

ESTADISTICA

2– Presentación de la información estadística.

2.1– Etapa para la realización de un estudio estadístico, clasificación de la presentación de los datos.

Las etapas para la realización de un estudio estadístico comienza con la recolección de datos, la definición de los intervalos.

La diferencia entre el mayor y el menor de los números, se llama recorrido o rango de los datos.

Posteriormente se construye la distribución de frecuencia y termina la etapa descriptiva.

Se elabora luego el modelo matemático teórico y luego se informa las conclusiones. Estas se expresan en términos de probabilidad de ocurrencia de cierto fenómeno.

El azar es la medida de muestra ignorancia.

Presentación de Datos

Una curva es una representación gráfica de las relaciones entre variables. En estadística se emplean muchos tipos de curvas, dependiendo de la naturaleza de los datos y del propósito para el que la curva ha sido proyectada.

Entre estas están los gráficos de barra, pictogramas, etc., que se conocen como grafos o diagramas.

Veremos esto mejor con un ejemplo.

Consideremos la tabla 1 de la población de los Estados Unidos (en millones) en los años 1840, 1850, ...1960 y representemos luego esos datos.

TABLA 1 – Población (en mill9 de EEUU1840–1960

1860	AÑO	1840	1850	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960
31,4	mill	17,1	23,2	39,8	50,2	62,9	76	92	105,7	122,8	131,7	151,1	179,3

Primer Método

En la fig. 1 la población se denota por la letra P que es la variable dependiente. Los puntos se sitúan en el sistema de coordenadas sacándolos de la tabla, se advierte que las unidades sobre los ejes no son iguales y esto está justificado, porque las variables son esencialmente distintas.

Posteriormente todos los puntos se unen por líneas rectas, puesto que no han información de la variación de la población en los años intermedios, el cero se representa en el eje vertical.

Esta representación se denomina Gráfico de Líneas .

Segundo Método (fig. 2)

La fig. 2 muestra un diagrama de barras o gráfico de barras. Las añaduras de las barras son iguales, no tienen significación en éste caso y pueden elegirse de cualquier tamaño con tal que se mantenga una separación entre barras. Los números en el tope de cada barra pueden omitirse si se desea.

Tercer Método (fig. 3)

- * * 17,1 mill * Población de los EEUU durante 1840–1960
- * * * 23,2 mill cada * representa 10.000.000 de individuos
- * * * * 31,4 mill
- * * * * * 39,8 mill
- * * * * * * 50,2 mill
- * * * * * * * 62,9 mill
- * * * * * * * * 76 mill
- * * * * * * * * * 92 mill
- * * * * * * * * * * 105,7 mill
- * * * * * * * * * * * 122,8 mill
- * * * * * * * * * * * * 131,7 mill
- * * * * * * * * * * * * * 151,1 mill

1960 * * * * * * * * * * * * * 179,3 mill

Los diagramas de este tipo se llaman pictografías o pictogramas y son utilizados a menudo para presentar datos estadísticos en una forllame la atención del público que lo vea.

Muchos de estos pictogramas presentan una gran originalidad e ingeniosidad en su diseño.

Los números a la derecha del pictograma pueden ser incluidos o no.

2.2– Concepto de variable y su clasificación. Tablas de Frecuencia. Distribución de Frecuencias. Histogramas. Polígono de Frecuencias. Frecuencias relativas y acumuladas.

Variable: es un símbolo tal como x, 4, h, que puede tomar un valor cualquiera de un conjunto determinado de ellos, llamado dominio de la variable.

Si la variable puede tomar solo un valor se llama constante, una variable que teoricamente puede tomar cualquier valor entre dos valores dados se llama continua, si no es así se llama discreta.

Ejemplo 1: El número de hijos de una familia es una variable discreta.

Ejemplo 2: La altura de un individuo es una variable continua.

Los datos que vienen definidos por una variable discreta o continua, se llaman datos discretos o datos continuos, respectivamente. El número de hijos de 1000 familias es un ejemplo de datos discretos, mientras que las alturas de 100 estudiantes universitarios es un ejemplo de datos continuos.

En general las medidas dan origen a datos continuos mientras que las enumeraciones o conteos originan datos discretos.

Tablas de Frecuencia: son las representaciones de las ordenaciones de los datos en orden creciente o decreciente. La diferencia entre el mayor y el menor de los números se llama rango de los datos, y cuando hay muchos datos algunos pueden repetirse.

Distribuciones de Frecuencias: cuando se dispone de gran número de datos es útil distribuirlos en clases o categorías y determinar el número de individuos pertenecientes a cada clase o categoría que es la Frecuencia de Clase.

Una ordenación tabular de los datos en clases, reunidas las clases con las frecuencias correspondientes a cada una, se conoce como una Distribución de Frecuencias o Tabla de Frecuencias.

Ejemplo: Tabla de Frecuencia – Altura de 100 estudiantes.

| ALTURA (Pulgadas) | NUMERO DE ESTUDIANTES |
|-------------------|-----------------------|
| 60 – 62 | 5 |
| 63 – 65 | 18 |
| 66 – 68 | 42 |
| 69 – 71 | 27 |
| 72 – 74 | 8 |
| TOTAL | 100 |

La primera clase comprende las alturas de 60 a 62 pulgadas, puesto que 5 estudiantes tienen una altura perteneciente a esa clase, la correspondiente frecuencia de clase es 5.

Los datos ordenados y requeridos como en la distribución de frecuencia anterior, se suelen llamar Datos Agrupados, aunque en el proceso de agrupamiento se pierden detalles del valor original de los datos, tiene la ventaja de presentarlos todos en un cuadro sencillo que facilita encontrar las relaciones entre ellos.

Intervalos de Clase y Límites de Clase

Un símbolo que define una clase tal como 60–62 de la tabla anterior, se conoce como intervalo de clase. Los números extremos 60 y 62 son los Límites de Clase, el número menor es el Límite inferior de la clase y el mayor es el Límite superior de la clase; Límite Real.

Los términos clase e intervalo de clase se usan a menudo indistintamente, aunque el intervalo de clase es realmente un símbolo para la clase.

Marca de Clase

La marca de clase es el punto medio del intervalo de clase y se obtiene sumando los límites superior e inferior de la clase y dividiendo por 2.

La marca de clase en el intervalo 60–62 es $(60+62)\%2=61$, también se denomina punto medio de clase.

Reglas Generales para formar las Distribuciones de Frecuencia.

- Determinar el mayor y el menor de los datos y encontrar el rango.
- Dividir el rango en un número conveniente de intervalos de clase del mismo tamaño. Los intervalos de clase se eligen de forma que las marcas de clase o puntos medios coinciden con los datos realmente observados. Esto tiende a reducir el error de agrupamiento en los análisis matemáticos posteriores.
- Determinar el número de observaciones que caen dentro de cada intervalo de clase, es decir encontrar las frecuencias de clase. Para esto se utiliza una hoja de conteo.

Ejemplo: En la siguiente tabla los pesos de 40 estudiantes de la State University se registran con aproximaciones de una libra, construir una distribución de frecuencias.

| | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|
| 138 | 164 | 150 | 132 | 144 | 125 | 149 | 157 |
| 146 | 158 | 140 | 147 | 136 | 148 | 152 | 144 |
| 168 | 126 | 138 | 176 | 163 | 119 | 154 | 165 |
| 146 | 173 | 142 | 147 | 135 | 153 | 140 | 135 |
| 161 | 145 | 135 | 142 | 150 | 156 | 145 | 128 |

Solución

- El peso mayor es 176 libras y el menor 119 libras, de modo que el rango es $176-119=57$ libras.
- 2- Si utilizamos 5 intervalos de clase, el tamaño de cada una será de $57/5=11$ aprox. Si utilizamos 20 intervalos de clase, el tamaño será de $57/20=3$ aprox. Una elección conveniente para el tamaño del intervalo es de 5 libras. También conviene elegir las marcas de clase en 120,125,130,135....libras. Así los intervalos de clase pueden ser 118-122, 123-127, 128-132.

La distribución de frecuencias pedida aparece en la tabla siguiente.

En la columna central, llamada columna de conteo se tabula las frecuencias de clase de la totalidad de los datos y generalmente se omite en la presentación final de la distribución de frecuencias.

| PESO(libras) | CONTEO | FRECUENCIA |
|--------------|--------|------------|
| 118-122 | 1 | 1 |
| 123-127 | 2 | 2 |
| 128-132 | 2 | 2 |
| 133-137 | 4 | 4 |
| 138-142 | 5 | 6 |
| 143-147 | 7 | 8 |
| 148-152 | 4 | 5 |
| 153-157 | 4 | 4 |
| 158-162 | 2 | 2 |
| 163-167 | 3 | 3 |
| 168-172 | 1 | 1 |
| 173-177 | 2 | 2 |
| | | 40 |

Otra distribución de frecuencias sobre el mismo evento estudiado se observa en la tabla siguiente, en la que se han distinguido solo 7 clases el intervalo es 9 libras.

Como se observa en esta última distribución los datos están más agrupados en ciertos intervalos de clase.

| PESO(libras) | CONTEO | FRECUENCIA |
|--------------|--------|------------|
| 118-126 | 3 | 3 |
| 127-135 | 5 | 5 |
| 136-144 | 9 | 9 |
| 145-153 | 12 | 12 |
| 154-162 | 5 | 5 |
| 163-171 | 4 | 4 |
| 172-180 | 2 | 2 |
| | | Total 40 |

Histograma, Polígono de Frecuencia. Frecuencias Relativas y Acumuladas.

Histograma y Polígonos de Frecuencia: son dos representaciones gráficas de las distribuciones de frecuencia.

