

$$\left.\begin{aligned} a_{11} &= a\, a_{20} + b\,\overline{x} \\ \overline{y} &= a\,\overline{x} + b \end{aligned}\right\}$$

Que son las *ecuaciones normales* de la recta de regresión de y sobre x, mínimo-cuadrática.

En el caso que $f_{ij} = 1$ para todo i y j, resulta:

$$\left.\begin{aligned} \sum y &= a\sum x + b\,N \\ \sum xy &= a\sum x^2 + b\sum x \end{aligned}\right\}$$

Los parámetros a y b, los determinaremos resolviendo el sistema anterior.

De la segunda ecuación se obtiene: $b = \overline{y} - a\,\overline{x}$

y sustituyendo en la primera ecuación resulta:

$$a_{11} = a\,a_{20} + (\overline{y} - a\,\overline{x})\,\overline{x}; \qquad a_{11} = a\,a_{20} + \overline{y}\,\overline{x} - a\,\overline{x}^2;$$

$$a_{11} - \overline{y}\,\overline{x} - a(a_{20} - \overline{x}^2) = 0.$$

Ahora bien , como $\begin{cases} a_{20} - \overline{x}^2 = s_x^2 \\ a_{11} - \overline{y}\,\overline{x} = s_{xy} \end{cases}$ Se tiene:

$$s_{xy} - a\,s_x^2 = 0 \quad \Longrightarrow \quad a = \frac{s_{xy}}{s_x^2}$$

Por tanto, acabamos de encontrar la pendiente de la recta de regresión. Como por la segunda de las ecuaciones normales sabemos que: $\overline{y} = a\,\overline{x} + b$, tenemos que la recta pasa por el punto $(\overline{x}\,,\,\overline{y})$

Podemos escribir la recta de regresión en la forma punto-pendiente.

Es decir:

$$\boxed{y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})} \qquad [1]$$

Que es la ecuación de la recta de regresión de y sobre x mínimo-cuadrática.

Del mismo modo, obtendremos la recta de regresión de x sobre y, mínimo-cuadrática, que será:

$$\boxed{x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})} \qquad [2]$$

La ecuación [1] permite calcular con cierta aproximación los valores de y conocidos los de x.

La ecuación [2] permite calcular con cierta aproximación los valores de x conocidos los de y.

A los coeficientes $\frac{s_{xy}}{s_x^2}$ y $\frac{s_{xy}}{s_y^2}$ los llamaremos *coeficientes de regresión.*

2.6 REGRESION PARABOLICA MINIMO CUADRATICA.- Supongamos que el diagrama de dispersión se condensa conforme a una parábola de 2° grado, de ecuación: $y = a\,x^2 + b\,x + c$.

El problema consiste en determinar los parámetros a, b, y c para poder conocer la ecuación de la parábola, para ello utilizaremos el método de mínimos cuadrados.

Luego, hemos de hacer mínima la expresión:

$$E = \sum f_{ij}(y_{ij} - a\,x_i^2 - b\,x_i - c)^2$$

$$\left.\begin{array}{l}
\dfrac{\partial E}{\partial a} = 2\sum f_{ij}(y_{ij} - a\,x_i^2 - b\,x_i - c)(-x_i^2) = 0 \\
\dfrac{\partial E}{\partial b} = 2\sum f_{ij}(y_{ij} - a\,x_i^2 - b\,x_i - c)(-x_i) = 0 \\
\dfrac{\partial E}{\partial c} = 2\sum f_{ij}(y_{ij} - a\,x_i^2 - b\,x_i - c)(-1) = 0
\end{array}\right\}$$

$$\left.\begin{array}{l}
\sum f_{ij}\,x_i^2\,y_{ij} = a\sum f_{ij}\,x_i^4 + b\sum f_{ij}\,x_i^3 + c\sum f_{ij}\,x_i^2 \\
\sum f_{ij}\,x_i\,y_{ij} = a\sum f_{ij}\,x_i^3 + b\sum f_{ij}\,x_i^2 + c\sum f_{ij}\,x_i \\
\sum f_{ij}\,y_{ij} = a\sum f_{ij}\,x_i^2 + b\sum f_{ij}\,x_i + c\sum f_{ij}
\end{array}\right\}$$

Estas son las ecuaciones normales de la parábola de regresión de y sobre x, mínimo-cuadrática.

En el caso particular que $f_{ij} = 1$, para todo i y j resulta:

$$\left.\begin{array}{l}
\sum y = a\sum x^2 + b\sum x + c\,N \\
\sum x\,y = a\sum x^3 + b\sum x^2 + c\sum x \\
\sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2
\end{array}\right\}$$

Este sistema se recuerda facilmente, observando que se puede obtener multiplicando la ecuación de partida:

$y = a\,x^2 + b\,x + c$, por 1, x, x^2, respectivamente y sumando los miembros de las ecuaciones respectivas.

Resolviendo el sistema formado por las ecuaciones normales, se obtienen los parámetros a, b y c.

Del mismo modo se obtendrian aproximaciones mínimo-cuadráticas respecto de otros tipos de lineas.

3. CORRELACION

3.1 CONCEPTO GENERAL DE CORRELACION.- Llamaremos *correlación* a la teoría que trata de estudiar la "dependencia" que existe entre las dos variables que intervienen en una distribución bidimensional.

3.2 DEPENDENCIA ALEATORIA.- Supongamos que tenemos en una urna 4 bolas numeradas con las cifras 1,2,3,4. Extraemos una bola y sin devolverla realizamos una segunda extracción.

Si representamos por x el número de la primera bola extraida, tendremos que si x = 3, la variable y únicamente podrá tomar los valores 1,2,4.

En este caso, se dice que la variable x *depende aleatoriamente* de la variable y.

Ahora bien, supongamos que extraemos una bola de la urna y antes de realizar una segunda extracción reintegramos la bola, con lo cual en este caso las pruebas son independientes.

Si representamos por x, el número de la primera bola extraida y por y el número de la segunda bola extraida, tendremos que si x = 3, la variable y podrá tomar los valores 1,2,3,4.

Es decir, en este caso, nada se puede decir acerca del valor de y, una vez conocido el de x.

A este tipo de situación, la llamaremos *independencia aleatoria* o también diremos que la variable y es aleatoriamente independiente con la variable x.

3.3 DEPENDENCIA FUNCIONAL.- Supongamos la misma situación anterior. Es decir, una urna que contiene 4 bolas numeradas con 1,2,3,4.

Representemos por x el número de la primera bola extraída y representemos por y el número 5 - x. En este caso conocido el valor de x, automaticamente sabemos el valor de y sin mas que sustituir en la función $y = 5 - x$.

Es decir, si $x = 3$, se sigue que $y = 2$.

Por lo tanto, diremos que la variable y depende funcionalmente de la variable x, o bien que ambas están en *dependencia funcional*.

Conviene resaltar qu el caso de dependencia aleatoria es un caso intermedio entre la dependencia funcional y la independencia aleatoria, ya que en la dependencia funcional el valor y queda perfectamente determinado al conocer x, y en la independencia aleatoria, el valor de x no nos proporciona ninguna información sobre el valor de y.

En cambio, en la dependencia aleatoria el valor de x nos permite conocer algo sobre el valor que puede tomar y.

Tanto la dependencia e independencia aleatoria, como la dependencia funcional, se presentan constantemente en las ciencias de la observación y experimentales.

3.4 COEFICIENTE DE CORRELACION LINEAL.- Llamaremos *coeficiente de correlación lineal* y representaremos por r, a la media geométrica de los coeficientes de regresión lineal.

Es decir:

$$r = \sqrt{\frac{s_{xy}}{s_x^2} \cdot \frac{s_{xy}}{s_y^2}} = \frac{s_{xy}}{s_x \, s_y}$$

Obsérvese que el coeficiente de correlación lineal es un número sin dimensión.

3.5 INVARIANCIA DEL COEFICIENTE DE CORRELACION LINEAL ANTE UN CAMBIO DE VARIABLE.- En 1.7 del presente capítulo, vimos que para casos de datos agrupados, resultaba muy cómodo realizar un cambio de variable del siguiente modo:

$$\left.\begin{array}{l} x_i = x_o + c\, u_i \\ y_j = y_o + k\, v_j \end{array}\right\} \Longrightarrow \left.\begin{array}{l} \overline{x} = x_o + c\, \overline{u} \\ \overline{y} = y_o + k\, \overline{v} \end{array}\right\}$$

Obteniendo que:
$$\left\{\begin{array}{l} s_x^2 = c^2\, s_u^2 \\ s_y^2 = k^2\, s_v^2 \\ s_{xy} = c\, k\, s_{uv} \end{array}\right.$$

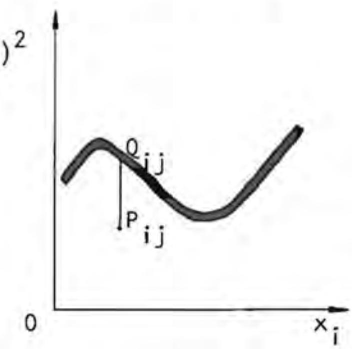
Por tanto:
$$r = \frac{s_{xy}}{s_x\, s_y} = \frac{c\, k\, s_{uv}}{c\, s_u\, k\, s_v} = \frac{s_{uv}}{s_u\, s_v}$$

Luego, el coeficiente de correlación lineal es invariante ante un cambio de variable.

3.6 ESTUDIO DEL VALOR DEL COEFICIENTE DE CORRELACION LINEAL A PARTIR DE LA VARIANZA RESIDUAL.

- Llamaremos *varianza residual* y representaremos por s_r^2, a la suma de los cuadrados de las desviaciones, es decir:

$$s_r^2 = \sum h_{ij} (\overline{P_{ij} Q_{ij}})^2$$



Para facilitar los cálculos, supongamos que realizamos un cambio de origen, con lo que $\overline{x} = \overline{y} = 0$.

Por tanto, la recta de regresión de y sobre x, quedará:

$$y = a\, x$$

Entonces, tendremos:

$$s_r^2 = \sum h_{ij} (\overline{P_{ij} Q_{ij}})^2 = \sum h_{ij} (y_j - a\, x_i)^2 =$$
$$= \sum h_{ij} (y_j^2 - 2\, a\, x_i y_j + a^2 x_i^2) =$$
$$= \sum h_{ij} y_j^2 - 2\, a \sum h_{ij} x_i y_j + a^2 \sum h_{ij} x_i^2 =$$
$$= s_y^2 - 2\, a\, s_{xy} + a^2\, s_x^2.$$

Sustituyendo a por su valor, es decir $a = \frac{s_{xy}}{s_x^2}$, resulta:

$$s_r^2 = s_y^2 - 2\frac{s_{xy}^2}{s_x^2} + \frac{s_{xy}^2}{s_x^2} = s_y^2 - \frac{s_{xy}^2}{s_x^2} = s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right) =$$

$$= s_y^2\,(1 - r^2) \geq 0.$$

Luego tenemos: $s_r^2 = s_y^2\,(1 - r^2)$.

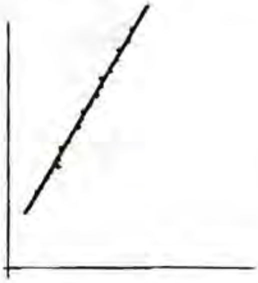
Por tanto,

$$r^2 = 1 - \frac{s_r^2}{s_y^2}$$

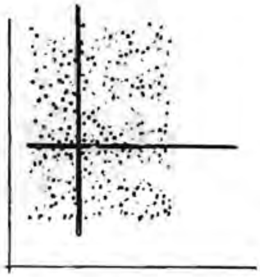
Como $s_r^2 \leq s_y^2$, resulta que: $r^2 \leq 1$

Es decir: $\boxed{-1 \leq r \leq 1}$

- 1°) Si $r = -1 \Longrightarrow s_r^2 = 0$, por tanto todos los puntos se encuentran sobre la recta de regresión y entre las dos variables existe una *dependencia funcional*.
- 2°) Si $-1 < r < 0$, las dos variables están tanto más correladas a medida que r se aproxima a -1. Por tanto, es un caso de *dependencia aleatoria*.
- 3°) Si $r = 0 \Longrightarrow s_r^2 = s_y^2$, en este caso no existe ninguna relación entre las dos variables. Es decir, ambas variables son *incorreladas*. Este es un caso de *independencia aleatoria*.
- 4°) Si $0 < r < 1$, las dos variables están tanto más correladas a medida que r se aproxima a 1. Por tanto, es un caso de *dependencia aleatoria*.
- 5°) Si $r = 1 \Longrightarrow s_r^2 = 0$, por tanto todos los puntos se encuentran sobre la recta de regresión y entre las dos variables existe una *dependencia funcional*.



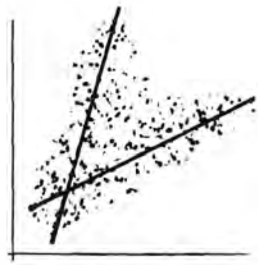
$r = +1$
Dependencia funcional



$r = 0$
Dependencia aleatoria

$r = -1$
Dependencia funcional

$-1 < r < 0$



$0 < r < +1$

Independencia aleatoria

****

PROBLEMAS RESUELTOS

5.1 *La tabla adjunta representa la población formada por 60 hombres adultos, respecto de dos caracteres, altura (x) y peso (y)*

y \ x	1,55-1,65	1,65-1,75	1,75-1,85
55-65	3	2	-
65-75	6	10	4
75-85	4	11	5
85-95	1	6	8

a) Estudiar la distribución marginal de la variable x.
b) Estudiar la distribución marginal de la variable y.
c) Media y varianza marginales de la variable x.
d) Media y varianza marginales de la variable y.

SOLUCION:

a) La distribución marginal de la x, es:

x_i	1,55-1,65	1,65-1,75	1,75-1,85
frecuencia absoluta	14	29	17
frecuencia relativa	14/60	29/60	17/60

b) La distribución marginal de la y, es:

y_i	frecuencia absoluta	frecuencia relativa
55 - 65	5	5/60
65 - 75	20	20/60
75 - 85	20	20/60
85 - 95	15	15/60

c) Media y varianza marginales de la x.

- Media marginal de la x:

Tomemos las marcas de clase de los intervalos.

$$\bar{x} = \frac{14.1,60 + 29.1,70 + 17.1,80}{60} = 1,70\dot{5}.$$

- Varianza marginal de la x: $s_x^2 = \frac{\sum f_i\, x_i^2}{N} - \bar{x}^2 =$

$$= \frac{14.1,60^2 + 29.1,70^2 + 17.1,80^2}{60} - (1,705)^2 = 0,005.$$

d) Media y varianza marginales de la y.

- Media marginal de la y:

$$\bar{y} = \frac{60.5 + 70.20 + 80.20 + 90.15}{60} = 77,5$$

- Varianza marginal de la y:

$$s_y^2 = \frac{\sum f_i y_i^2}{N} - \bar{y}^2 = \frac{60^2.5 + 70^2.20 + 80^2.20 + 90^2.15}{60} -$$

$$-77,5^2 = 85,41.$$

****

5.2 *Dada la variable estadística bidimensional discreta definida por la siguiente tabla:*

y \ x	3	5	
8	25	10	35
12	15	5	20
16	32	13	45
	72	28	100

Calcular:
a) Distribución condicionada de x para y = 12.
b) Distribución condicionada de y para x = 3.
c) Media y varianza condicionadas de la distribución del apartado a).

S O L U C I O N:

a) Distribución condicionada de x para y = 12.

	3	5
h(x/y = 12)	15/20	5/20

b) Distribución condicionada de y par x = 3.

	8	12	16
h(y/x = 3)	25/72	15/72	32/72

c) Media y varianza concidionadas de la distribución del apartado a).

$$\bar{x}\ (\ y = 12\ ) = \frac{3.15 + 5.5}{20} = 3,5$$

$$s_x^2\ (y = 12) = \frac{3^2.15 + 5^2.5}{20} - 3,5^2 = 0,75$$

****

5.3 *La tabla adjunta muestra la natalidad por cada 1000 individuos en España, durante los años 1969 - 1973.*

Años	*1969*	*1970*	*1971*	*1972*	*1973*
Natalidad por cada 1000 hab.	*56*	*59*	*64*	*65*	*73*

a) *Calcular el coeficiente de correlación lineal e interpretar su valor.*

b) *Cuantos nacimientos se esperan para el año 1978.*

SOLUCION:

x_i	y_i	u_i	v_i	u_i^2	v_i^2	u_iv_i
1969	56	-2	-8	4	64	16
1970	59	-1	-5	1	25	5
1971	64	0	0	0	0	0
1972	65	1	1	1	1	1
1973	73	2	11	4	121	22
		0	-1	10	211	44

Realizamos dos cambios de variable, ya que como sabemos el coeficiente de correlación lineal es invariante ante un cambio de origen.

$$\bar{u} = \frac{\sum u_i}{5} = 0, \qquad \bar{v} = \frac{\sum v_i}{5} = -\frac{1}{5}$$

$$s_u^2 = \frac{\sum (u_i - \bar{u})^2}{N} = \frac{\sum u_i^2}{N} - \bar{u}^2 = \frac{10}{5} - 0 = 2.$$

$$s_v^2 = \frac{\sum (v_i - \bar{v})^2}{N} = \frac{\sum v_i^2}{N} - \bar{v}^2 = \frac{211}{5} - \frac{1}{25} = 42,16.$$

$$s_{uv} = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{N} = \frac{\sum u_i v_i}{N} - \bar{u}\,\bar{v} = \frac{44}{5} - 0\left(\frac{-1}{5}\right) = 8,8.$$

Coeficiente de correlación lineal:

$$r = \frac{s_{uv}}{\sqrt{s_u^2\, s_v^2}} = \frac{8,8}{\sqrt{2.42,16}} = 0,95.$$

Como 0,95 es muy próximo a la unidad, la dependencia que existe entre ambas variables es practicamente funcional de no intervenir en los años próximos factores que alteren la relación de dependencia.

b) Cuantos nacimientos se esperan para el año 1978.

Calculemos la recta de regresión de y sobre x.

$$s_x^2 = c^2\, s_u^2 = 1^2.2 = 2; \qquad s_{xy} = c\; k\; s_{uv} = 1.1.8,8 = 8,8.$$

$$\left.\begin{array}{l} \bar{x} = x_o + c\,\bar{u} = 1971 + 1.0 = 1971 \\ \bar{y} = y_o + k\,\bar{v} = 64 + 1(-\frac{1}{5}) = 63,8 \end{array}\right\} \qquad \bar{x} = 1971;\ \bar{y} = 63,8$$

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}); \qquad y - 63,8 = \frac{8,8}{2}(x - 1971)$$

$$y = 4,4\,x - 8608,6$$

Por tanto, para x = 1978 resulta y = 94,6.

Es decir, para el año 1978, se esperan aproximadamente 94600 nacimientos, siempre que no se alteren en estos años los factores que hacen posible esta dependencia.

****

5.4 *Se determina la pérdida de actividad de un preparado hormonal en el curso del tiempo y se obtiene el siguiente resultado:*

Tiempo (en meses)	1	2	3	4	5
% Actividad restante	90	75	42	30	21

Se desea calcular:
a) La relación fundamental entre actividad y tiempo transcurrido.
b) El tiempo en meses que corresponde al 90% de actividad.
c) Cuando será nula la actividad.

S O L U C I O N:

Formemos en primer lugar la tabla de cálculo.

x_i	y_i	$x_i y_i$	x_i^2
1	90	90	1
2	75	150	4
3	42	126	9
4	30	120	16
5	21	105	25
15	258	591	55

Medias aritméticas: $\bar{x} = \frac{15}{5} = 3; \qquad \bar{y} = \frac{258}{5} = 51,6$

Covarianza: $s_{xy} = \frac{1}{N}(\sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + \bar{x}\,\bar{y}) =$

$= \frac{1}{5}(591 - 3.258 - 51,6.15 + 3.51,6) = -160,4.$

Varianza: $s_x^2 = \frac{1}{N}(\sum x_i^2 - \bar{x}^2) = \frac{1}{5}(55 - 3^2) = 9,2.$

a) Relación funcional entre actividad y tiempo transcurrido:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}); \qquad y - 51,6 = \frac{-160,4}{9,2}(x - 3).$$

$$\boxed{y = -17,4\,x + 104} \qquad [1]$$

b) Tiempo en meses que corresponde al 90% de actividad:

Sustituiremos $y = 90\%$ de actividad en la ecuación [1].

Obteniendo 0,8 meses, es decir, aproximadamente 24 dias.

c) ¿Cuando será nula la actividad?

Sustituiremos $y = 0$, correspondiendo la pérdida de la actividad hormonal a los 6 meses.

****

5.5 *Los datos siguientes representan los valores de dos constantes biológicas de 12 individuos de una cierta población:*

65	50	55	65	55	70	65	70	55	70	50	50	x
85	79	76	90	85	87	94	98	81	95	76	74	y

a) Cuando la constante biológica x tome el valor 80, ¿que valor se puede predecir para la variable y?.
b) Y cuando la variable y valga 100, ¿que valor esperado tendrá la variable x?
c) Decir qué tipo de dependencia hay entre las dos variables.

SOLUCION:

Formemos la correspondiente tabla de cálculo, ordenando previamente una de las variables.

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
50	74	- 10	- 11	100	121	110
50	76	- 10	- 9	100	81	90
50	79	- 10	- 6	100	36	60
55	76	- 5	- 9	25	81	45
55	81	- 5	- 4	25	16	20
55	85	- 5	0	25	0	0
65	85	5	0	25	0	0
65	90	5	5	25	25	25
65	94	5	9	25	81	45
70	87	10	2	100	4	20
70	95	10	10	100	100	100
70	98	10	13	100	169	130
720	1020	0	0	750	714	645

Medias aritméticas: $\bar{x} = \dfrac{\sum x_i}{N} = \dfrac{720}{12} = 60$; $\bar{y} = \dfrac{\sum y_i}{N} = \dfrac{1020}{12} = 85.$

Covarianza: $s_{xy} = \dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{N} = \dfrac{645}{12} = 53,75.$

Varianzas: $s_x^2 = \dfrac{\sum (x_i-\bar{x})^2}{N} = \dfrac{750}{12} = 62,5$; $s_y^2 = \dfrac{\sum (y_i-\bar{y})^2}{N} = 59,5.$

a) ¿Qué valor toma y para $x = 80$?

Calculemos la recta de regresión de y sobre x.

$y - \bar{y} = \dfrac{s_{xy}}{s_x^2}(x - \bar{x})$; $\qquad y - 85 = \dfrac{53,75}{62,5}(x - 60)$

$$y = 0{,}86\ x + 33{,}4$$

Por tanto, para $x = 80$, resulta $y = 0{,}86.80 + 33{,}4 = 102{,}2$.

b) ¿Qué valor se espera para x, cuando y = 100?.

Calculemos la recta de regresión de x sobre y.

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}); \qquad x - 60 = \frac{53{,}75}{59{,}5}(y - 85)$$

$$x = 0{,}90\ y - 16{,}8$$

Por tanto, para $y = 100$, resulta $x = 73{,}2$.

c) Dependencia que existe entre ambas variables.

Calculemos el coeficiente de correlación lineal.

$$r = \frac{s_{xy}}{s_x\ s_y} = \frac{53{,}75}{\sqrt{62{,}5.59{,}5}} = 0{,}88$$

Entre ambas variables existe una dependencia aleatoria muy fuerte, próxima a una dependencia funcional ya que r es muy próximo a la unidad.

****

5.6 *El número de bacterias por unidad de volumen presentes en un cultivo despues de x horas, viene representado en la tabla siguiente:*

Número de horas	*0*	*1*	*2*	*3*	*4*	*5*
Número de bacterias por unidad de volumen	*12*	*19*	*23*	*34*	*56*	*62*

a) *¿Cuantas horas habrán de pasar hasta tener 100 bacterias?.*
b) *¿Cuantas bacterias se espera tener al cabo de 6 horas?.*
c) *Calcular el coeficiente de correlación lineal e interpretar su valor.*

SOLUCION:

Representemos por x_i el número de horas y por y_i el número de bacterias por unidad de volumen.

Formemos la correspondiente tabla de cálculos.

x_i	y_i	x_iy_i	x_i^2	y_i^2
0	12	0	0	144
1	19	19	1	361
2	23	46	4	529
3	34	102	9	1156
4	56	224	16	3136
5	62	310	25	3844
15	206	701	55	9170

$$\bar{x} = \frac{15}{6} = 2,5; \qquad \bar{y} = \frac{206}{6} = 34,33.$$

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{N} = \frac{\sum x_i^2}{N} - \bar{x}^2 = \frac{55}{6} - 2,5^2 = 2,9.$$

$$s_y^2 = \frac{\sum (y_i - \bar{y})}{N} = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{9170}{6} - 34,3^2 = 349,5.$$

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum x_i y_i}{N} - \bar{x}\,\bar{y} = \frac{701}{6} - 2,5.34,3 = 31.$$

a) ¿Cuantas horas habrán de pasar hasta tener 100 bacterias?

Calculemos la recta de regresión de x sobre y.

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}); \qquad x - 2,5 = \frac{31}{349,5}(y - 34,3)$$

$$x = 0,088\ y - 0,55$$

Por tanto, para y = 100 bacterias, se obtiene x = 8,25.

Es decir a los 8 horas y 15 minutos habrá 100 bacterias en el cultivo.

b) ¿Cuantas bacterias se esperan tener al cabo de 6 horas?.

Formemos, en este caso, la recta de regresión de y sobre x.

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}); \qquad y - 34{,}33 = \frac{31}{2{,}9}(x - 2{,}5)$$

$$y = 10{,}68\ x + 7{,}61$$

Por tanto, para x = 6, resulta x = 72 bacterias.

c) Coeficiente de correlación lineal

$$r = \frac{s_{xy}}{s_x s_y} = \frac{31}{\sqrt{2{,}9 . 349{,}5}} = 0{,}97.$$

Al ser muy próximo a la unidad, existe entre las dos variables una dependencia aleatoria muy fuerte.

****

5.7 *En la siguiente tabla se recogen las medidas de la máxima de la presión sanguínea y la edad de 12 mujeres.*

Edad (años)	56	42	70	35	64	47	53	49	38	42	68	60
P.máx. (Mm Hg)	147	122	160	118	150	130	146	145	113	145	152	155

Estudiar la relación posible entre ambas variables y razonar el resultado.

S O L U C I O N:

Formemos la tablas de cálculos, teniendo en cuenta que conviene ordenar los datos respecto de una de las variables.

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
35	113	- 17	- 27	289	729	459
38	115	- 14	- 25	196	625	350
42	122	- 10	- 18	100	324	180
42	145	- 10	5	100	25	- 50
47	130	- 5	- 10	25	100	50
49	145	- 3	5	9	25	- 15
53	146	1	6	1	36	6
56	147	4	7	16	49	28
60	155	8	15	64	225	120
64	150	12	10	144	100	120
68	152	16	12	256	144	192
70	160	18	20	324	400	360
624	1680	- 10	0	1524	2782	1800

Medias aritméticas: $\bar{x} = \dfrac{\sum x_i}{N} = \dfrac{624}{12} = 52;\quad \bar{y} = \dfrac{1680}{12} = 140.$

Varianza de x: $s_x^2 = \dfrac{\sum (x_i - \bar{x})^2}{N} = \dfrac{1524}{12} = 127.$

Varianza de y: $s_y^2 = \dfrac{\sum (y_i - \bar{y})^2}{N} = \dfrac{2782}{12} = 232.$

Covarianza: $s_{xy} = \dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{N} = \dfrac{1800}{12} = 150.$

Calculemos las rectas de regresión.

a) Recta de regresión de y sobre x:

$y - 140 = \dfrac{150}{127}(x - 52) \qquad y = 1{,}2\,x + 78{,}5.$

b) Recta de regresión de x sobre y:

$x - 52 = \dfrac{150}{232}(y - 140) \qquad x = 0{,}64\,y - 38{,}5.$

c) Coeficiente de correlación lineal.

$$r = \frac{s_{xy}}{\sqrt{s_x^2\ s_y^2}} = \frac{150}{\sqrt{232.127}} = 0{,}87.$$

Como r es próximo a la unidad, la dependencia entre ambas variables es aleatoria positiva y bastante fuerte.

****

5.8 *Ajustar una parábola a los datos siguientes:*

x_i	50	52	54	56	58	60	62
y_i	31	50	77	96	112	130	162

SOLUCION:

Sea la parábola de ecuación $y = a\,x^2 + b\,x + c$.

Las ecuaciones normales, obtenidas por el procedimiento de mínimos cuadrados, son:

$$\begin{cases} \sum y = a \sum x^2 + b \sum x + c\, N \\ \sum y\, x = a \sum x^3 + b \sum x^2 + c \sum x \\ \sum y\, x^2 = a \sum x^4 + b \sum x^3 + c \sum x^2 \end{cases}$$

Para formar la tabla de cálculos, conviene realizar un cambio de variable para la x_i, con lo que se simplifica notablemente.

Sea $x_j = 55 + 2\, u_i$

x_i	y_i	u_i	u_i^2	u_i^3	u_i^4	$u_i y_i$	$u_i^2 y_i$
50	31	- 3	9	-27	81	- 93	279
52	50	- 2	4	- 8	16	-100	200
54	77	- 1	1	- 1	1	- 77	77
56	96	0	0	0	0	0	0
58	112	1	1	1	1	112	112
60	130	2	4	8	16	260	520
62	162	3	9	27	81	486	1458
	658	0	28	0	196	588	2646

Observamos que gracias al cambio de variable realizado, tenemos que $\sum x_i$ y $\sum x_i^3$, son nulas; por tanto, las ecuaciones normales quedan sensiblemente simplificadas de la siguiente forma:

$$\begin{cases} \sum y = a \sum u^2 + \qquad + c\, N \\ \sum u y = \qquad + b \sum u^2 \\ \sum u^2 y = a \sum u^4 \qquad + c \sum u^2 \end{cases} \Longrightarrow \begin{cases} 658 = 28\, a + 7\, c \\ 588 = 28\, b \\ 2646 = 196\, a + 28\, c \end{cases}$$

De donde $a = 0{,}16$; $b = 21$; $c = 93{,}33$

Y la parábola será:

$$y = 0{,}16\, u^2 + 21\, u + 93{,}3$$

Ahora bien, como $x = 55 + 2\, u$, se deduce $u = \dfrac{x - 55}{2}$

Sustituyendo este valor en la ecuación anterior, obtenemos el ajuste pedido:

$$\boxed{y = 0{,}04\, x^2 + 6{,}1\, x - 363{,}17}$$

****

5.9 *Los datos siguientes representan presiones diastólicas de la sangre tomadas en reposo. Los valores x denotan el tiempo en minutos transcurridos desde el comienzo del descanso y los valores y representan las presiones diastólicas.*

	0	5	10	15	20
66	1	2	0	1	2
68	3	2	1	0	1
70	0	1	0	1	2
72	1	2	1	2	1
74	3	1	2	1	2

a) *Calcular la recta de regresión de la y sobre la x.*
b) *Calcular la recta de regresión de la x sobre la y.*
c) *Calcular el coeficiente de correlación lineal y estudiar la dependencia que existe entre las dos variables.*

SOLUCION:

	u_x	-2	-1	0	1	2	f_y	$f_y u_y$	$f_y u_y^2$	$f u_x u_y$
u_y	y \ x	0	5	10	15	20				
-2	66	$1_{\boxed{4}}$	$2_{\boxed{4}}$	-	$1_{\boxed{2}}$	$2_{\boxed{-8}}$	6	-12	24	-2
-1	68	$3_{\boxed{6}}$	$2_{\boxed{2}}$	1_{0}	-	$1_{\boxed{-2}}$	7	- 7	7	6
0	70	-	$1_{\boxed{0}}$	-	$1_{\boxed{0}}$	$2_{\boxed{0}}$	4	0	0	0
1	72	$1_{\boxed{-2}}$	$2_{\boxed{-2}}$	1_{0}	$2_{\boxed{2}}$	$1_{\boxed{2}}$	7	7	7	0
2	74	$3_{\boxed{-12}}$	$1_{\boxed{-2}}$	2_{0}	$1_{\boxed{2}}$	$2_{\boxed{8}}$	9	18	36	-4
f_x		8	8	4	5	8	33	6	74	0
$f_x u_x$		-16	-8	0	5	16	-3			
$f_x u_x^2$		32	8	0	5	32	77			
$f u_x u_y$		-4	2	0	2	0	0			

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x}) \qquad \begin{cases} \overline{x} = x_o + c_x\,\overline{u}_x = 10 + 5\,\dfrac{-3}{33} = 9{,}5 \\ \overline{y} = y_o + c_y\,\overline{u}_y = 70 + 4\,\dfrac{6}{33} = 70{,}72 \end{cases}$$

$$s_x^2 = c_x^2\left(\frac{\sum f_x u_x^2}{N} - \left(\frac{\sum f_x u_x}{N}\right)^2\right) = 10^2\left(\frac{77}{33} - \left(\frac{-3}{33}\right)^2\right) = 229.$$

$$s_y^2 = c_y^2\left(\frac{\sum f_y u_y^2}{N} - \left(\frac{\sum f_y u_y}{N}\right)^2\right) = 4^2\left(\frac{74}{33} - \left(\frac{6}{33}\right)^2\right) = 35,3.$$

$$s_{xy} = c_{xy}\left(\frac{\sum f u_x u_y}{N} - \frac{\sum f_x u_x}{N}\frac{\sum f_y u_y}{N}\right) = 10.4\left(\frac{0}{33} - \frac{-3}{33}\,\frac{6}{33}\right) =$$

$$= -\,0,66.$$

a) Recta de regresión de y sobre x:

$$y - 70,72 = \frac{-0,66}{229}(x - 9,5); \qquad y = -0,003\,x + 70,7$$

b) Recta de regresión de x sobre y:

$$x - 9,5 = \frac{-0,66}{35,3}(y - 70,72) \qquad x = -0,02\,y + 10,8.$$

c) Coeficiente de correlación lineal:

$$r = \frac{s_{xy}}{\sqrt{s_x^2\, s_y^2}} = \frac{-\,0,66}{\sqrt{229.35,3}} = -\,0,007.$$

Al ser r muy próximo a cero, se puede afirmar que las variables x e y estan practicamente incorreladas, es decir se trata de un caso de independencia aleatoria.

****

5.10 *La tabla adjunta muestra la distribución de 113 individuos varones, donde x_i representa el perímetro torácico en cms. e y_i representa el peso en kgrs.*

a) *Calcular la recta de regresión de peso sobre perímetro torácico.*
b) *Calcular la recta de regresión de perímetro torácico sobre peso.*
c) *Estimar que peso corresponderá a un indivicuo que mide 105 cms de perímetro torácico.*
d) *Estimar que perímetro torácico corresponderá a un individuo que pesa 91 kgrs.*
e) *Calcular el coeficiente de correlación lineal y analizar el tipo de dependencia que existe entre ambas variables.*

	<75	75-79	80-84	85-89	90-94	95-100	>100
55-59	1		3				
60-64		1	2	7	1		1
65-69			2	4	8	6	4
70-74				5	13	10	11
> 75					7	12	15

S O L U C I O N:

	u_x	-3	-2	-1	0	1	2	3	f_y	$f_y u_y$	$f_y u_y^2$	$f u_x u_y$
u_y	y \ x	< 75	75-79	80-84	85-89	90-94	95-100	>100				
-2	55-59	1 $_{6}$		3 $_{6}$					4	-8	16	12
-1	60-64		1 $_{2}$	2 $_{2}$	7 $_{0}$	1 $_{-1}$		1 $_{-3}$	12	-12	12	0
0	65-69			2 $_{0}$	4 $_{0}$	8 $_{0}$	6 $_{0}$	4 $_{0}$	24	0	0	0
1	70-74				5 $_{0}$	13 $_{13}$	10 $_{20}$	11 $_{33}$	39	39	39	66
2	> 75					7 $_{14}$	12 $_{48}$	15 $_{90}$	34	68	136	152
f_x		1	1	7	16	29	28	31	**113**	87	203	230
$f_x u_x$		-3	-2	-7	0	29	56	93	166			
$f_x u_x^2$		9	4	7	0	29	112	279	440			
$f u_x u_y$		6	2	8	0	26	68	120	230			

$$\bar{x} = x_o + c_x \bar{u}_x = 87 + 5\,\frac{166}{113} = 94{,}35; \quad \bar{y} = y_o + c_y \bar{u}_y = 67 + 5\,\frac{87}{113} = 70{,}85.$$

$$s_x^2 = c_x^2\left(\frac{\sum f_x u_x^2}{N} - \left(\frac{\sum f_x u_x}{N}\right)^2\right) = 5^2\left(\frac{440}{113} - \left(\frac{166}{113}\right)^2\right) = 43{,}4.$$

$$s_y^2 = c_y^2\left(\frac{\sum f_y u_y^2}{N} - \left(\frac{\sum f_y u_y}{N}\right)^2\right) = 5^2\left(\frac{203}{113} - \left(\frac{87}{113}\right)^2\right) = 30.$$

$$s_{xy} = c_x c_y\left(\frac{\sum f u_x u_y}{N} - \frac{\sum f_x u_x}{N}\,\frac{\sum f_y u_y}{N}\right) = 5.5\left(\frac{230}{113} - \frac{166.87}{113.113}\right) = 22{,}6.$$

a) Recta de regresión de peso sobre perímetro torácico, es decir de y sobre x.

$$y - 70,85 = \frac{22,6}{43,4}(x - 94,35); \qquad y = 0,52 + 21,71 \qquad [1]$$

b) Recta de regresión de perímetro sobre peso, es decir de x sobre y

$$x - 94,35 = \frac{22,6}{30}(y - 70,85); \qquad x = 0,75\,y + 40,97 \qquad [2]$$

c) Peso (y) que corresponde a 105 cms de perímetro torácico:
Sustituyendo en [1] resulta: $y = 0,52.105 + 21,71 = 76,3$ Kgrs.

d) Perímetro torácico (x) que corresponde a 91 Kgrs.
Sustituyendo en [2] resulta: $x = 0,75.91 + 40,97 = 109,2$ cms.

e) El coeficiente de correlación lineal, es la media geométrica de los coeficientes de regresión lineal. Es decir:

$$r = \sqrt{0,52.0,75} = 0,62 \qquad \text{o también,}$$

$$r = \frac{s_{xy}}{\sqrt{s_x^2\, s_y^2}} = \frac{22,6}{\sqrt{43,4.30}} = 0,62$$

Luego las dos variables están en dependencia aleatoria.

****

6- estimación de parámetros

1. **INTRODUCCION**

2. **DEFINICIONES**

   2.1 Estimador
   2.2 Estimación
   2.3 Estimador por punto
   2.4 Estimación puntual
   2.5 Estimador por intervalo
   2.6 Estimación por intervalo
   2.7 Coeficiente de confianza
   2.8 Bondad de un estimador puntual
   2.9 Bondad de un estimador por intervalo

3. **ESTIMADORES POR PUNTO MAS USUALES**

   3.1 Estimador del parámetro p de la distribución binomial
   3.2 Estimador del parámetro λ de la distribución de Poisson
   3.3 Estimador del parámetro μ de la distribución normal
   3.4 Estimador del parámetro σ^2 de la distribución normal

4. **DISTRIBUCION EN EL MUESTREO DE ESTOS ESTIMADORES**

   4.1 Distribución en el muestreo de $\hat{p}$
   4.2 Distribución en el muestreo de $\hat{\lambda}$
   4.3 Distribución en el muestreo de $\hat{\mu}$
   4.4 Distribución en el muestreo de $\hat{\sigma}^2$

5. DISTRIBUCIONES ASOCIADAS AL ESTUDIO DE DOS POBLACIONES NORMALES E INDEPENDIENTES

- 5.1 Distribución de la diferencia de medias de muestras de dos poblaciones normales e independientes con desviaciones típicas conocidas.
- 5.2 Distribución de la diferencia de medias con desviaciones típicas desconocidas
  - 5.2.1 Muestras de tamaño grande
  - 5.2.2 Muestras de tamaño pequeño
    - 5.2.2.1 Desviaciones típicas desconocidas pero iguales
    - 5.2.2.2 Desviaciones típicas desconocidas y distintas
- 5.3 Distribución de la razón de varianzas

6. CONSTRUCCION DE INTERVALOS DE CONFIANZA

- 6.1 Intervalo de confianza para la media μ de una población normal
  - 6.1.1 σ conocida
  - 6.1.2 σ desconocida
    - 6.1.2.1 Muestras grandes
    - 6.1.2.2 Muestras pequeñas
- 6.2 Intervalo de confianza para la varianza σ^2 de una población normal
- 6.3 Intervalo de confianza para la desviación típica σ de una población normal
- 6.4 Intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones normales
  - 6.4.1 Desviaciones típicas conocidas
  - 6.4.2 Desviaciones típicas desconocidas
    - 6.4.2.1 Muestras grandes
    - 6.4.2.2 Muestras pequeñas
      - 6.4.2.2.1 Desviaciones típicas desconocidas pero iguales
      - 6.4.2.2.2 Desviaciones típicas distintas y desconocidas
- 6.5 Intervalo de confianza para la razón de varianzas de dos poblaciones normales
- 6.6 Intervalo de confianza para el parámetro p de una distribución binomial
- 6.7 Intervalo de confianza para la diferencia entre los parámetros p_1 y p_2 de dos distribuciones binomiales

6 - ESTIMACION DE PARAMETROS

1. INTRODUCCION

En la teoría del muestreo obtenemos información acerca de muestras extraidas de poblaciones conocidas, sin embargo, prácticamente es más interesante *inferir* información sobre una población basándonos en la información contenida en una muestra. Puesto que las poblaciones que hemos estudiado, binomial, Poisson, normal, etc, quedan determinadas por sus parámetros, podemos hacer inferencias sobre la población haciendo inferencias acerca de sus parámetros.

