CAPITULO SEIS

PROBABILIDAD

En la toma de decisiones para conseguir un objetivo no siempre se consigue un resultado. En la actividad diaria alcanzar una meta va acompañado de un riesgo. El riesgo tiene varias formas de expresión. Una de ellas es, que el esfuerzo que se realice puede ser vano. Otra que el esfuerzo pueda conducir, por desconocimiento de los diferentes impactos, a un resultado adverso. Una forma de medir el riesgo es a través del estudio de las probabilidades. Por ejemplo, se desea una cantidad de dinero para instalar una clínica para pacientes que sufren de problemas renales y por cada 1000 nuevos soles que se gaste se ganará 300 nuevos soles al año. La pregunta necesaria, es que tan probable es que se ganen los 300 nuevos soles. Analicemos este tema.

1. DEFINICIÓN

La probabilidad puede describirse como la ciencia de formular aseveraciones acerca de lo que ocurrirá cuando se tomen muestras de poblaciones conocidas.
El empleo de la probabilidad permite a quien toma decisiones, analizar los riesgos y minimizar el azar inherente, con información limitada, por ejemplo, al lanzar un nuevo producto o aceptar un embarque recién llegado que contenga partes defectuosas. Una probabilidad es un valor que está entre 0 y 1 que representa la posibilidad de lo que sucederá en un evento en particular.

EXPERIMENTO: Observación de alguna actividad o la acción de efectuar una medición.

EXPERIMENTO ALEATORIO (ε)

Es aquel que, al ser observado no se puede predecir con exactitud cuál será el resultado de la observación y pueden dividirse en dos clases: determinístico y no determinístico.
Se dice determinístico cuando los resultados del experimento están completamente determinados y pueden describirse mediante una fórmula matemática, mientras el no determinístico no puede predecirse con exactitud antes de realizar el experimento.

RESULTADO: Un acontecimiento final de un experimento.

EVENTO: Conjunto de uno o más resultados de un experimento

Ejemplo:

Experimento: lanzar dos dados y observar lo que cae.
Resultados posibles:

(1,1),(1,2),(1,3),(1,4) ,(1,5),(1,6)
(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)
(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)
(4,1),(4,2),(4,3),(4,4),(4,5(,(4,6)
(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)
(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)

Sean los eventos:
A: El número del segundo dado sea par.
B: El número del primer dado sea mayor que el del segundo.

2. ENFOQUES DE LA PROBABILIDAD

Son la Objetiva y Subjetiva. La probabilidad objetiva puede subdividirse en: probabilidad clásica o a priori, probabilidad como frecuencia relativa o probabilidad a posteriori.

2.1. DEFINICION CLASICA DE PROBABILIDAD

La definición clásica se basa en el supuesto de que todos los resultados posibles de un experimento aleatorio son **igualmente probables**, es decir, cada uno de los elementos del espacio muestral tiene la misma probabilidad de ocurrencia. Además sostiene que son **mutuamente excluyentes**, es decir, que no son comunes en resultado, debido a que no puede aparecer más de un par en forma simultánea, y si n_A de estos resultados tienen un atributo A, la probabilidad de A es la proporción de n_A con respecto a n (total de resultados posibles).
Se puede representar mediante la siguiente fórmula:

$$Probabilidad\ de\ un\ evento = \frac{Número\ de\ resultados\ favorables}{Número\ total\ de\ observaciones}$$

Ejemplo1: Una lotería consta de 1000 billetes. Un billete se premia S/. 100.00, 4 billetes de S/. 50.00, 10 billetes de S/. 20.00, 20 billetes de S/. 10.00, 165 billetes con S/. 5.00 y 400 billetes con S/. 1.00 cada uno. Los demás billetes no se premian. Se compra un billete, ¿Cuál es la probabilidad de ganar por lo menos S/. 10.00?

Solución:

El experimento aleatorio es "elegir un billete".
S = $\{B_1, B_2, \ldots, B_{1000}\}$, donde B_i representa el billete número i.
# total de observaciones = n(S) = 1000
Sea el evento A: "ganar por lo menos S/. 10.00"
Ganar al menos S/. 10.00, significa que se puede ganar S/.10.00, ó S/.20.00 ó S/. 50.00 ó S/.100.00. Es decir: n(A) = 20+10+4+1=35.

$$P(A) = \frac{n(A)}{n(S)} = \frac{35}{1000} = 0.0035$$

La probabilidad de ganar más de S/. 10.00 es 0.0035

Ejemplo: Las estimaciones poblacionales del INEI para el año 2000 determinaron que la población peruana era de 25'661,690, de los cuales 7'466.190 nacieron en el departamento de Lima. Calcular la probabilidad de que una persona que emigre al exterior nació en Lima.

Solución:

El experimento aleatorio es "persona que emigre al exterior".
S = $\{P_1, P_2, \ldots, P_{25`661,690}\}$, donde P_i representa la Persona i.
# Total de observaciones = n(S) = 25'661,690
Sea el evento A: "Persona que emigre al exterior nació en Lima".

n(A) = 7'466.190 (total de personas que nacieron en Lima)

$$P(A) = \frac{n(A)}{n(S)} = \frac{7'466,190}{25'661,690} = 0.291$$

Luego, la probabilidad de que la persona que emigre al exterior sea del departamento de Lima, para el año 2000, es de 0.291.

2.2. DEFINICIÓN DE PROBABILIDAD COMO FRECUENCIA RELATIVA

Si un experimento se repite n veces bajo las mismas condiciones y n_B de los resultados son favorables a un atributo B, el límite de n_B/n conforme n se vuelve grande, se define como la probabilidad del atributo B. Por ejemplo, en una fábrica mayormente se observan productos de mejor calidad mientras que los productos defectuosos se observan muy pocas veces, entonces la probabilidad de defectuosos se determinará en proporción al total de artículos producidos en dicha fábrica.

En términos de fórmula:

$$Probabilidad\ de\ que\ suceda\ un\ evento = \frac{Número\ de\ veces\ que\ el\ evento\ ocurrió\ en\ el\ pasado}{Número\ total\ de\ observaciones}$$

Ejemplo: Una muestra aleatoria de 10 fábricas que emplean un total de 10,000 personas, demostró que ocurrieron 500 accidentes de trabajo durante un período reciente de 12 meses. Hallar la probabilidad de un accidente de trabajo en una industria determinada.

Solución:
N = 10,000 personas que equivale al número de veces que se repite el experimento.
Sea el evento A: "un persona que sufrió un accidente de trabajo en la industria determinada"
Entonces n(A) = 500 y

$$P(A) = \frac{n(A)}{n} = \frac{500}{10,000} = 0.05$$

Por definición de frecuencia relativa, ya que este valor de la probabilidad, se basa en una muestra, por la tanto es una estimación del valor real desconocido. Observe, aquí se supone implícitamente que las formas de seguridad no han cambiado desde que se realizó el muestreo.

La probabilidad que una persona sufra un accidente de trabajo, en el año, en la industria, es 0.05.

Si se entrevistan a 100 personas en forma aleatoria es probable que 5 sufrieron un accidente de trabajo.

Ejemplo: Según la encuesta de hogares de Lima Metropolitana del año 2000 se ha obtenido el siguiente resultado. En 3 meses de observación a una muestra de 16,684 personas entrevistadas, 4,955 sufrieron una enfermedad o accidente. Hallar la probabilidad de elegir una persona haya sufrido una enfermedad o accidente

Solución:
N =16,684, número de personas entrevistadas.
Sea el evento A: "elegir una persona que halla sufrido una enfermedad o accidente".
n(A) = 4,955 (total de personas que sufrieron alguna enfermedad o accidente en la muestra)

$$P(A) = \frac{n(A)}{n(S)} = \frac{4,955}{16,684} = 0.297$$

La probabilidad de elegir una persona que haya sufrido alguna enfermedad o accidente es de 0.297.

2.3. PROBABILIDAD SUBJETIVA

Una probabilidad subjetiva se basa en cualquier información disponible. Se aplica sólo cuando no existe suficiente información para que sea utilizable otro método.

Ejemplo: La posibilidad de que un alumno obtenga una calificación de 20 en el curso de estadística.

3. DEFINICION AXIOMATICA DE PROBABILIDAD

ESPACIO MUESTRAL (Ω)

Es el conjunto de todos los posibles resultados de un experimento aleatorio y podemos describirlos con precisión, pueden ser finito, infinito numerable o infinito no numerable.

Espacio Muestral Finito

Se dice que un espacio muestral es finito cuando el resultado de un experimento aleatorio es contable, es decir, finito.

Ejemplo: El número de personal administrativo contratado en un hospital para 1999 constituye un espacio muestral finito, dado que el número nunca excederá a la cantidad programada para este año.

Ejemplo:

Sea: ε_1: la producción de un artículo por una determinada máquina.
Ω_1= {bueno, defectuoso}

Sea: ε_2: el crecimiento de un niño y anotar su sexo.
Ω_2= {varón, mujer}

Sea: ε_3: la elección de un ciudadano y anotar su grado de instrucción.
Ω_3= {primaria, secundaria superior}

Sea: ε_4: Lanzamiento de un dado sobre una superficie lisa y observar el número que aparece en la cara superior.
Ω_4= {1, 2, 3, 4, 5, 6}

Sea: ε_5 : se lanzo una moneda 3 veces y se cuenta el # de caras.
$\Omega_5 = \{0, 1, 2, 3\}$

Sea ε_6 : se lanza un dado hasta obtener por primera vez el 6.
$\Omega_6 = \{E, \overline{E}\,E, \overline{E}\,\overline{E}\,E, \overline{E}\,\overline{E}\,\overline{E}\,E,...\}$

Donde: E: Sale 6 en un tiro del dado
$\overline{E}$: No sale 6 en un tiro del dado

Sea ε_7: medir la resistencia a la tensión de una barra de acero.
$\Omega_7 = \{s / s \geq 0\}$

Según su naturaleza un **espacio muestral** puede ser numérico (como: Ω_4, Ω_5, Ω_7) o no numérico (como: Ω_1, Ω_2, Ω_3).
Según el número de elementos, puede ser **finito** (como: Ω_1, Ω_2, Ω_3, Ω_4, Ω_5).
Infinito numerable (como: Ω_6) e **infinito no numerable** (como: Ω_7).

4. REGLAS BASICAS DE PROBABILIDAD

4.1. REGLA DE ADICION

Para aplicar la regla especial de adición, los eventos deben ser mutuamente excluyentes. Esta regla se expresa con la fórmula siguiente:

$$\mathbf{P(A \cup B) = P(A) + P(B)}$$

Ejemplo: En el cuadro siguiente, se tiene información acerca de la población de mujeres en edad fértil para el año 2000. Si se escoge a una mujer para ser censada, hallar la probabilidad de que una mujer se encuentre en el grupo de 15 a 19 años de edad o en el grupo de 35 a 39 años de edad.

PERÚ: MUJERES EN EDAD FÉRTIL- 2000

Grupos de Edad	2000	
	Mujeres	Probabilidad de que una mujer se encuentre en un grupo de edad
TOTAL	**6,874,923**	**100.0**
15-19	1,331,836	0.194
20-24	1,268,424	0.185
25-29	1,126,802	0.164
30-34	989,498	0.144
35-39	859,297	0.125
40-44	710,789	0.103
45-49	588,277	0.086

FUENTE: INEI-Perú: Estimaciones y Proyecciones de la Población por Años Calendarios

Solución:
Total de observaciones: n(s) =6'874,923, total de mujeres en edad fértil.
Sea el evento A: "la mujer seleccionada tengan entre 15 y 19 años de edad"
Sea el evento B: "la mujer seleccionada tengan entre 35 y 39 años de edad"
Sea el evento A U B: "la mujer seleccionada se encuentre en el grupo de 15 a 19 años de edad o en el de 35 a 39 años de edad.

Como se puede observar, se trata de eventos mutuamente excluyentes. Por ello se usará la siguiente fórmula"

$$P(A \cup B) = P(A) + P(B)$$

Reemplazando datos:

$$P(A \cup B) = \frac{1'331,836}{6'874,923} + \frac{859,297}{6'874,923} = 0.194 + 0.125 = 0.319$$

La probabilidad de que al escoger a una mujer en edad fértil para ser censada, se encuentre en el grupo de 15-19 años de edad o en el de 35-39, es 0.319.
La interpretación en términos de la frecuencia relativa es: De cada 100 mujeres en edad fértil, 31 se encuentran en el grupo de 15 a 19 años de edad o en el de 35 a 39 años de edad.

