

CAPÍTULO I

Estadística Descriptiva

María Margarita Olivares M.

Abril 2004

1 INTRODUCCIÓN:

Si estamos interesados en conocer alguna característica de una población (ó conjunto de individuos u objetos) acerca de la que se quiere saber algún aspecto claramente definido, lo más completo sería estudiar la población entera. Pero este procedimiento requiere mucho tiempo y resulta muy costoso, así que normalmente nos conformamos con el conocimiento parcial de la población ó muestra, que elegida adecuadamente sea representativa de ésta.

El objetivo de la estadística es hacer inferencia (ó tomar decisiones ó hacer predicciones) acerca de una población, basándose en la información contenida en una muestra. Es decir, integrando el cálculo de probabilidades (que nace en el siglo XVII como teoría matemática de los juegos de azar) y la Estadística Descriptiva o ciencia del estado, del latín Status, que estudia la descripción de datos y tiene raíces más antiguas, se obtiene una ciencia (Estadística Matemática) que estudia cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos.

La estadística actúa como puente entre los modelos matemáticos y los fenómenos reales. Un modelo matemático es una abstracción simplificada de una realidad más compleja y siempre existirá cierta discrepancia entre lo observado y lo previsto por el modelo. La Estadística nos proporciona un método para evaluar estas discrepancias entre la realidad y la teoría. Su estudio es básico para todos aquellos que quieran trabajar en ciencia aplicada (economía, sociología, etc.).

En nuestra era cada aspecto de la actividad humana es medido e interpretado en términos estadísticos. El conocimiento básico de los métodos

estadísticos nos permitirá participar en los argumentos públicos basados en cifras y datos por lo que es un buen antídoto ante posibles manipulaciones.

Hay cinco elementos fundamentales en todo problema estadístico:

1. Definir claramente la pregunta que se desea responder acerca de la población.
2. Procedimiento de muestreo ó diseño del experimento.
3. Recolección y análisis de datos.
4. Hacer inferencia acerca de la población mediante una probabilidad.
5. Confiabilidad de la inferencia ó bondad del ajuste.

Cuando planteamos con claridad la pregunta que queremos responder acerca de la población, procedemos a la recolección de datos numéricos relacionados con el estudio que queremos hacer.

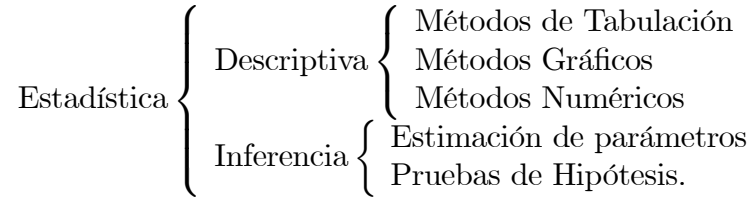
2 Estadística Descriptiva e Inferencia:

Una vez obtenido los datos debemos organizarlos, lo que se hace siguiendo ciertos métodos que constituyen la “Estadística Descriptiva”. Los métodos comúnmente usados son de tres tipos: Métodos de Tabulación, Métodos Gráficos y Métodos Numéricos.

Los primeros de ellos se constituyen a partir de la elaboración de tablas que incluyen los datos numéricos. Los métodos gráficos exigen la elaboración de gráficos, entre los cuales los más usados son los de barras, los circulares e histogramas. Los métodos numéricos consisten en obtener ciertas relaciones cuantitativas a partir de los datos.

Una vez realizado el estudio de los datos mediante los métodos de la Estadística Descriptiva, se trata entonces de inferir o sacar conclusiones sobre algunos aspectos de la población, que generalmente se refiere a la confirmación de alguna hipótesis, (prueba de hipótesis) o a la estimación de algún promedio numérico u otras características de la población (estimación de parámetros). Esta parte constituye lo que se conoce con el nombre de Estadística Inferencial o Inferencia Estadística.

El siguiente diagrama resume lo expuesto anteriormente:



3 CONCEPTOS BÁSICOS.

3.1 Estadística Descriptiva: Una Variable

Supongamos que tenemos una fuente de material radioactivo que emite partículas Alfa (α) y que definimos la variable aleatoria X como el número de partículas observadas en una pantalla, en un intervalo de tiempo t . Bajo ciertas hipótesis que idealizan el experimento, X tiene una distribución de Poisson de parámetro λt .

Si queremos calcular, por ejemplo, la probabilidad de que X sea mayor que 10 u otras características asociadas con la distribución tales como la esperanza, la varianza, etc., la respuesta dependerá del parámetro λ y del intervalo de tiempo t .

Para buscar un valor numérico de λ , dejamos el mundo de los modelos matemáticos teóricos y entramos en el mundo de las observaciones, es decir, observamos la emisión de partículas, obtenemos algunos valores numéricos de X y luego los utilizamos de alguna manera, a fin de obtener una información atinada del parámetro λ .

En general, un material estadístico que consiste en cierto número de observaciones

$$x_1, x_2, \dots, x_N$$

de una variable aleatoria X , dado en la forma original, en la que los N resultados aparecen en el orden en que se han observado, es muy difícil de examinar y por lo tanto no es adecuado para darnos información acerca de la variable X investigada.

El propósito de la Estadística Descriptiva es reemplazar el material observado por cantidades relativamente pocas en número, que representen el material total ó en otras palabras, que contenga tanta información como sea posible respecto a la variable X .

Tipos de variables:

Los tipos de variables que consideraremos, son:

1. Variables cualitativas o atributos: no toman valores numéricos y describen cualidades. Por ejemplo, clasificar una pieza como aceptable o defectuosa.
2. Variables cuantitativas discretas: toman sólo valores enteros, en muchos casos se limita a contar el número de veces que ocurre un suceso. Por ejemplo, número de compras de un producto en un mes.
3. Variables cuantitativas continuas: toman valores en un intervalo, corresponde a medir magnitudes continuas. Por ejemplo, tiempo entre la llegada de dos autobuses.

3.1.1 MUESTRA OBSERVADA:

Sea X una variable aleatoria asociada a cierto experimento. Si realizamos N veces el experimento, de manera independiente y bajo las mismas condiciones, obtendremos N valores numéricos, en caso de variables cuantitativas:

$$x_1, x_2, \dots, x_N$$

correspondientes a la variable aleatoria X . A estos resultados obtenidos se les llama muestra observada.

Cuando esta muestra no se somete a ninguna ordenación especial, se le denomina muestra bruta.

3.1.2 RECOPIACIÓN DE DATOS: Tablas de Frecuencias.

Los valores observados se suelen registrar en una lista. Si el número de observaciones no excede 20 o 30, por ejemplo, es posible darse una idea aproximada de la distribución, simplemente mediante la ordenación de los valores observados, escribiéndolos en una tabla, en orden creciente de magnitud. Con estos datos podemos hacer representaciones gráficas y calcular determinadas características numéricas.

Si el conjunto de datos es muy grande, resulta laborioso trabajar directamente con los valores individuales observados y entonces se lleva a cabo algún tipo de agrupación, como paso preliminar, antes de iniciar un nuevo tratamiento de los datos.

El procedimiento de agrupación es diferente según la variable aleatoria sea discreta o continua.

La presentación de los datos en forma agrupada implica alguna pérdida de información, pero permite apreciar mejor sus características.

Este agrupamiento se hace mediante las llamadas tablas de frecuencias. En la tablas de frecuencias, en lugar de mostrar individualmente todos los datos, se informa solamente cuántos de ellos están comprendidos entre determinados valores, llamados límites de clase. Las clases son intervalos cuyos extremos son los límites de clase. Generalmente las clases se escogen de igual longitud.

Una regla empírica que se suele aplicar consiste en escoger los intervalos de clase de tal forma que no haya menos de 10 ni más de 20 clases diferentes.

Veamos cómo se lleva a cabo la agrupación en cada uno de los casos:

1. **Caso Discreto:**

En este caso resulta conveniente hacer una tabla cuya primera columna contenga todos los valores observados y la segunda contenga la frecuencia con que han aparecido dichos valores o frecuencias absolutas. También se suele añadir una tercera columna que contiene la frecuencia relativa de los datos observados, a saber, la razón entre la frecuencia absoluta y el número total de observaciones.

Este tipo de agrupación se utiliza cuando el número total de valores observados no es muy grande, en caso contrario, en lugar de asignar una clase a cada valor observado, podemos considerar clases que contengan varias observaciones.

Ejemplo: se cuenta el número de glóbulos rojos en cada uno de los 169 compartimientos de un hemocitómetro. Cada uno de los compartimientos representa una observación y el número de glóbulos rojos en cada una de ellos es el valor observado correspondiente. De dicha

observación se obtiene la siguiente tabla de frecuencias:

Nº de Globulos rojos	Frecuencia Absoluta	Frecuencia Relativa
4	1	$\frac{1}{169}$
5	3	$\frac{3}{169}$
6	5	$\frac{5}{169}$
7	8	$\frac{8}{169}$
8	13	$\frac{13}{169}$
9	14	$\frac{14}{169}$
10	15	$\frac{15}{169}$
11	15	$\frac{15}{169}$
12	21	$\frac{21}{169}$
13	18	$\frac{18}{169}$
14	17	$\frac{17}{169}$
15	16	$\frac{16}{169}$
16	9	$\frac{9}{169}$
17	6	$\frac{6}{169}$
18	3	$\frac{3}{169}$
19	2	$\frac{2}{169}$
20	2	$\frac{2}{169}$
21	1	$\frac{1}{169}$
Total	169	1

Con esta información podemos hacer un gráfico por medio del llamado Histograma, que en este caso se elabora levantando una línea o barra sobre cada clase, de altura proporcional a la frecuencia correspondiente o a la frecuencia relativa.

2. Caso Continuo:

En el caso en que la variable aleatoria investigada es continua, la agrupación es algo más complicada, sin embargo, en general, se procede de la siguiente manera: se toma un intervalo adecuado de la recta real que contenga los N valores observados y se divide dicho intervalo en un cierto número de intervalos de clase.

Todas las observaciones que caen en una misma clase se agrupan y se cuentan, el número resultante es la frecuencia de clase correspondiente a dicho intervalo y después se procede a tabular. Para proceder a la elección de los límites de clase debemos conocer la “exactitud” de los

datos originales. Cuando la tabla de frecuencia ya ha sido elaborada debe ir acompañada de la exactitud de los datos.

Ilustraremos el procedimiento mediante algunos ejemplos:

- (a) Se preparan, con la misma mezcla setenta cilindros de concreto y se mide la resistencia a la compresión de cada uno de ellos. Los resultados originales o muestra bruta están dados con cuatro cifras enteras. La siguiente tabla representa la muestra bruta:

2860	3052	2940	3128	2865	3125	2881
2950	2911	2883	3027	2872	2942	3042
3128	2965	3109	2886	3045	3238	2965
3300	3193	2865	3298	2932	2782	3201
2961	2832	2950	3059	3052	3017	3001
3045	2944	3038	2968	2953	2998	3275
3185	3003	3317	2875	2820	2808	2973
2857	2903	2910	2957	2891	2899	2884
3015	3061	3097	3085	2975	3072	3115
3073	3169	3133	3152	3102	3251	2702

Procedamos a tabular estos resultados en una tabla de frecuencias: La diferencia entre el máximo y el mínimo valor observado es

$$3317 - 2702 = 615$$

(esta diferencia se llama rango de la muestra). Vamos a construir una tabla de once clases ($615/11 \simeq 60$), esta decisión es un tanto arbitraria, la longitud común de cada clase será de 60 unidades. Nuestro primer impulso sería tomar como clases los siguientes intervalos:

$$\begin{aligned} &(2700, 2760) \\ &(2760, 2820) \\ &(2820, 2880), \text{ etc.} \end{aligned}$$

pero tomando los límites de esta manera, no sabríamos en qué clase incluir los valores que coinciden con los límites de dichos intervalos, como por ejemplo 2820. Para evitar este tipo de ambigüedad podríamos tomar los siguientes intervalos:

$$\begin{aligned} &(2700, 2759) \\ &(2760, 2819) \\ &(2820, 2879), \text{ etc.} \end{aligned}$$

con esta elección, dejamos un hueco entre 2759 y 2760, etc., pero por la precisión de los datos sabemos que allí no hay observaciones, sin embargo, es preferible elegir como límites exactos de cada clase los puntos correspondientes a medias unidades de la última cifra significativa de los límites anteriores, es decir:

$$\begin{aligned} &(2699.5, 2759.5) \\ &(2759.5, 2819.5) \\ &(2819.5, 2879.5), \text{ etc.} \end{aligned}$$

en este caso estamos seguros de que ninguna observación caerá en un límite de clase.

Clase	Frecuencia Absoluta	Frecuencia relativa
(2699.5, 2759.5)	1	$\frac{1}{70} \simeq 0,0143$
(2759.5, 2819.5)	2	$\frac{2}{70} \simeq 0,0286$
(2819.5, 2879.5)	7	$\frac{7}{70} \simeq 0,1000$
(2879.5, 2939.5)	11	$\frac{11}{70} \simeq 0,1571$
(2939.5, 2999.5)	14	$\frac{14}{70} \simeq 0,2000$
(2999.5, 3059.5)	12	$\frac{12}{70} \simeq 0,1714$
(3059.5, 3119.5)	8	$\frac{8}{70} \simeq 0,1143$
(3119.5, 3179.5)	6	$\frac{6}{70} \simeq 0,0857$
(3179.5, 3239.5)	4	$\frac{4}{70} \simeq 0,0571$
(3239.5, 3299.5)	3	$\frac{3}{70} \simeq 0,0429$
(3299.5, 3359.5)	2	$\frac{2}{70} \simeq 0,0286$
Total	70	1,0000

Esta información puede ser representada gráficamente mediante un histograma, levantando sobre cada clase un rectángulo de altura proporcional a la frecuencia correspondiente o alternatively a la frecuencia relativa. Si se unen con segmentos las alturas de los rectángulos que constituyen el histograma, en las correspondientes marcas de clase, se obtiene una poligonal denominada polígono de frecuencias, el cual puede suavizarse mediante una curva suave.

- (b) Se determina el porcentaje de ceniza en una muestra de carbón, extraída de 250 vagones diferentes. Los datos originales son exactos hasta la segunda cifra decimal, representados en la siguiente

tabla de frecuencias:

Clase(% de ceniza)	Frecuencia Absoluta	Frecuencia Relativa
(9, 9.99)	1	$\frac{1}{250}$
(10, 10.99)	3	$\frac{3}{250}$
(11, 11.99)	3	$\frac{3}{250}$
(12, 12.99)	9	$\frac{9}{250}$
(13, 13.99)	13	$\frac{13}{250}$
(14, 14.99)	27	$\frac{27}{250}$
(15, 15.99)	28	$\frac{28}{250}$
(16, 16.99)	39	$\frac{39}{250}$
(17, 17.99)	42	$\frac{42}{250}$
(18, 18.99)	34	$\frac{34}{250}$
(19, 19.99)	19	$\frac{19}{250}$
(20, 20.99)	14	$\frac{14}{250}$
(21, 21.99)	10	$\frac{10}{250}$
(22, 22.99)	4	$\frac{4}{250}$
(23, 23.99)	3	$\frac{3}{250}$
(24, 24.99)	0	$\frac{0}{250} = 0$
(25, 25.99)	1	$\frac{1}{250}$
Total	250	1

Los datos se agrupan tal como aparecen en la tabla de forma que, por ejemplo, el intervalo de clase (14, 14.99) contenga todas las observaciones registradas con valor de 14 a 14.99, ambos inclusive.

Al agrupar los datos originales, si registramos una observación, por ejemplo, 14.27 con dos cifras decimales exactas, el valor realmente observado se encuentra entre 14.265 y 14.275. Los límites exactos de este intervalo de clase son 13.995 y 14.995. Si los datos hubiesen sido dados con una cifra decimal exacta, los intervalos de clase serían de la forma (14.0, 14.9) con límites exactos 13.95 y 14.95.

Cuando se utilizan los datos ya agrupados, para los cálculos, se supone que todas las observaciones que pertenecen a una clase dada, están situadas en el punto medio de dicha clase. Al hacer esta aproximación, se introduce un error que evidentemente se puede hacer tan pequeño como queramos, tomando los intervalos de clase suficientemente pequeños y reduciendo así la pérdida de información debida a la agrupación. Sin embargo esto aumenta el

largo de la tabla y nos hace perder algo de simplificación que es la razón de la agrupación. Como regla práctica se acostumbra tener un número de clases entre 10 y 20.

3.1.3 FRECUENCIAS ACUMULADAS.

La frecuencia absoluta acumulada es el número de observaciones menores o iguales a una cierta cantidad dada. El cociente entre frecuencia absoluta acumulada y el número de observaciones, es la frecuencia relativa acumulada.

3.1.4 EJEMPLO:

Consideremos la siguiente tabla de frecuencias:

Clases	Frecuencias Absolutas	Frecuencias Relativas
(100, 109.5)	2	$\frac{2}{26}$
(110, 119.5)	1	$\frac{1}{26}$
(120, 129.5)	6	$\frac{6}{26}$
(130, 139.5)	4	$\frac{4}{26}$
(140, 149.5)	6	$\frac{6}{26}$
(150, 159.5)	4	$\frac{4}{26}$
(160, 169.5)	0	$\frac{0}{26} = 0$
(170, 179.5)	1	$\frac{1}{26}$
(180, 189.5)	1	$\frac{1}{26}$
(190, 199.5)	0	0
(200, 209.5)	1	$\frac{1}{26}$
Total	26	1

A partir de ella construimos la tabla de frecuencias acumuladas:

Observaciones $\leq x$	Frec. Absol. Acumulada	Frec. Rel. Acumulada
100	0	0
110	2	$\frac{2}{26}$
120	3	$\frac{3}{26}$
130	9	$\frac{9}{26}$
140	13	$\frac{13}{26}$
150	19	$\frac{19}{26}$
160	23	$\frac{23}{26}$
170	23	$\frac{23}{26}$
180	24	$\frac{24}{26}$
190	25	$\frac{25}{26}$
200	25	$\frac{25}{26}$
210	26	$\frac{26}{26} = 1$

El gráfico resulta ser escalonado y creciente. Esta información se suele representar mediante las ojivas que son curvas equivalentes a polígonos de frecuencias acumuladas, suavizado.

3.1.5 MÉTODOS GRÁFICOS.

Para representar gráficamente los datos, existen, además del histograma, otros tipos de gráficos, tales como gráficos de sectores circulares, gráficos de barras, gráficos de líneas, pictogramas, polígonos de frecuencia u ojiva.

3.2 DESCRIPCIÓN NUMÉRICA DE DATOS.

Corresponde a medidas de centralización o dispersión, estos números ayudan a completar la información obtenida mediante las tabulaciones y gráficos.

Las medidas de centralización más usuales son: la media o promedio, la moda y la mediana.

Las medidas de dispersión más usuales son el rango, la varianza y la desviación estándar.

3.2.1 MEDIA OBSERVADA O PROMEDIO:

Se llama media observada de una muestra al promedio aritmético de las observaciones, es decir, si x_1, x_2, \dots, x_n son las observaciones individuales,

la media observada será:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

Esta fórmula sólo es aplicable cuando se ha conservado la muestra bruta. Si hemos perdido los datos originales y disponemos solamente de una tabla de frecuencias, identificamos todas las observaciones correspondientes a una clase con un valor único, denominado marca de clase, (en general se toma como marca de clase el punto medio del intervalo de clase); es decir, si y_i es la marca de de la i -ésima clase, f_i la frecuencia de esta clase y $\phi_i = \frac{f_i}{N}$ la frecuencia relativa de esta clase, podemos calcular la media observada mediante la fórmula:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^M y_i f_i = \sum_{i=1}^M y_i \phi_i.$$

donde M es el número de clases. Para simplificar los cálculos, puesto que de todas formas se trata de una aproximación, es permisible, cuando se han adoptado intervalos disjuntos, hacer coincidir los límites adyacentes. Por ejemplo, en los intervalos

$$\begin{aligned} &(2700, 2759) \\ &(2760, 2819) \end{aligned}$$

podemos tomar como marca de clase

$$\frac{2700 + 2760}{2} = 2730$$

en lugar de

$$\frac{2700 + 2759}{2} = 2729.5$$

3.2.2 VARIANZA Y DESVIACIÓN ESTÁNDAR OBSERVADA.

Se llama varianza observada de una muestra x_1, x_2, \dots, x_n al valor

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Algunas veces se prefiere trabajar con la varianza centrada, (como veremos más adelante tiene buenas propiedades), definida como

$$s_1^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

Note que

$$s^2 = \frac{N-1}{N} s_1^2$$

así, si $N \rightarrow \infty$, $s^2 = s_1^2$.

La desviación estándar observada es s y la centrada es s_1 . Cuando no disponemos de la muestra bruta y en su lugar contamos con la tabla de frecuencias, calculamos las varianzas de manera análoga al caso de la media, mediante las fórmulas:

$$\begin{aligned} s^2 &= \frac{1}{N} \sum_{i=1}^M f_i (y_i - \bar{x})^2 = \sum_{i=1}^M \phi_i (y_i - \bar{x})^2 \\ s_1^2 &= \frac{1}{N-1} \sum_{i=1}^M f_i (y_i - \bar{x})^2 = \frac{N}{N-1} \sum_{i=1}^M \phi_i (y_i - \bar{x})^2 \end{aligned}$$

donde M es el número de clases. La varianza y la desviación estándar miden el grado de dispersión de los datos alrededor de la media.

Para el cálculo de la varianza y la desviación estándar las siguientes fórmulas son útiles, éstas se obtienen fácilmente y se dejan como ejercicio:

$$\begin{aligned} s^2 &= \frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{x}^2 \\ s_1^2 &= \frac{1}{N-1} \left(\sum_{i=1}^N x_i^2 - N\bar{x}^2 \right) \end{aligned}$$

Propiedad de la Distribución Normal Estándar: Si Z tiene distribución $N(0, 1)$,

$$\mathbb{P}(-1 < Z < 1) = 0.6826$$

es decir, el 68,26% de las observaciones caen en el intervalo $(-1, 1)$, de manera análoga se tiene que el 95,44% de las observaciones caen en el intervalo $(-2, 2)$ y el 99,74% de las observaciones caen en el intervalo $(-3, 3)$.

Si la distribución de X es $N(\mu, \sigma)$, se obtiene que el 68,26% de las observaciones caen en el intervalo $(\mu - \sigma, \mu + \sigma)$, el 95,44% de las observaciones

caen en el intervalo $(\mu - 2\sigma, \mu + 2\sigma)$ y el 99,74% de las observaciones caen en el intervalo $(\mu - 3\sigma, \mu + 3\sigma)$.

De esta propiedad se obtiene la llamada regla empírica que se verifica en los casos en que el histograma correspondiente a las observaciones tiene forma de campana:

La Regla Empírica:

Dada una distribución de observaciones que es aproximadamente acampanada, el intervalo

1. $(\mu - \sigma, \mu + \sigma)$ contiene aproximadamente el 68% de las observaciones
2. $(\mu - 2\sigma, \mu + 2\sigma)$ contiene aproximadamente el 95% de las observaciones
3. $(\mu - 3\sigma, \mu + 3\sigma)$ contiene casi todas las observaciones.

EJEMPLOS:

1. Si observamos la tabla de datos correspondiente a la medida de la resistencia a la compresión de 70 cilindros de concreto, obtendremos:

$$\begin{aligned}\bar{x} &= 3010,8857 \\ s &= 133,84112 \\ s_1 &= 134,80794\end{aligned}$$

\bar{x} es una buena medida de la media o centro de los datos ya que 37 observaciones son menores que \bar{x} y 33 mayores. En el intervalo

$$(\bar{x} - s, \bar{x} + s) = (2887,04458; 3144,72682)$$

se encuentran 44 observaciones de las 70 observadas. En este caso \bar{x} y s describen los datos adecuadamente.

2. La siguiente tabla o muestra bruta, representa el ingreso anual en miles de dólares de 42 familias en un pueblo de E.E.U.U. (1977).

1,2	17,0	23,2	36,0	74,7	152,2	19,6
29,3	8,2	20,6	20,1	8,8	10,7	26,0
11,6	39,4	157,4	10,3	16,2	100,2	37,7
14,5	151,2	10,1	92,3	7,7	47,6	29,0
26,8	8,2	25,8	8,0	19,4	21,2	150,1
28,1	17,8	26,8	17,8	19,3	37,2	13,4

En la siguiente tabla de frecuencia correspondiente a esta muestra, se observa enseguida que el comportamiento de los datos es mucho más errático que en la tabla anterior:

Clases(\$)	Frecuencia Absoluta	Frecuencia Relativa
(100, 10000)	6	$\frac{6}{42} = 1/7 = 0,1429$
(10000, 20000)	13	$13/42 = 0,3095$
(20000, 30000)	11	$11/42 = 0,2619$
(30000, 50000)	5	$5/42 = 0,1190$
(50000, 160000)	7	$7/42 = 0,1667$

Si se representan estos datos en un histograma de frecuencias observarán que no es simétrico alrededor de ningún punto ya que tiene una cola larga hacia la derecha (sesgado hacia la derecha). Para estos datos

$$\begin{aligned}\bar{x} &= 37,28\$ \\ s &= 41,35\end{aligned}$$

este promedio no es un valor particularmente típico, de hecho, 32 de los 42 datos son menores que \bar{x} y sólo 10 son mayores, es decir, \bar{x} no es una buena medida de centramiento; el histograma tiene este gran sesgo a la derecha, (empuje del promedio a la derecha) de tal manera que 75% de las observaciones quedan a la izquierda del promedio. La diferencia grande entre los datos ejerce una gran influencia en el valor del promedio y lo hacen tener un valor no centrado, al igual que hace crecer la desviación estándar.

En resumen, para datos fuertemente sesgados (a la derecha o a la izquierda) \bar{x} , s ó s_1 pueden no ser los parámetros que describan el centro y dispersión de los datos, en este caso es conveniente definir otras medidas de centramiento.

3.2.3 RANGO DE LA MUESTRA.

Si

$$x_1, x_2, x_3, \dots, x_N$$

una muestra observada, definimos el rango de esta muestra como la diferencia entre la mayor y la menor de las observaciones:

$$R = \max_{1 \leq i \leq N} x_i - \min_{1 \leq i \leq N} x_i$$

3.2.4 Coeficiente de Variación o coeficiente de dispersión de la muestra:

El coeficiente de dispersión de la muestra observada expresa la magnitud de la dispersión con respecto a su media:

$$\frac{s}{\bar{x}} \text{ o alternativamente } \frac{s_1}{\bar{x}}$$

3.2.5 Momentos de orden n de una muestra observada:

Si

$$x_1, x_2, x_3, \dots, x_N$$

una muestra observada, definimos el momento de orden n de esta muestra como:

$$M_n = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^n$$

y si disponemos únicamente de la tabla de frecuencias:

$$M_n = \frac{1}{N} \sum_{i=1}^M f_i (y_i - \bar{x})^n = \sum_{i=1}^M \phi_i (y_i - \bar{x})^n$$

donde M es el número de clases, f_i es la frecuencia absoluta y ϕ_i la frecuencia relativa de la clase i .

3.2.6 Moda de la Muestra Observada.

Si

$$x_1, x_2, x_3, \dots, x_N$$

una muestra observada, llamamos moda de la muestra al valor que se presenta con mayor frecuencia. Si disponemos solamente de una tabla de frecuencias, tomaremos como moda el punto medio del intervalo de clase de mayor frecuencia.

3.2.7 Mediana de la Muestra Observada.

Si

$$x_1, x_2, x_3, \dots, x_N$$

una muestra observada, representamos por

$$x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(N)}$$

el mismo conjunto de datos ordenados de mayor a menor, es decir:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(N)}$$

Una nueva medida del centro del conjunto de datos está dada por la mediana m que es el valor central o promedio de los valores centrales de la muestra ordenada, es decir:

$$m = \begin{cases} x_{(\frac{N+1}{2})} & \text{si } N \text{ es impar} \\ \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} & \text{si } N \text{ es par} \end{cases}$$

Si ordenamos la tabla de sueldos del ejemplo anterior, de menor a mayor, como $n = 42$,

$$m = \frac{x_{(21)} + x_{(22)}}{2} = \frac{20,6 + 21,2}{2} = 20,9$$

m es una mejor medida del centro de los datos cuando estos son sesgados hacia un lado, m tiene la propiedad de que prácticamente la mitad de los datos está por debajo de m y la mitad por encima, de modo que en este sentido es una buena representación del centro.

Geométricamente, la mediana es el valor de la abcisa que corresponde a la vertical que divide el área encerrada por un histograma en dos partes iguales.

OBSERVACIÓN: La mediana de una variable aleatoria X es el punto $x \in \mathbb{R}$ tal que

$$\mathbb{P}(X > x) = \mathbb{P}(X \leq x) = \frac{1}{2}.$$

Cuando los datos están agrupados en una tabla de frecuencias, podemos calcular aproximadamente la mediana de la muestra observada mediante un método que describiremos a continuación (éste no es el único método, distintos métodos nos llevan a resultados diferentes, pero todos son valores aproximados de la mediana): Sea N el número de observaciones

1. Elegimos un intervalo de clase $[e_k, e_{k+1}]$ tale que

$$\sum_{i=1}^{k-1} f_i \leq \frac{N}{2}, \sum_{i=1}^k f_i > \frac{N}{2}$$

donde f_i representa la frecuencia del intervalo $[e_k, e_{k+1}]$.

2. Supondremos que las observaciones que caen en el intervalo $[e_k, e_{k+1}]$, están uniformemente distribuídas en dicho intervalo, es decir, si f_k es el número de observaciones en dicho intervalo y lo subdividimos en f_k subintervalos de igual longitud

$$L_k = \frac{e_{k+1} - e_k}{f_k}$$

supondremos que en cada subdivisión hay una sola observación:

- (a) Si N es impar nos gustaría aproximar la mediana por la observación que ocupa el lugar $\frac{N+1}{2}$, cuando la muestra bruta se ordena de menor a mayor, entonces añadimos a $\sum_{i=1}^{k-1} f_i$ la cantidad que falta para obtener $\frac{N+1}{2}$, es decir, hallamos k_0 tal que

$$\sum_{i=1}^{k-1} f_i + k_0 = \frac{N+1}{2}.$$

Por definición de $[e_k, e_{k+1}]$, $1 \leq k_0 \leq f_k$, entonces aproximamos la observación $\frac{N+1}{2}$ por un número en el intervalo

$$\left[e_k + (k_0 - 1) \frac{e_{k+1} - e_k}{f_k}, e_k + k_0 \frac{e_{k+1} - e_k}{f_k} \right].$$

- (b) Si N es par aproximamos la mediana por la observación que ocupa el lugar $\frac{N}{2} + 1$, cuando la muestra bruta se ordena de menor a mayor, para ello elegimos k_0 tal que

$$\sum_{i=1}^{k-1} f_i + k_0 = \frac{N}{2} + 1$$

Por definición de $[e_k, e_{k+1}]$, $1 \leq k_0 \leq f_k$, entonces aproximamos la mediana por un número en el intervalo

$$\left[e_k + (k_0 - 1) \frac{e_{k+1} - e_k}{f_k}, e_k + k_0 \frac{e_{k+1} - e_k}{f_k} \right]$$

Podríamos también aproximar la mediana por el valor

$$\frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} \cong e_k + (k_0 - 1) \frac{e_{k+1} - e_k}{f_k}$$

Observe que

$$\left[\frac{N}{2} + 1 \right] = \begin{cases} \frac{N}{2} + 1, & \text{si } N \text{ es par} \\ \frac{N+1}{2}, & \text{si } N \text{ es impar} \end{cases}$$

donde $[x]$ es parte entera de x .

EJEMPLO: Calculemos la mediana correspondiente a la tabla de frecuencias que describe el porcentaje de ceniza en una muestra de carbón:

$$N = 250, \frac{N}{2} = 125$$

Para hallar k tal que

$$\sum_{i=1}^{k-1} f_i \leq \frac{N}{2} = 125, \sum_{i=1}^k f_i > \frac{N}{2} = 125$$

acudimos a la tabla y sumamos las frecuencias de los intervalos de clase comenzando por el primero:

$$1 + 3 + 3 + 9 + 13 + 27 + 28 + 39 = 123 < \frac{N}{2}, 123 + 42 = 165 > \frac{N}{2} = 125$$

Luego, el intervalo donde se encuentra la mediana es $(17, 17.99)$, para simplificar los cálculos podemos considerar los límites exactos del intervalo, es decir, $(16.99, 17.99)$ cuya longitud es 1 :

$$e_{k+1} - e_k = 1, f_k = 42, e_k = 16.99, k = 9 \\ \sum_{i=1}^{k-1} f_i + k_0 = 123 + k_0 = \frac{N}{2} + 1 = 126, k_0 = 3$$

La mediana se encuentra en el intervalo

$$\left(16.99 + \frac{2}{42}, 16.99 + \frac{3}{42} \right), m \cong 17.05$$

3.2.8 PERCENTILES:

Como extensión de la idea de mediana (que divide los datos en dos partes iguales) podríamos pensar en aquellos valores que dividen a los datos en cuatro partes iguales aproximadamente, representados por $Q_i, i = 1, 2, 3$; los

cuales se llaman primero, segundo tercer cuartil, respectivamente, claramente Q_2 es la mediana.

Si denotamos por $Q_1 = x_{0.25}$, $Q_2 = x_{0.50}$, $Q_3 = x_{0.75}$ la notación nos dice el significado de cada uno de ellos, así, $x_{0.25}$ es un valor tal que aproximadamente el 25% de las observaciones están a su izquierda, similarmente para los otros casos.

Análogamente, los valores que dividen los datos en diez partes iguales se llaman deciles:

$$D_1 = x_{0.10}, D_2 = x_{0.20}, \dots, D_9 = x_{0.90}.$$

En algunas aplicaciones, especialmente cuando hay una gran cantidad de datos, es preferible usar percentiles (división de datos en cien partes iguales). El percentil P_p o percentil p -ésimo es el centil de $p\%$ y representa un número tomado entre las observaciones, ordenadas de menor a mayor tal que $p\%$ de la muestra está a la izquierda y el $(100 - p)\%$ está a la derecha.

Para hallar P_p procedemos de manera análoga al caso de la mediana:

1. Si disponemos de la muestra bruta ordenada en orden creciente, podemos calcular el centil de $p\%$ directamente: sea N el número de observaciones (en el caso de la mediana $p = 50$), el centil p es el dato tal que la cantidad de datos que están debajo de él es

$$\frac{pN}{100}$$

si esta cantidad es un entero, aproximamos

$$P_p = x_{(\frac{pN}{100}+1)}$$

o el punto medio de entre los valores $x_{(\frac{pN}{100})}$ y $x_{(\frac{pN}{100}+1)}$ (como lo hicimos en el caso de la mediana). Si esa cantidad no es un entero tomamos parte entera de $\frac{pN}{100} + 1$ y aproximamos

$$P_p = x_{([\frac{pN}{100}+1])}$$

2. Si no disponemos de la muestra bruta y contamos con la tabla de frecuencias, podemos proceder de la siguiente manera:

- (a) Elegimos un intervalo de clase $[e_k, e_{k+1}]$ tale que

$$\sum_{i=1}^{k-1} f_i \leq \frac{pN}{100}, \sum_{i=1}^k f_i > \frac{pN}{100}$$

donde f_i representa la frecuencia del intervalo $[e_k, e_{k+1}]$.