- El Histograma, consiste en una serie de rectángulos que tienen:

A) Sus bases sobre un eje horizontal (abscisa) con centros en las marcas de clase y longitud igual al tamaño de los intervalos de clase.

b) Superficies proporcionales a las frecuencias de clase. Si los rectángulos tienen su base de igual longitud por coincidir con iguales intervalos de clase, las alturas de los rectángulos son proporcionales a las frecuencias de clase y se acostumbra a tomar las alturas numéricamente iguales a las frecuencias de clase.

- El polígono de Frecuencias: es un gráfico de líneas trazado sobre las marcas de clase. Puede obtenerse uniendo los puntos medios de los techos de los rectángulos en el histograma.

El histograma y el polígono de frecuencia de la figura se construyeron con la distribución de frecuencias de las alturas de los estudiantes, cuya tabla de frecuencias hemos usado antes como ejemplo: se acostumbra a prolongar el polígono con PQ y RS hasta las marcas de clase inferior y superior inmediatas, que corresponderían a clases de frecuencia cero. En este caso la suma de las áreas de los rectángulos del histograma es igual al área total limitada por el polígono de frecuencias y el eje x.

- Distribuciones de Frecuencias Relativas.

La frecuencia relativa de una clase es la frecuencia de clase dividida por el total de frecuencias de todas las clases y se expresa generalmente como porcentaje.

Ejemplo: la frecuencia relativa de la clase 66-68 de la tabla de alturas es $42/100=42\%$. La suma de las frecuencias relativa de todas las clases es 100%.

Si las frecuencias de la tabla anterior se sustituyen por las correspondientes frecuencias relativas, la tabla resultante se llama Distribución de Frecuencias Relativas distribución porcentual o tabla de frecuencias relativas.

Estas representaciones de frecuencias relativas se obtiene directamente del histograma o del polígono de frecuencias con solo reemplazar la escala vertical de frecuencias a frecuencias relativas conservando el mismo diagrama.

Los gráficos que se obtienen se llaman histogramas de frecuencia relativas o histogramas porcentuales o polígonos de frecuencia relativa o polígono porcentual.

- Distribuciones de Frecuencias Acumuladas, Ojivas.

La frecuencia total de todos los valores menores que el límite real superior de una clase, de un intervalo de clase dado, se conoce como Frecuencia Acumulada, hasta ese intervalo de clase inclusive.

Ejemplo: En la tabla de alturas de estudiantes, la frecuencia acumulada del intervalo de clase 66–68 inclusive es $5+18+42=65$.

Esto significa que 65 estudiantes tienen alturas menores que 68,5 pulgadas.

Una tabla que representa las frecuencias acumuladas se llama Distribución de Frecuencias Acumuladas, tabla de frecuencias, acumuladas o distribución acumulada

Ejemplo: Veremos nuevamente el ejemplo de las alturas de los estudiantes.

| ALTURA (pulgadas) | NUMERO DE ESTUDIANTES |
|-------------------|-----------------------|
| Menor que 59,5 | 0 |
| 62,5 | 5 |
| 65,5 | 23 |
| 68,5 | 65 |
| 71,5 | 92 |
| 74,5 | 100 |

La curva de distribución de frecuencias acumuladas también se denomina Polígono de Frecuencias Acumuladas u Ojivas.

Las distribuciones de frecuencias acum. Pueden ser crecientes o del tipo Menor que o al inverso, pueden ser o más si se comienza desde el límite superior del rango.

También aquí se pueden usar los valores relativos.

- Curvas de Frecuencias – Ojivas suavizadas.

El conjunto de datos puede considerarse normalmente como perteneciente a una muestra extraída de una población grande. En ese caso el polígono de frecuencias tendrá los lados muy pequeños, que se aproximarán en conjunto a una curva. Las curvas de este tipo se llaman Curvas de Frecuencias o Curvas de Frecuencias Relativa, también se las denomina polígonos de frecuencia suavizada u ojivas suavizadas.

Tipos de curvas de frecuencia.

Las curvas de frecuencias presentan determinadas formas características que las distinguen como se indica en las siguientes figuras.

Simétrica sesgo a derecha

sesgo a izquierda forma de j

forma de j invertida forma de u

bimodal multimodal

Las curvas simétricas en la que el valor máximo equidista de los extremos se llama curva normal.

Ejemplo: según el histograma que representa a los 1546 estudiantes. Con los datos de la muestra se pide:

- Construir un polígono de frecuencias porcentual suavizada (curva de frecuencia).
- Construir la ojiva porcentual menor que suavizada.
- De los resultados de a y b, estimar el número de estudiantes de la universidad que tienen alturas entre 65 y 70 pulgadas. ¿Qué supuestos deberán hacerse?
- ¿Pueden los resultados utilizarse para estimar la proporción de estudiantes de los Estados Unidos que tengan alturas entre 65 y 70 pulgadas?

SOLUCIÓN

A y b– En las figuras siguientes, las líneas d trazos representan los polígonos de frecuencias y ojivas y se obtienen a partir de los datos dados en el histograma de la figura 5.

Es necesario construir el histograma de frecuencias relativas que en este caso es coincidente porque la muestra es de 100 estudiantes.

c– Si la muestra de 100 estudiantes es representativa de la población de los 1546 estudiantes, las curvas suavizadas de los apartados a y b pueden adoptarse como la curva de frecuencia porcentual y la ojiva, para esta población. Esta suposición es correcta solamente si la muestra se extrajo al azar, es decir, si todo estudiante tuvo la misma probabilidad de estar en la muestra.

Las alturas entre 65 y 70 pulgadas registradas con aproximación de una pulgada, realmente representan alturas entre 64,5 y 70,5 pulgadas, lo que se muestra en el área sombreada de la curva de frecuencia.

El porcentaje de estudiantes que tienen estas alturas puede sacarse dividiendo el área sombreada de la fig. 8 por el área total limitada por la curva suavizada y el eje de abscisas .

Sin embargo es más sencillo utilizar la fig. 9 que nos suscriba lo siguiente:

- porcentaje de estudiantes con alturas menores que 70,5 pulgadas: 82 %
- porcentaje de estudiantes con alturas menores que 64,5 pulgadas: 18 %

De modo que el porcentaje de estudiantes cuyas alturas están comprendidas entre 64,5 y 70,5 será la diferencia de ambos porcentajes: $82\% - 18\% = 64\%$

Luego el número de estudiantes de la universidad con alturas entre 65 y 70 pulgadas es 64%. Entonces para la población de 1546 estudiantes son 989 los que están entre los límites de alturas indicadas con una precisión de pilgadas.

Otra forma de expresar esto es que la probabilidad de que una persona elegida al azar entre los 1546 estudiantes tenga una altura entre 65 y 70 pulgadas es de 64%; a causa de su afinidad con la probabilidad, que veremos en capítulos posteriores, las curvas de frecuencia relativa se conocen también como curvas de probabilidad o distribución de probabilidad.

d– Se podría considerar la proporción encontrada como válida si se tuviera la certeza de que la muestra de 100 estudiantes es al azar respecto del total de los estudiantes universitarios de los Estados Unidos. Sin embargo, esto es improbable por varias razones: 1– Algunos pueden no haber alcanzado su altura máxima. 2– La

generación más joven puede tender a ser más alta que sus padres.

2.3- Distintos gráficos para el análisis de datos. Utilización de software aplicado.

En las hojas siguientes se observan distintos tipos de gráficos extraídos de documentación referida a la industria aeronáutica, como ejemplo:

3.1 Medidas de Posición para Datos Agrupados y sin Agrupar. Medias Aritméticas simples y ponderadas. Media Geométrica y Armónica. Mediana y Modo. Cuartiles y Deciles.

Un promedio es un valor típico de un conjunto de datos. Como tales valores suelen situarse en el centro del conjunto de datos ordenados por su magnitud, los promedios se conocen también como medidas de centralización.

3.1.1- Media Aritmética

La media aritmética, o media, es un conjunto de N números $x_1, x_2, x_3, \dots, x_n$ se representa por X y se define como:

Ejemplo: La media aritmética de los números 6, 3, 7, 10, 15 es:

Si los números $x_1, x_2, \dots, x_k$ aparecen $f_1, f_2, \dots, f_k$ veces cada uno la media aritmética será:

Dado que $N = \sum f_i$, o sea la frecuencia total.

Ejemplo: Si 6, 3, 7, 10, 15 aparecen con frecuencias 5, 4, 3, 2 y 1 respectivamente la media será:

3.1.2- Media Aritmética Ponderada

En ciertas circunstancias se asocian a los números $x_1, x_2, \dots, x_n$ ciertos factores o pesos $w_1, w_2, \dots, w_k$ que normalmente indican la importancia relativa de cada uno de los números.

En este caso la media será:

Y se denomina media ponderada, se observa una notable similitud con la media de números cuando estos se repiten.

Ejemplo: La evaluación de los alumnos del curso serán dos exámenes parciales y un examen final cuyo valor es equivalente a tres veces los exámenes parciales, siendo las calificaciones 7, 6 y 6 respectivamente, luego:

Propiedades de la Media Aritmética

a)- La suma algebraica de las desviaciones de un conjunto respecto de su media aritmética es cero.

b)- La suma de los cuadrados de las desviaciones de un conjunto de números x_i , respecto de cualquier número a es mínima solo si $a = X$.

c)- Si f_1 números tienen por media a_1 , f_2 números tienen por media $a_2, \dots, f_k$ números tienen por media a_k , entonces la media de todos los números es:

o sea una media ponderada de todas las medias.

d)– Si A es cualquier supuesta media aritmética y si $d_i = x_i - A$ son las desviaciones de x_i respecto de A, las ecuaciones 3.1 y 3.2 se convierten en:

Donde.....