REGLA GENERAL DE LA ADICIÓN
La regla general de la adición se utiliza para combinar eventos que no son mutuamente excluyentes:

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

Ejemplo: En el siguiente cuadro se presentan datos para el año 2000, se cuenta con las probabilidades referidas a los distintos niveles de pobreza. Si elige una persona al azar, calcular la probabilidad de que esa persona tenga un ingreso que se encuentra por debajo de la línea de pobreza o que tenga al menos una necesidad básica insatisfecha[1].

[1] *INDICADORES DE NBI:*
Con viviendas de características físicas inadecuadas (estera, quincha, madera, piso de tierra, improvisada, etc.)
Con viviendas hacinadas (más de 3 por habitación)
Sin servicios higiénicos.
Con niños que no asisten a la escuela.
Con alta dependencia económica (jefe del hogar con 2º de primaria y 3 personas por ocupado)

NIVELES DE POBREZA, 2000

	ABSOLUTO	PROBABILIDAD
Población total	**25,661,690**	
Población no pobre	11,958,347	0.534
P. con ingresos menores a la línea de pobreza	9,700,119	0.378
P. con al menos una necesidad básica insatisfecha (NBI)	10,033,721	0.391
P. con ingresos menores a la línea de pobreza y con al menos una NBI	6,030,497	0.235

Solución:
Total de observaciones: n(s) =25'661,690, total de la población.
Sea el evento A: "la persona se encuentre por debajo de la línea de pobreza"
Sea el evento B: "la persona se encuentre con al menos una necesidad básica insatisfecha"
Sea el evento $(\mathbf{A \cap B})$: "persona se encuentre por debajo de la línea de pobreza y que tenga necesidades básicas insatisfechas".

Como se puede observar, se trata de eventos que no son mutuamente excluyentes. Por ello se usará la siguiente fórmula

$$\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

Reemplazando se tiene:

$$\mathbf{P(A \cup B) = \frac{9{,}700{,}119}{25'661{,}690} + \frac{10{,}033{,}721}{25'661{,}690} - \frac{6{,}030{,}497}{25'661{,}690} = 0.378 + 0.391 - 0.235 = 0.534}$$

De cada 100 personas en el país para el año 2000, 53 personas eran pobres porque tenían una necesidad básica insatisfecha o porque tenían un ingreso muy bajo. En términos de probabilidades significa que: _ la probabilidad de que una persona perciba ingresos muy bajos o que tenga por lo menos una necesidad básica insatisfecha es 0.534.

La regla del complemento sirve para determinar la probabilidad de que ocurra un acontecimiento, restando del número uno, la probabilidad de que no suceda el acontecimiento:

$$\Omega = \{A_1, A_2, \ldots A_n\} \quad \text{ó} \quad \Omega = \{A, A'\}$$

$$P(\Omega) = 1 \quad \Rightarrow \quad P(A) + P(A') = 1$$

$$\boxed{P(A) = 1 - P(A')}$$

Ejemplo: En la siguiente tabla se muestra la población urbana y rural. La probabilidad de que una persona viva en el área urbana es de 0.723. A partir de este dato calcular la probabilidad de que una persona viva en el área rural.

PERÚ: POBLACION URBANA-RURAL,2000

	ABSOLUTO	PROBABILIDAD
Perú	**25,661,690**	
Urbana	18,553,402	0.723
Rural	7,108,288	

Solución:
Total de observaciones: n(s) = 25'661,690, total de la población.
Sea el evento A: "la persona viva en el área rural"
Sea el evento A': la persona no viva en el área rural", es decir que viva en el área urbana.
Tenemos entonces

$$\mathbf{P(A) = 1 - P(A') = 1 - \frac{18{,}553{,}402}{25{,}661{,}690} = 1 - 0.723 = 0.277}$$

La probabilidad de que una persona resida en el área rural es de 0277.
Por cada 100 peruanos 27 residen en el área rural.

Probabilidad conjunta: Probabilidad que mide la posibilidad de que dos o más eventos ocurran en forma simultánea.

4.2. INDEPENDENCIA DE EVENTOS
Dos o más eventos son independientes cuando la ocurrencia de uno no tiene efecto en la probabilidad de ocurrencia de cualquier otro.

4.3. REGLA DE MULTIPLICACION
Se usa para combinar eventos.
La regla especial de multiplicación requiere que los dos eventos A y B sean independientes. Se expresa de la forma siguiente:

$$\mathbf{P(A \cap B) = P(A) * P(B)}$$

REGLA GENERAL DE LA MULTIPLICACION (Teorema)
Se utiliza para determinar la probabilidad conjunta formada por todos los resultados comunes tanto en A como en B que ocurren al mismo tiempo, los cuales se asume como eventos no independientes y se denota por $P(A \cap B) = P(B/A)P(A)$. Este teorema permite incluir cualquier número de eventos que se encuentran en el espacio muestral.

5. PROBABILIDAD CONDICIONAL
Es la probabilidad de que ocurra un evento particular, dado que ocurrió otro evento. Es la diferencia que existe entre elegir al azar un artículo de un lote con o sin sustitución.

Ejemplo: cuando se elige un tipo de producto en un lote donde existen productos sanos y defectuosos. La probabilidad condicional de obtener un producto defectuoso (B) dado que en el lote por lo general existen productos sanos (A), se denota así: P (B/A).

B: El producto seleccionado sea defectuoso
A: El producto seleccionado sea no defectuoso o sano

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

Ejemplo: La población económicamente activa PEA (15 y más años) del Perú, para el año 2000 es de 10'387,225 personas, de los cuales 3'748,236 son mujeres y 2'866,953 residen en el área urbana. Calcular la probabilidad de que se escoja a una persona de la PEA que resida en el área urbana, dado que la persona escogida fue PEA mujer.
Solución: Se trata de calcular una probabilidad condicional.

La población económicamente activa es de: 10'387,225 = n
La población femenina económicamente activa es de: 3'748,236 = n(A)
La población femenina económicamente activa del área urbana es de: 2'866,953.

Sea
A: La persona escogida sea mujer de 15 y más años
B: La persona escogida sea mujer del área urbana

La probabilidad que la PEA sea mujer es: $P(A) = \frac{n(A)}{n} = \frac{3748,236}{10'387,225} = 0.361$

$A \cap B$: La persona escogida sea mujer de 15 y más años y que pertenezca al área urbana

La probabilidad de $P(A \cap B) = \frac{n(A \cap B)}{n} = \frac{2'866,953}{10'387,225}$

Dado que ha sido PEA mujer la persona elegida, la probabilidad de que resida en el área urbana es:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{2'866,953}{10'387,225}}{\frac{3'748,236}{10'387,225}} = \frac{2'866,953}{3'748,236} = 0.765$$

6. TEOREMA DE LA PROBABILIDAD TOTAL

Según la teoría de la probabilidad total, un evento está en función de un conjunto de particiones totalmente independientes, por lo tanto se puede decir que es la sumatoria de todas las particiones y se representa así:

$$A = A \cap B_1 \cup A \cap B_2 \cup \ldots \cup A \cap B_k$$

Lo importante es que todos los eventos $A\bigcap B_1,\ldots,A\bigcap B_k$ son parejas mutuamente excluyentes. Por lo tanto podemos aplicar la propiedad aditiva para este tipo de eventos.

$$\mathrm{P(A)} = \mathrm{P(A \bigcap B_1)} + \mathrm{P(A \bigcap B_2)} + \ldots + \mathrm{P(A \bigcap B_k)}$$

Sin embargo cada término $P(A\bigcap B_j)$ se puede expresar como $P(A/B_j)P(B_j)$, con lo cual obtenemos finalmente el teorema de la probabilidad total:

$$P(A) = P(A/B_1)P(B_1) + P(A/B_2)P(B_2)+\ldots+P(A/B_k)P(B_k)$$

7. DIAGRAMA DE ARBOL (O ARBORIGRAMAS)

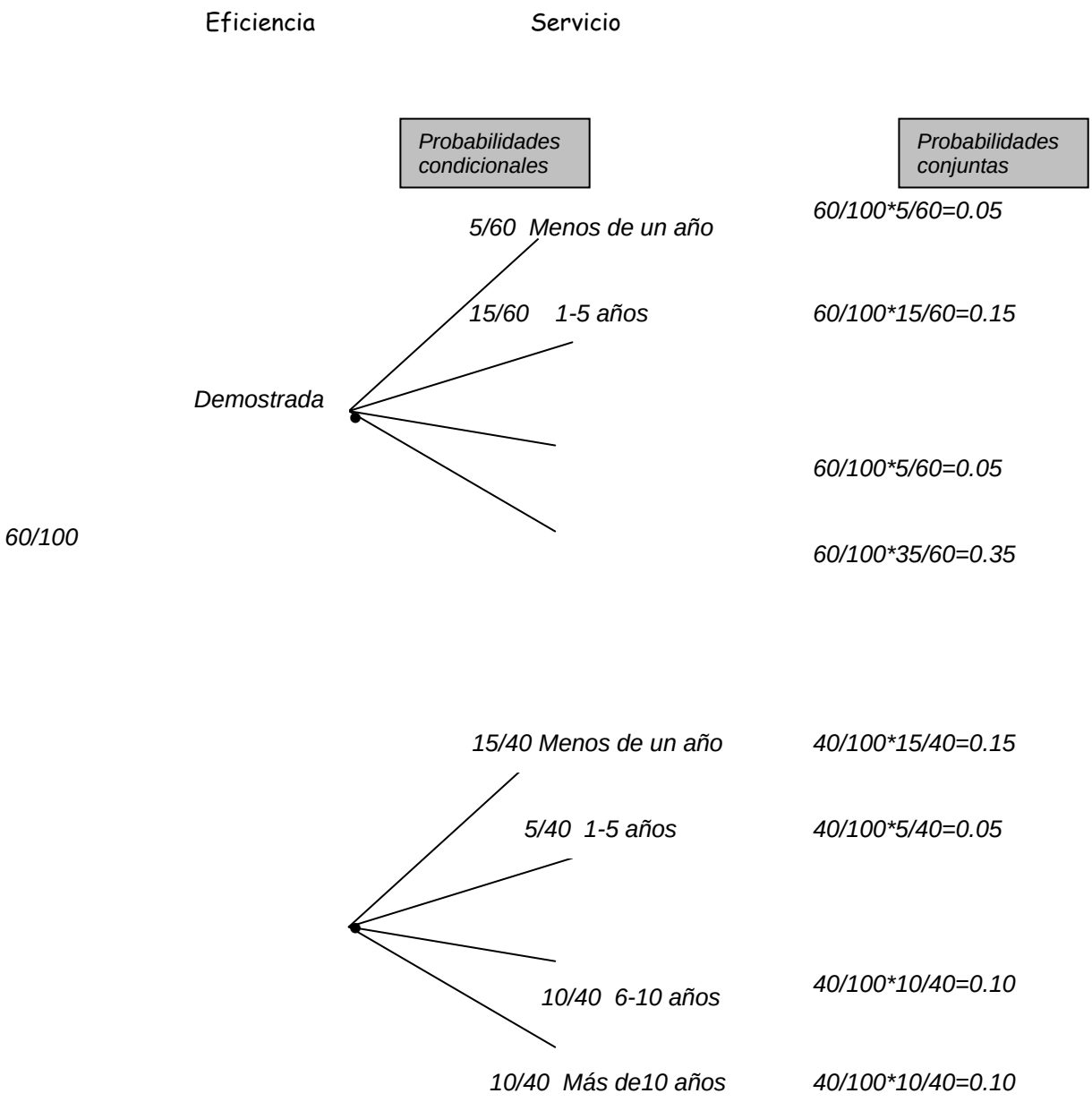
Este diagrama es muy útil para representar probabilidades condicionales y conjuntas. Un diagrama de árbol es particularmente útil para analizar decisiones de negocios, donde existen varias etapas para el problema.