Consideraremos dos métodos para hacer inferencia, que son los siguientes:

- Estimación
- Contraste de hipótesis

Cuando en una población desconocemos el valor de un parámetro y tomamos el valor que nos proporciona un estadístico para una muestra de esa población, como valor aproximado de dicho parámetro, estamos ante un problema de inferencia de la categoría de la estimación. Así por ejemplo, si en una población desconocemos su media y tomamos como valor de ese parámetro el que nos da el estadístico media muestral para una muestra particular, estamos ante un problema de estimación.

Cuando realizamos la estimación de un parámetro, es muy útil obtener una medida del posible error de esa estimación, o sea, señalar un intervalo de valores dentro del cual se tiene la confianza de que se encuentre el parámetro estimado.

Cuando por el contrario hacemos alguna afirmación sobre una población, sobre su forma o sobre el valor númerico de sus parámetros que contrastaremos mediante una muestra aleatoria extraida de la población, entonces estamos realizando inferencias de la categoría del contraste de hipótesis.

2. DEFINICIONES

2.1 ESTIMADOR.- Se llama *estimador* a una función de las observaciones muestrales, es decir, un estadístico que nos pérmite obtener un valor aproximado de alguna característica de la población. Por tanto, es una variable aleatoria en el muestreo, que tendrá su correspondiente función de distribución y ley de probabilidad.

2.2 ESTIMACION.- Llamaremos *estimación*, al valor del estimador para una muestra particular. Por tanto, es un número.

2.3 ESTIMADOR POR PUNTO.-Llamaremos *estimador por punto* a una función de una muestra extraida de una población, que nos permitirá obtener "un valor" aproximado de algún parámetro de esa población. Por tanto es una variable aleatoria que tiene su distribución en el muestreo.

2.4 ESTIMACION PUNTUAL.- Es el valor númerico que toma el estimador puntual para una muestra particular.

2.5 ESTIMADOR POR INTERVALO.- Llamaremos *estimador por intervalo* a dos funciones de la muestra aleatoria que nos permiten obtener los dos límites inferior y superior de dicho intervalo.

2.6 ESTIMACION POR INTERVALO.-Son los valores que toma el estimador por intervalo para una muestra particular.

2.7 COEFICIENTE DE CONFIANZA.-Se llama *coeficiente de confianza* a la probabilidad de que un estimador por intervalo cubra el verdadero valor del parámetro que se pretende estimar. Fijado un valor, por ejemplo 0,95 , como coeficiente de confianza para una muestra particular, obtendremos un intervalo de confianza que puede o no cubrir el valor del parámetro a estimar. Para muestras sucesivas, obtendriamos otros intervalos de confianza y si hicieramos una larga serie de determinaciones de estos intervalos, se cumpliría que el 95% de los mismos cubririan el valor del parámetro a estimar.

En la práctica, suele fijarse como coeficiente de confianza $1 - \alpha = 0,95$; $1 - \alpha = 0,99$; $1 - \alpha = 0,999$, según el grado de precisión que se desee.

2.8 BONDAD DE UN ESTIMADOR PUNTUAL.-Esta claro, que lo ideal sería que los valores de un estimador para las distintas muestras nos dieran un valor exacto del parámetro a estimar. Ahora bien, en general esto no es posible por lo que deberemos exigir a estos estimadores que cumplan ciertas condiciones.

Los criterios para medir la bondad de un estimador puntual son los siguientes:

1°) Que sea *insesgado*, es decir, que su media coincida con el valor del parámetro a estimar.

2°) Que sea *eficiente*, es decir, que siendo insesgado, su varianza sea mínima.

2.9 BONDAD DE UN ESTIMADOR POR INTERVALO.- *La bondad de un estimador por intervalo* viene dada por la mayor o menor longitud del intervalo para un nivel de confianza dado.

3. ESTIMADORES POR PUNTO MAS USUALES

A continuación vamos a dar una serie de funciones de una muestra aleatoria extraida de la población , que nos permitirán obtener un valor aproximado de los parámetros de las distribuciones que hemos estudiado en los capítulos anteriores.

3.1 ESTIMADOR DEL PARAMETRO p DE UNA DISTRIBUCION BINOMIAL.- Representaremos el estimador de p por $\hat{p}$, del mismo modo haremos para los demás estimadores.

Si en una muestra aleatoria de tamaño n, se ha presentado r veces un suceso dado, entonces $\hat{p} = \frac{r}{n}$

Por tanto, el estimador que tomamos de p es la frecuencia relativa del suceso favorable.

3.2 ESTIMADOR DEL PARAMETRO λ DE UNA DISTRIBUCION DE POISSON.-El estimador del parámetro λ de una distribución de Poisson, que representaremos por $\hat{\lambda}$, es la media aritmética de la la variable aleatoria X en la muestra.

Es decir: $$\hat{\lambda} = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Esto equivale a estimar la media poblacional a partir de la media muestral.

3.3 ESTIMADOR DEL PARAMETRO μ DE UNA DISTRIBUCION NORMAL.- Sea X una variable aleatoria que sigue una distribución $N(\mu , \sigma)$ y queremos encontrar un estimador para la media poblacional μ.

Como vimos al estudiar la distribución normal, este estimador será la media muestral.

Es decir: $$\hat{\mu} = \overline{x}$$

3.4 ESTIMADOR DEL PARAMETRO σ^2 DE UNA DISTRIBUCION NORMAL.-Análogamente a lo visto en el apartado anterior, sea X una variable aleatoria que sigue una distribución $N(\mu , \sigma)$. El estimador de σ^2 será: $$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

A $\hat{s}^2$, se le conoce con el nombre de *cuasivarianza muestral*.

En resumen: *Cuando n es grande, cualquiera que sea la ley de probabilidad de la variable X, tomaremos como estimador de la media poblacional la media muestral y como estimador de la varianza poblacional la cuasivarianza muestral.*

Todos estos estimadores son insesgados y de varianza mínima.

4. DISTRIBUCION EN EL MUESTREO DE ESTOS ESTIMADORES

Un estimador es una función de la variable aleatoria n-dimensio nal. Está claro, por tanto, que para cada muestra el estimador tomará un valor númerico, y si consideramos todas las posibles muestras que podemos extraer de la población, obtendremos un conjunto de valores del estimador que seguirá una determinada distribución; a esta distribución se la llama *distribución muestral o distribución en el muestreo.*

4.1 DISTRIBUCION EN EL MUESTREO DE $\hat{p}$.- Supuesto que el tamaño de la muestra n es "suficientemente grande", y que p no se acerca a 0 ni a 1, entonces el estimador $\hat{p}$ se aproxima a una distribución $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Tipificando la variable, se tiene que $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ se aproxima a una distribución $N(0, 1)$

Ahora bien si n no es grande o si p se acerca a 0 ó a 1, entonces deberemos calcular mediante la distribución binomial la probabilidad de la composición particular de cada muestra.

4.2 DISTRIBUCION EN EL MUESTREO DE $\hat{\lambda}$.- El estimador $\hat{\lambda}$ de la media de la distribución de Poisson, sigue otra distribución de Poisson de parámetro λ.

4.3 DISTRIBUCION EN EL MUESTREO DE $\hat{\mu}$.- Si $\hat{\mu}$ es el estimador de la media de una distribución $N(\mu, \sigma)$, su distribución en el muestreo es: $N(\mu, \sigma/\sqrt{n})$.

- Si σ es conocida, la variable tipificada $Z = \dfrac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}$ se distribuye segun una ley $N(0, 1)$ cualquiera que sea el tamaño n de la muestra.

- Si σ es desconocida, la estimaremos mediante la cuasivarianza muestral: $\hat{s}^2 = \dfrac{1}{n-1} \sum (x_i - \bar{x})^2$.

Si conocemos la varianza de la muestra s^2, tenemos que:

$$\hat{s}^2 = \frac{n}{n-1} s^2$$

La expresión $\dfrac{\hat{\mu} - \mu}{\hat{s}/\sqrt{n}}$ obedece a una distribución de Student con n-1 grados de libertad.

Como ya vimos al estudiar la distribución t de Student, si n es grande la expresión anterior se aproxima a una distribución N(0 , 1)

4.4 DISTRIBUCION EN EL MUESTREO DE $\hat{s}^2$.- Sea $\hat{s}^2$ el estimador que hemos construido para la varianza de una distribución $N(\mu , \sigma)$. Según vimos en el apartado 3.8 la expresión:

$$\frac{(n - 1)\ \hat{s}^2}{\sigma^2}$$

se distribuye según una χ^2 con (n - 1) grados de libertad.

5. DISTRIBUCIONES ASOCIADAS AL ESTUDIO DE DOS POBLACIONES NORMALES INDEPENDIENTES

5.1 DISTRIBUCION DE LA DIFERENCIA DE MEDIAS DE MUESTRAS DE DOS POBLACIONES NORMALES INDEPENDIENTES CON DESVIACIONES TIPICAS CONOCIDAS .- Sean X_1 una variable aleatoria que sigue una distribución $N(\mu_1 , \sigma_1)$ y X_2 otra variable aleatoria que sigue una distribución $N(\mu_2 , \sigma_2)$ e independiente con X_1. Elegimos una muestra aleatoria simple de tamaño n_1 de X_1 y otra muestra aleatoria simple de tamaño n_2 de X_2.

La distribución que sigue la diferencia de medias $\overline{x}_1 - \overline{x}_2$ siendo $\overline{x}_1$ y $\overline{x}_2$ las medias muestrales es:

$$N\left(\mu_1 - \mu_2 \ , \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Tipificando la variable, tendremos que

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$ sigue una N(0 , 1)

5.2 DISTRIBUCION DE LA DIFERENCIA DE MEDIAS EN EL CASO DE DESVIACIONES TIPICAS DESCONOCIDAS.

5.2.1 MUESTRAS DE TAMAÑO GRANDE.- En las mismas condiciones del apartado anterior, si n_1 y n_2 son suficientemente grandes y σ_1 y σ_2 desconocidas, las estimaremos a partir de $\hat{s}_1$ y $\hat{s}_2$, con lo que la expresión,

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

sigue aproximadamente una $N(0\ ,\ 1)$.

5.2.2 MUESTRAS DE TAMAÑO PEQUEÑO.

5.2.2.1 Desviaciones típicas desconocidas pero iguales

Es decir: $\sigma_1 = \sigma_2 = \sigma$

En este caso obtendremos una buena estimación de σ^2 a partir de la media aritmetica ponderada de las estimaciones $\hat{s}_1^2$ y $\hat{s}_2^2$:

$$s^2 = \frac{(n_1-1)\ \hat{s}_1^2 + (n_2-1)\ \hat{s}_2^2}{n_1 + n_2 - 2}$$

Y entonces se tiene que la expresión:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

sigue una distribución de Student con $n_1 + n_2 - 2$ grados de libertad.

5.2.2.2 Desviaciones típicas desconocidas y distintas

Estimamos σ_1^2 y σ_2^2 por $\hat{s}_1^2$ y $\hat{s}_2^2$ respectivamente, entonces la expresión,

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

sigue aproximadamente una distribución de Student con f grados de libertad, donde f es la *aproximación de Welch*, que se define del siguiente modo:

$$f = \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{s}_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{\hat{s}_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

tomando para el valor de f el entero más próximo.

5.3 DISTRIBUCION DE LA RAZON DE VARIANZAS.- La variable aleatoria definida por la expresión:

$$\frac{\hat{s}_1^2 \; / \; \sigma_1^2}{\hat{s}_2^2 \; / \; \sigma_2^2}$$

sigue una F de Snedecor con (n_1-1) , (n_2-1) grados de libertad.

6. CONSTRUCCION DE INTERVALOS DE CONFIANZA

Ya hemos visto, que un estimador es una función de las observaciones de la muestra que nos pérmite obtener un valor aproximado de alguna característica de la población.

El problema ahora, consiste en dar una medida de la aproximación de la estimación efectuada, razón por la que acabamos de estudiar la distribución en el muestreo de estos estimadores.

Así por ejemplo, si X es una variable aleatoria que sigue una distribución $N(\mu , \sigma)$ y sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria simple de tamaño n, sabemos que $\bar{x}$ es un estimador para la media μ . También hemos visto que $\bar{x}$ es una variable que sigue una distribución en el muestreo $N(\mu , \sigma/\sqrt{n})$. Al tipificar la variable tendremos que

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad \text{sigue una } N(0 , 1)$$

Si representamos por $z_{\alpha/2}$, el valor de la abcisa de una distribución $N(0 , 1)$ que deja a su derecha $\alpha/2$ de area, sien do $1-\alpha$ lo que hemos llamado coeficiente de confianza, tendremos que:

$$p\left[-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right] = 1-\alpha$$

De aqui obtenemos:

$$p\left[\bar{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Lo cual significa, que el intervalo $\left(\bar{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right)$

cubre el valor de μ con una probabilidad de $1-\alpha$.

Si en este momento tomamos una muestra particular que nos da para $\bar{x}$ un valor $\bar{x}_o$, diremos que

$$\left(\bar{x}_o - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\ ,\ \bar{x}_o + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right)$$

es un *intervalo de confianza*.

Este intervalo cubrirá o no el valor de μ, pero ya no tiene ningún sentido decir que hay una probabilidad de $1-\alpha$ de que lo cubra.

Otros estimadores seguirán distribuciones diferentes de la normal, por lo que representaremos por:

- $t_{\alpha/2,\,n}$.- El valor de la abcisa de una distribución de Student con n grados de libertad que deja a su derecha $\alpha/2$ de area.

- $\chi^2_{\alpha/2,\,n}$.- El valor de la abcisa de una distribución χ^2 con n grados de libertad que deja a su derecha $\alpha/2$ de area.

- $F_{\alpha/2,\,n_1,\,n_2}$.- El valor de la abcisa de una distribución F de Snedecor con n_1, n_2 grados de libertad que deja a su derecha $\alpha/2$ de area.

6.1 INTERVALO DE CONFIANZA PARA LA MEDIA μ DE UNA POBLACION NORMAL

6.1.1 σ conocida.

Ya vimos en 4.3 que $\dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}}$ sigue una N(0 , 1).

Si el coeficiente de confianza es $1 - \alpha$. tendremos:

$$p\left[-z_{\alpha/2} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right] = 1 - \alpha$$

$$p\left[\overline{x} - z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

que nos da el intervalo buscado: $I = \left(\overline{x} \mp z_{\alpha/2}\ \dfrac{\sigma}{\sqrt{n}}\right)$

6.1.2 <u>σ desconocida.</u>

Cuando σ era desconocida vimos que $\dfrac{\overline{x} - \mu}{\hat{s}\ /\sqrt{n}}$ sigue una distribución de Student con

n - 1 grados de libertad, pero si n es grande se aproxima a una N(0 , 1)

6.1.2.1 <u>Muestras grandes</u> Es decir para $n > 30$, tendremos:

$$p\left[-\ z_{\alpha/2} < \frac{\overline{x} - \mu}{\hat{s}\ /\sqrt{n}} < z_{\alpha/2}\right] = 1 - \alpha$$

de donde deducimos:

$$p\left[\overline{x} - z_{\alpha/2}\ \frac{\hat{s}}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\ \frac{\hat{s}}{\sqrt{n}}\right] = 1 - \alpha$$

con lo cual el intervalo buscado será:

$$I = \left(\overline{x} \mp z_{\alpha/2}\ \frac{\hat{s}}{\sqrt{n}}\right)$$

6.1.2.2 <u>Muestras pequeñas.</u>- Es decir $n < 30$, entonces

$$p\left[-t_{\alpha/2,\ n-1} < \frac{\overline{x} - \mu}{\hat{s}\ /\sqrt{n}} < t_{\alpha/2,\ n-1}\right] = 1 - \alpha$$

de donde deducimos:

$$p\left[\overline{x} - t_{\alpha/2,\ n-1}\ \frac{\hat{s}}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,\ n-1}\ \frac{\hat{s}}{\sqrt{n}}\right] = 1-\alpha$$

con lo cual el intervalo buscado será:

$$I = \left(\overline{x} \mp t_{\alpha/2,\ n-1}\ \frac{\hat{s}}{\sqrt{n}}\right)$$

6.2 INTERVALO DE CONFIANZA PARA LA VARIANZA σ^2 DE UNA POBLACION NORMAL.-

Vimos en 4.4 que la expresión:

$$\frac{(n-1)\,\hat{s}^2}{\sigma^2}$$

sigue una distribución χ^2 con n-1 grados de libertad.

Entonces, tendremos que:

$$p\left[\chi^2_{1-\frac{\alpha}{2},\,n-1} < \frac{(n-1)\hat{s}^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},\,n-1}\right] = 1-\alpha$$

de donde deducimos:

$$p\left[\frac{(n-1)\,\hat{s}^2}{\chi^2_{\frac{\alpha}{2},\,n-1}} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},\,n-1}}\right] = 1-\alpha$$

con lo que el intervalo buscado será:

$$I = \left(\frac{(n-1)\hat{s}^2}{\chi^2_{\frac{\alpha}{2},\,n-1}}\;,\;\frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},\,n-1}}\right)$$

6.3 INTERVALO DE CONFIANZA PARA LA DESVIACION TIPICA σ DE UNA POBLACION NORMAL.-

Como consecuencia del apartado anterior, será:

$$I = \left(\sqrt{\frac{(n-1)\hat{s}^2}{\chi^2_{\frac{\alpha}{2},\,n-1}}}\;,\;\sqrt{\frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},\,n-1}}}\right)$$

6.4 INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS $\mu_1 - \mu_2$ DE DOS POBLACIONES NORMALES.-

6.4.1 Desviaciones típicas conocidas.- Supongamos que tenemos dos poblaciones que siguen distribuciones $N(\mu_1\,,\,\sigma_1)$ y $N(\mu_2\,,\,\sigma_2)$, con σ_1 y σ_2 conocidas. Vimos es 5.1 que la expresión: $\dfrac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

sigue una distribución N(0 , 1). Por tanto, tendremos:

$$p\left[-z_{\alpha/2} < \frac{(\overline{x}_1-\overline{x}_2) - (\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+ \dfrac{\sigma_2^2}{n_2}}} < z_{\alpha/2}\right] = 1 - \alpha$$

de donde se deduce:

$$p\left[(\overline{x}_1-\overline{x}_2)-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1-\mu_2 < (\overline{x}_1-\overline{x}_2)+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] =$$

$= 1 - \alpha.$

Con lo que el intervalo buscado será:

$$I = \left((\overline{x}_1-\overline{x}_2)\mp z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

6.4.2 Desviaciones típicas desconocidas.

6.4.2.1. Muestras de tamaño grande .

$\dfrac{(\overline{x}_1-\overline{x}_2) - (\mu_1-\mu_2)}{\sqrt{\dfrac{\hat{s}_1^2}{n_1} + \dfrac{\hat{s}_2^2}{n_2}}}$ se aproxima a una N(0 , 1).

luego:

$$p\left[-z_{\frac{\alpha}{2}} < \frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\hat{s}_1^2}{n_1} + \dfrac{\hat{s}_2^2}{n_2}}} < z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

de donde deducimos:

$$p\left[(\bar{x}_1-\bar{x}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}}<\mu_1-\mu_2<(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}}\right]=1-\alpha$$

Con lo que el intervalo buscado será:

$$I=\left((\bar{x}_1-\bar{x}_2)\mp z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}}\right)$$

6.4.2.2 Muestras pequeñas.

6.4.2.2.1 Desviaciones típicas desconocidas pero iguales.

$\sigma_1=\sigma_2$. En 5.2.2.1, vimos que la expresión:

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

sigue una distribución t de Student

con n_1+n_2-2 grados de libertad. Entonces:

$$p\left[-t_{\frac{\alpha}{2},n_1+n_2-2}<\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<t_{\frac{\alpha}{2},n_1+n_2-2}\right]=$$

$=1-\alpha$. De donde se deduce:

$$p\left[(\bar{x}_1-\bar{x}_2)-t_{\frac{\alpha}{2},n_1+n_2-2}\,s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}<\mu_1-\mu_2<(\bar{x}_1-\bar{x}_2)+t_{\frac{\alpha}{2},n_1+n_2-2}\,s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]=$$

$=1-\alpha$

Con lo que el intervalo buscado será:

$$I=\left((\bar{x}_1-\bar{x}_2)\mp t_{\frac{\alpha}{2},n_1+n_2-2}\,s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

6.4.2.2.2 Desviaciones típicas distintas y desconocidas.

En 5.2.2.2 vimos que la expresión:

$$\frac{(\bar{x}_1-\bar{x}_2) - (\mu_1-\mu_2)}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

sigue aproximadamente una distribución t de Student con f grados de libertad, siendo f la aproximación de Welch vista en 5.2.2.2.

Entonces:

$$p\left[-t_{\frac{\alpha}{2}, f} < \frac{(\bar{x}_1-\bar{x}_2) - (\mu_1-\mu_2)}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} < t_{\frac{\alpha}{2}, f}\right] = 1 - \alpha$$

de donde se deduce:

$$p\left[(\bar{x}_1-\bar{x}_2)-t_{\frac{\alpha}{2},f}\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} < \mu_1-\mu_2 < (\bar{x}_1-\bar{x}_2)+t_{\frac{\alpha}{2},f}\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\right] =$$

$$= 1 - \alpha$$

con lo que el intervalo buscado será:

$$I = \left((\bar{x}_1-\bar{x}_2) \mp t_{\frac{\alpha}{2}, f}\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\right)$$

6.5 INTERVALO DE CONFIANZA PARA LA RAZON DE VARIANZAS DE DOS POBLACIONES NORMALES.- En 5.3, vimos que la expresión

$$\frac{\hat{s}_1^2/\sigma_1^2}{\hat{s}_2^2/\sigma_2^2}$$

sigue una distribución F de Snedecor con n_1-1, n_2-1 grados de libertad.

Entonces $$p\left[F_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{\hat{s}_1^2/\sigma_1^2}{\hat{s}_2^2/\sigma_2^2} < F_{\frac{\alpha}{2},n_1-1,n_2-1}\right] = 1-\alpha$$

de donde se deduce;

$$p\left[\frac{1}{F_{\frac{\alpha}{2},n_1-1,n_2-1}}\cdot\frac{\hat{s}_1^2}{\hat{s}_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}\cdot\frac{\hat{s}_1^2}{\hat{s}_2^2}\right] = 1-\alpha$$

Con lo que el intervalo buscado será:

$$I = \left(\frac{\hat{s}_1^2/\hat{s}_2^2}{F_{\frac{\alpha}{2},n_1-1,n_2-1}}\ ,\ \frac{\hat{s}_1^2/\hat{s}_2^2}{F_{1-\frac{\alpha}{2},\ n_1-1,n_2-1}}\right)$$

6.6 <u>INTERVALO DE CONFIANZA PARA EL PARAMETRO p DE UNA DISTRIBUCION BINOMIAL</u>.-

En 4.1, vimos que la expresión:

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

se aproxima a una distribución $N(0\ ,\ 1)$ para n grande y p no próximo a 0 o a 1.

Entonces $$p\left[-z_{\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

de donde se deduce:

$$p\left[\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right] = 1-\alpha$$

Si la muestra es grande, tomaremos como valor estimado muy aproximado de p, $\hat{p}$. Con lo que el intervalo será: $I = \left(\hat{p} \mp z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

En el caso de que no se cumplan las condiciones anteriores, se han desarrollado diversos métodos gráficos y analíticos para calcular el intervalo de confianza de p.

A continuación damos un intervalo de confianza para p:

$$I = \left(\frac{r}{r + (n-r+1)\, F_{\frac{\alpha}{2},\, 2(n-r+1),2r}} \;,\; \frac{(r+1)\,F_{\frac{\alpha}{2},2(r+1),2(n-r)}}{(n-r)+(r+1)\,F_{\frac{\alpha}{2},2(r+1),2(n-r)}} \right)$$

6.7 INTERVALO DE CONFIANZA PARA LA DIFERENCIA ENTRE LOS PARAMETROS p_1 y p_2 DE DOS DISTRIBUCIONES BINOMIALES.- En las mismas condiciones del apartado anterior, es decir, si los tamaños de las muestras n_1 y n_2 son grandes y $\hat{p}_1$ y $\hat{p}_2$ son los valores estimados de p_1 y p_2 respectivamente, entonces el intervalo de confianza será:

$$I = \left((\hat{p}_1-\hat{p}_2) \mp z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

****

PROBLEMAS RESUELTOS

6.1 *Se sospecha que el número de unidades que contiene cada dosis de un medicamento, no llegan a las 10.000 que se indican en el envase. El laboratorio que lo fabrica afirma que el contenido medio de la dosis es de 10.000 unidades. Para comprobarlo tomamos al azar 100 dosis y determinamos el número de unidades de cada una obteniendo de media 9.940 unidades y de desviación típica 120 unidades.*

Si suponemos que la distribución del número de dosis en la población es normal, ¿qué podemos decir acerca de la afirmación del laboratorio para un nivel de confianza del 99%?

SOLUCION:

El intervalo de confianza para la media μ de una población normal de σ desconocida y para muestras grandes, viene dado por la siguiente expresión:

$$\left(\bar{x} \mp z_{\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}} \right)$$

donde $\bar{x} = 9940$; $\alpha = 0{,}01$; $z_{\frac{\alpha}{2}} = 2{,}57$; $n = 100$

$\hat{s}$ podemos suponerlo igual a s para valores grandes de n, es decir para valores de $n > 30$. Por tanto $\hat{s} = s = 120$.

Sustituyendo estos datos en la expresión del intervalo, tendremos,

$$(9909{,}16 \quad , \quad 9970{,}84)$$

Como 10000 no queda comprendido dentro del intervalo, podemos asegurar con un nivel de confianza del 99%, que la afirmación del laboratorio es falsa.

****

6.2 *Queremos conocer, con un nivel de confianza del 95%, entre que valores expresados en días estará comprendida la permanencia media de los enfermos en un gran hospital. Para ello tomamos una muestra al azar de 300 enfermos y anotamos el número de dias de permanencia de cada uno de ellos. Calculada la media y la desviación típica de la muestra se obtuvo 8 y 12 dias respectivamente.*

SOLUCION:

Los límites de confianza de la permanencia media de los enfermos en el hospital vienen dados por: $\left(\bar{x} \mp z_{\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}}\right)$

donde $\bar{x} = 8$ dias; $\hat{s} = 12$ dias; $n = 300$; $\alpha = 0{,}05$

$z_{\alpha/2} = 1{,}96$.

Sustituyendo en la expresión anterior, resulta:

$$(6{,}64 \quad , \quad 9{,}36)$$

Luego la permanencia media de los enfermos de un gran hospital estará comprendida entre esos dias en el 95% de los casos.

****

6.3 *Se afirma que la estatura media de las personas adultas de una determinada región es de 1,80 metros. Queremos tener una confianza del 99% en saber si la afirmación anterior es correcta o errónea. Para ello tomamos una muestra al azar de 100 personas adultas a las que medimos sus altura, obteniendo de media 1,78 metros y de desviación típica 0,10 metros. Suponemos que la variable objeto de estudio es normal.*

SOLUCION:

El intervalo de confianza para la media de una población normal con σ desconocida y n grande, viene dado por la siguiente expresión:

$$\left(\bar{x} \mp z_{\frac{\alpha}{2}} \frac{\hat{s}}{\sqrt{n}} \right)$$

donde $\bar{x} = 1{,}78$; $\hat{s} = 0{,}10$; $n = 100$; $\alpha = 0{,}01$; $z_{\frac{\alpha}{2}} = 2{,}57$

Sustituyendo en la expresión anterior, resulta:

$$(1{,}754 \quad , \quad 1{,}806)$$

Por estar 1,80 comprendido dentro del intervalo, admitiremos la afirmación propuesta.

****

6.4 *Se quiere probar la efectividad de un antitérmico en reducir la temperatura. Para ello se tomó la temperatura de 10 niños de 4 años de edad afectados de gripe, antes y después de haberles suministrado el antitérmico y se obtuvieron las siguientes reducciones de temperatura: 1,2; 1,7; 1,6; 1,7; 1; 1; 1; 2,6; 3; 1.*

Suponiendo que la variable reducción de temperatura es normal, hallar un intervalo de confianza del 95% para la media.

SOLUCION:

El intervalo de confianza para la media de una población normal con σ desconocida y muestra pequeña, viene dado por la expresión:

$$\left(\bar{x} \mp t_{\frac{\alpha}{2},\, n-1} \frac{\hat{s}}{\sqrt{n}} \right)$$

donde $\bar{x} = 1{,}58$; $\hat{s} = 0{,}71$; $\alpha = 0{,}05$; $t_{\alpha/2,\, n-1} = 2{,}262$.

Sustituyendo en la expresión anterior, resulta:

$$(1{,}07 \quad , \quad 2{,}09)$$

Por tanto, el antitérmico es efectivo en la reducción de la temperatura al nivel del 95%.

****

6.5 *Una muestra de 20 cigarrillos de una cierta marca, son probados para estudiar el contenido de alquitran, obteniendo una media de 22 mlg. y una desviación típica de 4 mlg.*

Encontrar un intervalo de confianza del 95% para la media de la variable de respuesta elegida, suponiendo que la distribución de dicha variable en la población es normal.

S O L U C I O N:

El intervalo de confianza para la media de una población normal con σ desconocida y muestra pequeña viene dado por la expresión:

$$\left(\bar{x} \mp t_{\frac{\alpha}{2},\, n-1} \frac{\hat{s}}{\sqrt{n}}\right)$$

donde $\bar{x} = 22$; $\alpha = 0{,}05$; $t_{\alpha/2,\, n-1} = 2{,}093$; $\hat{s} = \sqrt{\frac{n}{n-1}}\, s = \sqrt{\frac{20}{19}}\, 4 = 4{,}1$.

Sustituyendo en la expresión anterior, resulta:

(20,08 , 23,92)

****

6.6 *En la tabla adjunta se representan los valores del pH, de una solución en 10 determinaciones diferentes:*
6,80; 6,78; 6,77; 6,80; 6,78; 6,80; 6,82; 6,81; 6,80; 6,79.

Suponiendo normal la distribución de la población de todas las determinaciones del pH de esa solución, encontrar un intervalo de confianza al 95% para la varianza poblacional.

S O L U C I O N:

El intervalo de confianza para la varianza poblacional viene dado por:

$$\left(\frac{(n-1)\hat{s}^2}{\chi^2_{\frac{\alpha}{2},\, n-1}} \; , \; \frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},\, n-1}}\right)$$

donde $\bar{x} = 6{,}795$; $\hat{s}^2 = 227{,}77$; $n = 10$; $\alpha = 0{,}05$;

$\chi^2_{\alpha/2,\, n-1} = 19{,}0$; $\chi^2_{1-\frac{\alpha}{2},\, n-1} = 2{,}70$.

Sustituyendo estos valores en la expresión anterior, resulta:

(107,89 , 759,26)

****

6.7 *Las edades en que se produce la muerte, para una muestra aleatoria de 19 individuos fallecidos de tuberculosis dan una media de 50 años y una desviación típica muestral de 6 años. Suponiendo normal la distribución, se pide:*

a) Estimaciones por puntos no sesgados de la media y la varianza.

b) Hallar un intervalo de confianza para la media al nivel del 99%.

c) Hallar un intervalo de confianza del 99% para la desviación típica.

SOLUCION:

a) Sabemos que el estimador insesgado de la media poblacional es la media muestral por tanto $\bar{x} = 50$.

Un estimador insesgado para la varianza poblacional es la cuasivarianza muestral, luego $\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{19}{18} 36 = 38$.

b) El intervalo de confianza para la media de una población normal σ desconocida y muestra pequeña, viene dado por:

$$\left(\bar{x} \mp t_{\frac{\alpha}{2},\, n-1} \frac{\hat{s}}{\sqrt{n}}\right)$$

donde $\alpha = 0{,}01$; $t_{\alpha/2,\, n-1} = 2{,}878$.

Sustituyendo estos valores en la expresión anterior, resulta:

$$(45{,}93 \quad , \quad 54{,}07)$$

c) El intervalo de confianza para σ viene dado por la expresión:

$$\left(\sqrt{\frac{(n-1)\,\hat{s}^2}{\chi^2_{\frac{\alpha}{2},\, n-1}}} \quad , \quad \sqrt{\frac{(n-1)\,\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},\, n-1}}}\right)$$

donde $\alpha = 0{,}01$; $\chi^2_{\alpha/2,\, n-1} = 37{,}156$;

$\chi^2_{1-\frac{\alpha}{2},\, n-1} = 6{,}265$.

Sustituyendo estos valores en la expresión anterior, resulta:

$$(4{,}29 \quad , \quad 10{,}45)$$

****

6.8 *En un Instituto de Investigaciones Dermatológicas se está investigando una afección cutanea de tipo cuncerígeno. Se eligen 20 ratas de una misma raza aleatoriamente y se les provoca el cancer citado; a continuación se les frota con un medicamento. Se elige como variable de respuesta, el número de horas que tarda el cancer en desaparecer. Se obtuvieron los siguientes resultados:*

$\bar{x} = 10$ *horas* *y* $s = 101$ *horas*

Se admite que la variable de respuesta sigue una distribución normal. Se pide:

a) Calcular el intervalo de confianza para la media de la variable de respuesta, al nivel del 90%.

b) Si $\sigma = 99$ horas, calcular un intervalo de confianza al 90% para la media de la variable de respuesta elegida.

c) ¿Qué tamaño de muestra se necesita para que al nivel de confianza del 95%, la longitud del intervalo sea de 5 horas, supuesto $\sigma = 99$ horas?

SOLUCION:

a) El intervalo de confianza para la media de una población normal de σ desconocida y muestras pequeñas, viene dado por:

$$\left(\bar{x} \mp t_{\frac{\alpha}{2},\, n-1} \frac{\hat{s}}{\sqrt{n}}\right)$$

donde $\alpha = 0{,}1$; $n = 20$; $t_{\alpha/2,\, n-1} = 1{,}729$; $\hat{s} = \sqrt{\frac{n}{n-1}}\, s = 103{,}62$.

Sustituyendo estos valores en la expresión anterior, resulta:

$$(-30{,}06 \quad , \quad 50{,}06)$$

b) El intervalo de confianza para la media de una población normal σ conocida, viene dado por:

$$\left(\bar{x} \mp z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

donde $z_{\alpha/2} = 1{,}64$. Sustituyendo obtenemos:

$$(-26{,}30 \quad , \quad 46{,}30)$$

c) La longitud del intervalo en este caso particular viene dada por:

$$\ell = 2\, z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

donde $\alpha = 0{,}05$; $z_{\alpha/2} = 1{,}96$; $\ell = 5$.

Sustituyendo en la expresión anterior se obtiene $n = 6025$.

****

6.9 *Las presiones sistólicas de dos grupos de niños, para el primero de los cuales sus padres son hipertensos y para el segundo normales, dan los siguientes valores:*

Grupo 1°: 100; 102; 96; 106; 110; 110; 120; 112; 112; 90;

Grupo 2°: 104; 88; 100; 98; 102; 92; 96; 100; 96; 96.

Suponiendo que las dos poblaciones son normales y de varianzas iguales y desconocidas, calcular un intervalo de confianza del 95% para la diferencia de medias.

SOLUCION:

Para el primer grupo de niños tenemos: $\bar{x}_1 = 105{,}8$; $\hat{s}_1 = 8{,}87$; $n_1 = 10$.

Para el segundo grupo tendremos: $\bar{x}_2 = 97{,}2$; $\hat{s}_2 = 4{,}78$; $n_2 = 10$.

El intervalo de confianza para la diferencia de medias viene dado por:

$$\left((\bar{x}_1 - \bar{x}_2) \mp t_{\frac{\alpha}{2},\, n_1+n_2-2} \; s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

siendo:

$$s = \sqrt{\frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2}}$$

donde $\alpha = 0{,}05$; $t_{\alpha/2,\, n_1+n_2-2} = 2{,}101$; $s = 7{,}124$

Por tanto, el intervalo de confianza será:

$$(1{,}91 \quad , \quad 15{,}29)$$

Como el intervalo de confianza no cubre el cero se deduce que la diferencia de medias es significativa.

****

6.10 *De una población de personas comparables con exceso de peso se seleccionan dos grupos A y B de 100 y 50 individuos respectivamente. A los individuos del grupo A, se les suministra una nueva dieta D_1 con la que sufren una pérdida media de peso al cabo de un mes de 7,9 Kgrs. con una desviación típica de 0,2 Kgrs. A los individuos del grupo B, se les suministra una dieta D_2 con la que sufren una pérdida media de peso al cabo de un mes de 6,8 Kgrs. con una desviación típica de 0,3 Kgrs.*

Hallar los límites de confianza del 95% para la diferencia del número medio de Kgrs.perdidos producido por el suministro de las dos dietas D_1 y D_2.

SOLUCION:

Los límites de confianza para la diferencia de medias, siendo las desviaciones típicas desconocidas y para muestras de tamaño grande, vienen dados por:

$$\left((\bar{x}_1-\bar{x}_2) \mp z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right)$$

donde $\bar{x}_1 = 7{,}9;$ $\hat{s}_1 = 0{,}2;$ $n_1 = 100;$ $\alpha = 0{,}05$

$\bar{x}_2 = 6{,}8;$ $\hat{s}_2 = 0{,}3;$ $n_2 = 50;$ $z_{\alpha/2} = 1{,}96.$

Sustituyendo estos valores en la expresión anterior, resulta:

$$(1{,}008 \quad , \quad 1{,}192)$$

Por tanto, la dieta D_1 es más efectiva para la pérdida de peso que la dieta D_2 ya que la media de peso perdido es superior en 1Kgr. con la D_1 que con la D_2.

****

6.11 *Con el fin de comparar el contenido medio de alquitrán por cigarrillo de dos marcas de tabaco, se tomaron dos muestras de 7 y 8 cigarrillos cada una respectivamente, y se midió su contenido de alquitrán que expresado en mlgrs, dió los siguientes valores:*

20, 24, 23, 22, 22, 20, 23.

19, 22, 20, 18, 20, 22, 20, 20.

Suponiendo que el contenido de alquitrán en ambas marcas de tabaco son normales y que las varianzas son iguales y desconocidas, encontrar un intervalo de confianza del 99% para la diferencia de medias.

SOLUCION:

El intervalo de confianza para la diferencia de medias, siendo las desviaciones típicas desconocidas pero iguales y muestras pequeñas, viene dado por:

$$\left((\bar{x}_1-\bar{x}_2) \mp t_{\frac{\alpha}{2},\, n_1+n_2-2} \; s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

siendo

$$s^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1 + n_2 - 2}$$

Para la 1ª muestra tenemos: $\bar{x}_1 = 22;$ $\hat{s}_1 = 1{,}527;$ $n_1 = 7.$

Para la 2ª muestra tenemos: $\bar{x}_2 = 20{,}125;$ $\hat{s}_2 = 1{,}356;$ $n_2 = 8.$

Por tanto, $s^2 = 2{,}067;$ $\alpha = 0{,}01;$ $t_{\alpha/2,\, n_1+n_2-2} = 3{,}012.$

El intervalo de confianza pedido será: (-1,526 , 5,276)

****

6.12 *Se dispone de dos medicamentos con propiedades reductoras del contenido de colesterol en la sangre. Aplicamos un medicamento a un grupo de 10 pacientes y obtenemos las siguientes reducciones expresadas en mg/100ml:*

8, 11, 1, 8, 3, 2, 7, 4, 4, 12

Aplicamos el 2º medicamento a un grupo de 12 pacientes y obtenemos las siguientes reducciones, también expresadas en mg/100ml:

4, 10, 6, 6, 7, 4, 2, 10, 1, 2 8, 6.

Encontrar un intervalo de confianza del 99% para la direncia de medias, admitiendo que las poblaciones de reducción de colesterol en la sangre aplicando ambos medicamentos son normales, independientes y de varianzas iguales y desconocidas.