- (b) Supondremos que las observaciones que caen en el intervalo $[e_k, e_{k+1}]$, están uniformemente distribuídas en dicho intervalo, es decir, si f_k es el número de observaciones en dicho intervalo y lo subdividimos en f_k subintervalos de igual longitud

$$L_k = \frac{e_{k+1} - e_k}{f_k}$$

supondremos que en cada subdivisión hay una sola observación.

- (c) Calculamos k_0 tal que

$$\sum_{i=1}^{k-1} f_i + k_0 = \left\lceil \frac{pN}{100} + 1 \right\rceil$$

entonces, elegimos P_p en el intervalo

$$\left(e_k + (k_0 - 1) \frac{e_{k+1} - e_k}{f_k}, e_k + k_0 \frac{e_{k+1} - e_k}{f_k} \right)$$

EJEMPLO: Las notas obtenidas por 1350 estudiantes en los exámenes de ingreso a la Universidad (en base a 100 puntos), en cierto año, aparece agrupado en la siguiente tabla de frecuencias:

Clases	Frecuencias
(0, 10)	2
(11, 20)	15
(21, 30)	75
(31, 40)	150
(41, 50)	302
(51, 60)	352
(61, 70)	287
(71, 80)	120
(81, 90)	42
(91, 100)	5
Total	1350

Cálculo de la Moda: el intervalo donde hay más observaciones es $(51, 60)$, tomamos como moda el valor

$$\frac{60 + 51}{2} = 55.5$$

Cálculo de la mediana: $N = 1350, \frac{N}{2} = 675, \frac{N}{2} + 1 = 676$:

$$\begin{aligned} 2 + 15 + 75 + 150 + 302 &= 544 < \frac{N}{2} = 675 \\ 544 + 352 &= 896 > 675, k = 6 \end{aligned}$$

Consideramos el intervalo de clase $[51, 60]$, para facilitar los cálculos tomamos en su lugar el intervalo $[50, 60]$ de longitud 10

$$\begin{aligned} e_{k+1} - e_k &= 10, f_6 = 352, \\ \sum_{i=1}^{k-1} f_i + k_0 &= 544 + k_0 = \frac{N}{2} + 1 = 676, k_0 = 132 \end{aligned}$$

La mediana estará en el intervalo

$$\left(e_k + (k_0 - 1) \frac{e_{k+1} - e_k}{f_k}, e_k + k_0 \frac{e_{k+1} - e_k}{f_k} \right) = (53.72, 53.75)$$

Si queremos aproximarla por un valor numérico, podemos tomar el punto medio del intervalo, a saber:

$$m = 53.73$$

Cálculo del centil 12% :

$$\begin{aligned} \frac{Np}{100} &= 162, \frac{Np}{100} + 1 = 163 \\ 2 + 15 + 75 &= 92 < 162 \\ 92 + 150 &> 162, k = 4. \end{aligned}$$

Tomamos el intervalo de clase: $[e_k, e_{k+1}] = [30, 40]$, $k_0 = 163 - 92 = 71$. El centil de 12% se encuentra en el intervalo $(34.66, 34.73)$ y podemos elegir

$$P_{12} = 34.70$$

es decir, que el 12% de las observaciones se hallan a la izquierda de 34.70.

3.3 Estadística Descriptiva (Dos Variables): Mínimos Cuadrados.

En muchos problemas obtenemos datos pareados (x_i, y_i) , no conocemos la distribución conjunta de las variables aleatorias correspondientes y al graficar estos datos tenemos la impresión de que una recta podría ser un buen ajuste para ellos, aunque los puntos no estén exactamente sobre una recta. Los problemas de este tipo, suelen manejarse por medio del método de los mínimos cuadrados que consiste en hallar la recta

$$y = ax + b$$

que mejor se ajusta a esos datos, para ello debemos calcular los parámetros a y b a partir de los datos, es decir:

Si nos dan un conjunto de datos pareados $\{(x_i, y_i); i = 1, 2, 3, \dots, n\}$, las estimaciones de mínimos cuadrados de los coeficientes a y b son los valores para los cuales la cantidad:

$$q(a, b) = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

es un mínimo. Al diferenciar parcialmente con respecto a a y a b y al igualar estas derivadas parciales a cero, se obtiene:

$$\begin{aligned}\frac{\partial q}{\partial a} &= (-2) \sum_{i=1}^n [y_i - (a + bx_i)] = 0 \\ \frac{\partial q}{\partial b} &= (-2) \sum_{i=1}^n x_i [y_i - (a + bx_i)] = 0\end{aligned}$$

que producen el siguiente sistema de ecuaciones:

$$\begin{aligned}\sum_{i=1}^n y_i &= an + b \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i &= a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2\end{aligned}$$

Al resolver ese sistema de ecuaciones se obtiene:

$$\begin{aligned}a &= \bar{y} - b\bar{x} \\ b &= \frac{S_{xy}}{S_{xx}}\end{aligned}$$

donde :

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

3.3.1 EJERCICIO:

Consideremos los siguientes datos acerca del número de horas de estudio de 10 personas para presentar un examen de inglés y sus calificaciones obtenidas en base a 100 puntos:

Horas de estudio (x)	Calificación en la prueba (y)
4	31
9	58
10	65
14	73
4	37
7	44
12	60
22	91
1	21
17	84

Grafique los datos y halle la ecuación de la recta que mejor se ajusta a estos datos, usando el Método de Mínimos Cuadrados.

3.4 Correlación:

Recuerde que si X e Y son dos variables aleatorias, el coeficiente de correlación de ellas se define como:

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(x)Var(Y)}}$$

este valor está en el intervalo $[-1, 1]$ y mide en cierto sentido el grado de dependencia lineal entre las variables, si $\rho = \pm 1$, con probabilidad uno, existe una dependencia lineal perfecta entre las variables. Si las variables son independientes $\rho = 0$, el recíproco es falso, salvo cuando la distribución conjunta de las variables es normal.

El coeficiente de correlación observado correspondiente a dos muestras aleatorias de X e Y respectivamente es:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

En la práctica para tener una idea estimada del grado de correlación de dos variables, se utilizan los llamados diagramas de dispersión ó nubes de puntos, que son los puntos correspondientes a los pares (x_i, y_i) , que representan las observaciones de ambas variables, representados en un plano cartesiano. Si $r = 0$ no existe relación lineal entre las variables, si $r < 0$ y cercano a -1 , existe cierta correlación lineal entre las variables y la mejor recta que aproxima los datos tiene pendiente negativa (es decreciente).

4 Función de Distribución Empírica.

Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidades. Cuando realizamos un experimento el conjunto de resultados de las observaciones sirve de material inicial para toda investigación estadística, en muchos casos corresponden a los valores experimentales $\{x_1, x_2, \dots, x_n\}$ de cierta variable aleatoria X . La distribución de esta variable $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$, B boreliano de \mathbb{R} , en general se desconoce al menos parcialmente.

Consideremos n repeticiones independientes de la variable aleatoria X , es decir, X_1, \dots, X_n es una sucesión de variables aleatorias independientes con la misma distribución que X .

Definamos

$$I_x(B) = \begin{cases} 1 & \text{si } x \in B \\ 0 & \text{si } x \notin B \end{cases}$$

si consideramos sobre el conjunto $\{x_1, x_2, \dots, x_n\}$ la distribución uniforme es decir la probabilidad \mathbb{P}_n definida como

$$\mathbb{P}_n(B) = \frac{1}{n} \sum_{i=1}^n I_{x_i}(B) = \frac{\text{card} \{i : x_i \in B\}}{n}$$

si en esta definición escribimos X_i en lugar de los resultados de la muestra, esa expresión será una variable aleatoria.

DEFINICIÓN: Sea X una variable aleatoria de función de distribución $F(x)$, X_1, \dots, X_n es una sucesión de variables aleatorias independientes con la misma distribución que X , definamos

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i}((-\infty, x]) = \frac{\text{card}\{i : X_i \leq x\}}{n}$$

esta expresión es una variable aleatoria que denominamos función de distribución empírica. También se puede expresar como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i) = \frac{\text{card}\{i : X_i \leq x\}}{n}$$

donde

$$I_{(-\infty, x]}(X_i) = \begin{cases} 1 & \text{si } X_i \leq x \\ 0 & \text{si } X_i > x \end{cases}$$

Teorema 1: Sea X una variable aleatoria de función de distribución $F(x)$, X_1, \dots, X_n es una sucesión de variables aleatorias independientes con la misma distribución que X ,

$$F_n(x) \xrightarrow{c.s} F(x), \forall x, n \rightarrow \infty$$

donde c.s significa, casi siempre y explícitamente expresa que la probabilidad \mathbb{P} del conjunto donde esto no ocurre es cero.

Demostración

La Ley fuerte de grandes números (Wiebe R. Pestman, Mathematical Statistics, Walter de Gruyter, Berlin, New York, 1998, Teorema VII.2.14) nos asegura este resultado pues $F_n(x)$ es el promedio de n variables aleatorias independientes de esperanza

$$\begin{aligned} \mathbb{E}(F_n(x)) &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n I_{X_i}((-\infty, x])\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(I_{X_i}((-\infty, x])) = \\ \frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_i \leq x) &= \frac{1}{n} \sum_{i=1}^n F(x) = F(x) \end{aligned}$$

■

Podemos “estimar” la función de distribución $F(x)$ por medio de la función de distribución empírica, la mayor distancia vertical entre las gráficas de

las funciones F_n y F está representada por la expresión $\sup_{x \in \mathbf{R}} |F_n(x) - F(x)|$, el teorema de Glivenko-Cantelli nos dice que esta expresión tiende a cero cuando $n \rightarrow \infty$, para casi todo $\omega \in \Omega$, es decir, el conjunto donde no hay convergencia tiene probabilidad cero.

Teorema 2: Si $n \rightarrow \infty$

$$\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \xrightarrow{c.s} 0, n \rightarrow \infty$$

este resultado se conoce como el Teorema de Glivenko-Cantelli.

Demostración:

Daremos una demostración para el caso F continua.

Sea $\varepsilon > 0$ arbitrariamente pequeño de forma tal que el número $N = \frac{1}{\varepsilon}$ sea un entero. La continuidad de F nos permite hallar números tales que

$$\begin{aligned} F(z_1) &= \frac{1}{N} = \varepsilon, F(z_k) = \frac{k}{N} = k\varepsilon, k = 1, 2, \dots, N-1; \\ \text{definimos } z_0 &= -\infty, z_N = \infty, \text{ así } F(z_0) = 0, F(z_N) = 1 \end{aligned}$$

Si $z \in (z_k, z_{k+1}]$ las relaciones siguientes son ciertas

$$\begin{aligned} F_n(z) - F(z) &\leq F_n(z_{k+1}) - F(z_k) = F_n(z_{k+1}) - F(z_{k+1}) + \varepsilon \\ F_n(z) - F(z) &\geq F_n(z_k) - F(z_{k+1}) \geq F_n(z_k) - F(z_k) - \varepsilon \end{aligned}$$

Consideremos los siguiente eventos

$$A_k = \{\omega \in \Omega : F_n(z_k) \rightarrow F(z_k), n \rightarrow \infty\}$$

por el teorema 1, $\mathbb{P}(A_k) = 1$. Sea $A = \bigcap_{k=0}^N A_k$, también $\mathbb{P}(A) = 1$. Si $\omega \in A$, existirá $n(\omega) : \text{si } n \geq n(\omega)$

$$|F_n(z_k) - F(z_k)| < \varepsilon, k = 0, 1, 2, \dots, N$$

este resultado junto a las desigualdades anteriores nos asegura que

$$-2\varepsilon \leq F_n(z_k) - F(z_k) - \varepsilon \leq F_n(z) - F(z) \leq F_n(z_{k+1}) - F(z_{k+1}) + \varepsilon \leq 2\varepsilon$$

de donde

$$\sup_z |F_n(z) - F(z)| \leq 2\varepsilon, \text{ si } n \geq n(\omega)$$

con probabilidad uno.

■

Para una demostración general se puede consultar Kai Lai Chung, *A course in Probability Theory*, Academic Press, capítulo 5, Teorema 5.5.1 o Wiebe R. Pestman, *Mathematical Statistics*, Walter de Gruyter, Berlin, New York, 1998, capítulo VII, Teorema VII.3.4.

Estadística Descriptiva.
Estadística
Práctica N° 1

1. Los siguientes datos indican el número de trabajadores que faltan a una fábrica en 50 días de trabajo:

13	5	13	37	10	16	2	11	6	12
8	19	21	12	11	7	7	9	16	18
3	11	19	6	15	10	14	10	7	24
11	3	6	10	4	6	32	9	12	7
29	12	9	10	8	20	15	5	17	10

Utilice las seis clases: 0-4, 5-9, 10-14, 15-19, 20-24, 25 ó mayor para construir una tabla de frecuencias absolutas y relativas. Dibujar el histograma. Construir la tabla de frecuencias acumuladas. Encontrar media muestral, desviación estandard, moda, mediana y cuartiles. Se cumple la regla empírica?

2. Los siguientes datos son los números de torsiones requeridas para doce barras cierta aleación: 33, 24, 39, 48, 26, 35, 38, 54, 23, 34, 29 y 37. Calcule:

- (a) media
- (b) s^2
- (c) la mediana
- (d) la moda
- (e) los cuartiles.
- (f) Se cumple la regla empírica?

3. Demuestre que

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

para una muestra x_1, x_2, \dots, x_n .

4. Si los datos se codifican de tal manera que $x_i = cu_i + a$, demuestre que

$$\bar{x} = c\bar{u} + a, \quad s_x = cs_u$$

para una muestra pareada $x_1, x_2, \dots, x_n; u_1, u_2, \dots, u_n$.

5. La efectividad de una nueva técnica para controlar un insecto que afecta un tipo de cultivo se puede medir contando el número de larvas del insecto halladas en cierta superficie de cultivo. Después de aplicar la técnica, se contaron las larvas en 40 áreas, obteniendo los datos siguientes:

5	0	2	40	27	3	0	22
14	0	4	19	38	2	5	16
0	7	42	15	39	0	2	0
29	26	14	0	3	27	32	20
3	0	17	35	29	12	16	6

- Elabore una tabla de frecuencias absolutas y relativas y haga los histogramas correspondientes.
 - Calcule las acumuladas absolutas y relativas y haga los histogramas correspondientes.
 - Se cumple la regla empírica?
 - En lugar de histogramas haga ahora gráficos de línea.
 - Encontrar media muestral, varianza, desviación estándar, moda, mediana y cuartiles de los datos.
6. Después de observar el tiempo de vida de 70 motores, se obtuvieron los siguientes datos:

Intervalos de años de funcionamiento	Número de motores
[0, 1)	30
[1, 2)	23
[2, 3)	6
[3, 4)	5
4 años o más	6

- Haga un histograma de frecuencias relativas.
- Se cumple la regla empírica?

- (c) En base al histograma de la parte a), qué distribución sospecha Ud. que podría tener la variable aleatoria T = tiempo de vida de un motor del tipo considerado?
- (d) Calcule aproximadamente, la media, desviación y mediana de estos datos.
7. La evidencia directa de la ley de gravitación universal de Newton la obtuvo Henry Cavendish (1731-1810). En el experimento se obtuvo la densidad (en el tiempo) de la tierra y se construyó la siguiente tabla:

5.36	5.29	5.58	5.65	5.57	5.53	5.62	5.29
5.44	5.34	5.79	5.10	5.27	5.39	5.42	5.47
5.63	5.34	5.46	5.30	5.75	5.68	5.85	

- (a) Calcular la media y la desviación estándar.
- (b) Calcular los cuartiles, graficar densidad contra tiempo.
- (c) Hay alguna tendencia obvia?
8. Las materias primas que se utilizan en la producción de una fibra sintética se almacenan en un sitio sin control de humedad. En 12 días, las mediciones de la humedad relativa del lugar del almacenamiento y el contenido de humedad de una muestra de la materia prima (en porcentajes ambas) producen los siguientes resultados:

Humedad Ambiente	46	53	37	42	43	29	60	44	41	48	33	40
Humedad en la materia prima	12	14	11	13	10	8	17	12	10	15	9	13

- (a) Ajuste una recta de mínimos cuadrados a partir de la cual podamos predecir el contenido de humedad de la materia prima en función de la humedad del lugar.
- (b) Utilice el resultado anterior para calcular el contenido de humedad de la materia prima cuando la humedad relativa es del 38%.
9. La siguiente tabla muestra las ventas (en miles de unidades) de una pequeña empresa de componentes electrónicos durante los últimos 10 años.

Año	1	2	3	4	5	6	7	8	9	10
Ventas	2,6	2,85	3,02	3,45	3,69	4,26	4,73	5,16	5,91	6,5

- (a) Sea $X = \text{año}$ y $Y = \text{ventas}$. Grafique la nube de puntos (x_i, y_i) .
- (b) Sea $X = \text{año}$ y $Y = \ln(\text{ventas})$. Grafique la nube de puntos (x_i, y_i) .
- (c) A cuál de las dos nubes anteriores cree Ud. que se ajusta mejor una recta?
- (d) Calcule las dos rectas de regresión y gráfíquelas.

Semestre Abril-Julio2004/MMOM.

REPASO DE DISTRIBUCIONES DE PROBABILIDADES

Estadística Práctica N° 2

1. El número de accidentes de trabajo en una fábrica sigue una distribución de Poisson. Se sabe que el promedio de accidentes en dicha fábrica mensualmente es de 3. Durante el mes pasado ocurrieron 6 accidentes. Se puede considerar que este número es excesivamente alto? (es decir, poco probable).
2. La experiencia ha demostrado que el 30% de las personas que contraen cierta enfermedad se logra curar. Una compañía farmacéutica ha desarrollado un medicamento para dicha enfermedad. Se eligen al azar 10 personas enfermas y se les administra el medicamento, 5 logran curarse, cuál es la probabilidad de este evento si se supone que la medicina no tuvo ningún efecto?. Qué opina Ud. del medicamento?
3. Un examen de selección múltiple tiene 15 preguntas, cada una de las cuales posee 5 respuestas posibles y de éstas sólo una es correcta. Si un estudiante contesta todas las preguntas al azar, cuál es la probabilidad de contestar correctamente al menos 10 preguntas?
4. El fabricante de una marca de pasta de dientes afirma que el 60% de los consumidores prefieren esa marca. Si entrevistamos a un grupo de personas escogidas al azar del grupo de consumidores de pasta de dientes, cuál es la probabilidad de tener que entrevistar al menos 5 personas para encontrar al primer consumidor que prefiere esa marca?
5. El número de errores que hace una mecanógrafa tiene una distribución de Poisson con una media de 4 errores por página. Cuál es la probabilidad de que una página escogida al azar tenga a lo sumo 4 errores?
6. Una fábrica utiliza un producto cuyo uso diario puede modelarse por medio de una distribución exponencial de parámetro 4 (esto es, la cantidad de producto utilizada en un día es una variable aleatoria exponencial de parámetro $\lambda = 4$, medida en toneladas). Cuántas toneladas de producto debe almacenar la fábrica para que la probabilidad de quedarse sin producto en un día dado sea sólo 0,05?.

7. Un defecto metabólico ocurre en aproximadamente 1 de cada 100 nacimientos. Si en un hospital nacen 4 niños en un día dado, calcule:
- (a) la probabilidad de que ninguno tenga el defecto
 - (b) la probabilidad de que a lo sumo uno de ellos tenga el defecto
 - (c) la probabilidad de que al menos uno de ellos tenga el defecto.
8. En un examen se plantean 10 preguntas a las que debe responderse con verdadero o falso. Un alumno aprobará el examen si al menos 7 de sus respuestas son acertadas.
- (a) Qué probabilidad de aprobar tiene un estudiante que responde todo al azar?
 - (b) Qué probabilidad de aprobar tiene un estudiante que sabe el 30%?

Semestre Abril-Julio 2004/MMOM

**Repaso de Desigualdad de Tchebysheff,
Distribución Normal, Teorema Central del Límite.
Estadística
Práctica N° 3**

1. Sea X una variable aleatoria con distribución normal de parámetros

$$\mu \in \mathbb{R}, \quad \sigma > 0$$

Demuestre que

- (a) $\mathbb{E}(X) = \mu, \quad \text{Var}(X) = \sigma^2$.
- (b) $Z = \frac{X - \mu}{\sigma}$ tiene distribución normal estándar.
2. Una línea aérea sabe que el 5% de las personas que hacen reservaciones en un cierto vuelo, al final no se presentan. Si la aerolínea vende 160 boletos para este vuelo, y sólo hay 155 asientos en el avión, cuál es la probabilidad de que todo pasajero con reservación que se presente al aeropuerto tenga un puesto en el vuelo?.
3. En una empresa se ha observado que el gasto semanal en mantenimiento y reparaciones es una variable aleatoria con distribución aproximadamente normal de media $\mu = Bs. 24000$ y desviación $\sigma = Bs.1200$. Cuánto debe presupuestarse semanalmente para mantenimiento y reparaciones para que el monto presupuestado sea excedido con una probabilidad de a lo sumo 0,1?
4. Un encuestador cree que el 20% de los votantes de una zona está a favor del candidato A . Si se escogen 24 votantes de la zona, aproxime la probabilidad de que la fracción de votantes de la muestra que favorece al candidato A , no difiera de la verdadera fracción (en toda la zona) en más de 0,06.
5. Una máquina se manda a reparar si una muestra de 100 artículos escogidos al azar de su gran producción diaria, revela un 15% ó mas de defectuosos. Si la máquina en realidad sólo produce un 10% de defectuosos, calcule aproximadamente la probabilidad de que la manden a reparar.

6. La vida activa de un cierto fármaco sigue una distribución $N(1200, 40)$ días. Se desea enviar un lote de medicamentos, de modo tal que la vida media del lote no sea inferior a 1180 días con probabilidad 0,95. Qué tamaño debe tener la muestra?
7. Encuentre una aproximación de la probabilidad, de que el número de veces que salga 1, esté comprendido entre 1900 y 2150 veces, al lanzar un dado perfecto 12000 veces.
8. Se toma una muestra al azar con reposición, a efectos de estimar la fracción p de hembras en una población. Encontrar un tamaño de muestra que asegure que la estimación se hará con un error de menos de 0,005, al menos con una probabilidad de 0,99.
9. Se desea estimar la probabilidad de falla p , en un proceso de producción, mediante la observación de n objetos producidos, elegidos independientemente. Se sabe que p está entre 0,1 y 0,3 por información previa. Halle el tamaño n de la muestra para que la probabilidad de que la frecuencia relativa de objetos fallados en la muestra, difiera del verdadero valor p en más de 0,01 sea menor que 0,05.
10. El porcentaje de individuos daltónicos de una población es P desconocido. Se desea estimar este porcentaje P a partir del porcentaje observado en una muestra de tamaño n . Calcular el tamaño que debe tener la muestra a fin de que el error cometido sea inferior al 1% con probabilidad 0,90 en los casos:
 - (a) No se sabe nada acerca de P .
 - (b) Se sabe que P es inferior al 16%.
11. Se ha observado que las notas de un examen de admisión siguen una distribución aproximadamente normal, de media 78 y varianza 36. (Las notas están entre 1 y 100).
 - (a) Si un grupo de estudiantes va a presentar dicho examen, qué porcentaje de ellos espera Ud. que obtenga notas entre 70 y 90?
 - (b)Cuál es la probabilidad de que una persona que tome el examen obtenga más de 72?

- (c) Suponga que los estudiantes cuyas notas se encuentran en el 10% superior de la distribución serán admitidos inmediatamente.Cuál debe ser la nota mínima que debe tener un estudiante para ser admitido inmediatamente?
12. La duración de un tipo de bombillos sigue una distribución Normal de media $\mu = 1000$ horas y desviación $\sigma = 100$ horas.
- Se desea enviar una muestra de bombillos de manera que la duración media de la muestra no difiera de μ en más de 50 horas con una probabilidad de 0,95.
- (a) Hallar el tamaño que debe tener la muestra.
- (b) Resuelva el problema, si se desconoce la distribución :
- Usando Tchebichev.
 - Usando el Teorema Central del Límite.
13. Una compañía tiene 90 ejecutivos. Supongamos que la probabilidad de que un ejecutivo necesite una secretaria al comenzar su día de trabajo es $\frac{1}{10}$. Si queremos que con un 95% de certeza haya una secretaria disponible para cada ejecutivo que la solicite, cuántas secretarias deberían contratarse para un centro secretarial que sirva al grupo de 90 ejecutivos?
14. Un fabricante de cereales afirma que el peso medio de una caja del cereal que vende es de 330,4 *grs.* con una desviación de 21 *grs.* Se desea verificar si su afirmación es cierta. Para esto se va a elegir una muestra aleatoria de cajas del cereal y calcular el peso promedio de la muestra. Cuántas cajas debemos tener en la muestra para que el peso promedio se encuentre a menos de 7 *grs.* de la verdadera media con una probabilidad de 0,99? (Suponga que la distribución del peso de cada caja es normal).
15. Si la probabilidad de que un individuo sufra una reacción alérgica por la inyección de cierto medicamento es de 0,001; calcule la probabilidad de que, de un total de 2000 individuos a quienes se inyectó el medicamento, más de 2 tengan una reacción alérgica.

Semestre Abril-Julio 2004/MMOM

Capítulo II

Inferencia Estadística:

Estimación Puntual de Parámetros.

María Margarita Olivares M.

Abril 2004

1 INTRODUCCIÓN:

Cuando se realiza un experimento aleatorio, los posibles resultados de dicho experimento se pueden pensar como una variable aleatoria.

En general, un material estadístico consiste en un número de observaciones x_1, x_2, \dots, x_N , obtenido a partir de N repeticiones independientes del experimento aleatorio relacionado con X . La estadística descriptiva reduce el material observado o muestra bruta, reemplazándolo por cantidades relativamente pocas en número que representen el material total y que contengan toda la información posible de la variable aleatoria X .

En el material estadístico raramente podemos incluir todas las observaciones que podríamos realizar teóricamente, por lo que este material se puede considerar como una muestra aleatoria simple o como una sucesión de variables aleatorias independientes, todas con la misma distribución, la cual está sujeta a fluctuaciones estadísticas ya que se obtendrían valores distintos x'_1, x'_2, \dots, x'_N , si realizáramos N nuevas observaciones.

Es decir, antes de realizar el experimento, los valores de X que se van a observar deben concebirse como N variables aleatorias

$$X_1, X_2, \dots, X_N,$$

independientes, idénticamente distribuidas, con la misma distribución de la variable aleatoria X .

El objetivo de la estadística es hacer inferencia acerca de una población basándose en la información contenida en una muestra, por ejemplo, tomar decisiones sobre la distribución de probabilidad de la variable aleatoria X y describir esa distribución basándose en la observación de esta variable aleatoria. Puesto que las distribuciones se caracterizan por medidas descriptivas numéricas, llamadas parámetros, la estadística se interesa en hacer inferencia acerca de los parámetros de las distribuciones de probabilidad. Algunos parámetros típicos son la media, la desviación estándar, el área bajo la distribución de probabilidad a partir de un valor de la variable aleatoria o el área entre dos valores de la variable.

Algunos ejemplos pueden aclarar esta idea:

1. El lavaplatos de un restaurante posee un certificado de garantía que expresa que de cada 100 platos que lava, sólo rompe 3. El primer día lava 500 platos y se le rompen 23, resulta creíble lo que expresa la garantía?
2. Se repite independientemente un experimento que puede dar lugar en cada repetición al resultado A con probabilidad p . Al cabo de 200 repeticiones, A ocurrió 22 veces. Se desea saber al menos aproximadamente el valor de p .
3. Una calculadora bolsillo tiene una rutina generadora de números aleatorios, que de acuerdo a lo que indica el fabricante, proporciona una sucesión de variables aleatorias independientes de distribución uniforme en $[0, 1]$. Se genera una sucesión x_1, x_2, \dots, x_N . A partir del conocimiento de ella, resulta aceptable la afirmación del fabricante?

En los ejemplos anteriores, planteamos problemas de estimación de parámetros y también de pruebas de hipótesis, los cuales analizaremos más adelante.

Supongamos que F es la función de distribución teórica de la variable aleatoria X , en observación; F en general contendrá uno o más parámetros tales como μ y σ en el caso de la distribución normal. Una vez conocidos los valores numéricos de estos parámetros, la variable aleatoria que estamos investigando queda completamente caracterizada.

Basándonos en las observaciones

$$x_1, x_2, \dots, x_N,$$

estimamos los valores numéricos de los parámetros de F , estos estimadores empíricos de los parámetros teóricos son a su vez variables aleatorias y como tales están sujetos a fluctuaciones estadísticas. Para tener una medida de la magnitud esperada de estas fluctuaciones y por lo tanto de la confianza que podemos depositar en los valores encontrados para los parámetros a partir de las observaciones, debemos deducir a partir de F , las distribuciones de nuestros estimadores.

Así pues, antes de que podamos resolver un problema estadístico dado, debemos, en primer lugar, establecer una hipótesis sobre la forma matemática de la función de distribución F .

A veces, por experiencias anteriores, se sabe que podemos suponer una determinada distribución, por ejemplo, la distribución normal. O bien, a partir de ciertas hipótesis que idealizan el experimento considerado podemos deducir su distribución, basándonos en las reglas conocidas de la teoría de probabilidad. Por ejemplo, cuando se cuenta el número de partículas α que se observan en una pantalla, emitidas por una sustancia radioactiva durante un tiempo t , bajo ciertas hipótesis simplificadoras, podemos suponer que la distribución es de Poisson de parámetro λt y justamente es el parámetro λ el valor que debemos estimar a partir de las observaciones.

1.1 DEFINICIONES:

1.1.1 MUESTRA ALEATORIA SIMPLE:

Es un vector aleatorio

$$\vec{X} = (X_1, X_2, \dots, X_N)$$

cuyas componentes son variables aleatorias independientes, idénticamente distribuidas, siendo N el tamaño de la muestra.

1.1.2 ESTADÍSTICO (o ESTADÍGRAFO)

Es toda función T de una muestra aleatoria $\vec{X} = (X_1, X_2, \dots, X_N)$ que a su vez resulte ser una variable aleatoria:

$$T_N = T(\vec{X}) = T(X_1, X_2, \dots, X_N)$$

(La función T debe ser lo suficientemente regular como para que $T(\vec{X})$ sea una variable aleatoria)

EJEMPLOS DE ESTADÍSTICOS:

1. Media Muestral Aleatoria:

$$\bar{X} = \frac{1}{N} (X_1 + X_2 + \cdots + X_N) = \frac{1}{N} \sum_{i=1}^N X_i$$

2. Varianza Muestral Aleatoria:

$$S^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

3. Varianza Muestral Centrada Aleatoria:

$$S_1^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

1.1.3 OBSERVACIÓN:

La distribución de estos estadísticos está determinada por la distribución teórica F de la variable aleatoria X .

1.2 ESTIMADOR:

Un estimador paramétrico o para simplificar, diremos simplemente, un estimador, es un estadístico cuyo valor observado intentamos usar para estimar el valor de un parámetro desconocido de la distribución teórica. (El enfoque paramétrico supone que la forma del modelo es conocida).

La media muestral y la varianza muestral aleatorias, como lo indica sus nombres, son estimadores de la media y la varianza de la distribución teórica.

Supongamos que $T_N = T(X_1, X_2, \dots, X_N)$ sea un estimador de un cierto parámetro β de una distribución teórica. La diferencia:

$$T_N - \beta = T(X_1, X_2, \dots, X_N) - \beta$$

se denomina Error de Estimación. Una buena forma de conseguir que T_N sea un buen estimador, es pedir que el error de estimación sea pequeño y

esto puede hacerse, por ejemplo, exigiendo que se cumplan condiciones tales como:

$$\mathbb{P}(|T(X_1, X_2, \dots, X_N) - \beta| > \delta) < \varepsilon$$

para valores pequeños de $\delta > 0, \varepsilon > 0$; o bien que

$$\mathbb{E}(|T(X_1, X_2, \dots, X_N) - \beta|^k) < c$$

para valores apropiados de las constantes $k > 0$, y $c > 0$ pequeño.

En particular, llamamos error cuadrático medio a la expresión:

$$\mathbb{E}(|T(X_1, X_2, \dots, X_N) - \beta|^2),$$

es deseable que un estimador tenga un error cuadrático medio pequeño.

A menudo, tienen interés, sobre todo técnico, las siguientes propiedades:

1. T es un Estimador Insesgado o Centrado de β cuando

$$\mathbb{E}(T(X_1, X_2, \dots, X_N)) = \beta$$

para todo $N \geq 1$. En este caso el error cuadrático medio coincide con la varianza.

A la diferencia

$$\mathbb{E}(T(X_1, X_2, \dots, X_N)) - \beta$$

se le llama sesgo de T .

2. EFICIENCIA RELATIVA: dos estimadores insesgados T_1 y T_2 , del mismo parámetro β , basados en las mismas observaciones, se suelen comparar utilizando la eficiencia relativa de T_2 con respecto a T_1 , la cual se define como el cociente

$$\frac{Var(T_1)}{Var(T_2)}.$$

Si este cociente es menor que 1, es decir, $Var(T_1) < Var(T_2)$ diremos que el estimador T_1 es más eficiente que T_2 .

3. ESTIMADOR CONSISTENTE:

Sea T un estimador del parámetro β y sea

$$T_n = T(X_1, X_2, \dots, X_n)$$

una sucesión de estimadores de β , que representan a T con base en la muestra de tamaño n .

Se dice que T es un estimador consistente si:

$$\lim_{N \rightarrow \infty} \mathbb{P}(|T(X_1, X_2, \dots, X_N) - \beta| \geq \varepsilon) = 0$$

(Este tipo de convergencia se llama convergencia en probabilidad del estimador al verdadero valor del parámetro).

EJERCICIOS:

1. Demostrar que si T es un estimador de β cuyo error cuadrático medio tiende a cero cuando $n \rightarrow \infty$, es consistente.
2. Si T es insesgado y $\text{Var}(T_n)$ tiende a cero cuando $n \rightarrow \infty$, entonces T es consistente.
3. Si $\lim_{N \rightarrow \infty} \mathbb{E}(T_N(X_1, X_2, \dots, X_N)) = \beta$ y $\lim_{N \rightarrow \infty} \text{Var}(T_N(X_1, X_2, \dots, X_N)) = 0$, entonces T es consistente.