Ejemplo:

EFICIENCIA Y AÑOS DE SERVICIOS DEL PERSONAL DE LA EMPRESA POLYSISTEMAS

EFICIENCIA	AÑOS DE SERVICIO				TOTAL
	Menos de 1 año	1 a 5 años	6 a 10 años	Más de 10 años	
DEMOSTRADA	5	15	5	35	60
NO DEMOSTRADA	15	5	10	10	40
	20	20	15	45	100

Diagrama de árbol que indica la eficiencia y los años de servicio

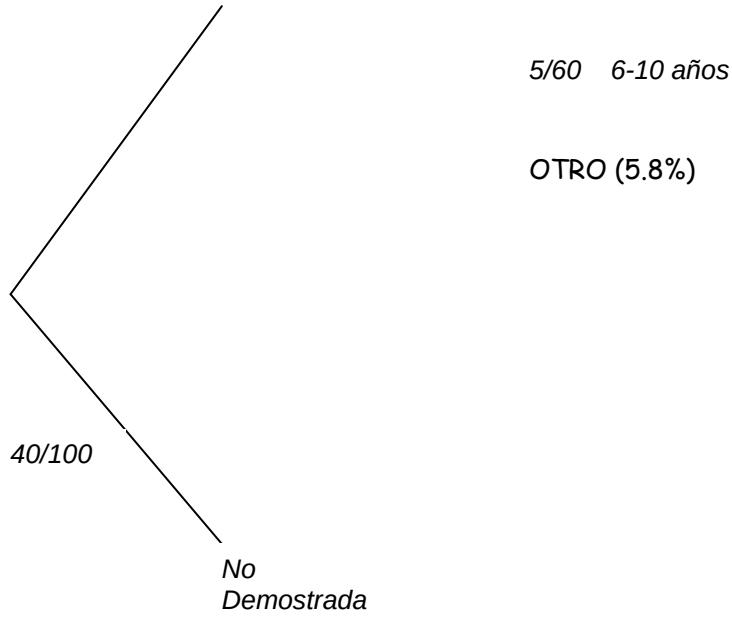


EJERCICIOS

1.- ¿Cuál es la diferencia entre un experimento y un evento?
Experimento: observación de los resultados de una actividad que se realiza.
Resultado: acontecimiento final de un experimento.

2.- ¿Es posible que una probabilidad asuma un valor de cero?
Si es posible, por ejemplo: la probabilidad que un reloj marque la hora 25, la probabilidad de que la tierra se convierta en agua, la probabilidad de que un pez camine, etc. Se sabe por regla general que $0 \leq P(x) \leq 1$; siendo x un evento cualquiera.

3.- El director general de una clínica expresará mañana a los accionistas su consideración de que la clínica debe fusionarse con otra institución del mismo ramo. Ha recibido seis cartas acerca de esa cuestión y está interesado en el número de personas que estén de



acuerdo con él.
a. ¿Cuál es el experimento?
b. ¿Cuáles son algunos de los eventos posibles?
c. Exprese dos posibles resultado.

a) E: cartas de accionistas que están de acuerdo con la fusión de la clínica con otra institución.
b) - De acuerdo un numero de 3 cartas.
- De acuerdo un numero de 2 cartas.
- De acuerdo un numero mayor a 4 cartas.
c) - 2 personas de acuerdo con la fusión de la clínica.
- 3 personas de acuerdo con la fusión de la clínica.

4.- Defina la expresión mutuamente excluyente con sus propias palabras.
Se dice eventos mutuamente excluyentes cuando dichos eventos no pueden suceder en forma simultánea, es decir, la ocurrencia de uno excluye la ocurrencia de otro.

5.- Según el "II Censo de infraestructura Sanitaria y Recursos del Sector Salud", en el Perú existían 9,658 médicos del Ministerio de Salud, de los cuales 733 trabajan en el área rural. En el departamento de Amazonas trabajaban 93 médicos, de los cuales 20 eran del área rural. Calcular la probabilidad de elegir un médico que trabaje en el departamento de Amazonas y que sea del área rural.

Solución:
total de observaciones: n(s) =9,658 médicos
total de médicos rurales =733 médicos
total de médicos en el departamento de Amazonas.

Sea el evento A: "el médico trabaje en el departamento de Amazonas"
Sea el evento B: "el médico trabaje en el área rural.
Sea el evento $(A \cap B)$: "el médico trabaje en el departamento de Amazonas y que sea del área rural".
La elección de un médico que trabaje en el departamento de Amazonas es independiente de elegir un médico que trabaje en el área rural dentro del departamento, entonces se aplica la siguiente fórmula:

$$P(A \cap B) = P(A) \times P(B) = \frac{93}{9,658} * \frac{20}{93} = 0.0021$$

De cada 1,000 médicos del Ministerio de Salud 2 trabajaban en el área rural del departamento de Amazonas.
La probabilidad de que un médico esté en el departamentote Amazonas y que trabaje en el área rural es 0.0021.

6.- Un estudiante está tomando dos cursos, estadística y matemática básica. La probabilidad de que sea aprobada en el curso de estadística es 0.60, y que pase el curso de matemática básica es 0.70. La probabilidad de que apruebe en ambas es 0.50. ¿Cuál es la probabilidad de que pase por lo menos en una?

7.- ¿Qué es una tabla de contingencia? ¿Qué indica?
Suponga que P(A)=0.40 y P (B/A)=0.30. ¿Cuál es la probabilidad conjunta de A y B?

9.- Un hospital tiene cuatro proveedores de materia prima. En la tabla que sigue se muestran las cantidades adquiridas de cada proveedor y el porcentaje de materia prima defectuosa que cada uno proporciona.

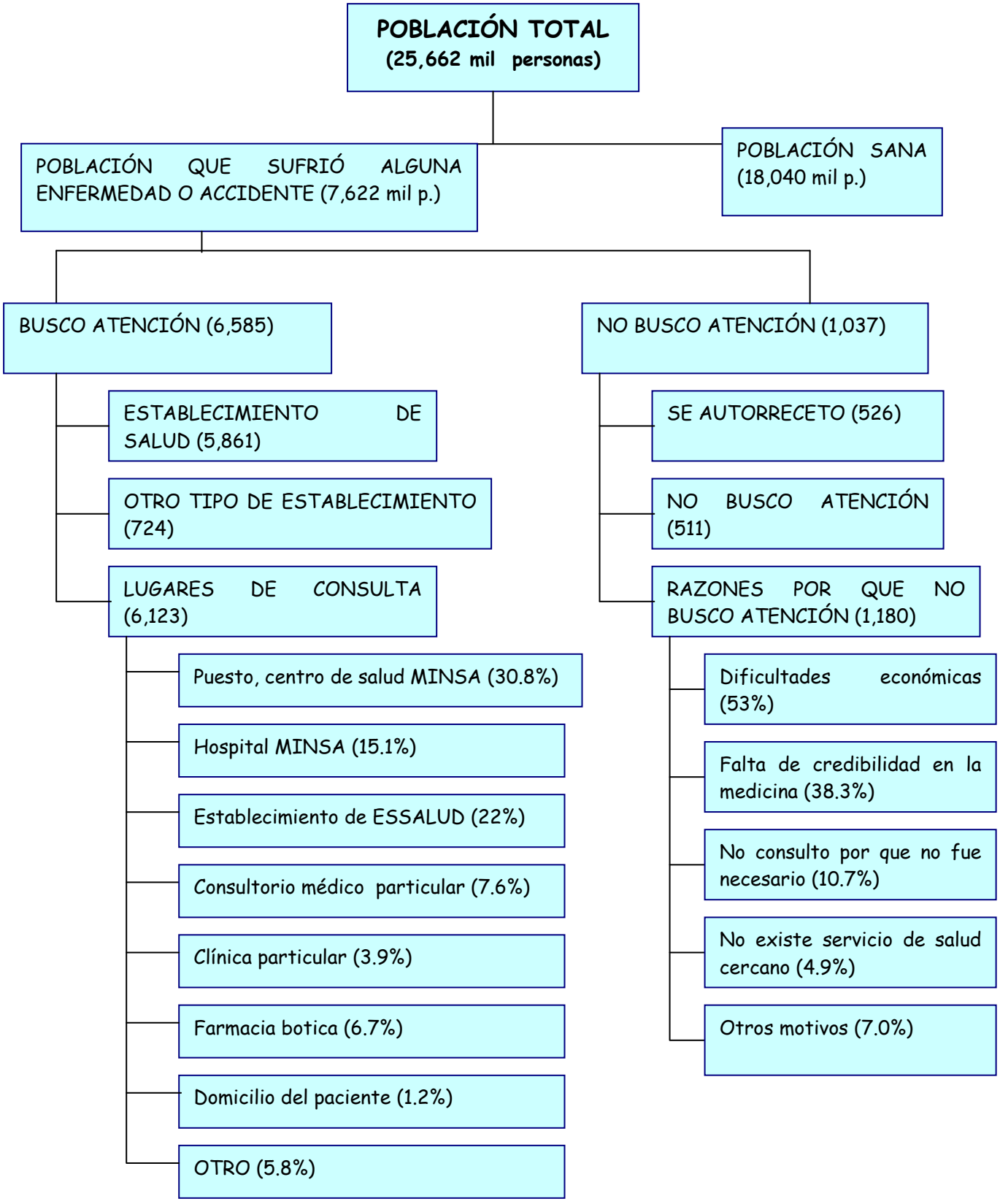
Proveedor	% Adquirido	% Defectuoso
Martinez Asociados	30.0	2.50
Asmat Mgf.	20.0	1.75
Millones SA	25.0	3.00
Garcia Ltda.	25.0	1.00

El material empleado esta mañana resultó defectuoso. ¿Cuál es la probabilidad de que se haya adquirido de la compañía Asmat Mgf?

10. En el diagrama siguiente se muestra algunos datos de la Encuesta Nacional de Hogares - IV trimestre 2000, éstos están referidos al estado de salud de la población.

En base a estos datos, determinar:

a. Calcular la probabilidad de que una persona este sana.
b. La probabilidad de que una persona no buscó atención si se sabe que sufrió una enfermedad o accidente.
c. Calcular la probabilidad que una persona haya respondido que no busco atención debido a que tenía dificultades económicas y por que tuvo algún otro motivo.
d. Se sabe que la probabilidad de que una persona no accedió a un servicio de Salud por que éste no queda cerca de su domicilio es de 0.049. calcular el total de personas que no accedieron a un servicio de salud por este motivo, si se sabe que el total de razones del por que no buscaron atención sumaban 1'180 miles.
e. La probabilidad de que una persona enferma se atendiera en un establecimiento de salud es 0.89005. Calcular la probabilidad de que una persona se atendiera en otro tipo de establecimiento.


POBLACIÓN TOTAL
(25,662 mil personas)
POBLACIÓN QUE SUFRIÓ ALGUNA ENFERMEDAD O ACCIDENTE (7,622 mil p.)
POBLACIÓN SANA (18,040 mil p.)
BUSCO ATENCIÓN (6,585)
NO BUSCO ATENCIÓN (1,037)
ESTABLECIMIENTO DE SALUD (5,861)
OTRO TIPO DE ESTABLECIMIENTO (724)
LUGARES DE CONSULTA (6,123)
Puesto, centro de salud MINSA (30.8%)
Hospital MINSA (15.1%)
Establecimiento de ESSALUD (22%)
Consultorio médico particular (7.6%)
Clínica particular (3.9%)
Farmacia botica (6.7%)
Domicilio del paciente (1.2%)
OTRO (5.8%)
SE AUTORRECETO (526)
NO BUSCO ATENCIÓN (511)
RAZONES POR QUE NO BUSCO ATENCIÓN (1,180)
Dificultades económicas (53%)
Falta de credibilidad en la medicina (38.3%)
No consulto por que no fue necesario (10.7%)
No existe servicio de salud cercano (4.9%)
Otros motivos (7.0%)

Solución:

a. Las observaciones totales n(S) = 25'662 (población total)
Sea el evento A: escoger una persona sana n(A) = 18'041

$$P(A) = \frac{n(A)}{n(S)} = \frac{18\,'041}{25\,'662} = 0.703$$

La probabilidad de que se escoja a una persona sana es de 0.703. Es decir que de cada 100 personas, 70 personas están sanas.

b. La población total es 25'662 mil habitantes.
Sea el evento A: "la persona haya sufrido un accidente" n(A) = 7'622
Sea el evento B: "la persona no buscó atención" n(B) = 1'037
Sea el evento C: "la persona escogida no buscó atención, dado que sufrió una enfermedad o un accidente"
Sea $A \cap B$: Sufre accidente y busca atención

La probabilidad que la persona haya sufrido un accidente: $P(A) = \frac{n(A)}{n} = \frac{7622}{25662} = 0.297$

La probabilidad: $P(A \cap B) = \frac{n(A \cap B)}{n} = \frac{1037}{25662} = 0.040$

La probabilidad de que una persona no busco atención médica, si se sabe que ésta sufrió un accidente es:

$$P(C) = P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1037}{25662}}{\frac{7622}{25662}} = \frac{1037}{7622} = 0.136$$

La probabilidad de que una persona no buscó atención dada que tuvo alguna enfermedad o sufrió un accidente es de 0.136. De cada 100 personas que sufrieron de algún accidente 13 no buscaron atención médica.

c. Del diagrama podemos ver el porcentaje de las razones del por que las personas no accedieron a un servicio de salud. El total no suma 100% debido a que algunas personas más de una respuesta.