S O L U C I O N:

El intervalo de confianza para la diferencia de medias, siendo las desviaciones típicas desconocidas pero iguales y en muestras pequeñas, viene dado por la expresión:

$$\left((\bar{x}_1 - \bar{x}_2) \mp t_{\frac{\alpha}{2},\, n_1+n_2-2}\; s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

siendo
$$s^2 = \frac{(n_1-1)\,\hat{s}_1^2 + (n_2-1)\,\hat{s}_2^2}{n_1 + n_2 - 2}$$

Para la 1ª muestra tenemos: $\bar{x}_1 = 6$; $\hat{s}_1 = 3,77$; $n_1 = 10$.

Para la 2ª muestra tenemos: $\bar{x}_2 = 5,5$; $\hat{s}_2 = 3$; $n_2 = 12$.

$s^2 = 11,3458$; $\alpha = 0,01$; $t_{\alpha/2,\, n_1+n_2-2} = 2,845$.

Sustituyendo estos valores en la expresion del intervalo, tenemos:

(-3,60 , 4,60)

****

6.13 *Con el fin de comparar dos drogas reductoras de la concentración de ácido úrico en la sangre, se aplica una de ellas a un grupo de 12 pacientes obteniéndose una reducción media de contenido de ácido úrico de 18 mg/100 ml con una desviación típica de 3 mg/100 ml, y aplicada la otra droga a un grupo de 10 pacientes se obtuvo una reducción media de 16 mg/100 ml, con una desviación típica de 4 mg/100 ml.*

Encontrar un intervalo de confianza del 95% para la diferencia de medias, admitiendo que las poblaciones de reducción de ácido úrico en la sangre aplicando ambas drogas son normales, independientes, varianzas desconocidas y en principio no iguales.

SOLUCION:

El intervalo de confianza viene dado por la expresión:

$$\left((\bar{x}_1-\bar{x}_2) \mp t_{\frac{\alpha}{2},f} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right)$$

siendo f el número entero más próximo al valor que nos dé la expresión:

$$f = \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{s}_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{\hat{s}_2^2}{n_2}\right)}{n_2+1}} - 2 = \frac{\left(\frac{9,82}{12} + \frac{17,78}{10}\right)^2}{\frac{\left(\frac{9,82}{12}\right)^2}{13} + \frac{\left(\frac{17,78}{10}\right)^2}{11}} - 2 = 17,89$$

luego $f = 18$.

Pues en la 1ª muestra tenemos: $\bar{x}_1 = 18$; $\hat{s}_1^2 = \frac{12}{11} 9 = 9,82$; $n_1 = 12$

y en la 2ª muestra tenemos: $\bar{x}_2 = 16$; $\hat{s}_2^2 = 10.16/9 = 17,78$; $n_2 = 10$.

$\alpha = 0,05$; $t_{\alpha/2,\ f} = 2,101$.

Sustituyendo estos valores en la expresión del intervalo, tendremos:

$$(-1,38 \quad , \quad 5,38)$$

****

6.14 *Se está haciendo un estudio sobre hipertensión. Se toma una muestra de 13 pacientes de una ciudad, que tienen una presión sistólica media $\bar{x}_1 = 167$ mm y $\hat{s}_1 = 28$ mm. En otra ciudad se toma otra muestra de 16 pacientes con media $\bar{x}_2 = 164,7$ mm y $\hat{s}_2 = 7$ mm.*

Calcular un intervalo de confianza del 95% para la diferencia de medias, suponiendo normales las distribuciones pero con varianzas distintas.

SOLUCION:

El intervalo de confianza para la diferencia de medias de dos poblaciones normales, muestras pequeñas y desviaciones típicas desconocidas y distintas, viene dado por:

$$\left((\bar{x}_1-\bar{x}_2) \mp t_{\frac{\alpha}{2},\ f} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right)$$

donde f es la aproximación de Welch, que para este caso es: f = 13.

Pues en la 1ª muestra tenemos: $\bar{x}_1 = 167$; $\hat{s}_1 = 28$; $n_1 = 13$

y en la 2ª muestra tenemos: $\bar{x}_2 = 164{,}7$; $\hat{s}_2 = 7$; $n_2 = 16$.

$\alpha = 0{,}05$. y $t_{\alpha/2,\ f} = 2{,}160$

Sustituyendo estos valores, obtenemos el intervalo pedido:

$$(-14{,}89 \quad , \quad 19{,}49)$$

****

6.15 *Una central de productos lácteos recibe diariamente la leche de dos granjas X e Y. Deseando estudiar la calidad de los productos recibidos se extraen dos muestras y se analiza el contenido de materia grasa, obteniendo los siguientes resultados:*

X	0,32; 0,29; 0,30; 0,28; 0,33; 0,31; 0,30; 0,29; 0,33; 0,32; 0,30; 0,29.
Y	0,28; 0,30; 0,32; 0,29; 0,31; 0,29; 0,33; 0,32; 0,29; 0,32; 0,31; 0,29; 0,32; 0,31; 0,32; 0,33.

a) Calcular las medias y las varianzas de las muestras.

b) Calcular los valores de las estimaciones centradas $\hat{s}^2_x$ y $\hat{s}^2_y$ de las varianzas desconocidas σ^2_x y σ^2_y.

c) Al nivel de confianza del 95%, determinar un intervalo de confianza para la desviación típica de las x, para la desviación típica de las y y para la razon de varianzas, suponiendo normal la variable estudiada, (contenido de materia grasa). en ambas poblaciones y con igual varianza e independientes.

SOLUCION:

a) Para la muestra X tenemos: $\bar{x} = 0{,}305$ y $s^2_x = 0{,}0031$.

Para la muestra Y tenemos: $\bar{y} = 0{,}308$ y $s^2_y = 0{,}0038$.

b) $$\left.\begin{array}{l} \hat{s}^2_x = \dfrac{12}{11}.0{,}0031 = 0{,}00338. \\ \hat{s}^2_y = \dfrac{16}{15}\,0{,}0038 = 0{,}00405. \end{array}\right\}$$

c) El intervalo de confianza para σ_x viene dado por:

$$\left(\sqrt{\frac{(n_x-1)\hat{s}^2_x}{\chi^2_{\frac{\alpha}{2},\ n_x-1}}}\ ,\ \sqrt{\frac{(n_x-1)\hat{s}^2_x}{\chi^2_{1-\frac{\alpha}{2},\ n_x-1}}}\right)$$

donde $n_x = 12$; $\alpha = 0{,}05$; $\chi^2_{\alpha/2,\ n_x-1} = 21{,}9$; $\chi^2_{1-\alpha/2,\ n_x-1} = 3{,}82$.

Sustituyendo estos valores en la expresión anterior, resulta:

$$(0{,}039 \quad , \quad 0{,}094)$$

Análogamente, el intervalo de confianza de σ_y, será:

$n_y = 16; \quad \alpha = 0{,}05; \quad \chi^2_{\alpha/2,\ n_y-1} = 27{,}5; \quad \chi^2_{1-\frac{\alpha}{2},n_y-1} = 6{,}26.$

$$(0{,}045 \quad , \quad 0{,}095)$$

El intervalo de confianza para σ_x^2/σ_y^2, viene dado por:

$$\left(\frac{\hat{s}_x^2\ /\hat{s}_y^2}{F_{\frac{\alpha}{2},\ n_x-1,\ n_y-1}} \quad , \quad \frac{\hat{s}_x^2\ /\hat{s}_y^2}{F_{1-\frac{\alpha}{2},\ n_x-1,\ n_y-1}} \right)$$

donde $F_{\frac{\alpha}{2},\ n_x-1,\ n_y-1} = 3{,}04; \quad F_{1-\frac{\alpha}{2},\ n_x-1,\ n_y-1} = 0{,}3.$

Por tanto, el intervalo pedido será:

$$(0{,}27 \quad , \quad 2{,}71)$$

****

6.16 *En un cruce de melanogaster hemos obtenido 77 moscas con alas verticales, de un total de 220.*

Estimar un intervalo de confianza del 95% para la proporción de moscas con alas verticales entre los individuos resultantes de un gran número de cruces como este.

SOLUCION:

Por ser el tamaño de la muestra, 220 bastante grande y el valor de p ni próximo a cero ni a uno, podemos tomar como intervalo de confianza para este parámetro, el siguiente:

$$\left(\hat{p} \mp z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

siendo $\hat{p}$, la proporción de individuos con alas verticales en la muestra que tomamos como valor aproximado de p.

$\hat{p} = \frac{77}{220} = 0{,}35; \quad \alpha = 0{,}05; \quad z_{\alpha/2} = 1{,}96.$

Luego el intervalo será: $(0{,}287 \quad , \quad 0{,}413)$.

6.17 *Una nueva droga ha curado 80 de 200 enfermos. Estímese un intervalo de confianza del 99% para la proporción de personas curadas si la nueva medicina se hubiese aplicado a una población constituida por todos los individuos con la misma enfermedad.*

SOLUCION:

Tenemos que $\hat{p} = \frac{80}{200} = 0,4; \qquad n = 200.$

En estas condiciones el intervalo de confianza para la proporción de personas curadas en la población viene dado por:

$$\left(\hat{p} \mp z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

siendo $\alpha = 0,01; \quad$ y $\quad z_{\alpha/2} = 2,57.$

Sustituyendo estos valores en la expresión anterior, obtenemos el intervalo pedido:

$$(0,311 \quad , \quad 0,489)$$

****

6.18 *Se sospecha que existe una diferencia significativa entre la proporción de hombres y la proporción de mujeres que contraen una determinada variedad de gripe. Para salir de dudas se toma una muestra aleatoria de 300 hombres de los cuales, 27 padecen o han padecido la gripe en un periodo de tiempo fijado. Análogamente tomamos una muestra de 400 mujeres de las cuales 32 padecen o padecieron la gripe. ¿Qué nos indican estos resultados?.*

SOLUCION:

El intervalo de confianza para la diferencia de proporciones p_1 y p_2 de hombres y mujeres que han contraido la enfermedad, al ser el tamaño de las muestras n_1 y n_2 grandes, viene dado por:

$$\left((\hat{p}_1-\hat{p}_2) \mp z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

siendo $\hat{p}_1$ y $\hat{p}_2$ los valores estimados de p_1 y p_2 respectivamente.

$$\hat{p}_1 = \frac{27}{300} = 0,09; \quad n_1 = 300; \quad \hat{p}_2 = \frac{32}{400} = 0,08; \quad n_2 = 400;$$

$$\alpha = 0,05; \quad z_{\alpha/2} = 1,96.$$

Sustituyendo estos valores en la expresión anterior, obtenemos el in tervalo pedido, que será: (-0,0319 , 0,0519)

La diferencia de proporciones puede ser negativa, nula o positiva, luego no existe diferencia significativa entre p_1 y p_2 y entonces nuestra sospecha no es cierta.

****

7- contraste de hipótesis

3.2.1 Muestras pequeñas
3.2.1.1 Contraste bilateral
3.2.1.2 Contraste unilateral
3.2.1.2.1 Hipótesis nula $H_o: \mu \leq \mu_0$
3.2.1.2.2 Hipótesis nula $H_o: \mu \geq \mu_0$
3.2.2 Muestras grandes
3.2.2.1 Contraste bilateral
3.2.2.2 Contraste unilateral
3.2.2.2.1 Hipótesis nula $H_o \quad \mu \leq \mu_0$
3.2.2.2.2 Hipótesis nula $H_o \quad \mu \geq \mu_0$

3.3 Contraste para la varianza de una población normal.
3.3.1 Contraste bilateral
3.3.2 Contraste unilateral
3.3.2.1 Hipótesis nula $H_o: \sigma^2 \leq \sigma_0^2$
3.3.2.2 Hipótesis nula $H_o: \sigma^2 \geq \sigma_0^2$

3.4 Contraste para la igualdad de medias de dos poblaciones normales
3.4.1 Conocidas las varianzas
3.4.1.1 Contraste bilateral
3.4.1.2 Contraste unilateral
3.4.1.2.1 Hipótesis nula $H_o: \mu_1 \leq \mu_2$
3.4.1.2.2 Hipótesis nula $H_o: \mu_1 \geq \mu_2$
3.4.2 Varianzas desconocidas
3.4.2.1 Muestras grandes
3.4.2.2 Muestras pequeñas y varianzas poblacionales iguales
3.4.2.2.1 Contraste bilateral
3.4.2.2.2 Contraste unilateral
3.4.2.2.2.1 Hipótesis nula $H_o: \mu_1 \leq \mu_2$
3.4.2.2.2.2 Hipótesis nula $H_o: \mu_1 \geq \mu_2$
3.4.2.3 Muestras pequeñas y varianzas poblacionales distintas
3.4.2.3.1 Contraste bilateral
3.4.2.3.2 Contraste unilateral
3.4.2.3.2.1 Hipótesis nula $H_o: \mu_1 \leq \mu_2$
3.4.2.3.2.2 Hipótesis nula $H_o: \mu_1 \geq \mu_2$

3.4.3 Datos apareados
    3.4.3.1 Muestras grandes
        3.4.3.1.1 Contraste bilateral
        3.4.3.1.2 Contraste unilateral
            3.4.3.1.2.1 Hipótesis nula H_o: $d \leqslant 0$
            3.4.3.1.2.2 Hipótesis nula H_o: $d \geqslant 0$
    3.4.3.2 Muestras pequeñas
        3.4.3.2.1 Contraste bilateral
        3.4.3.2.2 Contraste unilateral
            3.4.3.2.2.1 Hipótesis nula H_o: $d \leqslant 0$
            3.4.3.2.2.2 Hipótesis nula H_o: $d \geqslant 0$

3.5 Contraste de igualdad de varianzas de dos poblaciones normales
    3.5.1 Contraste bilateral
    3.5.2 Contraste unilateral

3.6 Contraste para el parámetro p de una distribución binomial
    3.6.1 Contraste bilateral
    3.6.2 Contraste unilateral

3.7 Contraste para la igualdad de parámetros de dos distribuciones binomiales
    3.7.1 Contraste bilateral
    3.7.2 Contraste unilateral

4. ANALOGIAS ENTRE CONTRASTES DE HIPOTESIS E INTERVALOS DE CONFIANZA

7 - CONTRASTE DE HIPOTESIS

1. INTRODUCCION

Cuando hacemos alguna afirmación sobre una población base, sobre su forma o sobre el valor númerico de uno o de más de sus parámetros, que se contrasta luego mediante una muestra aleatoria extraida de la población, estamos ante inferencias de la categoría del *contraste de hipótesis*.

El planteamiento general de un problema de contraste es el siguiente: se formula una *hipótesis* o conjetura acerca de la población y se trata de ver si como consecuencia de un conjunto de valores muestrales, debemos aceptar o rechazar la hipótesis formulada con unos *márgenes de error* previamente fijados.

Si los valores muestrales difieren mucho de los valores teóricos que cabría esperar bajo la hipótesis formulada, podriamos pensar en rechazar la hipótesis, pues diriamos que las diferencias son *significativas*. Estamos considerando una distribución teórica bajo la hipótesis formulada, una distribución de la muestra y por último una medida de la diferencia entre ambas mediante un *estadístico*. Para cada muestra el estadístico tomará un valor y tendrá una distribución en el muestreo. Podemos determinar un valor particular crítico de dicho estimador, tal que la probabilidad de que este estimador tome un valor mayor que el valor particular crítico, sea igual a un valor fijado que llamaremos *nivel de significación*.

A partir de este momento adoptamos la norma de aceptar la hipótesis formulada si la diferencia anterior es menor que el valor particular que hemos determinado y rechazado en el caso contrario.

2. DEFINICIONES

2.1 CONTRASTE DE HIPOTESIS.-Procedimiento estadístico mediante el que se investiga la verdad o falsedad de una hipótesis acerca de una población o poblaciones.

2.2 HIPOTESIS NULA H_o.- Es la hipótesis que se formula y que se quiere contrastar; es por tanto, la hipótesis que se acepta o se rechaza como consecuencia del contraste.

2.3 HIPOTESIS ALTERNATIVA H_a .-Cualquier otra hipótesis que difiera de la formulada y que nos situe frente a H_o, de forma que si se acepta H_o se rechaza H_a y si se acepta H_a se rechaza H_o.

2.4 ESTADISTICO DEL CONTRASTE.- Es una función de los valores muestrales. Es una variable aleatoria que seguirá una distribución de probabilidad dada. Toma un valor para cada muestra.

2.5 REGION CRITICA.- Al aplicar un contraste de hipótesis clasificamos los puntos del espacio muestral en dos regiones excluyentes y complementarias. La formada por los puntos tales que los valores del estadístico del contraste nos lleva a rechazar la hipotesis nula H_o, se llama *región crítica.*

2.6 REGION DE ACEPTACION.- Es la formada por el conjunto de puntos tales que los valores del estadístico del contraste nos lleva a aceptar la hipótesis nula H_o

2.7 ERROR DE TIPO I.- Es el que cometemos cuando rechazamos la hipotesis nula siendo verdadera.

2.8 ERROR DE TIPO II.-Es el que cometemos cuando aceptamos la hipotesis nula siendo falsa.

2.9 NIVEL DE SIGNIFICACION α .-Es la probabilidad que tenemos de cometer el error de tipo I. En general, lo representaremos por α y estará acotado entre 0 y 1.

Si fijamos un $\alpha = 0,05$, significa que rechazamos un 5% de las veces la hipótesis nula siendo cierta, parece lógico, fijar valores de α mas pequeños para limitar más este error, sin embargo al hacer esto aumentamos la probabilidad de cometer el error de tipo II.

Es usual, tomar como niveles de significación: $\alpha = 0,05$ $\alpha = 0,01$; $\alpha = 0,005$. en cuyo caso diremos que el resultado es *casi significativo, significativo, y muy significativo.* Sin embargo es necesario resaltar aquí, que más importante que decir si una hipótesis se acepta o se rechaza con un nivel de significación dado, es determinar un valor α a partir del cual la hipótesis comienza a rechazarse o a aceptarse.

2.10 POTENCIA DE UN CONTRASTE.- Probabilidad de rechazar la hipótesis nula cuando es falsa.

La idea de Neyman-Pearson es que fijado un nivel de significación α, elegiremos de entre los contrastes, aquel que haga mínimo el error de tipo II, o lo que es lo mismo, que la potencia del contraste sea máxima.

De ahora en adelante utilizaremos *contrastes de máxima potencia.*

2.11 CONTRASTE BILATERAL.- Diremos que estamos ante un *contraste bilateral*, cuando la región crítica está formada por dos conjuntos de puntos disjuntos. Corresponde al caso en el que hemos fijado como hipótesis nula que un parámetro poblacional tiene un determinado valor frente a la hipótesis alternativa de que el valor del parámetro es distinto.

2.12 CONTRASTE UNILATERAL.- Se dice así, cuando la región crítica está formada por un solo conjunto de puntos. Corresponde al caso en que hemos fijado como hipótesis nula que el parámetro poblacional toma valores, hasta o desde un valor extremo, frente a la alternativa de que el valor del parámetro es desde o hasta ese valor extremo.

3. FORMULAS PARA LOS CONTRASTES

3.1 CONTRASTE DE LA MEDIA DE UNA POBLACION NORMAL CUANDO SE CONOCE LA VARIANZA POBLACIONAL

3.1.1 Contraste bilateral.

Hipótesis nula H_o: $\mu = \mu_0$

Hipótesis alternativa H_a: $\mu \neq \mu_0$

Tomamos como estadístico la media muestral $\bar{x}$.

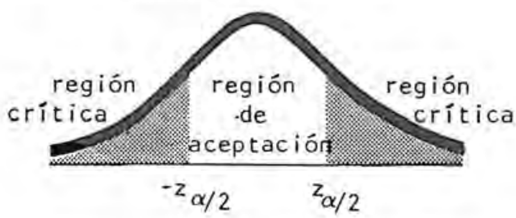
Este estadístico sigue una distribución $N(\mu\ ,\ \sigma/\sqrt{n})$ donde μ y σ son la media y la desviación típica poblacionales respectivamente y n es el tamaño de la muestra.

La distribución de la variable tipificada

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad \text{sigue una } N(0\ ,\ 1),$$

para $\mu = \mu_0$, o sea cuando la hipótesis es verdadera.

Si fijamos un nivel de significación α y la hipotesis es cierta, el valor de z, obtenido de una muestra real para el estadístico $\bar{x}$, se encontrará entre $-z_{\alpha/2}$ y $z_{\alpha/2}$, con una probabilidad de $1 - \alpha$.



El intervalo $(-z_{\alpha/2}\ ,\ z_{\alpha/2})$ es la región de aceptación.

Es decir aceptamos la hipótesis H_o si:

$$-z_{\alpha/2} \leq z \leq z_{\alpha/2}$$

Se acepta H_o si $\dfrac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$

Se rechaza H_o si $\dfrac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$

3.1.2 Contraste unilateral.

3.1.2.1 Hipótesis nula H_o: $\mu \leq \mu_0$

Hipótesis alternativa H_a: $\mu > \mu_0$

Según lo anteriormente expuesto, la región de aceptación es el intervalo $(-\infty \ , \ z_\alpha)$.

Aceptamos H_o si: $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha$.

Rechazamos H_o si: $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$.

3.1.2.2 Hipótesis nula H_o: $\mu \geq \mu_0$

Hipótesis alternativa H_a: $\mu < \mu_0$

En este caso la región de aceptación es el intervalo $(-z_\alpha \ , \ \infty)$.

Aceptamos H_o si : $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq - z_\alpha$.

Rechazamos H_o si : $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < - z_\alpha$.

3.2 CONTRASTE DE LA MEDIA DE UNA POBLACION NORMAL CUANDO NO SE CONOCE LA VARIANZA.-

3.2.1 Muestras pequeñas.

3.2.1.1 Contraste bilateral.

Hipótesis nula H_o: $\mu = \mu_0$

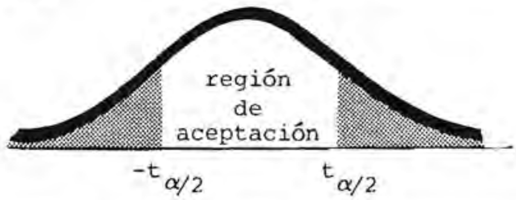
Hipótesis alternativa H_a: $\mu \neq \mu_0$

En este caso el estadístico

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \qquad \text{con} \quad \mu = \mu_0$$

se distribuye según una t de Student con n-1 grados de libertad, siendo $\bar{x}$ la media muestral, s la desviación típica muestral y n el tamaño de la muestra.

Si fijamos un nivel de significación α y la hipó tesis es cierta, el valor de t obtenido para una muestra real se encontrará entre $-t_{\alpha/2,\ n-1}$ y $t_{\alpha/2,\ n-1}$ con una probabilidad de $1-\alpha$.



Esta región de aceptación es el intervalo

$$\left(-t_{\frac{\alpha}{2},\ n-1}\ ,\ t_{\frac{\alpha}{2},\ n-1}\right).$$

Se acepta H_o si: $\dfrac{|\bar{x}-\mu_0|}{s/\sqrt{n-1}} \leq t_{\frac{\alpha}{2},\ n-1}$

Se rechaza H_o si: $\dfrac{|\bar{x}-\mu_0|}{s/\sqrt{n-1}} > t_{\frac{\alpha}{2},\ n-1}$

3.2.1.2 Contraste unilateral

3.2.1.2.1 Hipótesis nula $H_o: \mu \leq \mu_0$

Hipótesis alternativa $H_a: \mu > \mu_0$

Según lo expuesto anteriormente, la región de aceptación será el intervalo $(-\infty,\ t_{\alpha,\ n-1})$.

Se acepta H_o si: $\dfrac{\bar{x}-\mu_0}{s/\sqrt{n-1}} \leq t_{\alpha,\ n-1}$

Se rechaza H_o si: $\dfrac{\bar{x}-\mu_0}{s/\sqrt{n-1}} > t_{\alpha,\ n-1}$

3.2.1.2.2 Hipótesis nula $H_o: \mu \geq \mu_0$

Hipótesis alternativa $H_a: \mu < \mu_0$

La región de aceptación es el intervalo $(-t_{\alpha,\ n-1}\ ,\ \infty)$.

Se acepta H_o si: $\frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \geq -t_{\alpha,\ n-1}$

Se rechaza H_o si $\frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} < -t_{\alpha,\ n-1}$

3.2.2 Muestras grandes.

3.2.2.1 Contraste bilateral.

Hipótesis nula H_o: $\mu = \mu_0$

Hipótesis alternativa H_a: $\mu \neq \mu_0$

En este caso también utilizaremos $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ pero teniendo en cuenta que al no conocer σ y al ser la muestra grande, la estimamos mediante la desviación típica de la muestra con lo que nos queda:

$$z = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}}$$

siendo $\hat{s} = \sqrt{\frac{n}{n-1}}\, s$ y s la desviación típica de la muestra.

La región de aceptación es el intervalo:

$$(-z_{\alpha/2}\ ,\ z_{\alpha/2}).$$

Se acepta H_o si: $\frac{|\bar{x} - \mu_0|}{\hat{s}/\sqrt{n}} \leq z_{\alpha/2}$.

Se rechaza H_o si: $\frac{|\bar{x} - \mu_0|}{\hat{s}/\sqrt{n}} > z_{\alpha/2}$.

3.2.2.2 Contraste unilateral.

3.2.2.2.1 Hipótesis nula H_o: $\mu \leq \mu_0$

Hipótesis alternativa H_a: $\mu > \mu_0$

Según todo lo expuesto anteriormente, la región de aceptación es el intervalo $(-\infty\ ,\ z_\alpha)$.

Se acepta H_o si: $\frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} \leq z_\alpha$

Se rechaza H_o si: $\dfrac{\overline{x} - \mu_0}{\hat{s}/\sqrt{n}} > z_\alpha$

3.2.2.2.2 Hipótesis nula H_o: $\mu \geq \mu_0$

Hipótesis alternativa H_a: $\mu < \mu_0$

La región de aceptación viene definida por:

$$(-z_\alpha \ , \ \infty)$$

Se acepta H_o si: $\dfrac{\overline{x} - \mu_0}{\hat{s}/\sqrt{n}} \geq -z_\alpha$

Se rechaza H_o si: $\dfrac{\overline{x} - \mu_0}{\hat{s}/\sqrt{n}} < -z_\alpha$

3.3 CONTRASTE PARA LA VARIANZA DE UNA POBLACION NORMAL.-

3.3.1 Contraste bilateral.

Hipótesis nula H_o: $\sigma^2 = \sigma_0^2$

Hipótesis alternativa H_a: $\sigma^2 \neq \sigma_0^2$

En este caso el estadístico ns^2/σ_0^2, con $\sigma^2 = \sigma_0^2$ se distribuye según una χ^2con n-1 grados de libertad, siendo n el tamaño de la muestra y s^2 la varianza muestral.

La región de aceptación es el intervalo:

$$\left(\chi^2_{\frac{\alpha}{2},\, n-1} \ , \ \chi^2_{1-\frac{\alpha}{2},\, n-1}\right)$$

Se acepta H_o si: $\dfrac{ns^2}{\sigma_0^2} \in \left(\chi^2_{\frac{\alpha}{2},\, n-1} \ , \ \chi^2_{1-\frac{\alpha}{2},\, n-1}\right)$

3.3.2 Contraste unilateral.

3.3.2.1 Hipótesis nula H_o: $\sigma^2 \leq \sigma_0^2$

Hipótesis alternativa H_a: $\sigma^2 > \sigma_0^2$

La región de aceptación es el intervalo $(0 \ , \ \chi^2_{\alpha,\, n-1})$

Se acepta H_o si: $\dfrac{ns^2}{\sigma_0^2} < \chi^2_{\alpha,\, n-1}$

3.3.2.2 Hipótesis nula $H_o: \sigma^2 \geq \sigma_0^2$

Hipótesis alternativa $H_a: \sigma^2 < \sigma_0^2$

La región de aceptación es el intervalo

$$(\chi^2_{1-\alpha,\ n-1}\ ,\ \infty)$$

Se acepta H_o si: $\dfrac{n\ s^2}{\sigma_0^2} > \chi^2_{1-\alpha,\ n-1}$

3.4 CONTRASTE PARA LA IGUALDAD DE MEDIAS DE DOS POBLACIONES NORMALES

3.4.1 Conocidas las varianzas.

3.4.1.1 Contraste bilateral.

Hipótesis nula $H_o: \mu_1 = \mu_2$ ó $\mu_1 - \mu_2 = 0$

Hipótesis alternativa $H_a: \mu_1 \neq \mu_2$

De una población $N(\mu_1\ ,\ \sigma_1)$, obtenemos una muestra aleatoria simple:

$$x_{11},\ x_{12},\ .\ .\ .\ .\ ,\ x_{1n_1}$$

la media muestral $\overline{x}_1 = \dfrac{x_{11} + x_{12} + \ldots + x_{1n_1}}{n_1}$

se distribuye según una $N(\mu_1\ ,\ \sigma_1/n_1)$.

Si de una población $N(\mu_2\ ,\ \sigma_2)$ obtenemos otra muestra aleatoria simple:

$$x_{21},\ x_{22},\ .\ .\ .\ .\ ,x_{2n_2}$$

la media muestral $\overline{x}_2 = \dfrac{x_{21} + x_{22} + \ldots + x_{2n_2}}{n_2}$

se distribuye según una $N(\mu_2\ ,\ \sigma_2/n_2)$.

La diferencia $\overline{x}_1 - \overline{x}_2$ es una variable aleatoria, ya que $\overline{x}_1$ y $\overline{x}_2$ son variables aleatorias inde - pendientes. Además $\overline{x}_1 - \overline{x}_2$ se distribuye nor- malmente con media $\mu_1 - \mu_2$ y desviación típica

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

En este caso el estadístico

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

se distribuye según una $N(0\ ,\ 1)$ si $\mu_1 = \mu_2$.

Si fijamos un nivel de significación α y la hipótesis es cierta, z se encontrará entre $-z_{\alpha/2}$ y $z_{\alpha/2}$ con una probabilidad $1 - \alpha$.

Luego, la región de aceptación es:

$$(-z_{\alpha/2}\ ,\ z_{\alpha/2})\ .$$

Se acepta H_o si: $$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}$$

3.4.1.2 <u>Contraste unilateral</u>

- 3.4.1.2.1 Hipótesis nula $H_o:\ \mu_1 \leq \mu_2$

  Hipótesis alternativa $H_a:\ \mu_1 > \mu_2$

Según todo lo anterior, la región de aceptación es el intervalo $(-\infty\ ,\ z_\alpha)$.

Se acepta H_o si : $$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_\alpha .$$

- 3.4.1.2.2 Hipótesis nula $H_o:\ \mu_1 \geq \mu_2$

  Hipótesis alternativa $H_a:\ \mu_1 < \mu_2$

La región de aceptación en este caso, será:

$$(-z_\alpha\ ,\ \infty)\ .$$

Se acepta H_o si : $$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq -z_\alpha .$$

3.4.2 Varianzas desconocidas.

3.4.2.1 Muestras grandes.

Si las muestras son grandes, el contraste lo efectuaremos como indicabamos en el apartado 3.4.1, sustituyendo σ_1^2 y σ_2^2 por $\hat{s}_1^2$ y $\hat{s}_2^2$, que son los estimadores insesgados de σ_1^2 y σ_2^2 respectivamente.

$$\hat{s}_1^2 = \frac{n_1}{n_1 - 1} s_1^2 \qquad y \qquad \hat{s}_2^2 = \frac{n_2}{n_2 - 1} s_2^2$$

siendo s_1^2 y s_2^2 las varianzas muestrales.

3.4.2.2 Muestras pequeñas y varianzas poblacionales iguales.

3.4.2.2.1 Contraste bilateral.

Hipótesis nula $H_o: \mu_1 = \mu_2$

Hipótesis de alternativa $H_a: \mu_1 \neq \mu_2$

Como en este caso $\sigma_1 = \sigma_2 = \sigma$, el estadístico

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad con \qquad \mu_1 = \mu_2$$

sigue una distribución t de Student con n_1+n_2-2 grados de libertad, siendo

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

Utilizamos como estimador de σ, la media ponderada de $\hat{s}_1^2$ y $\hat{s}_2^2$, es decir:

$$s^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{(n_1-1) + (n_2-1)} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

La región de aceptación es:

$$(-t_{\frac{\alpha}{2}, n_1+n_2-2} \quad , \quad t_{\frac{\alpha}{2}, n_1+n_2-2})$$

Se acepta H_o si: $$\frac{|\bar{x}_1 - \bar{x}_2|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\frac{\alpha}{2},\, n_1+n_2-2}$$

y se rechaza en el caso contrario.

3.4.2.2.2 Contraste unilateral

- 3.4.2.2.2.1 Hipótesis nula H_o: $\mu_1 \leq \mu_2$
  Hipótesis alternativa
  H_a: $\mu_1 > \mu_2$

  La región de aceptación será:

  $$(-\infty \quad , \quad t_{\alpha,\, n_1+n_2-2})$$

  Se acepta H_o si se cumple:

  $$\frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{\alpha,\, n_1+n_2-2}$$

- 3.4.2.2.2.2 H_o: $\mu_1 \geq \mu_2$
  H_a: $\mu_1 < \mu_2$

  La región de aceptación será:

  $$(-t_{\alpha,\, n_1+n_2-2} \quad , \quad \infty)$$

  Se acepta H_o si se cumple:

  $$\frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > -t_{\alpha,\, n_1+n_2-2}$$

3.4.2.3 Muestras pequeñas y varianzas poblacionales distintas.

3.4.2.3.1 Contraste bilateral.

Hipótesis nula H_o: $\mu_1 = \mu_2$
Hipótesis alternativa H_a: $\mu_1 \neq \mu_2$

En este caso el estadístico:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$ si la hipótesis

es cierta sigue aproximadamente una distribución t de Student con f grados de libertad, siendo f la aproximación de Welch vista en el capítulo anterior.

La región de aceptación es el intervalo:

$$(-t_{\frac{\alpha}{2}, f} \quad , \quad t_{\frac{\alpha}{2}, f})$$

Por tanto, se acepta la hipótesis H_o si:

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \leq t_{\frac{\alpha}{2}, f}$$

3.4.2.3.2 Contraste unilateral.

3.4.2.3.2.1 $H_o: \quad \mu_1 \leq \mu_2$

$H_a: \quad \mu_1 > \mu_2$

La región de aceptación es el intervalo:

$$(-\infty \quad , \quad t_{\alpha, f})$$

Por tanto, se acepta la hipótesis H_o si:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} < t_{\alpha, f}$$

3.4.2.3.2.2 $H_o: \quad \mu_1 \geq \mu_2$

$H_a: \quad \mu_1 < \mu_2$

La región de aceptación es el intervalo:

$$(- t_{\alpha, f} \quad , \quad \infty)$$

Por tanto, se acepta la hipótesis H_o si:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} > - t_{\alpha,\, f}$$

3.4.3 Datos apareados.

A veces en el contraste de igualdad de medias de dos poblaciones normales, interesa formar pares de valores con elementos x correspondientes a la muestra de una población y los elementos muestrales y correspondientes a la otra población, considerando seguidamente la nueva variable: $d_i = x_i - y_i \qquad i = 1,2,\ldots,n$

3.4.3.1 Muestras grandes.

3.4.3.1.1 Contraste bilateral.

Hipótesis nula H_o: $d = 0$

Hipótesis alternativa H_a: $d \neq 0$

El estadístico:

$$z = \frac{\bar{d}}{\hat{s}_d/\sqrt{n}}$$

si la hipótesis es cierta, se distribuye según una N(0 , 1), siendo:

$$\bar{d} = \frac{1}{n}\sum d_i ; \qquad \hat{s}_d^2 = \frac{1}{n-1}\sum (d_i - \bar{d})^2$$

Si fijamos un nivel de significación y la hipótesis es cierta, z se encontrará entre $-z_{\alpha/2}$ y $z_{\alpha/2}$, con una probabilidad de $1-\alpha$.

La región de aceptación será:

$$(-z_{\alpha/2} \; , \; z_{\alpha/2})$$

Se acepta H_o si: $\frac{|\bar{d}|}{\hat{s}_d/\sqrt{n}} \leq z_{\alpha/2}$

3.4.3.1.2 Contraste unilateral.

3.4.3.1.2.1 H_o: $d \leq 0$

H_a: $d > 0$

Región de aceptación $(-\infty \quad , \quad z_\alpha)$

Se acepta la hipótesis H_o si:

$$\frac{\overline{d}}{\hat{s}_d/\sqrt{n}} \leq z_\alpha$$

- 3.4.3.1.2.2 H_o: $d \geq 0$

H_a: $d < 0$

Región de aceptación $(-z_\alpha \quad , \quad \infty)$

Se acepta la hipótesis H_o si:

$$\frac{\overline{d}}{\hat{s}_d/\sqrt{n}} \geq -z_\alpha$$

3.4.3.2 Muestras pequeñas.

3.4.3.2.1 Contraste bilateral.

Hipótesis nula H_o: $d = 0$

Hipótesis alternativa H_a: $d \neq 0$

El estadístico $t = \frac{\overline{d}}{\hat{s}_d/\sqrt{n}}$, si la hipótesis es cierta, se distribuye según una t de Student con n-1 grados de libertad.

La región de aceptación es:

$$\left(-t_{\frac{\alpha}{2}, n-1} \quad , \quad t_{\frac{\alpha}{2}, n-1}\right)$$

Se acepta la hipótesis H_o si:

$$\frac{|\overline{d}|}{\hat{s}_d/\sqrt{n}} \leq t_{\frac{\alpha}{2}, n-1}$$

3.4.3.2.2 Contraste unilateral

3.4.3.2.2.1 H_o: $d \leq 0$

H_a: $d > 0$

Región de aceptación $(-\infty \ , \ t_{\alpha, \ n-1})$.
Luego se acepta la hipótesis H_o, si:

$$\frac{\bar{d}}{\hat{s}_d/\sqrt{n}} \leq t_{\alpha, \ n-1}$$

3.4.3.2.2.2 H_o: $d \geq 0$

H_a: $d < 0$

Región de aceptación $(-\ t_{\alpha, \ n-1} \ , \ \infty)$.
Luego se acepta la hipótesis H_o, si

$$\frac{\bar{d}}{\hat{s}_d/\sqrt{n}} \geq -t_{\alpha, \ n-1}$$

3.5 CONTRASTE DE IGUALDAD DE VARIANZAS DE DOS POBLACIONES NORMALES.-

3.5.1 Contraste bilateral.

Hipótesis nula H_o: $\sigma_1^2 = \sigma_2^2$

Hipótesis alternativa H_a: $\sigma_1^2 \neq \sigma_2^2$

Si tomamos una muestra $x_{11}, x_{12}, \ldots, x_{1n_1}$ de una población $N(\mu_1 \ , \ \sigma_1)$ y otra $x_{21}, x_{22}, \ldots, x_{2n_2}$ de otra población $N(\mu_2 \ , \ \sigma_2)$, las cuasivarianzas muestrales serán:

$$\left.\begin{aligned} \hat{s}_1^2 &= \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 \\ \hat{s}_2^2 &= \frac{1}{n_2-1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 \end{aligned}\right\}$$

y la razón de las cuasivarianzas muestrales será:

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2} = \frac{(n_2-1) \sum (x_{1i} - \bar{x}_1)^2}{(n_1-1) \sum (x_{2j} - \bar{x}_2)^2}$$

F sigue una distribucion de Snedecor con (n_1-1) , (n_2-1) grados de libertad.

La región de aceptación será:

$$(F_{1-\frac{\alpha}{2}, n_1-1, n_2-1} \quad , \quad F_{\frac{\alpha}{2}, n_1-1, n_2-1}),$$

luego se acepta la hipótesis H_o, si la razón $\hat{s}_1^2/\hat{s}_2^2$ pertenece al intervalo anterior, en el caso contrario se rechazará la hipótesis H_o.

3.5.2 Contraste unilateral.

Hipótesis nula H_o: $\sigma_1^2 \leq \sigma_2^2$

Hipótesis alternativa H_a: $\sigma_1^2 > \sigma_2^2$

La región de aceptación es $(0 \quad , \quad F_{\alpha, n_1-1, n_2-1})$,

luego se acepta la hipótesis H_o, si la razón:

$$\frac{\hat{s}_1^2}{\hat{s}_2^2} \leq F_{\alpha, n_1-1, n_2-1}$$

3.6 CONTRASTE PARA EL PARAMETRO p DE UNA DISTRIBUCION BINOMIAL.-

3.6.1 Contraste bilateral.

Hipótesis nula H_o: $p = p_o$

Hipótesis alternativa H_a: $p \neq p_o$

Si tomamos una muestra de tamaño n, suficientemente grande y representamos por $\hat{p}$ la proporción de éxitos de dicha muestra, $\hat{p}$ sigue una distribución normal de media p y desviación típica $\sqrt{\frac{p(1-p)}{n}}$, siendo p la proporción de éxitos en la población.

El estadístico $z = \dfrac{\hat{p} - p_o}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}$, si la hipótesis es cierta, se distribuye según una $N(0 , 1)$.