1.3 Propiedades de la media y la varianza aleatorias:

(Como estimadores de la media y la varianza).

Sea $\vec{X} = (X_1, X_2, \dots, X_N)$ una muestra aleatoria de la variable X , con

$$\mu = \mathbb{E}(X_i), \sigma^2 = \text{Var}(X_i), i = 1, 2, \dots, N$$

1. $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ es un estimador insesgado y consistente de μ :

$$\mathbb{E}(\bar{X}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}(X_i) = \frac{1}{N} N \mu = \mu$$

$$\text{Var}(\bar{X}) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}(X_i) = \frac{1}{N^2} N \sigma^2 = \frac{\sigma^2}{N}$$

2. S_1^2 es un estimador insesgado de σ^2 :

Puesto que $S_1^2 = \frac{N}{N-1}S^2$, podemos calcular $\mathbb{E}(S^2)$ y deducir de allí la $\mathbb{E}(S_1^2) = \frac{N}{N-1}\mathbb{E}(S^2)$. Para calcular $\mathbb{E}(S^2)$, supongamos que:

$$\begin{aligned}\mathbb{E}(X_i) &= \mu, \mathbb{E}((X_i - \mu)^2) = \text{Var}(X_i) = \sigma^2, 1 \leq i \leq N. \\ \sum_{i=1}^N (X_i - \bar{X})^2 &= \sum_{i=1}^N (X_i - \mu - \bar{X} + \mu)^2 = \\ &= \sum_{i=1}^N (X_i - \mu)^2 + N(\mu - \bar{X})^2 + 2(\mu - \bar{X}) \sum_{i=1}^N (X_i - \mu) = \\ &= \sum_{i=1}^N (X_i - \mu)^2 + N(\mu - \bar{X})^2 - 2N(\mu - \bar{X})^2 = \sum_{i=1}^N (X_i - \mu)^2 - N(\mu - \bar{X})^2\end{aligned}$$

o equivalentemente

$$\sum_{i=1}^N (X_i - \mu)^2 = \sum_{i=1}^N (X_i - \bar{X})^2 + N(\bar{X} - \mu)^2$$

Este resultado tiene una interpretación importante pues descompone la variabilidad de los datos respecto a su media verdadera como suma de la variabilidad respecto a la media muestral y la variabilidad entre la media muestral y la verdadera.

Tomando esperanza:

$$\begin{aligned}N\sigma^2 &= \mathbb{E}(NS^2) + N\text{Var}(\bar{X}) \\ \mathbb{E}(S^2) &= \sigma^2 - \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2.\end{aligned}$$

De aquí se obtiene que:

$$\mathbb{E}(S_1^2) = \frac{N}{N-1}\mathbb{E}(S^2) = \sigma^2.$$

Note que S_1^2 es un estimador insesgado o centrado de la varianza, mientras que S^2 no lo es. Esta es la razón por la que se prefiere trabajar con S_1^2 en lugar de S^2 y por ésto S_1^2 recibe el nombre de varianza centrada o varianza muestral corregida, el divisor $n - 1$ se denomina “número de grados de libertad”.

Si llamamos residuo a

$$e_i = x_i - \bar{x}$$

entonces la varianza muestral centrada o corregida será

$$S_1^2 = \frac{1}{N-1} \sum_{i=1}^N e_i^2$$

cuando $N = 1$, $\bar{x} = x_1$ y antes de tomar la muestra podemos afirmar que $e_1 = 0$. No hay ningún grado de libertad. Si $N = 2$, tendremos que $e_1 = x_1 - \bar{x} = x_1 - \frac{x_1+x_2}{2} = \frac{x_1-x_2}{2} = -e_2$. Hay solamente un grado de libertad e_1 (o e_2). Dado un residuo el otro queda automáticamente fijado. En general, para cualquier tamaño muestral

$$\sum_{i=1}^N (x_i - \bar{x}) = \sum_{i=1}^N e_i = 0$$

antes de tomar la muestra solo hay $n-1$ residuos desconocidos porque el último siempre puede calcularse usando la expresión anterior. Diremos que disponemos de $n-1$ grados de libertad para calcular los residuos y por tanto la desviación típica de los datos.

3. S_1^2 y S^2 son consistentes, si $\mathbb{E}(X^4) < \infty$; se puede demostrar que:

$$Var(S^2) = \frac{\mu_4 - \mu_2^2}{N} - \frac{2(\mu_4 - 2\mu_2^2)}{N^2} + \frac{\mu_4 - 3\mu_2^2}{N^3}$$

donde

$$\mu_k = \mathbb{E}((X - \mu)^k)$$

es el k -ésimo momento centrado de la variable aleatoria X .

Este cálculo es bastante complicado, para una demostración se puede consultar Métodos Matemáticos de estadística de Harald Cramer, editorial Aguilar, Madrid.

Puesto que

$$\lim_{N \rightarrow \infty} \mathbb{E}(S^2) = \lim_{N \rightarrow \infty} \frac{N-1}{N} \sigma^2 = \sigma^2 \text{ y } \lim_{N \rightarrow \infty} Var(S^2) = 0$$

se deduce que S^2 es consistente y puesto que $\mathbb{E}(S_1^2) = \sigma^2$, se obtiene que:

$$\lim_{N \rightarrow \infty} Var(S_1^2) = \lim_{N \rightarrow \infty} \frac{N^2}{(N-1)^2} Var(S^2) = 0$$

también obtenemos que S_1^2 es consistente.

1.4 Método de Máxima Verosimilitud.

El método general más importante para hallar estimadores de los parámetros desconocidos de una distribución teórica se conoce con el nombre de método de máxima verosimilitud introducido por R. A. Fisher. Es un método sistemático que permite hallar estimadores puntuales de cualquier número de parámetros desconocidos de una distribución.

1.4.1 Función de Verosimilitud:

Se llama Función de Verosimilitud de una muestra observada a la densidad conjunta (o función de probabilidad conjunta en el caso discreto) de la muestra aleatoria X_1, X_2, \dots, X_N , considerada como función del parámetro o de los parámetros desconocidos. Es decir,

$$L(\beta_1, \beta_2, \dots, \beta_N) = f_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \beta_1, \beta_2, \dots, \beta_N)$$

en el caso de densidad y en el caso discreto:

$$L(\beta_1, \beta_2, \dots, \beta_N) = p_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \beta_1, \beta_2, \dots, \beta_N)$$

con

$$p_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \beta_1, \beta_2, \dots, \beta_k) = \mathbb{P}(X_i = x_i; i = 1, 2, \dots, N)$$

donde $\beta_j, j = 1, 2, \dots, k$, son los parámetros desconocidos de la distribución.

OBSERVACIÓN: La función de verosimilitud representa, en cierto sentido, la probabilidad de observar lo que realmente se observó.

El método consiste en elegir los parámetros β_j de manera que la probabilidad de observar lo que se observó sea máxima, es decir, se desea elegir los parámetros β_j de tal forma que maximicen la función de verosimilitud.

Si X es la variable aleatoria asociada al experimento y

$$\beta_j, j = 1, 2, \dots, k,$$

son los parámetros desconocidos de su distribución, denotando por

$$\vec{\beta} = (\beta_1, \beta_2, \dots, \beta_k),$$

si X_1, X_2, \dots, X_N es una muestra aleatoria de la variable aleatoria X entonces:

1. En el caso de densidad, si $f(x; \vec{\beta})$ es la densidad de X , y x_1, x_2, \dots, x_N representa la muestra observada, se tendrá que

$$L(x_1, x_2, \dots, x_N; \vec{\beta}) = f(x_1; \vec{\beta}) \cdot f(x_2; \vec{\beta}) \cdot \dots \cdot f(x_N; \vec{\beta})$$

2. En el caso discreto, si $g(x; \vec{\beta}) = \mathbb{P}(X = x)$ es la función de probabilidad de X , si x_1, x_2, \dots, x_N representa la muestra observada, $\xi_1, \xi_2, \dots, \xi_r$ con $1 \leq r \leq N$ son los valores distintos observados y f_1, f_2, \dots, f_r , son las frecuencias respectivas, con $\sum_{i=1}^r f_i = N$, se tendrá que

$$L(x_1, x_2, \dots, x_N; \vec{\beta}) = \left(g(\xi_1; \vec{\beta})\right)^{f_1} \cdot \left(g(\xi_2; \vec{\beta})\right)^{f_2} \cdot \dots \cdot \left(g(\xi_r; \vec{\beta})\right)^{f_r}$$

OBSERVACIÓN: Se quiere elegir

$$\vec{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$$

de modo que $L(x_1, x_2, \dots, x_N; \vec{\beta})$ sea máximo.

Note que $\frac{\partial L}{\partial \beta_i} = 0$ si y solo si $\frac{\partial \ln L}{\partial \beta_i} = 0$ ya que $\frac{\partial \ln L}{\partial \beta_i} = \frac{1}{L} \frac{\partial L}{\partial \beta_i}$. Luego, si L es derivable con respecto a los parámetros, los extremos se calculan trabajando con la función $\ln L$ ya que los cálculos son más simples.

EJEMPLOS:

1. Estimadores de máxima verosimilitud de la media μ y de la varianza σ^2 de una distribución normal: sea x_1, x_2, \dots, x_N una muestra observada de la distribución normal, queremos estimar los parámetros basándonos en esta muestra, por el método de máxima verosimilitud:

$$\begin{aligned} f(x_i; \mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right), \quad \vec{x} = (x_1, x_2, \dots, x_N) \\ L(\vec{x}; \mu, \sigma) &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) = \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2\right) \end{aligned}$$

tomando logaritmo neperiano:

$$\begin{aligned} l(\mu, \sigma) &= \ln L(\vec{x}; \mu, \sigma) = -N \ln \sigma - \frac{N}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 \\ \frac{\partial l(\mu, \sigma)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu) = 0 \Rightarrow \mu = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x} \\ \frac{\partial l(\mu, \sigma)}{\partial \sigma} &= -\frac{N}{\sigma} + \frac{1}{2\sigma^3} \sum_{i=1}^N (x_i - \mu)^2 = 0 \Rightarrow \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = s^2 \end{aligned}$$

Para verificar que estos valores maximizan la función $l(\mu, \sigma)$ se debe evaluar

$$\Delta = \begin{vmatrix} \frac{\partial^2 l(\mu, \sigma)}{\partial^2 \mu} & \frac{\partial^2 l(\mu, \sigma)}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l(\mu, \sigma)}{\partial \mu \partial \sigma} & \frac{\partial^2 l(\mu, \sigma)}{\partial^2 \sigma} \end{vmatrix}$$

en el punto (μ, σ) encontrado. (Método del Hessiano). Si $\Delta > 0$ y $\frac{\partial^2 l(\mu, \sigma)}{\partial^2 \mu} < 0$ evaluados ambos en los puntos

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x}, \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = s^2$$

encontrados, entonces concluimos que \bar{x} y s^2 realizan un máximo de $l(\mu, \sigma)$.

2. Sea X una variable con distribución uniforme en el intervalo $[0, b]$ con $b > 0$. Calculemos el estimador de máxima verosimilitud del parámetro b basándonos en una muestra x_1, x_2, \dots, x_N . La densidad de X es

$$f(x; b) = \frac{1}{b}, \quad x \in [0, b]$$

La función de verosimilitud es:

$$\begin{aligned} L(x_1, x_2, \dots, x_N; b) &= f(x_1; b) \cdot f(x_2; b) \cdot \dots \cdot f(x_N; b) = \\ &= \frac{1}{b^N}, \quad 0 \leq x_i \leq b, \text{ para todo } i = 1, 2, \dots, N. \end{aligned}$$

o expresado de otra forma:

$$L(x_1, x_2, \dots, x_N; b) = \frac{1}{b^N}, \quad \max x_i \in [0, b], 0 \leq \min x_i.$$

Al graficar la función L como función del parámetro b , se observa que el valor máximo se realiza en el valor

$$b = \max x_i$$

en este punto L no es derivable.

3. Estimador de máxima verosimilitud del parámetro $\lambda > 0$ de la distribución de Poisson, basado en una muestra x_1, x_2, \dots, x_N : Si X tiene distribución de Poisson, su rango es

$$\{0, 1, 2, \dots, k, (k+1), \dots\},$$

de estos valores solo un número finito estará representado en la muestra x_1, x_2, \dots, x_N . Sea $r = \max x_i$, entonces los valores $0, 1, 2, \dots, r$, estarán representados en la muestra con frecuencias f_i , $1 \leq i \leq r$, respectivamente, donde f_i puede ser eventualmente cero para algún $1 \leq i \leq r$, verifican:

$$\sum_{i=1}^r f_i = N.$$

Derivando el logaritmo neperiano de la función de verosimilitud e igualando a cero, se obtiene:

$$L(x_1, x_2, \dots, x_N; \lambda) = \prod_{i=0}^r \left(\frac{\lambda^i e^{-\lambda}}{i!} \right)^{f_i}, \quad l(\lambda) = \ln L(\vec{x}; \lambda) = \sum_{i=0}^r f_i \ln \left(\frac{\lambda^i e^{-\lambda}}{i!} \right)$$

$$\frac{\partial l(\lambda)}{\partial \lambda} = \sum_{i=0}^r f_i \left(\frac{i}{\lambda} - 1 \right) = 0 \Rightarrow \frac{1}{\lambda} \sum_{i=0}^r i f_i = \sum_{i=0}^r f_i = N$$

$$\text{de donde: } \lambda = \frac{1}{N} \sum_{i=0}^r i f_i = \frac{1}{N} \sum_{i=0}^r x_i = \bar{x}.$$

4. Supongamos que en cierto experimento se observa un suceso A cuya probabilidad p es desconocida. Hacemos N observaciones y observamos f veces A . La variable observada X tiene distribución de Bernoulli de parámetro p , siendo

$$\mathbb{P}(A) = \mathbb{P}(X = 1) = p.$$

La función de verosimilitud y la derivada de su logaritmo se obtiene fácilmente:

$$\begin{aligned} L(x_1, x_2, \dots, x_N; p) &= p^f (1-p)^{N-f} \\ l(p) &= \ln L(\vec{x}; p) = f \ln p + (N-f) \ln(1-p) \\ \frac{\partial l(p)}{\partial p} &= \frac{f}{p} + \frac{N-f}{1-p} = 0 \Rightarrow p = \frac{f}{N}. \end{aligned}$$

Es decir, la frecuencia relativa observada es el estimador de máxima verosimilitud de la probabilidad de que ocurra el suceso A .

EJERCICIOS: Halle los estimadores de máxima verosimilitud, basados en una muestra x_1, x_2, \dots, x_N , si la variable observada tienen distribución:

1. Exponencial de parámetro $\lambda > 0$.
2. Densidad $f(x; p) = px^{p-1}, 0 \leq x \leq 1, p > 0$.

1.5 Estimación Puntual: Método de los Momentos.

Sea X_1, X_2, \dots, X_N una muestra aleatoria de una variable aleatoria X cuya distribución teórica depende de uno o varios parámetros desconocidos. El método de los momentos para estimar los parámetros basándose en una observación x_1, x_2, \dots, x_N , es el más antiguo que se haya propuesto con este objeto, fue introducido por K. Pearson.

Consiste en igualar un número conveniente de momentos muestrales a los correspondientes momentos de la distribución, que son funciones de los parámetros desconocidos. Considerando tantos momentos como parámetros haya que estimar y resolviendo las ecuaciones resultantes respecto a dichos parámetros, se obtienen estimaciones de éstos. Este método da muchas veces lugar, en la práctica, a cálculos relativamente simples.

Así, por ejemplo, si X tiene densidad $f(x; \beta)$ dependiendo de un solo parámetro desconocido, se utiliza como estimador de β la solución de la ecuación

$$\mathbb{E}(X) = \frac{1}{N} \sum_{i=1}^N X_i = \bar{X}$$

donde

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x; \beta) dx$$

en el caso que esta ecuación tenga solución única. Si tiene infinitas soluciones, como suele suceder cuando la distribución teórica depende de k parámetros desconocidos, con $k \geq 2$, se agrega la ecuación

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x; \beta) dx = \frac{1}{N} \sum_{i=1}^N X_i^2.$$

Si ésta no es suficiente, se agrega la que corresponde al momento de tercer orden, y así sucesivamente, hasta determinar una solución única, si ésto es posible.

1.5.1 EJEMPLOS:

1. Distribución uniforme en $[0, b]$, con $b > 0$ desconocido: como por ejemplo
 - (a) Se escogen al azar números entre 0 y algún número desconocido.
 - (b) Tiempos de espera del autobús de las 8 A.M.

La densidad es

$$f(x; \beta) = \begin{cases} \frac{1}{b}, & x \in [0, b] \\ 0, & \text{si no.} \end{cases}$$

Tenemos que resolver la ecuación:

$$\mathbb{E}(X) = \frac{1}{N} \sum_{i=1}^N X_i = \bar{X}$$

pero la esperanza de una variable aleatoria de densidad uniforme de parámetros $(0, b)$ es:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x; \beta) dx = \frac{1}{b} \int_0^b x dx = \frac{b}{2}.$$

Igualando obtenemos:

$$\frac{b}{2} = \bar{X}, \text{ tomamos } \hat{b} = 2\bar{X}$$

El estimador \hat{b} de b , es insesgado, pues

$$\mathbb{E}(\hat{b}) = 2\mathbb{E}(\bar{X}) = \frac{2}{N} \sum_{i=1}^N \mathbb{E}(X_i) = \frac{2}{N} N \mathbb{E}(X) = 2 \frac{b}{2} = b.$$

El error medio cuadrático del estimador \hat{b} , por ser en este caso insesgado, coincide con su varianza:

$$\begin{aligned}\mathbb{E} \left((\hat{b} - b)^2 \right) &= \text{Var}(\hat{b}) = \text{Var}(2\bar{X}) = \frac{4}{N^2} \sum_{i=1}^N \text{Var}(X_i) = \\ \frac{4}{N} \text{Var}(X) &= \frac{4}{N} \frac{b^2}{12}\end{aligned}$$

por ser X uniforme en $[0, b]$. Por lo tanto

$$\text{Var}(\hat{b}) = \frac{b^2}{3N} \rightarrow 0 \text{ si } N \rightarrow \infty,$$

es decir, nuestro estimador $\hat{b} = 2\bar{X}$ del parámetro desconocido b es consistente.

Por el método de máxima verosimilitud, obtuvimos como estimador del parámetro b a $\hat{b} = \max(X_1, X_2, \dots, X_N)$. Veamos que este estimador no es insesgado; como tenemos que hallar su esperanza, debemos calcular antes su distribución:

$$\begin{aligned}\mathbb{P} \left(\hat{b} \leq x \right) &= \mathbb{P}(\max(X_1, X_2, \dots, X_N) \leq x) = \\ \mathbb{P}(X_1 \leq x, X_2 \leq x, \dots, X_N \leq x) &= \\ \prod_{i=1}^N \mathbb{P}(X_i \leq x) &= (\mathbb{P}(X \leq x))^N = (F(x; b))^N\end{aligned}$$

donde F es la función de distribución de X que es uniforme en el intervalo $[0, b]$. Es fácil calcular esta función de distribución para obtener:

$$F(x; b) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x}{b} & \text{si } x \in [0, b] \\ 1 & \text{si } x > b. \end{cases}$$

Así, la función de distribución del estimador, es:

$$\mathbb{P} \left(\hat{b} \leq x \right) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x^N}{b^N} & \text{si } x \in [0, b] \\ 1 & \text{si } x > b. \end{cases}$$

y derivando, obtenemos la densidad de \hat{b} :

$$f_{\hat{b}}(x) = \begin{cases} \frac{N}{b^N} x^{N-1} & \text{si } x \in [0, b] \\ 0 & \text{si no.} \end{cases}$$

La esperanza de esta distribución es:

$$\mathbb{E}(\hat{b}) = \frac{N}{N+1}b,$$

es decir, este estimador del parámetro b no es insesgado, pero sí lo es asintóticamente ya que

$$\lim_{N \rightarrow \infty} \mathbb{E}(\hat{b}) = \lim_{N \rightarrow \infty} \mathbb{E}(\max(X_1, X_2, \dots, X_N)) = \lim_{N \rightarrow \infty} \frac{N}{N+1}b = b.$$

Este estimador es consistente, en efecto:

$$\begin{aligned} \mathbb{E}(\hat{b}^2) &= \frac{N}{N+2}b^2, \quad \mathbb{E}(\hat{b}) = \frac{N}{N+1}b \\ \text{Var}(\hat{b}) &= \left(\frac{N}{N+2} - \frac{N^2}{(N+1)^2} \right) b^2 = \frac{Nb^2}{(N+2)(N+1)^2} \end{aligned}$$

por lo tanto su varianza tiende a cero cuando N tiende a infinito.

2. Dada una muestra aleatoria de distribución de Bernoulli de parámetro p , estimemos p por el método de los momentos y por el método de máxima verosimilitud:

(a) Método de Máxima Verosimilitud:

si X tiene distribución de Bernoulli de parámetro p , x_1, x_2, \dots, x_N es una muestra de la variable aleatoria X y f es la frecuencia correspondiente al número de unos presentes en la muestra, se tendrá:

$$\begin{aligned} L(x_1, x_2, \dots, x_N; p) &= p^f (1-p)^{N-f} \\ l(p) = \ln L(\vec{x}; p) &= f \ln p + (N-f) \ln(1-p) \\ \frac{\partial l(p)}{\partial p} &= \frac{f}{p} + \frac{N-f}{1-p} = 0 \Rightarrow p = \frac{f}{N}. \end{aligned}$$

De aquí se deduce que el estimador de máxima verosimilitud de p es:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^N X_i = \bar{X}.$$

(b) Método de los Momentos: debemos resolver la siguiente ecuación:

$$\mathbb{E}(X) = \frac{1}{N} \sum_{i=1}^N X_i$$

pero la esperanza de la distribución de Bernoulli de parámetro p , es p . Así, el estimador del parámetro p , en ambos casos es \bar{X} .

Este estimador es insesgado ya que

$$\mathbb{E}(\bar{X}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}(X_i) = p$$

y también es consistente pues:

$$Var(\bar{X}) = Var\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{1}{N^2} \sum_{i=1}^N Var(X_i) = \frac{p(p-1)}{N}$$

por lo tanto la varianza del estimador tiende a cero cuando N tiende a infinito.

EJERCICIO: Hallar el estimador del parámetro $\lambda > 0$, por el método de los momentos correspondiente a:

- a) La distribución exponencial.
- b) La distribución de Poisson.

1.6 Una cota inferior para el error cuadrático medio de un estimador: Desigualdad de Crámer-Rao.

Supongamos que X_1, X_2, \dots, X_N es una muestra aleatoria de distribución

$$F = F(x; \beta)$$

y que dicha distribución tiene una densidad $f(x, \beta)$ derivable respecto al parámetro $\beta \in \mathbb{R}$.

Denotemos por:

$$f(\vec{x}; \beta) = f_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \beta)$$

la densidad conjunta de la muestra aleatoria evaluada en el punto

$$\vec{x} = (x_1, x_2, \dots, x_N).$$

Supongamos también que la identidad

$$\int_{\mathbb{R}^N} f_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \beta) dx_1 dx_2 \dots dx_N = 1$$

puede derivarse respecto al parámetro β , bajo el signo de integral. Bajo esta hipótesis, podemos obtener la siguiente identidad:

$$0 = \int_{\mathbb{R}^N} \frac{\partial f(\vec{x}; \beta)}{\partial \beta} dx_1 dx_2 \dots dx_N = \int_{\mathbb{R}^N} \frac{1}{f(\vec{x}; \beta)} \frac{\partial f(\vec{x}; \beta)}{\partial \beta} f(\vec{x}; \beta) dx_1 dx_2 \dots dx_N = \\ \mathbb{E} \left(\frac{\partial \ln f(\vec{X}; \beta)}{\partial \beta} \right) = \mathbb{E} \left(\frac{\partial \ln L(\beta)}{\partial \beta} \right), \quad \vec{X} = (X_1, X_2, \dots, X_N).$$

Si

$$b(\beta) = \mathbb{E}(T(X_1, X_2, \dots, X_N) - \beta)$$

es el sesgo del estimador T del parámetro β y calculamos su derivada, se tendrá

$$1 + b'(\beta) = \frac{\partial}{\partial \beta} \mathbb{E}(T(X_1, X_2, \dots, X_N)) = \frac{\partial}{\partial \beta} \int_{\mathbb{R}^N} T(\vec{x}) f(\vec{x}; \beta) dx_1 dx_2 \dots dx_N$$

Si admitimos que esta última integral se puede derivar bajo el signo de integral respecto al parámetro β , obtenemos las siguientes igualdades:

$$1 + b'(\beta) = \int_{\mathbb{R}^N} T(\vec{x}) \frac{\partial}{\partial \beta} f(\vec{x}; \beta) dx_1 dx_2 \dots dx_N = \\ \int_{\mathbb{R}^N} T(\vec{x}) \frac{1}{f(\vec{x}; \beta)} \frac{\partial f(\vec{x}; \beta)}{\partial \beta} f(\vec{x}; \beta) dx_1 dx_2 \dots dx_N = \\ \int_{\mathbb{R}^N} T(\vec{x}) \frac{\partial \ln f(\vec{X}; \beta)}{\partial \beta} f(\vec{x}; \beta) dx_1 dx_2 \dots dx_N = \\ \mathbb{E} \left(T(\vec{X}) \frac{\partial \ln L(\beta)}{\partial \beta} \right) = \mathbb{E} \left(T(\vec{X}) \frac{\partial \ln L(\beta)}{\partial \beta} \right) - \beta \mathbb{E} \left(\frac{\partial \ln L(\beta)}{\partial \beta} \right) = \\ \mathbb{E} \left[\left(T(\vec{X}) - \beta \right) \frac{\partial \ln L(\beta)}{\partial \beta} \right] \\ \text{puesto que } \mathbb{E} \left(\frac{\partial \ln L(\beta)}{\partial \beta} \right) = 0.$$

1.6.1 Desigualdad de Crámer-Rao:

Si todas las derivaciones bajo el signo de integral son válidas (es cierto bajo la hipótesis de suficiente regularidad de la densidad f), se cumple la desigualdad:

$$(1 + b'(\beta))^2 \leq \mathbb{E} \left[\left(T(\vec{X}) - \beta \right)^2 \right] \mathbb{E} \left[\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right],$$

por la desigualdad de Cauchy-Schwarz, en particular, si T es insesgado

$$\begin{aligned} b(\beta) = 0, b'(\beta) = 0, \mathbb{E} \left[\left(T(\vec{X}) - \beta \right)^2 \right] &= \text{Var}(T(\vec{X})) \\ 1 &\leq \text{Var}(T(\vec{X})) \mathbb{E} \left[\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right]. \end{aligned}$$

Además, bajo condiciones de suficiente regularidad de la densidad f , se cumple

$$\mathbb{E} \left[\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right] = -\mathbb{E} \left[\frac{\partial^2 \ln L(\beta)}{\partial^2 \beta} \right]$$

ya que:

$$\begin{aligned} \frac{\partial^2 \ln L(\beta)}{\partial^2 \beta} &= \frac{\partial}{\partial \beta} \left(\frac{\partial \ln L(\beta)}{\partial \beta} \right) = \frac{\partial}{\partial \beta} \left(\frac{1}{L(\beta)} \frac{\partial L(\beta)}{\partial \beta} \right) = \\ &= \frac{1}{L(\beta)} \frac{\partial^2 L(\beta)}{\partial^2 \beta} - \left(\frac{1}{L(\beta)} \frac{\partial L(\beta)}{\partial \beta} \right)^2 = \frac{1}{L(\beta)} \frac{\partial^2 L(\beta)}{\partial^2 \beta} - \left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \\ \mathbb{E} \left[\frac{\partial^2 \ln L(\beta)}{\partial^2 \beta} \right] &= -\mathbb{E} \left[\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right] \quad \text{si } 0 = \int_{\mathbb{R}^N} \frac{\partial f(\vec{x}, \beta)}{\partial \beta} dx_1 dx_2 \cdots dx_N \end{aligned}$$

puesto que

$$\begin{aligned} \mathbb{E} \left[\frac{1}{L(\beta)} \frac{\partial^2 L(\beta)}{\partial^2 \beta} \right] &= \int_{\mathbb{R}^N} \frac{1}{f(\vec{x}, \beta)} \frac{\partial^2 f(\vec{x}, \beta)}{\partial^2 \beta} f(\vec{x}, \beta) dx_1 dx_2 \cdots dx_N = \\ \int_{\mathbb{R}^N} \frac{\partial^2 f(\vec{x}, \beta)}{\partial^2 \beta} dx_1 dx_2 \cdots dx_N &= \frac{\partial}{\partial \beta} \int_{\mathbb{R}^N} \frac{\partial f(\vec{x}, \beta)}{\partial \beta} dx_1 dx_2 \cdots dx_N = 0 \end{aligned}$$

si las derivaciones bajo el signo de integral son posibles.

En conclusión, la desigualdad obtenida se denomina desigualdad de Crámer-Rao:

$$\left[\mathbb{E} \left(\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right) \right]^{-1} \leq \mathbb{E} \left[\left(T(\vec{X}) - \beta \right)^2 \right]$$

OBSERVACIONES:

1. La desigualdad de Crámer-Rao proporciona una cota inferior del error cuadrático medio de un estimador. En particular, para los estimadores insesgados, proporciona una cota inferior para la varianza del estimador. Esta cota inferior no tiene por qué ser alcanzada, pero si se encuentra un estimador insesgado cuya varianza es:

$$\left[\mathbb{E} \left(\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right) \right]^{-1}$$

entonces la desigualdad expresa que se trata de un estimador de mínima varianza.

2. Valen resultados análogos a los anteriores cuando la distribución es discreta, reemplazando la densidad $f(\vec{x}, \beta)$ por la función

$$\mathbb{P}(\vec{X} = \vec{x}; \beta)$$

que representa la función de probabilidad de la variable aleatoria X .

3. Se llama eficiencia de un estimador insesgado al cociente de varianzas que define su eficiencia relativa respecto a un eventual estimador de varianza mínima, es decir:

$$\text{Eficiencia de } T = \frac{\left[\mathbb{E} \left(\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right) \right]^{-1}}{\text{Var}(T)}$$

4. Un estimador insesgado de varianza mínima tiene eficiencia igual a 1; tal estimador suele llamarse Estimador Eficiente. Si la sucesión de eficiencias de una sucesión de estimadores insesgados tiende a 1, la sucesión se dice que es asintóticamente eficiente.

EJEMPLO: El estimador de máxima verosimilitud del parámetro λ de la distribución de Poisson tiene varianza mínima, es decir, es un estimador eficiente; en efecto:

$$\ln L(\lambda) = \sum_{i=0}^r f_i \ln \left(\frac{\lambda^i e^{-\lambda}}{i!} \right)$$

donde f_i , $0 \leq i \leq r$ son las frecuencias de los valores $0, 1, 2, \dots, r$ representados en la muestra x_1, x_2, \dots, x_N con $\max(x_1, x_2, \dots, x_N) = r$. Derivando con respecto al parámetro e igualando a cero, obtenemos que el estimador de máxima verosimilitud del parámetro λ es

$$\hat{\lambda} = \bar{X} = \frac{1}{N} \sum_{i=1}^N X_i,$$

que es un estimador insesgado de λ pues

$$\mathbb{E}(\bar{X}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}(X_i) = \lambda,$$

además la cota de Crámer-Rao coincide con la varianza del estimador que es:

$$Var(\bar{X}) = \frac{\lambda}{N}$$

pues:

$$\mathbb{E} \left(\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right) = \frac{N^2}{\lambda^2} \mathbb{E} \left((\bar{X} - \lambda)^2 \right) = \frac{N^2}{\lambda^2} Var(\bar{X}) = \frac{N}{\lambda}$$

por lo tanto:

$$\left[\mathbb{E} \left(\left(\frac{\partial \ln L(\beta)}{\partial \beta} \right)^2 \right) \right]^{-1} = \frac{\lambda}{N} = Var(\bar{X}),$$

puesto que, $\hat{\lambda} = \bar{X}$ tiene eficiencia 1 es un estimador de varianza mínima.

1.7 Estadísticos Suficientes.

Sea X_1, X_2, \dots, X_N una muestra aleatoria cuya distribución es conocida y queremos estimar un parámetro θ de su distribución. Un estadístico $T_n = T(X_1, X_2, \dots, X_N)$ es suficiente para el parámetro desconocido θ si para todos los resultados posibles $T = t$ la distribución condicional de (X_1, X_2, \dots, X_N) dado $T = t$ no es función del parámetro θ . Es decir, toda la información acerca del parámetro θ que puede ser extraída de la muestra X_1, X_2, \dots, X_N está contenida en T .

Ejemplo:

Sean X_1, X_2 una muestra aleatoria con distribución de Poisson de parámetro λ . Consideremos el estadístico

$$T = 2X_1 + X_2$$

Este estadístico no es suficiente para λ :

$$\begin{aligned} \mathbb{P}((X_1, X_2) = (1, 1) \mid T = 3) &= \frac{\mathbb{P}((X_1, X_2)=(1,1), T=3)}{\mathbb{P}(T=3)} = \\ &= \frac{\mathbb{P}((X_1, X_2)=(1,1))}{\mathbb{P}(T=3)} = \\ &= \frac{\mathbb{P}(X_1=1)\mathbb{P}(X_2=1)}{\mathbb{P}(X_1=0)\mathbb{P}(X_2=3) + \mathbb{P}(X_1=1)\mathbb{P}(X_2=1)} = \\ &= \frac{e^{-2\lambda}\lambda^2}{e^{-2\lambda}\lambda^3/6 + e^{-2\lambda}\lambda^2} = \frac{6}{\lambda+6} \end{aligned}$$

Es muy útil conectar la idea de suficiencia con la de factorización de la función de verosimilitud asociada a una muestra, el siguiente teorema que solo enunciaremos nos da un criterio para la suficiencia relacionado con la función de verosimilitud:

Teorema de factorización:

Sea X_1, X_2, \dots, X_N una muestra aleatoria, $T_N = T(X_1, X_2, \dots, X_N)$ un estimador de un parámetro desconocido $\theta \in \Theta$ de la distribución. T es suficiente si y solo si existe $h : \mathbb{R}^n \rightarrow [0, \infty)$ que no depende de θ $\varphi : \Theta \times \mathbb{R} \rightarrow [0, \infty)$ tal que

$$L_\theta = h(x_1, \dots, x_N) \varphi(\theta, T(x_1, \dots, x_N))$$

donde L_θ es la función de verosimilitud de la muestra.