Si dividimos las razones en dos categorías, se tiene:
A: # de personas que contestaron tener dificultades económicas P(R1) = 0.530
B: # de personas que contestaron tener otros motivos P(R2) = 0.609

El total de personas que no accedieron a un servicio de salud representa el 100% (de los que sufrieron una enfermedad o un accidente)

Se trata de eventos que no son mutuamente excluyentes; aplicando la propiedad se tiene:

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B) \Rightarrow P(A \cap B) = P(A) + P(B) - P(A \cup B)}$$

Reemplazando datos, tenemos:

$$\boxed{P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.530 + 0.609 - 1 = 0.139}$$

La probabilidad de que una persona respondiera haber tenido más de un motivo para no buscar atención es de 0.139.

11. Según estudios del Ministerio de Salud, para el año 1999 se reportaron 49,393 casos de Malaria en el departamento de Loreto. Los niños menores a 1 año que sufrieron de esta enfermedad sumaban 671 casos y los 5 a 14 años de edad sumaban 12,763 casos. Determinar la probabilidad de que un niño escogido al azar tenga Malaria y que sea menor de un año o que tenga entre 5 a 14 años de edad.

12. Según la encuesta Nacional de Hogares ENAHO 99-IV Trimestre, para una población de 5'297,178 hogares, 3'572,912 estaban en el área urbana. Además se sabe que la probabilidad de que una casa sea independiente y que se encuentre en el área urbana es 0.811, la probabilidad de que una casa urbana sea de vecindad es 0.068 y para el caso del área rural se tiene las probabilidades de 0.9 y 0.011 respectivamente. Calcular la probabilidad de que el hogar encuestado sea del área rural, dado que fue una casa independiente.

EJERCICIOS APLICATIVOS AL TEMA DE LA MORTALIDAD

1. La probabilidad de morir (tasa bruta de mortalidad: $d^{\dagger}$) para el año 1999, fue de 0.00628. La población total para ese mismo año ascendió a 25'232,226. Calcular el número de defunciones ocurridas en ese año.

Sea P(A): Probabilidad de morir
n: población total = 25'232,226

$$P(A) = d^{1999} = \frac{\#\ Defunciones}{25'232,226} = 0.00628$$

$$\#\ Defunciones = 25'232,226 * 0.00628 = 158,500$$

Para el año 1999 se registraron 158,500 defunciones.

2. Para el año 1997 se tiene que el número de sobrevivientes de una muestra que alcanzaron los 25 años ascendió a 88,545 personas, y el número de muertes entre los 25 y 30 años de edad es de 1,268. Hallar la probabilidad de morir entre las 25 y 30 años de edad.

Sea:
P(A): Probabilidad de morir entre lo 25 y 30 años de edad ($_5q_{25}$)
n(A): 1,268

n: 88,545

Se parte del supuesto que las 1,268 defunciones corresponderían a la población de 88,545. Luego, la probabilidad de morir de una persona de edad x dentro de los "n" años siguientes, se calcula mediante la siguiente fórmula: $_n\mathbf{q_x} = \frac{_n\mathbf{d_x}}{\mathbf{l_x}}$

$$\mathbf{P(A)} = \frac{\mathbf{1,268}}{\mathbf{88,545}} = \mathbf{0.01432}$$

Entonces, la probabilidad de que una persona no llegue viva a los 30 años es 0.01432

3. Se sabe que la probabilidad de sobrevivir hasta la edad exacta de 10 años es de 0.99178, y el número de niños que alcanzaron la edad de 5 años fue de 92,049. Hallar el número de sobrevivientes a la edad de 10 años.

Lo que se desea calcular es de los 92,049 niños, cuantos llegan a cumplir la edad de 10 años (sobrevivientes)

Sea:
P(A): Probabilidad de sobrevivir hasta la edad exacta de 10 años de los sobrevivientes a la edad de 10 años (0.99178) ($_5p_5$)
n(A): niños que alcanzaron la edad de 10 años
n: 92,049

La probabilidad de sobrevivir a una determinada edad se calcula así: $_n\mathbf{p_x} = \frac{\mathbf{l_{x+n}}}{\mathbf{l_x}}$

$$\mathbf{P(A)} = \frac{\mathbf{n(A)}}{\mathbf{92,049}} = \mathbf{0.99178}$$
$$\mathbf{n(A)} = \mathbf{0.99178 * 92,049} = \mathbf{91,012}$$

Entonces el número de niños de 5 años (de una muestra de 92,049) que alcanzan vivos la edad de 10 años es de 91,012.

4. Completar el siguiente cuadro.
Donde:
x: representa la edad exacta
l(x): numero de sobrevivientes que alcanzan la edad exacta x
d(x,n): números de muertes entre las edades x y x+n años (n =5)
$_nq_x$: probabilidad de morir que tiene una persona de edad x dentro de los n años (n =5)
$_np_x$: probabilidad de sobrevivir que tienen una persona de edad exacta x llegue a la edad x+n (n =5).
Además se sabe que **q(x,n) + p(x,n) =1**

X	l(x)	d(x,n)	${}_nq_x$	${}_np_x$
35	85,575	2,241		
40				0.9655
45	80,462		0.0444	
50				0.9440
55	72,586			0.9309
60				0.9147

5. Conocidas las probabilidades de morir entre las edades exactas de 65 a 70 años de edad es de 0.10463, y entre las edades de 75 a 80 años de edad es de 0.17951, calcular el número de sobrevivientes a la edad exacta de 80 años. Se sabe que las defunciones registradas entre las personas que tenían 70 años y que no llegaron a cumplir los 75 años ascendieron a 7,282.

6. El número de defunciones registradas entre las edades de 5 a 9 años es 7,112. Además la población entre los 10 y 14 años de edad es de 2'821,096 y las defunciones en esa edad suman 3,315. Hallar la probabilidad de morir a la edad de 5 a 9 años o de 10 a 14 años de edad.

7. Para el año 1999, la probabilidad de morir por causa materna en el departamento de Ucayali fue de 0.0079. Además, el total de mujeres en edad fértil en ese departamento con respecto al país es de 1.62%. Sabiendo que la población peruana de mujeres en edad fértil es de 186,868 mujeres, calcular el número de muertes debido a una causa materna.

8. Para el año 2000 se registraron el nacimiento de 607,800 niños. La probabilidad de muerte infantil (dentro del primer año de vida) es de 0.039, además se sabe que la probabilidad de morir dentro de los 28 días o un mes de nacido es de 0.02. Hallar el número de nacidos que sobrevivieron al primer año de vida.

9. En el Perú para el año 1996 se registraron 26,972 muertes de niños menores a 1 año. El total de nacimientos para ese año fue de 611,600. A partir de estos datos calcular la probabilidad de que un niño no muera dentro del primer año de nacido.

10. La probabilidad de morir para el año 2000 fue de 0.00629 y la población estimada para ese mismo año fue de 25'661,690. Calcular el total de personas que murieron en ese año.

CAPITULO SIETE

TEOREMA DE BAYES

En este caso si $\boxed{B_1, B_2, \ldots, B_n}$ son n eventos mutuamente excluyentes de los cuales por lo menos uno debe ocurrir. Y se denota de la siguiente manera:

$$P(B_j / A) = \frac{P(A/B_j)P(B_j)}{\sum_{j=1}^{n} P(A/B_j)P(B_j)}\ ,\ j=1,2,\ldots n$$

Ejemplo: En una línea de producción hay dos procesos, A y B. En el proceso A hay un 20% de defectos y en B hay 25%. En una muestra de 300 productos hay 200 del proceso A y 100 del proceso B.

(a) Si se extrae un producto al azar, hallar la probabilidad que sea defectuoso.
(b) Si al extraer el producto resultó defectuoso, halle la probabilidad de que sea del proceso A.

Solución: Sean los siguientes eventos:

A: "el producto es del proceso A".
B: "el producto es del proceso B".
$\underline{D}$: "el producto es defectuoso".
$\overline{D}$: "el producto es no defectuoso".

Ω = **A** U **B**. Es decir, A y B forman una partición de Ω..

(a) Debemos calcular P[D]. Este evento se escribe D = **A**D U **B**D y por el teorema de probabilidad total es:

$$P[D] = P[\mathbf{A}D] + P[\mathbf{B}D] = P[\mathbf{A}]\,P[D\backslash\mathbf{A}] + P[\mathbf{B}]P[D\backslash\mathbf{B}]$$

$$= \frac{2}{3}(0.20) + \frac{1}{3}(0.25) = \frac{65}{300}.$$

(b) Por el teorema de Bayes se tiene:

$$P[\mathbf{A}\backslash D] = \frac{P[A]P[D\backslash A]}{P[D]} = \frac{(2/3)(0.20\ )}{65/300} = 0.615$$

Ejemplo: Según un estudio realizado, para una muestra de 1,357 personas, se obtuvo los siguientes: las personas que fuman eran 1350, las personas que fuman y tienen cáncer pulmonar eran 133 y las personas que no fuman y tienen cáncer pulmonar eran 3. Calcular la probabilidad de que una persona fume si se sabe que tiene cáncer pulmonar.

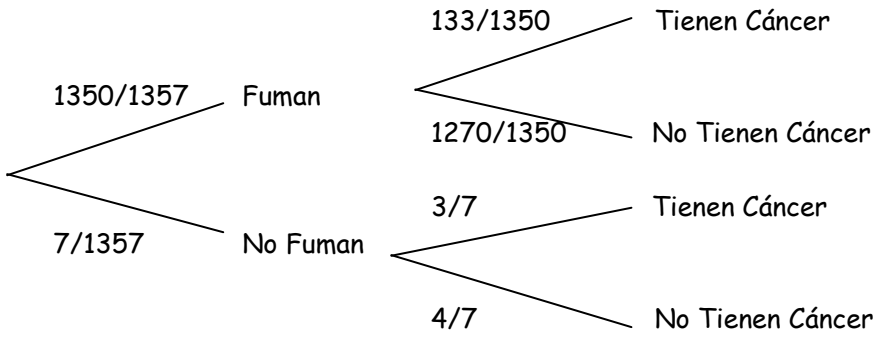
Solución: Se tiene los siguientes datos:

El total de observaciones es 1,357

A: persona fumadora n(A) = 1350

B: persona no fumadora n(B) = 7

Realizando el diagrama de árbol, tenemos:



Lo que se desea calcular es la probabilidad de que la persona escogida sea fumadora, si se sabe que padece de cáncer pulmonar.

Aplicando el teorema de Bayes, tenemos:

$$P(B_j / A) = \frac{P(A/B_j)P(B_j)}{\sum_{j=1}^{n} P(A/B_j)P(B_j)} = \frac{\frac{1350}{1357} * \frac{133}{1350}}{\frac{1350}{1357} * \frac{133}{1350} + \frac{7}{1350} * \frac{3}{7}} = 0.95$$

La probabilidad de que una persona sea fumadora, si se sabe que tiene cáncer pulmonar es 0.95.

Ejercicio 1: Se tiene los siguientes datos: La población peruana en el año 2000 ascendió a 25'661,690, la población masculina ascendía a 12'726,385, por otro lado la PEA total era de 10'387,225 y la PEA femenina era de 3'748,236. ¿Cuál es la probabilidad de que una persona escogida al azar sea varón, si se sabe que no pertenece a la PEA? Analizar el resultado.

APLICACION DE LAS PROBABILIDADES

MEDIDA DE ASOCIACIÓN ENTRE EL FACTOR DE RIESGO Y LA MORTALIDAD INFANTIL

La asociación entre el factor de riesgo y la variable dependiente se mide a través del ODDS Ratio, que es una medida del grado de asociación entre dos variables categóricas. Dentro de un modelo de regresión logística indica el factor de riesgo siempre que su valor sea mayor que 1.

Su cálculo se basa en la comparación del producto de las frecuencias en la diagonal principal de una tabla de doble entrada como la siguiente:

	Mortalidad	
	< 50%	>= 50%
	0.00	1.00
Años de estudio (Categorizado)	0	1
0	10	5
1	6	9

ODDS RATIO (OR) = (10x9)/ (5x6) = 3

En este caso los años de estudio (menos de 5 años) ofrece un riesgo 3 veces mayor respecto a la tasa de mortalidad mayor al 50%.