Las ecuaciones 5 y 6 se pueden resumir como

Utilizar el método clave para hallar la media aritmética de las alturas de los 100 estudiantes de la Universidad.

Solución:

| Marca de Clase | Desviaciones $d = x - A$ | Frecuencias | $f \cdot d$ |
|----------------|--------------------------|-------------|-------------|
| 61 | -6 | 5 | -30 |
| 64 | -3 | 18 | -54 |
| 67 | 0 | 42 | 0 |
| 70 | 3 | 27 | 81 |
| 73 | 6 | 8 | 48 |
| | | $= f = 100$ | $fd = 45$ |

21– Indicando por $d_j = x_j - A$, las desviaciones de cualquier marca de clase x_i respecto de una marca de clase A en una distribución de frecuencias. Demostrar c, si tiene a) las desviaciones son todas multiples de c, es decir $d_j = cn_j$ donde $n_j = 0, \pm 1, \pm 2, \dots$

- La media aritmética puede ser obtenida de la fórmula
- De la tabla vemos que las desviaciones son multiples de $c=3$ (el tamaño del intervalo). Entonces dos marcas de clase cualquiera difieren en:

Que es multiplo de c.

- Hemos visto que las desviaciones de todas las marcas de clase, respecto de una dada son multiples de c. Entonces:

Esto es equivalente a que se saca de sin más que sustituir de dos por cn.

Media Aritmética calculada a partir de datos agrupados.

Cuando los datos se presentan mediante una distribución de frecuencia, todos los valores caen dentro de los intervalos de clase dados.

Las expresiones 3 (-2) y (3,6) son válidas para tales datos agrupados, si se forma x_i como la marca de clase y f_i su correspondiente frecuencia de clase, A una marca de clase cualquiera y $d_i = x_i - A$. Las desviaciones de x_i respecto de A.

Si todos los intervalos de clase tienen igual tamaño c, las desviaciones $d_i = x_i - A$, pueden expresarse como $c u_i$, donde u_i puede ser un número entero positivo, negativo o cero, luego la fórmula 3.6 queda

La que resulta equivalente a la ecuación . Este se llama método clave para el cálculo de la media y por ser muy corto y práctico debería usarse siempre.

3.1.2– Mediana

La mediana de una colección de datos ordenados en orden de magnitud es el valor medio o la media aritmética de los dos valores medios.

Ejemplo1: Dados los números 3, 4, 4, 1, 6, 8, 8, 8, 10, que tienen mediana 6.

Ejemplo2: Dados los números 5, 5, 7, 9, 11, 12, 15, 18 ($\frac{1}{2}(9+11)=10$) geométricamente, la mediana es la abscisa que corresponde a la vertical que divide el histograma en dos partes de igual área.

Para los datos agrupados la mediana se obtiene por interpolación y viene dada por:

Donde: L1: límite real inferior de la clase mediana, es decir la que contiene la mediana.

M: Número total de datos (frecuencia total).

(f)1: la suma de las frecuencias de todas las clases por debajo de la clase mediana.

Feudiana: frecuencia de la clase mediana, tamaño del intervalo de clase mediana.

Moda

Definimos la moda como la abscisa X del punto de intersección P de las líneas QS y RT, sean $x=L1$ y $X=U1$ los límites reales inferior y superior de la clase modal y A1 y A2 representan el exceso de frecuencia de la clase modal sobre las contiguas a ella.

PQR y PST son triángulos semejantes y ahí tenemos:

Punto que $U1=L1+c$ (tamaño del intervalo de clase)

La moda de una serie de números es aquel valor que se presenta su mayor frecuencia. La moda puede no existir y si existe puede no ser única.

Ejemplo1: El conjunto 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 tiene por moda 9.

Ejemplo2: El conjunto 2, 3, 4, 4, 4, 3, 5, 7, 7, 7, 9 tiene dos modas 4 y 7 y se llama bimodal

En el caso de datos agrupados donde se ha construido una curva de f para ajustar los datos, la moda será el valor de x correspondiente al máximo de la curva. A veces se representa como X.

Dada una distribución de frecuencias a un histograma la moda puede sacarse de la fórmula:

Donde: L1: límite real inferior de la clase de la clase modal.

A1: exceso de la frecuencia modal sobre la frecuencia de clase contigua inferior.

A2: exceso de la frecuencia de clase modal sobre la frecuencia de clase superior.

C: tamaño de la clase modal

Relación Empírica entre Media, Mediana y Moda

Para curvas de frecuencias, unimodales, que sean levemente sesgadas se cumple la siguiente relación empírica: $Media - Moda = 3 (Mediana - Moda)$.

Para curvas simétricas Media, Mediana y Moda, coinciden.

Media Geométrica (G)

La media geométrica de una serie N de números $x_1, x_2, \dots, x_n$ es la raíz N – suma del producto de los números.

Ejemplo: Dados 2, 4, 8 la media geométrica es

En la práctica G se calcula por logaritmos.

Cuando los datos están agrupados, es decir cuando los números $x_1, x_2, \dots, x_n$ se presentan con frecuencia $f_1, f_2, \dots, f_k$ donde $f_1 + f_2 + \dots + f_k = N$ es la frecuencia total.

- Hallar la media geométrica de los números. (G)
- Obtener una expresión para $\log G$
- Utilizar los datos hallados para encontrar la media geométrica de los datos agrupados en distribución de frecuencia.

a)

G calculada de esta manera se llama media geométrica ponderada.

b) Tomando logaritmo a la expresión anterior tenemos

para que la expresión sea válida los números deben ser positivos, pues de otro modo los logaritmos no están definidos.

Se observa además de la expresión hallada que el logaritmo de la media geométrica de una serie de números positivos la media aritmética de los logaritmos de los números.

- Esta expresión nos permite hallar la media geométrica para datos agrupados, si $x_1, x_2, \dots, x_n$ son marcas de clase y $f_1, f_2, \dots, f_k$ las frecuencias de clase.

3.1.6– Media Armónica (H)

La media armónica H de una serie de N números $x_1, x_2, \dots, x_n$ es la recíproca de la media aritmética de los recíprocos de los números.

En su expresión matemática es:

Ejemplo: La media armónica de los números 2, 4, 8 es $H = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{7/8} = 3,43$

En la práctica es más fácil recordar que

Para el caso de datos agrupados, el cálculo es análogo al realizado con la media geométrica.

3.1.7– Relación entre las Medias Aritméticas, Geométricas y Armónica.

La media geométrica de una serie de números positivos $x_1, x_2, \dots, x_n$ es menor o igual que su media aritmética, pero mayor o igual que la media armónica.

La igualdad se da únicamente cuando $x_1, x_2, \dots, x_n$ son idénticos.

3.1.8 Cuartiles y Deciles

Hemos visto la definición de mediana que decía que si una serie de datos se colocan en orden según su magnitud, el valor medio (o la media aritmética de los dos valores medios), que divide al conjunto de datos en dos partes iguales es la mediana. Extrafolando este concepto, podemos pensar en los valores que dividen a los datos en cuatro partes iguales. Estos valores representados por Q1, Q2, Q3, Q4 se llaman primero, segundo, tercero y cuarto, cuartil respectivamente. De ellos el Q2 es la mediana.

Analogamente los valores que dividen los datos en diez partes iguales, se denominan deciles y se representan por D1, D2, D5.....Dq donde el D5 es la mediana.

Percentiles: dividiendo los datos en 100 partes iguales.

Ejemplo: Dados los salarios de 65 empleados de la cía. P y R como se muestra en la distribución de frecuencias dadas.

Se pide hallar: a) Los cuartiles Q1, Q2, Q3

b) Los deciles D1, D2, D3..Dq

| SALARIOS \$ | NÚMERO DE EMPLEADOS |
|-----------------|---------------------|
| 50,00 – 59,99 | 8 |
| 60,00 – 69,99 | 10 |
| 70,00 – 79,99 | 16 |
| 80,00 – 89,99 | 14 |
| 90,00 – 99,99 | 10 |
| 100,00 – 109,99 | 5 |
| 110,00 – 119,99 | 2 |
| TOTAL | 65 |

Solución: a) El primer cuartil Q1 es el valor que se obtiene para $N/4=65/4=16,25$ de los casos, comenzando de la clase primera (lo más baja).

La primera clase tiene 8 casos y la segunda 10, entonces de la segunda tomamos 8,25 casos y por el método de interpolación lineal tenemos: $Q1 = 59,99 + 8,25/10(10,00) = \$ 68,25$

El segundo cuartil Q2 es aquel valor que se obtiene para $2N/4 = N/2 = 65/2 = 32,5$ de los casos.

Puesto que las dos primeras clases comprenden 18 casos tomaremos $32,5-18=14,5$ casos de la tercera clase, entonces:

$Q2 = 69,99 + 14,5/16 (10,00) = 79,06$ que es además la mediana.

El tercer cuartil Q3 es el valor que se define para $3N/4$ de los casos o sea $3(65/4) = 48,75$, puesto que las cuatro primeras.

4.1 Conceptos básicos. Espacio Muestral y Eventos. Definiciones. Probabilidad de eventos compuestos. Reglas de Adición y Multiplicación.

Hemos visto al comienzo de este curso que el objetivo de la Estadística es hacer inferencia acerca de una

población basándose en la información proporcionada por una muestra. Como la muestra sólo da una información parcial de la población, se necesita un mecanismo que permita alcanzar el objetivo, ese mecanismo es la probabilidad.

4.1.1– Espacio Muestral y Eventos

Los datos obtenidos al observar eventos, son la base de todo el análisis posterior. Al trabajo de recolección de datos se lo denomina genéricamente experimento.

Ejemplo: 1– Registrar el ingreso anual de los trabajadores. 2– Registrar el valor de una acción de la bolsa.