La región de aceptación es $(-z_{\alpha/2} \quad , \quad z_{\alpha/2})$,

por tanto se acepta la hipótesis H_o, si:

$$\frac{|\hat{p} - p_o|}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\alpha/2}$$

3.6.2 Contraste unilateral.

Hipótesis nula H_o: $p \leq p_o$

Hipótesis alternativa H_a: $p > p_o$

Teniendo en cuenta lo dicho para el contraste bilateral, tendremos como región de aceptación $(-\infty\ ,\ z_\alpha)$, por tanto se acepta la hipótesis H_o, si:

$$\frac{\hat{p} - p_o}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \leq z_\alpha$$

3.7 CONTRASTE PARA LA IGUALDAD DE PARAMETROS DE DOS DISTRIBUCIONES BINOMIALES.-

3.7.1 Contraste bilateral.

Hipótesis nula H_o: $p_1 = p_2$

Hipótesis alternativa H_a: $p_1 \neq p_2$

Sean $\hat{p}_1$ y $\hat{p}_2$ las proporciones de éxitos en dos grandes muestras de tamaños n_1 y n_2 respectivamente, tomadas de dos poblaciones que tienen como proporciones de éxitos p_1 y p_2. Si la hipótesis $p_1 = p_2$ es cierta, el estadístico

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

sigue una distribución $N(0\ ,\ 1)$.

La región de aceptación es : $(-z_{\alpha/2}\ ,\ z_{\alpha/2})$,

luego se acepta la hipótesis H_o, si:

$$|z| \leq z_{\alpha/2}$$

3.7.2 Contraste unilateral

Hipótesis nula H_o: $p_1 \leq p_2$

Hipótesis alternativa H_a: $p_1 > p_2$

La región de aceptación es $(-\infty\ ,\ z_\alpha)$, por tanto se acepta la hipótesis H_o, si:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_\alpha$$

4. ANALOGIAS ENTRE CONTRASTES DE HIPOTESIS E INTERVALOS DE CONFIANZA

Existe una gran relación entre el intervalo de confianza para un parámetro de una distribución y un contraste de hipótesis relativo al mismo, así, si formulamos la hipótesis de que la media μ de una distribución toma un determinado valor μ_o, obtenido un intervalo de confianza para una muestra particular cuando dicho intervalo no cubre el valor μ_o, equivale a rechazar la hipótesis de que $\mu = \mu_o$.

El error de tipo I, se corresponde con la probabilidad de que el intervalo de confianza no cubra el valor del parámetro.

El error de tipo II, se corresponde con la probabilidad de que el intervalo de confianza cubra valores erroneos.

En los resultados de los contrastes realizados no hemos indicado la curva característica de operación, gráficos que nos muestran las probabilidades de errores de tipo II bajo diferentes hipótesis; luego según esto, conseguimos mejor información con un intervalo de confianza que con un contraste de hipótesis.

Si hubieramos tenido en cuenta el error de tipo II, habriamos determinado un tamaño necesario de la muestra para que la potencia del contraste fuera mayor que un cierto valor, en correspondencia con el tamaño de la muestra necesaria para limitar la longitud de los intervalos de confianza.

****

PROBLEMAS RESUELTOS

7.1 *Al recepcionista de una fábrica azucarera le llegan continuamente cargamentos de remolacha para los que quiere contrastar el contenido o riqueza en azucar. El sabe que la riqueza en azucar se distribuye normalmente con media $\mu_0 = 18\%$ para el cargamento de regadio y con media superior para el de secano, siendo en ambos casos la desviación típica igual a 6%.*

Toma una muestra de 20 cargamentos y desea calcular el valor crítico para un nivel de significación $\alpha = 0,05$ y la regla de decisión del contraste.

SOLUCION:

Hipótesis nula $H_0: \mu = \mu_0$

Hipótesis alternativa $H_a: \mu > \mu_0$

El estadístico del contraste es: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \in N(0\,,1)$, para $\mu = \mu_0$, o sea cuando la hipotesis nula H_0 es cierta.

Aceptaremos H_0, si z es menor que z_α.

$\alpha = 0{,}05$; $z_\alpha = 1{,}64$. Para aceptar H_0 tiene que ocurrir que:

$$\frac{\bar{x} - 18}{6/\sqrt{20}} < 1{,}64 \implies \bar{x} < 20{,}20\%$$

Luego 20,20% es el punto crítico.

La regla de decisión del contraste es:

$\bar{x} < 20{,}20\%$ cargamento de regadio.

$\bar{x} \geqslant 20{,}20\%$ cargamento de secano.

****

7.2 *Tomamos 10 tubos y determinamos el contenido en grs. de ungüento de cada uno de ellos, obteniendo los siguientes resultados:*

tubo	1	2	3	4	5	6	7	8	9	10
contenido en grs	5,2	4,9	5	5,1	5,2	4,8	4,9	5,3	4,6	5,4

Sabemos por otras muchas determinaciones que la desviación típica de la población es de 0,10 gramos y queremos averiguar si los valores anteriores son compatibles con la media $\mu = 5$ grs. para la población, supuesta esta normal.

SOLUCION

Hipótesis nula H_o: $\mu = \mu_o$

Hipótesis alternativa H_a: $\mu \neq \mu_o$

En este caso σ es conocida, por tanto se acepta la hipótesis H_o si se verifica

$$\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

$\bar{x} = 5,04$; $\mu_o = 5$; $\sigma = 0,10$; $n = 10$.

Tomamos $\alpha = 0,05$ de donde: $z_{\alpha/2} = 1,96$.

Sustituyendo estos valores en la expresión anterior, resulta:

$$1,26 < 1,96$$

Al ser esto cierto, podemos concluir que la muestra es realmente compatible con la poblacion en el 95% de los casos.

****

7.3 *En un anuncio publicitario se indica que un determinado tipo de agua reduce peso. Doce individuos que decidieron tomar dicha marca de agua en sustitución de la que habitualmente tomaban, manteniendo intacta el resto de la dieta alimenticia sufrieron las siguientes variaciones de peso al cabo de cierto tiempo:*

Individuo	1	2	3	4	5	6	7	8	9	10	11	12
Variación de peso.	0,2	0	1	0,6	-0,5	-0,6	-1	0,6	1	0,5	-0,4	-0,5

Teniendo en cuenta estos datos se puede admitir que el anuncio es correcto.

SOLUCION:

Hacemos la hipótesis de que la media de peso perdido es cero para la población, que sería lo que sucedería si el agua citada no sirviera para reducir el peso.

Hipótesis nula H_o: $\mu = \mu_o$

Hipótesis alternativa H_a: $\mu \neq \mu_o$

Aceptaremos la hipótesis H_o si se verifica:

$$\frac{|\bar{x} - \mu_0|}{s/\sqrt{n-1}} \leq t_{\frac{\alpha}{2}, n-1}$$

$\bar{x} = 0{,}075$; $s = 0{,}643$; $n = 12$; $\mu_o = 0.$

Si tomamos $\alpha = 0{,}05$ resulta que: $t_{\frac{\alpha}{2}, n-1} = 2{,}201.$

Sustituyendo estos valores en la expresión anterior, tenemos que:

$$0{,}40 < 2{,}201$$

Por tanto, la hipótesis de que el agua reduce peso es falsa, ya que se admite la hipótesis H_o de que la media de peso reducido es nula.

****

7.4 *En un preparado alimenticio infantil se especifica que según análisis garantizado el contenido mínimo de proteinas es del 42%. Tratamos de comprobar esta especificación y para ello tomamos 10 preparados que analizamos para determinar su contenido en proteinas, obteniendo de media 40% y desviación típica 3,5%.*

Es correcta la especificación citada para un nivel de significación $\alpha = 0{,}05$, suponiendo normal la distribución de la variable, contenido proteínico.

SOLUCION:

Se trata de un contraste unilateral de la media de una población normal de varianza desconocida.

Hipótesis nula H_o: $\mu \geq \mu_o$

Hipótesis alternativa H_a: $\mu < \mu_o$

Se acepta H_o si se verifica:

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \geq -t_{\alpha, n-1}$$

donde si fijamos $\alpha = 0{,}05$, resulta: $-t_{\alpha, n-1} = -2{,}262$

Sustituyendo estos valores obtenemos que:

$$-1{,}714 > -2{,}262$$

Luego aceptaremos la hipótesis H_o, por lo que admitiremos como correcta la especificación del preparado acerca del contenido proteínico.

7.5 *El nivel medio de protrombina en una población normal es conocido y resulta ser aproximadamente de 20 mg/100ml de plasma. Se toma una muestra de 40 pacientes que se sabe que tienen deficiencia de vitamina K.*

Los resultados son: $\bar{x}$ = 18,5 mg/100ml y s = 4 mg/100ml

¿Es la muestra comparable con la población, con un nivel de significacion del 0,05?.

SOLUCION:

Queremos contrastar la hipótesis de que la media μ de la población a la que pertenece la muestra es igual a μ_0, frente a la hipótesis alternativa de que $\mu \neq \mu_0$.

Por ser el tamaño de la muestra grande y σ desconocido, el estadístico del contraste será: $\dfrac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}}$, aceptando la hipótesis nula H_0 cuando el valor absoluto del citado estadístico sea menor o igual que $z_{\alpha/2}$.

Como $\mu_0 = 20$; $\hat{s} = 4$; $\alpha = 0,05$; $z_{\alpha/2} = 1,96$.

Sustituyendo tendremos que:

$$2,37 > 1,96$$

por tanto se rechaza la hipótesis nula H_0, luego la muestra no es comparable con la población.

****

7.6 *En el equipo de análisis que acompaña a los acuarios para la determinación de la dureza del agua de los mismos en tanto por ciento, se indica que la varianza de las determinaciones es igual o menor que el 5%.*

Llevamos a cabo 20 determinaciones de la dureza del agua del acuario y obtenemos una varianza para las mismas igual a 6%.

Si la variable, determinación de la dureza del agua, es normal, ¿aceptaremos la indicación con un nivel de significación de $\alpha = 0,01$?.

SOLUCION:

Hipótesis nula H_0: $\sigma^2 \leq \sigma_0^2$

Hipótesis alternativa H_a: $\sigma^2 > \sigma_0^2$

Aceptaremos H_0 si se verifica:

$$\frac{n\,s^2}{\sigma_0^2} < \chi^2_{\alpha,n-1}$$

Para $\alpha = 0{,}01$ se tiene que: $\chi^2_{\alpha\,,\,n-1} = 36{,}19.$

$\frac{ns^2}{\sigma_0^2} = 28{,}8.$ Sustituyendo resulta: $28{,}8 < 36{,}19.$

Aceptaremos la indicación que especifica el equipo de análisis.

****

7.7 *A los 100 alumnos de una clase se les separa en dos grupos, aquellos que practican habitualmente un deporte y los que no práctican ninguno, formando cada grupo 60 y 40 alumnos respectivamente. Les medimos la altura y obtenemos para el primer grupo: media 1,80 mtrs. y desviación típica 0,08 mtrs. y para el segundo grupo: media 1,76 mtrs y desviación típica 0,10 mtrs.*

¿Se puede afirmar con un nivel de confianza del 95% que los alumnos que practican algún deporte son mas altos?

Supóngase que la altura es una variable aleatoria que se distribuye en la población anterior normalmente.

S O L U C I O N:

Se trata de un contraste unilateral para la igualdad de medias con varianzas desconocidas y muestras grandes.

Hipótesis nula H_o: $\mu_1 \geq \mu_2$

Hipótesis alternativa H_a: $\mu_1 < \mu_2$

Aceptaremos la hipótesis H_o si:
$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \geq -z_\alpha$$

Sustituyendo los valores dados en el enunciado, obtenemos el valor de este estadístico que resulta ser: 2,118.

Para $\alpha = 0{,}05$ se tiene que: $-z_\alpha = -1{,}64.$

Luego, por ser: $2{,}118 > -1{,}64$, se acepta la hipótesis H_o, de donde puede admitirse que los alumnos que practican habitualmente algún deporte son más altos.

****

7.8 *El nivel medio de nistanina en un ungüento se supone cono-*

cido a traves de una larga serie de análisis e igual a 2,5 millones de unidades por gramo. Su distribución sigue la ley normal.

Se toma una muestra de 16 ungüentos, que nos dan un contenido medio de nistanina de 2,8 millones de unidades por gr. con una desviación típica de 0,4 millones de unidades por gramo.

¿Son estos resultados compatibles con la media supuesta para la población?.

SOLUCION:

Hipótesis nula H_o: $\mu = \mu_o$

Hipótesis alternativa H_a: $\mu \neq \mu_o$

Aceptaremos la hipótesis H_o si se verifica:

$$\frac{|\bar{x} - \mu_0|}{s/\sqrt{n-1}} \leq t_{\frac{\alpha}{2}, n-1}$$

Sustituyendo tendremos:

$$\frac{|2,8 - 2,5|}{0,4/\sqrt{15}} = 2,90$$

Para $\alpha = 0,05$ se tiene que: $t_{\frac{\alpha}{2}, n-1} = 2,131$

Por ser 2,90 > 2,131, se rechaza la hipótesis nula H_o y podemos concluir que la muestra no es compatible con la población.

****

7.9 *Se están utilizando normalmente en una granja avícola dos tipos de piensos compuestos A y B. Queriendo averiguar si la media de engorde con ambos piensos es idéntica, para un nivel de significación del 5%, se alimenta a 20 aves durante cierto tiempo con el pienso A y se obtiene una ganancia media de peso por ave de 0,4 Kgrs. con una desviación típica de 0,2 Kgrs.*

Simultaneamente a otras 20 aves se las alimenta con el pienso B y se obtiene un engorde medio de 0,5 Kgrs. con una desviación típica de 0,3 Kgrs.

Suponemos que las variables objeto de estudio, engorde con cada uno de los dos piensos, son normales y con la misma varianza σ^2.

Comparense los dos piensos compuestos.

SOLUCION:

Queremos comparar dos medias μ_A y μ_B. Las poblaciones son normales con la misma varianza σ^2 desconocida, que debemos estimar. Las muestras son pequeñas.

Hipótesis nula H_o: $\mu_A = \mu_B$

Hipótesis alternativa H_a: $\mu_A \neq \mu_B$

Aceptaremos la hipótesis nula H_o si se verifica:

$$\frac{|\bar{x}_A - \bar{x}_B|}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \leq t_{\frac{\alpha}{2},\, n_A+n_B-2}$$

siendo

$$s^2 = \frac{n_A s_A^2 + n_B s_B^2}{n_A + n_B - 2} = \frac{20.0{,}04 + 20.0{,}09}{20 + 20 - 2} = 0{,}068$$

$$\frac{|0{,}4 - 0{,}5|}{0{,}26\sqrt{\frac{1}{20} + \frac{1}{20}}} = 1{,}216$$

Para $\alpha = 0{,}05$ se tiene que: $t_{\frac{\alpha}{2},\, n_A + n_B - 2} = 2{,}025$.

Ahora bien, como $1{,}216 < 2{,}025$, aceptamos la hipótesis H_o y por tanto concluimos que la media de engorde con los dos piensos compuestos es la misma.

****

7.10 *90 inadaptados son ingresados en centros de rehabilitación. De ellos 50 en el Centro A, donde consiguieron su recuperación en un plazo medio de 150 días, con una desviación típica de 30 días. Los 40 restantes fueron ingresados en otro Centro B, donde se recuperaron en 160 días con una desviación típica de 25 días.*

a) ¿Se puede considerar que el Centro de Rehabilitación A es más adecuado que el B, para conseguir una recuperación más rápida?

b) ¿Bajo qué hipótesis?

SOLUCION:

a) Hipótesis nula H_o: $\mu_A \leq \mu_B$

Hipótesis alternativa H_a: $\mu_A > \mu_B$

Aceptaremos H_o si: $$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \leq z_\alpha$$

Sustituyendo los datos del enunciado resulta:

$$\frac{150 - 160}{\sqrt{\frac{900}{50} + \frac{625}{40}}} = -1{,}72$$

Tomamos $\alpha = 0{,}05$ por lo que: $z_\alpha = 1{,}64$.

luego por ser $-1{,}72 < 1{,}64$ aceptamos la hipótesis H_o, con lo cual podemos considerar el centro A mas adecuado que el B.

b) Lo anteriormente expuesto es cierto, si la variable objeto de estudio se distribuye según la ley normal.

****

7.11 *Se quieren comparar dos poblaciones de rana pipiens aisladas geográficamente. Para ello se toman dos muestras de ambas poblaciones y se les mide la longitud del cuerpo expresado en milimetros, obteniendose los siguientes resultados:*

$\bar{x}_1 = 74;\quad s_1^2 = 225;\quad n_1 = 42.$

$\bar{x}_2 = 78;\quad s_2^2 = 169;\quad n_2 = 56.$

a) Contrastar la hipótesis de igualdad de medias con un nivel de significación del 5%.

b) ¿Qué hipótesis debemos formular?

SOLUCION:

Queremos comparar dos medias μ_1 y μ_2. No se conocen las varianzas σ_1^2 y σ_2^2. Las muestras son grandes.

Hipótesis nula H_o: $\mu_1 = \mu_2$

Hipótesis alternativa H_a: $\mu_1 \neq \mu_2$

Aceptaremos H_o si: $$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \leq z_{\alpha/2}$$

siendo $\hat{s}_1^2$ y $\hat{s}_2^2$ los estimadores insesgados de σ_1^2 y σ_2^2 respectivamente.

$$\hat{s}_1^2 = \frac{42}{41}\,225 = 230,49$$

$$\hat{s}_2^2 = \frac{56}{55}\,169 = 172,07$$

Por ser las muestras grandes podiamos haber tomado como valores aproximados de $\hat{s}_1^2$ y $\hat{s}_2^2$ los de s_1^2 y s_2^2 respectivamente.

a) Sustituyendo:

$$\frac{|74 - 78|}{\sqrt{\dfrac{230,48}{42} + \dfrac{172,07}{56}}} = 1,37.$$

Tomando $\alpha = 0,05$ resulta que: $z_{\alpha/2} = 1,96$.

Por ser $1,37 < 1,96$ se acepta la hipótesis de igualdad de medias.

b) Todo lo anteriormente expuesto no es válido si la variable objeto de estudio, longitud del cuerpo de la rana pipiens, no se distribuye según la ley normal, que es la hipótesis que debemos formular.

****

7.12 *Un laboratorio farmaceútico fabrica dos tipos de somníferos A y B. Se toman dos grupos análogos de enfermos de insomnio formados por 80 y 100 inidividuos respectivamente, suministrando a los enfermos del primer grupo el somnífero A y a los del segundo grupo, el B.*

El número medio de horas de sueño para los enfermos del primer grupo fué de 7,84 con una desviación típica de 0,90 y para los del segundo grupo fué de 6,90 y 1,30 respectivamente.

a) ¿Se puede decir que la diferencia entre los números medios de horas de sueño es significativa?

b) ¿Qué hipótesis debemos formular?.

SOLUCION:

a) Queremos comparar dos medias μ_1 y μ_2.

No conocemos las varianzas σ_1^2 y σ_2^2; y las muestras tomadas son grandes.

Hipótesis nula H_o: $\mu_1 = \mu_2$

Hipótesis alternativa H_a: $\mu_1 \neq \mu_2$

Aceptaremos H_o si: $$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \leq z_{\alpha/2}$$

Sustituyendo los valores del enunciado, tendremos:

$$\frac{|7,84 - 6,90|}{\sqrt{\frac{0,81}{80} + \frac{1,69}{100}}} = 5,72.$$

Tomamos $\alpha = 0,05$ por tanto: $z_{\alpha/2} = 1,96.$

Por ser $5,72 > 1,96$ rechazamos la hipótesis H_o, luego existe una diferencia significativa entre los números medios de horas de sueño con ambos somniferos.

b) La conclusión anterior no es válida si no se cumple que la duración del sueño con ambos somniferos es una variable aleatoria normal.

****

7.13 *Se sabe que la duración de una determinada enfermedad sigue la ley normal. Para la curación de dicha enfermedad se aplica un determinado antibiótico. Se desea comparar la duración de la enfermedad según que al enfermo se le haya aplicado o no, en otra ocasión dicho antibiótico.*

Observamos 6 enfermos a los que no se había aplicado anteriormente el antibiótico y la duración media de la enfermedad ha sido de 12 dias y 5 enfermos a los que sí se había aplicado, obteniendo una duración de la enfermedad de 15 dias.

La estimación común de la varianza es 16.

¿Qué podemos afirmar acerca de la duración de la enfermedad para un nivel de significación $\alpha = 0,01$*?*

SOLUCION:

Queremos comparar dos medias μ_1 y μ_2. Las muestras son pequeñas.

Hipótesis nula H_o: $\mu_1 = \mu_2$

Hipótesis alternativa H_a: $\mu_1 \neq \mu_2$

Aceptaremos H_o si: $$\frac{|\bar{x}_1 - \bar{x}_2|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\frac{\alpha}{2},\, n_1 + n_2 - 2}$$

donde $s^2 = 16;$ $\bar{x}_1 = 12;$ $n_1 = 6;$ $\bar{x}_2 = 15;$ $n_2 = 5.$

Sustituyendo en la expresión anterior, resulta que el estadístico es igual a: 1,238.

Tomando $\alpha = 0{,}01$ se tiene que : $t_{\alpha/2,\ n_1+n_2-2} = 3{,}250.$

Ahora bien, como $1{,}238 < 3{,}250$ admitiremos la hipótesis.

Es decir, la duración media de la enfermedad es la misma para los enfermos a los que se les ha aplicado anteriormente el antibiótico que para los que no .

****

7.14 *Un instituto de dietética requiere comparar dos dietas. Se selecciona una muestra de 50 individuos, al azar de una población de personas comparables con exceso de peso.*

A 25 personas se les suministra la dieta A y a los 25 restantes la dieta B. Las pérdidas de peso expresadas en kgrs. al cabo de una semana, son las siguientes:

Dieta A: $\bar{x}_A = 4{,}3$ *Kgrs.* $s_A = 1{,}4$ *kgrs.*

Dieta B: $\bar{x}_B = 3{,}6$ *Kgrs.* $s_B = 1{,}1$ *Kgrs.*

Contrástese la hipótesis de igualdad de medias con un nivel de significación de 0,05, suponiendo normal la variable pérdida de peso.

S O L U C I O N:

Debemos en primer lugar contrastar la hipótesis de igualdad de varianzas.

Hipótesis nula H_o: $\sigma_A^2 = \sigma_B^2$

Hipótesis alternativa H_a: $\sigma_A^2 = \sigma_B^2$

Aceptaremos la hipótesis H_o si:

$$\frac{\hat{s}_A^2}{\hat{s}_B^2} \in \left(F_{1-\frac{\alpha}{2},\ n_A-1,\ n_B-1}\ ,\ F_{\frac{\alpha}{2},\ n_A-1,\ n_B-1}\right)$$

$$\left.\begin{aligned} \hat{s}_A^2 &= \frac{25}{24}\,1{,}96 = 2{,}04 \\ \hat{s}_B^2 &= \frac{25}{24}\,1{,}21 = 1{,}26 \end{aligned}\right\} \quad \frac{\hat{s}_A^2}{\hat{s}_B^2} = \frac{2{,}04}{1{,}26} = 1{,}62.$$

$\alpha = 0{,}05;\qquad F_{0{,}975,\ 24,\ 24} = 0{,}44;\qquad F_{0{,}025,\ 24,\ 24} = 2{,}27.$

Como 1,62 pertenece al intervalo (0,44 , 2,27), aceptamos la hipótesis acerca de que las varianzas son iguales.

Contrastemos ahora la hipótesis de igualdad de medias de dos poblaciones normales de varianzas desconocidas pero iguales y muestras pequeñas.

Hipótesis nula H_o: $\mu_A = \mu_B$

Hipótesis alternativa H_a: $\mu_A \neq \mu_B$

Aceptaremos la hipótesis H_o, si se verifica:

$$\frac{|\bar{x}_A - \bar{x}_B|}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \leq t_{\frac{\alpha}{2},\ n_A + n_B - 2}$$

siendo $$s^2 = \frac{(n_A - 1)\ \hat{s}_A^2 + (n_B - 1)\ \hat{s}_B^2}{(n_A - 1) + (n_B - 1)} = \frac{24.2{,}04 + 24.1{,}26}{24 + 24} = 1{,}65.$$

Sustituyendo resulta: $$\frac{|4{,}3 - 3{,}6|}{1{,}2845\sqrt{\frac{1}{25} + \frac{1}{25}}} = 1{,}93.$$

Tomando $\alpha = 0{,}05$ se tiene que: $t_{\frac{\alpha}{2},\ 48} = 2{,}01.$

Por ser $1{,}93 < 2{,}01$ admitimos la hipótesis acerca de la igualdad de medias en la reducción de peso para las dietas A y B.

****

7.15 *En una empresa de fundición se recibe periodicamente mineral de hierro procedente de dos yacimientos distintos A y B.*

Para estudiar la calidad del mineral recibido se extraen dos muestras y se analiza la riqueza en hierro, obteniendo los siguientes resultados en tanto por ciento:

A :	43	45	42	35	37	38	33	38	41	43		
B :	39	36	35	37	40	39	40	38	35	39	38	34

Suponiendo normal la distribución de la riqueza del mine *ral en ambos yacimientos, se puede admitir que la diferencia de varianzas es significativa a un nivel $\alpha = 0,05$.*

SOLUCION:

Vamos a contrastar la igualdad de varianzas de dos poblaciones normales.

Hipótesis nula H_o: $\sigma_A^2 = \sigma_B^2$

Hipótesis alternativa H_a: $\sigma_A^2 \neq \sigma_B^2$

Aceptaremos la hipótesis H_o si $\hat{s}_A^2 / \hat{s}_B^2$ pertenece al intervalo:

$$\left(F_{1-\frac{\alpha}{2},\, n_A-1,\, n_B-1} \qquad F_{\frac{\alpha}{2},\, n_A-1,\, n_B-1}\right)$$

$$\left.\begin{array}{l} \hat{s}_A^2 = \dfrac{1}{n_A-1}\sum (x_{Ai}-\bar{x}_A)^2 = 15,17 \\[2ex] \hat{s}_B^2 = \dfrac{1}{n_B-1}\sum (x_{Bj}-\bar{x}_B)^2 = 4,27 \end{array}\right\} \qquad \left\{\begin{array}{l} \bar{x}_A = 39,5 \\ \bar{x}_B = 37,5 \end{array}\right.$$

Tomando $\alpha = 0,05$; $F_{1-\frac{\alpha}{2},\, 9,\, 11} = 0,25$; $F_{\frac{\alpha}{2},\, 9,\, 11} = 3,59$.

Sustituyendo estos valores, resulta que:

$$\frac{\hat{s}_A^2}{\hat{s}_B^2} = \frac{15,17}{4,27} = 3,55$$

Como 3,55 pertenece al intervalo (0,25 3,59), no es significativa la diferencia entre las varianzas.

****

7.16 *Se ignora la proporción de familias numerosas y con el fin de determinar dicha proporción se toma una muestra de 800 familias siendo la proporción observada 0,18.*

Formulamos la hipótesis nula H_o: $p = 0,20$ frente a la alternativa H_a: $p \neq 0,20$ y queremos contrastarla para un nivel de significación $\alpha = 0,05$.

SOLUCION:

Aceptaremos H_o si se verifica:

$$\frac{|\hat{p} - p_o|}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\alpha/2}$$

donde $\hat{p} = 0{,}18$; $n = 800$; $p_o = 0{,}20$;

Si tomamos $\alpha = 0{,}05$ se tiene que: $z_{\alpha/2} = 1{,}96$.

Sustituyendo estos valores en la desigualdad anterior, resulta:

$$1{,}47 < 1{,}96$$

Admitimos la hipótesis nula, es decir, la proporción desconocida p se toma igual a 0,20.

****

7.17 *Un laboratorio farmacológico afirma que un analgésico que fabrica es efectivo en la eliminación del dolor de estómago durante 12 horas en un 95% de los casos.*

Se tomó una muestra de 150 individuos que padecían de dolor de estómago y se les suministró el analgésico citado, siendo efectivo para 132 individuos. ¿Qué podemos opinar acerca de la afirmación del laboratorio para un nivel de significación $\alpha = 0{,}05$?

SOLUCION:

Hacemos el siguiente contraste unilateral:

Hipótesis nula H_o: $p \geq p_o$

Hipótesis alternativa H_a: $p < p_o$

Aceptaremos H_o si se verifica:

$$\frac{\hat{p} - p_o}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} > -z_\alpha$$

donde $p_o = 0{,}95$; $\hat{p} = \frac{132}{150} = 0{,}88$; $\alpha = 0{,}05$; $-z_\alpha = -1{,}64$.

Sustituyendo estos valores, en la desigualdad anterior, resulta:

$$-2{,}638 < -1{,}64$$

Luego, tenemos que rechazar la hipotesis nula H_o, por lo que podemos afirmar que la manifestación hecha por el laboratorio no es correcta.

****

7.18 *Se sospecha que añadiendo al tratamiento habitual para* *la curación de una determinada enfermedad un medicamento A, se consigue mayor número de curaciones.*

Tomamos dos grupos de enfermos de 100 individuos cada uno. A un grupo se le suministra el medicamento A y se curan 60 enfermos y al otro grupo no se le suministra, curandose 55.

¿Es efectivo el tratamiento A en la curación de la enfermedad, al nivel de significación de $\alpha = 0,01$?

SOLUCION:

Hipótesis nula H_o: $p_1 = p_2$

Hipótesis alternativa H_a: $p_1 \neq p_2$

Aceptaremos H_o, si se verifica:

$$\frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_{\alpha/2}$$

donde $\hat{p}_1 = \frac{60}{100} = 0,6$; $\hat{p}_2 = \frac{55}{100} = 0,55$; $n_1 = 100$; $n_2 = 100$

Fijando $\alpha = 0,01$ se tiene que: $z_{\alpha/2} = 2,57$.

Sustituyendo estos datos en la desigualdad anterior, tendremos:

$$0,716 < 2,57$$

Aceptaremos la hipótesis nula, es decir que podemos afirmar que el medicamento A no consigue un mayor número de curaciones.

****

7.19 *Se quiere comprobar la efectividad de una vacuna contra una enfermedad. Para ello se suministró la vacuna a 100 animales y se les comparó con un grupo testigo de otros 100. A los 200 se les contagió la enfermedad. Entre los vacunados murieron 8 como resultado de la enfermedad, y del grupo testigo hubo 20 muertos.*

¿Podemos concluir que la vacuna es eficaz en reducir la tasa de mortalidad? Utilícese un nivel de significación del 0,05.

SOLUCION:

Queremos comparar dos proporciones p_1 y p_2. Las muestras son grandes.

Hipótesis nula H_o: $p_1 \leq p_2$

Hipótesis alternativa H_a: $p_1 > p_2$

Aceptaremos la hipótesis H_o, si se verifica:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_\alpha$$

donde $\hat{p}_1 = 8/100 = 0{,}08$; $\hat{p}_2 = 20/100 = 0{,}20$; $n_1 = 100$; $n_2 = 100$.

Para $\alpha = 0{,}05$ se tiene que: $z_\alpha = 1{,}64$.

Sustituyendo estos valores en la expresión anterior, se obtiene:

$$-2{,}48 < 1{,}64$$

Admitiremos la hipótesis H_o, o lo que es lo mismo podremos afirmar que la vacuna es eficaz para reducir la tasa de mortalidad.

****

7.20 *Ante una epidemia de eimeria, un granjero sospecha que el número de aves jóvenes muertas es inferior que el de aves adultas*

Para salir de dudas, aisla una muestra de 200 aves jóvenes y al cabo de un mes se han muerto 58, igualmente aisla 150 aves adultas muriendo en el mismo periodo 36.

¿Qué conclusión puede sacar el granjero, con un nivel de significación del 5%?

SOLUCION:

Queremos comparar dos proporciones p_1 y p_2. Las muestras son grandes.

Hipótesis nula H_o: $p_1 \leq p_2$

Hipótesis alternativa H_a: $p_1 > p_2$

Aceptaremos la hipotesis H_o, si se verifica:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_\alpha$$

donde $\hat{p}_1 = 58/200 = 0{,}29$; $\hat{p}_2 = 36/150 = 0{,}24$; $n_1 = 200$; $n_2 = 150$.

Para $\alpha = 0{,}05$ se tiene que: $z_\alpha = 1{,}64$.

Sustituyendo estos valores en la expresión anterior tendremos:

$1{,}055 < 1{,}64$, luego la sospecha del granjero es cierta.

****

8-aplicaciones de la χ^2

8 - APLICACIONES DE LA χ^2

1. INTRODUCCION

Hemos estudiado algunos aspectos o planteamientos de un problema que, de forma general, trata de formular leyes para una población a partir de consideraciones y estudios sobre una muestra de dicha población.

Aspectos nuevos de este problema que vamos a tratar, son los siguientes:

- *Conformidad* de una distribución experimental y una distribución teórica.
- *Dependencia o independencia* entre caracteres cualitativos.
- Contraste de *homogeneidad* de un conjunto de muestras, caso de un carácter cualitativo.

Hemos creido conveniente agrupar estos temas en el mismo capítulo, ya que todos ellos los podemos considerar como aplicaciones notables de la χ^2.

2. CONFORMIDAD DE UNA DISTRIBUCION EXPERIMENTAL Y UNA DISTRIBUCION TEORICA

Los problemas de *conformidad* se presentan en términos de confrontar los resultados de un experimento con una teoría. Se trata de ver en qué medida los resultados proporcionados por una muestra están o no de acuerdo con una teoría que suponemos se cumple para la población.

Esto hicimos cuando comparabamos dos medias o los parámetros de dos distribuciones binomiales; en este caso, tratamos de comparar una distribución de frecuencias experimentales con una distribución teórica.

Tomamos una muestra de la población y estudiamos cierto carácter cuantitativo o cualitativo, lo que nos conduce a tener un conjunto - de datos o *serie estadística*.

Si el carácter estudiado es cualitativo, se clasifican los resultados obtenidos en una tabla de frecuencias, y si el caracter es cuantitativo, se puede construir un histograma de frecuencias y calcular media, mediana, moda, desviación típica, etc.

Tratamos de sustituir la distribución experimental, histograma o tabla de frecuencias por una distribución teórica.

La distribución teórica se deduce de una serie de consideraciones teóricas sobre la población estudiada o viene sugerida por la - forma del histograma o por los valores de la tabla de frecuencias.

Escogida la distribución teórica y estimados los parámetros poblacionales a partir de los datos de la muestra, nos preguntaremos, si la distribución teórica representa o no a la distribción empírica.

Sea N el tamaño de la muestra que agruparemos en k clases y sea f_i la frecuencia absoluta de la clase i.

Sea p_i la probabilidad que teoricamente asociamos a la clase i, de donde la frecuencia absoluta teórica asociada a la clase i, será: Np_i.

Vamos a hacer la hipótesis H, consistente en suponer que la distribución teórica escogida está conforme con la distribución empírica, y que por lo tanto, las desviaciones entre las frecuencias observadas y las teóricas son únicamente debidas al azar en la toma de la muestra.

Una medida de esas desviaciones fué estudiada por Pearson, antes incluso que el desarrollo de la teoría general sobre contrastes de hipótesis, construyendo el siguiente estadístico:

$$\chi^2 = \frac{(f_1 - Np_1)^2}{Np_1} + \frac{(f_2 - Np_2)^2}{Np_2} + \ldots + \frac{(f_k - Np_k)^2}{Np_k}$$

como medida de dichas desviaciones.

Este estadístico sigue una distribución χ^2, con k-1 grados de libertad.

El valor del estadístico es más grande a medida que la distribución experimental se separa más de la distribución teórica, pero así mismo puede ser grande ajustándose bien la distribución experimental a la teórica, si el número de términos de la suma es grande.

¿A partir de qúe discrepancia dejaremos de aceptar la hipótesis H?.

Como dijimos en el capítulo anterior, determinaremos un valor "particular" crítico de χ^2, que llamaremos χ^2_α, el cual no se deberá pasar, si las discrepancias entre las frecuencias experimental y teóricas son debidas únicamente al azar en la toma de la muestra.

Aceptaremos H, si $\chi^2 < \chi^2_\alpha$; si por el contrario $\chi^2 \geq \chi^2_\alpha$ rechazaremos la hipótesis.

Del estudio teórico de la distribución χ^2 se deduce que para poder aplicar correctamente el test, las frecuencias de las diferentes clases deben ser grandes, practicamente no inferiores a 5.

Una noción muy importante a tener en cuenta en último lugar es la del *número de grados de libertad*.

Apuntamos como causa del incremento del valor del estadístico χ^2, el número de clases o términos de la suma, este número se puede fijar libremente y es por definición el número de grados de libertad.

Entre las frecuencias teóricas existe la relación:

$$N\,p_1 + N\,p_2 + \ldots + N\,p_k = N$$

luego el número de grados de libertad será k-1.

Toda relación suplementaria que añadamos a las frecuencias teóricas, nos conduce a reducir en una unidad el número de grados de - libertad. Así, si tomamos la media y la varianza de la distribución experimental como media y varianza de la distribución teórica, redu ciremos en 2 unidades el número de grados de libertad que serán:

$$k - 1 - 2 = k - 3.$$

3. RELACIONES DE DEPENDENCIA E INDEPENDENCIA ENTRE CARACTERES CUALITATIVOS

Hasta ahora hemos venido estudiando en una muestra un solo carácter, pero no hemos dicho nada acerca del estudio en una única muestra de dos o más caracteres simultaneamente y en la que uno de ellos puede a su vez admitir dos o más estados diferentes.

Por ejemplo, en una muestra de individuos podemos considerar el color de la piel y el color del pelo simultaneamente, admitiendo cada variable varios estados, así la piel puede ser oscura o clara y el pelo negro, castaño o rubio.

Una serie de datos, clasificados con arreglo a dos o más características, constituye lo que se llama una *tabla de contingencia.*

El ejemplo anterior puede reflejarse en una tabla de contingencia de la siguiente forma:

		PELO			
		negro	castaño	rubio	total
PIEL	oscura	20	20	5	45
	clara	5	10	10	25
	total	25	30	15	70

Ante un problema de este tipo, nos surge inmediatamente la pregunta de si existe o no relación entre los caracteres estudiados.

En el ejemplo anterior, nos preguntamos si la piel clara es más frecuente en los individuos rubios o por el contrario ambos caracteres son independientes.

Si un carácter A admite n estados y un carácter B admite m estados, habrá un total de m.n clases diferentes, definida cada una por un estado particular de cada carácter.

En el ejemplo habrá 2.3 = 6 clases, siendo una de ellas la definida por los estados "piel clara, pelo castaño".

Decimos que dos caracteres A y B *son independientes*, cuando la proporción de individuos de cada uno de los estados del carácter A es la misma para las distintas clases que determinan el estado del carácter B.

En el ejemplo citado, si el color del pelo fuera independiente del color de la piel, deberíamos esperar las siguientes proporciones.

De los 25 individuos de pelo negro:

$$\begin{cases} \dfrac{45}{70} & \text{deberían tener la piel oscura} \\[2ex] \dfrac{25}{70} & \text{deberían tener la piel clara.} \end{cases}$$

Análogas proporciones para los individuos de pelo castaño o rubio.

Según esto, el número de individuos teóricos, que tendrían:

$$\begin{cases} \text{Pelo negro y piel oscura} & = 25\,\dfrac{45}{70} \\ \text{Pelo castaño y piel oscura} & = 30\,\dfrac{45}{70} \\ \text{Pelo rubio y piel oscura} & = 15\,\dfrac{45}{70} \end{cases}$$

Como vemos, estos números no coinciden con los que figuran en la tabla, pero es evidente que aunque los caracteres A y B fueran independientes en la población, tampoco coincidirían como consecuencia de las fluctuaciones debidas al azar. El problema, por tanto, estará en determinar si las diferencias observadas son suficientemente importantes como para pensar que nuestra hipótesis, de que los dos caracteres A y B son independientes, no es cierta.

Obsérvese que el contrastar esta hipótesis de independencia de dos caracteres, es el mismo problema que comparar una distribución teórica con una distribución empírica, que vimos en la pregunta anterior, ya que por un lado tendremos unas frecuencias de cada clase observadas y por otro lado unas frecuencias teóricas calculadas a partir de la hipótesis de independencia de los caracteres.

La representación general, correspondiente a los caracteres A y B, vendrá expresada por una tabla del siguiente tipo:

A \ B	B_1	B_2	. . .	B_m
A_1	f_{11}	f_{12}	. . .	f_{1m}
A_2	f_{21}	f_{22}	. . .	f_{2m}
$\vdots$	$\vdots$	$\vdots$	. . .	$\vdots$
A_n	f_{n1}	f_{n2}	. . .	f_{nm}

f_{ij} representa la frecuencia absoluta observada del par (A_i,B_j).

$\sum_i f_{ij}$ representa el total de la fila i.

$\sum_j f_{ij}$ representa el total de la columna j.

$\sum_i \sum_j f_{ij}$ representa el total general.