ANÁLISIS DE UN FACTOR DE RIESGO

Ilustremos esto con un ejemplo, tomando el departamento de Huancavelica, que es uno de los que presenta mayor nivel de mortalidad infantil. Una muestra hipotética de individuos estudiada analiza el Factor de Riesgo, nivel de educación de la madre y su efecto en la Mortalidad Infantil. El mismo, se recoge en una tabla de la forma:

A	B
C	D

		Mortalidad Infantil (M)		
		Sí	No	
Factor de Riesgo (F)	Sí	75	305	380
	No	14	606	620
		89	911	1000

De esta observación se deduce que existen 75 casos de Mortalidad de niños menores de un año, (mortalidad infantil) por cada 380 individuos que presentan el factor de riesgo (ser analfabeto), mientras que existen 14 casos de Mortalidad Infantil por cada 620 individuos que no lo presentan. Si estas frecuencias relativas pueden ser asimiladas a probabilidades por tratarse de una gran muestra, la probabilidad de morir de un niño menor de un año en un hogar, presentando el factor "madre analfabeta", será:

$$P(M/F) = \frac{75}{380} = 0{,}197$$

Mientras que la probabilidad de serlo, no presentando el factor, será:

$$P(M/F) = \frac{14}{620} = 0.022$$

Por consiguiente, se puede decir que habrá más de ocho veces el número de casos de mortalidad infantil cuando existe el factor de riesgo, que cuando no. Pues bien, a esta relación:

$$RR = \frac{P(M/F)}{P(M/F)} = \frac{0,197}{0,022} = 8.95$$

Se denomina riesgo relativo (RR) del factor.

APLICANDO EL ODDS RATIO EN EL EJEMPLO DE HUANCAVELICA

En el caso del ejemplo de Huancavelica podemos obtener el ODDS Ratio de la siguiente forma:
Con riesgo, existen 75 casos de mortalidad infantil por cada 305 niños que no fallecen (75/305=0,245 mortalidad / no mortalidad).

Sin riesgo, existen 14 casos de mortalidad infantil por cada 606 niños que no fallecen (14/606=0,023) mortalidad/ no mortalidad).

Por tanto, con riesgo, habrá $\frac{75/305}{14/606}$ = 10.64 veces más niños muertos menores de un año, que sin riesgo.

Es decir, se observa que el ODDS Ratio es una "razón de proporciones" de presencia de mortalidad infantil por no mortalidad infantil, entre los que presentan el factor y los que no lo presentan. Puesto que igualmente puede expresarse en la forma (75 x 606)/ (14 x 305)=10.64 la odds ratio también se denomina "razón de productos cruzados"

Otra forma de expresar el riesgo relativo y el Odds ratio es con la siguiente tabla:

	Casos Mortalidad Infantil	**Casos de No Mortalidad Infantil**	**Personas-año Riesgo**	**Tasa x 1000 Personas-Año**	**Riesgo Relativo**	**ODDS RATIO**
ANALFABETISMO (Presenta factor de riesgo)	75	305	380	0.197 (75/ 380)	8.95 $\frac{75/380}{14/620}$	10.64 $\frac{75/305}{14/606}$
ALFABETISMO (No presenta factor de riesgo)	14	606	620	0.022 (14 / 620)	Referente	Referente
Total	**89**	**911**	**1000**			

0.197 representa los casos de mortalidad infantil con respecto al analfabetismo.
0.022 representa los casos de mortalidad infantil con respecto al alfabetismo.

En este cuadro el alfabetismo se toma como base de comparación, por eso se le denomina REFERENTE. Es decir, se está comparando la mortalidad infantil cuando se presenta el factor de riesgo (analfabetismo) respecto al alfabetismo.

ALGUNOS PRINCIPIOS DE CONTEO

Existen tres reglas de conteo que son útiles para determinar el número total de modos o formas en que pueden ocurrir eventos.

1. La regla de la **multiplicación** establece que si existen m formas que un evento pueda ocurrir, y n formas en que otro pueda ocurrir, también existirán entonces mn modos en el cual los dos eventos puedan suceder.

2. Una **permutación** es un arreglo en el cual el orden de los objetos es importante.

1	2	3	.....	r	...	n
			.....		...	

$$n\mathrm{P}_r = \frac{n!}{(n-r)!}$$

3. Una **combinación** es un arreglo donde el orden de los objetos no es importante.

1	2	3	.....	r	...	n
			.....		...	

$$_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

1. DIFERENCIA ENTRE UNA PERMUTACION Y UNA COMBINACION

En una permutación el orden de los objetos para cada posible resultado es diferente, mientras en una combinación no importa el orden de los objetos.

2. VARIABLES ALEATORIA UNIDIMENSIONALES

Habiendo considerado las distribuciones de frecuencias de conjuntos de datos y los fundamentos de la probabilidad, ya podemos combinar estas ideas para elaborar Distribuciones de Probabilidad, las cuales de asemejan a las distribuciones de frecuencias

relativas. La diferencia básica entre estos dos tipos de distribuciones es el uso de la variable aleatoria.
La variable aleatoria de una distribución de probabilidad corresponde a la variable respuesta de una distribución de frecuencias.

3. ¿QUÉ ES UNA DISTRIBUCION PROBABILISTICA?

Es la enumeración de todos los resultados de un experimento junto con la probabilidad asociada a cada uno.

4. VARIABLE ALEATORIA

Es un valor numérico determinado por el resultado de un experimento aleatorio al azar y puede tomar distintos valores.
Sea ε un experimento y sea Ω el espacio muestral asociado a él. Una función X que asigna a cada punto muestral w es un número real X(w) y se llama variable aleatoria.
Simbólicamente:

$$X: \Omega \rightarrow \qquad R_X \subset \Re \ , R_X \neq \phi$$

Ejemplos:

1.- Sea la variable aleatoria X el número de llamadas telefónicas recibidas diariamente por una compañía, la cual puede tomar valores entre 0 y algún número grande.
2.- En un estudio sobre la composición familiar, sea la v.a. X el número de hijos por familia, la cual puede tomar valores entre 0 y n.
3.- Al hacer disparos a un blanco, sea la v.a. X que indica el número de aciertos.
En general nos interesamos en los posibles valores de X.

Ejemplo: Al lanzar dos monedas se tiene Ω = {CC, CS, SC, SS}.
Definimos la v.a. X como el número de caras obtenidas en los dos lanzamientos. Por lo tanto, X(CC) = 2, X(CS) = X(SC)=1, X(SS)= 0.
Así, R_X = {0, 1, 2} es el recorrido de la v.a. X.

Nota: al referirnos a las variables aleatorias usaremos letras mayúsculas como X,Y,Z, etc. Cuando hablemos del **valor** de esas variables aleatorias emplearemos letras minúsculas como x,y,z.
Esta variable aleatoria puede ser discreta o continua.

4.1. VARIABLE ALEATORIA DISCRETA

Es la variable que sólo puede tomar ciertos valores claramente definidos y distantes, que es el resultado de contar algún elemento de interés.
Se dice también que una variable aleatoria X es discreta si el conjunto de valores de X, R_X, es finito o infinito numerable, es decir, $R_X=\{x_1,x_2,...\}$, con cada resultado posible de x_i asociamos un número $p(x_i) = P(X=x_i)$, llamado probabilidad de x_i. La función $p(x_i)$, i=1,2,... deben de cumplir las siguientes condiciones:

(1) $p(x_i) \geq 0$, siendo xi un evento cualquiera

(2) $\sum_{i=1}^{\infty} p(x_i) = 1$

La función $p(x_i)$ se llama función de probabilidad de la v.a. X. La colección de pares (x_i , $p(x_i)$), i=1,2,..., se llama distribución de probabilidad de X.
La distribución de probabilidad de una v.a. discreta X permite estudiar completamente a la variable aleatoria y se puede representar por una fórmula, una tabla o una gráfica que indique las probabilidades $p(x_i)$ correspondientes a cada uno de los valores de X.

Ejemplo:

Un capataz en una fábrica tiene 3 hombres y 3 mujeres laborando para él. Desea elegir dos trabajadores para una labor especial y decide seleccionarlos al azar para no introducir algún sesgo en su selección. Sea X el número de mujeres seleccionadas. Encuentre la distribución de probabilidad de X.

Solución:

El capataz puede escoger dos de seis trabajadores de $\binom{6}{2}$ =15 maneras.

Por lo tanto, Ω contiene 15 puntos muestrales, en forma de pares, igualmente probables, los valores de X son: 0, 1, 2.
La función de probabilidad en cada valor de X es:

$$p(0) = P(X=0) = \frac{\binom{3}{0}\binom{3}{2}}{15} = \frac{3}{15} = 1/5$$

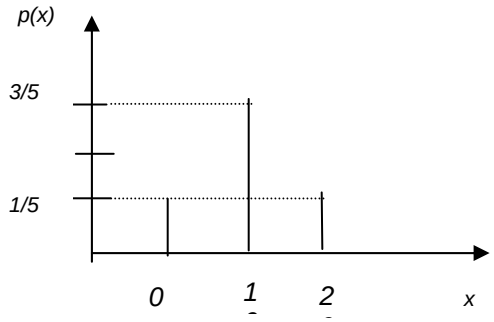
$$p(1) = P(X=1) = \frac{\binom{3}{1}\binom{3}{1}}{15} = \frac{9}{15} = 3/5$$

$$p(2) = P(X=2) = \frac{\binom{3}{2}\binom{3}{0}}{15} = \frac{3}{15} = 1/5$$

La distribución de probabilidad de X se da en la tabla siguiente:

x	p(x)
0	1/5
1	3/5
2	1/5

y la representación gráfica es como aparece a continuación:



La distribución de probabilidad de X también se puede representar por medio de una fórmula. En este caso sería como sigue:

$$P(x) = \frac{\binom{3}{x}\binom{3}{2-x}}{\binom{6}{2}}, \; x = 0,1,2.$$

Las distribuciones de probabilidad son modelos que se utilizan para representar distribuciones empíricas.

Entre las variables aleatorias discretas tenemos: Binomial, Bernolli, Geométrica, Hipergeométrica, Poisson.

4.2. VARIABLE ALEATORIA CONTINUA

Se dice que una variable aleatoria es continua, si se puede tomar cualquiera de los valores de un intervalo.
Ejemplo: La edad, el peso de una persona, el tiempo que dura una bujía, la resistencia a la rotura de una tela de algodón.

Formalmente, una v.a. X es continua si es posible encontrar una función f(x) no negativa que cumple las siguientes propiedades:

(1) $f(x) \geq 0$

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

(3) $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$, donde $a,b \in \Re$, $a<b$

La función f describe la manera como se comporta la variable se llama <u>función de densidad</u> de la v.a. X. El conjunto $\{R_X, f(x)\}$ se llama <u>Distribución de probabilidad</u> de la v.a. X, y contiene toda la información necesaria para estudiar completamente a la v.a. X.
Como consecuencia de la propiedad (3), la probabilidad de una variable alatoria continua tome un valor x_o es cero, puesto que $P(X=x_o) = \int_{x_o}^{x_o} f(x)dx = 0$

Nota: f(x) no representa la probabilidad de nada. Sólo cuando la función se integra entre dos límites produce una probabilidad.

Las variables aleatorias continuas son: La Uniforme, La Normal, La Exponencial, La Ji-Cuadrado.

Ejemplos:

1.- Sea la v.a. contínua X con función de densidad f dada por:

$$f(x)=\begin{cases}2x \quad , \quad 0<x<1\\ 0, \text{ en otro lugar}\end{cases}$$

Claramente, f(x) ≥ 0 y $\int_0^{X_o} f(x)dx = \int_0^1 2xdx = x^2\Big|_0^1 = 1$

Para calcular P(X ≤ ½), debemos evaluar la integral $\int_0^{1/2} 2xdx = x^2\Big|_0^{1/2} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$

2.- Dada $f(y) = \begin{cases}Cy^2 \quad , \quad 0<y<2\\ 0, \text{ en otro lugar}\end{cases}$, encontrar el valor de C

Para que f(y) sea una función de densidad válida buscaremos el valor de C tal que: $\int_0^2 Cy^2dy = 1$, según la propiedad (2) de f. Integrando:

$$\int_0^2 Cy^2dy = C\,\frac{y^3}{3}\Big|_0^2 = C\left[\frac{2^3}{3}+\frac{0^3}{3}\right]$$

Entonces: $C\left[\frac{2^3}{3}+\frac{0^3}{3}\right]=1$

Por lo tanto:

$$C[8/3 + 0] = 1$$

Despejando: C = 3/8.