Ejemplo:

Sea X_1, X_2, \dots, X_N una muestra aleatoria de distribución exponencial de parámetro θ . Sea

$$T_N = \frac{X_1 + X_2 + \dots + X_N}{N}$$

donde la densidad de X_i viene dada por

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}; x > 0$$

Así la función de verosimilitud del parámetro desconocido es

$$L_\theta = \frac{1}{\theta^N} e^{-\frac{x_1 + \dots + x_N}{\theta}} = \frac{1}{\theta^N} e^{-\frac{N\bar{x}}{\theta}}; x_1, x_2, \dots, x_N > 0$$

el estadístico $T = \bar{X}$ es suficiente para θ , defina

$$\begin{aligned}\varphi(\theta, t) &= \frac{1}{\theta^N} e^{\frac{-Nt}{\theta}}, t \geq 0 \\ h &= 1\end{aligned}$$

así

$$L_\theta = h(x_1, \dots, x_N) \varphi(\theta, T(x_1, \dots, x_N))$$

■

Observación: Los estadísticos suficientes no son únicos en el sentido que si $\sum_{i=1}^n x_i$ es suficiente para λ en un modelo de Poisson, también lo será $\frac{1}{n} \sum_{i=1}^n x_i$ o $\frac{1}{8} \sum_{i=1}^n x_i$. Algunas de estas funciones tendrán buenas propiedades como estimadores del parámetro entonces las llamaremos estimadores suficientes.

**DISTRIBUCIONES DE PROBABILIDAD.
ESTADÍSTICA
PRÁCTICA N^o4**

1. Determine la varianza de la distribución de Poisson basándose en su función generatriz de momentos (o transformada geométrica).
2. Sea X una variable aleatoria de densidad exponencial, de parámetro $\lambda = 1$.

Determine la función de densidad de

$$Y = X^3$$

3. Sean X e Y dos variables aleatorias independientes, f y g las densidades de X e Y respectivamente, con $X > 0$. Calcule la densidad de la variable aleatoria

$$Z = \frac{Y}{X}$$

(Sugerencia: Expresa $F_Z(z) = \mathbb{P}\left(\frac{Y}{X} \leq z\right)$ como una integral doble y luego derive respecto a z).

4. DISTRIBUCIÓN NORMAL BIDIMENSIONAL:

Diremos que dos variables aleatorias X e Y tienen distribución normal conjunta, si su función de densidad conjunta viene dada por:

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_1)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right]$$

donde ρ es el coeficiente de correlación entre X e Y , σ_1^2 es la varianza de X , σ_2^2 es la varianza de Y , μ_1 es la esperanza de X y μ_2 es la esperanza de Y .

(a) Sea

$$C = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

la matriz de covarianza de X e Y , donde $\sigma_{12} = \text{Cov}(X, Y)$. Calcule la inversa de C .

(b) Verifique que

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho}} \exp \left[-\frac{1}{2} (zC^{-1}z^t) \right]$$

donde $z = ((x - \mu_1), (y - \mu_1))$ y z^t es la traspuesta de z y C^{-1} es la inversa de C .

(c) Calcule las densidades marginales de X e Y .

(d) Si X e Y son independientes, el coeficiente de correlación es cero. El recíproco no es cierto, en general. Demuestre que si X e Y tienen distribución conjunta normal y $\rho = 0$, entonces X e Y son independientes.

5. DISTRIBUCIÓN NORMAL MULTIDIMENSIONAL:

Sea $\vec{X} = (X_1, X_2, X_3, \dots, X_n)$ un vector aleatorio. Si la matriz de covarianzas C de \vec{X} tiene determinante distinto de cero, diremos que la distribución de \vec{X} es normal de dimensión n , cuando la densidad conjunta de $X_1, X_2, X_3, \dots, X_n$ viene dada por:

$$f(x_1, x_2, x_3, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\Delta}} \exp \left[-\frac{1}{2} (zC^{-1}z^t) \right]$$

donde Δ es el determinante de la matriz de covarianzas C ,

$$z = ((x_1 - \mu_1), (x_2 - \mu_2), \dots, (x_n - \mu_n)), \mu_i = \mathbb{E}(X_i), i = 1, 2, \dots, n$$

y C^{-1} es la inversa de C .

(a) Si $\sigma_{ij} = Cov(X_i X_j) = 0$, para i distinto de j , entonces C es una matriz diagonal tal que

$$Diag(C) = (\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$$

Calcule C^{-1} .

(b) Demuestre que si $\sigma_{ij} = Cov(X_i X_j) = 0$, para i distinto de j , entonces $X_1, X_2, X_3, \dots, X_n$ son mutuamente independientes.

Observación: la parte b) establece que un vector normal aleatorio tiene componentes mutuamente independientes sí y sólo si $Cov(X_i X_j) = 0$ para i distinto de j . Es decir, si y sólo si la matriz de covarianzas es diagonal.

6. Sea X una variable aleatoria con distribución normal $N(\mu, \sigma^2)$.

(a) Demuestre que su transformada de Laplace viene dada por:

$$L(s; \mu, \sigma) = e^{-\mu s} e^{\frac{\sigma^2 s^2}{2}},$$

deduzca de aquí su Transformada de Fourier (o Función Característica).

(b) Sea $Z = X^2$, donde X es normal $N(0, 1)$. Calcule la transformada de Laplace de Z y deduzca que su distribución es gamma de parámetros $p = \frac{1}{2}$, $\lambda = \frac{1}{2}$; es decir, que su función de densidad viene dada por:

$$f_Z(z) = \frac{\sqrt{1/2} z^{-1/2} e^{-\frac{z}{2}}}{\Gamma(1/2)}$$

donde $\Gamma(1/2) = \sqrt{\pi}$. Puede calcular la densidad de Z sin calcular su transformada de Laplace?

(c) Sean $X_1, X_2, X_3, \dots, X_n$ variables aleatorias independientes e idénticamente distribuidas, normales $N(0, 1)$. Sea

$$Y = X_1^2 + X_2^2 + X_3^2 + \dots + X_n^2.$$

Calcule la transformada de Laplace de Y . Deduzca que Y tiene distribución gamma de parámetros $p = \frac{n}{2}$, $\lambda = \frac{1}{2}$, es decir, que la densidad de Y viene dada por:

$$f_Y(y) = \frac{(1/2)^{\frac{n}{2}} y^{-(n-2)/2} e^{-\frac{y}{2}}}{\Gamma(n/2)}$$

Esta distribución recibe el nombre de chi-cuadrado (χ^2) con n grados de libertad.

7. Sea X una variable aleatoria $N(0, 1)$. Y una variable aleatoria chi-cuadrado (χ^2) con n grados de libertad. Si X e Y son independientes, definimos T como:

$$T = \sqrt{n} \frac{X}{\sqrt{Y}}.$$

Demuestre que la densidad de T viene dada por:

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(n/2)} \frac{1}{((1 + \frac{t^2}{n})^{\frac{n+1}{2}})}, \quad t \in (-\infty, \infty)$$

La distribución de T recibe el nombre de distribución t -Student con n grados de libertad (o de parámetro n). Note que f_T es simétrica alrededor del origen ($\mathbb{E}(T) = 0$).

Sugerencia: calcule $F_t(t) = \mathbb{P}(T \leq t) = \mathbb{P}\left(X \leq \frac{t}{\sqrt{n}}\sqrt{Y}\right)$, utilizando la densidad conjunta de X e Y , luego derive respecto a t .

8. Sean $X_1, X_2, X_3, \dots, X_n$ variables aleatorias independientes, normales $N(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots, n$. Demuestre que la variable aleatoria:

$$X = a_1X_1 + a_2X_2 + a_3X_3 + \dots + a_nX_n$$

tiene distribución $N(\mu, \sigma^2)$, donde

$$\mu = a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + \dots + a_n\mu_n, \sigma^2 = \sum_{i=1}^n a_i^2\sigma_i^2$$

9. Sean $X_1, X_2, X_3, \dots, X_n$ variables aleatorias independientes e idénticamente distribuidas, normales $N(\mu, \sigma^2)$.

- (a) Sea $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Demuestre que \bar{X} tiene distribución normal $N(\mu, \sigma^2/n)$.

Sugerencia: Calcule la transformada de Fourier de \bar{X} .

- (b) Sea $Z = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$. Demuestre que Z tiene distribución \mathcal{X}^2 con n grados de libertad.

- (c) Sea $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$. Verifique que:

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{nS^2}{\sigma^2}.$$

El objetivo de este ejercicio es demostrar que la variable aleatoria $\frac{nS^2}{\sigma^2}$, tiene distribución \mathcal{X}^2 con $n - 1$ grados de libertad. Observe que en la definición de S^2 si sustituimos las variables aleatorias X_i por $Z_i = X_i - \mu$, el valor de S^2 no varía.

Definamos las siguientes variables aleatorias:

$$\begin{aligned} Y_1 &= \frac{1}{\sqrt{1 \cdot 2}}(Z_1 - Z_2) \\ Y_2 &= \frac{1}{\sqrt{2 \cdot 3}}(Z_1 + Z_2 - 2Z_3) \cdots \\ Y_{n-1} &= \frac{1}{\sqrt{(n-1) \cdot n}}(Z_1 + Z_2 + Z_3 + \cdots + Z_{n-1} - (n-1)Z_n) \\ Y_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i. \end{aligned}$$

- i. Demuestre que las variables aleatorias $Y_1, Y_2, Y_3, \dots, Y_n$ tienen distribución $N(0, \sigma^2)$.
- ii. Verifique, calculando las correlaciones entre ellas, que son independientes.
- iii. Establezca por inducción que

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Z_i^2$$

(La transformación definida en c , deja invariante el origen).

- iv. Demuestre que $nS^2 = \sum_{i=1}^{n-1} Y_i^2$.
 - v. Deduzca que $\frac{nS^2}{\sigma^2}$ tiene distribución χ^2 con $n-1$ grados de libertad.
- (d) Note que $\bar{X} = \frac{1}{\sqrt{n}}Y_n + \mu$. Deduzca que S^2 y \bar{X} son independientes.
- (e) Sea $S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- i. Verifique que:

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)S_1^2}{\sigma^2}$$

- ii. Sean

$$X = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma}; Y = \frac{(n-1)S_1^2}{\sigma^2}.$$

Verifique que X es normal estándar e Y es χ^2 con $n-1$ grados de libertad. Además X e Y son independientes.

- iii. Verifique que:

$$\frac{\bar{X} - \mu}{S_1} \sqrt{n} = \frac{\sqrt{n-1}X}{\sqrt{Y}}$$

y deduzca de esta igualdad que la distribución de

$$\frac{\bar{X} - \mu}{S_1} \sqrt{n}$$

es t - student con $(n - 1)$ grados de libertad.

Abril-Julio 2004/MMOM

ESTIMADORES PUNTUALES
ESTADÍSTICA -PRÁCTICA N° 5

1. Sea $\hat{\theta}$ un estimador de un parámetro θ . Sea b el sesgo de $\hat{\theta}$, es decir,

$$b = \mathbb{E} \left(\hat{\theta} \right) - \theta.$$

Demuestre que el error cuadrático medio de $\hat{\theta}$ es igual a:

$$Var(\hat{\theta}) + b^2.$$

2. Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados de un parámetro θ tales que

$$Var(\hat{\theta}_1) = \sigma_1^2, Var(\hat{\theta}_2) = \sigma_2^2$$

Demuestre que si a es un número real, entonces:

$$\hat{\theta}_3 = a\hat{\theta}_1 + (1 - a)\hat{\theta}_2$$

es un estimador insesgado de θ . Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son independientes, cómo se debe escoger a para minimizar la varianza de $\hat{\theta}_3$?

3. Sea Y_1, Y_2, Y_3 una muestra aleatoria simple de una distribución exponencial de densidad:

$$f(y) = \begin{cases} \frac{1}{\theta} e^{-\frac{y}{\theta}}, & y > 0 \\ 0 & \text{si no} \end{cases}$$

Considere los 5 siguientes estimadores de θ :

$$\hat{\theta}_1 = Y_1, \hat{\theta}_2 = \frac{Y_1 + Y_2}{2}, \hat{\theta}_3 = \frac{Y_1 + 2Y_2}{3}, \hat{\theta}_4 = \min(Y_1, Y_2, Y_3), \hat{\theta}_5 = \bar{Y}$$

- (a) Cuáles estimadores son sesgados?
(b) Entre estos estimadores, cuál es el que tiene la varianza más pequeña?

- (c) Hallar la eficiencia relativa de $\hat{\theta}_1$ respecto a $\hat{\theta}_5$, la de $\hat{\theta}_2$ y $\hat{\theta}_3$ respecto a $\hat{\theta}_5$.
4. El número de fallas por semanas de un cierto tipo de mini-computadoras es una variable aleatoria Y con distribución de Poisson de parámetro λ . Se dispone de una muestra aleatoria simple $Y_1, Y_2, Y_3, \dots, Y_n$ de Y .
- (a) Sugiera dos estimadores insesgados para λ .
- (b) El costo semanal de reparación de estas fallas es la v.a.

$$C = 3Y + Y^2$$

Demuestre que

$$\mathbb{E}(C) = 4\lambda + \lambda^2$$

- (c) Obtenga una función de $Y_1, Y_2, Y_3, \dots, Y_n$, que sea un estimador insesgado de $\mathbb{E}(C)$.
5. Sea $X_1, X_2, X_3, \dots, X_n$ una muestra aleatoria simple de una distribución de Bernoulli de parámetro p .
- (a) Demuestre que \bar{X} es un estimador insesgado de p .
- (b) Considere el estimador

$$n\bar{X}(1 - \bar{X})$$

¿Es éste un estimador insesgado de la varianza de la distribución?

- (c) Modifique adecuadamente el estimador anterior para obtener un estimador insesgado de la varianza.
6. Sea $X_1, X_2, X_3, \dots, X_n$ una muestra aleatoria simple de una distribución normal $N(\mu_1, \sigma_1^2)$ y $Y_1, Y_2, Y_3, \dots, Y_m$ una muestra aleatoria simple de una distribución normal $N(\mu_2, \sigma_2^2)$, supongamos que las X_i, Y_j , son independientes entre sí.
- (a) Considere el estadístico $\bar{X} - \bar{Y}$ y encuentre su distribución, su media y su varianza. Es éste un estimador insesgado de $\mu_1 - \mu_2$?

(b) Sean:

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, S_2^2 = \frac{1}{m-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$S_p = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

S_p se puede interpretar como una “ponderación” de S_1^2 y S_2^2 . En el caso $n = m$, $S_p = \frac{S_1^2 + S_2^2}{2}$. Demuestre que si $\sigma_1 = \sigma_2 = \sigma$, entonces S_p es un estimador insesgado de σ^2 .

7. Se utiliza el siguiente procedimiento para evitar respuestas falsas a preguntas delicadas en una encuesta. Sea A una pregunta delicada (por ejemplo, evade Ud. el pago de impuestos?). Sea B una pregunta inocua (por ejemplo, su cédula de identidad termina en un número par?). Se le pide al sujeto que lance una moneda en secreto; si sale cara, contesta la pregunta A , si sale sello contesta la pregunta B . El encuestador recibe una sola respuesta (sí o no) y no sabe a qué pregunta corresponde. Si esta encuesta se realiza a 1000 sujetos y 600 de ellos contestaron “sí”, qué porcentaje de individuos se estima que evade impuesto?
8. Sea $X_1, X_2, X_3, \dots, X_n$ una muestra aleatoria simple de una distribución $N(\mu, \sigma)$. Consideremos los estimadores de μ del tipo siguiente:

$$U = a_1 X_1 + a_2 X_2 + a_3 X_3 + \dots + a_n X_n$$

donde $a_1, a_2, a_3, \dots, a_n$ son números reales. Determine los a_i para que U sea un estimador insesgado y tenga varianza mínima.

9. Demuestre que si $\hat{\theta}$ es un estimador insesgado de θ y si $Var(\hat{\theta})$ no es igual a cero, entonces $\hat{\theta}^2$ no es un estimador insesgado de θ^2 .
10. Demuestre que la media muestral \bar{X} es un estimador insesgado de varianza mínima del parámetro λ de una población de Poisson.

M.M.O.M./Abril 2004

Métodos de Estimación.
Estadística
Práctica N°6

1. Dada una muestra aleatoria simple de tamaño n de una variable aleatoria X , calcular el estimador de máxima verosimilitud y de los momentos cuando X tiene las siguientes distribuciones:

- (a) Bernoulli de parámetro p .
- (b) Poisson de parámetro λ .
- (c) Exponencial de parámetro λ .
- (d) $N(\mu, \sigma^2)$ con μ y σ desconocido.
- (e) $N(\mu, \sigma^2)$ con μ conocido y σ desconocido.
- (f) $N(\mu, \sigma^2)$ con μ desconocido y σ conocido.

Hallar en cada caso las propiedades del estimador obtenido: sesgo, consistencia, eficiencia.

2. Hallar por el método de los momentos los estimadores de α y β para la función Gamma $\Gamma(\alpha, \beta)$.
3. La muestra 1.3, 0.6, 1.7, 2.2, 0.3, 1.1 proviene de una distribución uniforme en $[0, b]$. Encontrar los valores numéricos de los estimadores de b obtenidos por los métodos de máxima verosimilitud y de los momentos.
4. El número de defectos congénitos de un individuo en una cierta población sigue una distribución de Poisson de parámetro λ . De una muestra de $n = 50$ individuos de la población, se observaron los siguientes datos:

Nºde defectos	1	2	3	4
Frecuencias	31	15	4	0

Hallar el estimador de máxima verosimilitud del parámetro λ .

5. Una variable aleatoria discreta toma los valores 0, 1 y 2 con:

$$\mathbb{P}(0) = p^2, \mathbb{P}(1) = 2p(1 - p), \mathbb{P}(2) = (1 - p)^2,$$

donde $0 < p < 1$ es un parámetro desconocido. Hallar la estimación máximo verosímil de p a partir de una muestra de tamaño $n = 100$ en la que se ha presentado 23 veces el 0, 52 veces el 1 y 25 veces el 2.

6. Para determinar la verdadera proporción p de artículos defectuosos en un lote grande, se revisan uno tras otro artículos de éste, hasta encontrar 10 defectuosos. Si tuvimos que revisar 168 artículos antes de parar, cuál es el estimador de máxima verosimilitud de p ?
7. Sea Y_1, Y_2, \dots, Y_n una muestra aleatoria simple de una distribución uniforme en el intervalo $[0, 2\theta + 1]$, con $\theta > 0$ es decir, con densidad:

$$f(y) = \begin{cases} \frac{1}{2\theta+1}, & 0 < y < 2\theta + 1 \\ 0 & , \quad y \leq 0 \end{cases}$$

Encuentre el estimador de máxima verosimilitud de θ .

8. Sea Y_1, Y_2, \dots, Y_n una muestra aleatoria simple de una distribución con la siguiente densidad:

$$f(y) = \begin{cases} \frac{2(\theta-y)}{\theta^2}, & 0 < y < \theta \\ 0 & , \quad y \leq 0 \end{cases}$$

Determinar el valor de θ por el método de los momentos.

9. Supongamos que X_1, X_2, \dots, X_n y Y_1, Y_2, \dots, Y_n son muestras aleatorias simples de dos poblaciones normales independientes con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 respectivamente. Determinar el estimador de máxima verosimilitud de $\mu_1 - \mu_2$.

Sugerencia: Considere la muestra $Z_i = X_i - Y_i$, halle su distribución y escriba su función de verosimilitud.

10. Una máquina se puede averiar por dos razones A y B . Se desea estimar la probabilidad de avería diaria de cada tipo sabiendo que:
 - (a) La probabilidad de avería tipo A es doble que la de B .
 - (b) No existen otros tipos de avería posible.
 - (c) Se han observado 30 días con el resultado siguiente: 2 averías tipo A , 3 tipo B y 25 días sin avería.

Respuesta: $p_A = \frac{1}{18}, p_B = \frac{2}{18}, p_C = \text{probabilidad de no avería} = \frac{15}{18}$.

11. Demostrar que \bar{X} es un estadístico suficiente para λ en un modelo de Poisson.

12. Sea $f(x) = \alpha x^{\alpha-1}, x \in (0, 1)$. Hallar el estimador de máximo verosímil para α y verificar que es un estadístico suficiente.
13. Basándose en la igualdad

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n (\bar{x} - \mu)^2$$

demuestre que la función de verosimilitud de un modelo normal $N(\mu, \sigma^2)$ se puede escribir

$$L(\mu, \sigma) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp \left[-\frac{n}{2\sigma^2} (s^2 + (\bar{x} - \mu)^2) \right]$$

concluyendo que el estadístico vectorial $T_n = T(X_1, \dots, X_n) = (\bar{X}, S^2)$ es suficiente para

$(\mu, \sigma) \in \Theta \subset \mathbb{R} \times \mathbb{R}^+$ (una familia paramétrica). Sugerencia: tome $h(x_1, \dots, x_n) = 1$, para obtener

$$L(\mu, \sigma) = h(x_1, \dots, x_n) \varphi(\mu, \sigma, T(x_1, \dots, x_n))$$

$$\varphi : \Theta \times \mathbb{R}^2 \rightarrow \mathbb{R}$$

MMOM/2004

Capítulo III

Intervalos de Confianza

María Margarita Olivares

Mayo 2004

1 Estimación por Intervalos de Confianza.

1.1 Introducción y conceptos básicos.

Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de una variable aleatoria X cuya distribución depende de uno o varios parámetros desconocidos. Si $T(X_1, X_2, \dots, X_n)$ es un estimador puntual de β (un parámetro desconocido de la distribución), dada una observación x_1, x_2, \dots, x_n , estimamos β por

$$T(x_1, x_2, \dots, x_n).$$

La estimación puntual tiene como inconveniente que al dar como estimación del parámetro desconocido un único valor, la estimación no coincidirá con el verdadero valor del parámetro, luego, para que ésto tenga interés práctico debemos conocer el grado de precisión de la estimación o conformarnos con que la equivocación no sea muy grande.

El concepto de estimación por Intervalo de Confianza proporciona una respuesta a esta cuestión.

Ejemplo:

Supongamos que una tienda mantiene registros sobre el número de unidades de cierto producto que vende mensualmente. El conocimiento de la demanda promedio es importante para el mantenimiento del inventario. Se supondrá que no hay fluctuaciones en la temporada.

En los últimos 36 meses, en base a los datos, se tiene una media $\bar{x} = 200$ unidades. Es decir, esta media es una estimación puntual de un parámetro desconocido, el cual representa la demanda promedio de este producto en

la tienda. Del conocimiento de este estimador no podemos responder por ejemplo, la siguiente pregunta: ‘¿podría la demanda media desconocida no ser mayor a 250 ni menor a 150 unidades?’, pues no se tiene alguna indicación del posible error en la estimación puntual, por eso, no podemos responder esta pregunta. El error en el estimado puntual se mide en términos de la variación (varianza) muestral del correspondiente estimador.

Si por ejemplo, la desviación estándar de la media \bar{X} es 60 unidades, valiéndonos del teorema central del límite podemos argumentar que \bar{X} tiene una distribución aproximadamente $N(\mu, \sigma = 60)$, de aquí se deduce que la probabilidad de que \bar{X} se encuentre a dos desviaciones estándar de la verdadera media es aproximadamente igual a 0,95. Es decir, para n grande,

$$\mathbb{P}\left(\left|\bar{X} - \mu\right| < 120\right) = 0,95$$

o equivalentemente

$$\mathbb{P}\left(\bar{X} - 120 < \mu < \bar{X} + 120\right) = 0,95$$

si sustituimos el valor de la media muestral $\bar{x} = 200$ obtenido se tiene que $\mu \in (80, 320)$ lo cual sugiere que la demanda podría ser tan grande como 250 y tan pequeña como 150 unidades siempre que $\sqrt{Var(\bar{X})} = 60$.

Observe que $(\bar{X} - 120, \bar{X} + 120)$ es un intervalo aleatorio y la probabilidad de que la verdadera media esté allí es de 0,95. Si se obtuviesen 100 muestras del mismo tamaño en forma repetida de una población y para cada muestra se calculan las medias observadas y sustituimos \bar{x} en el intervalo aleatorio debe esperarse que 95 de ellos contengan el verdadero valor de la media μ desconocida. El intervalo específico $(80, 320)$ no es más que una realización del intervalo aleatorio $(\bar{X} - 120, \bar{X} + 120)$ en base a los datos de una sola muestra en la cual sustituimos $\bar{x} = 200$. El valor de probabilidad 0,95 se refiere sólo al intervalo aleatorio $(\bar{X} - 120, \bar{X} + 120)$, es incorrecto decir que la probabilidad de que $\mu \in (80, 320)$ es de 0,95 pues aquí sólo hay constantes, nada es aleatorio, lo que se puede decir que con una confianza de 0,95 podemos decir que $\mu \in (80, 320)$, lo cual es una alta confianza. Así que no es correcto escribir $\mathbb{P}(80 < \mu < 320) = 0,95$, en algunos textos lo escriben abusando del lenguaje probabilístico pero debe entenderse como se explicó arriba. De acuerdo a estas aclaratorias, el intervalo $(80, 320)$ recibe el nombre de intervalo de confianza del 95% para μ .

1.1.1 Definición:

Intervalo de Confianza al nivel $1 - \alpha$ para el parámetro β construido a partir de la muestra X_1, X_2, \dots, X_n , es una pareja de estadísticos (T_1, T_2) , que cumplan la propiedad:

$$\mathbb{P}(T_1(X_1, X_2, \dots, X_n) \leq \beta \leq T_2(X_1, X_2, \dots, X_n)) = 1 - \alpha$$

1.1.2 Definición:

Región de confianza al nivel $1 - \alpha$ para el parámetro β construida a partir de la muestra X_1, X_2, \dots, X_n es una aplicación C que a cada punto

$$\vec{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

asocia el subconjunto $C(\vec{x})$ del espacio de parámetros, de modo que

$$\mathbb{P}(\beta \in C(X_1, X_2, \dots, X_n)) = 1 - \alpha$$

Nótese que $C(X_1, X_2, \dots, X_n)$ es un conjunto aleatorio. En el caso en que el espacio de parámetros esté contenido en \mathbb{R} , si este conjunto es un intervalo, la región de confianza se llama intervalo de confianza y esta definición coincide con la dada anteriormente.

1.1.3 Método práctico para la construcción de intervalos de confianza.

1. Obtener una variable aleatoria $g(X_1, X_2, \dots, X_n; \beta)$ que sea función de la muestra y del parámetro desconocido cuya distribución sea conocida e independiente de β .
2. Obtener un conjunto D en el rango de g para el cual valga

$$\mathbb{P}(g(X_1, X_2, \dots, X_n; \beta) \in D) = 1 - \alpha$$

3. Expresar el suceso $[g(X_1, X_2, \dots, X_n; \beta) \in D]$ en la forma equivalente

$$[\beta \in C(X_1, X_2, \dots, X_n)]$$

La región $C(X_1, X_2, \dots, X_n)$ es una región de confianza al nivel $1 - \alpha$ para β , si ésta resulta ser un intervalo, entonces es un intervalo de confianza.

1.1.4 Observaciones:

1. Si

$$C(X_1, X_2, \dots, X_n) = (T_1(X_1, X_2, \dots, X_n), T_2(X_1, X_2, \dots, X_n))$$

es un intervalo de confianza para β al nivel $1 - \alpha$ a menudo se dice que constituye un intervalo de confianza de $100(1 - \alpha)\%$ o con coeficiente de confianza $100(1 - \alpha)\%$.

2. El coeficiente de confianza $1 - \alpha$ es un valor que elige el experimentador. Suele tomarse $\alpha = 0,10$ (ó $0,05$ ó $0,01$), es decir $100(1 - \alpha)\%$ será 90% ó respectivamente 95% ó 99% .
3. El coeficiente de confianza representa la probabilidad ó proporción de veces que los intervalos conocidos contendrán el verdadero valor de β .
4. El valor desconocido β es constante, mientras que los intervalos obtenidos son aleatorios puesto que sus extremos dependen de la muestra aleatoria. Así, un intervalo de confianza del 95% (ó una estimación por intervalos al nivel $0,95$) no es un intervalo fijo que contiene a β con probabilidad $0,95$, ésto no tendría sentido pues si el intervalo es fijo el parámetro estará o no en él (la probabilidad sería cero ó uno). Luego, al evaluar los extremos del intervalo a partir de la muestra observada x_1, x_2, \dots, x_n , lo que podemos decir es que si se repitieran las n observaciones varias veces al menos el 95% de las veces acertaremos, es decir, β estará en el intervalo obtenido.
5. Si podemos elegir entre varios métodos que dan lugar a diferentes intervalos de confianza para un parámetro desconocido β , trataremos de elegir el que nos proporcione el intervalo de menor longitud.
6. Si hay varios parámetros desconocidos hallaremos regiones de confianza o intervalos de confianza para cada uno de ellos.

1.2 Estimación por intervalos de confianza para la media μ :

1.2.1 Para la distribución $N(\mu, \sigma)$ cuando σ es conocido:

Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de la variable aleatoria X de distribución $N(\mu, \sigma)$, cuando σ es conocido.

Para un nivel de confianza $1 - \alpha$ consideremos el estadístico:

$$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma}$$

el cual tiene distribución normal $N(0, 1)$. Sea $z_{\frac{\alpha}{2}}$ el valor tal que si Z es normal estándar satisface:

$$\mathbb{P}(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

Si α es conocido podemos buscar este valor en la tabla de la distribución normal. Puesto que la distribución normal es simétrica alrededor del origen, se tiene que :

$$\mathbb{P}(Z < -z_{\frac{\alpha}{2}}) = \frac{\alpha}{2},$$

por lo tanto

$$\mathbb{P}(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

Si sustituimos el valor de Z en la expresión anterior y despejando μ , obtenemos:

$$\mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

es decir, con probabilidad $1 - \alpha$, μ se encuentra en el intervalo aleatorio

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}\right);$$

dicho de otra forma: para cada muestra x_1, x_2, \dots, x_n , el intervalo que se obtien como estimación es

$$I = \left(\bar{x} - \frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}\right)$$

con nivel $1 - \alpha$, donde $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Es habitual expresar este resultado diciendo que : “la estimación de μ es \bar{x} con error de $\pm \frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}$ ” para un nivel de confianza $1 - \alpha$.

1.2.2 Ejemplos:

1. A partir de una muestra aleatoria simple X_1, X_2, \dots, X_n de la distribución $N(\mu, 1)$ construir un intervalo de confianza para μ al nivel 95% siendo la media observada

$$\bar{x} = 3,05; \quad n = 100.$$

Usando el procedimiento anteriormente desarrollado, se obtienen que el intervalo de confianza al nivel 95% es:

$$\left(\bar{X} - \frac{z_{0,025}}{10}, \bar{X} + \frac{z_{0,025}}{10} \right)$$

donde $z_{0,025}$ se obtiene en la tabla de la distribución normal, hallando:

$$\mathbb{P}(Z > z_{0,025}) = 1 - \mathbb{P}(Z \leq z_{0,025}) = \frac{\alpha}{2} = 0,025$$

de donde

$$\mathbb{P}(Z \leq z_{0,025}) = 1 - 0,025 = 0,975.$$

De aquí se obtiene que

$$z_{0,025} = 1,96.$$

Luego el intervalo de confianza es

$$\left(\bar{X} - 0,196, \bar{X} + 0,196 \right)$$

Basándonos en nuestras observaciones, sustituimos \bar{X} por su valor observado para obtener como estimación del intervalo

$$I = (2.85, 3.25)$$

al nivel de confianza de 95%.

2. Si en el ejemplo anterior se desea estimar el tamaño necesario de la muestra de manera tal que con probabilidad $1 - \alpha = 0,95$ la media muestral \bar{X} se encuentre en un intervalo igual a $\varepsilon = 0,20$ unidades alrededor de la media μ de la distribución se procede de la siguiente manera: puesto que en este caso

$$\mathbb{P}\left(-\frac{1}{\sqrt{n}}z_{\frac{\alpha}{2}} < \bar{X} - \mu < \frac{1}{\sqrt{n}}z_{\frac{\alpha}{2}}\right) = 1 - \alpha = 0.95$$

y lo que se quiere es que

$$\bar{X} \in (\mu - 0.20, \mu + 0.20)$$

con probabilidad 0.95, tomamos $\varepsilon = \frac{1}{\sqrt{n}} z_{\frac{\alpha}{2}} = 0.20$, hemos calculado en este caso $z_{\frac{\alpha}{2}}$ y obtuvimos 1.96, despejando n en la última ecuación se tiene que

$$n = \left(\frac{1.96}{0.20} \right)^2 = 96.04$$

Si tomamos $n = 97$ podemos confiar en que si estimamos μ por medio de \bar{X} el error será menor que 0.20 para el nivel 0.95.

1.2.3 Para la distribución $N(\mu, \sigma)$ cuando σ es desconocido:

Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de la variable aleatoria X de distribución $N(\mu, \sigma)$, cuando σ es desconocido.

En este caso se sabe que

$$\frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \text{ es } N(0, 1)$$

$$\frac{(n-1)S_1^2}{\sigma^2} \text{ es } \chi_{n-1}^2$$

y que S_1^2 y \bar{X} son independientes, por lo tanto el estadístico:

$$\frac{(\bar{X} - \mu)}{S_1/\sqrt{n}} = \frac{(\bar{X} - \mu)}{S/\sqrt{n-1}}$$

tiene distribución t de Studente con $n - 1$ grados de libertad, donde

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Recordemos que la densidad correspondiente a esta distribución es simétrica alrededor del origen. Es fácil ver que si

$$T_{n-1} = \frac{(\bar{X} - \mu)}{S_1/\sqrt{n}}$$

$$\mathbb{E}(T_{n-1}) = 0, \text{Var}(T_{n-1}) = \frac{n}{n-2}, \lim_{n \rightarrow \infty} \text{Var}(T_{n-1}) = 1.$$

Usaremos una notación análoga al caso anterior, llamando $t_{n-1, \frac{\alpha}{2}}$ el valor tal que si T_{n-1} es t -Student con $n - 1$ grados de libertad

$$\mathbb{P}(T_{n-1} > t_{n-1, \frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

entonces se tendrá que

$$\begin{aligned} \mathbb{P}\left(-t_{n-1, \frac{\alpha}{2}} < \frac{(\bar{X} - \mu)}{S_1/\sqrt{n}} < t_{n-1, \frac{\alpha}{2}}\right) &= 1 - \alpha \\ \mathbb{P}\left(\bar{X} - \frac{S_1}{\sqrt{n}}t_{n-1, \frac{\alpha}{2}} < \mu < \bar{X} + \frac{S_1}{\sqrt{n}}t_{n-1, \frac{\alpha}{2}}\right) &= 1 - \alpha \end{aligned}$$

siendo, el intervalo aleatorio

$$\left(\bar{X} - \frac{S_1}{\sqrt{n}}t_{n-1, \frac{\alpha}{2}}, \bar{X} + \frac{S_1}{\sqrt{n}}t_{n-1, \frac{\alpha}{2}}\right)$$

el intervalo de confianza para μ al nivel $1 - \alpha$.