Al mismo tiempo, correspondiente a cada frecuencia observada f_{ij} en una tabla de contingencia, hay una frecuencia teórica, que representaremos por e_{ij} y viene dada por la siguiente expresión:

$$e_{ij} = \frac{\sum_i f_{ij} \sum_j f_{ij}}{\sum_i \sum_j f_{ij}}$$

Definimos el estadístico $\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$

que se aproxima a una distribución χ^2 con $\nu = (n-1)(m-1)$ grados de libertad, si $e_{ij} > 5$. $\forall i,j$.

La condición de que las frecuencias teóricas o esperadas deban ser mayores que 5, nos obligará a agrupar filas o columnas contiguas Este puede ser un gran problema, para el cual, la solución más sencilla será la *corrección de Yates*, que consiste en restar 0,5 a cada una de las diferencias entre las frecuencias observadas y las teóricas, antes de elevar al cuadrado.

Es decir:

$$\chi^2 \text{ (corregido)} = \sum_i \sum_j \frac{(|f_{ij} - e_{ij}| - 0,5)^2}{e_{ij}}$$

Una vez calculado el valor del estadístico χ^2, determinaremos una valor χ^2_α, que calcularemos en función del número de grados de libertad ν y del nivel de significación fijado α, acudiendo a las tablas de la χ^2 de Pearson. Si $\chi^2 \leq \chi^2_\alpha$ diremos que los caracteres A y B son independientes.

Para tablas de contingencia del tipo 2×2 y 2×3, pueden deducirse fácilmente las fórmulas que damos a continuación para hallar el valor del estadístico χ^2 sin calcular las frecuencias teóricas o esperadas.

<u>TABLA 2 × 2</u>

$$\chi^2 = \frac{N\,(f_{11}f_{22} - f_{12}f_{21})^2}{(f_{11} + f_{12})(f_{21} + f_{22})(f_{11} + f_{21})(f_{12} + f_{22})}$$

siendo $N = f_{11} + f_{12} + f_{21} + f_{22}$

$$\chi^2 \text{ (corregida)} = \frac{N(|f_{11}f_{22} - f_{12}f_{21}| - N/2)^2}{(f_{11} + f_{12})(f_{21} + f_{22})(f_{11} + f_{21})(f_{12} + f_{22})}$$

Del mismo modo para una tabla de contingencia 2×3, será:

TABLA 2×3

$$\chi^2 = \frac{N}{f_{11}+f_{12}+f_{13}}\left[\frac{f_{11}^2}{f_{11}+f_{21}}+\frac{f_{12}^2}{f_{12}+f_{22}}+\frac{f_{13}^2}{f_{13}+f_{23}}\right]+$$

$$+\frac{N}{f_{21}+f_{22}+f_{23}}\left[\frac{f_{21}^2}{f_{11}+f_{21}}+\frac{f_{22}^2}{f_{12}+f_{22}}+\frac{f_{23}^2}{f_{13}+f_{23}}\right]-N$$

4. CONTRASTE DE HOMOGENEIDAD DE VARIAS MUESTRAS

Los problemas de homogeneidad se presentan en términos de ver si dos o más muestras pueden ser consideradas como procedentes de una misma población, o por el contrario, si las diferencias que se presentan son tan grandes que no pueden ser imputables únicamente al muestreo, con lo que consideraremos que las muestras no son homogeneas.

Un ejemplo de este tipo de problemas, consiste en estudiar si son homogeneos respecto a los resultados tres tratamientos A, B, y C aplicados a un cierto tipo de enfermos con el fin de conseguir su curación.

Si se tratara de estudiar dos tratamientos, el problema ya fué estudiado en el capítulo anterior, cuando comparabamos dos porcentajes.

Ahora bien , para el caso de tres tratamientos, podriamos compararlos dos a dos, pero en este momento lo que queremos es contrastar de una vez la homogeneidad del conjunto.

Para ello consideremos un cierto carácter cualitativo A.

Tomemos k muestras con $n_1, n_2, \ldots, n_k$ elementos cada una de ellas y sean $a_1, a_2, \ldots, a_k$ el número de elementos que posee el carácter A. El número de elementos que no poseen dicho carácter será:

$$n_1 - a_1, \quad n_2 - a_2, \quad \ldots \quad , n_k - a_k$$

Como hemos dicho, el problema de homogeneidad consiste en comprobar si las k muestras han sido extraidas de la misma población.

Hacemos la hipótesis H de que las k muestras proceden de la misma población.

La mejor estimación del porcentaje de elementos de la población que poseen el carácter A, será:

$$p = \frac{a_1 + a_2 + \ldots + a_k}{n_1 + n_2 + \ldots + n_k}$$

Bajo la hipótesis H, el número de elementos teóricos de cada muestra que poseen el carácter A, será:

$$x_1 = p\, n_1; \quad x_2 = p\, n_2; \quad , \ldots , \quad x_k = p\, n_k$$

y los que no lo poseen:

$$n_1 - x_1; \quad n_2 - x_2; \quad , \ldots , n_k - x_k$$

Tenemos unas frecuencias empíricas $a_1, a_2, \ldots, a_k, n_1-a_1, n_2-a_2, \ldots n_k-a_k$ y unas frecuencias teóricas $x_1, x_2, \ldots x_k, n_1-x_1, n_2-x_2, \ldots n_k-x_k$ y tratamos de ver si las diferencias entre las mismas son imputables únicamente al muestreo.

Obsérvese que hemos llegado, al igual que anteriormente, a comparar una distribución experimental con una distribución teórica, problema que ya ha sido estudiado.

La diferencia entre las frecuencias, la mediremos por:

$$\chi^2 = \frac{(a_1-x_1)^2}{x_1} + \frac{(a_2-x_2)^2}{x_2} + \ldots + \frac{(a_k-x_k)^2}{x_k} + \frac{\left[(n_1-a_1)-(n_1-x_1)\right]^2}{n_1-x_1} + \ldots$$
$$+ \ldots + \frac{\left[(n_k-a_k)-(n_k-x_k)\right]^2}{n_k-x_k}$$

Para un nivel de significación dado α y para $\nu = k-1$ grados de libertad, buscamos en las tablas de la χ^2 de Pearson el valor χ^2_α.

Si $\chi^2 < \chi^2_\alpha$, aceptamos la hipótesis H de homogeneidad de las muestras.

****

PROBLEMAS RESUELTOS

8.1 *¿Se puede admitir la distribución uniforme de edades en una gran población de la que hemos tomado una muestra aleatoria de 100 individuos y hemos obtenido los siguientes resultados:*

Edades en años	*N°de individuos*
< 15	16
15 - 30	22
30 - 45	20
45 - 60	19
> 60	23

S O L U C I O N:

La probabilidad teórica de cada clase según la ley de distribución uniforme, es igual a la unidad dividida por el número de clases, es decir: $\frac{1}{5} = 0,20.$

Formamos la siguiente tabla:

Edades	frecuencia observada f_i	probabilidad teórica p_i	frecuencia teórica Np_i
< 15	16	0,20	20
15 - 30	22	0,20	20
30 - 45	20	0,20	20
45 - 60	19	0,20	20
> 60	23	0,20	20

$$\chi^2 = \sum \frac{(f_i - N\,p_i)^2}{N\,p_i}$$

$$\chi^2 = \frac{(16-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(23-20)^2}{20} = 1,5$$

El número de grados de libertad es 5-1 = 4. Tomamos el nivel de significación $\alpha = 0,05$ y buscamos en la tabla de la "chi-cuadrado" de Pearson: $\chi^2_\alpha = 9,49$

Por ser $\chi^2 < \chi^2_\alpha$, la diferencia entre la distribución empírica y la ley de distribución uniforme no es significativa, o lo que es lo mismo, admitiremos que la distribución de individuos por edades es uniforme en la población considerada.

****

8.2 *Se ha observado el número de hijos varones en 1000 familias, cada una de ellas con 5 hijos, obteniéndose los siguientes resultados:*

Nº de hijos varones	*Nº de familias*
0	31
1	168
2	319
3	308
4	150
5	24

Ajustar una distribución binomial y estudiar la bondad del ajuste.

S O L U C I O N:

Sea X la variable que representa el número de hijos varones en una familia;

$$p(X = k) = \binom{n}{k} p^k q^{n-k} \quad ;\text{con } n = 5$$

p y q son las probabilidades de que un hijo sea varon o hembra respectivamente.

El número medio de hijos varones observados, es:

$$\bar{x} = \frac{31.0 + 168.1 + 319.2 + 308.3 + 150.4 + 24.5}{1000} = 2,455.$$

Como la media de la distribucion es n.p, identificando, tendremos:

$$2,455 = 5.p \qquad p = 0,49$$

Por tanto, la distribución binomial ajustada, viene dada por:

$$p(X = k) = \binom{5}{k} 0,49^k \, 0,51^{5-k}$$

Para estudiar la bondad del ajuste, determinaremos las frecuencias teóricas o esperadas que recogemos en la siguiente tabla:

nºde hijos varones	frecuencia observada f_i	$p_i = p(X=k)$	frecuencia teórica
0	31	0,0345	34,5
1	168	0,1657	165,7
2	319	0,3185	318,5
3	308	0,3060	306,0
4	150	0,1470	147,0
5	24	0,0283	28,3

$$\chi^2 = \sum_i \frac{(f_i - N\,p_i)^2}{N\,p_i}$$

$$\chi^2 = \frac{(31-34,5)^2}{34,5} + \frac{(168-165,7)^2}{165,7} + \frac{(319-318,5)^2}{318,5} + \frac{(308-306)^2}{306} +$$

$$+ \frac{(150-147)^2}{147} + \frac{(24-28,3)^2}{28,3} = \mathbf{1,115}$$

El número de grados de libertad es: $6 - 1 - 1 = 4$.

Tomamos como nivel de significación $\alpha = 0,05$, por lo que $\chi^2_\alpha = 9,49$

Por ser $1,115 < 9,49$, podemos afirmar que el ajuste es bueno

****

8.3 *A lo largo de 540 días se anota el número de accidentes mortales de tráfico que se producen en una ciudad, obteniéndose los resultados de la tabla adjunta:*

nºde accidentes mortales por día	*nºde días*
0	132
1	195
2	120
3	60
4	24
5	9

a) ¿Qué distribución podremos ajustar y por qué?.
b) Estudiar la bondad del ajuste.
c) ¿Cuántos días se producirán dos accidentes mortales en un año?

S O L U C I O N:

a) Comencemos calculando la media y la varianza del número de accidentes por día.

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = 1,4; \qquad s^2 = \frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2 = 1,417.$$

Como vimos al estudiar la distribución de Poisson, sabemos que la media es igual a la varianza, luego teniendo en cuenta lo obtenido, la distribución de Poisson nos dará una buena aproximación a la distribución empírica considerada.

La estimación insesgada del parámetro λ de una distribución de Poisson, es $\bar{x}$, luego:

$$p(X=k) = e^{-1,4} \frac{1,4^k}{k!}$$

b) Vamos a determinar la bondad del ajuste mediante un contraste de la "chi-cuadrado" de Pearson.

$$\chi^2 = \sum \frac{(f_i - N\,p_i)^2}{N\,p_i}$$

Para ello formemos la siguiente tabla:

x_i	f_i	p_i	$N\,p_i$
0	132	0,2466	133,164
1	195	0,3452	186,408
2	120	0,2417	130,518
3	60	0,1128	60,912
4	24	0,0395	21,330
5	9	0,0111	5,994

$$\chi^2 = \frac{(132-133,164)^2}{133,164} + \frac{(195-186,408)^2}{186,408} + \frac{(120-130,518)^2}{130,518} +$$

$$+ \frac{(60-60,912)^2}{60,912} + \frac{(24-21,330)^2}{21,330} + \frac{(9-5,994)^2}{5,994} = \mathbf{3,109}$$

El número de grados de libertad es $6-1-1 = 4$.

Tomamos el nivel de significación $\alpha = 0,05$ y buscamos en la tabla de la "chi-cuadrado" de Pearson, obteniendo $\chi^2_\alpha = \mathbf{9,49}$

Como $\chi^2 < \chi^2_\alpha$, entonces la diferencia entre la distribución empírica y la distribución de Poisson de parámetro $\lambda = 1,4$ no es significativa.

c) $p(X = 2) = 0,2417$. por tanto $365.0,2417 = 88,22$ luego teóricamente se producirán en 88 dias dos accidentes mortales durante un año.

****

8.4 *Se mide el número de partículas α que llegan a una determinada zona procedentes de una sustancia radioactiva en un corto espacio de tiempo siempre igual, anotándose los resultados en la siguiente tabla:*

nº de partículas	0	1	2	3	4	5	6
nº de periodos de tiempo	269	325	207	82	28	7	2

a) Ajustar una distribución de Poisson
b) Calcular la probabilidad de que lleguen a dicha superficie 0, 1, 2, 3, ..., 6 partículas α.
c) Verificar el ajuste mediante un contraste de la χ^2.

SOLUCION:

a) El número medio de partículas α , por periodo de tiempo es:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{325 + 414 + 246 + 112 + 35 + 12}{920} = 1,24.$$

La varianza empírica s^2, es:

$$s^2 = \frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2 = 1,56.$$

Como la media y la varianza son casi iguales, entonces la distribución de Poisson sugerida en el enunciado nos puede proporcionar un buen ajuste.

Por otra parte, sabemos que un estimador insesgado para el parámetro desconocido λ , es la media muestral. Luego $\hat{\lambda} = 1,24$.

Por tanto $\quad p(X = r) = e^{-1,24} \dfrac{1,24^r}{r!}$

b) Las probabilidades de $r = 0, 1, 2, \dots, 6.$ son:

r	0	1	2	3	4	5	6
p(X = r)	,2898	,3586	,2222	,0919	,0285	,0070	,0014

c) Estudiemos la bondad del ajuste.

Como los efectivos deben ser todos superiores a 5 para poder hacer el contraste de la χ^2, debemos agrupar las dos últimas clases, obteniendo la siguiente tabla:

r	f_i	p_i	$N p_i$	f_i-Np_i	$(f_i-Np_i)^2$	$\frac{(f_i-Np_i)^2}{Np_i}$
0	269	0,2898	267	2	4	0,0149
1	325	0,3586	330	-5	25	0,0757
2	207	0,2222	204	3	9	0,0441
3	82	0,0919	85	-3	9	0,1058
4	28	0,0285	26	2	4	0,1538
5y6	9	0,0084	7	2	4	0,5714
						0,9657

Luego $\chi^2 = 0,9657$

El número de grados de libertad es: $\nu = 6-1-1 = 4$

Tomamos como nivel de significación $\alpha = 0,05$, por tanto, de la tabla obtenemos el valor: $\chi^2_\alpha = 9,49$

Como $\chi^2 < \chi^2_\alpha$, aceptamos la hipótesis de que la distribución empírica se ajusta bien mediante la ley de Poisson de parámetro $\lambda = 1,24$.

****

8.5 *Tomamos una muestra de 650 análisis de sangre realizados en un laboratorio clínico y anotamos el número de eritrocitos por milímetro cúbico de sangre. Los resultados agrupados en siete clases son los que figuran en la tabla adjunta.*

a) *¿Se puede admitir que el número de eritrocitos se distribuyen normalmente?*

b) *Calcular la probabilidad de que el número de eritrocitos en millones, esté comprendido entre 4,5 y 5,5.*

nº de eritrocitos en millones	*nº de análisis*
menos de 2,5	*8*
2,5 - 3,5	*52*
3,5 - 4,5	*140*
4,5 - 5,5	*210*
5,5 - 6,5	*160*
6,5 - 7,5	*70*
7,5 y más	*10*

SOLUCION:

a) Vamos a ajustar una distribución normal a la distribución empírica dada, para determinar a continuación la bondad del ajuste.

La media y varianza muestrales son:

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{3312}{650} = 5{,}09. \qquad s^2 = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = 1{,}45.$$

La estimación insesgada de μ es $\bar{x}$ y la de σ^2 es $\hat{s}^2$, siendo:

$$\hat{s}^2 = \frac{N}{N-1}\, s^2$$

aunque al ser N grande $\hat{s}^2 \simeq s^2$

La distribución normal, a la que puede ajustarse esta distribución teórica será: $N(5{,}09\ ,\ 1{,}45)$.

Por tanto, las probabilidades teóricas asociadas a cada clase, las recogemos en la siguiente tabla:

clases	x_i	f_i	$f_i x_i$	$f_i x_i^2$	p_i	Np_i
1,5-2,5	2	8	16	32	0,0138	8,97
2,5-3,5	3	52	156	468	0,0765	49,72
3,5-4,5	4	140	560	2240	0,2168	140,92
4,5-5,5	5	210	1050	5250	0,3208	208,52
5,5-6,5	6	160	960	5760	0,2476	160,94
6,5-7,5	7	70	490	3430	0,1003	65,19
7,5-8,5	8	10	80	640	0,0206	13,39
T O T A L E S		650	3312	17820		

$$\chi^2 = \sum \frac{(f_i - Np_i)^2}{Np_i}$$

$$\chi^2 = \frac{(8-8,97)^2}{8,97} + \frac{(52-49,72)^2}{49,72} + \frac{(140-140,92)^2}{140,92} + \frac{(210-208,52)^2}{208,52} +$$

$$+ \frac{(160-160,94)^2}{160,94} + \frac{(70-65,19)^2}{65,19} + \frac{(10-13,39)^2}{13,39} = \mathbf{1,44}.$$

El número de grados de libertad es: $\nu = 7-1-2 = 4$.

Para el nivel de significación $\alpha = 0,05$, tenemos $\chi^2_\alpha = 9,49$.

Por ser $\chi^2 < \chi^2_\alpha$, admitiremos que la variable número de eritrocitos, se distribuye según una $N(5,09 \ , \ 1,45)$

b) $p(4,5 \leq X \leq 5,5) = 0,3208$. Este valor ya fué calculado anteriormente para determinar la bondad del ajuste.

****

8.6 *Una central de transformación de productos lacteos produce un preparado para niños en edad de lactancia. Se analiza el contenido de materia grasa de los mismos. Para ello se utiliza una muestra de 742 preparados y los resultados agrupados en 8 clases están expresados en la siguiente tabla:*

0,255 - 0,285	6
0,285 - 0,315	38
0,315 - 0,345	66
0,345 - 0,375	131
0,375 - 0,405	240
0,405 - 0,435	162
0,435 - 0,465	84
0,465 - 0,495	15

a) *¿Se puede admitir que le contenido de materia grasa se distribuye normalmente?.*

b) *Calcular las probabilidades teóricas asociadas a cada intervalo.*

c) *Contrastar la bondad del ajuste mediante la* χ^2.

SOLUCION:

a) La media es: $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i} = \dfrac{288,60}{742} = 0,39.$

La varianza muestral es:

$$s^2 = \frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2 = \frac{113{,}5692}{742} - 0{,}15 = 0{,}0018; \qquad s = 0{,}042.$$

clases	x_i	f_i	$f_i x_i$	$f_i x_i^2$
0,255 - 0,285	0,27	6	1,62	0,4374
0,285 - 0,315	0,30	38	11,40	3,4200
0,315 - 0,345	0,33	66	21,78	7,1874
0,345 - 0,375	0,36	131	47,16	16,9776
0,375 - 0,405	0,39	240	93,60	36,5040
0,405 - 0,435	0,42	162	68,04	28,5768
0,435 - 0,465	0,45	84	37,80	17,0100
0,465 - 0,495	0,48	15	7,20	3,4560
T O T A L E S		742	288,60	113,5692

Si construimos el correspondiente histograma de frecuencias, nos sugiere un ajuste por medio de una distribucion $N(\mu\ ,\ \sigma)$. Una estimación puntual insesgada de μ es $\bar{x}$ y de σ^2 es $\hat{s}^2$ pero como en este caso la muestra es grande, podemos identificar $\hat{s}^2 = s^2$.

Por tanto la distribución será: $N(0{,}39\ ,\ 0{,}042)$

b) Las probabilidades teóricas asociadas se calculan por medio de la tabla de la $N(0\ ,\ 1)$.

$$p(0{,}255 \leq X \leq 0{,}285) = p\left[\frac{0{,}255-0{,}39}{0{,}042} \leq Z \leq \frac{0{,}285-0{,}39}{0{,}042}\right] =$$

$$= 0{,}9993 - 0{,}9938 = 0{,}0055.$$

Del mismo modo tendremos:

$p(0{,}285 \leq X \leq 0{,}315) = 0{,}0313.$

$p(0{,}315 \leq X \leq 0{,}345) = 0{,}1048.$

$p(0{,}345 \leq X \leq 0{,}375) = 0{,}2171.$

$p(0{,}375 \leq X \leq 0{,}405) = 0{,}2812.$

$p(0{,}405 \leq X \leq 0{,}435) = 0{,}2171.$

$p(0{,}435 \leq X \leq 0{,}465) = 0{,}1048.$

$p(0{,}465 \leq X \leq 0{,}495) = 0{,}0313.$

c) Formemos el estadístico del contraste:

$$\chi^2 = \sum \frac{(f_i - N\,p_i)^2}{N\,p_i} = \mathbf{25{,}8022.}$$

El número de grados de libertad es: $\nu = 8-1-2 = 5$.

clases	f_i	p_i	Np_i	f_i-Np_i	$(f_i-Np_i)^2$	$\frac{(f_i-Np_i)^2}{Np_i}$
0,255-0,285	6	0,0055	4,0810	1,9190	3,6825	0,9023
0,285-0,315	38	0,0313	23,2246	14,7754	217,6805	9,3728
0,315-0,345	66	0,1048	77,7616	-11,7616	138,3352	1,7789
0,345-0,375	131	0,2171	161,0882	-30,0882	905,2997	5,6199
0,375-0,405	240	0,2812	208,6504	31,3496	982,7974	4,7102
0,405-0,435	162	0,2171	161,0882	0,9118	0,8313	0,0051
0,435-0,465	84	0,1048	77,7616	6,2384	38,9176	0,5004
0,465-0,495	15	0,0313	23,2246	-8,2246	67,6440	2,9126
						25,8022

Tomamos un nivel de significación del 0,05, de donde:

$$\chi^2_\alpha = 11{,}07$$

Como $\chi^2 > \chi^2_\alpha$, no podemos afirmar que el contenido de grasa siga una distribución normal N(0,39 , 0,042).

****

8.7 *Tiramos un dado 720 veces y obtenemos los siguientes resultados:*

x_i	f_i
1	116
2	120
3	115
4	120
5	125
6	124

Contrástese la hipótesis de que el dado está bien construido al nivel de significación $\alpha = 0{,}01$.

S O L U C I O N:

Formulamos la hipótesis H_o " el dado está bien construido".

Según H_o, la probabilidad teórica para cada cara del dado será: $\frac{1}{6}$, por tanto las frecuencias esperadas serán, para todo valor de x_i:

$$\text{frecuencia esperada de } x_i = 720.\frac{1}{6} = 120.$$

$$\chi^2 = \frac{(116-120)^2}{120} + \frac{(120-120)^2}{120} + \frac{(115-120)^2}{120} + \frac{(120-120)^2}{120} + \frac{(125-120)^2}{120} + \frac{(124-120)^2}{120} = \mathbf{0{,}683}$$

El número de grados de libertad es: $\nu = 6-1 = 5$.

Para $\alpha = 0{,}01$ se tiene que: $\chi^2_\alpha = 15{,}09$.

Como $\chi^2 < \chi^2_\alpha$, admitimos la hipótesis de que el dado está bien construido.

****

8.8 *Para comprobar las leyes de Mendel, cruzamos guisantes obteniendo los siguientes resultados:*

	amarillos lisos	*amarillos rugosos*	*verdes lisos*	*verdes rugosos*
número de guisantes	*262*	*91*	*86*	*31*

Las proporciones esperadas son 9, 3, 3, 1.

¿Contradicen estos resultados obtenidos, las proporciones esperadas?. Estudiese al nivel de significación del 1%.

S O L U C I O N:

El número teórico de guisantes de cada uno de los dos tipos citados según las proporciones conocidas, será:

$$\begin{cases} \text{amarillos lisos} & \frac{9}{16}\ 470 = 264{,}375. \\ \text{amarillos rugosos} & \frac{3}{16}.470 = 88{,}125. \\ \text{verdes lisos} & \frac{3}{16}.470 = 88{,}125. \\ \text{verdes rugosos} & \frac{1}{16}.470 = 29{,}375 \end{cases}$$

Según esto, el estadístico χ^2, será:

$$\chi^2 = \frac{(262-264,375)^2}{264,375} + \frac{(91-88,125)^2}{88,125} + \frac{(86-81,125)^2}{81,125} + \frac{(31-29,375)^2}{29,375} =$$

$$= \mathbf{0,498}$$

El número de grados de libertad es: $\nu = 4 - 1 = 3$.

Para $\alpha = 0,01$, tenemos $\chi^2_\alpha = 11,24$.

Como $\chi^2 < \chi^2_\alpha$ podemos afirmar que los resultados experimentales, estan de acuerdo con las proporciones teóricas proporcionadas por las leyes de Mendel.

****

8.9 *Ante una epidemia se desea contrastar si padecer la enfermedad es independiente de una vacunación previa. Para ello, se toma una muestra de 300 personas y se obtuvieron los siguientes resultados:*

ENFERMOS / VACUNADOS	SI	NO	
SI	40	110	150
NO	52	98	150
	92	208	300

S O L U C I O N:

Hacemos la hipótesis de que la vacunación no tiene ninguna influencia a la hora de padecer la enfermedad, o lo que es lo mismo que es independiente padecer la enfermedad del hecho de estar vacunado.

Las frecuencias teóricas o esperadas bajo la hipótesis H_o, serán:

$$p = \frac{92}{300}$$

Por tanto, la tabla de frecuencias teóricas será:

ENFERMOS / VACUNADOS	SI	NO
SI	46	104
NO	46	104

$$\chi^2 = \frac{(40-46)^2}{46} + \frac{(110-104)^2}{104} + \frac{(52-46)^2}{46} + \frac{(98-104)^2}{104} = \mathbf{2{,}257}.$$

El número de grados de libertad es: $\nu = 1$.

Tomando $\alpha = 0{,}05$, tenemos $\chi^2_\alpha = 3{,}84$.

Como $\chi^2 < \chi^2_\alpha$, aceptamos la hipótesis, es decir, el padecer la enfermedad es independiente del estar vacunado.

Aplicando la correción de Yates, habriamos obtenido:

$$\chi^2(\text{corregida}) = \frac{(|40-46|-0{,}5)^2}{46} + \frac{(|110-104|-0{,}5)^2}{104} + \frac{(|52-46|-0{,}5)^2}{46} +$$
$$+ \frac{(|98-104|-0{,}5)^2}{104} = \mathbf{1{,}897}.$$

Por supuesto, que también así aceptamos la hipótesis.

De una manera general diremos, que la corrección de Yates siempre disminuye el valor de χ^2.

****

8.10 *Los resultados de una encuesta realizada con el fin de determinar, si la edad de los individuos influye a la hora de contraer una enfermedad, fueron los siguientes:*

EDAD	Contraen la enfermedad	
	SI	NO
menos de 15 años	38	44
15 - 30	45	28
30 - 45	30	54
45 - 60	22	62
más de 60 años	20	57
	155	245

¿Se puede admitir la hipótesis de que el número de individuos que contraen la enfermedad, es independiente de la edad?

SOLUCION:

Bajo la hipótesis H_0, de que la edad no tiene ninguna importancia a la hora de contraer una enfermedad, cabria esperar las siguientes frecuencias teóricas que recogemos en la tabla adjunta, teniendo en cuenta que: $p = \frac{155}{155+245} = \frac{155}{400}$

y lo visto en el apartado 3 del presente capítulo.

Por ejemplo, la frecuencia teórica de los contraen la enfermedad y tienen menos de 15 años, será: $\frac{82}{400}.155 = 31,775.$

Analogamente obtendriamos las demas frecuencias teóricas.

EDADES	Contraen la enfermedad	
	SI	NO
Menos de 15 años	31,775	50,225
De 15 a 30	28,2875	44,7125
De 30 a 45	32,55	51,45
De 45 a 60	32,55	51,45
Mas de 60 años	29,8375	47,1625

$$\chi^2 = \frac{(38-31,775)^2}{31,775} + \frac{(45-28,2875)^2}{28,2875} + \frac{(30-32,55)^2}{32,55} + \frac{(22-32,55)^2}{32,55} +$$

$$+ \frac{(20-29,8375)^2}{29,8375} + \frac{(44-50,225)^2}{50,225} + \frac{(28-44,7125)^2}{44,7125} + \frac{(54-51,45)^2}{51,45} +$$

$$+ \frac{(62-51,45)^2}{51,45} + \frac{(57-47,1625)^2}{47,1625} = \mathbf{29,316}.$$

El número de grados de libertad es: $\nu = 5-1 = 4$.

Tomando $\alpha = 0,05$, tenemos $\chi^2_\alpha = 9,49$.

Como $\chi^2 > \chi^2_\alpha$, rechazamos la hipótesis H_o, luego el número de personas que contraen la enfermedad no es independiente de la edad.

****

8.11 *A la mitad de los 160 enfermos de un hospital se les somete a un determinado tratamiento adicional T, contabilizándose al cabo de un cierto tiempo que entre estos se han recuperado 63, mientras que de los no tratados solamente se han recuperado 57.*

Contrástese la hipótesis de que la curación es independiente de la aplicación del tratamiento T, al nivel de significación $\alpha = 0,05$.

SOLUCION:

Los datos del enunciado vamos a expresarlos en forma de tabla de contingencia.

	CURADOS	NO CURADOS	TOTAL
CON TRATAMIENTO T	63	17	80
SIN TRATAMIENTO T	57	23	80
TOTAL	120	40	160

Si hacemos la hipótesis H, de que el tratamiento T no tiene ninguna influencia a efectos de curación, es equivalente a afirmar que es independiente la curación de la aplicación del tratamiento.
Por tanto, las frecuencias esperadas bajo la hipótesis H, serán:

	CURADOS	NO CURADOS	TOTAL
CON TRATAMIENTO T	60	20	80
SIN TRATAMIENTO T	60	20	80
TOTAL	120	40	160

$$\chi^2 = \frac{(63-60)^2}{60} + \frac{(57-60)^2}{60} + \frac{(17-20)^2}{20} + \frac{(23-20)^2}{20} = \mathbf{1,2}$$

Buscamos en las tablas de la χ^2 para $\alpha = 0,05$ y $\nu = 1$, obteniendo: $\chi^2_\alpha = \mathbf{3,84}$

Luego, como $\chi^2 < \chi^2_\alpha$, aceptaremos la hipótesis, es decir, el tratamiento no tiene ninguna influencia a efectos de curación.

También podiamos calcular el valor de χ^2, sin necesidad de obtener previamente las frecuencias teóricas, como vimos es la página 255 para tablas 2×2.

$$\chi^2 = \frac{N(f_{11}f_{22}-f_{12}f_{21})^2}{(f_{11}+f_{12})(f_{21}+f_{22})(f_{11}+f_{21})(f_{12}+f_{22})} = \frac{160(63.23-17.57)^2}{80.80.120.40} = 1,2$$

****

8.12 *Contrástese la hipótesis H_0 de que la proporción de estudiantes suspendidos en BIOESTADISTICA, por los cuatro profesores de dicha*

	CURADOS	NO CURADOS	TOTAL
CON TRATAMIENTO T	63	17	80
SIN TRATAMIENTO T	57	23	80
TOTAL	120	40	160

Si hacemos la hipótesis H, de que el tratamiento T no tiene ninguna influencia a efectos de curación, es equivalente a afirmar que es independiente la curación de la aplicación del tratamiento.
Por tanto, las frecuencias esperadas bajo la hipótesis H, serán:

	CURADOS	NO CURADOS	TOTAL
CON TRATAMIENTO T	60	20	80
SIN TRATAMIENTO T	60	20	80
TOTAL	120	40	160

$$\chi^2 = \frac{(63-60)^2}{60} + \frac{(57-60)^2}{60} + \frac{(17-20)^2}{20} + \frac{(23-20)^2}{20} = \mathbf{1{,}2}$$

Buscamos en las tablas de la χ^2 para $\alpha = 0{,}05$ y $\nu = 1$, obteniendo: $\chi^2_\alpha = \mathbf{3{,}84}$

Luego, como $\chi^2 < \chi^2_\alpha$, aceptaremos la hipótesis, es decir, el tratamiento no tiene ninguna influencia a efectos de curación.

También podiamos calcular el valor de χ^2, sin necesidad de obtener previamente las frecuencias teóricas, como vimos es la página 255 para tablas 2×2.

$$\chi^2 = \frac{N(f_{11}f_{22}-f_{12}f_{21})^2}{(f_{11}+f_{12})(f_{21}+f_{22})(f_{11}+f_{21})(f_{12}+f_{22})} = \frac{160(63.23-17.57)^2}{80.80.120.40} = 1{,}2$$

****

8.12 *Contrástese la hipótesis H_0 de que la proporción de estudiantes suspendidos en BIOESTADISTICA, por los cuatro profesores de dicha*

	CURADOS	NO CURADOS	TOTAL
CON TRATAMIENTO T	63	17	80
SIN TRATAMIENTO T	57	23	80
TOTAL	120	40	160

Si hacemos la hipótesis H, de que el tratamiento T no tiene ninguna influencia a efectos de curación, es equivalente a afirmar que es independiente la curación de la aplicación del tratamiento.
Por tanto, las frecuencias esperadas bajo la hipótesis H, serán:

	CURADOS	NO CURADOS	TOTAL
CON TRATAMIENTO T	60	20	80
SIN TRATAMIENTO T	60	20	80
TOTAL	120	40	160

$$\chi^2 = \frac{(63-60)^2}{60} + \frac{(57-60)^2}{60} + \frac{(17-20)^2}{20} + \frac{(23-20)^2}{20} = \mathbf{1{,}2}$$

Buscamos en las tablas de la χ^2 para $\alpha = 0{,}05$ y $\nu = 1$, obteniendo: $\chi^2_\alpha = \mathbf{3{,}84}$

Luego, como $\chi^2 < \chi^2_\alpha$, aceptaremos la hipótesis, es decir, el tratamiento no tiene ninguna influencia a efectos de curación.

También podiamos calcular el valor de χ^2, sin necesidad de obtener previamente las frecuencias teóricas, como vimos es la página 255 para **tablas** 2×2.

$$\chi^2 = \frac{N(f_{11}f_{22}-f_{12}f_{21})^2}{(f_{11}+f_{12})(f_{21}+f_{22})(f_{11}+f_{21})(f_{12}+f_{22})} = \frac{160(63.23-17.57)^2}{80.80.120.40} = 1{,}2$$

****

8.12 *Contrástese la hipótesis H_0 de que la proporción de estudiantes suspendidos en BIOESTADISTICA, por los cuatro profesores de dicha*

8.13 *El rendimiento de la cosecha de un cereal se considera muy bueno, si la producción es superior a 25 Kgrs. por area de cultivo, bueno si es superior a 15 Kgrs. y malo si no llega a 15 kgrs.*

Se hacen 30 determinaciones del rendimiento en otras tantas parcelas donde se ha sembrado cereal de tipo A, y 30 determinaciones en parcelas donde se sembró un cereal tipo B.

Los resultados son los siguientes:

Tipo de cereal / Rendimiento	A	B
MUY BUENO	10	12
BUENO	14	10
MALO	6	8

¿Son igualmente efectivos para el cultivo los dos tipos de cereales A y B?

SOLUCION:

Formulamos la hipótesis H_0 de que los dos tipos de cereales A y B nos producen el mismo rendimiento en su cultivo.

Según esto:

- La probabilidad de que una parcela tenga un rendimiento muy bueno es: $\dfrac{10 + 12}{60} = 0{,}36.$
- La probabilidad de que una parcela tenga un rendimiento bueno es: $\dfrac{14 + 10}{60} = 0{,}40.$
- La probabilidad de que una parcela tenga un rendimiento malo es: $\dfrac{6 + 8}{60} = 0{,}23.$

Se trata de un contraste de homogeneidad. En función de lo anterior las frecuencias teóricas o esperadas, serán las que figuran en la tabla de la página siguiente

Como todas las frecuencias teóricas son superiores a 5, no es preciso tener que agrupar ninguna clase, de manera que podemos calcular directamente el estadístico χ^2.

asignatura fué la misma, para un nivel de significación del 5%, si una vez entregadas las actas, los resultados fueron:

	Prof A	*Prof B*	*Prof C*	*Prof D*	*total*
APROBADOS	*150*	*141*	*168*	*152*	*611*
SUSPENSOS	*30*	*40*	*32*	*37*	*139*
total	*180*	*181*	*200*	*189*	*750*

S O L U C I O N:

Bajo la hipótesis H_o, la proporción de suspensos será: $p = \frac{139}{750}$

por tanto la frecuencia teórica o esperada correspondiente a suspensos por el Profesor A., será:

$$\frac{180.139}{750} = 33{,}36$$

y análogamente calculariamos las demás.

La tabla de frecuencias teóricas será:

	Prof A	Prof B	Prof C	Prof D
APROBADOS	146,64	147,46	162,94	153,98
SUSPENSOS	33,36	33,54	37,06	35,02

$$\chi^2 = \frac{(150-146{,}64)^2}{146{,}64} + \frac{(141-147{,}46)^2}{147{,}46} + \frac{(168-162{,}94)^2}{162{,}94} + \frac{(152-153{,}98)^2}{153{,}98} +$$

$$+ \frac{(30-33{,}36)^2}{33{,}36} + \frac{(40-33{,}54)^2}{33{,}54} + \frac{(32-37{,}06)^2}{37{,}06} + \frac{(37-35{,}02)^2}{35{,}02} = \mathbf{2{,}928}$$

El número de grados de libertad es: $\nu = 4-1 = 3$.

Tomando $\alpha = 0{,}05$, se tiene que $\chi^2_\alpha = \mathbf{7{,}81}$.

Luego, al ser $\chi^2 < \chi^2_\alpha$, aceptaremos la hipótesis, es decir, admitiremos que las proporciones de suspensos de los distintos profesores son iguales.

****

FRECUENCIAS TEORICAS

Rendimiento \ Tipo de cereal	A	B
MUY BUENO	11	11
BUENO	12	12
MALO	7	7

El estadístico χ^2 será:

$$\chi^2 = \frac{(10-11)^2}{11} + \frac{(14-12)^2}{12} + \frac{(6-7)^2}{7} + \frac{(12-11)^2}{11} + \frac{(10-12)^2}{12} + \frac{(8-7)^2}{7} =$$

$$= \mathbf{1,134}$$

El número de grados de libertad es: $\nu = 3-1 = 2$.

Tomamos como nivel de significación $\alpha = 0,05$, de donde $\chi^2_\alpha = 5,99$.

Por ser $\chi^2 < \chi^2_\alpha$, no existe diferencia significativa entre la distribución teórica y la observada, es decir, el rendimiento de muy bueno bueno y malo es independiente del tipo de cereal sembrado A o B.

****

8.14 *La observación de 302 individuos procedentes de la segunda generación de padres de raza pura*

	FENOTIPO AL	FENOTIPO VL	FENOTIPO AR	FENOTIPO VR
número de individuos	161	61	63	17

diferenciados por dos pares de caracteres alelomorfos AALL, VVRR, en los que los caracteres A y L son dominantes, han dado los resultados que figuran en la tabla.

¿Se puede considerar que el crecimiento ha seguido las leyes de Mendel con un nivel de confianza del 95%?

SOLUCION:

Las probabilidades teóricas de cada fenotipo, suponiendo que se cumplan las leyes de Mendel, son las que figuran en la presente tabla: junto con el número teórico de cada fenotipo:

	FENOTIPO AL	FENOTIPO VL	FENOTIPO AR	FENOTIPO VR
probabilidad	9/16	3/16	3/16	1/16
número teórico de individuos	169	57	57	19

$$\chi^2 = \frac{(161-169)^2}{169} + \frac{(61-57)^2}{57} + \frac{(63-57)^2}{57} + \frac{(17-19)^2}{19} = \mathbf{1{,}50}$$

El número de grados de libertad es: $\nu = 4-1 = 3$.

Tomando $\alpha = 0{,}05$, resulta $\chi^2_\alpha = \mathbf{7{,}81}$.

Por ser $\chi^2 < \chi^2_\alpha$, aceptamos la hipótesis de que el crecimiento se ha hecho siguiendo las leyes de Mendel.

****

8.15 *En la tabla adjunta se reflejan las notas de BIOESTADISTICA y FISIOLOGIA de una muestra de 100 alumnos.*

		BIOESTADISTICA			
		Sobres.	Notable	Aprobad.	Suspens.
FISIOLOGIA	Sobres.	10	6	4	4
	Notable	6	4	8	6
	Aprobado	4	18	3	4
	Suspenso	6	2	6	9

Para un nivel de significación $\alpha = 0{,}01$, ¿son independientes las calificaciones obtenidas en ambas asignaturas?