Podemos evaluar $P(1 \le Y \le 2) = \int_1^2 \frac{3}{8}y^2dy = \left(\frac{3}{8}\right)\frac{y^3}{3}\Big|_1^2 = \frac{3}{8}*\frac{y^3}{3}\Big|_1^2 = \frac{8}{8}-\frac{1}{8} = \frac{7}{8}$

3.- Un vendedor de kerosene tiene un tanque de 150 galones que se llena al principio de cada semana. Su demanda semanal tiene una frecuencia relativa que crece constantemente desde 0 hasta 100 galones y permanece constante entre 100 y 150 galones. Si Y denota la demanda semanal en ciertos galones, la frecuencia relativa de la demanda se puede representar por:

$$f(y) = \begin{cases} y, & 0 \le y \le 1\\ 1, & 1 \le y \le 1.5\\ 0, & \text{en otro lugar}\end{cases}$$

Calcular: P(0 ≤ y ≤ 0.5), P(0 ≤ y ≤ 1.2)

Solución:

$P(0 \le y \le 0.5) = \int_0^{0.5} ydy = \frac{y^2}{2}\Big|_0^{0.5}$ =0.125

$P(0 \le y \le 1.2) = \int_0^1 ydy + \int_1^{1.2} ydy$ = 0.5 + 0.22 = 0.72

5. MEDIA Y VARIANCIA DE UNA VARIABLE ALEATORIA

5.1. MEDIA

Sea X una variable aleatoria **discreta** con función de probabilidad $p(x_i)$. Entonces, el valor esperado de X (media o esperanza matemática de X), E(X), está definido por:

$$E(X) = \sum_i x_i\, p(x_i)$$

Si $p(x_i)$ es una caracterización exacta de la distribución de frecuencias de la población, entonces $E(X)=\mu$, que es la media de la población.

Ejemplo: Consideremos una variable aleatoria discreta X, que puede tomar los valores 0, 1, 2 con distribución de probabilidad dada por:

x	p(x)
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

Entonces:

$$\mu = E(x) = (0*1/4) + (1*1/2) + (2*1/4) = 1$$

Es el valor alrededor del cual se sitúan los valores de x.

Propiedades de E(X)

1) Sea c una constante. Entonces E(c)=c
2) E(cX) = cE(x), siendo c una constante.
3) Sean X e Y dos variables aleatorias cualesquiera. Entonces:
   E(X+Y) = E(X) +E(Y).
4) $E(X \pm c) = E(X) \pm c$, donde c es una constante.
5) $E[(x-u)^2] = E[x^2 - 2ux + u^2] = E(x^2) - 2uE(x) + E(u^2) = E(x^2) - u^2$

Ejemplo: Utilizando la propiedad 5, calcular Var(Y) del ejemplo anterior.
Del el ejemplo anterior se tenía que la media $\mu = 1$ y por tanto:

$E(y^2) = \sum_{y=0}^{3} y^2\, p(y)$ = $(0)^2(1/4) + (1)^2(1/2) + (2)^2(1/4) = 1.75$

Luego:

$$\sigma^2 = E(Y^2) - u^2 = 1.75 - (1)^2 = 0.75$$

5.2. VARIANCIA Y DESVIACION ESTANDAR

La varianza de una variable aleatoria X está definida como el valor esperado de $(u - x)^2$. Es decir:

$$\text{Var}(X) = E[(X - u)^2] = \sum \left[(X - \mu)^2 \, p(x_i) \right]$$

La desviación estándar de X es la raíz cuadrada positiva de Var(X).

Si $p(x_i)$ es una caracterización exacta de la distribución de frecuencias de una población, entonces $E(X)=\mu$, $VAR(x)=\sigma^2$ es la varianza de la población y σ es la desviación estándar de la población.

Los pasos de cálculo son:
1. Restar la media de cada valor y elevar el cuadrado la diferencia.
2. Multiplicar cada diferencia al cuadrado por su probabilidad.
3. Sumar los productos resultantes para llegar a la variancia.

Ejemplo: Encontrar la media, la varianza y la desviación estándar de la variable aleatoria Y, cuya distribución de probabilidad es:

y	p(y)
0	1/8
1	1/4
2	3/8
3	1/4

Entonces:

$\mu = E(Y) = \sum_{y=0}^{3} yp(y)$ = (0*1/8) + (1*1/4) + (2*3/8) + (3*1/4) = 1.75

$\sigma^2 = E[(X - u)^2] = \sum_{y=0}^{3} (y - \mu)p(y)$

$=(0-1.75)^2(1/8)+ (1-1.75)^2(1/4)+(2-1.75)^2(3/8)+(3-1.75)^2(1/4)$

$= 0.9375$

$\sigma = \sqrt{\sigma^2} = \sqrt{0.9375} = 0.97$

EJERCICIOS

1. Resuelva lo siguiente:
   a. 40!/35!
   b. ${}_7P_4$
   c. ${}_5C_2$

2. En una encuesta médica, aleatoriamente se seleccionaron 4 de 10 personas disponibles. ¿Cuántos grupos de diferentes de 4 son posibles?

3. Una empresa de mensajería con viajes durante la noche, debe incluir cinco ciudades en su recorrido. ¿Cuántas rutas diferentes posibles suponiendo que no importa el orden en que las ciudades se incluyen en el recorrido?

4. Una organización nacional de encuestas ha elaborado 15 preguntas destinadas a evaluar la eficiencia de un Hospital estatal. El entrevistador seleccionará 10 de tales preguntas. ¿Cuántos diferentes arreglos existen para el orden de las 10 preguntas seleccionadas?

5. Describa las características de una distribución probabilística discreta

6. Determine la media y la variancia de la siguiente distribución probabilística discreta.

x	P(X)
2	0.50
8	0.30
10	0.20

7. El director de admisiones de la universidad de Ingeniería, estimó como sigue la admisión de estudiantes para el semestre de otoño con base en pasadas experiencias.

Admisión	Probabilidad
1000	0.60
1200	0.30
1500	0.10

   ¿Cuál es el número esperado de alumnos admitidos para el semestre de otoño?. Evalúe la variancia y la desviación estándar.

8. La producción de un analgésico, en u.f. está distribuida según una función continúa:
   $F(x) = 5*(1 - x)$ donde $0 \leq x \leq 1$
   Calcular:

a. La producción media
b. La desviación Típica
c. El coeficiente de variación
d. La moda.

CAPITULO OCHO

DISTRIBUCION PROBABILISTICA BINOMIAL

Es una distribución de probabilidades discreta y tiene las siguientes características:

1. Un resultado de un experimento se clasifica en una de dos categorías mutuamente excluyentes que son éxito o fracaso
2. Los datos recopilados son resultados de conteos.
3. La probabilidad de un éxito permanece igual para cada ensayo. Lo mismo sucede con la probabilidad de fracaso.
4. Los ensayos son independientes, lo cual significa que el resultado de un ensayo no afecta al resultado de algún otro.
5. Una probabilidad binomial se determina como sigue:

$$P(r) = {}_nC_r p^r q^{n-r}$$

6. La media se calcula como sigue:

$$\mu = np$$

7. La variancia es:

$$\sigma^2 = np(1-p)$$

Nota: Cuando n = 1 la distribución binomial corresponde a la distribución Bernoulli con parámetro p.

Ejemplo:

Todos los días se seleccionan, de manera aleatoria, 15 unidades de un proceso de manufactura con el propósito de verificar el porcentaje de unidades defectuosas en la producción. Con base en información pasada, la probabilidad de tener una unidad defectuosa es de 0.05. La gerencia ha decidido detener la producción cada vez que una muestra de 15 unidades tenga dos o más defectuosas. ¿Cuál es la probabilidad de que, en cualquier día, la producción se detenga?

Solución:

Si el modelo apropiado para esta situación es la distribución binomial, se puede suponer que las 15 unidades que se seleccionan al día, constituyen un conjunto de ensayos independientes de manera tal que la probabilidad de tener una unidad defectuosa es 0.05 entre ensayos. Sea X el número de unidades defectuosas que se encuentran entre las 15. Para n = 15 y p = 0.05, la probabilidad de que la producción se detenga es igual a la probabilidad de que X sea igual o mayor que dos. De esta manera:

$$P(X \geq 2) = 1\, P(X \geq 1) = 1 - P(X < 2)$$

$$= 1 - [P(X = 0) + P(X = 1) = 0.1709$$

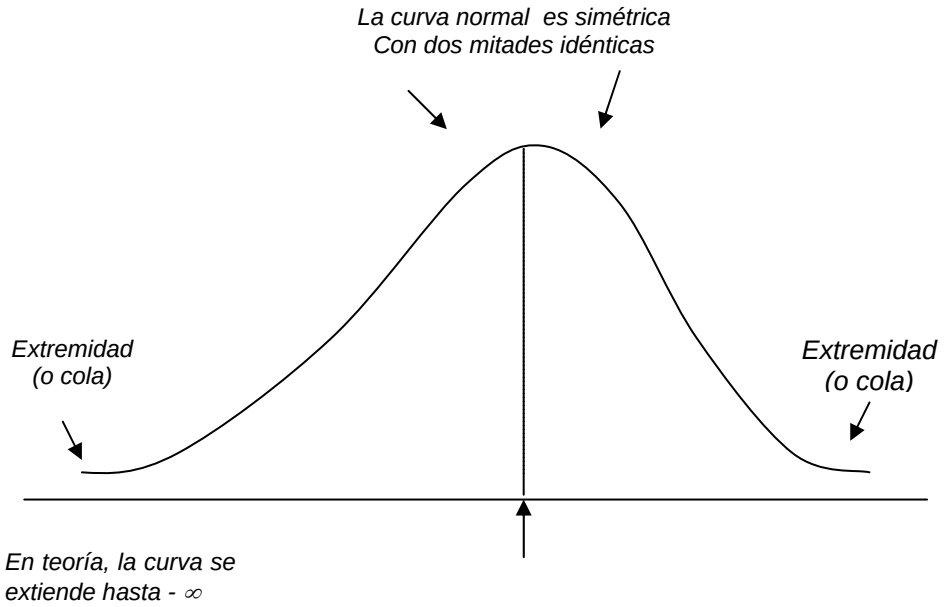
DISTRIBUCION DE PROBABILIDAD NORMAL

LA DISTRIBUCION NORMAL

Se dice que la variable X tiene una distribución normal con parámetros μ y σ^2, y se escribe $X \sim N(\mu,\sigma^2)$, si su función de densidad es:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$, con $-\infty<\mu<\infty$, $\sigma>0$, $x\in\Re$

Cada par (μ,σ^2), da lugar a una distribución diferente y cuando se tiene valores dados de μ y σ^2 se determina completamente a la distribución normal de interés.



Este modelo que vamos a emplear para representar la distribución de ciertos variables continuas, en poblaciones inmensamente grandes tiene las siguientes características:

1.- Es acampanada y la media, la mediana y la moda son iguales.
2.- La distribución probabilística es simétrica con respecto a la media
3.- La curva normal decrece uniformemente en ambas direcciones a partir del valor central. Es asintótica, lo que significa que la curva se acerca cada vez más al eje x, pero en realidad nunca llega a tocarlo.
4.- Es descrita completamente por la media y la desviación estándar.
5.- Existe una familia de distribuciones normales. Cada vez que cambian la media o la desviación estándar, se origina una nueva distribución normal.
6.- La variable asume todos los valores reales, es decir va de $-\infty$ a ∞

DISTRIBUCION DE PROBABILIDAD NORMAL ESTANDAR

La distribución normal estándar es un caso especial de la distribución normal.

Sea Z una v.a con media 0 y desviación estándar de 1, es decir, Z ~ N (0,1), entonces Z es una variable con distribución Normal Estándar.
Cualquier distribución normal puede convertirse a una distribución normal estándar mediante la siguiente fórmula:

$$z = \frac{x - \mu}{\sigma}$$, donde X ~ N (μ,σ^2)

Z ~ N (0,1)

Donde:

- x: es el valor de cualquier observación específica de la distribución N
- μ: es la media de la distribución N
- σ: es la desviación estándar de la distribución N

Estandarizando una distribución normal podemos apreciar la distancia de la media en unidades de la desviación estándar.