De hecho, cada experimento produce uno o varios resultados posibles que se llaman eventos y se denotan con letras mayúsculas.

Se pueden distinguir eventos simples y eventos compuestos, un evento que no puede ser descompuesto se llama evento simple. Si cada evento simple se le asocia un punto llamado punto muestral, al conjunto de todos los puntos muestrales de un experimento se lo denomina espacio muestral, y se representa por la letra S; se dice entonces que S es la totalidad de los puntos muestrales.

Esta situación puede representarse gráficamente en un diagrama de Venn.

4.1.2– Se observa que un evento con términos de puntos muestrales es una colección de esos puntos muestrales.

¿Cómo se decide si un punto muestral se incluye o no en un evento?

Simplemente se verifica si la ocurrencia del punto muestral implica la ocurrencia del evento.

Definición: un evento es una colección particular de puntos muestrales. Desde un punto de vista práctico, se piensa en la fracción de veces que resulta el evento A, como la probabilidad de A.

Dicho de otra manera, si un experimento se repite un número grande de veces (N) y de estas el evento A ocurre N_a veces, la probabilidad de A es: $P(a) = \frac{N_a}{N}$

N

Esta es la forma práctica de ver la probabilidad como frecuencia relativa.

Desde el punto de vista matemático, se toman las probabilidades como dadas produciéndose así un modelo para la población.

Ejemplo: Se supone un número grande de lanzamientos de un dado dónde las probabilidades de cada una de sus caras es igual: $P(E_1) = P(E_2) = \dots = P(E_6) = 1/6$

Esto supone el dado perfectamente balanceado, existe un dado perfectamente balanceado? Probablemente no, pero para los propósitos prácticos, si el desbalance no es significativo, las probabilidades de los puntos muestrales estarán muy cerca de $1/6$, y entonces el modelo es válido.

Para completar la descripción del modelo se le agregan las siguientes condiciones:

A cada punto del espacio muestral se le asigna un número llamado la probabilidad de E_i , que se indica como $P(E_i)$ tal que:

- $C < 0 = P(E_i) < 0 =$ para cada i
- $E P(E_i) = 1$ desde E es la suma de todos los puntos del espacio muestral.

Siendo un evento determinado una colección particular en puntos muestrales, vale lo siguiente definición: La probabilidad de un evento A , es igual a la suma de las probabilidades de los puntos muestrales de A

El método es claro y fácil de aplicar, sin embargo se forma lento y tedioso cuando el espacio muestral es grande.

Veremos a continuación algunas propiedades que permiten agilizar el cálculo.

4.1.3– Probabilidad de Eventos Compuestos

Como su nombre lo indica los eventos compuestos están formados por la composición de dos o más eventos.

Con la ayuda de la teoría de conjuntos veremos que la composición puede hacerse de dos formas, una unión o una intersección.

Unión: Si A y B son dos eventos en el espacio muestral S , la unión de A y B es el evento que contiene a todos los puntos muestrales de A o de B o de ambos. Se indica $A \cup B$.

Gráficamente se representa por el área sombreada en el diagrama de Venn.

Intersección: si A y B son los datos en el espacio muestral S , la intersección de A y B es el evento que contiene a los puntos muestrales que están en ambos A y B , se representa como AB o también.

Ejemplo: Volveremos sobre el ejemplo del lanzamiento de dos monedas, referimos los eventos así.

Evento A = al menos una cara Evento B = al menos una cruz

$E_1 = CC$ $E_3 = XC$

$E_2 = CX$ $E_4 = XX$

Ejemplo1: Calcular la probabilidad de observar un número impar en el lanzamiento de un dado.

El evento A número impar incluye los puntos muestrales E_1, E_3, E_5 luego:

$$P(A) = P(E_1) + P(E_3) + P(E_5) = 1/6 + 1/6 + 1/6 = 1/2$$

Ejemplo2: Calcular la probabilidad de observar una cara en el lanzamiento de dos monedas. Se construye el espacio muestral indicando por C y X cada cara o cruz respectivamente.

| Evento | 1ra. Moneda | 2da. Moneda | $P(E_i)$ |
|--------|-------------|-------------|----------|
| E_1 | C | C | $1/4$ |
| E_2 | C | X | $1/4$ |
| E_3 | X | C | $1/4$ |
| E_4 | X | X | $1/4$ |

El evento muestral A , incluye los puntos muestrales E_2 y E_3 únicamente, luego.

$$P(A) = P(E2) + P(E3) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Procedimiento de cálculo de la Probabilidad de un Evento.

- Definir el experimento
- Hacer un listado de los eventos simples y verificarlos. Esto define el espacio muestral.
- Asignar probabilidades a los puntos muestrales de S, verificando que $\sum P(E_i) = 1$
- Definir el evento A, como una colección particular de puntos muestrales.
- Encontrar $P(A)$ sumando las probabilidades de los puntos de a.

Evento A = E1, E2 y E3

Evento B = E2, E3 y E4

Evento AB = E1, E2, E3 E4 = S

4.1.4– Relaciones entre Eventos

Definimos tres relaciones entre eventos. Eventos complementarios, eventos independientes y eventos mutuamente excluyentes.

Complementarios

El complementario de un evento A, es la colección de todos aquellos puntos muestrales, que están en el espacio muestral S pero no están en A.

El complemento de A se denomina A' , luego sabiendo que $\sum P(E_i) = 1$ tenemos que $P(A) + P(A') = 1$, relación muy útil para encontrar $P(A')$ si se conoce $P(A)$

Independientes

Dos eventos se denominan independientes si la ocurrencia o no ocurrencia de uno de ellos, no cambia la probabilidad de ocurrencia del otro.

En el ejemplo de las monedas cada lanzamiento es independiente de los otros E1, E2, E3, E4

Mutuamente Excluyentes

Dos eventos se dicen mutuamente excluyentes si la ocurrencia de uno de ellos implica la no ocurrencia del otro.

Expresado de otro modo, dos eventos A y B son mutuamente excluyente si AB no contienen ningún punto muestral.

Ejemplo: En el lanzamiento de una moneda el evento cara implica que el evento cruz no puede ocurrir.

4.1.5– Probabilidad Condicional

Con frecuencia dos eventos se encuentran relacionados de manera tal que la ocurrencia de uno de ellos depende de si otro a ocurrido o no.

Este tema esta ligado con el hecho de que hay sucesos dependientes de otros. Es decir se trata de saber cual es

la probabilidad de que ocurra el evento B, si ocurrió el evento A. Esta probabilidad se denomina condicional de B dado que A ha sucedido y se indica $P(B|A)$.

Si los eventos son independientes entonces $P(A,B) = P(A)$, los eventos son dependientes entonces se aplica la siguiente definición. La probabilidad condicional de B dado que A ha ocurrido es, $P(B|A) = P(AB) / P(A)$

La probabilidad condicional de A dado que B ha ocurrido es, $P(A|B) = P(AB) / P(B)$

Ejemplo: Calcular $P(A|B)$ para el experimento de lanzar un dado, definiendo los siguientes eventos:

Evento A= se observa un número impar E1, E3, E5

Evento B= se observa un número menor que 4 (E1, E2, E3)

Dado que B ha ocurrido, sólo interesan los puntos muestrales E1, E2, y E3 que ocurren en la misma frecuencia, de estos E1 y E3 implican al evento A.

Luego $P(A|B) = P(AB) / P(B) = 1/3 / 1/2 = 2/3$

4.1.6- Ley de Mutiplicación de la Probabilidad

Dados dos eventos A y B, la probabilidad de intersección AB es:

$$P(AB) = P(A) P(A|B) = P(B) P(B|A)$$

Si A y B son independientes, $P(AB) = P(A) P(B)$

Esta ley proporciona una fórmula para calcular la probabilidad de la intersección.

En general, si se tienen n eventos independientes A, B, C, ..., H, cuyas probabilidades son $P(A), P(B), \dots, P(H)$, entonces la probabilidad de ocurrencia de los n eventos es, $P(A) P(B) P(C) \dots P(H)$.

Ejemplo1: Si la probabilidad de que A viva 20 años es 0,7 y la probabilidad de que ambos vivan 20 años es: $P(AB) = P(A) P(B) = 0,7 \cdot 0,5 = 0,35$

Ejemplo2: Supongamos una caja que contenga 3 bolas negras y 2 bolas blancas.

Evento A: extraer una bola negra

Evento B: extraer una bola negra, en segundo término.

Sin reemplazamiento A y B son eventos dependientes

$P(A) = 3/5$ Es la probabilidad de que la primera bola sea negra

$P(B|A) = 2/4 = 1/2$ Es la probabilidad de que la segunda bola sea negra, dado que la primera fue negra, luego la probabilidad de que ambas sean negras:

$$P(AB) = P(A) P(B|A) = 3/5 \cdot 1/2 = 3/10$$

4.1.7- Ley de adición de la Probabilidad

La probabilidad de la unión AUB es $P(A \cup B) = P(A) + P(B) - P(AB)$

En el caso en que A y B sean eventos mutuamente excluyentes $P(A \cup B) = P(A) + P(B)$ ya que $P(AB) = 0$

Esta ley consiste en el modelo. La suma de $P(A) + P(B)$ considera la suma de todos los puntos muestrales de AUB pero cuenta dos veces los puntos muestrales de la intersección AB, por ello al restar $P(AB)$ se obtiene el resultado correcto.

Generalizando esto si A, B, ..., H son eventos mutuamente excluyentes, con probabilidad $P(A), P(B), \dots, P(H)$, entonces la probabilidad de ocurrencia de A o B o H es la suma de las probabilidades $P(A) + P(B) + \dots + P(H)$

Ejemplo1: Si A es el evento extraer un as de la baraja de cartas, y B es el evento extraer un rey.