SOLUCION:

Formulamos la hipótesis H_o de que las notas de FISIOLOGIA son independientes de las obtenidas en BIOESTADISTICA.

Según esta hipótesis:

- La probabilidad de obtener sobresaliente en FISIOLOGIA, es:
$$\frac{10 + 6 + 4 + 4}{100} = 0,24$$
- La probabilidad de obtener notable en FSIOLOGIA, es:
$$\frac{6 + 4 + 8 + 6}{100} = 0,24$$
- La probabilidad de obtener aprobado en FISIOLOGIA, es:
$$\frac{4 + 18 + 3 + 4}{100} = 0,29$$
- La probabilidad de obtener suspenso en FISIOLOGIA, es:
$$\frac{6 + 2 + 6 + 9}{100} = 0,23$$

Según estas probabilidades, las frecuencias teóricas serán las expresadas en la siguiente tabla:

		BIOESTADISTICA			
		Sobres.	Notable	Aprobad.	Suspens.
FISIOLOGIA	Sobresal.	6,24	7,2	5,04	5,52
	Notable	6,24	7,2	5,04	5,52
	Aprobado	7,54	8,7	6,09	6,67
	Suspenso	5,98	6,9	4,83	5,29

$$\chi^2 = \frac{(10-6,24)^2}{6,24} + \frac{(6-7,2)^2}{7,2} + \frac{(4-5,04)^2}{5,04} + \frac{(4-5,52)^2}{5,52} + \frac{(6-6,24)^2}{6,24} +$$
$$+ \frac{(4-7,2)^2}{7,2} + \frac{(8-5,04)^2}{5,04} + \frac{(6-5,52)^2}{5,52} + \frac{(4-7,54)^2}{7,54} + \frac{(18-8,7)^2}{8,7} +$$
$$+ \frac{(3-6,09)^2}{6,09} + \frac{(4-6,67)^2}{6,67} + \frac{(6-5,98)^2}{5,98} + \frac{(2-6,9)^2}{6,9} + \frac{(6-4,83)^2}{4,83} +$$
$$+ \frac{(9-5,29)^2}{5,29} = \mathbf{26,915}$$

El número de grados de libertad es: $\nu = (4-1).(4-1) = 9$.

Tomando $\alpha = 0,01$ se tiene que $\chi^2_\alpha = \mathbf{21,66}$.

Al ser $\chi^2 > \chi^2_\alpha$, rechazamos la hipótesis, es decir las calificaciones en FISIOLOGIA y BIOESTADISTICA no son independientes.

8.16 *Para curar una cierta enfermedad, se sabe que existe 5 tratamientos diferentes. Aplicados por separado, cada uno, a un grupo de enfermos que padecen esa enfermedad, se han observado los siguientes resultados:*

	CURADOS	NO CURADOS	TOTAL
TRATAMIENTO A	61	15	76
TRATAMIENTO B	50	14	64
TRATAMIENTO C	63	18	81
TRATAMIENTO D	66	23	89
TRATAMIENTO E	60	30	90
TOTAL	300	100	400

¿Se puede considerar que la eficacia de los 5 tratamientos es la misma, con un nivel de confianza del 95%?.

SOLUCION:

Formulamos la hipótesis H_0, de que la eficacia de los cinco tratamientos es la misma. En este caso, la mejor estimación del porcentaje teórico de curación será: $p = \frac{300}{400} = 0{,}75$.

El número teórico de enfermos curados con cada tratamiento son los que figuran en la siguiente tabla:

FRECUENCIAS TEORICAS

	CURADOS	NO CURADOS
TRATAMIENTO A	57	19
TRATAMIENTO B	48	16
TRATAMIENTO C	60,75	20,25
TRATAMIENTO D	66,75	22,25
TRATAMIENTO E	67,50	22,50

Luego el estadístico χ^2 tomará el siguiente valor:

$$\chi^2 = \frac{(61-57)^2}{57} + \frac{(50-48)^2}{48} + \frac{(63-60,75)^2}{60,75} + \frac{(66-66,75)^2}{66,75} + \frac{(60-67,50)^2}{67,50} +$$

$$= \frac{(15-19)^2}{19} + \frac{(14-16)^2}{16} + \frac{(18-20,25)^2}{20,25} + \frac{(23-22,25)^2}{22,25} + \frac{(30-22,50)^2}{22,50} =$$

$$= \mathbf{5,156}$$

El número de grados de libertad es: $\nu = 5-1 = 4$.

Tomando $\alpha = 0,05$, se tiene que $\chi^2_\alpha = \mathbf{9,49}$.

Al ser $\chi^2 < \chi^2_\alpha$, aceptaremos la hipótesis H_0, es decir, admitiremos que los cinco tratamientos sí tienen la misma eficacia.

****

9-análisis de la varianza

1. INTRODUCCION
2. ANALISIS DE LA VARIANZA CON UN FACTOR DE VARIACION
3. ANALISIS DE LA VARIANZA CON DOS FACTORES INDEPENDIENTES DE VARIACION

9- ANALISIS DE LA VARIANZA

1. INTRODUCCION

Dentro de los problemas de homogeneidad, y para el caso de caracteres cuantitativos, estudiaremos *el analisis de la varianza.*

Problemas típicos del análisis de la varianza, son:

- Se aplican 4 somníferos A,B,C,D a cuatro grupos de individuos y se toma como variable efecto los tiempos que tardan en dormirse. Nos preguntamos: ¿los efectos medios de los cuatro somníferos son iguales?

- Se siembran 5 tipos de maiz y se someten al mismo tratamiento de humedad, temperatura y composición del suelo, midiéndose a continuación los rendimientos. ¿Serán los rendimientos medios iguales?.

El comparar los efectos medios para dos sonníferos, o los rendimientos de dos tipos de maiz, ya lo hicimos mediante una t de Student en el capítulo anterior. Podriamos pensar, en comparar, por ejemplo, el somnífero A con el B, luego con el C, con el D, el B con el C, etc. Pero al hacer una gran número de comparaciones por separado, nos encontrariamos cierto número de diferencias significativas sin que fueran tales, sino debidas únicamente al azar en la toma de las muestras.

El análisis de la varianza es precisamente el método que nos pérmite superar los problemas de diferencias significativas cuando se hacen varias comparaciones y así mismo, nos permitirá contestar si existe diferencia entre los efectos o los rendimientos medios de los ejemplos anteriores.

2. ANALISIS DE LA VARIANZA CON UN FACTOR DE VARIACION

Representamos en una tabla el conjunto de observaciones x_{ij} correspondientes a k grupos, i representa el grupo y j el elemento dentro del grupo.

Es decir:

GRUPOS	OBSERVACIONES	TAMAÑO MUESTRA	MEDIAS MUESTR.	g.l	MEDIAS POBLAC.	VARIANZ. MUESTR.
1	$x_{11}, x_{12}, \ldots, x_{1n_1}$	n_1	$\overline{x}_1$	n_1-1	μ_1	s_1^2
2	$x_{21}, x_{22}, \ldots, x_{2n_2}$	n_2	$\overline{x}_2$	n_2-1	μ_2	s_2^2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$x_{k1}, x_{k2}, \ldots, x_{kn_k}$	n_k	$\overline{x}_k$	n_k-1	μ_k	s_k^2

Suponemos que las observaciones del grupo i proceden de una población normal $N(\mu_i, \sigma)$, en donde σ no depende de i, es decir, todas las poblaciones se suponen normales con igual varianza e independientes.

Vamos a contrastar la hipótesis H, de que todas las medias μ_i son iguales:

$$\mu_1 = \mu_2 = \ldots = \mu_k$$

Representaremos por N, el número total de observaciones:

$$N = n_1 + n_2 + \ldots + n_k$$

y por $\overline{X}$, la media de todas las observaciones.

El análisis de la varianza consiste en la obtención de dos estimaciones independientes de la varianza σ^2 común de las k poblaciones.

Una estimación de σ^2, viene dada por la varianza *dentro de los grupos o varianza residual*, que se obtiene calculando la media de las varianzas de cada grupo ponderadas por el número de grados de libertad de cada una de ellas.

$$s_1^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{(n_1-1) + (n_2-1) + \ldots + (n_k-1)}$$

s_1^2 es un estimador insesgado de σ^2, independiente de la hipótesis que se haga sobre las medias μ_i.

La otra estimación de σ^2, viene dada por la *varianza entre grupos:*

$$s_2^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{X})^2$$

que es un estimador insesgado de σ^2 si: $\mu_1 = \mu_2 = \ldots = \mu_k$

La significación de la diferencia entre estas dos estimaciones nos la proporciona el cociente: $\dfrac{s_2^2}{s_1^2}$

Este cociente, en caso de que las medias de las poblaciones sean iguales, es una F de Snedecor con (k-1 , N-k) grados de libertad.

Fijado un nivel de significación α , aceptaremos la hipótesis de igualdad de medias, si la razón:

$$F = \frac{s_2^2}{s_1^2}\text{ , es tal que,} \qquad F < F_{\alpha,\, k-1,\, N-k}$$

PREPARACION DE LOS CALCULOS.-

Para calcular de una manera ordenada el valor de F, es conveniente disponer las operaciones como se reflejan en el cuadro adjunto, el cual nos pérmite sistematizar y simplificar los cálculos.

GRUPOS	OBSERVACIONES	TAMAÑO MUESTRAS	S_i	SC_i	S_i^2/n_i
1	$x_{11}, x_{12}, \ldots, x_{1n_1}$	n_1	S_1	SC_1	S_1^2/n_1
2	$x_{21}, x_{22}, \ldots, x_{2n_2}$	n_2	S_2	SC_2	S_2^2/n_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
k	$x_{k1}, x_{k2}, \ldots, x_{kn_k}$	n_k	S_k	SC_k	S_k^2/n_k
		$\sum n_i = N$	$\sum S_i$	$\sum SC_i$	$\sum S_i^2/n_i$

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = \sum \frac{S_i^2}{n_i} - \frac{(\sum S_i)^2}{N}$	k-1	$s_2^2 = \frac{E}{k-1}$
DENTRO DE LOS GRUPOS	$D = T - E$	N-k	$s_1^2 = \frac{D}{N-k}$
TOTAL	$T = \sum SC_i - \frac{(\sum S_i)^2}{N}$	$F = s_2^2 \,/\, s_1^2$	

S_i es la suma de las observaciones del grupo i: $S_i = \sum_j x_{ij}$

$S\,C_i$ es la suma de los cuadrados de las observaciones del grupo i;

$$S\,C_i = \sum_j x_{ij}^2$$

S C D es la suma de los cuadrados de las desviaciones.

3. ANALISIS DE LA VARIANZA CON DOS FACTORES INDEPENDIENTES DE VARIACION

Supongamos que tenemos un número m de máquinas que fabrican cierto producto, partiendo de un número n de partidas de materia prima.

En cada unidad de producto fabricado, podemos considerar su terminación final. Cada unidad de producto x_{ij} procederá de una máquina i y de una partida de materia prima j, o sea, dependerá de dos factores. Manteniendo uno fijo y variando el otro, podremos representar los resultados en una tabla:

MAQUINAS \ PARTIDAS	1	2	. . .	n
1	x_{11}	x_{12}	. . .	x_{1n}
2	x_{21}	x_{22}	. . .	x_{2n}
.	.	.	. . .	.
.	.	.	. . .	.
.	.	.	. . .	.
m	x_{m1}	x_{m2}	. . .	x_{mn}

La fila i, corresponde a los productos de la máquina i y cada una de las partidas de materia prima.

La columna j, corresponde a los productos de la materia prima j fabricados por cada una de las máquinas.

En total tenemos m.n observaciones x_{ij}.

Suponemos que las variables x_{ij} son independientes y siguen distribuciones $N(\mu_{ij}, \sigma)$.

En estas condiciones, cabe preguntarse si la terminación final media del producto es la misma para las m máquinas y si difiere dicha terminación con la utilización de las diferentes partidas de materia prima.

Consiste este problema en contrastar simultaneamente las dos hipótesis:

H_1: de igualdad de medias de las filas.

H_2: de igualdad de medias de las columnas.

En general, si tenemos dos factores A y B independientes que admiten las modalidades $A_1, A_2, A_3, \ldots, A_n$ y $B_1, B_2, B_3, \ldots, B_m$ respectivamente, podremos ordenar los datos del siguiente modo:

B \ A	A_1	A_2	. . .	A_n	MEDIAS MUESTRALES	MEDIAS POBLACIONALES
B_1	x_{11}	x_{12}		x_{1n}	f_1	μ_{B_1}
B_2	x_{21}	x_{22}	. . .	x_{2n}	f_2	μ_{B_2}
.	.	.	. . .	.	.	.
.	.	.	. . .	.	.	.
.	.	.	. . .	.	.	.
B_m	x_{m1}	x_{m2}	. . .	x_{mn}	f_m	μ_{B_m}
MEDIAS MUESTRALES	c_1	c_2	. . .	c_n	$\overline{X}$	
MEDIAS POBLACIONALES	μ_{A_1}	μ_{A_2}	. . .	μ_{A_n}		

Tratamos de contrastar:

$$\left.\begin{array}{ll} H_1 : & \mu_{B_1} = \mu_{B_2} = \ldots = \mu_{B_m} \\ H_2 : & \mu_{A_1} = \mu_{A_2} = \ldots = \mu_{A_n} \end{array}\right\}$$

Hemos representado por $\overline{X}$ la media total de todas las observaciones y por f_i y c_j las medias de filas y columnas respectivamente.

El análisis de la varianza, consiste en la obtención de cuatro estimaciones de la varianza σ^2 común de las poblaciones.

★ Varianza total: $$s^2 = \frac{\sum_i \sum_j (x_{ij} - \overline{X})^2}{m.n-1}$$

★ Varianza entre filas: $$s_f^2 = \frac{n \sum_i (f_i - \overline{X})^2}{m-1}$$

Esta varianza entre filas, que representamos por s_f^2, es un estimador insesgado de σ^2 si: $\mu_{B_1} = \mu_{B_2} = \ldots = \mu_{B_m}$

★ Varianza entre columnas: $$s_c^2 = \frac{m \sum_j (c_j - \overline{X})^2}{n-1}$$

Esta varianza entre columnas, que representamos por s_c^2, es un estimador insesgado de σ^2 si: $\mu_{A_1} = \mu_{A_2} = \ldots = \mu_{A_n}$

★ Varianza residual: $$s_r^2 = \frac{\sum_i \sum_j d_{ij}^2}{(m-1)(n-1)}$$

siendo $d_{ij} = x_{ij} - f_i - c_j + \overline{X}$

Esta varianza residual, es la varianza una vez eliminado el efecto de filas y columnas, y es también un estimador insesgado de σ^2.

La significación de la diferencia entre estas estimaciones, nos la proporcionan los siguientes cocientes:

a) $F_f = \dfrac{s_f^2}{s_r^2}$, que es una F de Snedecor con (m-1) y (m-1)(n-1) grados de libertad, si se cumple: $\mu_{B_1} = \mu_{B_2} = \ldots = \mu_{B_m}$

b) $F_c = \dfrac{s_c^2}{s_r^2}$, que es una F de Snedecor con (n-1) y (m-1)(n-1) grados de libertad, si se cumple: $\mu_{A_1} = \mu_{A_2} = \ldots = \mu_{A_n}$

Fijado un nivel de significación α, aceptaremos la hipótesis H_1 si se verifica:

$$F_f < F_{\alpha,\ (m-1),\ (m-1)(n-1)}$$

Aceptaremos la hipótesis H_2 si se verifica:

$$F_c < F_{\alpha,\ (n-1),\ (m-1)(n-1)}$$

<u>PREPARACION DE LOS CALCULOS:</u> Con el fin de simplificar y sistematizar los cálculos dispondremos las operaciones como se especifica en el cuadro adjunto:

B \ A	A_1	A_2	. . .	A_n	F_i	f_i	$F_i.f_i$	$\sum_j x_{ij}^2$
B_1	x_{11}	x_{12}	. . .	x_{1n}	F_1	f_1	$F_1.f_1$	$\sum_j x_{1j}^2$
B_2	x_{21}	x_{22}	. . .	x_{2n}	F_2	f_2	$F_2.f_2$	$\sum_j x_{2j}^2$
⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮
B_m	x_{m1}	x_{m2}	. . .	x_{mn}	F_m	f_m	$F_m.f_m$	$\sum_j x_{mj}^2$
C_j	C_1	C_2	. . .	C_n	X		$\sum_i F_i.f_i$	$\sum_i \sum_j x_{ij}^2$
c_j	c_1	c_2	. . .	c_n				
$C_j.c_j$	$C_1.c_1$	$C_2.c_2$	. . .	$C_n.c_n$	$\sum_j C_j.c_j$			

Donde:

F_i es la suma de las observaciones de la fila i.

f_i es la media de las observaciones de la fila i.

C_j es la suma de las observaciones de la columna j.

c_j es la media de las observaciones de la columna j.

X es la suma de todas las observaciones.

N es el número total de observaciones.

$\overline{X}$ es la media de todas las observaciones.

Las expresiones anteriores de los valores de F_c y F_f se transforman en las siguientes:

$$F_f = \frac{\frac{1}{m-1}\left(\sum_i F_i . f_i - X.\overline{X}\right)}{\frac{1}{(m-1)(n-1)}\left(\sum_i \sum_j x_{ij}^2 - \sum_j C_j . c_j - \sum_i F_i f_i + X.\overline{X}\right)}$$

$$F_c = \frac{\frac{1}{n-1}\left(\sum_j C_j . c_j - X.\overline{X}\right)}{\frac{1}{(m-1)(n-1)}\left(\sum_i \sum_j x_{ij}^2 - \sum_j C_j . c_j - \sum_i F_i f_i + X.\overline{X}\right)}$$

****

PROBLEMAS RESUELTOS

9.1 *Se les midió la tensión arterial a tres grupos A, B, y C de animales de padres hipertensos, normales e hipotensos, con el fin de ver si la tensión arterial era un factor hereditario.*

Los resultados fueron:

GRUPOS	PRESION ARTERIAL
A	94, 89, 87, 99, 97, 98.
B	81, 86, 92, 91, 94, 84.
C	92, 78, 82, 88, 85, 85.

Suponiendo que las tres poblaciones siguen la ley normal, son independientes y de igual varianza, contrástese la hipótesis de igualdad de medias al nivel de significación del 5%.

SOLUCION:

Para el cálculo del valor del estadístico F del contraste, dispondremos las operaciones como se indica en el cuadro adjunto. Para simplificar los cálculos restaremos 90 unidades a los valores de la presión arterial observadas.

GRUPOS	OBSERVACIONES	n_i	S_i	SC_i	S_i^2/n_i
A	4, -1, -3, 9, 7, 8	6	24	220	96
B	-9, -4, 2, 1, 4, -6	6	-12	154	24
C	2, -12, -8, -2, -5, -5	6	-30	266	150
T O T A L E S		18	-18	640	270

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = 270 - \frac{(-18)^2}{18} = 252$	2	$s_2^2 = 126$
DENTRO DE LOS GRUPOS	$D = 370$	15	$s_1^2 = 24,66$
	$T = 640 - \frac{(-18)^2}{18} = 622$		

$$F = \frac{s_2^2}{s_1^2} = \frac{126}{24,66} = 5,108$$

Para $\alpha = 0,05$ se tiene en las tablas de la F de Snedecor que:

$$F_{0,05;\ 2,\ 15} = 3,6823$$

Al ser $F > F_{0,05;\ 2,\ 15}$ se rechaza la hipótesis de la igualdad de medias de tensiones arteriales.

****

9.2 *Se quiere conocer la resistencia de tres razas de moscas a un determinado insecticida. Se elige como variable de respuesta la tasa de mortalidad expresada en tanto por ciento. Para cada raza de moscas realizamos 8 determinaciones del tanto por ciento de mortalidad. Aplicado el insecticida los resultados fueron:*

R A Z A S

		A	B	C
	1	60	58	61
	2	55	60	44
	3	52	50	48
DETERMINACIONES	4	40	48	60
	5	62	61	58
	6	38	45	46
	7	46	49	52
	8	50	51	53

¿Que conclusión puede sacarse y bajo qué supuestos?.

S O L U C I O N:

Suponemos:

- que la variable, tasa de mortalidad, sigue la ley normal en las tres razas de moscas;
- que las varianzas de las tres poblaciones son iguales;
- que las muestras han sido obtenidas aleatoriamente.

Vamos a contrastar la hipótesis H, de que las medias de tasas de mortalidad son iguales para las tres poblaciones.

$$\mu_1 = \mu_2 = \mu_3.$$

Para el cálculo práctico del valor de F, restamos al valor de cada observación 50.

GRUPOS	OBSERVACIONES	n_i	S_i	SC_i	S_i^2/n_i
A	10, 5, 2, -10, 12, -12, -4, 0	8	3	533	1,125
B	8, 10, 0, -2, 11, -5 , -1 1	8	22	316	60,5
C	11, -6, -2, 10, 8, -4, 2 3	8	22	354	60,5
T O T A L E S		24	47	1203	122,125

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = 122,125 - \frac{47^2}{24} = 30,083$	2	$s_2^2=15,0415$
DENTRO DE LOS GRUPOS	$D = 1080,875$	21	$s_1^2=51,4702$
	$T = 1203 - \frac{47^2}{24} = 1110,958$		

$$F = \frac{S_2^2}{S_1^2} = \frac{15,0415}{51,4702} = \mathbf{0,292}$$

Tomando $\alpha = 0,05$ se tiene que: $F_{0,05;\ 2,\ 21} = \mathbf{3,4668}$.

Al ser $F < F_{0,05;\ 2,\ 21}$ aceptamos la hipótesis H, es decir, aceptamos que la tasa media de mortalidad es la misma para los tres tipos de mosca a un nivel de significación del 5%.

****

9.3 *Un fumador muy preocupado por el contenido medio de nicotina de los cigarrillos que habitualmente fuma, y que corresponden a tres marcas distintas de las existentes en el mercado, decide realizar una serie de análisis en cigarrillos tomados al azar. Los contenidos de nicotina expresados en miligramos fueron:*

MARCAS	CONTENIDO NICOTINA EN MILIGRAMOS
A	28, 20, 26, 30, 32, 34, 25, 23
B	29, 31, 28, 30, 24, 26, 25, 25, 28, 28
C	32, 30, 28, 30, 25, 27, 30, 26, 24, 22

¿Que conclusión se puede sacar a la vista de estos resultados, si suponemos que el contenido de nicotina sigue la ley normal y que la varianza es la misma para las tres poblaciones?

SOLUCION:

Vamos a cambiar el origen de las observaciones, pues el valor de F permanece invariante, para ello restaremos 30 unidades al valor de cada observación.

GRUPOS	OBSERVACIONES	g.l	S_i	SC_i	S_i^2/n_i
A	-2, -10, -4, 0, 2, 4, -5, -7	8	-22	214	60,5
B	-1, 1, -2, 0, -6, -4, -5, -5, -2, -2	10	-26	116	67,6
C	2, 0, -2, 0, -5, -3, 0, -4, -6, -8	10	-26	158	67,6
TOTALES		28	-74	488	195,7

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = 195{,}7 - \frac{(-74)^2}{28} = 0{,}12858$	2	$S_2^2 = 0{,}06429$
DENTRO DE LOS GRUPOS	$D = 292{,}3$	25	$S_1^2 = 11{,}692$
	$T = 488 - \frac{(-74)^2}{28} = 292{,}42858$		

Entonces, el valor del estadístico F será:

$$F = \frac{0{,}06429}{11{,}692} = 0{,}00549$$

Fijamos arbitrariamente un valor $\alpha = 0{,}05$, y la región crítica vendrá dada por los valores de F: $F > F_{0{,}05;\ 2,\ 25} = 3{,}3852$.

Por ser el valor calculado $F < F_{0{,}05;\ 2,\ 25}$ obtenemos:

que si el contenido medio de nicotina no fuera el mismo para las tres marcas de cigarrillos, los resultados del experimento no se lo prueban suficientemente al fumador.

****

9.4 *Un Profesor Ayudante de clases prácticas de la asignatura de Bioestadística, imparte clase a cuatro grupos que supondremos formados por alumnos escogidos aleatoriamente.*

Para ver el mismo programa, utiliza para cada grupo un metodo distinto, de la siguiente forma:

- *grupo A : El Profesor explica la parte teórica y hace él mismo los problemas en la pizarra.*
- *grupo B : El Profesor explica la parte teórica pero los problemas los hacen los alumnos en la pizarra.*
- *grupo C : Los alumnos exponen la parte teórica en la pizarra y el Profesor los problemas.*
- *grupo D : El Profesor solo hace de guía, de modo que la teoría y los problemas los exponen los alumnos en la pizarra.*

Al final de curso, los cuatro grupos realizan el mismo examen obteniendo las calificaciones de la tabla.

¿Qué conclusiones pueden sacarse?

GRUPOS	calificaciones
A	8, 7, 7, 6, 3, 8, 2, 5, 5, 4, 5, 10
B	6, 4, 3, 9, 8, 7, 3, 2, 6, 5, 5, 5
C	9, 10, 8, 5, 2, 3, 6, 9, 10, 4, 2, 0
D	8, 7, 6, 4, 3, 5, 6, 4, 3, 5, 9, 8

S O L U C I O N:

Vamos a contrastar la hipótesis H de que las notas medias son iguales para los cuatro métodos de enseñanza.

A cada nota vamos a restarle 5 puntos.

GRUPOS	OBSERVACIONES	g.l	S_i	SC_i	S_i^2/n_i
A	3, 2, 2, 1, -2, 3, -3, 0, 0, -1, 0, 5	12	10	66	8,33
B	1, -1, -2, 4, 3, 2, -2, -3, 1, 0, 0, 0	12	3	49	0,75
C	4, 5, 3, 0, -3, -2, 1, 4, 5, -1, -3, -5	12	8	140	5,33
D	3, 2, 1, -1, -2, 0, 1, -1, -2, 0, 4, 3	12	8	50	5,33
T O T A L E S		48	29	305	19,74

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = 19{,}74 - \dfrac{29^2}{48} = 2{,}225$	3	$S_2^2 = 0{,}7416$
DENTRO DE LOS GRUPOS	$D = 285{,}254$	44	$S_1^2 = 6{,}4830$
	$T = 305 - \dfrac{29^2}{48} = 287{,}479$		

$$F = \frac{0{,}7416}{6{,}4830} = \mathbf{0{,}114}$$

Tomando $\alpha = 0{,}05$ se tiene que $F_{0,05;\ 3,\ 44} = \mathbf{2{,}8387}$.

Al ser $F < F_{0,05;\ 3,\ 44}$ aceptaremos la hipótesis H.

Por tanto, aceptaremos que las notas medias son iguales para los cuatro métodos de enseñanza. Si las notas medias no son iguales, las calificaciones del examen final no nos proporcionan suficiente información.

****

9.5 *Realizamos un experimento para determinar la cantidad media de droga que se deposita en las vísceras de tres tipos de animales de raza diferente. Tomamos tres muestras aleatorias de 10 animales de cada raza y les suministramos la citada droga, sacrificándolos después. Los contenidos de droga en las vísceras son los siguientes:*

RAZAS	CONTENIDO DE DROGA
A	69, 72, 71, 72, 72, 73, 71, 68, 69, 75
B	71, 75, 72, 72, 74, 76, 73, 70, 75, 76
C	71, 74, 72, 70, 75, 74, 73, 68, 73, 74

Suponemos que la variable de respuesta elegida es normal en cada una de las tres poblaciones y con igual varianza.

¿Justifican los resultados anteriores que el contenido medio de droga en las vísceras es el mismo en las tres poblaciones?

S O L U C I O N:

Vamos a contrastar la hipótesis H de que el contenido medio de droga en las vísceras es la misma para las tres poblaciones.

Para el cálculo práctico restaremos 70 unidades a cada observación.

GRUPOS	OBSERVACIONES	g.l	S_i	SC_i	S_i^2/n_i
A	-1, 2, 1, 2, 2, 3, 1,-2,-1, 5	10	12	54	14,4
B	1, 5, 2, 2, 4, 6, 3, 0, 5, 6	10	34	156	115,6
C	1, 4, 2, 0, 5, 4, 3,-2, 3, 4	10	24	100	57,6
T O T A L E S		30	70	310	187,6

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = 187{,}6 - \frac{70^2}{30} = 24{,}266$	2	$s_2^2 = 12{,}133$
DENTRO DE LOS GRUPOS	$D = 122{,}40$	27	$s_1^2 = 4{,}5333$
	$T = 310 - \frac{70^2}{30} = 146{,}666$		

$$F = \frac{12{,}133}{4{,}5333} = \mathbf{2{,}6763}$$

Para $\alpha = 0{,}05$, se tiene que: $F_{0{,}05,\ 2,\ 27} = \mathbf{3{,}3541}$.

Al ser $F < F_{0{,}05;\ 2,\ 27}$ aceptaremos la hipótesis H, es decir el contenido medio de droga en las vísceras es la misma para las tres poblaciones.

****

9.6 *Se consideran cuatro medicamentos todos ellos indicados para reducir el contenido de colesterol en la sangre. Queriendo contrastar si existe diferencia significativa entre los cuatro medicamentos citados, se suministra cada uno de ellos a un grupo de enfermos comparables.*

Se mide en mg/100ml la reducción de colesterol habida. Suponiendo que esta variable se distribuye normalmente y con igual varianza en cada una de las poblaciones, sacar las conclusiones pertinentes.

MEDICAMENTO	REDUCCION DEL CONTENIDO DE COLESTER
1	8, 12, 11, 6. 10, 2, 8, 7, 9, 11
2	4, 6, 8, 8, 7, 3, 6, 9, 4, 6
3	5, 5, 6, 7, 5, 8, 6, 2
4	0, 3, 5, 2, 1, 0, 4, 2
5	3, 4, 3, 8, 6, 10, 4, 8

S O L U C I O N:

★ Formulamos la hipótesis H, de que no existe diferencia en la reducción media de colesterol para los cuatro medicamentos.

★ Disponemos los cálculos en el cuadro adjunto:

GRUPOS	OBSERVACIONES	n_i	S_i	SC_i	S_i^2/n_i
1	8,12,11, 6,10, 2, 8, 7, 9,11	10	84	784	705,6
2	4, 6, 8, 8, 7, 3, 6, 9, 4, 6	10	61	407	372,1
3	5, 5, 6, 7, 5, 8, 6, 2	8	44	264	242,-
4	0, 3, 5, 2, 1, 0, 4, 2	8	17	59	36,125
5	3, 4, 3, 8, 6,10, 4, 8	8	46	314	264,5
T O T A L E S		44	252	1828	1620,325

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = 1620,325 - \frac{252^2}{44} = 177,0523$	4	$s_2^2 = 44,263$
DENTRO DE LOS GRUPOS	$D = 207,675$	39	$s_1^2 = 5,325$
	$T = 1828 - \frac{252^2}{44} = 384,7273$		

$$F = \frac{s_2^2}{s_1^2} = \mathbf{8,3123.}$$

★ Para $\alpha = 0,05$, resulta $F_{0,05;\ 4,\ 39} = \mathbf{2,6060}$.

★ Al ser $F > F_{0,05;\ 4,\ 39}$ rechazamos la hipótesis H, es decir, que las observaciones nos prueban que la reducción media de colesterol **no** es la misma para los cuatro medicamentos.

****

9.7 *Un médico quiere contrastar si existe diferencia significativa entre tres tipos de somníferos de fuerte potencia. A tal fin suministra la droga a pacientes comparables. Se registraron los tiempos en segundos, transcurridos desde la administración de la droga hasta que el paciente se duerme. Los resultados fueron:*

SOMNIFERO A	SOMNIFERO B	SOMNIFERO C
222	326	263
300	275	260
262	218	299
264	207	221
200	272	198
211	268	211
267	308	266
326	227	319

Contrástese la hipótesis de igualdad del efecto medio de los tres somníferos, con un nivel de significación $\alpha = 0,05$.

SOLUCION:

Con el fin de facilitar los cálculos vamos a restar 250 segundos a las observaciones. Disponiendo los cálculos en el cuadro adjunto:

GRUPOS	OBSERVACIONES	n_i	S_i	SC_i	S_i^2/n_i
A	-28, 50, 12, 14,-50,-39, 17, 76	8	52	13710	338
B	76, 25,-32,-43, 22, 18, 58,-23	8	101	13975	1275,125
C	13, 10, 49,-29,-52,-39, 16, 69	8	37	12753	171,125
TOTALES		24	190	40438	1784,25

VARIACION	S C D	g.l	
ENTRE GRUPOS	$E = 1784,25 - \frac{190^2}{24} = 280,0834$	2	$s_2^2 = 140,0417$
DENTRO DE LOS GRUPOS	$D = 38653,749$	21	$s_1^2 = 1840,6547$
	$T = 40438 - \frac{190^2}{24} = 38933,833$		

$$F = \frac{140,0417}{1840,6547} = \mathbf{0,076}$$

Para $\alpha = 0,05$ se tiene que: $F_{0,05;\ 2,\ 21} = \mathbf{3,4668}$.

Al ser $F < F_{0,05;\ 2,\ 21}$ aceptaremos la hipótesis H, es decir, admitimos que el efecto medio de los tres somniferos es el mismo.

****

9.8 *La siguiente tabla representa los pesos ganados en gramos por cuatro razas de conejos al suministrarles cuatro tipos de tratamientos a base de complejos vitamínicos.*

RAZAS / TRATAMIENTOS	A	B	C	D
1	94	89	87	98
2	92	86	86	92
3	92	86	84	87
4	91	92	83	88

Contrástese la hipótesis de igualdad de medias de pesos ganados para los cuatro tipos de tratamientos, y de igualdad de medias de pesos ganados para las cuatro razas de conejos, al nivel de significación del 5%.

S O L U C I O N:

Suponemos que las variables x_{ij}, peso ganado por el tratamiento i y la raza de conejo j, son independientes y normales con igual varianza todas ellas.

Contrastaremos las hipótesis:

H_1: de igualdad de medias de pesos ganados para los cuatro tipos de tratamientos.

H_2: de igualdad de medias de pesos ganados para las cuatro razas de conejos.

Para sistematizar y simplificar los cálculos de los estadísticos F_f y F_c del contraste, dispondremos las operaciones como se especifica en el cuadro adjunto.

Restaremos a cada observación 90 gramos ya que este cambio de origen de las observaciones no afecta a los valores F_f y F_c.

RAZAS / TRATAMIENTOS	A	B	C	D	F_i	f_i	$F_i . f_i$	$\sum_j x_{ij}^2$
1	4	-1	-3	8	8	2	16	90
2	2	-4	-4	2	-4	-1	4	40
3	2	-4	-6	-3	-11	-2,75	30,25	65
4	1	2	-7	-2	-6	-1,5	9	59
C_j	9	-7	-20	5	-13		59,25	254
c_j	2,25	-1,75	-5	1,25				
$C_j . c_j$	20,25	12,25	100	6,25	138,75			

Entonces: $\overline{X} = -\dfrac{13}{16}$.

$$F_f = \frac{\frac{1}{4-1}\left[59,25 - (-13)\left(\frac{-13}{16}\right)\right]}{\frac{1}{4-1}\,\frac{1}{4-1}\left[254-138,75-59,25+(-13)\left(\frac{-13}{16}\right)\right]} = \frac{16,229166}{7,395833} = \mathbf{2,194}$$

$$F_c = \frac{\frac{1}{4-1}\left[138,75 - (-13)\left(\frac{-13}{16}\right)\right]}{\frac{1}{4-1}\,\frac{1}{4-1}\left[254-138,75-59,25+(-13)\left(\frac{-13}{16}\right)\right]} = \frac{42,729166}{7,395833} = \mathbf{5,777}$$

Tomando $\alpha = 0,05$ tenemos que: $F_{0,05;\,3,\,9} = \mathbf{3,8626}$

- Como $F_f < F_{0,05;\,3,\,9}$ aceptaremos H_1, de igualdad de medias de pesos ganados para los cuatro tratamientos.

- Como $F_c > F_{0,05;\,3,9}$ rechazaremos H_2, de igualdad de pesos medios ganados para las cuatro razas de conejos.

****

9.9 *Se desea contrastar la hipótesis de igualdad de rendimientos medios para 3 tipos de terrenos diferentes simultaneamente con la hipótesis de igualdad de rendimientos medios para cinco variedades de* *siembra de maiz. Se toman al azar 5 areas de cada uno de los terrenos y se siembra cada area con un tipo de maiz. Cuando se recoge el maiz, se anotan los rendimientos, que son los siguientes expresados en Kgrs. por area:*

VARIEDAD DE MAIZ / TIPO DE TERRENO	A	B	C	D	E
1	280	300	310	270	330
2	250	240	260	270	240
3	310	304	290	280	300

S O L U C I O N:

Suponemos que las variables x_{ij}, rendimientos para el tipo de terreno i y la variedad de maiz j, son independientes y $N(\mu_{ij}, \sigma)$.

Vamos a contrastar simultaneamente las hipótesis:

H_1: de igualdad de rendimientos medios para los tres tipos de terreno.

H_2: de igualdad de rendimientos medios para las cinco variedades de maiz.

Con el fin de simplificar, restaremos a cada observacion 300.

VARIEDAD DE MAIZ / TIPO DE TERRENO	A	B	C	D	E	F_i	f_i	$F_i.f_i$	$\sum_j x_{ij}^2$
1	-20	0	10	-30	30	-10	-2	20	2300
2	-50	-60	-40	-30	-60	-240	-48	11520	12200
3	10	4	-10	-20	0	-16	-3,2	51,2	616
C_j	-60	-56	-40	-80	-30	-266		11591,2	15116
c_j	-20	-18,66	-13,33	-26,66	-10				
$C_j.c_j$	1200	1045,33	533,33	2133,33	300	5211,99			

Donde: $\overline{X} = -\dfrac{266}{15}$

$$F_f = \frac{\frac{1}{3-1}\left[11591,2 - (-266)\left(\frac{-266}{15}\right)\right]}{\frac{1}{3-1}\,\frac{1}{5-1}\left[15116 - 5211,99 - 11591,2 + (-266)\left(\frac{-266}{15}\right)\right]} =$$

$$= \frac{3437,0667}{378,73325} = 9,075$$

$$F_c = \frac{\frac{1}{5-1}\left[5211,99 - (-266)\left(\frac{-266}{15}\right)\right]}{\frac{1}{3-1}\,\frac{1}{5-1}\left[15116 - 5211,99 - 11591,2 + (-266)\left(\frac{-266}{15}\right)\right]} =$$

$$= \frac{123,73}{378,73325} = \mathbf{0,326}$$

Fijando $\alpha = 0,05$ se tiene que $F_{0,05,\ 2,\ 8} = \mathbf{4,459}$ y

$F_{0,05,\ 4,\ 8} = \mathbf{3,8378}$.

- Al ser $F_f > F_{0,05;\ 2,\ 8}$ rechazamos la hipótesis H_1, de igualdad de rendimientos medios para los tres tipos de terrenos.

- Al ser $F_c < F_{0,05;\ 4,\ 8}$ aceptamos la hipótesis H_2, de igualdad de rendimientos medios para los cinco tipos de maiz.

****

9.10 *Una fábrica dedicada a la producción de termómetros clínicos, utiliza cuatro máquinas de diferente marca, que son manejadas cada una por un operario de los cuatro que tiene la fábrica.*

Los números de unidades producidas diariamente en centenas, por operario y máquina, son los siguientes:

OPERARIO \ MAQUINA	A	B	C	D
1	14	9	7	8
2	12	11	10	9
3	16	8	8	11
4	14	8	6	10

Realizar el análisis de la varianza correspondiente, especificando supuestos y conclusiones.

SOLUCION:

Suponemos:

- Que las variables x_{ij}, número de centenas de termómetros fabricados en un día por el operario i, con la máquina j, son independientes.
  Los operarios manejan con la misma habilidad cada una de las cuatro marcas de máquinas.
- Que las variables x_{ij} son $N(\mu_{ij}, \sigma)$.

Bajo estos supuestos, resolver este problema por el método de análisis de la varianza con dos factores independientes de variación, consiste en contrastar simultaneamente las siguientes hipótesis:

H_1: de igualdad de fabricación media para cada uno de los cuatro operarios.

H_2: de igualdad de fabricación media para cada marca de máquina.

Para simplificar los cálculos, restaremos a cada observación 10 unidades.