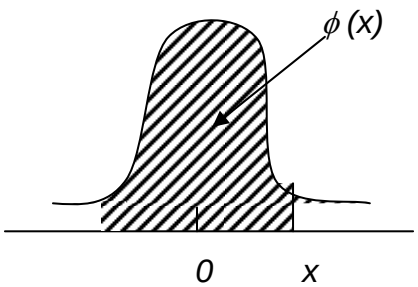
La función de densidad de la distribución de probabilidad normal estándar será:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}x^2}$$, $x \in \Re$;

La función de distribución Acumulada es:

$$\Phi(x) = P[X \le x] = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Gráficamente:



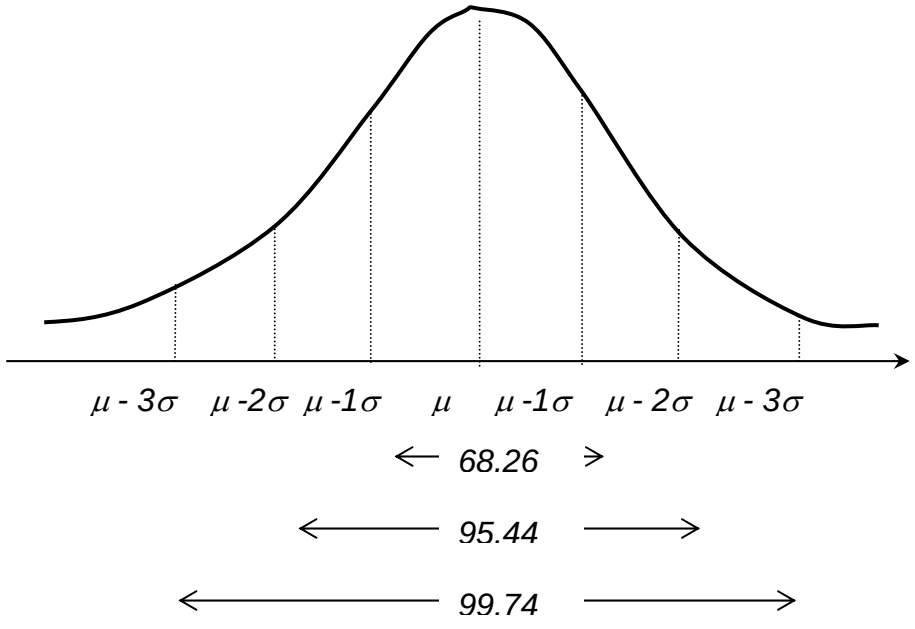
AREAS BAJO LA CURVA

Las áreas bajo la curva normal generalmente se utilizan en tres áreas que son:

1.- Aproximadamente 68% del área bajo la curva normal está dentro de más una y menos una desviación estándar respecto de la media. Esto puede expresarse como $\mu \pm 1\sigma$.

2.- Aproximadamente 95% del área bajo la curva normal está dentro de más dos y menos dos desviaciones estándares respecto a la media, lo que se expresa $\mu \pm 2\sigma$.

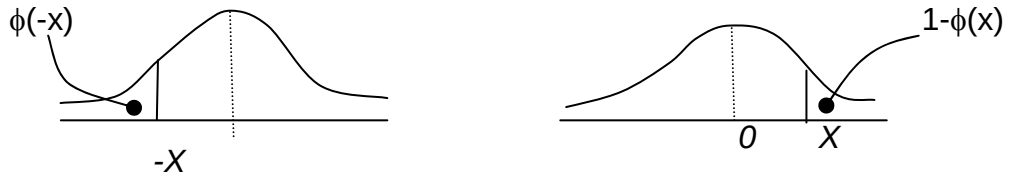
3.- Prácticamente toda el área (99.74%) bajo la curva normal está dentro de tres desviaciones estándares respecto de la media (a uno y otro lado), lo cual se escribe $\mu \pm 3\sigma$.



Otras Propiedades:

1.- Si x tiene distribución N(0,1), entonces para todo x real positivo se cumple:

$$\phi(-x)=1-\phi(x)$$



2.- Si la variable X tiene distribución $N(\mu,\sigma^2)$, entonces la variable Z, definida por $Z = \frac{x-\mu}{\sigma}$, tiene distribución N(0,1) esta propiedad indica lo siguiente: cualquiera que sea los valores de los parámetros de la distribución $N(\mu,\sigma^2)$, ella puede ser transformada a una N(0,1). Según la transformación Z anterior, las distribuciones probabilidades correspondientes a X pueden ser calculados a partir de la distribución de la variable $z = \frac{x-\mu}{\sigma}$, a la que se denomina variable normal estandarizada. Al proceso de transformación aplicado se le denomina "estandarización".

3.- $E(x)=\mu$ y $Var(x)=\sigma^2$

4.- Si la var. X tiene distribución $N(\mu,\sigma^2)$, entonces la variable $Y= ax + b$ tiene distribución $N(a\mu+b, a^2\sigma^2)$

TABLA DE LA DISTRIBUCION NORMAL

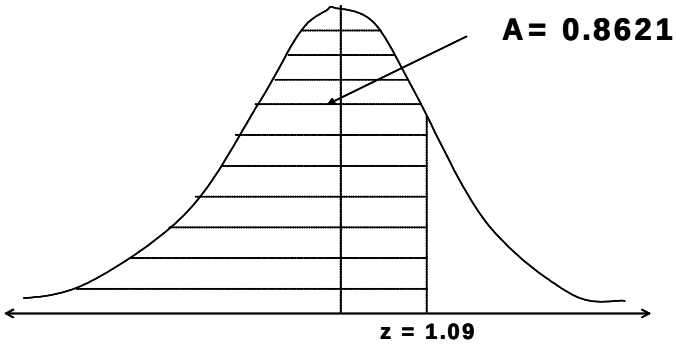
Para valores de x, que varían a intervalos de un centésimo, generalmente desde 0 hasta 3.49, el cuerpo de la tabla presenta valores de $\phi(x)$. Esta tabla se usa de dos maneras :

Uso directo: dado x se halla $\phi(x)$.
Uso inverso: dado $\phi(x)$, hallar x.

X	0.00 0.01 0.02 0.03.....0.09
0.0	
0.1	..... $\phi(0.11)= 0.0438$
.	
.	
1.0	.....$\phi(1.09)= 0.3621$
1.1	..... $\phi(1.13)= 0.3708$
.	
.	
.	
3.4	.....$\phi(3.42)= 0.4997$

Donde:

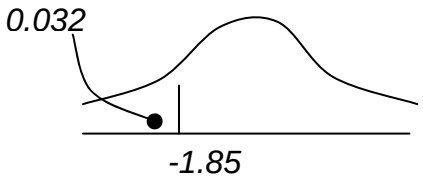
$\phi(0.11)=0.5438$, $\phi(1.09)=0.8621$, $\phi(1.13)=0.8708$, $\phi(3.42)=0.9997$



$\phi(1.09)= 0.8621$

Ejemplos: sea X ~N(0.1). dado x, hallar $\phi(x)$

1.- Para x= -1.85, $\phi(-1.85)= 1 - \phi(1.85)= 1 - 0.9678=0.0322$
Gráficamente:



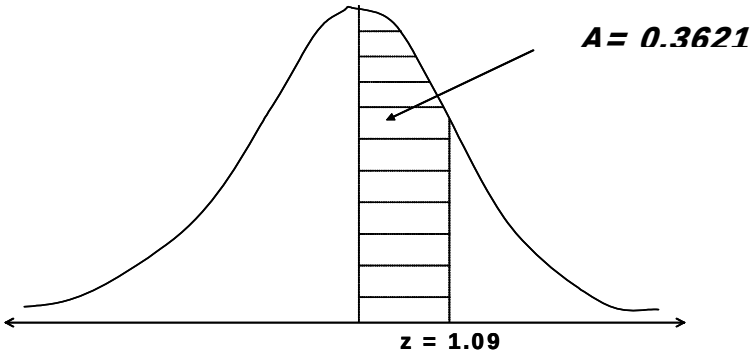
2.- Para x= 2, $P(X \leq 2)= \phi(2)= 0.9772$

3.- $P(0 \leq X \leq 2)= \phi(2) - \phi(0)= 0.9772-0.50=0.4772$

OTRA FORMA DE VER LA TABLA:

X	0.00 0.01 0.02 0.03.....0.09
0.0	
0.1	..... $\phi(0.11)= 0.0438$
.	
.	
1.0	.....$\phi(1.09)= 0.3621$
1.1	..... $\phi(1.13)= 0.3708$
.	
.	
.	
3.4	.....$\phi(3.42)= 0.4997$

Donde: $\phi(0.11)=0.0438$, $\phi(1.09)=0.3621$, $\phi(1.13)=0.3708$, $\phi(3.42)=0.4997$



$\phi(1.09)= 0.3621$

APLICACIONES DE LA DISTRIBUCION NORMAL

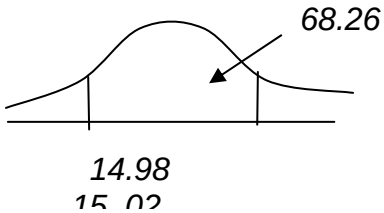
Ejemplo 1: Se sabe que el diámetro de ciertos rodamientos producidos por una máquina, sigue una distribución normal con media μ=15 cm. Y σ=0.02 cm. para que el rodamiento sea considerado como no defectuoso, su diámetro debe variar entre 14.98 y 15.02 cm.

a. ¿Cuál es la probabilidad de que un rodamiento escogido al azar sea defectuoso?
b. ¿Qué probabilidad existe de que en un una caja de 50 rodamientos producidos hayan exactamente tres rodamientos defectuosos?

Solución:

Sea la variable aleatoria X que indica la longitud de los diámetros
a. La probabilidad de que cada rodamiento sea no defectuoso es:

$$P[14.98 \le X \le 15.02] = P[-1 \le Z \le 1] = 0.6826$$



Luego la probabilidad de que un rodamiento sea defectuoso es:

1 - 0.6826=0.3174

b. Sea la variable aleatoria Y que denota el número de rodamientos defectuosos en una caja de 50 unidades. Entonces, Y tiene distribución binomial con parámetros n=50 y p=0.3174.

Luego, $P[Y = 3] = \binom{50}{3}(0.3174)^3(0.6826)^{47} \approx 0$

Ejemplo 2: El tiempo de duración de un foco de luz está normalmente distribuido, con una duración media de 800 horas y una desviación estándar de 200 horas. Se compran 500 de

estos focos. ¿Cuál es la distribución de probabilidad del número de focos que estarán en servicio después de 1000 horas?

Solución:

Sea la variable X: Tiempo de duración de los focos, $X \sim N(800,200^2)$. Que el foco esté en servicio después de 1000 horas, significa que el tiempo de duración X, sea mayor que 1000 horas.

Entonces, $$P(X > 1000) = 1 - P(X \leq 1000) = 1 - P\left[\frac{X - 800}{200} \leq \frac{1000 - 800}{200}\right]$$

$$= 1 - P(Z \leq 1) = 1 - \phi(1) = 1 - 0.8413 = 0.1587$$

De otro lado, si la producción de focos es muy grande, los 500 focos que se compran constituyen ensayos independientes de Bemoulli, con probabilidad de éxito igual P(x>1000). Entonces, el número de focos que duran más de 1000 horas es una nueva variable aleatoria, Y, con distribución binomial cuyos parámetros son n=500 y p=0.1587.

EJERCICIOS RESUELTOS

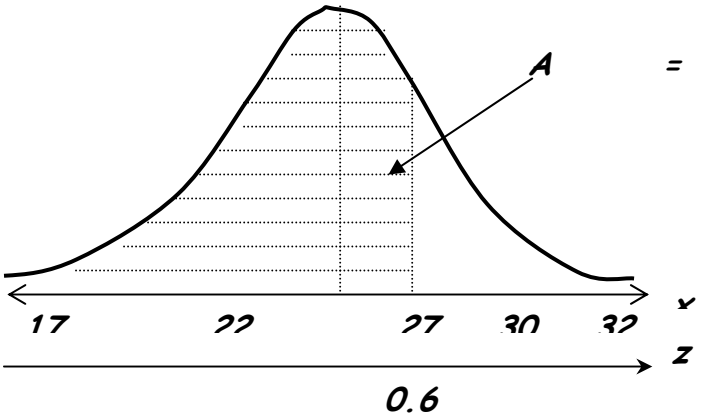
Ejercicio 1: El tiempo empleado de ir de un Centro de Salud al Hospital por la ruta A se distribuye normalmente con media igual a 27 y desviación típica igual a 5; mientras que por la ruta B, la distribución es normal con media igual a 30 y desviación típica igual a 2. ¿Qué ruta conviene utilizar si se dispone de: a. 30 minutos? b. 34 minutos? c. En cuál de las rutas se tiene la mayor probabilidad de llegar antes de 30 minutos?