Así, a cada muestra x_1, x_2, \dots, x_n , le corresponde la estimación por intervalo

$$I = \left(\bar{x} - \frac{s_1}{\sqrt{n}}t_{n-1, \frac{\alpha}{2}}, \bar{x} + \frac{s_1}{\sqrt{n}}t_{n-1, \frac{\alpha}{2}}\right)$$

al nivel $1 - \alpha$. A este nivel “la estimación de μ ” es \bar{x} con error de $\pm \frac{s_1}{\sqrt{n}}t_{n-1, \frac{\alpha}{2}}$.

1.3 Estimación por intervalos de confianza para σ^2 :

1.3.1 Para la distribución $N(\mu, \sigma)$ cuando μ es desconocido:

Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de la variable aleatoria X de distribución $N(\mu, \sigma)$, con μ y σ son desconocidos. Fijemos un nivel $1 - \alpha$ y consideremos el estadístico

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{(n-1)S_1^2}{\sigma^2},$$

el cual tiene distribución chi- cuadrado con $n - 1$ grados de libertad. Esta densidad no es simétrica alrededor del cero. Queremos hallar un intervalo tal que

$$\mathbb{P}(a < \mathcal{X}_{n-1}^2 < b) = 1 - \alpha, \mathbb{P}(\mathcal{X}_{n-1}^2 \geq b) = \mathbb{P}(\mathcal{X}_{n-1}^2 \leq a) = \frac{\alpha}{2}$$

de aquí se deduce que $a = x_{n-1, 1-\frac{\alpha}{2}}^2, b = x_{n-1, \frac{\alpha}{2}}^2$. Así tendremos:

$$\mathbb{P} \left(x_{n-1, 1-\frac{\alpha}{2}}^2 < \frac{(n-1)S_1^2}{\sigma^2} < x_{n-1, \frac{\alpha}{2}}^2 \right) = 1 - \alpha$$

despejando σ^2 obtenemos:

$$\mathbb{P} \left(\frac{(n-1)S_1^2}{x_{n-1, 1-\frac{\alpha}{2}}^2} < \sigma^2 < \frac{(n-1)S_1^2}{x_{n-1, \frac{\alpha}{2}}^2} \right) = 1 - \alpha$$

obteniéndose como intervalo aleatorio de confianza para σ^2

$$\left(\frac{(n-1)S_1^2}{x_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)S_1^2}{x_{n-1, \frac{\alpha}{2}}^2} \right)$$

al nivel $1 - \alpha$. Si queremos estimar σ^2 con S^2 en lugar de $S_1^2 = \frac{n}{n-1}S^2$, obtendremos el intervalo:

$$\left(\frac{nS^2}{x_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{nS^2}{x_{n-1, \frac{\alpha}{2}}^2} \right).$$

1.3.2 Para la distribución $N(\mu, \sigma)$ cuando μ es conocido:

En este caso podemos hallar el intervalo de confianza para σ como lo hicimos para μ cuando σ es conocido, pero en este caso despejamos σ en lugar de μ . Se recomienda estimar σ usando el estadístico del caso anterior el cual no depende de μ .

1.4 Intervalos de Confianza para $\mu_1 - \mu_2$ en el caso Normal:

Un problema importante es la comparación de las medias de dos poblaciones. Por ejemplo, podría desearse comparar la efectividad de dos métodos de enseñanza. Para ésto, los estudiantes se dividirían en dos grupos, con uno de ellos se usaría el método A y con el otro el método B . Entonces habría que hacer inferencias respecto a la diferencia en el aprovechamiento promedio de los estudiantes, medido en base a alguna prueba.

En este caso se consideran dos poblaciones, la primera con media μ_1 y varianza σ_1^2 y la segunda con media μ_2 y varianza σ_2^2 .

Se extrae una muestra aleatoria de n_1 observaciones de la primera población y una muestra aleatoria de n_2 observaciones de la segunda población y se supone que las muestras se extraen independientemente.

Sean X_1, X_2, \dots, X_n la muestra aleatoria de la primera población con distribución $N(\mu_1, \sigma_1)$ y Y_1, Y_2, \dots, Y_n la muestra aleatoria de la segunda población con distribución $N(\mu_2, \sigma_2)$, $\bar{X} - \bar{Y}$ es un estimador puntual de $\mu_1 - \mu_2$, insesgado pues:

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_1 - \mu_2$$

y consistente, pues

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \rightarrow 0, n_1, n_2 \rightarrow \infty.$$

1.4.1 Para la distribución Normal cuando σ_1 y σ_2 son conocidos, al nivel $1 - \alpha$:

El estadístico

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tiene distribución $N(0, 1)$. Tomando $z_{\frac{\alpha}{2}}$ tal que

$$\mathbb{P}(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

se tiene que

$$\mathbb{P}(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

despejando $\mu_1 - \mu_2$, obtenemos:

$$\mathbb{P}\left(\bar{X} - \bar{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha.$$

De aquí que un intervalo aleatorio de confianza para $\mu_1 - \mu_2$ si σ_1 y σ_2 son conocidos, al nivel $1 - \alpha$ es:

$$\left(\bar{X} - \bar{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

1.4.2 Ejemplo:

Se compara la calidad de dos tipos de neumáticos para automóviles probando muestras de $n_1 = n_2 = 100$ neumáticos de cada tipo. Para ésto se observa el número de kilómetros recorridos hasta que se produce un desgaste que se define como una cierta cantidad de desgaste. Se supone $\sigma_1^2 = 1.400.000$, $\sigma_2^2 = 1.960.000$ ambos conocidos y que la distribución de los recorridos es normal. Como resultados de las pruebas se obtuvo $\bar{x} = 26.400Kms.$, $\bar{y} = 25.100Kms.$

Estimemos mediante un intervalo de confianza de 0.99 la diferencia de las medias:

$$1 - \alpha = 0.99, \alpha = 0.01, \frac{\alpha}{2} = 0.005, z_{\frac{\alpha}{2}} = 2,5758$$
$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 183,0303; \bar{X} - \bar{Y} = 1300$$

Así obtenemos como intervalo para $\mu_1 - \mu_2$

$$(827,8481; 1772,1519)$$

al estimar $\mu_1 - \mu_2$ por $\bar{X} - \bar{Y}$ se comete un error de estimación de

$$\pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \pm 472,1519Kms.$$

1.4.3 Para la distribución Normal cuando σ_1 y σ_2 son desconocidos pero iguales, al nivel $1 - \alpha$:

Si

$$\sigma_1 = \sigma_2 = \sigma, X_1, X_2, \dots, X_n \text{ e } Y_1, Y_2, \dots, Y_n$$

son dos muestras independientes de distribuciones $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ respectivamente, consideramos el estadístico:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} S_p}$$

con

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

siendo

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2.$$

El estadístico T tiene distribución t -Student con $n_1 + n_2 - 2$ grados de libertad. Un intervalo de confianza para $\mu_1 - \mu_2$, en este caso se obtiene haciendo:

$$\mathbb{P}\left(-t_{n_1+n_2-2, \frac{\alpha}{2}} \leq T \leq t_{n_1+n_2-2, \frac{\alpha}{2}}\right) = 1 - \alpha$$

donde

$$\mathbb{P}\left(T_{n_1+n_2-2} \geq t_{n_1+n_2-2, \frac{\alpha}{2}}\right) = \frac{\alpha}{2}.$$

Despejando $\mu_1 - \mu_2$ obtenemos:

$$\mathbb{P}\left(\bar{X} - \bar{Y} - S_p t_{n_1+n_2-2, \frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X} - \bar{Y} + S_p t_{n_1+n_2-2, \frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

De aquí se deduce que el intervalo de confianza aleatorio para $\mu_1 - \mu_2$ cuando $\sigma_1 = \sigma_2 = \sigma$ es desconocido será

$$\left(\bar{X} - \bar{Y} - S_p t_{n_1+n_2-2, \frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + S_p t_{n_1+n_2-2, \frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right).$$

1.4.4 Observación:

Si n es grande ($n \geq 30$), las tablas de la distribución t con n grados de libertad dan la aproximación normal. De allí que si $n_1 + n_2 - 2 \geq 30$, resolver el problema para $\sigma_1 = \sigma_2 = \sigma$ desconocido es equivalente a usar el intervalo de confianza para $\mu_1 - \mu_2$ con σ_1 y σ_2 conocidos e iguales estimándolo con S_p que es un estimador insesgado de σ . Observe que

$$S_p t_{n_1+n_2-2, \frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = t_{n_1+n_2-2, \frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

y si $n_1 + n_2 - 2 \geq 30$, $t_{n_1+n_2-2, \frac{\alpha}{2}} \cong z_{\frac{\alpha}{2}}$.

1.4.5 Preliminar: Distribución F o de Snedecor.

Si Y_1 y Y_2 son variables aleatorias independientes de distribución Chi-cuadrado con n_1 y n_2 grados de libertad, respectivamente, la distribución del cociente

$$Z = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}}$$

se denomina distribución F o de Snedecor.

Se deja como ejercicio, utilizar las técnicas conocidas del curso de probabilidades para demostrar que la densidad viene dada por:

$$f_F(z) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{z^{\frac{n_1}{2}-1}}{\left(1 + \frac{n_1 z}{n_2}\right)^{\frac{n_1+n_2}{2}}}, \text{ si } z \geq 0$$

Los parámetros n_1 y n_2 suelen denominarse grados de libertad de la distribución.

Los fractiles de esta distribución se encuentran en las tablas. Es decir si denotamos por $F_{n_1, n_2} = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}}$, hallarán

$$\mathbb{P}(F_{n_1, n_2} \geq f_{n_1, n_2; \alpha}) = \alpha$$

para α pequeños. Por lo tanto si queremos hallar

$$\mathbb{P}(F_{n_1, n_2} \geq f_{n_1, n_2; 1-\alpha}) = 1 - \alpha$$

con α pequeño, como $1-\alpha$ es grande no podrán utilizar la tabla directamente.

Note que $F_{n_1, n_2} \geq f_{n_1, n_2; 1-\alpha}$ equivale a $\frac{\frac{Y_2}{n_2}}{\frac{Y_1}{n_1}} \leq \frac{1}{f_{n_1, n_2; 1-\alpha}}$ por lo tanto

$$1 - \alpha = \mathbb{P}(F_{n_1, n_2} \geq f_{n_1, n_2; 1-\alpha}) = \mathbb{P}\left(F_{n_2, n_1} \leq \frac{1}{f_{n_1, n_2; 1-\alpha}}\right) =$$

$$1 - \mathbb{P}\left(F_{n_2, n_1} \geq \frac{1}{f_{n_1, n_2; 1-\alpha}}\right), \text{ por lo tanto } \mathbb{P}\left(F_{n_2, n_1} \geq \frac{1}{f_{n_1, n_2; 1-\alpha}}\right) = \alpha$$

de donde se concluye que $f_{n_1, n_2; 1-\alpha} = \frac{1}{f_{n_2, n_1; \alpha}}$

1.5 Intervalo de confianza para $\frac{\sigma_1}{\sigma_2}$ cuando μ_1 y μ_2 no se conocen:

Consideramos el estadístico

$$F_{n_1-1, n_2-2} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\frac{(n_1-1) \frac{S_1^2}{\sigma_1^2}}{(n_1-1)}}{\frac{(n_2-1) \frac{S_2^2}{\sigma_2^2}}{(n_2-1)}}$$

que tiene distribución F con parámetros $n_1 - 1$ y $n_2 - 1$ (del numerador y del denominador respectivamente), pues $(n_1 - 1) \frac{S_1^2}{\sigma_1^2}$ tiene distribución Chi-cuadrado con $n_1 - 1$ grados de libertad y $(n_2 - 1) \frac{S_2^2}{\sigma_2^2}$ tiene distribución Chi-cuadrado con $n_2 - 1$ grados de libertad donde S_1^2 y S_2^2 son independientes. Se obtiene el siguiente intervalo de confianza para el cociente de las varianzas $\frac{\sigma_1^2}{\sigma_2^2}$:

$$I = \left(\frac{S_1^2}{S_2^2} \frac{1}{f_{n_1-1, n_2-1, \frac{\alpha}{2}}}, \frac{S_1^2}{S_2^2} \cdot f_{n_2-1, n_1-1, \frac{\alpha}{2}} \right)$$

al nivel $1 - \alpha$, donde $f_{n_2-1, n_1-1, \frac{\alpha}{2}} = \frac{1}{f_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}}$ y

$$\mathbb{P} \left(f_{n_1-1, n_2-1, 1-\frac{\alpha}{2}} < \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} < f_{n_1-1, n_2-1, \frac{\alpha}{2}} \right) = 1 - \alpha$$

1.6 Intervalos de Confianza para muestras grandes:

1.6.1 Para la proporción p de la Distribución de Bernoulli:

Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de la variable aleatoria X de distribución de Bernoulli de parámetro p . Queremos construir un intervalo de confianza para p al nivel $1 - \alpha$.

En definitiva lo que se quiere estimar es la proporción de elementos de una población que tienen una determinada característica (por ejemplo, proporción de machos en una población ó proporción de éxitos).

En este caso la muestra no es normal, pero si la muestra es grande podemos valernos del teorema central del límite para tener una aproximación Normal de \bar{X} . Es decir,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

es un estadístico insesgado y consistente de p , ya que

$$\mathbb{E}(\bar{X}) = p, \text{Var}(\bar{X}) = \frac{p(1-p)}{n}$$

así el estadístico

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

es aproximadamente normal estándar.

1.6.2 Si $\sigma = p(1 - p)$ es conocido:

En este caso no hay nada que estimar, pues se conoce p .

1.6.3 Si $\sigma = p(1 - p)$ es desconocido:

La distribución de

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

es normal estándar si n es grande. Si en lugar de éste estadístico, estimamos la varianza por $\hat{p}(1 - \hat{p})$ donde $\hat{p} = \bar{X}$, se puede demostrar que

$$\frac{\bar{X} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

para n grande, es aproximadamente normal estándar. Observe que

$$\frac{\bar{X} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \frac{\sqrt{\frac{p(1-p)}{n}}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

y $\frac{\sqrt{\frac{p(1-p)}{n}}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \rightarrow 1$ probabilidad, pues $\hat{p} = \bar{X}$ es un estimador consistente de p .

De aquí se deduce que:

$$\mathbb{P} \left(-z_{\frac{\alpha}{2}} < \frac{\bar{X} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

y de aquí:

$$\mathbb{P} \left(\bar{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = 1 - \alpha$$

El intervalo obtendo al $(1 - \alpha)100\%$ de confiabilidad, viene dado por:

$$I = \left(\bar{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

donde $\hat{p} = \bar{x}$. El error que se comete al estimar $p = \bar{X}$, que es la frecuencia relativa de éxitos es de $\pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, donde $z_{\frac{\alpha}{2}}$ verifica:

$$\mathbb{P}(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

si Z tiene distribución normal estándar.

1.6.4 Comparación de proporciones:

Sea X_1, X_2, \dots, X_{n_1} es una muestra aleatoria de una variable aleatoria X con distribución $B(p_1)$ y Y_1, Y_2, \dots, Y_{n_2} es una muestra aleatoria de una variable aleatoria Y con distribución $B(p_2)$, X e Y independientes, n_1 y n_2 grandes un intervalo de confianza para $p_1 - p_2$ al nivel $1 - \alpha$ es:

$$\left(\bar{X} - \bar{Y} - z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n_1} + \frac{\bar{Y}(1-\bar{Y})}{n_2}}, \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n_1} + \frac{\bar{Y}(1-\bar{Y})}{n_2}} \right)$$

1.6.5 Para el parámetro λ de la distribución de Poisson.

Si X_1, X_2, \dots, X_{n_1} es una muestra aleatoria de una variable aleatoria X con distribución de Poisson de parámetro λ y queremos construir un intervalo de confianza al nivel $1 - \alpha$ para λ , cuando n es grande, por el teorema central del límite, se tiene que:

$$\frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

es aproximadamente normal estándar. De aquí se obtiene que:

$$\mathbb{P} \left(-z_{\frac{\alpha}{2}} < \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} < z_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

y despejando el parámetro desconocido:

$$\mathbb{P} \left(\bar{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\lambda}{n}} < \lambda < \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\lambda}{n}} \right) = 1 - \alpha$$

de esta igualdad, usando el método gráfico para resolver inecuaciones, se obtiene:

$$I = \left(\bar{X} + \frac{z_{\frac{\alpha}{2}}^2}{n} - \frac{1}{2} z_{\frac{\alpha}{2}} \sqrt{\frac{4\bar{X}}{n} + \frac{z_{\frac{\alpha}{2}}^2}{n^2}}, \bar{X} + \frac{z_{\frac{\alpha}{2}}^2}{n} + \frac{1}{2} z_{\frac{\alpha}{2}} \sqrt{\frac{4\bar{X}}{n} + \frac{z_{\frac{\alpha}{2}}^2}{n^2}} \right)$$

en la práctica, se opera como en el caso de la distribución de Bernoulli, es decir, estimamos la varianza por $\hat{\lambda} = \bar{x}$ y usando la aproximación normal para n grande podemos suponer que:

$$\frac{\bar{X} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}}$$

es normal estándar. Obtenemos así, un intervalo de confianza más simple:

$$I = \left(\bar{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\lambda}}{n}} < \lambda < \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\lambda}}{n}} \right).$$

1.6.6 Ejercicio:

Hallar un intervalo para la media β de la distribución exponencial de parámetro $\frac{1}{\beta}$, es decir, de densidad:

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x > 0 \\ 0 & \text{si no.} \end{cases}$$

1.7 Construcción de Intervalos de Confianza a partir de la Desigualdad de Tchebychev.

Si Y es una variable aleatoria se puede probar que si $\varepsilon > 0$,

$$\mathbb{P}(|Y| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \mathbb{E}(Y^2)$$

sea β un parámetro desconocido de la distribución de Y y $\hat{\beta}$ su estimador.

Si $R(\hat{\beta}) = \mathbb{E} \left[\left(\hat{\beta} - \beta \right)^2 \right]^{\frac{1}{2}}$ y $Y = \frac{\hat{\beta} - \beta}{R(\hat{\beta})}$ obtenemos que

$$\frac{1}{\varepsilon^2} \mathbb{E} (Y^2) = \frac{1}{\varepsilon^2}, \text{ pues } \mathbb{E} (Y^2) = 1$$

de aquí se obtiene

$$\mathbb{P} \left(\left| \frac{\hat{\beta} - \beta}{R(\hat{\beta})} \right| \leq \varepsilon \right) \geq 1 - \frac{1}{\varepsilon^2}$$

despejando el parámetro desconocido

$$\mathbb{P} \left(\hat{\beta} - \varepsilon R(\hat{\beta}) \leq \beta \leq \hat{\beta} + \varepsilon R(\hat{\beta}) \right) \geq 1 - \frac{1}{\varepsilon^2}.$$

Este intervalo de confianza está basado en el estimador $\hat{\beta}$ y su error cuadrático medio. La desventaja práctica de este método es que se obtienen intervalos demasiado grandes.

1.8 Otro método de construcción de intervalos de confianza:

Si se conoce la distribución del estimador $\hat{\beta}$ y esta depende del parámetro β a estimar, se pueden hallar $a(\beta)$ y $b(\beta)$ tales que

$$\mathbb{P} \left(a(\beta) \leq \hat{\beta} \leq b(\beta) \right) = 1 - \alpha$$

y despejando el parámetro se obtiene el intervalo de confianza.

1.8.1 Ejemplo: Distribución uniforme en $[0, \beta]$.

Recordemos que el estimador de máxima verosimilitud de β es

$$\max(X_1, X_2, \dots, X_n)$$

cuya esperanza es $\frac{n}{n+1}\beta$; si tomamos $\hat{\beta} = \frac{n+1}{n} \max(X_1, X_2, \dots, X_n)$, $\hat{\beta}$ es insesgado.

La densidad de $\hat{\beta}$ es:

$$f_{\hat{\beta}}(x) = \begin{cases} \frac{n^{n+1}}{(n+1)^n} \frac{x^{n-1}}{\beta^n}, & 0 \leq x \leq \frac{n+1}{n}\beta \\ 0 & \text{si no} \end{cases}$$

Supongamos que queremos hallar a y b tales que

$$\mathbb{P}\left(a \leq \hat{\beta} \leq b\right) = 1 - \alpha = 0.90$$

Hallamos a y b verificando las siguientes ecuaciones:

$$\mathbb{P}\left(\hat{\beta} \leq a\right) = 0.05, \quad \mathbb{P}\left(b \geq \hat{\beta}\right) = 0.05$$

obtenemos como solución

$$\mathbb{P}\left((0.05)^{\frac{1}{n}} \frac{n+1}{n} \beta \leq \hat{\beta} \leq (0.95)^{\frac{1}{n}} \frac{n+1}{n} \beta\right) = 0.90$$

de donde:

$$\mathbb{P}\left(\frac{\max(X_1, X_2, \dots, X_n)}{(0.95)^{\frac{1}{n}}} \leq \beta \leq \frac{\max(X_1, X_2, \dots, X_n)}{(0.05)^{\frac{1}{n}}}\right) = 0.90$$

Observe que si estimamos por $\max(X_1, X_2, \dots, X_n)$ se obtiene el mismo intervalo de confianza para β .

1.9 Datos Apareados:

Si $(X_1, Y_1), \dots, (X_n, Y_n)$ es una muestra aleatoria simple de (X, Y) , y $D = X - Y$, si D es normal $N(\mu, \sigma^2)$, entonces D_1, \dots, D_n es una muestra aleatoria simple de D , $D_i = X_i - Y_i$, con distribución normal $N(\mu, \sigma^2)$. Para obtener intervalos de confianza para μ y σ^2 se procede como en el caso de una muestra unidimensional, es decir:

1.9.1 Intervalo de Confianza para μ al nivel $1 - \alpha$:

Si σ es conocido:

$$I = \left(\bar{D} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{D} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

Si σ es desconocido:

$$I = \left(\bar{D} - t_{n-1, \frac{\alpha}{2}} \frac{S_1}{\sqrt{n}}, \bar{D} + t_{n-1, \frac{\alpha}{2}} \frac{S_1}{\sqrt{n}}\right)$$

1.9.2 Intervalo de Confianza para σ^2 al nivel $1 - \alpha$:

$$I = \left(\frac{(n-1)S_1^2}{\chi_{n-1, \frac{\alpha}{2}}^2}, \frac{(n-1)S_1^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \right)$$

donde

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2.$$

Note que si

$$\begin{aligned} \mathbb{E}(X_i) &= \mu_1, \text{Var}(X_i) = \sigma_1^2, \mathbb{E}(Y_i) = \mu_2, \text{Var}(Y_i) = \sigma_2^2 \\ \mathbb{E}(X_i - Y_i) &= \mu_1 - \mu_2 = \mu, \text{Var}(X_i - Y_i) = \sigma_1^2 + \sigma_2^2 = \sigma^2 \end{aligned}$$

si X e Y son independientes. Si no lo son

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\text{Cov}(X, Y) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2.$$

Ejemplo de una muestra con datos apareados.

Se quiere saber si un fármaco tiene el efecto colateral de llevar la presión sistólica sanguínea. Se seleccionan en forma aleatoria n personas de diferentes edades y condiciones de salud. En condiciones controladas de laboratorio se toma la presión sanguínea a cada persona, luego se les administra el fármaco y nuevamente se toma la presión. Obtenemos así observaciones apareadas: $(X_1, Y_1), \dots, (X_n, Y_n)$ de presiones sanguíneas de cada sujeto, antes y después de administrar el fármaco.

Intervalos de Confianza (Parte I)

Estadística- Práctica N° 7

1. Los datos que a continuación se dan son los pesos en gramos del contenido de 16 cajas de cereal que se seleccionaron de un proceso de llenado, con el propósito de verificar el peso promedio:

506	508	499	497
503	504	510	512
514	493	496	506
502	509	496	505

Si el peso de cada caja es una variable aleatoria normal con una desviación estándar igual a $\sigma = 5\text{grs.}$, obtenga los intervalos de confianza estimados del 90%, 95% y 99% para la media de llenada del proceso.

2. Una finca cuadrada tiene lado L . Suponga que, si medimos esta longitud L , la medición, debido a diversos errores, sigue una distribución $N(L, 1)$.

En 100 mediciones se ha obtenido una media muestral de 325mts. . Si usamos ésta como estimación de L , obtenga un intervalo de confianza de 95% para L . Calcule el mínimo tamaño que debe tener la muestra para que en la estimación de L se cometa un error máximo de 0.1mts. con una probabilidad de 0.95.

3. Se hace un envío de latas de conserva, de las que se afirma que el peso medio es de 1000grs. . Se examina una muestra de 5 latas, obteniéndose los siguientes pesos:

995	992	1005	998	1000
-----	-----	------	-----	------

en gramos. Encuentre un intervalo de confianza del 95% para el peso medio de las latas. En base al intervalo obtenido, aceptaría Ud. la afirmación de que $\mu = 1000\text{grs.}$?

4. Un fabricante de cauchos asegura que un tipo de caucho tiene una vida útil media de aproximadamente 43000 millas. Para verificar esta afirmación se prueban 10 cauchos en una rueda de prueba que simula las

condiciones normales de rodaje. Los tiempos de vida que se obtienen (en miles de millas) son:

42 36 46 43 41 35 43 45 40 39

Calcule el intervalo de confianza apropiado y decida si estos datos confirman o no la afirmación del fabricante. Hágalo al nivel $(1 - \alpha)$ para los valores de α igual a 0.2, 0.1, 0.05 y 0.01. Discuta los resultados.

5. Halle un intervalo de confianza al nivel $(1 - \alpha)$ para σ correspondiente a una distribución normal con media $\mu = 0$ y σ desconocido a partir de una muestra simple X_1, \dots, X_n de esta distribución.
6. Sea X el número de horas semanales que trabajan los ingenieros de una gran planta industrial. Suponemos X normal y queremos estimar su varianza. Para esto se selecciona una muestra de 21 ingenieros de la planta y se observa que la desviación de la muestra es de 7 horas. Calcule el intervalo de confianza del 90% para la verdadera desviación.
7. Los errores aleatorios que se cometen en las pesadas de una balanza siguen una distribución $N(\mu, \sigma^2)$. Se realizan 9 pesadas, comprobándose los siguientes errores (en *mgs.*) :

-0.07 0.3 1.8 -0.1 2.0 2.3 0.62 0.12 1.4

- (a) Encuentre un intervalo de confianza del 95% para μ y un intervalo de confianza del 95% para σ^2 .
 - (b) Aceptaría Ud. la afirmación del fabricante de balanzas, de que los errores cometidos en las pesadas siguen una distribución $N(0, 1)$?
8. Se administraron 2 nuevos medicamentos a pacientes con padecimiento cardíaco. El primer medicamento bajó la presión sanguínea de 16 pacientes en un promedio de 11 puntos con una desviación de 6. El segundo medicamento bajó la presión sanguínea de otros 20 pacientes en un promedio de 12 puntos con una desviación de 8. Determine un intervalo de confianza del 99% para la diferencia en la reducción media de la presión sanguínea, al suponer que las mediciones tienen distribuciones normales con varianzas iguales. Qué se puede concluir?

9. Una compañía de seguros trabaja con 2 talleres, A y B . La compañía sospecha que el taller A tiende a cobrar más que el taller B . Para verificar esto, se pide a ambos talleres que hagan un avalúo del costo para reparar 15 vehículos, los mismos en ambos talleres, obteniéndose (en miles de bolívares) los siguientes datos:

Vehículo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Taller A	76	102	95	13	30	63	53	62	22	48	113	121	69	76	84
Taller B	73	91	84	15	27	58	49	53	20	42	110	110	61	67	75

A los niveles de confianza 90, 95 y 99%, decida si estos datos presentan o no evidencia para aceptar que el taller A tiende a sobre facturar.

10. Un problema importante es el de fijar criterios acerca de la cantidad de sustancias químicas tóxicas que se puede tolerar en aguas de ríos y lagos. Una medida común de toxicidad de una sustancia es la concentración de ésta que es capaz de matar el 50% de una especie de peces en prueba en un tiempo determinado. Esta medida se denomina $CL50$ (concentración letal que mata el 50%).

- (a) Para insecticida DDT y una cierta especie de peces, las medidas de $CL50$ (en partes por millón) en 12 experimentos fueron:

16 5 21 19 10 5 8 2 7 4 9 2

Suponiendo normalidad, estime la media de $CL50$ para DDT , con un coeficiente de confianza del 90%.

- (b) Otro insecticida común, el Diazinon, dió las siguientes medidas para su $CL50$, en tres experimentos:

7.8 1.6 1.3

Encuentre un intervalo de confianza del 90% para la diferencia de las medias de $CL50$ para el DDT y el Diazinon (suponga igualdad de varianzas). Se puede concluir aquí que algunos de los insecticidas es menos letal?

- (c) Se realizaron 7 experimentos mas para medir el $CL50$ del Diazinon obteniendo los siguientes resultados:

4.6 1.2 10.5 6.3 1.5 7.1 1.8

Repita la parte anterior tomando los 10 datos para el Diazinon.

11. Un fabricante asegura a una compañía que le compra un producto regularmente, que el porcentaje de productos defectuosos no es mayor dle 5% . Para comprobar la afirmación del fabricante, la compañía selecciona de su inventario, 200 unidades de este producto y las prueba. Se descubren un total de 19 unidades defectuosas en la muestra, deberá sospechar la compañía de la afirmación del fabricante? Utilice 95% de confiabilidad.
12. La Cámara de Comercio de una ciudad se encuentra interesada en estimar la cantidad de dinero promedio que gasta la gente que asiste a convenciones, calculando alojamiento, comida y entretenimiento por día. De las distintas convenciones se seleccionaron 16 personas y se les preguntó la cantidad que gastaban por día. Se obtuvo la siguiente información en dólares:

150	175	163	148	142	189	135	174
168	152	158	184	134	146	155	163

si suponemos que la cantidad gastada en un día es una variable aleatoria con distribución norma, obtenga los intervalos de confianza estimados del 90, 95 y 98% para la cantidad promedio real.

MMOM/04

Intervalos de Confianza (II)
Estadística. Práctica N°8

1. El tiempo de compras de una muestra aleatoria de 64 clientes de un supermercado dió un promedio de 33 minutos con una desviación de 16 minutos. Estime el verdadero tiempo de compras por cliente con un intervalo de confianza del 90%. Suponga distribución normal.
2. Si se utiliza \bar{X} como estimador la media μ , demuestre que podemos tener el $(1 - \alpha)100\%$ de confianza en que el valor absoluto del error $|\bar{X} - \mu|$, no excederá una cantidad prefijada e , cuando el valor de la muestra sea

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{e} \right)^2.$$

Suponga distribución normal.

3. Supongamos que tenemos una muestra grande de una distribución de Bernoulli de parámetro p . Demuestre que podemos tener por lo menos el $(1 - \alpha)100\%$ de confianza en que el valor absoluto del error $|\hat{p} - p|$, cuando utilizamos la proporción de la muestra para estimar p , no excederá una cantidad e especificada cuando el tamaño de la muestra sea

$$n = \frac{z_{\frac{\alpha}{2}}^2}{4e^2}.$$

4. En una cierta población se desea conocer la proporción de individuos alérgicos al polen de Acacias. En 100 individuos escogidos al azar se observaron 10 alérgicos.
 1. Hallar el intervalo de confianza del 95% de la proporción pedida.
 2. Cuántos individuos se deberían observar para que, con probabilidad 0.95, el error máximo en la estimación de alérgicos sea del 1%.
 3. Conteste la pregunta anterior si Ud. no dispone de ninguna muestra ya examinada de la población.
5. Un remedio tópico para la calvicie, el Minoxidil, se administró a un grupo de 310 hombres calvos, de los cuales el 32% observó crecimiento de cabello

nuevo. Simultáneamente se administró un placebo a un grupo de 309 hombres calvos, de los cuales 20% observó crecimiento de cabello nuevo. A partir de estos datos, al nivel de confianza del 95%, qué puede decir de la efectividad del Minoxidil?

6. Se supone que el número de erratas por página de un cierto libro sigue una distribución de Poisson. Elegidas al azar 95 páginas se obtuvo:

Número de erratas	0	1	2	3	4	5
Número de páginas	40	30	15	7	2	1

Hallar el intervalo de confianza del 90% para el número medio de erratas por página en el libro.

1. Usando el Teorema Central del Límite.
 2. Usando la desigualdad de Chebychev.
7. Sea X una variable aleatoria con distribución Gamma de parámetro $p = 2$ y λ desconocido, es decir la densidad de X es:

$$f(x) = \frac{\lambda^2}{\Gamma(2)} x e^{-\lambda x}, \quad x > 0.$$

1. Si $Y = 2\lambda X$, pruebe que Y tiene distribución chi-cuadrado con 4 grados de libertad.
 2. Utilice la parte anterior para hallar un intervalo de confianza del 90% para λ .
8. Se han medido los siguientes valores (en miles de personas) para la audiencia de un programa de TV en distintos días: 521,742,593,635,788,717,606,639,666,624. Construir un intervalo de confianza para la audiencia media y otro para la varianza, bajo hipótesis de normalidad, al 99% de confianza.
9. El número diario de piezas fabricadas por una máquina A en cinco días ha sido: 50,48,53,60,37; mientras que en esos mismos días una máquina B ha hecho: 40,51,62,55 y 64. Se pide:
1. Construir un intervalo al 95% de confianza para la diferencia de las medias entre ambas máquinas, suponiendo la misma varianza y distribución normal.

2. Construir un intervalo para el cociente de las varianzas al 95% de confianza. ¿Es posible suponer que las varianzas son iguales?
 3. Utilizando el mismo estimador de la varianza común hallado en la primera parte de este ejercicio, halle el tamaño muestral n de ambas muestras para que al nivel de 95%, la longitud del intervalo de confianza para la diferencia de las medias sea 8 unidades.
10. Suponga que se mide un objeto independientemente con dos procedimientos de medidas diferentes. Sean L_1 y L_2 las longitudes medidas obtenidas con cada método. Si cada método está correctamente calibrado podemos suponer que $\mathbb{E}(L_1) = \mathbb{E}(L_2) = L$, la longitud verdadera. Los métodos no tienen necesariamente la misma exactitud, si medimos la exactitud mediante la varianza, entonces $\text{Var}(L_1) \neq \text{Var}(L_2)$. Si $Z = aL_1 + (1 - a)L_2$ como nuestro estimador de L , es inmediato verificar que es insesgado. ¿Para qué valor de $a \in (0, 1)$ es mínima la varianza de Z ?
11. Una muestra de 400 candidatos políticos, 200 escogidos al azar en el este y 200 en el oeste, se clasificó teniendo en cuenta si el candidato tuvo respaldo de un sindicato nacional y si el candidato ganó la elección. Un resumen de los datos es el siguiente:

	Oeste	Este
Ganadores respaldados por el sindicato	120	142

Encuentre un intervalo de confianza del 95% para la diferencia entre las proporciones de ganadores respaldados por el sindicato, en el oeste y en el este.