MAQUINA / OPERARIO	A	B	C	D	F_i	f_i	$F_i.f_i$	$\sum_j x_{ij}^2$
1	4	-1	-3	-2	-2	-0,5	1	30
2	2	1	0	-1	2	0,5	1	6
3	6	-2	-2	1	3	0,75	2,25	45
4	4	-2	-4	0	-2	-0,5	1	36
C_j	16	-4	-9	-2	1		5,25	117
c_j	4	-1	-2,25	-0,5				
$C_j.c_j$	64	4	20,25	1	89,25			

$$\bar{x} = \frac{1}{16}$$

$$F_f = \frac{\frac{1}{4-1}\left[5,25 - 1.\frac{1}{16}\right]}{\frac{1}{4-1}\,\frac{1}{4-1}\left[117 - 89,25 - 5,25 + 1.\frac{1}{16}\right]} = \mathbf{0,6897.}$$

tomado el nivel de significación $\alpha = 0,05$, se tiene que:

$F_{0,05;\ 3,\ 9} = \mathbf{3,8626}$. Como $F_f < F_{0,05;\ 3,\ 9}$, aceptamos la hipótesis H_1, de igualdad de medias de centenas fabricadas por cada uno de los cuatro operarios.

$$F_c = \frac{\frac{1}{4-1}\left[89,25 - 1.\frac{1}{16}\right]}{\frac{1}{4-1}\,\frac{1}{4-1}\left[117 - 89,25 - 5,25 + 1.\frac{1}{16}\right]} = \mathbf{11,8587.}$$

tomado el nivel de significación $\alpha = 0,05$, se tiene que:

$F_{0,05;\ 3,\ 9} = \mathbf{3,8626}$. Como $F_c > F_{0,05;\ 3,\ 9}$ rechazaremos la hipótesis H_2.

Podemos concluir diciendo que las marcas de las máquinas influyen significativamente en la producción.

****

9.11 *Se quieren comparar tres complejos vitamínicos; para ello se reunen 4 conjuntos de trillizos de un año de edad. Cada niño de una familia dada, recibió al azar, uno de los tres regímenes vitamínicos durante dos años consecutivos. Suponiendo que la variable de respuesta era el crecimiento y los datos expresados en décimas de kgrs. los siguientes:*

		COMPLEJO VITAMINICO		
		A	B	C
FAMILIA	1	68	70	72
	2	74	71	69
	3	66	67	73
	4	70	68	68

realícese el análisis de la varianza consiguiente.

SOLUCION:

Suponemos que las variables x_{ij} son independientes y $N(\mu_{ij}, \sigma)$. Contrastaremos simultaneamente las hipótesis:

H_1: de igualdad de crecimiento medio para cada familia.

H_2: de igualdad de crecimiento medio para cada complejo vitamínico.

Para facilitar los cálculos restaremos 70 decimas de kilogramo a cada observación.

COMPLEJO VITAMINICO / FAMILIA	A	B	C	F_i	f_i	$F_i \cdot f_i$	$\sum_j x_{ij}^2$
1	-2	O	2	O	O	O	8
2	4	1	-1	4	1,33	5,33	18
3	-4	-3	3	-4	-1,33	5,33	34
4	O	-2	-2	-4	-1,33	5,33	8
C_j	-2	-4	2	-4		15,99	68
c_j	-0,5	-1	0,5				
$C_j \cdot c_j$	1	4	1	6			

$$\bar{x} = \frac{-4}{12}$$

$$F_f = \frac{\frac{1}{4-1}\left[15{,}99 - (-4)\left(\frac{-4}{12}\right)\right]}{\frac{1}{4-1}\,\frac{1}{3-1}\left[68 - 6 - 15{,}99 + (-4)\left(\frac{-4}{12}\right)\right]} = \mathbf{0{,}6197.}$$

Fijamos un nivel de significación igual a 0,05, entonces tenemos que: $F_{0,05,\ 3,\ 6} = \mathbf{4{,}7571.}$

Como $F_f < F_{0,05;\ 3,\ 6}$ aceptaremos la hipótesis H_1.

$$F_c = \frac{\frac{1}{3-1}\left[6 - (-4)\left(\frac{-4}{12}\right)\right]}{\frac{1}{4-1}\,\frac{1}{3-1}\left[68 - 6 - 15{,}99 + (-4)\left(\frac{-4}{12}\right)\right]} = \mathbf{0{,}2957.}$$

Si tomamos $\alpha = 0{,}05$, se tiene que $F_{0{,}05;\,2,\,6} = 5{,}1433$.

Al ser $F_c < F_{0{,}05,\,2,\,6}$ aceptaremos la hipótesis H_2.

****

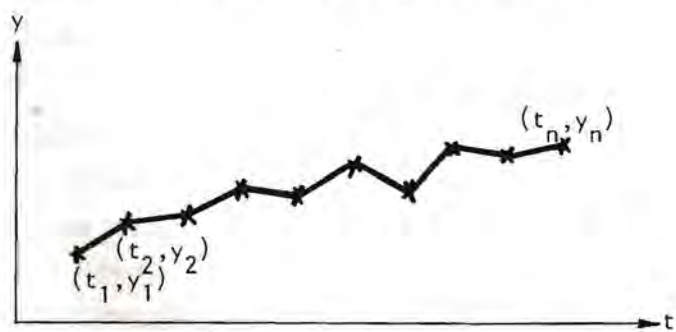
10- series cronológicas

10-SERIES CRONOLOGICAS

1. CONCEPTO

Una *serie cronológica* se define como un conjunto de observaciones cuantitativas de un fenómeno, referidas a un intervalo de tiempo tomado como unidad.

Los valores $y_1, y_2, \ldots, y_n$ que toma un fenómeno en los momentos t_1 $t_2, \ldots t_n$ constituyen una serie cronológica. Suelen representarse en un sitema de ejes cartesianos, figurando en el eje de abcisas los tiempos y en el eje de ordenadas la magnitud del fenómeno.



Ejemplos de series cronológicas son: el número de niños nacidos vivos en España por mes, durante un cierto número de meses o la tasa de mortalidad anual durante un número de años, y en general cualquier magnitud ligada a un momento de tiempo.

2. ANALISIS DE UNA SERIE CRONOLOGICA

En el estudio de una serie cronológica nos limitaremos a determinar las variaciones en el periodo total de tiempo para el que se dispone de observaciones, o también predeciremos su comportamiento en el futuro.

Las variaciones de la serie para el periodo en que se dispone de observaciones don debidas a multitud de causas superpuestas, entre las que consideraremos:

2.1 TENDENCIA SECULAR.- Se llama así, a la tendencia durante un periodo de tiempo largo, es decir, se trata de un movimiento suave, regular y a largo plazo. Las variaciones súbitas y frecuentes tanto en magnitud absoluta como en rapidez de aumento o disminución, no son compatibles con la idea de tendencia secular. Esta tendencia indica la marcha general y persistente del fenómeno si se elimaran los efectos de fuerzas accidentales hasta quedar tan solo el efecto del crecimiento normal.

2.2 FLUCTUACIONES PERIODICAS.- La tendencia general de la serie viene determinada por la superposición a la tendencia secular de numerosas fluctuaciones que pueden ser regulares o irregulares, debiles o intensas, debidas a fuerzas de carácter periódico o accidental. Entre las de carácter periódico tenemos:

2.2.1 VARIACIONES ESTACIONALES.- Son oscilaciones a corto plazo y que se repiten con carácter periódico.El periodo de repetición más general suele ser de un año y la variación suele deberse a las estaciones, de ahí el nombre.
A veces, el periodo de repetición puede ser un mes, una semana e incluso un día.

2.2.2 FLUCTUACIONES CICLICAS.- Son variaciones con un periodo menos definido que las anteriores, pero que sin embargo presentan cierto grado de regularidad. Incluso la amplitud de los periodos podrá variar aunque tendrá que ser de una forma regular para ser consideradas como tales. Las fluctuaciones cíclicas suelen corresponder a periodos de intensificación y de depresión del fenómeno.

2.3 VARIACIONES ACCIDENTALES.- Son fluctuaciones fortuitas e irregulares. Se deben a sucesos que influyen sobre el valor de la variable en una fecha dada, modificando los efectos de los movimientos seculares y de los factores estacionales y cíclicos.

El estudio de una serie cronológica nos exige separar y estudiar en la medida de lo posible todas estas componentes.

3. METODOS DE ESTUDIO DE LA TENDENCIA SECULAR

3.1 METODO GRAFICO.- Es el más sencillo y rápido pero es demasiado subjetivo, dependiendo la bondad del método de la persona que lo aplique y del conocimiento que tenga de la serie.

Consiste en representar gráficamente la serie y después trazar a ojo una linea suave que pase entre las observaciones.

3.2 METODO ANALITICO.- Se trata de ajustar por los métodos conocidos, una función matemática de ecuación $y = f(t)$, que describa lo mejor posible la tendencia secular de la serie.

En primer lugar deberemos elegir la función de ajuste, que deberá ser sencilla. Las más usuales son:

Recta:	$y = a + b\,t$
Parábola de 2° grado:	$y = a + b\,t + c\,t^2$
Exponencial:	$y = a\,b^t$

La linea recta nos dará un buen ajuste cuando la serie varie aproximadamente en progresión aritmética, y la función exponencial nos servirá cuando la serie varie en progresión geométrica. De todos modos la representación gráfica de la serie nos ayudará a la elección de la función de ajuste.

Una vez escogida la función determinaremos sus parámetros, aplicando el método de los mínimos cuadrados, obtenemos en cada caso las siguientes ecuaciones:

- Recta: $y = a + b\,t$

Ecuaciones normales
$$\begin{cases} \sum y_i = a N + b \sum t_i \\ \sum y_i t_i = a \sum t_i + b \sum t_i^2 \end{cases}$$

- Parábola: $y = a + b\,t + c\,t^2$

Ecuaciones normales
$$\begin{cases} \sum y_i = a N + b \sum t_i + c \sum t_i^2 \\ \sum y_i t_i = a \sum t_i + b \sum t_i^2 + c \sum t_i^3 \\ \sum y_i t_i^2 = a \sum t_i^2 + b \sum t_i^3 + c \sum t_i^4 \end{cases}$$

- Exponencial: $y = a\,b^t$

  Una vez tomados logaritmos en los dos miembros de la ecuación tendremos: $\log y = \log a + t \log b$, las ecuaciones normales serán totalmente análogas al caso de la recta.

Si al representar la serie, esta muestra un cambio brusco en

Si al representar la serie,esta muestra un cambio brusco en su tendencia, en lugar de complicar la función de ajuste, tomaremos una función para los datos anteriores al cambio y otra para los posteriores.

3.3 METODO DE LAS MEDIAS MOVILES.- Este método consiste en fijar un periodo que comprenda N observaciones y calcular las sucesivas medias aritméticas desplazando una a una las observaciones.

Así,en la serie $y_1, y_2, \ldots$, calcularíamos:

$$\frac{y_1+y_2+\ldots+y_N}{N}, \quad \frac{y_2+y_3+\ldots+y_{N+1}}{N}, \quad \frac{y_3+y_4+\ldots+y_{N+2}}{N}, \ldots$$

Y estos valores los tomaremos como los de la tendencia para el centro del periodo considerado.

Si el número de observaciones N es impar, el centro del periodo coincidirá con uno de la serie; si N es par deberemos modificar los tiempos de la serie, razón por la que se toma N impar.

Cuanto mayor sea la amplitud del periodo que consideremos, proporcionaremos mayor amplitud a la curva; ahora bien, aunque subjetiva, la elección del periodo no es arbitraria.

Vamos a indicar algunos -actores que nos determinen dicha amplitud:

- Si existen fluctuaciones cíclicas con relación a una linea recta y se mantiene constante la amplitud del ciclo y la magnitud de la fluctuación, entonces todo promedio móvil cuyo periodo sea igual al del ciclo o a un múltiplo de este, nos dará una linea recta, o sea, una representación perfecta de la tendencia. Si hubieramos tomado un periodo mayor o menor que el del ciclo, habriamos obtenido no una linea recta, sino un nuevo ciclo de igual periodo, pero de menor amplitud las fluctuaciones.
- En la mayoría de los casos no se darán las condiciones anteriores, pues el periodo del ciclo variará, entonces se debe elegir un periodo igual o mayor que la longitud media del ciclo.
- Si el periodo permanece constante pero varía la amplitud del ciclo, entonces tomando un periodo para el promedio igual al del ciclo o a un múltiplo de éste, obtendremos una linea caracterizada por unos ciclos de igual periodo, pero de amplitud mínima.
- Cuando varia al periodo y la amplitud del ciclo, el promedio móvil que proporciona la mejor representación de la tendencia es aquel, cuyo periodo es igual a la longitud media del ciclo o a un múltiplo de esta.

Si la tendencia se aparta de la forma rectilinea, el error que supone el empleo de la media móvil, aumenta a medida que crece el periodo considerado. Es pues necesario emplear un periodo corto. Si la curva a la que se adapta la tendencia es cóncava, la media móvil excederá al valor real de la tendencia; si es convexa, la media móvil será siempre inferior.

En la práctica se presentan todos estos casos combinados y es preciso para la elección del periodo hacer un esrudio detallado de los datos.

4. ESTUDIO DE LAS VARIACIONES ESTACIONALES

Existen muchos métodos para determinar las variaciones estacionales, pero unicamente vamos a considerar los dos siguientes:

- de las razones a la media móvil, y
- promediando relaciones a la tendencia.

4.1 METODO DE LAS RAZONES A LA MEDIA MOVIL.- Antes de exponer el método haremos algunas consideraciones que nos permitan estudiar mejor el problema:

De vez en cuando hablaremos de años e incluso de meses sin que ello presuponga que el año es un año del calendario, ni el mes un mes del calendario sino un divisor del año.

Por un año denominaremos la longitud del periodo. Si este está mal tomado, obtendremos que no hay variación estacional con ese periodo.

La representación gráfica nos ayudará a la determinación del periodo.

A continuación vamos a detallar en que consiste este método.

En primer lugar obtenemos la media móvil para un periodo de 12 meses. Al ser el periodo un número par, la media cae entre dos unidades de tiempo.

Calculamos con las medias anteriores una nueva media de periodo dos meses con el fin de centrar las medias móviles, es decir, hacerlas corresponder a las unidades de tiempo de la serie original.

Si el periodo está bien escogido, habremos eliminado las variaciones estacionales y las variaciones accidentales, ya que la media móvil suaviza las observaciones. La nueva serie contiene la tendencia secular y las fluctuaciones cíclicas.

Bajo la hipótesis multiplicativa de que la serie cronológica es el resultado del producto de la tendencia secular, variación estacional, fluctuaciones cíclicas y variaciones accidentales, si dividimos los datos de la serie original por las correspondientes medias móviles, los resultados contendrán la variación estacional y la accidental.

Cuando la variación accidental es poco significativa, los resultados anteriores expresan bastante bien la variación estacional y podemos tomar esos datos como *índices de variación estacional de cada año*.

En el caso de que estos índices permanezcan estables al pasar de un año a otro, podremos calcular lo que llamaremos *índices generales de variación estacional*, calculando la media aritmética de cada mes para todos los meses de los años considerados.

Si un índice para un mes de un año determinado destaca sobre los de los demás años, lo eliminamos antes de calcular la media aritmética. También se puede emplear la mediana o incluso la media geométrica.

Si la serie no es estable, nos quedaríamos con los índices de variación estacional de cada año.

4.2 CALCULO DE LOS INDICES DE VARIACION ESTACIONAL PROMEDIANDO RELACIONES A LA TENDENCIA.- Este método es esencialmente idéntico al que acabamos de exponer, la única diferencia estriba en que en vez de considerar los valores de las medias móviles de doce meses centrada, consideraremos los valores de la tendencia obtenidos a partir de la curva que nos representa la tendencia secular y a continuación expresaremos los valores de la serie como porcentaje de los de la tendencia.

6. ESTUDIO DE LAS FLUCTUACIONES CICLICAS

Asi como en el estudio de la tendencia secular o de las variaciones estacionales, los métodos que hemos expuesto son muy generales y de gran aceptación, en el estudio de las fluctuaciones cíclicas hay gran diversidad de métodos y ninguno es aceptado como realmente satisfactorio.

Tanto las variaciones estacionales como las fluctuaciones cíclicas pueden no ser únicas, superponiendose varias en una misma serie cronológica.

Bajo la hipótesis multiplicativa que anteriormente citamos, eliminaremos sucesivamente los factores variación estacional, tendencia secular y variaciones accidentales para quedarnos así con las fluctuaciones cíclicas.

Comenzaremos desestacionalizando la serie para lo que dividiremos cada observación por el índice de variación estacional correspondiente. De esta manera eliminamos las variaciones estacionales.

En segundo lugar eliminaremos la tendencia secular dividiendo cada valor anterior de la serie desestacionalizada por el correspondiente valor de la tendencia.

Por último, eliminaremos las variaciones accidentales calculando medias móviles en la serie anterior para un periodo de tres o de cinco meses.

La nueva serie obtenida recoge unicamente las variaciones cíclicas de modo que si presenta una cierta periodicidad de ciclos, podremos calcular unos índices de variación cíclica análogamente a como hicimos para el caso de índices de variación estacional.

Hemos de decir que estas consideraciones son muy delicadas, ya que la hipotesis multiplicativa en que nos hemos basado, es una hipótesis muy simplificada y por otro lado cada ciclo suele tener unas características muy particulares.

Las variaciones cíclicas calculadas, nos sirven perfectamente para estudiar la serie en el periodo a que se refieren las observaciones, pero será muy arriesgado hacer predicciones para el futuro.

Para dar una visión más general a este capítulo, diremos que algunos estadísticos prefieren considerar la serie cronológica como suma de la tendencia secular, variación estacional, fluctuaciones cíclicas y variaciones accidentales.

Bajo esta hipótesis aditiva el estudio de las fluctuaciones cíclicas se haría como se refleja en la siguiente tabla:

y_i	t_i	$p_i = \frac{y_i}{t_i} 100$	$c_i = p_i - e_i$
y_1	t_1	$p_1 = \frac{y_1}{t_1} 100$	$c_1 = p_1 - e_1$
y_2	t_2	$p_2 = \frac{y_2}{t_2} 100$	$c_2 = p_2 - e_2$
.	.	.	.
.	.	.	.
.	.	.	.
y_n	t_n	$p_n = \frac{y_n}{t_n} 100$	$c_n = p_n - e_n$

donde hemos representado por:

y_i la serie dada.

t_i los valores de la tendencia.

p_i el porcentaje que cada tendencia representa en función de su valor teórico correspondiente.

e_i las variaciones estacionales.

c_i las variaciones cíclicas.

Por último, otro método que presupone ambas hipótesis consiste en:

- Multiplicar los valores de la tendencia por los indices de variación estacional en forma relativa. A estos valores los denominaremos tendencia corregida por la estacional.

- Restar a cada dato de la serie dada el valor de su tendencia corregida por la estacional, resultado que denominaremos desviación del valor real respecto de la tendencia corregida por la estacional.

- El cociente de cada desviación del valor real respecto de la tendencia corregida por la estacional, entre la tendencia corregida por la estacional, nos da las fluctuaciones cíclicas.

7. ESTUDIO DE LAS VARIACIONES ACCIDENTALES

En general estas variaciones suelen ser poco importantes y su estudio no suele tener más interés que el de probar que se trata de una componente aleatoria que tiende a una distribución normal. Si a esta variable le aplicamos un test de normalidad, nos permitirá tener cierta confianza en el estudio realizado para conocer las componentes de la serie.

Bajo la hipótesis multiplicativa, la determinación de las variaciones accidentales la haremos dividiendo por las fluctuaciones cíclicas los valores de la serie desestacionalizada dividida por los correspondientes valores de la tendencia.

8. PREDICCION

Concluido el estudio de la serie para el periodo de tiempo en que se dispone de observaciones, vamos a tratar de *predecir* su comportamiento en el futuro.

Resolveremos el problema prediciendo cada componente de la serie, si bien este método no es muy adecuado cuando esta no tiene un comportamiento regular.

La predicción de la tendencia la haremos apoyándonos en la ecuación de la tendencia secular $y = f(t)$. Obtendremos valores y' para el futuro dando valores a t en la ecuación anterior $y' = f(t)$.

La predicción de la variable estacional, si esta es estable, consistirá en tomar para el futuro los índices generales de variación estacional calculados para el periodo observado. Si la serie no es estable, podriamos suavizar dichos índices tomando medias móviles. Estos indices suavizados los extrapolariamos y tendriamos el problema resuelto.

Por lo que se refiere a las fluctuaciones cíclicas, dada la irregularidad de esta componente es muy dificil predecirla.

Las variaciones accidentales, por ser aleatorias es imposible predecirlas.

PROBLEMA RESUELTO

10.1 *Sea la serie que figura en la tabla I y que representa el número de niños nacidos vivos cada mes en España en los años 1957 a 1966.*

TABLA I: Número total de niños nacidos vivos cada mes en España en los años 1957 a 1966.

TABLA I

M E S	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
Enero.......	53575	56686	55609	55085	55145	57276	56275	57106	54748	53875
Febrero	50777	51612	52566	54670	51484	51644	52369	56003	52047	50555
Marzo.......	55489	57702	56818	56714	55696	56669	56951	58861	59441	57828
Abril.......	54392	56006	56153	55275	55359	52896	56961	59076	56702	54647
Mayo........	54598	54530	53985	58485	55629	54215	57226	57472	55349	57727
Junio.......	50968	51222	52234	53773	52139	53807	52014	57769	55063	55160
Julio.......	53668	48752	53031	54215	53521	54093	57192	61191	56105	56028
Agosto......	51656	47873	52680	55334	53629	53924	55286	56267	56715	57430
Septiembre..	54393	55515	53580	54227	53144	52936	55774	58392	58083	56919
Octubre.....	54978	56869	54726	53625	53581	55445	55704	57740	56337	55345
Noviembre...	50871	52950	50927	50156	52327	51869	51864	53471	53398	52081
Diciembre...	54088	56514	54851	52978	53959	54906	54901	55360	53761	54136

En el análisis de esta serie vamos a determinar:

A) *Tendencia secular.*
B) *Variaciones estacionales.*
C) *Fluctuaciones cíclicas.*

S O L U C I O N:

A) Estudio de la tendencia secular.

TABLA II: Medias móviles de tres meses.

Meses	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
Enero		54128	54896	54869	53202	54293	54517	56003	54052	52730
Febrero	53280	55333	54998	55490	54108	55196	55198	57323	55412	54086
Marzo	53553	55106	55179	55553	54180	53736	55427	57980	56063	54343
Abril	54826	56079	55652	56825	55561	54593	57046	58470	57164	56734
Mayo	53319	53919	54124	55844	54376	53639	55400	58106	55705	55845
Junio	53078	51501	53083	55491	53763	54038	55477	58811	55506	56305
Julio	52097	49282	52648	54441	53096	53941	54831	58409	55961	56206
Agosto	53239	50713	53097	54592	53431	53651	56084	58617	56934	56792
Septiembre	53676	53419	53662	54395	53451	54102	55588	57466	57045	56565
Octubre	53414	55111	53078	52269	53)17	53417	54447	56534	55930	54782
Noviembre	53312	55444	53501	52253	53289	54073	54156	55524	54499	53854
Diciembre	53881	55024	53621	52760	54521	54350	54624	54526	53678	

TABLA III: Medias móviles de siete meses.

Meses	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
Enero		54563	55354	54606	53553	54050	54968	56217	55643	54350
Febrero		54499	54942	55143	53839	54141	55222	56470	55301	54549
Marzo		54549	54840	55550	54061	54352	55243	57314	55530	54801
Abril	53352	53787	54342	55459	54139	54371	55570	58213	55636	55125
Mayo	53079	52528	53924	55495	53922	53892	55429	58093	55917	55633
Junio	53596	53086	54069	55432	54159	54077	55915	58434	56779	56542
Julio	53523	52967	53770	54991	53857	53902	55737	58272	56336	56179
Agosto	53020	52530	53023	54260	53424	53755	55009	57471	55864	55812
Septiembre	52947	52813	53147	53473	53185	53854	54677	57169	55637	55299
Octubre	53764	53440	53554	53669	53919	54207	55404	56737	55467	
Noviembre	53469	53985	53788	53339	53651	53961	55234	55431	54675	
Diciembre	54333	55263	54364	53391	54085	54393	55745	55884	54834	

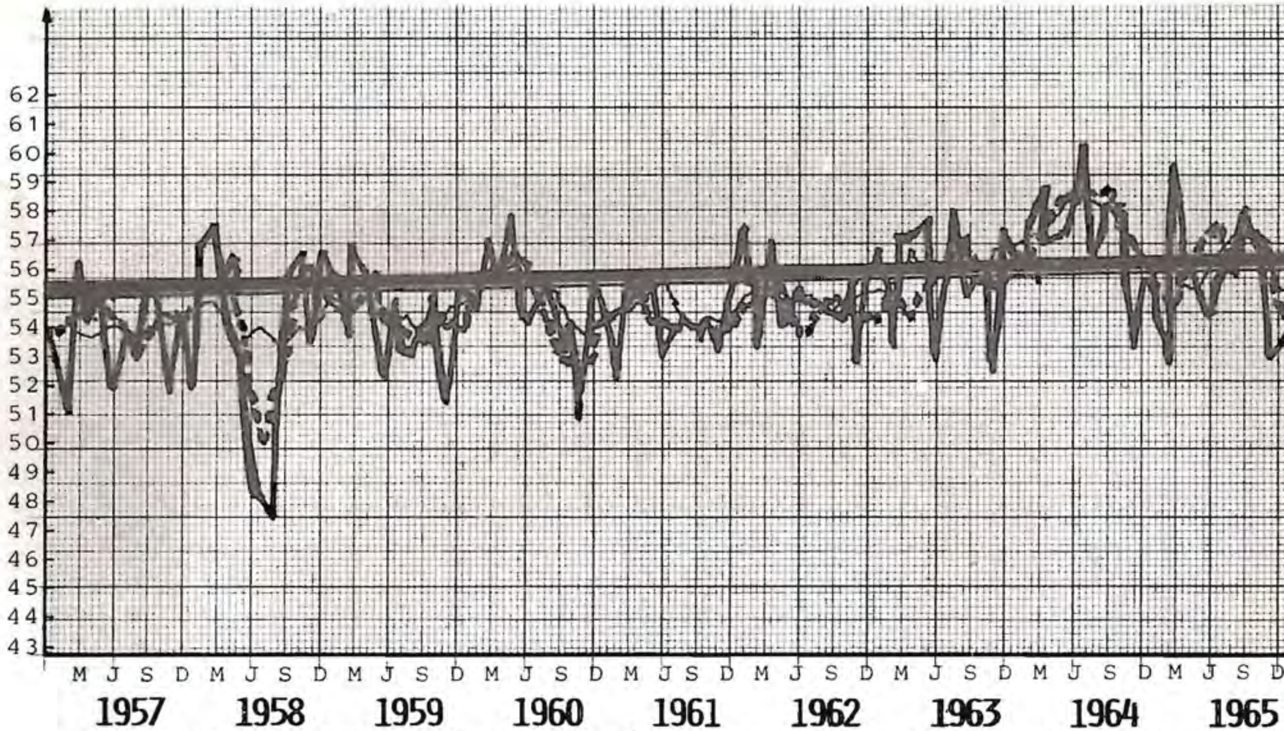


Fig. 1: Representación gráfica, del número de niños nacidos vivos en España, de los
móviles de 3 y 7 meses y de la tendencia.

:número de niños nacidos vivos
:promedio móvil de 3 meses
:promedio móvil de 7 meses
:tendencia

a) Metodo de las medias móviles:

Veanse las tablas II y III y su representación grafica en la Fig.1.

Hemos tomado un número impar de meses para facilitar el trabajo, pero si hubieramos tomado un número par hubieramos centralizado la media considerando medias móviles bimensuales. (Al estudiar las variaciones estacionales haremos uso de esto).

Es evidente que al obtener medias móviles, lo que hacemos es obtener una media suave, en la que se atenuan las influencias de las fluctuaciones que separan los valores de la tendencia general.

b) Método analítico:

Para ajustar una curva a los datos efectuaremos los siguientes pasos:

- Elegir la función de ajuste.
- Determinar los parámetros.

La observación de la representación gráfica de los datos nos permite intuir como éstos muestran una tendencia que puede ser representada por una linea recta, luego escogeremos como función de ajuste la ecuación: $y = a + b\,t$.

Para la determinación de los parámetros a y b, emplearemos el método de los mínimos cuadrados, obteniendo el siguiente sistema de ecuaciones normales:

$$\begin{cases} \sum y_i = N a + b \sum t_i \\ \sum y_i t_i = a \sum t_i + b \sum t_i^2 \end{cases}$$

Cuando, como en este caso, los valores de t son consecutivos, las ecuaciones anteriores se simplifican ya que tomando como origen un valor medio, se tiene $\sum t_i = 0$, obteniendose el siguiente sistema:

$$\begin{cases} \sum y_i = N\, a \\ \sum y_i t_i = b \sum t_i^2 \end{cases}$$

Para simplificar los cálculos hacemos el siguiente cambio de variable: $x_i = y_i - h$, donde h = 55000. Con en este cambio, la segunda ecuación del sistema anterior quedará de la siguiente forma:

$$\sum t_i (x_i + h) = \sum x_i t_i + h \sum t_i = \sum x_i t_i, \text{ ya que } \sum t_i = 0$$

La determinación de $\sum x_i t_i$, figura en la tabla IV.

Como sabemos que la suma de los cuadrados de los n primeros números naturales, viene dada por la expresión:

$$\frac{2\,n^3 + 3\,n^2 + n}{6}$$

Obtenemos el siguiente sistema:

$$\begin{cases} 6563379 = 119\ a \\ 3466267 = 1137640\ b \end{cases}$$

de donde, a = 55154,4 y b = 3,04.

Por tanto, la ecuación de la recta que mejor se adapta a la tendencia de los datos es: $\boxed{y = 55154{,}4 + 3{,}04\,t}$ Vease su representación gráfica en la Fig 1 y téngase en cuenta el cambio de escalas.

TABLA IV

Año y mes	t_i	x_i	$x_i . t_i$	Año y mes	t_i	x_i	$x_i . t_i$
1957				**1960**			
Enero	-59	-1425	84075	Enero	-23	85	-1955
Febrero	-58	-4223	244934	Febrero	-22	-330	7260
Marzo	-57	489	-27863	Marzo	-21	1714	-35994
Abril	-56	-608	34048	Abril	-20	275	-5500
Mayo	-55	-402	22110	Mayo	-19	3485	-66215
Junio	-54	-4032	217728	Junio	-18	-1227	22086
Julio	-53	-1332	70596	Julio	-17	-785	13345
Agosto	-52	-3344	173888	Agosto	-16	334	-5344
Septiembre	-51	-607	30957	Septiembre	-15	-773	11595
Octubre	-50	-22	1100	Octubre	-14	-1375	19250
Noviembre	-49	-4129	202321	Noviembre	-13	-4844	62972
Diciembre	-48	-912	43776	Diciembre	-12	-2022	24264
1958				**1961**			
Enero	-47	1686	-79242	Enero	-11	145	-1595
Febrero	-46	-3388	155848	Febrero	-10	-3516	35160
Marzo	-45	2702	-121590	Marzo	-9	696	-6264
Abril	-44	1006	-44264	Abril	-8	359	-2872
Mayo	-43	-470	20210	Mayo	-7	629	-4403
Junio	-42	-3778	158706	Junio	-6	-2861	17166
Julio	-41	-6250	256250	Julio	-5	-1479	7395
Agosto	-40	-7127	285080	Agosto	-4	-1371	5484
Septiembre	-39	515	20085	Septiembre	-3	-1856	5568
Octubre	-38	1869	-71022	Octubre	-2	-1419	2838
Noviembre	-37	-2050	75850	Noviembre	-1	-2673	2673
Diciembre	-36	1514	-54504	Diciembre	0	0	0
1959				**1962**			
Enero	-35	609	-21315	Enero	1	2276	2276
Febrero	-34	-2434	-82756	Febrero	2	-3356	-6712
Marzo	-33	1818	-59994	Marzo	3	1669	5007
Abril	-32	1153	-36896	Abril	4	-2104	-8416
Mayo	-31	1015	31465	Mayo	5	-785	-3925
Junio	-30	-2766	82980	Junio	6	-1193	-7158
Julio	-29	-1969	57101	Julio	7	-907	-6349
Agosto	-28	-2320	64960	Agosto	8	-1076	-8608
Septiembre	-27	-1420	38340	Septiembre	9	-2064	-18576
Octubre	-26	-274	7124	Octubre	10	445	4450
Noviembre	-25	-4073	101825	Noviembre	11	-3131	-34441
Diciembre	-24	-149	3576	Diciembre	12	-94	-1128

TABLA IV (Continuación)

Año y mes	t_i	x_i	$x_i . t_i$	Año y mes	t_i	x_i	$x_i . t_i$
1963				**1965**			
Enero	13	1275	16575	Enero	37	252	-9324
Febrero	14	-2631	-36834	Febrero	38	-2953	-112214
Marzo	15	1951	29265	Marzo	39	4441	173199
Abril	16	1961	31376	Abril	40	1702	68080
Mayo	17	2226	37842	Mayo	41	349	14309
Junio	18	-2986	-53748	Junio	42	63	2646
Julio	19	2192	41648	Julio	43	1105	47515
Agosto	20	286	5720	Agosto	44	1715	75460
Septiembre	21	774	16254	Septiembre	45	3083	138735
Octubre	22	704	15488	Octubre	46	1337	61502
Noviembre	23	-3136	-72128	Noviembre	47	-1602	-75294
Diciembre	24	-99	-2376	Diciembre	48	1239	-59472
1964				**1966**			
Enero	25	2106	52650	Enero	49	-1125	-55125
Febrero	26	1003	26078	Febrero	50	-4445	-22225
Marzo	27	3861	104247	Marzo	51	2848	144228
Abril	28	4076	114128	Abril	52	-353	-18356
Mayo	29	2472	71688	Mayo	53	2727	144531
Junio	30	2769	83070	Junio	54	160	8640
Julio	31	6191	191921	Julio	55	1028	56540
Agosto	32	1267	40544	Agosto	56	2430	13608
Septiembre	33	3392	111936	Septiembre	57	1919	109383
Octubre	34	2740	93160	Octubre	58	345	20010
Noviembre	35	-1529	-53515	Noviembre	59	-2919	-172221
Diciembre	36	360	12960	Diciembre			

B) Media de las variaciones estacionales.

a) Método de las razones a la media móvil.

Las fluctuaciones periódicas ocurren dentro de un periodo constante de 12 meses, lo que nos pérmite emplear medias móviles de periodo 12.

Obtendremos medias móviles de doce meses centradas, que ajustaremos luego por otra media móvil de dos meses centrada, y hallaremos la relación existente entre los datos reales mensuales y las medias moviles correspondientes al mismo mes. Por último calcularemos la media de los porcentajes correspondientes a un mismo mes del año y tomaremos esa media como índice de variación estacional.

La media de los doce valores mensuales, cuando se centra, coincide con el 1 de julio, la media de doce meses, tomando como primer mes febrero coincide con el 1 de agosto. Si hallamos la media de estos datos corresponderá al 15 de julio. Estos valores son los que figuran en la tabla V.

TABLA V: Media móvil de doce meses centrada y ajustada por otra media móvil de dos meses centrada.

Meses	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
Enero		53746	53998	54531	53804	53913	54710	56583	56102	55347
Febrero		53383	54377	54691	53704	53949	54896	56791	55908	55374
Marzo		53273	54497	54829	53588	53953	55021	56941	55914	55350
Abril		53399	54327	54810	53541	54022	55200	57135	55842	55264
Mayo		53564	54153	54732	53629	54081	55211	57287	55781	55168
Junio		53752	53999	54622	53760	54101	55210	57373	55712	55128
Julio	53417	53808	53908	54547	53890	54098	55244	57294	55609	
Agosto	53582	53802	53973	54417	53985	54087	55430	57031	55511	
Septiembre	53709	53805	54056	54242	54032	54128	55661	56890	55382	
Octubre	53868	53774	54015	54203	53970	54309	55829	56815	55229	
Noviembre	53933	53758	54166	54088	53809	54604	55928	56627	55243	
Diciembre	53940	53778	54418	53901	53819	54655	56177	56426	55346	

Expresamos ahora los datos de la serie como porcentaje de los correspondientes medias móviles que acabamos de obtener. Estos porcentajes son los que figuran en la tabla VI.

TABLA VI

Meses	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
Enero		105,4	102,9	101,0	102,4	106,2	102,8	100,9	97,5	97,3
Febrero		96,6	96,7	99,9	95,8	95,7	95,4	98,6	93,1	91,3
Marzo		108,3	104,2	103,4	103,9	105,0	103,5	103,4	106,3	104,5
Abril		104,9	103,3	100,8	103,4	97,9	103,2	103,4	101,5	98,9
Mayo		101,8	99,7	106,8	103,7	100,2	103,6	100,3	99,2	104,6
Junio		95,3	96,7	98,4	96,9	99,4	94,2	100,6	98,8	100,0
Julio	100,4	90,6	98,4	99,4	99,3	99,9	103,5	106,8	100,8	
Agosto	96,4	88,9	97,6	101,7	99,3	99,7	99,7	98,6	102,1	
Septiembre	101,2	103,1	99,1	99,9	98,3	97,7	100,2	102,6	104,8	
Octubre	102,1	105,7	101,3	98,9	99,3	102,9	99,7	101,6	102,0	
Noviembre	94,3	98,5	94,0	92,7	97,2	94,9	92,7	94,4	96,6	
Diciembre	102,1	105,1	100,8	98,3	100,2	100,4	97,7	98,1	97,1	

Los nueve porcentajes obtenidos para cada mes lo promediamos y así obtenemos el índice de variación estacional. Estos datos figuran en la tabla VII

TABLA VII: Indices de variación estacional calculados a partir de las medias móviles.

Meses	medias aritméticas	medianas sin corregir	medianas corregidas
Enero	101,8	102,4	102,3
Febrero	95,9	95,8	95,7
Marzo	104,7	104,2	104,1
Abril	101,9	103,2	103,
Mayo	102,2	101,8	101,7
Junio	97,8	98,4	98,3
Julio	99,9	99,9	99,9
Agosto	98,5	99,3	99,1
Septiembre	100,8	100,2	100,1
Octubre	101,5	101,6	101,5
Noviembre	95,	94,4	94,3
Diciembre	100,	100,2	100,1

Al promediar hemos hallado la media aritmética y la mediana y vemos que en el caso de la mediana el promedio se aparta algo de 100 por lo que ajustaremos los índices mensuales a fin de que su promedio sea igual a 100. Para ello dividimos por 1,00116.

b) Cálculo de los índices de variación estacional promediando relaciones a la tendencia.

Anteriormente obtuvimos una recta de ecuación:

$$y = 55154{,}4 + 3{,}04\ t$$

con origen el día 15 de diciembre de 1961.

El valor de la tendencia para cada mes, centrado en el 15 de dicho mes, se obtiene sin más que ir dando valores a t en la ecuación anterior. Estos valores son los que figuran en la tabla VIII.

TABLA VIII : Valores de la tendencia del número de niños nacidos vivos en España en los años 1957 a 1966.

TABLA VIII

Meses	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
Enero	54975	55012	55048	55085	55121	55157	55193	55230	55266	55303
Febrero	54978	55015	55051	55088	55124	55160	55196	55233	55269	55306
Marzo	54981	55018	55054	55091	55127	55163	55199	55236	55272	55309
Abril	54984	55021	55057	55094	55130	55166	55202	55239	55275	55312
Mayo	54987	55024	55060	55097	55133	55169	55205	55242	55278	55315
Junio	54990	55027	55063	55100	55136	55172	55208	55245	55281	55318
Julio	54993	55030	55066	55103	55139	55175	55211	55248	55284	55321
Agosto	54996	55033	55069	55106	55142	55178	55214	55251	55287	55324
Septiembre	54999	55036	55072	55109	55145	55181	55217	55254	55290	55327
Octubre	55002	55039	55075	55112	55148	55184	55220	55257	55293	55330
Noviembre	55005	55042	55078	55115	55151	55187	55223	55260	55296	55333
Diciembre	55008	55045	55081	55118	55154	55190	55226	55263	55299	55336

Teniendo ya los valores de la tendencia para cada mes, expresaremos los valores reales como porcentaje de sus correspondientes valores de la tendencia. Estos resultados figuran en la tabla IX.

TABLA IX: Relación en porcentaje de los valores de la serie a los avlores de la tendencia.

TABLA IX

Meses	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
Enero	97,4	103,0	101,0	100,0	100,0	103,8	101,9	103,4	99,1	97,4
Febrero	92,3	93,8	95,4	99,2	93,4	93,6	94,8	101,3	94,1	91,4
Marzo	100,9	104,8	103,2	102,9	101,0	102,7	103,1	106,5	107,5	104,5
Abril	98,9	101,7	101,9	100,3	100,4	95,8	103,2	106,9	102,5	98,8
Mayo	99,2	99,1	98,0	106,1	100,9	98,2	103,6	104,0	100,1	104,3
Junio	92,6	93,1	94,8	97,6	94,5	97,5	94,2	104,5	99,6	99,7
Julio	97,6	88,6	96,3	98,4	97,0	98,0	103,6	110,7	101,4	101,3
Agosto	93,9	86,9	95,6	100,4	97,2	97,7	100,1	101,8	102,5	103,8
Septiembre	98,9	100,8	97,3	98,4	96,3	95,9	101,0	105,6	105,0	102,8
Octubre	99,9	103,3	99,3	97,3	97,1	100,4	100,8	104,5	101,9	100,0
Noviembre	92,4	96,2	92,4	91,0	94,8	93,9	93,9	96,7	96,5	94,1
Diciembre	98,3	102,6	99,6	96,1	97,8	99,5	99,4	100,1	97,2	97,8

La media de cuatro de dichos valores centrales, nos sirve mejor para la determinación del índice de variación estacional que la mediana o la media aritmética. Vease Tabla X.