$P(A) = 4/52 = 1/13$ y $P(B) = 4/52 = 1/13$

Luego la probabilidad de extraer un as o un rey es una extracción

$P(A+B) = P(A) + P(B) = 1/13 + 1/13 = 2/13$ puesto que ambos no pueden extraerse en una sola extracción. Son eventos mutuamente excluyentes.

- MUESTREO

6.1- Teoría del Muestreo

La teoría del muestreo permite establecer relaciones existentes entre una población y las muestras extraídas de la misma.

Por ejemplo, permite estimar cantidades desconocidas de la población, frecuentemente llamados parámetros poblacionales, a partir del conocimiento de las correspondientes cantidades muestrales (media muestral, varianza) y viceversa, conocidos los datos de la población la teoría del muestreo es también útil para determinar si las diferencias que se pueden observar entre dos muestras son debidas a causas aleatorias o fortuitas o si son atribuibles a alguna causa.

Este tipo de aplicaciones se denomina ensayos de hipótesis de significación y tienen gran importancia en la teoría de la decisión.

Ejemplo1: Con el propósito de decidir si se aceptan los lotes de mercaderías que envía una determinada fábrica, se lleva a cabo un procedimiento que consiste en seleccionar 10 artículos al azar de cada lote determinar el número de defectuosos.

Un lote se rechazará siempre que se encuentren 2 o más artículos defectuosos de los 10 seleccionados.

Se supone que el número de artículos de cada lote es grande y que cada lote contiene 5% de defectuosos.

¿Cuál es la probabilidad de aceptar los lotes 2 y cuál la de rechazarlo?

En este ejemplo el objetivo consiste en eliminar aquellos lotes que tienen una proporción alta de artículos defectuosos.

Naturalmente el comprador no está interesado en pagar artículos defectuosos y solo acepta aquellos lotes que no contienen una proporción mayor de $p = 0,05$ unidades defectuosas.

Esto es lo mismo que decir un buen lote se define como aquel para el cual la proporción de artículos defectuosos p es menor que 0,05 y un lote en el que $p > 0,05$ es inaceptable.

Se trata de detectar los lotes inaceptables y devolverlos.

Una forma de detectar lotes inaceptables es examinar cada uno de los artículos. Este procedimiento aplicado a lotes de 1000 cojinetes es muy costoso.

Un procedimiento menos costoso y más efectivo consiste en aceptar o rechazar un lote sobre la base de una muestra.

Es decir se obtiene una muestra de N artículos de un lote, se inspeccionan sus elementos y se registra el número Y de defectuosos.

El lote se acepta si el número de defectuosos es menor o igual que el número a seleccionado de antemano.

Si Y excede a a , el lote se rechaza y se devuelve.

Todo el plan de muestreo está definido por el tamaño de la muestra n y el número a .

El cambio de cualquiera de esos números modifica las características del plan, que se miden en términos de la probabilidad de aceptar o rechazar un lote que contenga una determinada proporción de artículos defectuosos como se muestra en el ejemplo dado.

La bondad del plan de muestreo se caracteriza por el cálculo de la probabilidad de aceptar para diferentes valores de la proporción de defectuosos.

El resultado se llama Curva característica de Operación del plan de muestreo.

Para que el funcionamiento del método sea correcto es deseable que la probabilidad de aceptar lotes con baja proporción de defectuosos sea alta y disminuya con el incremento de estos.

6.2- Muestreo Probabilístico y No Probabilístico.

Para que las conclusiones de la teoría del muestreo e inferencia estadística sean válidas, las muestras deben elegirse de forma que sean representativas de la población.

El estudio sobre los métodos de muestreo y los problemas asociados, se conocen como diseño de experimentos.

El proceso mediante el cual se extrae una muestra representativa de una población se llama muestreo al azar, y de acuerdo con ello, cada miembro de la población tiene la misma posibilidad de ser elegido (incluido) en la muestra.

Una forma típica de aplicar muestreo al azar es usar números aleatorios dados en una tabla.

Si se extrae un número de una curva, se puede volver o no el número a una curva antes de realizar una segunda extracción. En el primer caso, un mismo número puede salir varias veces, mientras que en el segundo un número determinado solo puede salir una vez.

Estos dos casos diferentes se conocen como Muestreo con reemplazamiento y sin reemplazamiento.

Por otra parte las poblaciones pueden ser finitas o infinitas.

Por ejemplo: si se extraen sucesivamente 10 bolas sin remplazamiento de una urna que contiene 100 u, esta formando una muestra de una población finita, mientras que si se lanza una moneda al aire 50 veces anotándose el número de caras, se esta mostrando una población infinita.

Una población finita en la que se realiza muestreo con remplazamiento puede considerarse teóricamente como una población infinita, puesto que puede extraerse cualquier número de muestras son agotar la población. Además hay mucho casos de poblaciones finitas de gran número que pueden considerarse como infinitas.

6.3- Distribuciones Muestrales

Si se consideran las posibles muestras de tamaño N que pueden extraerse de una población dada.

Para cada muestra se puede calcular una estadística, tal como la media la desviación standard, etc, que será de distinto valor de una muestra a otra.

Ejemplo: Calcule la probabilidad de aceptar un lote para un plan de muestreo en el que $n=6$ y $a=0$, para proporciones de defectuosos $p=1, 3$ y 5 .

Haga un bosquejo de la curva.

Solución. Para captar el lote la muestra no debe contener defectuosos, luego

En estos valores y dado que la probabilidad de captar debe ser 1 si $p=0$ y debe ser 0 si $p=1$, puede obtenerse un bosquejo de la curva característica de operación.

Un plan de muestreo puede verse como una red o malla de filtro, los embarques y su curva característica describe en cierta forma el tamaño de los agujeros.

Un incremento del tamaño de la muestra produce una mayor información y aumenta el poder desiniminatorio del proceso de decisión.

El procedimiento así descrito es un ejemplo de inferencia estadística, ya que se toman decisiones en relación con la proporción de artículos defectuosos p , de un lote.

La curva característica de operación del plan de muestreo es una medida de la bondad de este procedimiento.

De esta forma se obtiene una distribución del parámetro estadístico que se conoce como distribución muestra.

A su vez para cada distribución se puede calcular la media, la desviación standard, etc. Se habla entonces de la media y la desviación standard de la distribución muestral de medias, etc.

Distribución Muestral de Medias

Supóngase que de una población finita se extrae todas las posibles muestras sin remplazamiento de tamaño N , siendo N_p el tamaño de la población y $N_p > N$.

Si llamamos a la media y la desviación standard de la distribución muestral de medias, siendo la media y la desviación estándar de la población

Respectivamente tendremos.

(1)

Si la población es infinita o el muestreo es con reemplazamiento se tiene:

(2)

Distribución muestral Muestral de Proporciones

Supóngase una población infinita y que la probabilidad de ocurrencia de un evento es p , mientras que la probabilidad de que no ocurra es $q=1-p$.

Por ejemplo: la población puede ser todos los posibles lanzamientos de una moneda en los que la probabilidad del evento cara es $p=1/2$. Se consideran todas las posibles muestras de tamaño N extraídas de la población y para cada muestra se determina la proporción p de éxito. En el ejemplo de la moneda p sería la proporción de caras aparecidas en los N lanzamientos.

Entonces se define una distribución muestral de proporciones cuya media y desviaciones standard vienen dadas por:

(3)

Estas expresiones pueden obtenerse de las (1) sustituyendo

Se nota en este caso que la población se distribuye binomialmente y las ecuaciones (3) se obtienen dividiendo por N , la media y la desviación standard de la binomial.

Distribución muestral de Diferencias y Sumas

Supóngase que se tiene dos poblaciones, para cada muestra de tamaño N_1 extraída de la primera población se calcula el parámetro estadístico S_1 , esto proporciona una distribución muestral cuya media y desviación standard vienen dadas por respectivamente. Análogamente para cada muestra de tamaño N_2 extraída de la segunda población se calcula el parámetro estadístico S_2 cuya media y desviación standard son . De todas las posibles combinaciones de estas muestras de las dos poblaciones se pueden obtener una distribución de las diferencias S_1-S_2 que se conoce como distribución muestral de diferencias de los parámetros estadísticos.

La media y la varianza de esta distribución se indican por respectivamente y su expresión es:

(4)

Siempre que las muestras sean independientes.

Si S_1 y S_2 son las medias muestrales de las dos poblaciones que se expresan

Entonces la distribución muestral de las diferencias de medias para poblaciones infinitas con medias y desviaciones standard respectivamente se expresan como sigue:

(5)

que esta basado en el conjunto de ecuaciones (2).

Estas ecuaciones son válidas para poblaciones finitas sin el muestreo es con reemplazamiento.

Para el caso de poblaciones finitas en las que el muestreo se realiza sin reemplazamiento, se obtienen resultados similares partiendo de las ecuaciones (1).

Si el caso se trata de diferencias de distribuciones muestrales de proporciones de dos poblaciones con distribución binomial y parámetros p_1, q_1 y p_2, q_2 respectivamente. Entonces S_1 y S_2 corresponden a las proporciones de éxito p_1 y p_2 y las ecuaciones (4) dan los resultados siguientes:

A veces es útil considerar la distribución muestral de la suma de estadísticas. La media y la desviación standard de esta distribución vienen dadas por:

Todo lo anterior supone que las muestras son independientes.

6.4- Tamaño de la Muestra

Hemos observado al considerar los distintos tipos de distribuciones muestrales que algunas expresiones son diferentes para tamaños diferentes de muestras.