MMOM/04

Capítulo IV

Prueba de Hipótesis

María Margarita Olivares

Mayo 2004

1 Introducción:

En los capítulos anteriores hemos resuelto el problema de obtener medidas aproximadas de parámetros de los que depende la distribución de un conjunto de observaciones. Ahora queremos tomar decisiones acerca de esos mismos parámetros. El tipo de decisión que vamos a considerar es el siguiente

“sabemos que el parámetro β pertenece a un cierto espacio de parámetros B . Dado un subconjunto B_0 de B queremos decidir si resulta razonable admitir que $\beta \in B_0$ ó si debemos rechazar esta suposición”,

es decir, consideraremos el estudio de la prueba o contraste de una hipótesis.

Una hipótesis estadística es una afirmación con respecto a alguna característica desconocida de una población de interés. La esencia de probar una hipótesis estadística es el decidir si la afirmación se encuentra apoyada por la evidencia experimental que se obtiene a través de una muestra aleatoria.

Ilustremos con un ejemplo la noción de una hipótesis estadística: “se tiene interés en probar la eficiencia de una nueva vacuna contra la gripe o resfriado común”, para ello se inyectan 10 personas con el suero y se observan por el periodo de un año. Ocho personas pasaron el invierno sin resfriado, ¿resulta razonable admitir la eficiencia del nuevo suero basándose en el resultado experimental obtenido?.

El razonamiento que se utiliza en una prueba de hipótesis estadística es parecido al procedimiento que se usa en una corte judicial: “al juzgar

un hombre por robo, la corte siempre asume que el acusado es inocente, mientras no se pruebe lo contrario. El fiscal trata por todos los medios de presentar evidencias que contradigan la hipótesis de inocencia del acusado y de conseguir su condena”. En el problemas estadístico del ejemplo, la vacuna es el acusado, la hipótesis de prueba, llamada hipótesis nula, es que la vacuna no es efectiva. El experimentador hace papel de fiscal, está convencido de que su vacuna es efectiva y trata de usar la evidencia contenida en la muestra para rechazar la hipótesis nula.

Intuitivamente procederíamos seleccionando el número Y de personas que no contraen el resfriado, como una medida de la cantidad de evidencia en la muestra. Si Y resulta muy pequeño se tendría poca evidencia para rechazar o apoyar la hipótesis nula, si Y resulta grande, rechazaríamos la hipótesis nula, concluyendo así que la vacuna es efectiva. De hecho, si la hipótesis nula fuese cierta (y la vacuna fuese ineficaz) la probabilidad de pasar el invierno sin resfriado sería $p = \frac{1}{2}$, el promedio $\mathbb{E}(Y) = np = 5$. Existen valores de Y tales como $Y = 10$ ó $Y = 0, 1, 2, 3, 4, 5$ que permitiría a cualquier persona tomar una decisión intuitiva respecto al rechazo o no de la hipótesis nula (vacuna ineficaz). Pero si $Y = 7, 8, 9$, ¿qué se podría decir?. Ya sea que usemos un procedimiento objetivo o subjetivo, seleccionaríamos aquel que tenga menor probabilidad de tomar una decisión incorrecta. Un instrumento para tomar decisiones, es llamado, estadístico de prueba. En nuestro ejemplo, el número de personas no afectadas por el resfriado es suficiente como estadístico de prueba:

$$Y = n^{\circ} \text{ de personas no afectadas por el resfriado}$$

el Rango de Y es

$$R(Y) = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Estos valores se dividen en dos grupos: uno llamado región de rechazo y el otro región de aceptación. Después de realizar el experimento, si Y toma un valor en la región de rechazo, la hipótesis nula es rechazada, de lo contrario, el experimentador no rechaza la hipótesis ya que la muestra no presenta suficiente evidencia como para rechazar la hipótesis. (Posteriormente veremos que se rechaza o se acepta la hipótesis nula solo si los riesgos de una decisión errónea son conocidos y pequeños).

Supongamos que en nuestro ejemplo hemos escogido

$$\{Y = 8 \text{ ó } 9 \text{ ó } 10\}$$

como región de rechazo y los otros valores como región de aceptación, puesto que observamos $Y = 8$ personas sin resfriado, rechazamos la hipótesis nula de que la vacuna es ineficaz y concluimos que la probabilidad de pasar el invierno sin ningún resfriado es mayor que $p = \frac{1}{2}$ cuando se usa la vacuna.

Podríamos preguntarnos: ¿cuál es la probabilidad de rechazar la hipótesis nula siendo cierta?. Esto no es más que la probabilidad del evento

$$\{Y = 8 \text{ ó } 9 \text{ ó } 10\}$$

si $p = \frac{1}{2}$ y esta probabilidad es

$$\mathbb{P}(Y = 8 \text{ ó } Y = 9 \text{ ó } Y = 10) = \sum_{k=8}^{10} \binom{10}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{10-k} = 0,055$$

Puesto que hemos decidido rechazar la hipótesis nula y la probabilidad de error es pequeña, podemos estar razonablemente seguros de haber tomado una decisión correcta.

El fabricante de la vacuna se enfrenta a dos tipos de errores posibles:

1. Error de tipo I: rechazar la hipótesis nula y concluir equivocadamente que la vacuna es eficaz (esto podría conducir a desarrollar programas costosos de producción y grandes pérdidas financieras).
2. Error de tipo II: no rechazar la hipótesis nula y concluir equivocadamente que la vacuna es ineficaz (aquí el productor perdería las utilidades potenciales derivadas de la venta de la vacuna).

La probabilidad de estos dos tipos de errores, α y β respectivamente, se relacionan inversamente. Un aumento del tamaño de la muestra proporciona más información para tomar la decisión y por tanto reduce tanto α como β . El experimentador selecciona los valores α y β y luego, la región de rechazo; el tamaño de la muestra se escoge de acuerdo a estas probabilidades.

En nuestro ejemplo, llamando H_0 la hipótesis nula y H_1 la hipótesis alternativa

$$H_0 : p = 0.5$$

$$H_1 : p > 0.5$$

$$\alpha = \mathbb{P}(\text{Rechazar } H_0 \text{ siendo cierta}) = \mathbb{P}(Y = 8, 9, 10 / p = 0.5) = 0.055$$

$$\beta_p = \mathbb{P}(\text{Aceptar } H_0 \text{ siendo falsa, para } p > 0.5)$$

$$\text{en particular, } \beta_p = \mathbb{P}(\text{Aceptar } H_0 \text{ siendo falsa, para } p = 0.9) =$$

$$\mathbb{P}(0, 1, 2, 3, 4, 5, 6, 7 / p = 0.9) = 0.07$$

Estos valores

$$\alpha = 0.055, \beta_{0.9} = 0.07$$

dan una medida de los riesgos de cometer alguno de los errores posibles para esta prueba.

2 Conceptos Básicos:

Una prueba estadística implica cuatro elementos:

1. La hipótesis nula
2. La hipótesis alternativa
3. El estadístico de prueba
4. La región de rechazo

2.1 La Hipótesis Nula:

Se indica simbólicamente por H_0 y establece la hipótesis sometida a prueba. Así H_0 especifica valores para uno ó más parámetros de la población. En el ejemplo anterior, la hipótesis nula es $p = 0.5$, y puesto que en este caso H_0 especifica un solo valor para el parámetro p , recibe el nombre de hipótesis simple, de otra forma se conoce como hipótesis compuesta, por ejemplo si $H_0 : p \leq 0.5$.

2.2 La Hipótesis Alternativa:

Se indica simbólicamente como H_1 la cual debe reflejar el valor posible ó intervalo de valores del parámetro de interés, si la hipótesis nula es falsa, así, la hipótesis alternativa representa una forma de negación de la hipótesis nula. H_1 puede ser simple o compuesta. En nuestro ejemplo, H_1 es $p > 0.5$ (es una hipótesis compuesta). A pesar de que no se pretende generalizar, en muchas ocasiones es deseable establecer una hipótesis nula que sea más específica que la alternativa. De esta manera la hipótesis nula es generalmente simple y mientras que la alternativa es compuesta.

2.3 El Estadístico de Prueba:

La decisión de rechazar o de aceptar la hipótesis nula se basa en la información contenida en una muestra extraída de la población de interés. Los valores de la muestra se usan para calcular un sólo número, este número actúa como ente que toma decisiones y lo llamamos estadístico de prueba.

2.4 Región de Rechazo:

El conjunto de valores o rango del estadístico de prueba se divide en dos conjuntos ó regiones. Si el estadístico de prueba calculado a partir de una muestra específica asume un valor dentro de la región de rechazo, entonces la hipótesis nula es rechazada y decidimos a favor de la alternativa. Si este valor cae fuera de la región de rechazo (cae en la región de aceptación), no se rechaza H_0 lo cual no debe interpretarse como equivalente a que H_0 deba ser aceptada sino que simplemente, la información experimental que se tiene no conduce a rechazarla.

De todo lo anterior se deduce que dadas las hipótesis a probar: H_0 y H_1 y dado el conjunto de variables sobre la que se basa la prueba, diseñar esa prueba consiste en elegir la región de rechazo.

2.5 Tipos de Errores:

Al llevar a cabo la prueba de una hipótesis H_0 contra H_1 , el resultado puede ser rechazar H_0 o bien no rechazar H_0 . Por otra parte puede ser que en la realidad la hipótesis H_0 se cumpla o que no se cumpla. Denotamos por R la región de rechazo y por $T(X_1, X_2, \dots, X_n)$ el estadístico de prueba.

1. Error de primera especie o tipo I: Cuando H_0 se cumple y nuestra decisión es rechazarla, cometemos un error llamado de primera especie o tipo I. La probabilidad de cometer este error la denotaremos por:

$$\alpha = \mathbb{P}(T(X_1, X_2, \dots, X_n) \in R \mid H_0) = \mathbb{P}_{H_0}(T(X_1, X_2, \dots, X_n) \in R)$$

2. Error de segunda especie o tipo II: Cuando H_0 no se cumple y no la rechazamos, cometemos un error llamado de segunda especie o de tipo II. La probabilidad de cometer este error la denotamos por

$$\beta = \mathbb{P}(T(X_1, X_2, \dots, X_n) \in R^c \mid H_1) = \mathbb{P}_{H_1}(T(X_1, X_2, \dots, X_n) \in R^c)$$

donde R^c es el complemento de la región de rechazo, es decir, es la región de aceptación.

2.6 Potencia de la Prueba:

$$\Pi = 1 - \beta = \mathbb{P}_{H_1}(T(X_1, X_2, \dots, X_n) \in R)$$

es la potencia de la prueba, representa la probabilidad de no cometer un error de segunda especie.

2.6.1 Observaciones:

1. Las notaciones

$$\mathbb{P}(T(X_1, X_2, \dots, X_n) \in R \mid H_0) \text{ y } \mathbb{P}(T(X_1, X_2, \dots, X_n) \in R \mid H_1)$$

no corresponden a probabilidades condicionales sino a probabilidades correspondientes a un valor del parámetro compatible con cada una de las hipótesis, para evitar confusión es preferible utilizar las notaciones $\mathbb{P}_{H_1}(T(X_1, X_2, \dots, X_n) \in R^c)$ y $\mathbb{P}_{H_0}(T(X_1, X_2, \dots, X_n) \in R)$ sin embargo utilizaremos las dos notaciones indistintamente..

2. Si H_0 es simple, por ejemplo $H_0 : p = p_0$

$$\alpha = \mathbb{P}(T(X_1, X_2, \dots, X_n) \in R \mid p = p_0)$$

es el nivel de la prueba. Si H_0 es compuesta, por ejemplo, $p \in [a, b]$, se suele llamar el nivel de la prueba a :

$$\sup_{p \in [a, b]} \mathbb{P}(T(X_1, X_2, \dots, X_n) \in R \mid p) = \alpha = \sup_{p \in [a, b]} \alpha(p)$$

3. Si H_1 es simple, por ejemplo $H_1 : p = p_1$

$$\Pi = \mathbb{P}(T(X_1, X_2, \dots, X_n) \in R \mid p = p_1)$$

es la potencia de la prueba. Si H_1 es compuesta, por ejemplo, $H_1 : p \in [c, d]$, se tendrá una función potencia de la prueba

$$\Pi(p) = \mathbb{P}(T(X_1, X_2, \dots, X_n) \in R \mid p)$$

para cada p compatible con H_1 (en este caso, $p \in [c, d]$).

La siguiente tabla muestra las cuatro situaciones que pueden darse según se cumpla o no H_0 y según decidamos o no rechazarla:

		Rechazar H_0 $T(X_1, X_2, \dots, X_n) \in R$	No rechazar H_0 $T(X_1, X_2, \dots, X_n) \notin R$
Realidad	Se cumple H_0	Error I(prob. α)	Probabilidad $1 - \alpha$
	No se cumple H_0	Probabilidad $1 - \beta = \Pi$	Error II(prob. β)

Regresemos a la analogía del proceso judicial para proporcionar una idea más clara sobre el problema. Si la hipótesis nula es “el acusado es inocente”, entonces, con toda seguridad, la alternativa es “el acusado es culpable”. El rechazo de la hipótesis nula implicaría que el juicio ha sido capaz de proporcionar suficiente evidencia para garantizar el veredicto de “culpable”.

Por otro lado, si el juicio no presenta evidencia sustancial, el veredicto será “inocente”. Esta decisión no implica necesariamente que el acusado sea inocente, más bien, hace énfasis en la falta de evidencia necesaria para condenar al acusado. Por lo tanto, un veredicto de culpable es más fuerte que un veredicto de inocente, lo cual surge del principio judicial generalmente aceptado: “es peor condenar a un inocente que dejar ir a un culpable”. Si el veredicto es culpable, se desea tener un grado muy alto de seguridad de que no se va a condenar a una persona inocente, es decir, que el error tipo I proveniente de rechazar H_0 siendo cierta tenga muy baja probabilidad, por lo tanto, en muchas situaciones el error tipo I se considera como un error más grave que el error tipo II.

En las pruebas de hipótesis estadísticas el enfoque general es aceptar la premisa de que el error de tipo I es mucho más serio que el tipo II y formular la hipótesis nula y alternativa de acuerdo a esto.

Como resultado se tiene que muchas veces se selecciona con anticipación el tamaño máximo del error de tipo I que puede tolerarse y se intenta definir un procedimiento de prueba que minimice el error de tipo II. En otras palabras, no es posible seleccionar tanto α como β y diseñar una prueba estadística para probar H_0 contra H_1 , dada una muestra aleatoria de tamaño n .

Un principio sencillo y razonable al obtener reglas de decisión para una prueba de hipótesis estadística es seleccionar aquel procedimiento de prueba que tenga el tamaño más pequeño para el error del tipo II (máxima potencia), entre todos los procedimientos que tengan el mismo tamaño para el error del tipo I. En este contexto debe notarse que el valor de α no puede hacerse arbitrariamente pequeño, sin que se incremente el valor de β . En otras palabras, para una muestra de tamaño n dado, el tamaño del error de

tipo II normalmente aumentará conforme el error de tipo I disminuya. Lo que en general se hace en la práctica, es ajustar el tamaño del error de tipo I cambiando la región de rechazo del estadístico de prueba para así alcanzar un balance satisfactorio entre los tamaños de los errores, siempre y cuando se tome en cuenta el máximo tamaño del error de tipo I que puede tolerarse en una situación particular.

2.7 Valor Crítico del Estadístico de Prueba:

Si $T(X_1, \dots, X_n)$ es el estadístico de prueba, el valor c de $T(X_1, \dots, X_n)$ que separa la región de rechazo y aceptación (ó valores c_i $i = 1, 2, \dots, k$; según la región tenga uno o varios puntos frontera) se llama valor crítico (respectivamente críticos) del estadístico de prueba.

2.7.1 Tipos de Regiones Críticas o de Rechazo:

Supongamos que H_0 es simple: $\mu = \mu_0$

1. Si H_1 es $\mu > \mu_0$ ó H_1 es $\mu < \mu_0$ alternativamente la prueba es unilateral.
2. Si H_1 es $\mu > \mu_0$ ó $\mu < \mu_0$ simultáneamente la prueba es bilateral.

La prueba estadística unilateral debe formarse sólo si el valor del parámetro en uno de los lados no tiene sentido para el investigador. (Así, en el ejemplo de la vacuna antiresfriado $H_1: p > 0.5$ es una prueba unilateral, pues la alternativa que interesa es que aumente la probabilidad de pasar el invierno sin resfriado). Si el investigador no está seguro, debe realizarse una prueba bilateral.

Una prueba bilateral da lugar a regiones críticas bilaterales, que en forma general serán simétricas, las dos partes de la región se seleccionan de tal manera que la probabilidad de caer en cada una de ellas sea la misma.

2.8 Procedimiento para algunas pruebas de Hipótesis:

Supongamos que X_1, \dots, X_n es una muestra aleatoria de X de distribución $N(\mu, \sigma^2)$:

1. Para μ

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \text{ ó } \mu < \mu_0 \text{ (prueba bilateral)}$$

(a) Si σ es conocido, utilizamos el estadístico de prueba

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim N(0, 1) \text{ bajo } H_0$$

Si α es el error de primera especie, la región de rechazo se obtiene hallando $z_{\frac{\alpha}{2}}$ tal que

$$\mathbb{P}(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

La región de rechazo viene dada por:

$$(-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$$

Si la prueba es unilateral, por ejemplo, $H_1 : \mu > \mu_0$, la región de rechazo se halla encontrando z tal que

$$\mathbb{P}(Z > z) = \alpha$$

la región de rechazo viene dada por

$$(z_{\alpha}, \infty).$$

Si la prueba es unilateral izquierda, por ejemplo, $H_1 : \mu < \mu_0$, la región de rechazo se halla encontrando z tal que

$$\mathbb{P}(Z < z) = \alpha$$

o equivalentemente

$$\mathbb{P}(Z > z) = \alpha$$

puesto que la distribución normal estándar es simétrica respecto al origen la región de rechazo viene dada por

$$(-\infty, -z_{\alpha})$$

2.8.1 P-valor

Observe que en el caso de la prueba bilateral, la región de rechazo es de la forma

$$\left\{ (x_1, \dots, x_n) : \left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \right| > c \right\}$$

hemos ajustado c de manera que

$$\mathbb{P}_{H_0} \left(\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \right| > c \right) = \alpha$$

El procedimiento de decisión es el siguiente

$$\begin{aligned} \text{Si } \left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \right| &> c, \text{ asumimos } H_1 \\ \text{Si } \left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \right| &\leq c, \text{ asumimos } H_0 \end{aligned}$$

la función $u \rightarrow \mathbb{P}_{H_0} \left(\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \right| > u \right)$ es estrictamente decreciente y en $u = c$ tiene el valor α y se cumple

$$\mathbb{P}_{H_0} \left(\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \right| > u \right) \leq \alpha \Leftrightarrow u \geq c$$

si denotamos por T_{obs} una observación del estadístico de prueba $\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$ y esta satisface

$$\mathbb{P}_{H_0} \left(\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \right| > |T_{obs}| \right) \leq \alpha$$

rechazamos H_0 .

La expresión $\mathbb{P} \left(\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \right| > |T_{obs}| \right)$ es llamada p -valor asociado a la observación T_{obs} o nivel de significación de lo observado.

En el caso de prueba unilateral derecha el p -valor es

$$\mathbb{P}_{H_0} \left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > T_{obs} \right)$$

y en el caso de la prueba unilateral izquierda el p -valor corresponde a

$$\mathbb{P}_{H_0} \left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < T_{obs} \right)$$

En todos los casos se toma la siguiente decisión:

$$\begin{aligned} \text{Si el } p\text{-valor} &\leq \alpha \text{ asumimos } H_1 \\ \text{Si el } p\text{-valor} &> \alpha \text{ asumimos } H_0 \end{aligned}$$

Esta técnica expuesta para hallar el p-valor asociado a una prueba estadística se generaliza fácilmente a las pruebas que se desarrollarán más adelante sustituyendo el estadístico de prueba por el correspondiente en cada caso. Se debe señalar que hay que ser cuidadoso con aquellas distribuciones que no sean simétricas respecto al origen (chi-cuadrado, distribución F)

1. (a) **EJEMPLO:** Se desea probar al 95% de certeza la hipótesis nula siguiente:

una población estadística normal tiene media cero, basándose en una muestra de tamaño $n = 9$ y suponiendo que $\sigma = 1$ (es conocida), contra la alternativa que la media es positiva.

Hallamos la región crítica al nivel $\alpha = 1 - 0.95 = 0.05$

$$H_0 : \mu = 0 \quad H_1 : \mu > 0$$

que es una prueba unilateral. El estadístico de prueba es:

$$Z = \frac{\sqrt{n}\bar{X}}{\sigma} = \sqrt{9}\bar{X} = 3\bar{X}$$

que tiene distribución $N(0, 1)$ bajo la hipótesis nula. Queremos hallar la región de rechazo a partir de:

$$\mathbb{P}(Z > z_\alpha) = \alpha = 0.05$$

de aquí se deduce que

$$z_\alpha = 1,6448$$

Si $Z > 1,6448$ rechazamos la hipótesis nula, esto equivale a

$$\bar{X} > \frac{1,6448}{3} = 0,5483$$

La región de rechazo se puede expresar como

$$\bar{X} > 0,5483$$

El error de tipo I es $\alpha = 0.05$

Si $\mu = \mu_1 > 0$, el error de tipo II se calcula a partir de

$$\begin{aligned} \beta(\mu_1) &= \mathbb{P}(\bar{X} \leq 0,5483 \mid \mu = \mu_1) = \\ &\mathbb{P}\left(\frac{\bar{X} - \mu_1}{1/\sqrt{9}} \leq 3 \cdot (0,5483 - \mu_1) \mid \mu = \mu_1\right) \end{aligned}$$

si en particular $\mu_1 = 1$

$$\begin{aligned}\beta &= \mathbb{P}(\bar{X} \leq 0,5483 \mid \mu = 1) = \\ &\mathbb{P}\left(\frac{\bar{X} - 1}{1/\sqrt{9}} \leq -1,3551 \mid \mu = \mu_1\right)\end{aligned}$$

donde $\frac{\bar{X}-1}{1/\sqrt{9}}$ tiene distribución $N(0, 1)$ si $\mu_1 = 1$, de aquí obtenemos $\beta = 0,0877$.

(b) Si σ es desconocido, utilizamos el estadístico de prueba

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S_1} \sim t\text{-Student con } n-1 \text{ grados de libertad, bajo } H_0$$

La región de rechazo bilateral vendrá dada por:

$$(-\infty, -t_{n-1, \frac{\alpha}{2}}) \cup (t_{n-1, \frac{\alpha}{2}}, \infty)$$

siendo α el nivel de significación de la prueba.

EJEMPLO: El fabricante de cierto tipo de baterías anuncian que estas tienen una vida promedio de 2000 horas. Se desea probar esta hipótesis al nivel 95% de confianza basándose en el hecho de que 6 baterías tomadas al azar tuvieron vidas de:

2005, 1980, 1920, 2010, 1975, 1950 horas.

Supondremos que la duración es normal. En este caso tendremos:

$$H_0 : \mu = 2000 \quad H_1 : \mu < 2000$$

(es una prueba unilateral). $\alpha = 1 - 0.95 = 0.05$, hallemos $t_{5, \alpha}$ tal que

$$\mathbb{P}(T > t_{5, \alpha}) = \alpha = 0.05$$

La región de rechazo unilateral es:

$$(-\infty, -t_{5, \alpha}) = (-\infty, -2.015)$$

Usando la muestra obtenemos:

$$\bar{X} = 1973,3333; \quad S_1 = 34,009804; \quad T = -1,9206145 \notin R$$

No hay evidencia como para rechazar la hipótesis al 95% de nivel de significación.

El p - *valor* es

$$\mathbb{P}(T_5 < -1,9206145) = 0,05642$$

que es el nivel de significación de lo observado. Si hubiésemos tomado $\alpha = 0,06$, se rechazaría la hipótesis nula. Podríamos rechazar al 94%.

2. Para σ^2 :

$$H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \sigma^2 > \sigma_0^2 \text{ ó } \sigma^2 < \sigma_0^2$$

El estadístico de prueba en este caso es:

$$\chi^2 = \frac{(n-1)S_1^2}{\sigma_0^2}$$

que tiene distribución Chi-cuadrado con $n-1$ grados de libertad, bajo la hipótesis nula. Al nivel de significación α se obtiene como región de rechazo bilateral

$$\left(0, x_{n-1, 1-\frac{\alpha}{2}}^2\right) \cup \left(x_{n-1, \frac{\alpha}{2}}^2, \infty\right).$$

EJEMPLO: Dadas las observaciones:

$$1.4, 1.9, 1.3, 1.4, 1.5$$

Probar

$$H_0 : \sigma^2 = \sigma_0^2 = 0.05 \quad H_1 : \sigma^2 > \sigma_0^2 = 0.05$$

con nivel $\alpha = 0.05$. De la muestra se obtiene:

$$S_1^2 = 0.055, \bar{X} = 1.5, n-1 = 4, \chi^2 = 4.4$$

Puesto que la prueba es unilateral, tomamos como región de rechazo la que proviene de:

$$\mathbb{P}(\chi^2 > x_{4,\alpha}^2) = \alpha = 0.05$$

Usando la tabla, obtenemos $x_{4,\alpha}^2 = 9,488$, puesto que el valor observado $\chi^2 = 4.4$ no cae en la región de rechazo al 5% de nivel, no se puede rechazar la hipótesis nula. El p - *valor* es

$$\mathbb{P}(\chi^2 > 4, 4) = 0,35457$$

Si $\sigma^2 = 0.08$, la potencia

$$\Pi = \mathbb{P}(\mathcal{X}^2 > x_{4,\alpha}^2 \mid \sigma^2 = 0.08)$$

se puede calcular de la siguiente manera: la región de rechazo es

$$R = \left[\mathcal{X}^2 = \frac{(n-1)S_1^2}{\sigma_0^2} > 9,488 = x_{4,\alpha}^2 \right]$$

puesto que si suponemos que la varianza es $\sigma^2 = 0.08$ ya no será cierto que $\frac{(n-1)S_1^2}{\sigma_0^2}$ es chi-cuadrado con 4 grados de libertad, por lo tanto debemos cambiar σ_0^2 por σ^2 , es decir:

$$\left[\frac{(n-1)S_1^2}{\sigma_0^2} > 9,488 \right] = \left[\frac{(n-1)S_1^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2} 9,488 \right] = \left[\frac{(n-1)S_1^2}{\sigma^2} > 5.93 \right]$$

Bajo la hipótesis alternativa $\sigma^2 = 0.08$, $\frac{(n-1)S_1^2}{\sigma^2}$ tiene distribución Chi-cuadrado con 4 grados de libertad, usando la tabla se obtiene que:

$$\mathbb{P} \left(\frac{(n-1)S_1^2}{\sigma^2} > 5.93 \mid \sigma^2 = 0.08 \right) = 0.20$$

2.8.2 POBLACIONES NORMALES E INDEPENDIENTES:

Sea $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, X_1, \dots, X_{n_1} una muestra aleatoria simple de X y Y_1, \dots, Y_{n_2} una muestra aleatoria simple de Y . Las muestras X e Y son independientes.

Comparación de Medias:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 > \mu_2 \text{ ó } \mu_1 < \mu_2$$

1. Si σ_1 y σ_2 son conocidos, el estadístico de prueba es:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

bajo la hipótesis nula, este estadístico tiene distribución $N(0, 1)$, al nivel α se toma como región de aceptación aquella que satisface

$$\mathbb{P} \left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

obteniéndose como región de rechazo

$$R = (-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$$

2. $\sigma_1 = \sigma_2 = \sigma$ son desconocidos, el estadístico de prueba es:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t - Student \text{ con } n_1 + n_2 - 2$$

grados de libertad, bajo la hipótesis nula. Al nivel α , la región de rechazo bilateral es:

$$R = (-\infty, -t_{n-1, \frac{\alpha}{2}}) \cup (t_{n-1, \frac{\alpha}{2}}, \infty).$$

Comparación de Varianzas:

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 > \sigma_2^2 \text{ ó } \sigma_1^2 < \sigma_2^2$$

El estadístico de prueba es:

$$F = \frac{\frac{(n_1-1)S_1^2/\sigma_1^2}{(n_1-1)}}{\frac{(n_2-1)S_2^2/\sigma_2^2}{(n_2-1)}} = \frac{S_1^2}{S_2^2}$$

bajo la hipótesis nula, F tiene distribución F con parámetros n_1-1 y n_2-1 . Al nivel α la región de aceptación se obtiene a partir de:

$$\mathbb{P}(f_{n_1-1, n_2-1, 1-\frac{\alpha}{2}} \leq F \leq f_{n_1-1, n_2-1, \frac{\alpha}{2}}) = 1 - \alpha$$

donde $f_{n_1-1, n_2-1, 1-\frac{\alpha}{2}} = \frac{1}{f_{n_2-1, n_1-1, \frac{\alpha}{2}}}$. La región de rechazo bilateral viene dada por:

$$R = (0, f_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}) \cup (f_{n_1-1, n_2-1, \frac{\alpha}{2}}, \infty)$$

2.8.3 Datos Apareados:

Comparación de medias, muestras dependientes apareadas:

Supongamos que queremos comparar dos marcas de neumáticos. Una posibilidad es poner durante k kilómetros la marca A en n_1 neumáticos y la marca B en n_2 neumáticos, medir los desgastes medios (\bar{X}, \bar{Y}) y aplicar la comparación de medias para dos muestras independientes y normales. Si la variabilidad de la población es muy grande, el valor observado del estadístico de prueba podría ser muy pequeño y no detectaríamos la diferencia entre las medias a menos que esta diferencia sea enorme.

Lo que ocurre es que la diferencia de los desgastes de los cauchos depende de muchos factores que no controlamos: tipo de conducción, conductor, superficie, etc. Al no controlar estos factores, la variabilidad muestral será tan grande que nos impedirá observar posibles diferencias entre las medias. Una solución es disponer en cada vehículo dos neumáticos A y dos B y medir diferencias de desgastes en el mismo vehículo, al ser la variabilidad de estas diferencias mucho menor, tendremos en general un mejor contraste.

La clave de este procedimiento es disponer de medidas por pares tomadas en condiciones muy semejantes, de manera que a priori las dos unidades experimentales (ruedas en el ejemplo) que comparamos sean lo mas iguales posibles. Así la variabilidad de las diferencias será pequeña y podemos identificar más fácilmente cambios.

Supongamos que elegimos $2n$ unidades homogéneas por pares (ruedas del mismo coche, personas de iguales características, objetos de iguales propiedades, etc.). Si $(X_1, Y_1), \dots, (X_n, Y_n)$ es una muestra aleatoria simple de (X, Y) , y $D = X - Y$, si D es normal $N(\mu, \sigma^2)$, entonces D_1, \dots, D_n es una muestra aleatoria simple de D , $D_i = X_i - Y_i$, con distribución normal $N(\mu, \sigma^2)$. Si $\mathbb{E}(X_i) = \mu_1$, $\mathbb{E}(Y_i) = \mu_2$; $\mathbb{E}(D_i) = \mu_1 - \mu_2$ y si no hay diferencia entre las medias la esperanza de D_i será cero. Supongamos igualdad de las varianzas de X_i y Y_i , denotemos por σ_0^2 esta varianza común, entonces $\text{Var}(D_i) = \sigma_0^2 + \sigma_0^2 - 2\rho\sigma_0\sigma_0$, ρ es el coeficiente de correlación de las dos variables (X, Y) , de aquí,

$$\sigma^2 = \text{Var}(D_i) = 2\sigma_0^2(1 - \rho)$$

si las dos medidas que comparamos, (los desgastes de los neumáticos en un mismo coche, en nuestro ejemplo), son análogas, ρ será positivo y cercano a 1, la variabilidad será menor que si tomamos muestras independientes.

Para la prueba de hipótesis (bilateral)

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

usamos el estadístico

$$Z = \frac{\bar{D}}{\frac{\sigma}{\sqrt{n}}}$$

que bajo la hipótesis nula tiene distribución normal estándar, si se conoce la

varianza. Si no conocemos la varianza, nos valemos del estadístico

$$T = \frac{\bar{D}}{\frac{S_1}{\sqrt{n}}}$$

que tiene distribución t-de Student, con $n - 1$ grados de libertad y donde

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

las regiones de rechazo serán respectivamente

$$\begin{aligned} & (-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty) \\ & (-\infty, -t_{n-1; \frac{\alpha}{2}}) \cup (t_{n-1; \frac{\alpha}{2}}, \infty) \end{aligned}$$

para las pruebas unilaterales se procede como se ha hecho anteriormente.

Ejemplo:

En el ejemplo de los cauchos se tienen los siguientes datos apareados

<i>Automóvil</i>	<i>Neumático A</i>	<i>Neumático B</i>	$D = A - B$
1	10,6	10,2	0,4
2	9,8	9,4	0,4
3	12,3	11,8	0,5
4	9,7	9,1	0,6
5	8,8	8,3	0,5

obtenemos los siguientes valores

$$\begin{aligned} \bar{D} &= 0,48 \\ S_1 &= 0,0837 \end{aligned}$$

(si hubiésemos considerado dos muestras independientes, se hubiese obtenido $S_p = 1,32$). El estadístico de prueba tiene distribución T-estudent con 4 grados de libertad

$$T = \frac{\bar{D}}{\frac{S_1}{\sqrt{n}}}$$

nos da como valor observado $T_{obs} = 0,57$, no se rechaza la hipótesis nula al 95% de confianza.