TABLA X: Indice de variación estacional del número de niños nacidos vivos cada mes en España, basados en la relación de porcentaje de los valores de la serie a los valores de la tendencia.

TABLA X

Meses	números índices sin ajustar (basados en cuatro datos centrales)	números índices corregidos (basados en cuatro datos centrales)
Enero	100,7	101,8
Febrero	94,1	95,1
Marzo	103,4	104,5
Abril	101,1	102,2
Mayo	100,9	102,0
Junio	96,1	97,2
Julio	98,8	99,9
Agosto	98,8	99,9
Septiembre	98,8	100,9
Octubre	100,3	101,4
Noviembre	94,2	95,2
Diciembre	98,8	99,9
	98,917	100,-

C) Medida de las fluctuaciones cíclicas.

Los cálculos se recogen directamente en la tabla XI.

TABLA XI: Medida de las fluctuaciones cíclicas

TABLA XI

Años y Meses	Valor de la tendencia T	Indice estacional (en forma relativa) E	Tendencia corregida por la estacional T.E	Valor real R	Desviación del valor real respecto de la tendencia corregida por la estacional R-T.E	Desviación en porcentaje del valor real respecto de la tendencia corregida por la estacional $\frac{R-T.E}{T.E}$
1957						
Enero	54975	1,018	55964	53575	-2389	-4,3
Febrero	54978	0,951	52284	50777	-1507	-2,8
Marzo	54981	1,045	57455	55489	-1966	-3,4
Abril	54984	1,022	56194	54392	-1802	-3,2
Mayo	54987	1,020	56087	54598	-1489	-2,6
Junio	54990	0,972	53440	50968	-2472	-4,6
Julio	54993	0,999	54938	53668	-1270	-2,3
Agosto	54996	0,999	54941	51656	-3285	-5,9
Septiembre	54999	1,009	55494	54393	-1101	-1,9
Octubre	55002	1,014	55772	54978	- 794	-1,4
Noviembre	55005	0,952	52365	50871	-1494	-2,8
Diciembre	55008	0,999	54953	54088	- 865	-1,5
1958						
Enero	55012	1,018	56002	56686	684	1,2
Febreo	55015	0,951	52319	51612	-707	-1,3
Marzo	55018	1,045	57494	57702	208	0,3
Abril	55021	1,022	56231	56006	-225	-0,4
Mayo	55024	1,020	56124	54530	-1594	-2,8
Junio	55027	0,972	53486	51222	-2264	-4,2
Julio	55030	0,999	54975	48752	-6223	-11,3
Agosto	55033	0,999	54978	47873	-7105	-12,9
Septiembre	55036	1,009	55531	55515	-16	-0,03
Octubre	55039	1,014	55810	56869	1059	1,9
Noviembre	55042	0,952	52400	52950	550	1,0
Diciembre	55045	0,999	54990	56514	1524	2,7

TABLA XI (Continuación)

Año y mes	T	E	T.E	R	R-T.E	$\frac{R-T.E}{T.E}$
1959						
Enero	55048	1,018	56039	55609	-430	-0,7
Febrero	55051	0,951	52353	52566	213	0,4
Marzo	55054	1,045	57531	56818	-713	-1,2
Abril	55057	1,022	56268	56153	-115	-0,2
Mayo	55060	1,02o	56161	53985	-2176	-3,9
Junio	55063	0,972	53521	52234	-1287	-2,4
Julio	55066	0,999	55011	53031	-1980	-3,6
Agosto	55069	0,999	55014	52680	-2334	-4,2
Septiembre	55072	1,099	55568	53580	-1988	-3,5
Octubre	55075	1,014	55846	54726	-1120	-2,0
Noviembre	55078	0,952	52434	50927	-1507	-2,8
Diciembre	55081	0,999	55026	54851	- 175	-0,3
1960						
Enero	55085	1,018	56076	55085	- 991	-1,7
Febrero	55088	0,951	52389	54670	2281	4,3
Marzo	55091	1,045	57570	56714	- 856	-1,5
Abril	55094	1,022	56306	55275	-1031	-1,8
Mayo	55097	1,020	56199	58485	2286	4,0
Junio	55100	0,972	53557	53773	216	0,4
Julio	55103	0,999	55048	54215	- 833	-1,5
Agosto	55106	0,999	55052	55334	282	0,5
Septiembre	55109	1,009	55605	54227	-1378	-2,4
Octubre	55112	1,014	55884	53625	-2259	-4,0
Noviembre	55115	0,952	52469	50156	-2313	-4,4
Diciembre	55118	0,999	55063	52978	-2085	-3,7
1961						
Enero	55121	1,018	56113	55145	- 968	-1,7
Febrero	55124	0,951	52424	51484	- 940	-1,7
Marzo	55127	1,045	57608	55696	-1912	-3,3
Abril	55130	1,022	56343	55359	- 984	-1,7
Mayo	55133	1,020	56136	55629	- 607	-1,1
Junio	55136	0,972	53594	52139	-1455	-2,7
Julio	55139	0,999	55084	53521	-1563	-2,8
Agosto	55142	0,999	55088	53629	-1459	-2,6
Septiembre	55145	1,009	55641	53144	-2497	-4,4
Octubre	55148	1,014	55920	53581	-2339	-4,2
Noviembre	55151	0,952	52504	52327	- 177	-0,3
Diciembre	55154	0,999	55099	53959	-1140	-2,1

TABLA XI (Continuación)

Año y mes	T	E	T.E	R	R-T.E	$\frac{R-T.E}{T.E}$
1962						
Enero	55157	1,018	56150	57276	1126	2,0
Febrero	55160	0,951	52459	51644	- 815	-1,5
Marzo	55163	1,045	57645	56669	- 976	-1,7
Abril	55166	1,022	56380	52896	-3484	-6,1
Mayo	55169	1,020	56273	54215	-2058	-3,6
Junio	55172	0,972	53628	53807	179	0,3
Julio	55175	0,999	55120	54093	-1027	-1,8
Agosto	55178	0,999	55124	53924	-1200	-2,2
Septiembre	55181	1,009	55677	52931	-2741	-4,9
Octubre	55184	1,014	55956	55445	- 511	-0,9
Noviembre	55187	0,952	52539	51869	- 670	-1,2
Diciembre	55190	0,999	55135	54906	- 229	-0,4
1963						
Enero	55193	1,018	56187	56275	88	0,1
Febrero	55196	0,951	52494	52369	- 125	-0,2
Marzo	55199	1,045	57683	56951	- 732	-1,3
Abril	55202	1,022	56417	56961	544	0,9
Mayo	55205	1,020	56310	57226	916	1,6
Junio	55208	0,972	53663	52014	-1649	-3,7
Julio	55211	0,999	55156	57192	2036	3,7
Agosto	55214	0,999	55160	55286	126	0,2
Septiembre	55217	1,009	55714	55774	60	0,1
Octubre	55220	1,014	55993	55704	- 289	-0,5
Noviembre	55223	0,952	52574	51864	- 710	-1,3
Diciembre	55226	0,999	55171	54901	- 270	-0,5
1964						
Enero	55230	1,018	56224	57106	882	1,5
Febrero	55233	0,951	52529	56003	3474	6,6
Marzo	55236	1,045	57721	58861	1140	1,9
Abril	55239	1,022	56454	59076	2622	4,6
Mayo	55242	1,020	56347	57472	1125	1,9
Junio	55245	0,972	53699	57769	4070	7,5
Julio	55248	0,999	55193	61191	5998	10,8
Agosto	55251	0,999	55197	56267	1070	1,9
Septiembre	55254	1,009	55751	58392	2641	4,7
Octubre	55257	1,014	56031	57740	1709	3,0
Noviembre	55260	0,952	52610	53471	861	1,6
Diciembre	55263	0,999	55208	55360	152	0,3

TABLA XI (Continuación)

Mes y año	T	E	T.E	R	R-T.E	$\frac{R-T.E}{T.E}$
1965						
Enero	55266	1,018	56261	54748	-1513	-2,6
Febrero	55269	0,951	52564	52047	- 517	-0,9
Marzo	55272	1,045	57758	59441	1683	2,9
Abril	55275	1,022	56491	56702	211	0,4
Mayo	55278	1,020	56384	55349	-1035	-1,8
Junio	55281	0,972	53734	55063	1329	2,5
Julio	55284	0,999	55229	56105	876	1,5
Agosto	55287	0,999	55233	56715	1482	2,7
Septiembre	55290	1,009	55787	58083	2296	4,1
Octubre	55293	1,014	56068	56337	269	0,4
Noviembre	55296	0,952	52645	53398	753	1,2
Diciembre	55299	0,999	55244	53761	-1483	-2,6
1966						
Enero	55303	1,018	56299	53875	-2424	-4,3
Febrero	55306	0,951	52600	50555	-2045	-3,9
Marzo	55309	1,045	57796	57828	32	0,05
Abril	55312	1,022	56529	54647	-1882	-3,3
Mayo	55315	1,020	56422	57727	1305	2,3
Junio	55318	0,972	53770	55160	1390	2,6
Julio	55321	0,999	55266	56028	762	1,3
Agosto	55324	0,999	55270	57430	2160	3,9
Septiembre	55327	1,009	55824	56919	1095	1,9
Octubre	55330	1,014	56106	55345	-761	-1,3
Noviembre	55333	0,952	52680	52081	-599	-1,1
Diciembre	55336	0,999	55281	54136	-1145	-2,1

Los valores de los ciclos figuran en las dos últimas columnas de la tabla.

RESUMEN

Con lo que acabamos de hacer queda completo el estudio de la serie cronológica dada.

Tal y como nos habiamos propuesto, hemos determinado, en primer lugar la tendencia secular, o tendencia durante un periodo de tiempo largo de la serie.

Empleamos para ello dos procedimientos:

1°) Método de las medias móviles de tres y siete meses.

2°) Método analítico.

Con el primero los resultados son los que figuran en las tablas II y III, y con el segundo obtuvimos que la tendencia se ajustaba a la recta de ecuación:

$$y = 55154,4 + 3,04\ t$$

A continuación medimos las fluctuaciones de carácter estacional y cíclico. Definimos y determinamos primero, un índice de variación estacional por dos procedimientos:

a) Mediante el empleo de medias móviles de doce datos, por tener estas variaciones un periodo constante de doce meses.

b) Promediando relaciones a la tendencia.

Por ambos procedimientos, como era de esperar, obtuvimos los mismos resultados que son los que figuran en las tablas VII y X.

A continuación determinamos las fluctuaciones cíclicas definiendo un valor que venia dado por el de la tendencia, corregido por el índice de variación estacional; la desviación existente entre ese valor y el real que nos permitió determinar esas fluctuaciones y los resultados son los que figuran en la tabla XI.

****

bibliografía

ANDERSON, R.L., BANCROFT, T.A., *Statiscal Theory in Research.* Mc-Graw-Hill Book Company, Inc New York, 1952.

AZORIN POCH, F., *Curso de muestreo y aplicaciones.* Instituto Nacional de Estadística, Madrid,1962.

AZOULAY, E., COHEN, D., *Cours et exercices de Statistique.* Société d'edition d'enseignement supérieur, Paris, 1967.

BANCROFT, H., *Introducción a la estadística biológica.* Atika S.A., Madrid, 1966.

BLISS, C.I., *The calculation of the dosage-mortality curve.* Journal of the American Statistical Association, (vol. 39), 1944.

CALOT, G., *Curso de Estadística Descriptiva.* Paraninfo, Madrid,1970.

CRAMER, H., *Métodos matemáticos de Estadística.* Editorial Aguilar S.A. Madrid, 1953.

CRAMER, H., *Elementos de la teoría de probabilidades y aplicaciones* Editorial Aguilar S.A. Madrid, 1958.

CROXTON, G., *Medical Statistics.* University of Chicago Press, 1962.

DAVID, F.N., *Probability theory for Statistical Methods.* Cambridge University Press, New York, 1949.

DIXON & MASSEY. *Introducción al Análisis Estadístico.* Mc-Graw-Hill. Ediciones del Castillo, 1966.

EISENHART, C., & WILSON, P.W., *Statistical methods and control in bacteriology.* Bacteriological Reviews. (vol. 7), 1943.

FELLER, W., *An introduction to probability theory and its applications,* (vol.1). John Wiley & Sons, Inc. New York, 1957.

FINNEY, D.J., *Statistical methods in biological Assay*. Charles Griffin & Co Ltd. Londres, 1952.

FISHER, R.A., YATES, F., *Statistical tables for Biological, Agricultural and Medical Research*. 4ª ed. Oliver & Boyd Ltd. Edimburgo, 1953.

GIRAULT, M., *Calcul des Probabilités en vue des applications*. Dunod. Paris, 1960.

GMURMAN, V.E., *Teoría de las probabilidades y la estadística matemática*. Editorial Mir. Moscu, 1974.

GOLDSTEIN, A., *Bioestatistics*. Mac Millan and Co Limited. Londres, 1964.

HILL, A.B., *Principles of Medical Statistics*. Oxford University Press. New York, 1955.

LABROUSSE, C., *Estadística, ejercicios resueltos*. (vols. 1,2,3). Editorial Paraninfo. Madrid, 1968.

LAMOTTE, M., *Estadística biológica*. Toray-Masson.S.A. Barcelona 1965.

MAINLAN, D., *Elementary Medical Statistics*. W.B. Saunders Company. Filadelfia, 1952.

MATHER, K., *Análisis Estadístico en Biología*. Editorial Paraninfo. Madrid, 1971.

MOOD, A.M., *Introducción a la teoría estadística*. Editorial Aguilar S.A. Madrid, 1955.

NEYMAN, J., *First Course in Probability and Statistics*. Henry Holt and Company Inc. New York, 1950.

PARZEN, E., *Modern Probability Theory and its Applications*. John Wiley & Sons. 1960.

RAO, F., *Advanced Statistical Methods in Biometric Research*. John Wilwy & Sons. 1952.

RENYI, A., *Calcul des Probabilités*. Dunod. Paris, 1966.

RIOS, S., *Métodos estadísticos*. Ediciones del Castillo S.A. Madrid, 1971.

RIOS, S., *Análisis estadístico aplicado*. Editorial Paraninfo. Madrid, 1972.

RIOS, S., *Análisis estadístico-matemático*. Madrid, 1965.

RIOS, S., *Iniciación estadística*. Madrid, 1958.

RIOS, S., *Ejercicios de estadística*. Ediciones I.C.E. Madrid, 1974.

TUCKER, H.G., *Introducción a la teoría matemática de las probabilidades y la estadística*. Vicens-Vives, 1974.

WILKS, S.S., *Análisis estadístico elemental*. Consejo Superior de Investigaciones Cientificas. Madrid, 1952.

apéndice

TABLA I : DISTRIBUCION BINOMIAL

TABLA II : DISTRIBUCION DE POISSON

TABLA III : DISTRIBUCION NORMAL

TABLA IV : DISTRIBUCION χ^2 DE PEARSON

TABLA V : DISTRIBUCION t DE STUDENT

TABLA VI : DISTRIBUCION F DE FISHER-SNEDECOR

Los autores expresan su agradecimiento a **R.A.FISHER** y **F.YATES** por la utilización de: *Statistical tables for Biological, Agricultural and Medical Research.*

TABLA I

DISTRIBUCION BINOMIAL

$$p(X = r) = \binom{n}{r} p^r q^{n-r}$$

n	r	.01	.05	.10	.15	.20	.25	.30	$\frac{1}{3}$	.35	.40	.45	.49	.50
2	0	.9801	.9025	.8100	.7225	.6400	.5625	.4900	.4444	.4225	.3600	.3025	.2601	.2500
	1	.0198	.0950	.1800	.2550	.3200	.3750	.4200	.4444	.4550	.4800	.4950	.4998	.5000
	2	.0001	.0025	.0100	.0225	.0400	.0625	.0900	.1111	.1225	.1600	.2025	.2401	.2500
3	0	.9703	.8574	.7290	.6141	.5120	.4219	.3430	.2963	.2746	.2160	.1664	.1327	.1250
	1	.0294	.1354	.2430	.3251	.3840	.4219	.4410	.4444	.4436	.4320	.4084	.3823	.3750
	2	.0003	.0071	.0270	.0574	.0960	.1406	.1890	.2222	.2389	.2880	.3341	.3674	.3750
	3	.0000	.0001	.0010	.0034	.0080	.0156	.0270	.0370	.0429	.0640	.0911	.1176	.1250
4	0	.9606	.8145	.6561	.5220	.4096	.3164	.2401	.1975	.1785	.1296	.0915	.0677	.0625
	1	.0388	.1715	.2916	.3685	.4096	.4219	.4116	.3951	.3845	.3456	.2995	.2600	.2500
	2	.0006	.0135	.0486	.0975	.1536	.2109	.2646	.2963	.3105	.3456	.3675	.3747	.3750
	3	.0000	.0005	.0036	.0115	.0256	.0469	.0756	.0988	.1115	.1536	.2005	.2400	.2500
	4	.0000	.0000	.0001	.0005	.0016	.0039	.0081	.0123	.0150	.0256	.0410	.0576	.0625
5	0	.9510	.7738	.5905	.4437	.3277	.2373	.1681	.1317	.1160	.0778	.0503	.0345	.0312
	1	.0480	.2036	.3280	.3915	.4096	.3955	.3602	.3292	.3124	.2592	.2059	.1657	.1562
	2	.0010	.0214	.0729	.1382	.2048	.2637	.3087	.3292	.3364	.3456	.3369	.3185	.3125
	3	.0000	.0011	.0081	.0244	.0512	.0879	.1323	.1646	.1811	.2304	.2757	.3060	.3125
	4	.0000	.0000	.0004	.0022	.0064	.0146	.0284	.0412	.0488	.0768	.1128	.1470	.1562
	5	.0000	.0000	.0000	.0001	.0003	.0010	.0024	.0041	.0053	.0102	.0185	.0283	.0312
6	0	.9415	.7351	.5314	.3771	.2621	.1780	.1176	.0878	.0754	.0467	.0277	.0176	.0156
	1	.0571	.2321	.3543	.3993	.3932	.3560	.3025	.2634	.2437	.1866	.1359	.1014	.0938
	2	.0014	.0305	.0984	.1762	.2458	.2966	.3241	.3292	.3280	.3110	.2780	.2437	.2344
	3	.0000	.0021	.0146	.0415	.0819	.1318	.1852	.2195	.2355	.2765	.3032	.3121	.3125
	4	.0000	.0001	.0012	.0055	.0154	.0330	.0595	.0823	.0951	.1382	.1861	.2249	.2344
	5	.0000	.0000	.0001	.0004	.0015	.0044	.0102	.0165	.0205	.0369	.0609	.0864	.0938
	6	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0014	.0018	.0041	.0083	.0139	.0156
7	0	.9321	.6983	.4783	.3206	.2097	.1335	.0824	.0585	.0490	.0280	.0152	.0090	.0078
	1	.0659	.2573	.3720	.3960	.3670	.3115	.2471	.2048	.1848	.1306	.0872	.0603	.0547
	2	.0020	.0406	.1240	.2097	.2753	.3115	.3177	.3073	.2985	.2613	.2140	.1740	.1641
	3	.0000	.0036	.0230	.0617	.1147	.1730	.2269	.2561	.2679	.2903	.2918	.2786	.2734
	4	.0000	.0002	.0026	.0109	.0287	.0577	.0972	.1280	.1442	.1935	.2388	.2676	.2734
	5	.0000	.0000	.0002	.0012	.0043	.0115	.0250	.0384	.0466	.0774	.1172	.1543	.1641
	6	.0000	.0000	.0000	.0001	.0004	.0013	.0036	.0064	.0084	.0172	.0320	.0494	.0547
	7	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0006	.0016	.0037	.0068	.0078
8	0	.9227	.6634	.4305	.2725	.1678	.1001	.0576	.0390	.0319	.0168	.0084	.0046	.0039
	1	.0746	.2793	.3826	.3847	.3355	.2670	.1977	.1561	.1373	.0896	.0548	.0352	.0312
	2	.0026	.0515	.1488	.2376	.2936	.3115	.2965	.2731	.2587	.2090	.1569	.1183	.1094
	3	.0001	.0054	.0331	.0839	.1468	.2076	.2541	.2731	.2786	.2787	.2568	.2273	.2188
	4	.0000	.0004	.0046	.0185	.0459	.0865	.1361	.1707	.1875	.2322	.2627	.2730	.2734
	5	.0000	.0000	.0004	.0026	.0092	.0231	.0467	.0683	.0808	.1239	.1719	.2098	.2188
	6	.0000	.0000	.0000	.0002	.0011	.0038	.0100	.0171	.0217	.0413	.0703	.1008	.1094
	7	.0000	.0000	.0000	.0000	.0001	.0004	.0012	.0024	.0033	.0079	.0164	.0277	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0002	.0007	.0017	.0033	.0039
9	0	.9135	.6302	.3874	.2316	.1342	.0751	.0404	.0260	.0207	.0101	.0046	.0023	.0020
	1	.0830	.2985	.3874	.3679	.3020	.2253	.1556	.1171	.1004	.0605	.0339	.0202	.0176
	2	.0034	.0629	.1722	.2597	.3020	.3003	.2668	.2341	.2162	.1612	.1110	.0776	.0703
	3	.0001	.0077	.0446	.1069	.1762	.2336	.2668	.2731	.2716	.2508	.2119	.1739	.1641
	4	.0000	.0006	.0074	.0283	.0661	.1168	.1715	.2048	.2194	.2508	.2600	.2506	.2461
	5	.0000	.0000	.0008	.0050	.0165	.0389	.0735	.1024	.1181	.1672	.2128	.2408	.2461
	6	.0000	.0000	.0001	.0006	.0028	.0087	.0210	.0341	.0424	.0743	.1160	.1542	.1641
	7	.0000	.0000	.0000	.0000	.0003	.0012	.0039	.0073	.0098	.0212	.0407	.0635	.0703
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0009	.0013	.0035	.0083	.0153	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003	.0008	.0016	.0020
10	0	.9044	.5987	.3487	.1969	.1074	.0563	.0282	.0173	.0135	.0060	.0025	.0012	.0010
	1	.0914	.3151	.3874	.3474	.2684	.1877	.1211	.0867	.0725	.0403	.0207	.0114	.0098
	2	.0042	.0746	.1937	.2759	.3020	.2816	.2335	.1951	.1757	.1209	.0763	.0495	.0439
	3	.0001	.0105	.0574	.1298	.2013	.2503	.2668	.2601	.2522	.2150	.1665	.1267	.1172
	4	.0000	.0010	.0112	.0401	.0881	.1460	.2001	.2276	.2377	.2508	.2384	.2130	.2051
	5	.0000	.0001	.0015	.0085	.0264	.0584	.1029	.1366	.1536	.2007	.2340	.2456	.2461
	6	.0000	.0000	.0001	.0012	.0055	.0162	.0368	.0569	.0689	.1115	.1596	.1966	.2051
	7	.0000	.0000	.0000	.0001	.0008	.0031	.0090	.0163	.0212	.0425	.0746	.1080	.1172
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0014	.0030	.0043	.0106	.0229	.0389	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0005	.0016	.0042	.0083	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0010

TABLA II

DISTRIBUCION DE POISSON

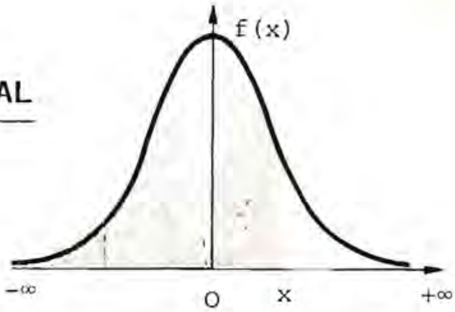
$$p(X = r) = e^{-\lambda} \frac{\lambda^r}{r!}$$

λ \ r	0	1	2	3	4	5	6	7	8	9	10	11	12
.1	.9048	.0905	.0045	.0002	.0000								
.2	.8187	.1637	.0164	.0011	.0001	.0000							
.3	.7408	.2222	.0333	.0033	.0002	.0000							
.4	.6703	.2681	.0536	.0072	.0007	.0001	.0000						
.5	.6065	.3033	.0758	.0126	.0016	.0002	.0000						
.6	.5488	.3293	.0988	.0198	.0030	.0004	.0000						
.7	.4966	.3476	.1217	.0284	.0050	.0007	.0001	.0000					
.8	.4493	.3595	.1438	.0383	.0077	.0012	.0002	.0000					
.9	.4066	.3659	.1647	.0494	.0111	.0020	.0003	.0000					
1.0	.3679	.3679	.1839	.0613	.0153	.0031	.0005	.0001	.0000				
1.1	.3329	.3662	.2014	.0738	.0203	.0045	.0008	.0001	.0000				
1.2	.3012	.3614	.2169	.0867	.0260	.0062	.0012	.0002	.0000				
1.3	.2725	.3543	.2303	.0998	.0324	.0084	.0018	.0003	.0001	.0000			
1.4	.2466	.3452	.2417	.1128	.0395	.0111	.0026	.0005	.0001	.0000			
1.5	.2231	.3347	.2510	.1255	.0471	.0141	.0035	.0008	.0001	.0000			
1.6	.2019	.3230	.2584	.1378	.0551	.0176	.0047	.0011	.0002	.0000			
1.7	.1827	.3106	.2640	.1496	.0636	.0216	.0061	.0015	.0003	.0001	.0000		
1.8	.1653	.2975	.2678	.1607	.0723	.0260	.0078	.0020	.0005	.0001	.0000		
1.9	.1496	.2842	.2700	.1710	.0812	.0309	.0098	.0027	.0006	.0001	.0000		
2.0	.1353	.2707	.2707	.1804	.0902	.0361	.0120	.0034	.0009	.0002	.0000		
2.2	.1108	.2438	.2681	.1966	.1082	.0476	.0174	.0055	.0015	.0004	.0001	.0000	
2.4	.0907	.2177	.2613	.2090	.1254	.0602	.0241	.0083	.0025	.0007	.0002	.0000	
2.6	.0743	.1931	.2510	.2176	.1414	.0735	.0319	.0118	.0038	.0011	.0003	.0001	.0000
2.8	.0608	.1703	.2384	.2225	.1557	.0872	.0407	.0163	.0057	.0018	.0005	.0001	.0000
3.0	.0498	.1494	.2240	.2240	.1680	.1008	.0504	.0216	.0081	.0027	.0008	.0002	.0001
3.2	.0408	.1304	.2087	.2226	.1781	.1140	.0608	.0278	.0111	.0040	.0013	.0004	.0001
3.4	.0334	.1135	.1929	.2186	.1858	.1264	.0716	.0348	.0148	.0056	.0019	.0006	.0002
3.6	.0273	.0984	.1771	.2125	.1912	.1377	.0826	.0425	.0191	.0076	.0028	.0009	.0003
3.8	.0224	.0850	.1615	.2046	.1944	.1477	.0936	.0508	.0241	.0102	.0039	.0013	.0004
4.0	.0183	.0733	.1465	.1954	.1954	.1563	.1042	.0595	.0298	.0132	.0053	.0019	.0006
5.0	.0067	.0337	.0842	.1404	.1755	.1755	.1462	.1044	.0653	.0363	.0181	.0082	.0034
6.0	.0025	.0149	.0446	.0892	.1339	.1606	.1606	.1377	.1033	.0688	.0413	.0225	.0113
7.0	.0009	.0064	.0223	.0521	.0912	.1277	.1490	1490	.1304	.1014	.0710	.0452	.0264
8.0	.0003	.0027	.0107	.0286	.0573	.0916	.1221	.1396	.1396	.1241	.0993	.0722	.0481
9.0	.0001	.0011	.0050	.0150	.0337	.0607	.0911	.1171	.1318	.1381	.1186	.0970	.0728
10.0	.0000	.0005	.0023	.0076	.0189	.0378	.0631	.0901	.1126	.1251	.1251	.1137	.0948

λ \ r	13	14	15	16	17	18	19	20	21	22	23	24
5.0	.0013	.0005	.0002									
6.0	.0052	.0022	.0009	.0003	.0001							
7.0	.0142	.0071	.0033	.0014	.0006	.0002	.0001					
8.0	.0296	.0169	.0090	.0045	.0021	.0009	.0004	.0002	.0001			
9.0	.0504	.0324	.0194	.0109	.0058	.0029	.0014	.0006	.0003	.0001		
10.0	.0729	.0521	.0347	.0217	.0128	.0071	.0037	.0019	.0009	.0004	.0002	.0001

TABLA III

DISTRIBUCION NORMAL



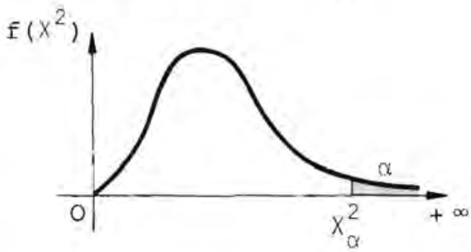
$$F(x) = p(X \leqslant x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}\, dt.$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

TABLA IV

DISTRIBUCION χ^2 DE PEARSON

$$p(\chi^2 \geqslant \chi^2_\alpha) = \alpha$$



n \ α	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010
1	0,031 6	0,039 8	0,023 9	0,015 8	2,71	3,84	5,02	6,63
2	0,02	0,05	0,10	0,21	4,60	5,99	7,38	9,21
3	0,12	0,22	0,35	0,58	6,25	7,81	9,35	11,24
4	0,30	0,48	0,71	1,06	7,78	9,49	11,1	13,28
5	0,55	0,83	1,15	1,61	9,24	11,07	12,8	15,09
6	0,87	1,24	1,64	2,20	10,64	12,59	14,0	16,81
7	1,24	1,69	2,17	2,83	12,02	14,07	16,0	18,47
8	1,65	2,18	2,73	3,49	13,36	15,51	17,5	20,09
9	2,09	2,70	3,33	4,17	14,68	16,92	19,0	21,66
10	2,56	3,25	3,94	4,86	15,99	18,31	20,5	23,21
11	3,05	3,82	4,57	5,58	17,27	19,67	21,9	24,72
12	3,57	4,40	5,23	6,30	18,55	21,03	23,3	26,22
13	4,11	5,01	5,89	7,04	19,81	22,36	24,7	27,69
14	4,66	5,63	6,57	7,79	21,06	23,68	26,1	29,14
15	5,23	6,26	7,26	8,55	22,31	25,00	27,5	30,58
16	5,81	6,81	7,96	9,31	23,54	26,30	28,8	32,00
17	6,41	7,56	8,87	10,08	24,77	27,59	30,2	33,41
18	7,01	8,23	9,39	10,86	25,99	28,87	31,3	34,80
19	7,63	8,91	10,1	11,65	27,20	30,14	32,9	36,19
20	8,26	9,59	10,9	12,44	28,41	31,41	34,2	37,57
21	8,90	10,3	11,6	13,24	29,61	32,67	35,5	38,93
22	9,54	11,0	12,3	14,04	30,81	33,92	36,8	40,29
23	10,2	11,7	13,1	14,85	32,01	35,17	38,1	41,64
24	10,9	12,4	13,8	15,66	33,20	36,41	39,4	42,98
25	11,5	13,1	14,6	16,47	34,38	37,65	40,6	44,31
26	12,2	13,8	15,4	17,29	35,56	38,88	41,9	45,64
27	12,9	14,6	16,2	18,11	36,74	40,11	43,2	46,96
28	13,6	15,3	16,9	18,94	37,92	41,34	44,5	48,28
29	14,3	16,0	17,7	19,77	39,09	42,56	45,7	49,59
30	15,0	16,8	18,5	20,60	40,26	43,77	47,0	50,89

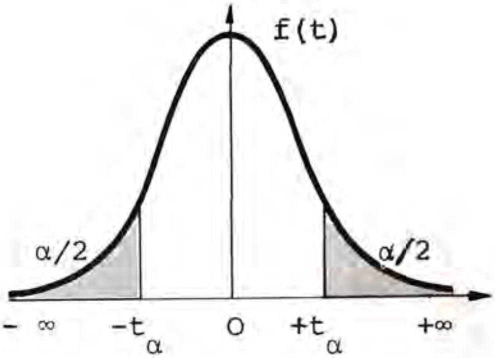
Cuando n es superior a 30, se utiliza la tabla de la distribución N(0 , 1) (tabla III), con:

$$t = \sqrt{2\chi^2} - \sqrt{2n-1}$$

TABLA V

DISTRIBUCION t DE STUDENT

$p(t \geqslant t_\alpha) = \alpha$

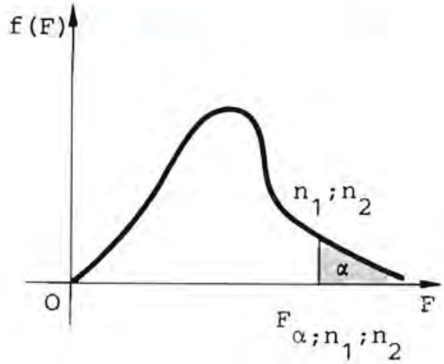


n \ α	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,957
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,127	0,257	0,391	0,533	0,683	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
> 30	0,125 66	0,253 35	0,385 32	0,524 40	0,674 49	0,841 62	1,036 43	1,281 55	1,644 85	1,959 96	2,326 34	2,575 82

TABLA VI

DISTRIBUCION F DE FISHER – SNEDECOR

$p(F \geqslant F_{\alpha}) = \alpha$



n_2 \ n_1	1		2		3		4		5	
	α=0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
1	161,4	4 052	199,5	4 999	215,7	5 403	224,8	5 625	230,2	5 764
2	18,51	98,49	19,00	99,00	19,16	99,17	19,25	99,25	19,30	99,30
3	10,13	34,12	9,55	30,81	9,28	29,46	9,12	28,71	9,01	28,24
4	7,71	21,20	6,94	18,00	6,59	16,69	6,39	15,98	6,26	15,52
5	6,61	16,26	5,79	13.27	5,41	12,06	5,19	11,39	5,05	10,97
6	5,99	13,74	5,14	10,91	4,76	9,78	4,53	9,15	4,39	8,75
7	5,59	12,25	4,74	9,55	4,35	8,45	4,12	7,85	3,97	7,45
8	5,32	11,26	4,46	8,65	4,07	7,59	3,84	7,01	3,69	6,63
9	5,12	10,56	4,26	8,02	3,86	6,99	3,63	6,42	3,48	6,06
10	4,96	10,04	4,10	7,56	3,71	6,55	3,48	5,99	3,33	5,64
11	4,84	9,65	3,98	7,20	3,59	6,22	3,36	5,67	3,20	5,32
12	4,75	9,33	3,88	6,93	3,49	5,95	3,26	5,41	3,11	5,06
13	4,67	9,07	3,80	6,70	3,41	5,74	3,18	5,20	3,02	4,86
14	4,60	8,86	3,74	6,51	3,34	5,56	3,11	5,03	2,96	4,69
15	4,54	8,68	3,68	6,36	3,29	5,42	3,06	4,89	2,90	4,56
16	4,49	8,53	3,63	6,23	3,24	5,29	3,01	4,77	2,85	4,44
17	4,45	8,40	3,59	6,11	3,20	5,18	2,96	4,67	2,81	4,34
18	4,41	8,28	3,55	6,01	3,16	5,09	2,93	4,58	2,77	4,25
19	4,38	8,18	3,52	5,93	3,13	5,01	2,90	4,50	2,74	4,17
20	4,35	8,10	3,49	5,85	3,10	4,94	2,87	4,43	2,71	4,10
21	4,32	8,02	3,47	5,78	3,07	4,87	2,84	4,37	2,68	4,04
22	4,30	7,94	3,44	5,72	3,05	4,82	2,82	4,31	2,66	3,99
23	4,28	7,88	3,42	5,66	3,03	4,76	2,80	4,26	2,64	3,94
24	4,26	7,82	3,40	5,61	3,01	4,72	2,78	4,22	2,62	3,90
25	4,24	7,77	3,38	5,57	2,99	4,68	2,76	4,18	2,60	3,86
26	4,22	7,72	3,37	5,53	2,98	4,64	2,74	4,14	2,59	3,82
27	4,21	7,68	3,35	5,49	2,96	4,60	2,73	4,11	2,57	3,78
28	4.20	7,64	3,34	5,45	2,95	4,57	2,71	4,07	2,56	3,75
29	4,18	7,60	3,33	5,42	2,93	4,54	2,70	4,04	2,54	3,73
30	4,17	7,56	3,32	5,39	2,92	4,51	2,69	4,02	2,53	3,70
40	4,08	7,31	3,23	5,18	2,84	4,31	2,61	3,83	2,45	3,51
60	4,00	7,08	3,15	4,98	2,76	4,13	2,52	3,65	2,37	3,34
120	3,92	6,85	3,07	4,79	2,68	3,95	2,45	3,48	2,29	3,17
∞	3,84	6,64	2,99	4,60	2,60	3,78	2,37	3,32	2,21	3,02

TABLA VI (Continuación)

DISTRIBUCION F DE FISHER - SNEDECOR

n_2 \ n_1	6		8		12		24		∞	
	α=0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
1	234,0	5 859	238,9	5 981	243,9	6 106	249,0	6 234	254,3	6 366
2	19,33	99,33	19,37	99,36	19,41	99,42	19,45	99,46	19,50	99,50
3	8,94	27,91	8,84	27,49	8,74	27,05	8,64	26,60	8,53	26,12
4	6,16	15,21	6,04	14,80	5,91	14,37	5,77	13,93	5,63	13,46
5	4,95	10,67	4,82	10,27	4,68	9,89	4,53	9,47	4,36	9,02
6	4,28	8,47	4,15	8,10	4,00	7,72	3,84	7,31	3,67	6,88
7	3,87	7,19	3,73	6,84	3,57	6,47	3,41	6,07	3,23	5,65
8	3,58	6,37	3,44	6,03	3,28	5,67	3,12	5,28	2,93	4,86
9	3,37	5,80	3,23	5,47	3,07	5,11	2,90	4,73	2,71	4,31
10	3,22	5,39	3,07	5,06	2,91	4,71	2,74	4,33	2,54	3,91
11	3,09	5,07	2,95	4,74	2,79	4,40	2,61	4,02	2,40	3,60
12	3,00	4,82	2,85	4,50	2,69	4,16	2,50	3,78	2,30	3,36
13	2,92	4,62	2,77	4,30	2,60	3,96	2,42	3,59	2,21	3,16
14	2,85	4,46	2,70	4,14	2,53	3,80	2,35	3,43	2,13	3,00
15	2,79	4,32	2,64	4,00	2,48	3,67	2,29	3,29	2,07	2,87
16	2,74	4,20	2,59	3,89	2,42	3,55	2,24	3,18	2,01	2,75
17	2,70	4,10	2,55	3,79	2,38	3,45	2,19	3,08	1,96	2,65
18	2,66	4,01	2,51	3,71	2,34	3,37	2,15	3,00	1,92	2,57
19	2,63	3,94	2,48	3,63	2,31	3,30	2,11	2,92	1,88	2,49
20	2,60	3,87	2,45	3,56	2,28	3,23	2,08	2,86	1,84	2,42
21	2,57	3,81	2,42	3,51	2,25	3,17	2,05	2,80	1,81	2,36
22	2,55	3,76	2,40	2,45	2,23	3,12	2,03	2,75	1,78	2,31
23	2,53	3,71	2,38	3,41	2,20	3,07	2,00	2,70	1,76	2,26
24	2,51	3,67	2,36	3,36	2,18	3,03	1,98	2,66	1,73	2,21
25	2,49	3,63	2,34	3,32	2,16	2,99	1,96	2,62	1,71	2,17
26	2,47	3,59	2,32	3,29	2,15	2,96	1,95	2,58	1,69	2,13
27	2,46	3,56	2,30	3,26	2,13	2,93	1,93	2,55	1,67	2,10
28	2,44.	3,53	2,29	3,23	2,12	2,90	1,91	2,52	1,65	2,06
29	2,43	3,50	2,28	3,20	2,10	2,87	1,90	2,49	1,64	2,03
30	2,42	3,47	2,27	3,17	2,09	2,84	1,89	2,47	1,62	2,01
40	2,34	3,29	2,18	2,99	2,00	2,66	1,79	2,29	1,51	1,80
60	2,25	3,12	2,10	2,82	1,92	2,50	1,70	2,12	1,39	1,60
120	2,17	2,96	2,01	2,66	1,83	2,34	1,61	1,95	1,25	1,38
∞	2,09	2,80	1,94	2,51	1,75	2,18	1,52	1,79	1.00	1.00