Por ejemplo: para la distribución muestral de medias, cuando o sea valores considerados grandes, la distribución muestral de medias se aproxima a una distribución normal con media y desviación standard independientemente de la población de que se trate (siempre que la media y la varianza poblacional sean finitas y el tamaño de la población sea al menos dos veces el tamaño de la muestra).

En el caso que la población se distribuya normalmente la distribución muestral de medias se distribuye también normalmente, incluso para pequeños valores de N ($N \geq 30$).

En el caso de la distribución muestral de proporciones, cuya media y desviación vienen dadas por las ecuaciones vistas para grandes valores de

La distribución muestral se aproxima mucho a una distribución normal. La población se distribuye binomialmente.

6.5- Teorema Central Límite

Enunciado: si se extraen muestras de tamaño n de una población con media finita y desviación σ , entonces si es grande, la media muestral tiene una distribución aproximadamente normal con media μ y desviación $\sigma/\sqrt{n}$. La aproximación es mayor a medida que n crece.

Es decir que para el caso de :

El teorema central del límite, establece que a medida que n crece, la distribución de la suma de las mediciones muestrales tienden a una distribución normal, con media μ y desviación standard $\sigma/\sqrt{n}$.

Ambas distribuciones tienen la misma media y la desviación standard de la distribución de medias muestrales es igual a la de la población dividida por $\sqrt{n}$. Por lo tanto a medida que n crece, la dispersión de la distribución de las medias muestrales se hace considerablemente menor que la dispersión de la población.

La importancia del teorema, abarca dos aspectos: Primero, explica porque algunas mediciones tienen una distribución aproximadamente normal, como por ejemplo la estatura de un individuo o la posibilidad de una pieza defectuosa en una producción, siempre que en ambos casos el fenómeno dependa de variables aleatorias.

La segunda y más importante contribución del teorema central del límite radica en la Inferencia Estadística.

Muchos de los estimadores que se usan para hacer inferencias cerca de los parámetros de la población son sumas o promedios de mediciones muestrales.

Cuando el tamaño de la muestra es suficientemente grande se espera que el estimador tenga una distribución aproximadamente normal, en un muestreo repetitivo de acuerdo con el teorema central del límite.

Con respecto al tamaño de la muestra n no existe una solución general y el valor adecuado de n dependerá de la distribución de probabilidad de la población y del uso que se le quiera dar a la aproximación.

6-Error – Standard

La desviación standard de la distribución muestral de un parámetro estadístico

Se conoce también como su error standard = desvío standard.

En la tabla de la hoja siguiente se han resumido los errores standard de distribuciones muestrales para diversos estadísticos en condiciones de muestreo aleatorio sin reemplazamiento para una población infinita o con reemplazamiento para una población finita. Se agregan notas, que indican las condiciones para las cuales los resultados son válidos, así como otros comentarios de interés.

Es de notar que si el tamaño de la muestra N es bastante grande, las distribuciones muestrales son normales o casi normales, por eso: Los resultados descriptos se conocen como Métodos para Grandes Muestras.

Cuando $N > 30$ las muestras se llaman pequeñas y se aplican otros procedimientos de cálculo.

Lo mismo que en el problema 1 para el caso de muestreo sin reemplazamiento.

Solución

a) y b) como en el problema 1

c) Hay muestras de tamaño dos que pueden extraerse sin reemplazamiento de la población, estas son: (2,3); (2,6); (2,8); (2,11); (3,6); (3,8); (3,11); (6,8); (6,11); (8,11)

Las correspondientes medias muestrales son: 2,5; 4,0; 5,0; 6,5; 4,5; 5,5; 7,0; 8,5; 9,5 y la media de la distribución muestral de medias es:

Se observa que ya que vemos que el segundo miembro es $10,8/2 = (5 - 2/5 - 1) = 4,05$ que es el mismo valor obtenido anteriormente.

Quinientos cojinetes de bolas tienen un peso medio de 5,02 onzas y una desviación típica de 0,30 onzas.

Hallar la probabilidad de que una muestra al azar de 100 cojinetes elegidos entre este grupo tenga un peso total.

- comprendido entre 496 y 500 onzas
- de más de 510 onzas

Solución

Para la distribución muestral de medias

- El peso total se encontrará entre 496 y 500 onzas si la media de los 100 cojinetes se encuentre entre 4,96 y 5,0 onzas.
- Hay que calcular cuanto es 4.96 en unidades standard lo que se ha llamado Z.

4,96 ____

5,00 ____

La probabilidad pedida es igual al área bajo la curva entre los valores de Z hallados.

- Ejercicios como ejemplos

1– Una población se compone de los cinco números 2, 3, 6, 8, 11. Considerar todas las muestras posibles de tamaño dos que puedan extraerse con reemplazamiento de esa población.

Hallar a) La media de la población

b) La desviación standard de la población

- La media de la distribución muestral de medias
- La desviación standard de la distribución muestral de medias, es decir el error standard de medias.

Solución

- Por el primer principio de conteo hay $5 \times 5 = 25$ muestras de tamaño dos, que pueden extraerse con reemplazamiento.

(2,2) (2,3) (2,6) (2,8) (2,11)

(3,2) (3,3) (3,6) (3,8) (3,11)

(6,2) (6,3) (6,6) (6,8) (6,11)

(11,2) (11,3) (11,6) (11,8) (11,11)

Las correspondientes medias muestrales son:

(1)

2,0 2,5 4,0 5,0 6,5

2,5 3,0 4,5 5,5 7,0

4,0 4,5 6,0 7,0 8,5

6,0 5,5 7,0 8,0 9,5

6,5 7,0 8,5 9,5 11,0

y la media de la distribución muestral de medias es

suma de medias muestrales

comprobamos que

d) La varianza de la distribución muestral de media se obtiene restando el valor de la media 6 de cada número de (1), elevando al cuadrado cada diferencia, sumando los 25 números así obtenidos y dividiendo por 25. El resultado final es:

luego

observamos que

b) Cuánto vale la probabilidad pedida?

En unidades Z tenemos $(0,02 + 0,00125) = \underline{0,02 + 0,00125 - 0,02}$

0,007

La probabilidad pedida = área bajo la curva normal a la izquierda de $Z=0,18=0,5+0,0714=0,5714$

También puede utilizarse el segundo método del apartado (a)

- Las bombillas eléctricas de una fábrica A, tienen una duración media de 1400 horas con una desviación standard de 200 horas, mientras que las del otro fabricante B, tienen una duración total media de 1200 horas con una desviación standard de 100 horas. Si se toman muestras al azar de 125 bombillas de cada fabricante, ¿Cuál es la probabilidad de que las bombillas de A tengan una duración media que sea al menos – 160 horas o más; y que las del fabricante B, 250 horas más que las de B.

Solución

Se determinará a las medias de las muestras A y B, respectivamente. Entonces:

Y la variable standard para la diferencia de medias es:

- La diferencia de 160 horas en unidades standard es: $160 - 200/20 = -2$, la probabilidad pedida es:

$p = \text{área bajo la curva normal a la derecha de } Z = -2 = 0,5 + 0,4772 = 2,50$

La probabilidad pedida es:

$P = (\text{área bajo la curva normal a la derecha de } Z = 2,5) = 0,5 - 0,4938 = 0,0062$

$P = (\text{área entre } Z = -2,22 \text{ y } Z = -0,74) = (\text{área entre } Z = 2,22 \text{ y } Z = 0) - (\text{área entre}$

$Z = -0,74 \text{ y } Z = 0) = 0,4868 - 0,2704 = 0,2164 = \text{Probabilidad}$

- El peso total excederá de 510 onzas si el peso medio de los 100 cojinetes sobrepasa las 5,10 onzas.

5,10 en unidades standard = $Z_{5,10} = (5,10 - 5,02/0,027) = 2,96$

La probabilidad pedida será:

$P = \text{área a la derecha de } Z = 2,96 = (\text{área a la derecha de } Z = 0) - (\text{área entre } Z=0 \text{ y } Z = 2,96) = 0,5 - 0,4985 = 0,0015 = \text{Probabilidad}$

Es decir que solamente se darán 3 casos cada 2000 de obtener una muestra de 100 cojinetes con un peso total superior a 510 onzas.

3- Se ha encontrado que el 2% de las piezas producidas por cierta máquina son defectuosas. ¿Cuál es la probabilidad de que una partida de 400 piezas sean defectuosas?

a) 3% o más b) 2% o menos

Solución

- Mediante la corrección para variables discretas $1/2N = 1/800 = 0,00125$ se tiene $(0,03 - 0,00125)$ en unidades standard = $0,03 - 0,00125 - 0,02 = 1,25$

0,007

La probabilidad pedida es el área bajo la curva normal a la derecha de $Z = 1,25$

$P = (\text{área bajo la curva } Z = 1,25) = 0,1056$

Si no se hubiese aplicado la corrección se hubiese obtenido 0,0764

Otro método

3% de 400 = 12 piezas defectuosas, considerando variable continua 12 equivalente a 11.5 o más.

Por otra parte

Luego 11,5 en unidades = $(11,5 - 8) / 2,8 = 1,25$ y como antes la probabilidad tipificadas pedida es 0,1056 de la tabla.

RESUMEN DEL CAPITULO 6

- Muestra aleatoria: todos tienen igual probabilidad. Asegura estimación insesgada.
- Media población; desvío standard población; N tamaño de la muestra.
- Al crecer N y se aproxima a normal se trabaja con tabla Gauss
- Proporciones son similares a las muestras la semejanza no es accidental.
- Muestreo sin reemplazo es más eficiente que el muestreo con reemplazo, reduce el EE de o de P por el factor Np población, N muestra.
- El método Monte Carlo de simulación de muestreos confirma los resultados teóricos como la regla de la aproximación normal. Se adapta especialmente a computadoras.