2.8.4 PROPORCIONES (Muestras Grandes):

- A) Si X es de Bernoulli de parámetro p , X_1, \dots, X_n una muestra aleatoria simple de X :

$$H_0 : p = p_0, \quad H_1 : p > p_0 \text{ ó } p < p_0$$

El estadístico de prueba es:

$$Z = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1)$$

bajo la hipótesis nula, si n es grande. La región de rechazo bilaterales:

$$R = (-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$$

- B) Si X tiene distribución de Poisson de parámetro λ y X_1, \dots, X_n una muestra aleatoria simple de X :

$$H_0 : \lambda = \lambda_0, \quad H_1 : \lambda > \lambda_0 \text{ ó } \lambda < \lambda_0$$

El estadístico de prueba es:

$$Z = \frac{\bar{X} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} \sim N(0, 1)$$

bajo la hipótesis nula. La región de rechazo bilateral viene dada por:

$$R = (-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$$

2.8.5 Comparación de proporciones:

X_1, \dots, X_{n_1} una muestra aleatoria simple de X y Y_1, \dots, Y_{n_2} una muestra aleatoria simple de Y . Las muestras X e Y son independientes de distribución de Bernoulli de parámetros p_1 y p_2 respectivamente:

$$H_0 : p = p_1 = p_2, \quad H_1 : p_1 > p_2 \text{ ó } p_1 < p_2$$

Si denotamos por

$$\bar{p} = \frac{n_1 \bar{X} + n_2 \bar{Y}}{n_1 + n_2},$$

\bar{p} es un estimador insesgado de p bajo la hipótesis nula, por lo tanto es un estimador de la varianza $p(1 - p)$. Se usa el siguiente estadístico de prueba

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\bar{p}(1 - \bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

bajo la hipótesis nula, para muestras grandes, es aproximadamente $N(0, 1)$. Se tiene como región de rechazo:

$$R = (-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty).$$

2.9 Las mejores pruebas.

Una buena prueba estadística es aquella que para un nivel de significación dado o tolerable por el investigador conduzca a una máxima potencia (tenga el menor error de segunda especie). La respuesta a este problema no la da el teorema conocido como Lema de Neyman-Pearson, el cual enunciaremos para hipótesis simples, su demostración se encuentra mas allá de los objetivos de este curso.

Lema de Neyman- Pearson

Sea X_1, \dots, X_n es una muestra aleatoria de tamaño n de una población cuya función de densidad (o función de probabilidad según el caso) es $f(x; \theta)$, se consideran las hipótesis nulas y alternativas siguientes:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

en donde θ_0 y θ_1 se especifican. Si α es el tamaño máximo del error de tipo I que se puede tolerar, entonces la mejor prueba para H_0 contra H_1 es aquella que tiene máxima potencia entre todas las pruebas cuyo error tipo I no sea mayor a α .

Si existe una región crítica C de tamaño α y una constante positiva k tal que

$$\begin{aligned} \frac{L_0(x_1, \dots, x_n; \theta_0)}{L_1(x_1, \dots, x_n; \theta_1)} &\leq k \text{ interior } C \\ \frac{L_0(x_1, \dots, x_n; \theta_0)}{L_1(x_1, \dots, x_n; \theta_1)} &\geq k \text{ exterior } C \end{aligned}$$

entonces C es la mejor región crítica de tamaño α para probar H_0 contra H_1 , en donde L_0 y L_1 son las funciones de verosimilitud relativas a cada una de las hipótesis.

Ejemplo:

Sea X_1, \dots, X_n es una muestra aleatoria de tamaño n de una población normal de media μ desconocida y varianza σ^2 conocida. Determinar la mejor región crítica de tamaño α para probar

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &= \mu_1 \end{aligned}$$

en donde $\mu_1 > \mu_0$. Las funciones de verosimilitud bajo cada hipótesis son:

$$\begin{aligned} L_0(x_1, \dots, x_n; \mu_0) &= (\sqrt{2\pi}\sigma)^{-n} \exp \left[- \sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma^2 \right] \\ L_1(x_1, \dots, x_n; \mu_1) &= (\sqrt{2\pi}\sigma)^{-n} \exp \left[- \sum_{i=1}^n (x_i - \mu_1)^2 / 2\sigma^2 \right] \end{aligned}$$

de acuerdo al lema, la mejor región crítica es aquella para la cual

$$\frac{\exp \left[- \sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma^2 \right]}{\exp \left[- \sum_{i=1}^n (x_i - \mu_1)^2 / 2\sigma^2 \right]} \leq k$$

si tomamos logaritmo después de desarrollar, se puede expresar como

$$\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2 \leq 2\sigma^2 \ln(k)$$

desarrollando y simplificando obtenemos

$$\bar{x} \geq \frac{n(\mu_1^2 - \mu_0^2) - 2\sigma^2 \ln(k)}{2(\mu_1 - \mu_0)}$$

esta expresión define la mejor región crítica para esta prueba, donde $\mu_1 > \mu_0$, de manera sencilla, la mejor región crítica es el extremo derecho de la distribución de muestreo de \bar{X} bajo la hipótesis nula, es decir, dado α , el valor crítico \bar{x}_0 puede encontrarse mediante una adecuada elección de la constante $k > 0$, de manera tal que

$$\mathbb{P}(\bar{X} \geq \bar{x}_0 \mid \mu = \mu_0) = \alpha$$

si en particular $\alpha = 0,05$, como bajo H_0 , \bar{X} tiene distribución normal con media μ_0 y varianza σ^2 entonces $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ es normal estándar y valiéndonos de la tabla para esta distribución se obtiene

$$\mathbb{P}\left(Z \geq \frac{\bar{x}_0 - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_0\right) = 0.05$$

con $\frac{\bar{x}_0 - \mu_0}{\sigma/\sqrt{n}} = 1,645$ o equivalentemente $\bar{x}_0 = \frac{1,645\sigma}{\sqrt{n}} + \mu_0$, la hipótesis H_0 :

$\mu = \mu_0$ se rechazará a favor de $H_1 : \mu = \mu_1 > \mu_0$ cada vez que \bar{X} sea $\geq \frac{1,645\sigma}{\sqrt{n}} + \mu_0$. Note que esta mejor región crítica no depende de $\mu_1 > \mu_0$, por lo que esta región crítica recibe el nombre de región (o prueba) uniformemente más potente para probar $H_0 : \mu = \mu_0$ contra $H_1 : \mu = \mu_1 > \mu_0$.

Observe que esta región coincide con la expuesta anteriormente para este tipo de prueba de hipótesis unilateral en este caso específico. Hemos enunciado este lema y expuesto este ejemplo de manera de “justificar” en cierto modo las regiones críticas seleccionadas para cada uno de los casos presentados en esta guía.

Práctica N°9
Estadística

1. Para probar la hipótesis de que una moneda está bien hecha, se toma la siguiente regla de decisión: se acepta la hipótesis si el número de caras obtenido en una serie de 100 lanzamientos se encuentra entre 40 y 60 (ambos inclusive). De otro modo se rechaza.

- (a) Hallar la probabilidad de rechazar la hipótesis cuando en realidad es cierta.
- (b) Cuál es la probabilidad de aceptar la hipótesis de que la moneda está bien hecha cuando la probabilidad real de obtener cara es $p = 0.7$?
- (c) Denotemos por β la probabilidad de la parte anterior (probabilidad de error de tipo II). El siguiente cuadro muestra los valores de β correspondientes a distintos valores de p :

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	0.00	0.00	0.0192	0.504	0.9642	0.504	0.0192	0.00	0.00

Haga un gráfico de β y de $1 - \beta$ en función de p . Cómo deberían ser los gráficos ideales?

2. Se recibe un envío de latas de conserva, de las que se afirma que el peso medio son 1000 gramos. Examinamos una muestra de 5 de estas latas, obteniendo que el peso medio de la muestra es de 995 grs. Al nivel de confianza 95% se puede aceptar que el peso medio es 100gramos?. Cuál es el máximo valor para el nivel de significación que permitiría aceptar la hipótesis de la empresa de latas de conserva?(Este valor se denomina p valor o valor de significación observado). Suponga distribución normal.
3. Una compañía está interesada en determinar si el promedio de Kms. rodados por vehículos en un mes, de su flota de vehículos asignados a vendedores, ha aumentado por encima del promedio usual de 2600 Kms. Supongamos que la desviación σ es conocida e igual a 35 Kms. Después de examinar 400 vehículos de la flota, en un mes dado, se encontró que el promedio de rodaje de éstos, fue de 2640 Kms. (es decir, hubo un aumento de 1,5% en el rodaje medio mensual para los vehículos de la muestra).

- (a) Al nivel $\alpha = 0,05$ y en base a la muestra, se puede aceptar que el rodaje medio de la flota ha aumentado?
- (b) Calcule el p valor o valor de significación observado de esta flota y discuta el resultado. ¿Son los datos suficientemente significativos para rechazar

$$H_0 : \mu = 2600$$

contra

$$H_1 : \mu > 2600?$$

- (c) ¿Qué pasaría si observásemos un aumento de 1,5% en el rodaje medio mensual para una muestra de 100 vehículos? ¿Qué pasaría si esa muestra fuese de 25 vehículos?
4. Dos investigadores efectúan el conteo de colonias de bacterias en 10 placas de Petri. Los datos obtenidos por cada uno fueron:

Inv.1	63	249	292	79	161	397	118	93	94	163
Inv.2	80	198	323	97	181	416	139	112	118	161

Para $\alpha = 0,05$ ¿qué puede decirse acerca de los métodos de conteo?

5. Un odontólogo afirma que el 40% de los niños de 10 años presentan indicios de caries dental. Se tomó una muestra de 100 niños y se observó que 32 de ellos presentan indicios de caries. Probar la hipótesis del dentista al nivel $\alpha = 0,10$. ¿Cuál es el p valor en este caso? ¿Para qué valores de α se aceptaría y para qué valores se rechazaría la hipótesis del odontólogo?
6. Para estudiar el efecto de la aspirina sobre la coagulación sanguínea, se midió el tiempo de protombina (que es una prueba relacionada con la velocidad de coagulación) a 12 hombres adultos, antes y tres horas después de administrarles 650 mgrs. de aspirina, obteniéndose los siguientes datos, (los cuales indican tiempo en segundo). Suponga distribución normal.

sujeto	1	2	3	4	5	6	7	8	9	10	11	12
Antes	12,3	12,0	12,0	13,0	13,0	12,5	11,3	11,8	11,5	11,0	11,0	11,3
Después	12,0	12,3	12,5	12,0	13,0	12,5	10,3	11,3	11,5	11,5	11,0	11,5

- (a) Para $\alpha = 0,05$ diga si estos datos permiten concluir que la administración de 650 mgs. de aspirina tiene algún efecto sobre el tiempo de protombina. Hágalo de dos maneras:
- Por medio de una prueba de hipótesis bilateral.
 - Por medio de un intervalo de confianza.
- (b) Utilice la tabla para obtener el p valor aproximado.
7. Los electro encefalogramas muestran las fluctuaciones de la actividad eléctrica en el cerebro. Hay diversos tipos de ondas cerebrales: una de ellas son las ondas alfa, cuya frecuencia varía entre 8 y 13 ciclos por segundo. Un grupo de médicos canadienses realizó un estudio acerca de los efectos de la privación sensorial en los patrones de emisión de ondas alfa. Se estudiaron 20 presos, que se dividieron en dos grupos de 10. Cada individuo del primer grupo fue recluido en celda solitaria, los del segundo en celdas usuales. Después de 7 días se midieron las ondas alfa, obteniendo:

Celdas usuales	10,7	10,7	10,4	10,9	10,5	10,3	9,6	11,1	11,2	11,4
Celdas solitarias	9,6	10,4	9,7	10,3	9,2	9,3	9,9	9,5	9,0	10,9

Aparentemente hay un descenso en la frecuencia de las ondas alfa y aumento en la variabilidad cuando las personas se recluyen en soledad. Pruebe si esta diferencia de frecuencias y variabilidad son significativas al nivel $\alpha = 0,05$. Suponga distribución normal.

8. Se considera que el 20% de las personas de una población tienen una determinada característica genética. Sin embargo en un estudio realizado en 100 personas de la población se encontró sólo 15 con la característica. Para $\alpha = 0.01$ ¿aceptaríamos la hipótesis de que la proporción de personas con la característica dada es inferior al 20%? ¿Qué nivel de significación tienen los datos? Cambiaría la respuesta si en una población de 200 personas hubiésemos hallado 30 con la característica?
9. En un estudio diseñado para ver si una dieta controlada puede retardar los efectos de la arterioesclerosis, se hizo un seguimiento a 846 personas elegidas al azar de una población. La mitad de ellas siguió una dieta prefijada y a la otra mitad se le permitió alimentarse como deseara. Al final de 8 años, 66 personas del primer grupo murió de un infarto al

miocardio o infarto cerebral. En el segundo grupo (grupo de control), murieron 93 personas por la misma causa. Haga los análisis estadísticos apropiados.

10. La duración media de una muestra de 12 bombillos es 1250 horas, con una desviación de 115 horas. Se cambia el material del filamento por otro nuevo y entonces de una muestra de 12 bombillos se obtuvo una duración media de 1340 horas, con una desviación de 106 horas. Para $\alpha = 0,05$
 - (a) ¿Puede aceptarse que la varianza no ha cambiado?
 - (b) ¿Ha aumentado la duración media de los bombillos?
11. Los registros de un hospital muestran que de una muestra aleatoria de 1000 hombres que ingresaron al servicio de hospitalización, 52 presentaban problemas cardiacos, mientras que en una muestra aleatoria de 1000 mujeres que ingresaron a este servicio, 23 presentaron problemas cardiacos.
 - (a) ¿Estos datos presentan suficiente evidencia, al nivel $\alpha = 0,05$ para concluir que la proporción de hombres que ingresan al hospital por enfermedades cardiacas es superior a la proporción de mujeres que ingresan al hospital por el mismo motivo?
 - (b) ¿Cuál es el p valor?
 - (c) En base a la respuesta anterior, ¿qué decidiría Ud. para $\alpha = 0,01$?
12. Un proceso químico ha producido en promedio, 800 toneladas por día. Las producciones diarias durante la semana pasada fueron, en toneladas: (suponga distribución normal)

785, 805, 790, 793, 802

- (a) ¿Indican estos datos que la producción promedio fue menor que 800 toneladas y que por lo tanto, hay algo malo en el proceso? Realice una prueba al nivel de significación de 5%.
- (b) Halle el p valor o nivel de significación observado

- (c) Encuentre un intervalo de confianza del 90% para la producción media.

13. Se obtienen dos muestras aleatorias cada una de 11 observaciones, de poblaciones normales con medias μ_1 y μ_2 respectivamente y varianza común σ^2 . Las medias muestrales y las varianzas muestrales centradas son las siguientes:

$$\begin{aligned}\bar{y}_1 &= 60,4 \\ S_1^2 &= 31,40 \\ \bar{y}_2 &= 65,3 \\ S_2^2 &= 44,8\end{aligned}$$

- (a) Al nivel de significación $\alpha = 0,1$ ¿presentan los datos suficiente evidencia para concluir que hay diferencia entre las medias de las poblaciones?
- (b) Halle el p valor o nivel de significación observado.
- (c) Encuentre un intervalo de confianza del 90% para la diferencia de las medias.
14. Se observa durante 20 días la temperatura de operación de dos hornos para el secado de pintura asociado con dos líneas de producción. Suponga distribución normal. Las medias muestrales y las varianzas muestrales centradas son:

$$\begin{aligned}\bar{y}_1 &= 164 \\ S_1^2 &= 81 \\ \bar{y}_2 &= 168 \\ S_2^2 &= 172\end{aligned}$$

¿Presentan los datos suficiente evidencia para concluir que hay diferencia en la variabilidad de la temperatura de los dos hornos? Pruebe la hipótesis $\sigma_1^2 = \sigma_2^2$ contra $\sigma_2^2 > \sigma_1^2$ con un nivel de significación $\alpha = 0,1$. Suponga distribución normal.

15. El fabricante de una máquina para envasar jabón en polvo, afirma que la máquina puede cargar las cajas con un peso dado, con una amplitud (variación) no mayor de $2/5$ de onza. Se encontró que la media y la varianza centrada de una muestra de 8 cajas de tres libras eran iguales a 3,1 y 0,018 respectivamente. Suponga distribución normal.

- (a) Pruebe la hipótesis de que la varianza de la población de pesos es $\sigma^2 = 0,01$ contra $\sigma^2 > 0,01$. Use $\alpha = 0,05$.
- (b) Encuentre un intervalo de confianza para σ^2 y para la media al 90% de confianza.
16. Durante un periodo de 15 días se registran los precios de cierre de dos tipos de acciones. Las medias muestrales y las varianzas centradas muestrales son las siguientes: (suponga distribución normal)

$$\bar{y}_1 = 40,33$$

$$S_1^2 = 1,64$$

$$\bar{y}_2 = 42,54$$

$$S_2^2 = 2,96$$

- (a) ¿Presentan los datos suficiente evidencia para concluir que hay diferencia en la variabilidad de las dos acciones para las poblaciones asociadas con las dos muestras?
- (b) Encuentre un intervalo de confianza para el cociente de las dos varianzas poblacionales al 95% de confianza.

MMOM/2004

Capítulo V

Bondad de Ajuste-Prueba Chi-Cuadrado
Tablas de Contingencia con dos criterios de
clasificación.

Análisis de varianza: comparación de más de
dos medias

Mara Margarita Olivares M.

Junio 2004

0.1 Prueba Chi-cuadrado: Bondad de Ajuste

0.1.1 Introducción:

Hasta ahora todas las situaciones que hemos examinado han tenido como suposición básica que los datos que se tienen provienen de una distribución dada que depende de uno o varios parámetros desconocidos los cuales se pueden estimar por medio de un número, un intervalo de confianza o hacer pruebas de hipótesis referente a ellos.

Si tenemos un conjunto de valores muestrales x_1, x_2, \dots, x_n , correspondiente a una muestra aleatoria simple X_1, X_2, \dots, X_n y se desea saber si hay motivos razonables para considerar la distribución de esta muestra, como una distribución de probabilidad dada, es importante tener criterios para decidir si efectivamente es razonable suponer, basándose en los resultados experimentales, acerca de la veracidad de la hipótesis formulada.

A partir de las observaciones podemos trazar una curva de frecuencias acumuladas (o un histograma) y compararla con la función de distribución de la hipótesis (o función de probabilidad o densidad, según la variable sea discreta o continua) y obtener así una idea, al menos cualitativa de la coincidencia entre ambas distribuciones.

Sin embargo, es necesario, para dar un veredicto preciso, introducir alguna medida cuantitativa del grado de desviación que muestran los datos respecto a la distribución hipotética. Si esta medida excede "algún límite adecuado fijo" debemos rechazar la hipótesis y viceversa.

Tal medida de la desviación se puede definir de diversas formas, nosotros estudiaremos una de ellas: la prueba Chi-Cuadrado introducida por K.Pearson.

Las pruebas que tratan este tipo de problemas, se llaman pruebas de "Bondad de Ajuste".

0.1.2 Caso Discreto:

Supongamos que X es discreta y se realizan n observaciones del experimento en investigación. Sea $\xi_1, \xi_2, \dots, \xi_k$ con $(k \leq n)$, el número de observaciones distintas de la variable X , f_1, f_2, \dots, f_k son las frecuencias correspondientes, es decir, f_i es el número de observaciones iguales a ξ_i . (Eventualmente $f_i = 0$ para algún i).

Sea

$$p_i = \mathbb{P}(X = \xi_i), i = 1, 2, \dots, k; \sum_{i=1}^k p_i = 1$$

la distribución hipotética, la cual suponemos totalmente especificada, es decir, en su expresión no aparecen parámetros desconocidos.

Sea

$$\frac{f_i}{n}$$

el estimador de máxima verosimilitud de p_i , $f_1 + f_2 + \dots, f_k = n$.

Observaciones:

- a) Si n está fijo, f_i es el número de veces que aparece ξ_i en n repeticiones del experimento y p_i representa la probabilidad de obtener ξ_i , luego, f_i tiene distribución binomial de parámetros (n, p_i) , donde $\mathbb{E}(f_i) = np_i$.

La diferencia $f_i - np_i$ mide la desviación entre las frecuencias observadas y las frecuencias esperadas.

K. Pearson demostró que si tomamos

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i} = \sum_{i=1}^k \frac{f_i^2}{np_i} - n \quad (\text{si } \sum_{i=1}^k p_i = 1)$$

obtenemos una medida de la desviación cuyas propiedades son particularmente sencillas:

Se puede demostrar que si $n \rightarrow \infty, (np_i \geq 5)$, el estadístico

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$$

bajo la hipótesis que

$$p_i = \mathbb{P}(X = \xi_i), i = 1, 2, \dots, k$$

es la verdadera distribución (sin parámetros desconocidos) tiene distribución Chi-Cuadrado con $k - 1$ grados de libertad.

(Para una demostración ver : Métodos Matemáticos de Estadística, de Harald Cramer, Cap.XXX, 30.1.; un esbozo de la demostración en Mathematical Statistics, an introduction, Wiebe R. Pestman; una demostración rigurosa en Wilks, S.S. Mathematical Statistics, Jhon Wiley & Sons, Inc., New York 1962)

- b) Si los valores $p_i = \mathbb{P}(X = \xi_i), i = 1, 2, \dots, k$ fueron obtenidos estimando r parámetros desconocidos de la distribución hipotética \mathbb{P} , la expresión

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$$

cuando $n \rightarrow \infty$ ($np_i \geq 5$) tiene distribución Chi-Cuadrado con $k-1-r$ grados de libertad. (Métodos Matemáticos de Estadística, de Harald Cramer, Cap.XXX, 30.3).

Toma de Decisión:

Nosotros queremos que f_i esté cercano a np_i , es decir, que el valor observado

$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$$

esté cercano a cero, procedemos fijando el nivel α de significación de la prueba, hallamos $x_{\alpha, k-1-r}^2$ a partir de

$$\mathbb{P}(\chi^2 > x_{\alpha, k-1-r}^2) = \alpha$$

que representa la probabilidad de rechazar la hipótesis nula, siendo cierta ó probabilidad de cometer un error de primera especie.

Si $\chi_{obs}^2 > x_{\alpha, k-1-r}^2$ rechazamos la hipótesis nula.

Ejemplo:

Después de lanzar un dado 300 veces, se han obtenido las siguientes frecuencias:

cara	1	2	3	4	5	6
frecuencias	43	49	56	45	66	41

al nivel $\alpha = 0.05$, se puede decir que el dado está bien construido?.

1. Si el dado está bien construido debe suceder que

$$H_0 : \frac{1}{6} = \mathbb{P}(X = i), i = 1, 2, \dots, 6$$

En este caso, $\mathbb{E}(f_i) = np_i = 300 \cdot \frac{1}{6} = 50$.

Evaluamos

$$\chi^2_{obs} = \sum_{i=1}^k \frac{(f_i - 50)^2}{50} = 8.96$$

La distribución del estadístico es Chi-Cuadrado con $k - 1 = 5$ grados de libertad. Al buscar en la tabla hallamos que

$$\chi^2_{5;0.05} = 11.07$$

por lo tanto al nivel de 5% aceptamos H_0 pues $\chi^2 < \chi^2_{5;0.05}$ y concluimos que el dado está bien construido.

Para hallar el p -valor debemos calcular $\mathbb{P}(\chi^2_5 > 8.96)$ es algo mayor a 0,10.

0.1.3 Caso Continuo:

Esta es una simple generalización del caso discreto. Se procede de la siguiente manera: sean x_1, \dots, x_n n observaciones de la variable aleatoria X , las cuales tabulamos en una tabla de frecuencias, si k es el número de intervalos de clase, I_i , es el i -ésimo intervalo tal que:

$$\sum_{i=1}^k \mathbb{P}(I_i) = 1,$$

f_i el número de observaciones que caen en I_i , denotando por $p_i = \mathbb{P}(I_i)$ la probabilidad teórica, el estadístico de prueba será:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$$

si $n \rightarrow \infty (np_i \geq 5)$, su distribución es Chi-Cuadrado con $k - 1 - r$ grados de libertad, donde r es el número de parámetros estimados en la distribución teórica.

Si no hay parámetros estimados, en este caso, $r = 0$, la distribución del estadístico será χ^2 con $k - 1$ grados de libertad.

Ejemplos:

1. Un generador de números aleatorios produjo $n = 100$ números, los cuales aparecen tabulados en la siguiente tabla:

Clases	0.0 → 0.099	0.1 → 0.199	0.2 → 0.299	0.3 → 0.399	0.4 → 0.499
Frecuencias	7	14	8	16	6
Clases	0.5 → 0.599	0. → 0.699	0.75 → 0.799	0.85 → 0.8599	0.9 → 0.999
Frecuencias	13	17	4	10	5

Queremos probar al nivel de confianza 99% ($\alpha = 0.01$) la hipótesis que dichos números provienen de una distribución uniforme en $[0, 1]$.

La longitud de cada intervalo de clase I_i es

$$|I_i| = 0.099 \simeq 0.1 = p_i = \mathbb{P}(I_i)$$

si \mathbb{P} es uniforme en $[0, 1]$; $n = 100$, $np_i = 10 \geq 5$, $i = 1, 2, \dots, 10$,

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}; \chi_{obs}^2 = 20$$

tiene aproximadamente distribución Chi-Cuadrado con nueve grados de libertad, en la tabla hallamos que

$$\chi_{9;0.01}^2 = 21,666;$$

luego, al nivel de confianza de 99% aceptamos la hipótesis. Si $\alpha = 0.05$ se rechaza la hipótesis nula pues

$$\chi_{9;0.05}^2 = 16,916;$$

el p valor está entre esos dos niveles de significación, por lo que se rechaza la hipótesis nula.

2. Los resultados del peso en gramos de 570 niños nacidos en un cierto

hospital están tabulados en la siguiente tabla:

Clases	(0, 2400)	(2401, 2600)	(2601, 2800)	(2801, 3000)
Frecuencias	10	13	19	60
Clases	(3001, 3200)	(3201, 3400)	(3401, 3600)	(3601, 3800)
Frecuencias	61	72	92	80
Clases	(3801, 4000)	(4001, 4200)	(4201, 4400)	
Frecuencias	66	48	21	
Clases	(4401, 4600)	(4601, 4800)	4801 →	
Frecuencias	9	15	4	

Queremos probar el ajuste de estos datos a una distribución normal por medios de una prueba Chi-Cuadrado al 95% y 99% de confianza.

Estimamos $\mu = \bar{X} = 3540$; $\sigma^2 = S_1^2 = 283,240$, (la primera y la última clase están abiertas, arbitrariamente hemos tomado en ellas como marcas de clase los valores 1900 y 5100 gramos). Si suponemos $\mu = 3540$; $\sigma^2 = 283,240$, calculamos

$$p_i = \mathbb{P}(I_i)$$

obtendremos:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - 570p_i)^2}{570p_i}; \chi_{obs}^2 = 24,283.$$

Puesto que hemos estimado dos parámetros, la distribución de nuestro estadístico de prueba es asintóticamente Chi-Cuadrado con 11 grados de libertad, obteniendo

$$\chi_{11;0.05}^2 = 19,675; \chi_{11;0.01}^2 = 24,725$$

así el p -valor está entre esos dos niveles por lo que rechazamos la hipótesis nula.

0.1.4 Tablas de Contingencia con dos criterios de clasificación:

Un problema frecuente en el análisis de datos enumerativos es el de la independencia de dos métodos de clasificación de los sucesos observados. Por

ejemplo, clasificamos los defectos de los muebles producidos en una planta de fabricación, primero, de acuerdo al tipo de defecto y segundo, de acuerdo al turno de producción. Lo que deseamos investigar es una posible dependencia entre las dos clasificaciones. Varían las proporciones de los diversos tipos de defectos de un turno a otro?. Por ejemplo, se observa un total de $n = 309$ muebles con defectos y se clasifican en cuatro tipos de defectos : A, B, C, D . Al mismo tiempo, cada mueble se identifica de acuerdo al turno de producción en el que es fabricado.

Tabla de Contingencia

Turnos	Defecto A	Defecto B	Defecto C	Defecto D	Total
1	15(22.51)	21(20.99)	45(38.94)	13(11.56)	94
2	26(22.99)	31(21.44)	34(39.77)	5(11.81)	96
3	33(28.50)	17(26.57)	49(49.29)	20(14.63)	119
Total	74	69	128	38	309

Denotamos por p_A la probabilidad de que el defecto sea del tipo A, análogamente para p_B, p_C, p_D ; estas probabilidades las llamaremos probabilidades de las columnas de la tabla y se satisface:

$$p_A + p_B + p_C + p_D = 1$$

Análogamente $p_i, i = 1, 2, 3$ es la probabilidad de que ocurra un defecto en el turno i (probabilidad de la fila i) donde:

$$p_1 + p_2 + p_3 = 1.$$

Si las clasificaciones son independientes, entonces la probabilidad correspondiente a una celda debe ser el producto de las probabilidades de la fila y de la columna correspondiente a dicha celda. Por ejemplo, la probabilidad de que un defecto particular ocurra en el turno 1 y sea del tipo A debe ser $p_1 p_A$.

La hipótesis nula se refiere a la independencia de las dos clasificaciones. No se especifican los valores numéricos de las probabilidades de las celdas. Por lo tanto, debemos estimar las probabilidades de las filas y de las columnas para poder estimar las frecuencias de celdas esperadas. Los estimadores de máxima verosimilitud de las probabilidades correspondientes a las columnas, son:

$$\begin{aligned} \hat{p}_A &= \frac{c_1}{n} = \frac{74}{309} & \hat{p}_B &= \frac{c_2}{n} = \frac{69}{309} \\ \hat{p}_C &= \frac{c_3}{n} = \frac{128}{309} & \hat{p}_D &= \frac{c_4}{n} = \frac{38}{309} \end{aligned}$$

donde c_i , $i = 1, 2, 3, 4$ es la frecuencia observada de la columna i . Similarmente,

$$\hat{p}_1 = \frac{r_1}{n} = \frac{94}{309} \quad \hat{p}_2 = \frac{r_2}{n} = \frac{96}{309} \quad \hat{p}_3 = \frac{r_3}{n} = \frac{119}{309}$$

son los estimadores de las probabilidades correspondientes a las filas, r_i , $i = 1, 2, 3$ es la frecuencia observada de la fila i .

Si $n_{i,j}$ es la frecuencia observada de la celda que se encuentra en la fila i y la columna j de la tabla de contingencia, entonces, la estimación del valor esperado de n_{ij} es en particular para n_{11}

$$\hat{E}(n_{11}) = n \hat{p}_1 \hat{p}_A = n \frac{r_1}{n} \frac{c_1}{n} = \frac{r_1 c_1}{n},$$

en general,

$$\hat{E}(n_{ij}) = \frac{r_i c_j}{n}, i = 1, 2, 3; j = 1, 2, 3, 4.$$

En nuestro ejemplo, hemos colocado los cálculos de las frecuencias esperadas entre paréntesis, en la tabla de contingencia. El estadístico de prueba, en general, es

$$\chi^2 = \sum_{j=1}^c \sum_{i=1}^r \frac{\left(n_{ij} - \hat{E}(n_{ij}) \right)^2}{\hat{E}(n_{ij})} \sim \chi_{(c-1)(r-1)}^2,$$

donde c es el número de columnas y r es el número de filas.

Observación: El número de grados de libertad debería ser rc menos 1 por la restricción

$$\sum_{j=1}^c \sum_{i=1}^r n_{ij} = n,$$

y debemos restar el número total de estimaciones, es decir, por cada estimación, $(r-1)$ en total por las filas, ya que la r -ésima queda determinada por las primeras $(r-1)$, análogamente, por cada estimación, $(c-1)$ en total por las columnas, se obtiene el número de grados de libertad del estimador:

$$rc - 1 - (r-1) - (c-1) = (r-1)(c-1).$$

En nuestro ejemplo

$$\chi^2 = \sum_{j=1}^4 \sum_{i=1}^3 \frac{\left(n_{ij} - \hat{E}(n_{ij}) \right)^2}{\hat{E}(n_{ij})} \sim \chi_6^2$$

el valor observado del estimador es

$$\chi_{obs}^2 = 19.18$$

A un nivel $\alpha = 0,05$ la región de rechazo viene dada por

$$(\chi_{0.05;6}^2, \infty)$$

donde

$$\mathbb{P}(\chi_6^2 > \chi_{0.05;6}^2) = 0,05$$

Utilizando la tabla se obtiene que $\chi_{0.05;6}^2 = 12,60$ como $\chi_{obs}^2 = 19.18$ cae en la región de rechazo se rechaza la hipótesis nula, es decir, se concluye que no hay independencia entre el turno y el tipo de defecto. El p -valor se calcula hallando

$$\mathbb{P}(\chi_6^2 > 19,18)$$

valiéndonos de la tabla se obtiene un p -valor menor que 0.005, mucho menor que 0,05.

0.1.5 Anova

Análisis de Varianza: Para introducir el método de Análisis de Varianza (ANOVA) vamos a estudiar un ejemplo sencillo:

Supongamos que el número de horas de sueño de los miembros de una familia está dada por:

Adultos	8.4	7.7	7.9
Niños	9.8	9.9	10.3

Queremos constatar si la variación (diferencia entre las medias), es debida a la edad ó no es significativa esa diferencia.

$$\begin{aligned}\bar{y}_1 &= \frac{8.4+7.7+7.9}{3} = 8 \text{ (media del grupo } i = 1 \text{ de adultos)} \\ \bar{y}_2 &= \frac{9.8+9.9+10.3}{3} = 10 \text{ (media del grupo } i = 2 \text{ de niños)}\end{aligned}$$

Si y_{ij} es la observación número j del grupo i :

$$y_{ij} = \bar{y}_i + (y_{ij} - \bar{y}_i)$$

Hagamos una tabla que compare cada resultado con la media de su grupo:

	$j = 1$	$j = 2$	$j = 3$
Adultos ($i = 1$)	$8 + 0.4$	$8 - 0.3$	$8 - 0.1$
Niños ($i = 2$)	$10 - 0.2$	$10 - 0.1$	$10 + 0.3$

Observe que tenemos dos grupos, cada uno con medias diferentes. La media de toda la muestra (uniendo los dos grupos) es:

$$\bar{y} = \frac{8.4 + 7.7 + 7.9 + 9.8 + 9.9 + 10.3}{6} = 9$$

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

Hagamos una tabla que muestre la variación de la media de cada grupo con la media general:

	$j = 1$	$j = 2$	$j = 3$
Adultos ($i = 1$)	$9 - 1 + 0.4$	$9 - 1 - 0.3$	$9 - 1 - 0.1$
Niños ($i = 2$)	$9 + 1 - 0.2$	$9 + 1 - 0.1$	$9 + 1 + 0.3$

donde

\bar{y} es la media general

$(\bar{y}_i - \bar{y})$ compara la media de cada grupo con la media general

$(y_{ij} - \bar{y}_i)$ variación de cada individuo respecto a la media de su grupo

sumamos $i = 1, 2$ (número de grupos), $j = 1, 2, 3$ (observaciones en cada grupo):

$$\sum_{i=1}^2 \sum_{j=1}^3 (y_{ij} - \bar{y})^2 = \sum_{i=1}^2 \sum_{j=1}^3 (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^2 \sum_{j=1}^3 (y_{ij} - \bar{y}_i)^2$$

porque

$$2 \sum_{i=1}^2 \sum_{j=1}^3 (\bar{y}_i - \bar{y})(y_{ij} - \bar{y}_i) = 0.$$

ya que $\sum_{j=1}^3 (y_{ij} - \bar{y}_i) = 3\bar{y}_i - 3\bar{y}_i = 0$.