Ejercicio Monte Carlo

Registro de Ausentismo de 100 empleados (población)

| Distribución de la Población | | | |
|------------------------------|------------|-------------------------------|-------------------|
| Días Ausencia | Frecuencia | Frecuencia Relativa
$p(x)$ | Números asociados |
| 0 | 30 | 30 | 01 - 30 |

| | | | |
|---|----|----|---------|
| 1 | 26 | 26 | 31 – 56 |
| 2 | 22 | 22 | 57 – 78 |
| 3 | 12 | 12 | 79 – 90 |
| 4 | 7 | 07 | 91 – 97 |
| 5 | 2 | 02 | 98 – 99 |
| 6 | 1 | 06 | 00 |

M= 100 1.00

a) Calcular de la población

- Extraer sucesivas muestras de tamaño $N=5$ tomando como base los mismos números aleatorios. Calcular la media y la desviación de cada muestra.
- Calcular la media y la desviación de la distribución muestral de medias, comparar con los valores poblacionales calculados.
- Verificar fluctúa en forma normal alrededor de .

TRABAJO PRACTICO N° 5

- Supóngase que las alturas de 3000 estudiantes de una universidad se distribuye normalmente con media 68,0 pulgadas y desviación 3,0 pulgadas.
- Si se toman 0 muestras de 25 estudiantes cada una, ¿Cuál será la media y la desviación esperada de la distribución muestral de medias resultante si el muestreo se hizo: a) entre 66,8 y 68,3 pulgadas; b) menor que 66,4 pulgadas.
- Los resultados de una elección demostraron que un cierto candidato obtuvo 46% de los votos. Determinar la probabilidad de que de a) 200 individuos elegidos al azar de la población. B) 1000 votantes se hubiesen obtenido una mayoría de votos para dicho candidato.
- Los cojinetes de una determinada casa pesan 0,50 onzas, con una desviación de 0,02 onzas, ¿Cuál es la probabilidad de que dos lotes de 1000 cojinetes cada una difieran un peso superior a 2 onzas?
- A y B juegan a cara y cruz, lanzando cada uno 50 monedas. A ganará el juego si consigue 5 o más caras que B de otro modo gana B. Determinar la proporción en contra de que A gane un juego determinado.
- La desviación de los pesos de una población muy grande de estudiantes es 10,0 libras. Se extraen muestras de 200 estudiantes cada una, de la población y se calculan las desviaciones standard de las alturas de cada muestra.

Hallar (a) la media (b) la desviación de la distribución muestral de las desviaciones.

- ¿Que porcentaje de muestras en el problema anterior tendrían desviaciones standard a) mayor de 11,0 libras, b) menor de 8,0 libras?
- Una población de 7 números tiene una media de 40 y una desviación standard de 3. Si se extraen muestras de tamaño 5 de esa población y se calcula la varianza de cada muestra, hallar la media de la distribución muestral de varianzas en el muestreo es. A) con reemplazamiento, b) sin reemplazamiento.
- Se ha encontrado que el tiempo de servicio que se requiere por persona en una caja bancaria tiene una distribución aproximadamente normal con media

Segundos y segundos. ¿Cuál es la probabilidad de que una persona elegida al azar; a) requiera menos de 100 segundos para terminar sus transacciones, b) pase entre dos y tres minutos en la caja bancaria.

R: a) 0,2514 b) 0.4538

- De los 8 hoteles ubicados en una zona turística, puede decirse que tres de ellos son mediocres en sus servicios.

Un agente de viajes elige al azar dos hoteles para dos clientes que planean pasar sus vacaciones en esa zona.

¿Cuál es la probabilidad de que al menos uno de los clientes se hospede en una de los hoteles mediocres?

R: 0,6429

TRABAJO PRACTICO N°3

Probabilidad – Capítulo 4

- Determinar la probabilidad p en cada uno de los siguientes eventos.
- La aparición de al menos una cara, en dos lanzamientos de una moneda.
- La aparición de un as, el diez de diamantes o el dos de corazones en una sola extracción de una baraja de 52 cartas.
- La obtención de 7 puntos en una sola tirada de un par de dados.
- De una caja que contiene 6 bolas rojas, 4 blancas y 5 azules se extrae una al azar. Determinar la probabilidad de que sea:

a– Roja b– Blanca c– Azul d– No roja e– Roja o Blanca

- Se extraen sucesivamente tres bolas de la caja del problema 2. Hallar la probabilidad de que sean extraídas en el orden roja, blanca, azul; si las extracciones son:
- con reemplazamiento
- sin reemplazamiento
- Hallar la probabilidad de obtener al menos con 4 en dos lanzamientos de un dado.

5– Utilizando un espacio muestral, demostrar que:

- $P(A + B) = P(A) + P(B) - P(AB)$
- $P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$

CAPITULO 10: REGRESIÓN Y CORRELACION

Análisis de Regresión y de Correlación. Regresión Lineal. Ecuación de Regresión y Método de Mínimos Cuadrados, análisis y propiedades. Predicciones usando la recta de regresión, calidad. Coeficientes de Correlación y Determinación. Análisis de Regresión no Lineal, Conceptos.

10.1 Regresión y Correlación

Relación entre Variables

Muy a menudo en la práctica se encuentra que existe más relación entre dos o más variables. Por ejemplo, los pesos de hombres adultos dependen en cierto modo de sus alturas. Las circunferencias de los dependen de los radios. La presión de una masa dada de gas depende de su temperatura.

Frecuentemente se necesita expresar una relación mediante una ecuación matemática que ligue las variables.

Curva de Ajuste

Para llegar a encontrar la ecuación que relaciona las variables, el primer paso es la recolección de datos que muestran los valores correspondientes de las variables.

Por ejemplo, supongamos que x e y representan la altura y el peso de hombres adultos.

Una muestra de N individuos daría $x_1, x_2, \dots, x_n$ alturas y los pesos correspondientes serían $y_1, y_2, \dots, y_n$.

El paso siguiente es representar los puntos correspondientes a cada elemento de la muestra $(x_1, y_1); (x_2, y_2) \dots (x_n, y_n)$ en un sistema de coordenadas cartesianas.

Este sistema de puntos resultante se denomina Diagrama de Dispersión.

Teniendo el diagrama de dispersión es posible encontrar una curva que se aproxima a los datos. Esta curva recibe el nombre de Curva de Aproximación.

A los datos se aproximan a una línea recta, se dice que entre las variables existe una relación lineal. Si entre las variables existe una relación que no es lineal (que no se puede aproximar por una recta sino por una curva).

Se dice que entre las variables hay una relación no lineal.

El problema general de encontrar ecuaciones de curvas e aproximación que se ajusten al conjunto de datos se denomina buscar la curva de ajuste.

Las curvas de ajuste pueden ser varias y de diversos tipos. Entre las curvas polimoniales tenemos:

- a) $y = \text{Recta}$
- b) $y = \text{Parabola de } 2^\circ \text{ grado}$
- c) $y = \text{Cúbica}$
- d) $y = \text{Cuártica}$
- e) $y = \text{Curva de grado } n$

Otras curvas pueden usarse según el caso:

- f) $y = \text{Exponencial}$
- g) $y = \text{Geométrica}$
- h) $y = \text{Curva de Gompertz}$

Para facilitar el trabajo con este tipo de curvas se usa una escala logarítmica para graficar.

- La Línea Recta

El tipo más sencillo de curva de aproximación es la línea recta, cuya ecuación puede escribirse:

O sea, que dados los puntos (x_1, y_1) y (x_2, y_2) de la curva, las constantes a_0 y a_1 pueden determinarse.

La ecuación de la línea resultante puede describirse:

Donde m es la pendiente de la recta y la constante a_0 , es el valor de y cuando $x=0$.

El sistema de ecuaciones para la recta de mínimos cuadrados puede extraerse de la siguiente forma.

Los valores de y de la recta de mínimos cuadrados, son los correspondientes a los valores $x_1, x_2 \dots x_n$ o sea, sin

Por otra parte los valores reales de las ordenadas de los puntos muestrales son $y_1, y_2, \dots y_n$.

Entonces la condición impuesta por la definición es la siguiente:

La condición de S mínimo, se cumple cuando las derivadas parciales de S con respecto a a_1 y a_2 sean iguales a cero, es decir:

Las constantes a_0 y a_1 se obtienen de estas expresiones y son:

El trabajo necesario para encontrar la recta de mínimos cuadrados puede simplificarse transformando los datos, de forma que y

Entonces la ecuación de la recta de mínimos cuadrados puede escribirse:

En particular si x es tal que es decir la ecuación queda:

De estas ecuaciones se deduce inmediatamente que la recta de mínimos cuadrados pasa por el punto que es la recta de curvas de regresión de y sobre x , en este caso se llama frecuentemente Recta de Tendencia, Curva de Tendencia y se utiliza para fines de estimación, predicción o pronóstico.

10.7 Problemas que abarcan más de dos Variables

Veremos el ejemplo de una relación de tres variables x, y, z que puede estar dada por la ecuación: $Z = a_0 + a_1 x + a_2 y$ y que se denomina ecuación lineal entre las variables x, y, z .

En un sistema de coordenadas cartesianas bidimensional esta ecuación representa un plano y los puntos de la muestra pueden o no estar muy distantes de ese plano.

Por extensión del método de los mínimos cuadrados se puede hablar en este caso de plano de mínimos cuadrados.

Si se está estimando la variable z , a partir de valores x e y dados, este sería el plano de regresión de z sobre x e y .

Los problemas normales correspondientes al plano de mínimos cuadrados están dados por:

Los problemas de este tipo se denominan problemas de regresión múltiple.

10.8 Ejemplos de Aplicación

1) – La tabla siguiente, muestra las alturas con aproximación de una pulgada y los pesos con aproximación de una libra, de una muestra de 12 estudiantes extraídos al azar entre los estudiantes de primer año de un colegio.

A – A partir de una recta de mínimos cuadrados siendo x la variable independiente.

B – Siendo y la variable independiente.

| | | | | | | | | | | | | |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Altura X | 70 | 63 | 72 | 60 | 66 | 70 | 74 | 65 | 62 | 67 | 65 | 65 |
| Peso Y | 155 | 150 | 180 | 135 | 156 | 168 | 178 | 160 | 132 | 145 | 139 | 152 |

Solución

A – El trabajo necesario para el cálculo de las sumas se puede ordenar en una tabla, como la siguiente:

| Altura X | Peso Y | | | | | |
|----------|--------|-------|--------|---------|-------|--------|
| 70 | 155 | 3,2 | 0,8 | 2,56 | 10,24 | 0,64 |
| 63 | 150 | – 3,8 | – 4,2 | 15,96 | 14,44 | 17,64 |
| 72 | 180 | 5,2 | 25,8 | 134,16 | 27,04 | 665,64 |
| 60 | 135 | – 6,8 | – 19,2 | 130,56 | 46,24 | 368,64 |
| 66 | 156 | – 0,8 | 1,8 | – 1144 | 0,64 | 3,24 |
| 70 | 168 | 3,2 | 13,8 | 44,16 | 10,24 | 190,44 |
| 74 | 178 | 7,2 | 23,8 | 171,36 | 51,84 | 566,44 |
| 65 | 160 | – 1,8 | 5,8 | – 10,44 | 3,24 | 33,64 |
| 62 | 132 | – 4,8 | – 22,2 | 106,56 | 23,04 | 492,84 |
| 67 | 145 | 0,2 | – 9,2 | – 184 | 0,04 | 84,64 |
| 65 | 139 | – 1,8 | – 15,2 | 27,36 | 3,24 | 231,04 |
| 68 | 152 | 1,2 | – 2,2 | – 1,64 | 1144 | 4,84 |

Luego la recta de mínimos cuadrados pedida es:

O sea

Que puede escribirse también:

Esta ecuación de la recta, encontrada por el método de mínimos cuadrados se llama recta de regresión y y sobre x, y se utiliza para estimar los valores de y a partir de valores dados de x.

B – Si la variable independiente es x, la recta pedida será:

Que puede escribirse:

Que se denomina recta de regresión de x sobre y, y se usa para estimar valores de x a partir de valores dados de y.

a)

- Primer Método

Se utiliza la ecuación donde . El trabajo puede ordenarse en una tabla.

| Año | | | | | | |
|------|----|-------|-----|--------|----|--------|
| 1946 | 0 | 66,6 | - 5 | - 28,4 | 25 | 142,0 |
| 1947 | 1 | 84,9 | - 4 | - 10,1 | 16 | 40,4 |
| 1948 | 2 | 88,6 | - 3 | - 6,4 | 9 | 19,2 |
| 1949 | 3 | 78,0 | - 2 | - 17,0 | 4 | 34,0 |
| 1950 | 4 | 96,8 | - 1 | 1,8 | 1 | - 1,8 |
| 1951 | 5 | 105,2 | 0 | 10,2 | 0 | 0 |
| 1952 | 6 | 93,2 | 1 | - 1,8 | 1 | - 1,8 |
| 1953 | 7 | 111,6 | 2 | 16,5 | 4 | 33,2 |
| 1954 | 8 | 88,3 | 3 | - 6,7 | 9 | - 20,1 |
| 1955 | 9 | 117,0 | 4 | 22,0 | 16 | 88,0 |
| 1956 | 10 | 115,2 | 5 | 20,2 | 25 | 101,0 |

La ecuación planteada queda como sigue:

Que puede escribirse también como:

Donde el origen $x = 0$ es el año 1946 y las unidades de x son 1 año.

El gráfico de esa recta se llama a veces, recta de tendencia y aparece a trazos en la figura. La ecuación se llama a menudo ecuación de tendencia, y los valores de y calculados para diferentes valores de x se llaman valores de tendencia.

• Segundo Método

Si se asignan valores de x a los años 1946 – 1956, la ecuación de la recta de mínimos cuadrados puede escribirse:

En los datos y cálculos realizados en el ejemplo 1, representar graficamente ambas rectas y estimar:

A – El peso de un estudiante cuya altura es de 63 pulgadas.

B – La altura de un estudiante cuyo peso es de 168 libras.

Solución

A – Para estimar y de x se emplea la recta de regresión de y sobre x .

$$Y = 3,22x - 60,9, \text{ Entonces si } x = 63 \quad Y = 3,22 \cdot 63 - 60,9 = 142$$

B – Para estimar x de y se emplea la recta de regresión de x sobre y .

$$X = 31,0 + 0,232y$$

Luego para $Y = 168,06$ la altura

$$X = 31,0 + 0,232 \cdot 168 = 70,0 \text{ pulgadas}$$

3 – Ejemplo de Aplicación a Series de Tiempo

La producción de acero en los EEUU en mill de

toneladas cortas (1 tonelada corta = 2000 libras), durante los años 1946 a 1956 se muestra en la tabla.

A – Representar los datos.

B – Hallar la ecuación de la recta de mínimos cuadrados que se ajuste a los datos.

C – Estimar la producción de acero para los años 1957 y 1958 y comparar con los valores 112,7 y 89,6 mill.

D – Estimar la producción de acero durante los años 1945 y 1944 y comparar con los valores reales de 79,6 y 89,6 mill de toneladas cortas.

| Año | Producción Acero |
|------|------------------|
| 1946 | 66,6 |
| 1947 | 84,9 |
| 1948 | 88,6 |
| 1949 | 78,0 |
| 1950 | 96,8 |
| 1951 | 195,2 |
| 1952 | 93,2 |
| 1953 | 111,6 |
| 1954 | 88,3 |
| 1955 | 117,0 |
| 1956 | 115,2 |

Trabajo Práctico

5 – Hallar:

- la desviación típica de x
- la desviación típica de y
- la varianza de x
- la varianza de y
- la covarianza de x e y para los datos del problema 4.

6 – Comprobar la fórmula $r = S_{xy}/S_x S_y$ para los datos del problema 4.

7 – La siguiente tabla muestra la edad x y la presión sanguínea y de 12 mujeres.

- hallar el coeficiente de correlación entre x e y.
- Determinar la ecuación de regresión de mínimos cuadrados de y sobre x.
- Estimar la presión sanguínea de una mujer de 45 años con el 99 % de confianza.

| Edad (X) | 56 | 42 | 72 | 36 | 63 | 47 | 55 | 49 | 38 | 42 | 68 | 60 |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Presión (Y) | 147 | 145 | 160 | 118 | 149 | 128 | 150 | 145 | 115 | 140 | 152 | 155 |

8 – Hallar el coeficiente de correlación entre los valores de alturas y los pesos de 300 hombres de EEUU,

dados en la siguiente tabla de frecuencias.

| | 59 – 62 | 63 – 66 | 67 – 70 | 71 – 74 | 75 – 78 |
|-----------|---------|---------|---------|---------|---------|
| 90 – 109 | 2 | 1 | | | |
| 110 – 129 | 7 | 8 | 4 | 2 | |
| 130 – 149 | 5 | 15 | 22 | 7 | 1 |
| 150 – 169 | 2 | 12 | 63 | 19 | 5 |
| 170 – 189 | | 7 | 28 | 32 | 12 |
| 190 – 209 | | 2 | 10 | 20 | 7 |
| 210 – 229 | | | 1 | 4 | 2 |

9 – a) Hallar la ecuación de regresión de mínimos cuadrados de y sobre x para los datos del problema anterior.

b) Estimar los pesos de dos hombres cuyas alturas sean 64 y 72 pulgadas respectivamente.

- Estimar con el 95 % de confianza los pesos indicados en b.

Trabajo Práctico

1 – La producción de cigarros puros en EEUU durante los años 1945 – 1954 aparece en la tabla.

- Representar los datos.
- Hallar la ecuación de la recta de mínimos cuadrados que ajuste los datos.
- Estimar la producción de cigarro en el año 1955.

| Año | 1945 | 1946 | 1947 | 1948 | 1949 | 1950 | 1951 | 1952 | 1953 | 1954 |
|----------------------|------|------|------|------|------|------|------|------|------|------|
| Cigarros mill | 98,2 | 92,3 | 80,0 | 89,1 | 83,5 | 68,9 | 69,2 | 67,1 | 58,3 | 61,2 |

2 – Ajuste una recta de mínimos cuadrados a los datos de la tabla adjunta tomando: a) x como variable independiente, b) y como variable independiente.

Representar los datos y las rectas de mínimos cuadrados en el mismo sistema de ejes, coordenads.

| | | | | | | |
|---|---|---|---|---|---|----|
| X | 3 | 5 | 6 | 8 | 9 | 11 |
| Y | 2 | 3 | 4 | 6 | 5 | 8 |

3 – El tiempo total para detener un automóvil después de percibir un peligro se compone del tiempo de reacción, más el tiempo de frenada. La tabla adjunta muestra las de parada en pies D que marcha a la velocidad V (mmillas por hora), desde el instate en que se observa el peligro.

- Representar los datos
- Ajustar una parábola de mínimos cuadrados de la forma $D =$
- Estimar D cuando $V = 45$ mill/hs y 80 mill/hs

| | | | | | | |
|---|----|----|-----|-----|-----|-----|
| V | 20 | 30 | 40 | 50 | 60 | 70 |
| D | 54 | 90 | 138 | 206 | 292 | 396 |

4 – Hallar el coeficiente de correlación lineal entre las variables x e y presentadas en la tabla siguiente