Esta descomposición es la idea básica del ANOVA, (Análisis de Varianza en inglés), si $N = n_1 + n_2$, n_i es el número de observaciones del grupo i

$$\tilde{S}_2^2 = \frac{1}{N-2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

es un estimador puntual de la varianza σ^2 de la muestra Y_{ij} (se supone que todas las variables tienen la misma varianza). Es fácil ver que

$$\sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^2 n_i (\bar{y}_i - \bar{y})^2.$$

En general: si $(Y_{ij}), i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$; (k es el número de poblaciones o grupos), $Y_{ij} \sim N(m_i, \sigma^2)$:

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \text{ estima } m_i$$

$$\tilde{S}_k^2 = \frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \text{ es un estimador de } \sigma^2, N = n_1 + n_2 + \dots, n_k.$$

Se quiere hacer la siguiente prueba de hipótesis:

$$H_0 : m_1 = m_2 = \dots = m_k \quad H_1 : \text{ existen al menos dos medias diferentes}$$

Bajo la hipótesis nula

$$\tilde{S}^2 = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

estima σ^2 y

$$F = \frac{\tilde{S}^2}{\tilde{S}_k^2} \sim F_{k-1, N-k}$$

Una discrepancia con la hipótesis nula queda indicada por un valor grande de F , ya que el numerador (variabilidad de la media de cada grupo con la media general), cuando la hipótesis nula es falsa, será en promedio más grande que el denominador (variabilidad dentro de cada grupo) por lo que la región de rechazo para un α dado será:

$$[F > f_{k-1, N-k, \alpha}]$$

donde

$$\mathbb{P}([F > f_{k-1, N-k, \alpha}]) = \alpha.$$

En nuestro ejemplo, $k = 2, n_1 = n_2 = 3, N = 6, \tilde{S}^2 = 6, \tilde{S}_2^2 = 0.1, F = 60$, si $\alpha = 0.01$, $f_{1,4,\alpha} = 21.20$ por lo que se rechaza la hipótesis a este nivel. El p valor es 0.015 el cual representa

$$\mathbb{P}([F > 60])$$

Observe que en nuestro ejemplo sólo hay dos grupos, este contraste es idéntico al de la prueba T- de student hecha para comparar dos medias, se puede demostrar que el cociente $\frac{\tilde{S}^2}{S_2^2}$ el cuadrado del estadístico T , $n_1 + n_2 - 2$ grados de libertad.

El método que hemos expuesto, se denomina, Análisis de varianza con un sólo factor o clasificación simple. Fue inventado por Fisher (1925) con el objetivo de descomponer la variabilidad de un experimento (variabilidad total) en componentes independientes que puedan asignarse a diferentes causas. Por ejemplo, si queremos comparar el rendimiento de k máquinas medido por su producción diaria. Existen diversos factores que pueden influir en la producción diaria de cada máquina (aunque trabajen en condiciones idénticas), por ejemplo, pureza de la materia prima, desajustes aleatorios de la máquina, temperatura de funcionamiento, habilidad del operario, etc. Si medimos durante n_i días la producción diaria de la máquina i

$$\sum_{i=1}^k n_i = N \text{ es el total de datos}$$

Si y_{ij} es la producción diaria de la máquina i en el día j , el objetivo del análisis es

1. comprobar si todas las máquinas son idénticas respecto a la producción media diaria
2. Si las máquinas no tienen la misma producción media, estimar la producción media de cada una.

El análisis de varianza formula esta situación mediante un modelo matemático, nosotros tratamos sólo el modelo con un solo factor.

La motivación es la siguiente: ” comparación de medias de poblaciones normales”. Hemos estudiado el problema de comparar dos medias de poblaciones normales, cuando hay igual varianza e independencia de las dos muestras, usando una prueba T de Student, ya sea construyendo un intervalo

de confianza o haciendo una prueba de hipótesis. Mediante el análisis de varianza se generaliza este problema a la comparación de medias de k poblaciones a partir de muestras independientes de tamaños n_1, \dots, n_k . Se trata de hallar una prueba para

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \dots = \mu_k \\ H_1 &: \mu_i \neq \mu_j \text{ para algún } i \neq j \end{aligned}$$

donde Y_{ij} tiene distribución $N(\mu_i, \sigma^2)$; $i = 1, 2, \dots, k$; $j = 1, 2, \dots, n_i$.

El inconveniente de hacer esta prueba dos a dos es que el error de primera especie se incrementa por cada prueba. Fisher desarrolló este método para comparar más de dos medias, comparando la variabilidad interna de los grupos con la variabilidad entre grupos.

Bajo hipótesis nula, Y_{ij} es normal $N(\mu, \sigma^2)$, se tiene la siguiente descomposición de la variabilidad total

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

donde, $\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$, el estadístico de prueba bajo la hipótesis nula y suponiendo igualdad de varianzas es

$$F = \frac{\tilde{S}^2}{\tilde{S}_k^2} \sim F_{k-1, N-k}$$

donde

$$\tilde{S}_k^2 = \frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

representa la variabilidad interna de los grupos y

$$\tilde{S}^2 = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

la variabilidad entre grupos. Valores grandes de F indican que la hipótesis nula no es verdadera.

Observaciones: Los resultados del contraste F en la prueba ANOVA son sustancialmente válidos aunque los datos no sean normales, en ese sentido se dice que es una técnica "robusta" frente a desviaciones de la normalidad.

El efecto de desigualdad de las varianzas en los grupos sobre el contraste F y los contrastes de medias dependen de que el número de observaciones en cada grupo sea igual o muy distinto. Si todos los grupos tienen el mismo número de observaciones, el contraste F es igualmente exacto aunque las varianzas sean distintas. Es decir, podemos despreocuparnos de las varianzas a efectos de contrastes de medias, siempre que haya aproximadamente el mismo número de observaciones por grupo, en caso contrario, diferencias entre las varianzas pueden ser graves.

Prueba χ^2 —Bondad de Ajuste- Independencia
Análisis de Varianza (ANOVA)
Práctica #10
Estadística

1. Después de lanzar un dado 300 veces, se han obtenido las siguientes frecuencias:

Lado	1	2	3	4	5	6
Frecuencia	43	49	56	45	66	41

Al nivel de significación, $\alpha = 0.05$, se puede afirmar que el dado es regular?

2. En el transcurso de dos horas, el número de llamadas por minuto solicitadas a una central telefónica fue:

Llamadas/minuto	0	1	2	3	4	5	6
Frecuencia	6	18	32	35	17	10	2

Al nivel de significación, $\alpha = 0.05$, se puede aceptar que el número de llamadas por minuto sigue una distribución de Poisson?.

3. Se clasificaron 1000 individuos de una población según el sexo y según sean normales o daltónicos:

	masculino	femenino
normales	442	514
daltónicos	38	6

En base a un modelo genético, las probabilidades deben ser:

	masculino	femenino
normales	$\frac{p}{2}$	$\frac{p^2}{2} + pq$
daltónicos	$\frac{q}{2}$	$\frac{q^2}{2}$

donde $q = 1 - p$ = proporción de genes defectuosos en la población. A partir de la muestra se ha estimado que $q = 0.087$ concuerdan los datos con el modelo?

4. El gerente de una planta industrial quiere determinar si el número de empleados que asisten al consultorio médico de la planta está distribuido en forma equitativa durante los 5 días de la semana. Con base

a una muestra aleatoria se observó durante 4 semanas de trabajo, el siguiente número de consultas:

Lunes	Martes	Miércoles	Jueves	Viernes
49	35	32	39	45

Al nivel de significación, $\alpha = 0.05$, existe alguna razón para creer que el número de empleados que asisten al consultorio médico se encuentra distribuido equitativamente durante los cinco días de la semana?

5. Los datos que se dan a continuación corresponden al lapso de tiempo, medido en minutos, necesario para que cada uno de 50 clientes de un banco, lleve a cabo una transacción:

2.3	0.2	2.9	0.4	2.8
2.4	4.4	5.8	2.8	3.3
3.3	4.7	2.5	5.6	9.5
1.8	4.7	2.5	5.6	9.5
1.8	4.7	0.7	6.2	1.2
7.8	0.8	0.9	0.4	1.3
3.1	3.7	7.2	1.6	1.9
2.4	4.6	3.8	1.5	2.7
0.4	1.3	1.1	5.5	3.4
4.2	1.2	0.5	6.8	5.2
6.3	7.6	1.4	0.5	1.4

Al nivel de significación, $\alpha = 0.05$, se puede concluir que los tiempos están exponencialmente distribuidos con parámetro $\lambda = 3.2$ minutos? La densidad exponencial es:

$$f(c) = \frac{\exp(-\frac{x}{\lambda})}{\lambda}, x > 0.$$

6. Un sicólogo desea conocer la relación entre los síntomas "deterioros-sicogénicos del pensamiento y depresión". En una muestra de 100 individuos obtuvo los siguientes datos

	Deterioros Si	Sicogénicos No
Depresión Si	38	31
Depresión No	9	22

Con el nivel de confianza del 95%, existe relación entre ambos síntomas?

7. Una fábrica de automóviles quiere averiguar si la preferencia por un modelo tiene relación con el sexo de los clientes. En una muestra aleatoria de 2000 posibles clientes se observaron las siguientes preferencias:

	Modelo A	Modelo B	Modelo C
Mujeres	340	400	260
Hombres	350	270	380

Existe relación al nivel $\alpha = 0.01$?

8. Para los datos que se dan en la siguiente tabla, pruebe si hay independencia entre la capacidad de una persona para la matemática y su interés en la estadística. Utilice el nivel de significación $\alpha = 0.01$.

	Cap.Mat.baja	Cap.Mat.promedio	Cap.Mat.alta
Interés Est.Baja	63	42	15
Interés Est.promedio	58	61	31
Interés Est.alto	14	47	29

9. Un total de 6000 estudiantes de escuela elemental fueron clasificados de acuerdo a su condición social y a su ubicación en dos tipos de programas educacionales. Los resultados se dan en la tabla. Proporcionan los datos suficiente evidencia para concluir que hay dependencia entre la condición social y la ubicación en los programas educacionales? Use $\alpha = 0.01$

	Cond.Social A	Cond.Social B	Total
Prog.Educativo A	163	117	280
Prog.Educativo B	1477	4243	5720
Total	1640	4360	6000

10. Cuatro grupos de estudiantes se sometieron a técnicas de enseñanza diferentes y se examinaron al final de un período específico de tiempo. Debido a las bajas en los grupos experimentales (por enfermedad, transferencias, etc.) el número de estudiantes en los grupos no fue el mismo. Se tiene la siguiente tabla:

Técnica:1	65, 87, 73, 79, 61, 69
Técnica:2	75, 69, 83, 81, 72, 79, 90
Técnica:3	59, 78, 67, 62, 76
Técnica:4	94, 89, 80, 88

Presentan los datos suficiente evidencia para concluir que hay diferencias en el rendimiento medio correspondientes a las cuatro técnicas? (Realice un análisis de varianza)

11. Un sicólogo clínico quería comparar tres métodos para reducir los niveles de hostilidad en estudiantes universitarios. Cada prueba psicológica (PNH) fue usada para medir el grado de hostilidad. Las puntuaciones altas en esta prueba se usaron como indicación de gran hostilidad. En el experimento se usaron 11 estudiantes que obtuvieron puntuaciones altas y muy cercanas entre sí. De los 11 estudiantes se seleccionaron 5 al azar y se trataron con el método *A*, de los 6 restantes se tomaron tres al azar y se trataron con el método *B* y el resto se trató con el método *C*. Todos los tratamientos se realizaron durante un semestre. Cada estudiante tomó la prueba PNH nuevamente al final del semestre, con los resultados siguientes:

Método: <i>A</i>	73, 83, 76, 68, 80
Método: <i>B</i>	54, 74, 71
Método: <i>C</i>	79, 98, 87

- (a) Realice un análisis de varianza para este experimento.
 - (b) Presentan los datos suficiente evidencia para concluir que hay diferencias entre las respuestas medias de los estudiantes de los tres métodos, después del tratamiento?
12. Se efectúa un experimento para determinar el efecto de la edad sobre el ritmo cardiaco, cuando una persona es sometida a una cantidad específica de ejercicio. Para esto, se seleccionaron 10 varones de los 4 grupos de edades: 10-19-20-39-40-59-60-69. Cada individuo accionó un molino a una velocidad específica durante 12 minutos y se anotó el aumento del ritmo cardiaco (diferencia antes y después del ejercicio), en latidos

por minuto. Los datos se muestran en la tabla.

	Edad			
	[10 – 19]	[20 – 39]	[40 – 59]	[60 – 69]
	29	24	37	28
	33	27	25	29
	26	33	22	34
	27	31	33	36
	39	21	28	21
	35	28	26	20
	33	24	30	25
	29	34	34	24
	36	21	27	33
	22	32	33	32
Total	309	275	295	282

- ¿Presentan los datos suficiente evidencia que indique que hay diferencia entre el aumento medio del ritmo cardiaco de los 4 grupos de edades?. Use $\alpha = 0.05$.
- Encuentre un intervalo de confianza la 90% para la diferencia entre el aumento medio del ritmo cardiaco del grupo de 10-19 años y el del grupo de 60-69 años.
- Encuentre un intervalo de confianza del 90% para el aumento medio del ritmo cardiaco del grupo de 20-39 años.
- Aproximadamente, cuántas personas se necesitarían en cada grupo si se desea estimar la media de un grupo con un error no mayor de 2 latidos por segundo, con una probabilidad igual a 0,95?.

MMOM/04

Capítulo VI

Recta de Regresión lineal

Prof. María Margarita Olivares

Julio 2004

1 Recta de mínimos cuadrados.

En muchos problemas obtenemos datos pareados (x_i, y_i) , no conocemos la distribución conjunta de las variables aleatorias correspondientes y al graficar estos datos tenemos la impresión de que una recta podría ser un buen ajuste para ellos, aunque los puntos no estén exactamente sobre una recta. Problemas de este tipo, suelen manejarse por medio del método de los mínimos cuadrados que consiste en hallar la recta

$$y = ax + b$$

que mejor se ajusta a esos datos, para ello debemos calcular los parámetros a y b a partir de los datos, es decir:

si nos dan un conjunto de datos pareados $\{(x_i, y_i); i = 1, 2, 3, \dots, n\}$, las estimaciones de mínimos cuadrados de los coeficientes a y b son los valores para los cuales la cantidad:

$$q(a, b) = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

es un mínimo. Al diferenciar parcialmente con respecto a a y a b y al igualar estas derivadas parciales a cero:

$$\begin{aligned}\frac{\partial q}{\partial a} &= (-2) \sum_{i=1}^n [y_i - (a + bx_i)] = 0 \\ \frac{\partial q}{\partial b} &= (-2) \sum_{i=1}^n x_i [y_i - (a + bx_i)] = 0\end{aligned}$$

que producen el siguiente sistema de ecuaciones:

$$\begin{aligned}\sum_{i=1}^n y_i &= an + b \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i &= a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2\end{aligned}$$

Al resolver ese sistema de ecuaciones se obtiene:

$$\begin{aligned}a &= \bar{y} - b\bar{x} \\ b &= \frac{S_{xy}}{S_{xx}}\end{aligned}$$

donde :

$$\begin{aligned}S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \\ S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)\end{aligned}$$

2 Regresión Lineal Simple.

Introducción:

En este curso hemos estudiado el siguiente tipo de problema:

se tiene una variable aleatoria Y (que denominaremos respuesta) cuya distribución se supone conocida, excepto por uno o varios parámetros. En particular se estudiaron los siguientes casos:

1. Y con distribución de Bernoulli de parámetro p desconocido, donde

$$\mathbb{P}(Y = 1) = p, \mathbb{P}(Y = 0) = 1 - p, \mathbb{E}(Y) = p$$

2. Y con distribución de Poisson de parámetro λ , $\mathbb{E}(Y) = \lambda$ es desconocido.
3. Y es normal de parámetros desconocidos μ, σ , con $\mathbb{E}(Y) = \mu$, $Var(Y) = \sigma^2$.

y en base a una muestra y_1, \dots, y_n de la variable aleatoria Y se desea estimar dichos parámetros. Como es observable, uno de los parámetros básicos a observar es la media de la variable aleatoria.

Se estudiaron tres maneras de estimar:

- (a) Estimación puntual
- (b) Estimación por intervalos de confianza
- (c) Pruebas de hipótesis.

En cualquiera de estos casos lo primero que hay que hacer es encontrar un estadístico para estimar el parámetro (obtener un estimador del parámetro).

Hagamos incapié en el caso Y es normal de parámetros desconocidos μ, σ , con $\mathbb{E}(Y) = \mu, Var(Y) = \sigma^2$ que es el que se asemeja más al problema que estamos tratando de estudiar. En este caso podemos estimar la media μ por

$$\bar{y} = \frac{y_1 + \cdots + y_n}{n}$$

o promedio de valores de la muestra y estimamos σ^2 por

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

o variabilidad centrada de la muestra alrededor de su media, en ambos casos los estadísticos son insesgados.

Suponer que Y tiene distribución normal $N(\mu, \sigma^2)$ es equivalente a suponer que

$$Y = \mu + \varepsilon$$

donde ε es una variable aleatoria normal de media cero y varianza σ^2 .

Con los estadísticos \bar{Y} y S_1^2 o con $S^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$ se construye los llamados estadísticos pivotaes, por ejemplo, si σ es conocido, el estadístico pivotal para μ es

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

que tiene distribución normal estándar.

A partir de los estadísticos pivotaes obtenemos intervalos de confianza, podemos hacer pruebas de hipótesis.

Generalización del problema de estimación de una variable aleatoria .

Supongamos que tenemos una variable aleatoria Y (respuesta) que depende de una o varias variables x (o factores). Estos factores se pueden dividir en dos grupos:

1. Un factor o un grupo de factores (x_1, \dots, x_k) conocidos al observar Y . Suponemos que la dependencia de Y de estos factores tiene una forma funcional conocida.
2. Otro grupo de factores que pueden ser conocidos o no al observar Y , no sabemos necesariamente cómo depende de ellos, pero suponemos que cada uno influye en la respuesta Y , en pequeña magnitud.

En resumen, suponemos que

$$Y = f(x_1, \dots, x_k) + \varepsilon$$

siendo ε una variable aleatoria y en el caso de un solo factor

$$Y = f(x) + \varepsilon$$

Ejemplos:

1. Y = beneficio anual obtenido por una corporación
 X = gastos de publicidad.
2. Y = aumento de peso, en un mes, de cierta raza de ganado que se alimenta con cierto tipo de alimento.
 X_1 = peso inicial del animal.
 X_2 = cantidad de alimento consumido diariamente.
 X_3 = contenido de proteínas del alimento.
 X_4 = contenido de agua del alimento.
 X_5 = contenido de carbohidratos del alimento.
 $X_6 = \begin{cases} 1 & \text{el animal estuvo sano en el mes} \\ 0 & \text{el animal estuvo enfermo en el mes} \end{cases}$
3. Y = nota obtenida por un estudiante de ingeniería en matemática I
 X = nota obtenida por un estudiante en la prueba de admisión.

En todos estos casos vamos a querer estimar la media de Y como función de X o de X_1, \dots, X_k , queremos contestar preguntas como la siguiente:

1. ¿Cuál es el beneficio medio de la corporación para un gasto dado de publicidad $X = x$?
2. ¿Cuál es el aumento de peso promedio para los animales que no se enfermaron en un mes y que recibieron un alimento que contenía 15% de proteínas, 10% de agua y el resto de carbohidratos?

Estos modelos

$$Y = f(x_1, \dots, x_k) + \varepsilon$$

que relacionan una variable aleatoria dependiente Y con una o varias variables independientes x_1, \dots, x_k no aleatorias se denominan modelos de regresión. Si tenemos una sola variable independiente

$$Y = f(x) + \varepsilon$$

hablamos de modelos de regresión simple. Si tenemos varias variables independientes hablamos de regresión múltiple.

El modelo que estudiaremos es el de regresión lineal simple, nos limitaremos al caso en que f es una función lineal de x :

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

donde ε es una variable aleatoria, se estudiará la dependencia lineal de una variable aleatoria Y (respuesta) respecto de una variable explicativa x . En general, la palabra "lineal" se refiere a los parámetros, es decir, ningún parámetro en un modelo lineal, aparecerá con exponente o multiplicado o dividido entre otro parámetro. Es decir los siguientes modelos son también lineales en los parámetros:

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \\ Y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon \\ Y &= \beta_0 + \beta_1 \ln(x) + \varepsilon \\ Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_3 + \varepsilon \end{aligned}$$

pero el modelo $Y = \beta_0 \exp(\beta_1 x) + \varepsilon$ no es lineal en los parámetros. Nosotros estudiaremos sólo el caso $f(x) = \beta_0 + \beta_1 x$.

Hipótesis Básicas:

Se observan n variables respuestas (aleatorias) $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$ que corresponden a cada variable no aleatoria x_i (variable independiente explicativa). Las variables Y_i son independientes estocásticamente

(independencia en el sentido aleatorio o probabilístico), es decir, ε_i y ε_j son independientes para $i \neq j$. ε_i tiene distribución $N(0, \sigma^2)$. Todas las observaciones tienen la misma varianza, $Var(Y_i) = \sigma^2$; $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$. Por lo que estimar la media de las variables respuestas es equivalente a estimar los parámetros de la recta de regresión.

Método de Mínimos Cuadrados:

Para obtener los estimadores de mínimos cuadrados de los parámetros β_0, β_1 se considera la desviación de la observación Y_i de su media y se determinan los valores de los parámetros que minimizan la suma de los cuadrados de estas desviaciones. La i -ésima desviación o i -ésimo error es

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$$

y la suma de los cuadrados de los errores es

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$$

diferenciando con respecto a los parámetros e igualando a cero obtenemos las siguientes ecuaciones

$$\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \end{aligned}$$

que conducen a obtener los estimadores buscados, el estimador mínimo cuadrado de β_0 es

$$B_0 = \bar{y} - B_1 \bar{x}$$

con

$$B_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(x, y)}{S_x^2}$$

donde

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}); S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

es el estimador de mínimos cuadrados de β_1 correspondiente a las observaciones (x_i, y_i) . Finalmente, dados los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ la recta de regresión estimada para el modelo es

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

donde \hat{Y}_i es el estimador para la media de la observación Y_i , que corresponde al valor x_i de la variable de predicción.

Observación:

Como notación, utilizaremos $\hat{\beta}_1$ y $\hat{\beta}_0$ para referirnos al estimador de β_1 y β_0 respectivamente y B_1 y B_0 para los estimadores observados, es decir, evaluados en las observaciones.

Si sustituimos la relación entre los parámetros

$$B_0 = \bar{y} - B_1 \bar{x}$$

en la recta de regresión estimada, obtenemos

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1(x_i - \bar{x})$$

La diferencia entre el valor observado y el estimado se denomina *iésimo* residuo o residual:

$$e_i = y_i - \hat{y}_i$$

y son estimadores de la variable aleatoria no observable ε_i , proporcionan importante información de lo que puede faltar en el modelo de regresión estimado. La varianza σ^2 en general no se conoce y se estima por la varianza residual (su raíz cuadrada es la desviación estándar residual)

$$S_R^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

se usa $n-2$ pues se pierden dos grados de libertad al estimar los dos parámetros β_0, β_1 . Este estimador de la varianza es insesgado si el modelo de regresión es correcto.

Se cumplen las siguientes propiedades: (se deja como ejercicio la verificación)

1. $\sum_{i=1}^n e_i = 0$

$$2. \sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$$

$$3. \sum_{i=1}^n x_i e_i = 0$$

Estimación por Máxima Verosimilitud para el Modelo Lineal Simple.

Para hallar los estimadores mínimos cuadrados de los parámetros no se necesitó utilizar la distribución de los errores aleatorios ε_i . Bajo nuestras hipótesis es posible obtener los estimadores de máxima verosimilitud ya que Y_i son variables aleatorias independientes de distribución $N(\beta_0 + \beta_1 x_i, \sigma^2)$ dado que es una función lineal de una variable aleatoria ε_i con distribución normal. La función de verosimilitud viene dada por

$$L(y_1, \dots, y_n; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right]$$

donde

$$\ln(L(\beta_0, \beta_1, \sigma^2)) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

al tomar las derivadas parciales con respecto a cada uno de los parámetros se obtendrá que los estimadores de β_0, β_1 son los mismos hallados por mínimos cuadrados y para σ^2 se obtiene

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

que no es insesgado pero para valores grandes de n no se diferencia mucho del S_R^2 que si es insesgado.

Los estimadores de máxima verosimilitud tienen buenas propiedades, son consistentes, suficientes y tienen varianza mínima es por eso que mostramos que los estimadores hallados para los parámetros coinciden con los de máxima verosimilitud.

Interpretación Geométrica de la estimación.

El método de mínimos cuadrados admite una simple interpretación geométrica. Expresando vectorialmente las observaciones, definamos los siguientes vectores filas:

$$\begin{array}{ll} \mathbf{Y}' = (y_1, \dots, y_n) & \mathbf{1}' = (1, \dots, 1) \\ \mathbf{X}' = (x_1, \dots, x_n) & \boldsymbol{\varepsilon}' = (\varepsilon_1, \dots, \varepsilon_n) \end{array}$$

El modelo postulado es: $\mathbf{Y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{X} + \boldsymbol{\varepsilon}$ $\mathbf{e}' = (e_1, \dots, e_n)$

Estimar el modelo por mínimos cuadrados equivale a encontrar constantes B_0, B_1 tales que el módulo del vector de residuos

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

sea mínimo, es decir, se trata de encontrar un vector $\hat{\mathbf{Y}}$ en el plano definido por los vectores $\mathbf{1}$ y \mathbf{X} tales que su distancia al vector \mathbf{Y} sea mínima, la solución es la proyección ortogonal del vector \mathbf{Y} sobre este plano, cualquier otra elección nos daría un vector residual \mathbf{e} de módulo mayor. El vector de residuos \mathbf{e} será perpendicular a todos los vectores del plano, en particular su producto escalar con los vectores $\mathbf{1}$ y \mathbf{X} es cero obteniéndose así las ecuaciones llamadas normales o mínimo- cuadráticas

$$\begin{aligned} \mathbf{e}'\mathbf{1} &= \sum_{i=1}^n e_i = 0 \\ \mathbf{e}'\mathbf{X} &= \sum_{i=1}^n e_i x_i = 0 \end{aligned}$$

observe que al reemplazar

$$\mathbf{e}' = \left(\mathbf{Y} - \hat{\mathbf{Y}} \right)'$$

con $\hat{\mathbf{Y}} = B_0 \mathbf{1} + B_1 \mathbf{X}$, B_0 es la estimación de β_0 y B_1 es la de β_1 obtendremos las ecuaciones

$$\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \end{aligned}$$

que permitieron hallar los estimadores mínimo-cuadrado.

Propiedades de los Estimadores:

1. **Coefficiente de Regresión** $\hat{\beta}_1$

Este estimador se puede expresar como

$$\hat{\beta}_1 = \sum_{i=1}^n w_i Y_i$$

con $w_i = \frac{(x_i - \bar{x})}{nS_x^2}$ pues

$$B_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{nS_x^2}$$

así, $\hat{\beta}_1$ es combinación lineal de variables Y_i normales, por lo tanto tiene distribución normal. Observe que

$$\sum_{i=1}^n w_i = 0, \sum_{i=1}^n w_i^2 = \frac{1}{nS_x^2}$$

y que si denotamos por $p_i = \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$ la pendiente de la recta que une (x_i, y_i) con (\bar{x}, \bar{y})

$$B_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 p_i}{nS_x^2} = \sum_{i=1}^n d_i p_i$$

es una ponderación de las pendientes $p_i = \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$ con pesos que dependen de la distancia relativa de cada punto x_i y el centro de todos ellos, note que

$$d_i = \frac{(x_i - \bar{x})^2}{nS_x^2} = w_i(x_i - \bar{x}) \geq 0, \sum_{i=1}^n d_i = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{nS_x^2} = 1.$$

El estimador $\hat{\beta}_1$ es insesgado, en efecto

$$\mathbb{E}(\hat{\beta}_1) = \sum w_i \mathbb{E}(Y_i) = \beta_0 \sum w_i + \beta_1 \sum w_i x_i = \beta_1$$

Calculemos la varianza de $\hat{\beta}_1$

$$\text{Var}(\hat{\beta}_1) = \sum w_i^2 \text{Var}(Y_i) = \sigma^2 \frac{1}{n S_x^2}$$

por lo tanto $\hat{\beta}_1$ tiene distribución $N\left(\beta_1, \sigma^2 \frac{1}{n S_x^2}\right)$, y así, es un estimador insesgado y consistente del coeficiente de regresión.

2. Estimador $\hat{\beta}_0$ (ordenada en el origen)

Es preferible escribir la recta de regresión en función del estimador del coeficiente de regresión, pues a veces la relación observada no tiene sentido para $x = 0$

$$\hat{y} = \bar{y} + B_1(x - \bar{x})$$

esta expresión pone de manifiesto que la relación construida es válida en un entorno del punto (\bar{x}, \bar{y}) que representa el centro de las observaciones que se utiliza para construir el modelo. Estudiaremos las propiedades del estimador $\hat{\beta}_0$ ya que puede ser de interés:

$$B_0 = \bar{y} - B_1 \bar{x} = \frac{1}{n} \sum y_i - \bar{x} \sum w_i y_i = \sum \left(\frac{1}{n} - \bar{x} w_i \right) y_i = \sum r_i y_i$$

es decir

$$\hat{\beta}_0 = \sum r_i Y_i$$

por lo tanto también obtenemos que $\hat{\beta}_0$ tiene distribución normal.

$\hat{\beta}_0$ es insesgado, en efecto

$$\mathbb{E}(\hat{\beta}_0) = \sum \left(\frac{1}{n} - \bar{x} w_i \right) (\beta_0 + \beta_1 x_i) = \beta_0 (1 - \bar{x} \sum w_i) + \beta_1 \bar{x} - \beta_1 \bar{x} \sum w_i x_i = \beta_0$$

Hallemos la $Var(\hat{\beta}_0)$:

$$Var(\hat{\beta}_0) = \sum r_i^2 \sigma^2 = \sigma^2 \sum \left(\frac{1}{n} - \bar{x}w_i \right)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2} \right)$$

note que $\frac{\sigma^2}{n}$ es el error de la estimación de \bar{Y} pues $Var(\bar{Y}) = \frac{\sigma^2}{n}$; el error de estimación de la pendiente de la recta de regresión viene dado por $Var(\hat{\beta}_1) = \sigma^2 \frac{1}{nS_x^2}$ el cual aparece reflejado en la $Var(\hat{\beta}_0)$ y se transmite a la ordenada del origen en función de lo alejado que esté \bar{x} del origen.

Se concluye que $\hat{\beta}_0$ tiene distribución $N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)\right)$ siendo así insesgado y consistente..

Observaciones:

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ no son independientes, aceptaremos sin demostración que

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{nS_x^2}$$

y que \bar{Y} y $\hat{\beta}_1$ son independientes

3. Varianza residual \hat{S}_R^2 (estimador de σ^2)

Recordemos que

$$S_R^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{y}_i \right)^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

se puede demostrar que $\frac{(n-2)\hat{S}_R^2}{\sigma^2}$ tiene distribución χ_{n-2}^2 (chi-cuadrado con $n-2$ grados de libertad), la esperanza de esta distribución es $n-2$ y la varianza es $2(n-2)$ por lo que se concluye que

$$\mathbb{E}\left(\hat{S}_R^2\right) = \frac{\sigma^2}{n-2} \mathbb{E}\left(\frac{(n-2)\hat{S}_R^2}{\sigma^2}\right) = \sigma^2$$

y que

$$Var(\hat{S}_R^2) = \frac{\sigma^4}{(n-2)^2} Var\left(\frac{(n-2)\hat{S}_R^2}{\sigma^2}\right) = \frac{2\sigma^4}{(n-2)}$$

así \hat{S}_R^2 es un estimador insesgado y consistente de σ^2 .

Hemos denotado por \hat{S}_R^2 el estimador insesgado de σ^2 en función de las variables aleatorias Y_i y S_R^2 cuando este estimador es evaluado en las observaciones.

Inferencia sobre los parámetros: intervalos de confianza y pruebas de hipótesis.

A partir de las distribuciones de los estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{S}_R^2$ de los parámetros podemos construir estadísticos que nos permitan hacer inferencias, el siguiente cuadro muestra los estadísticos que nos permiten hacer el contraste

parámetro	β_0	β_1	σ^2
estimador	$\hat{\beta}_0$	$\hat{\beta}_1$	\hat{S}_R^2
estadístico	$T_{n-2} = \frac{\sqrt{n}(\hat{\beta}_0 - \beta_0)}{S_R \sqrt{1 + \frac{\bar{x}^2}{S_x^2}}}$	$T_{n-2} = \frac{\sqrt{n}(\hat{\beta}_1 - \beta_1) S_x}{S_R}$	$\frac{(n-2)\hat{S}_R^2}{\sigma^2} \sim \chi_{n-2}^2$

Coefficiente de determinación o R^2

El coeficiente de determinación se define como

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

el numerador (VE) es la variabilidad explicada por la regresión y el denominador (VT) la variabilidad total. En el caso que estamos tratando, tenemos una recta de regresión lineal simple así:

$$VE = \sum (\hat{y}_i - \bar{y})^2 = \sum (\bar{y} + B_1(x_i - \bar{x}) - \bar{y})^2 = B_1^2 \sum (x_i - \bar{x})^2 = B_1^2 n S_x^2$$

$$VT = n S_y^2$$

por lo tanto

$$R^2 = \frac{B_1^2 n S_x^2}{n S_y^2} = \frac{B_1^2 S_x^2}{S_y^2} = \frac{(cov(x, y))^2 S_x^2}{S_x^4 S_y^2} = \frac{(cov(x, y))^2}{S_x^2 S_y^2} = r^2$$

donde r es el coeficiente de correlación asociado a (x_i, y_i) el cual se define como

$$r = \frac{cov(x, y)}{S_x S_y}$$

es decir, en el caso de regresión lineal simple, cuando se trata de una recta de regresión, el coeficiente de determinación coincide con el coeficiente de correlación lineal.

Se concluye que si la regresión o relación lineal entre x e y es exacta, $|r| = 1$.

Si no existe relación entre las variables ($\mathbb{E} \left(\hat{\beta}_1 \right) \sim 0$, \hat{y}_i es próximo a \bar{y}) en este caso $r \sim 0$.

La definición de R^2 es general, $R^2 = r^2$ (coeficiente de correlación de Pearson) solo en el caso de regresión lineal simple, cuando la regresión es una recta.

El coeficiente de correlación se usa para comparar rectas de regresión entre sí, pero se debe evitar su uso indiscriminado. Dos rectas de regresión pueden tener la misma eficacia predictiva y los mismos errores de estimación y sin embargo tener diferentes coeficientes de correlación.