Solución: **Ruta A: μ= 27 σ= 5** **Ruta B: μ= 30 σ= 2**

a. ¿Qué ruta conviene utilizar si se dispone de 30 minutos?

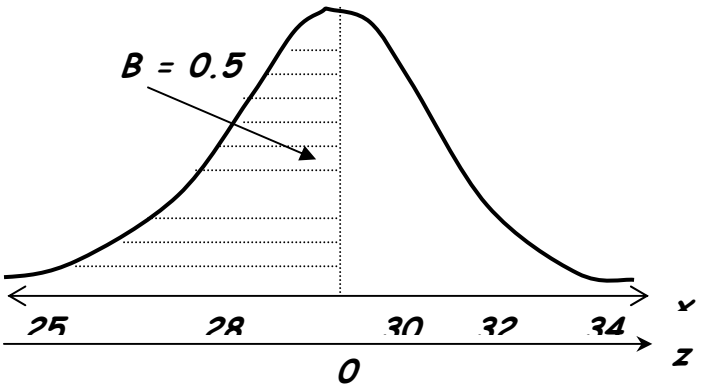
$$z_A = \frac{30-27}{5} = \frac{3}{5} = 0.6$$

P(x<30) = P(z<0.6) =0.5000+A(0.6) = 0.50000+0.22575 =0.72575 P(x<30) = 72.58%



$$z_B = \frac{30-30}{2} = \frac{0}{2} = 0$$

P(x<30) = P(z<0) =0.50

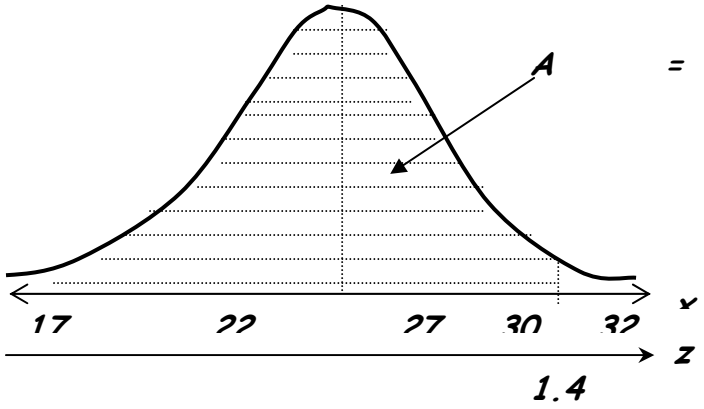


Como interesa el porcentaje favorable mayor para P(x<30) que significa llegar temprano, en este caso conviene elegir la ruta A que da un valor para P (x<30) 0.7528 = 72.58%

b. ¿Qué ruta conviene utilizar si se dispone de 34 minutos?

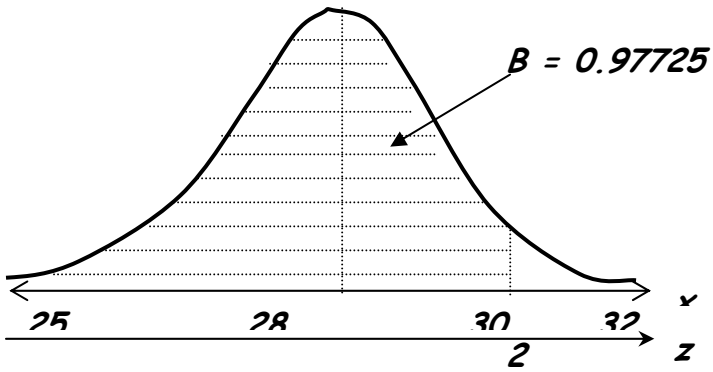
$$z_A = \frac{34-27}{5} = 1.4$$

P (x<34) = P (z<1.4) = 0.5000 + 0.41924 = 0.91924



$$z_B = \frac{34-30}{2} = \frac{4}{2} = 2$$

P (x<34) = P (z<2.0) =0.5000 + 0.47725 = 0.97725



En este caso, elegimos la ruta B, por que representa el porcentaje favorable mayor para llegar temprano, ya que da un valor P (x<34) = 0.9772 = 97.72%

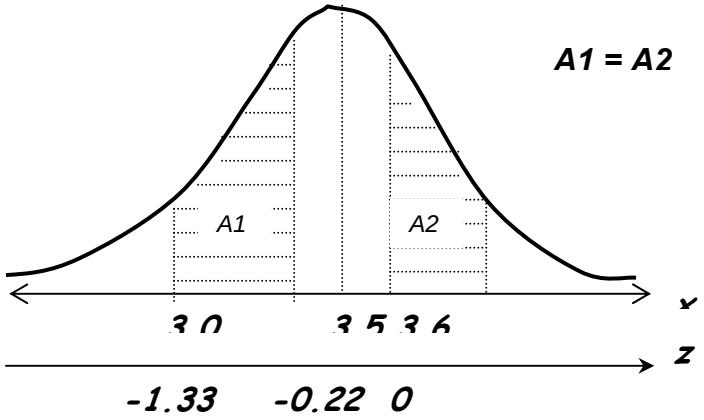
Ejercicio 2: En cierta clínica, el salario medio de los médicos es de $ 3.60 por hora y la desviación estándar es de 45 centavos de dólar. Si se supone que los salarios tienen una distribución normal, ¿Qué porcentaje de de médicos percibe salarios entre 3.00 y 3.500 por hora?

Solución:

Se sabe que: **μ= 3.6** **σ= 0.45**

$$z = \frac{3.00-3.60}{0.45} = \frac{-0.60}{0.45} = \frac{-4}{3} = -1.33 \qquad z = \frac{3.50-3.60}{0.45} = \frac{-0.10}{0.45} = \frac{-2}{9} = -0.22$$

La solución es el área comprendida entre los valores de z = -0.22 y z = -1.33. Por simetría podemos calcular esa área:



Área pedida: A (1.33) – A (0.22) = 0.4082 – 0.0871 = 0.3211 = 32.11%

Ejercicio 3: En una población de 3,428 adultos, la distribución de las estaturas es aproximadamente normal, con media 140 centímetros y desviación estándar de 25 centímetros. Calcule el número de dichas personas con estatura:

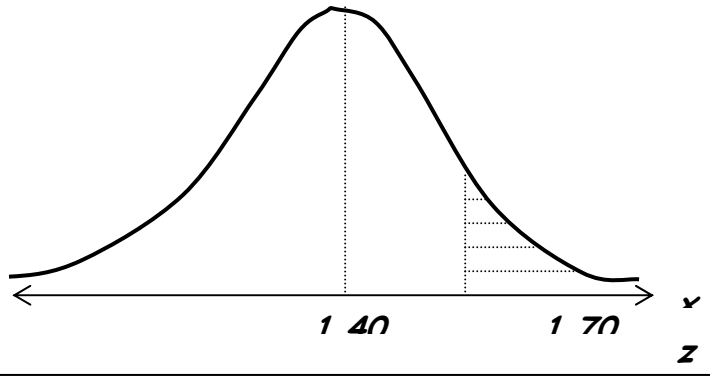
a. Superior a 170 centímetros
b. Inferior a 90 centímetros
c. Comprendida entre 1 metro y 1.50 metros
d. Comprendida entre 1.80 m. y 190 cms.
e. Comprendida entre 1 m. y 130 cms.
f. ¿Entre que valores queda ubicado el 40% central? El 68.26? el 95.44? y el 99.74%?

Solución:
Población adultos: 3,428 **μ= 140 cm.** σ= **25 cm.**

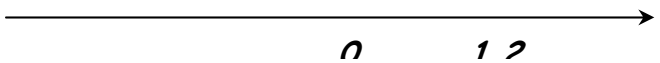
a. Superior a 170 centímetros

$$z = \frac{170-140}{25} = \frac{30}{25} = 1.2$$

El área que buscamos será: 0.5 – A(1.2)
A (1.2) = 0.38493
0.5000 - 0.3849 = 0.1151 = 11.51%
El 11.51% de 3428 es 395 personas, es decir que hay 395 personas con estatura superior a 1.70 cm.



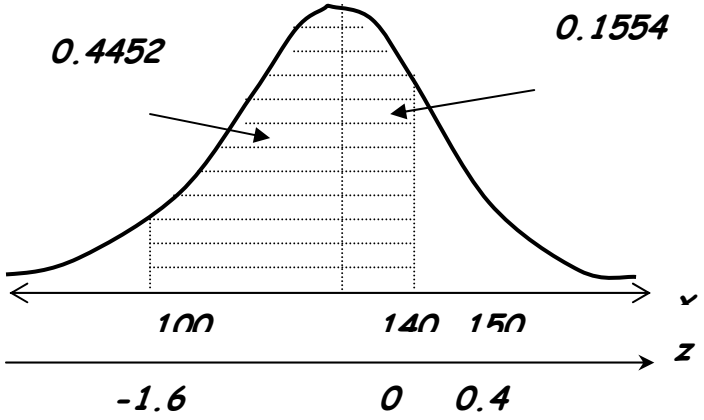
c. Comprendida entre 1 m. y 1.5 m.



P (1 < x < 1.5) = ?

$$z = \frac{100-140}{25} = \frac{-40}{25} = -1.6 \qquad z = \frac{150-140}{25} = \frac{10}{25} = 0.4$$

El área que buscamos es: A (-1.6) + A (0.4) = A (1.6) + A (0.4)



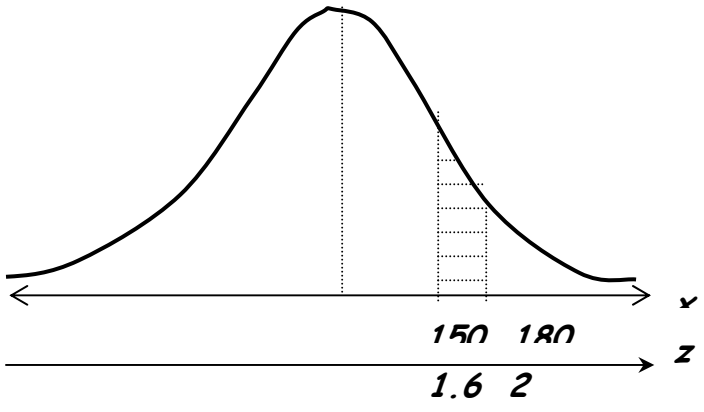
El área buscada es = 0.4452 + 0.1554 = 0.6006 = 60.06%
El 60.60% (2,059 personas) tienen una estatura comprendida entre 1m. y 1.5 m.

d. Comprendida entre 1.8 m. y 1.9 m.

P (1.8 < x < 1.9) = ?

$$z = \frac{180-140}{25} = \frac{40}{25} = 1.6 \qquad z = \frac{190-140}{25} = \frac{50}{25} = 2$$

El área que necesitamos encontrar es: A (2) - A (1.6)



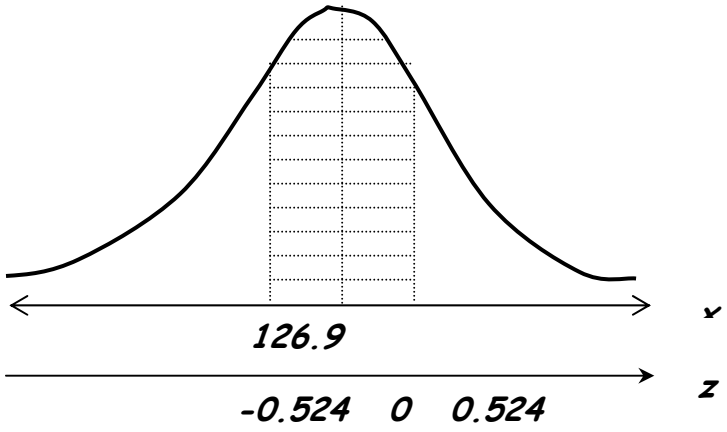
El área buscada es = 0.47725 - 0.44520 = 0.03205 = 3.2%
El 3.2% (110 personas) tienen una estatura comprendida entre 1.8m. y 1.9 m.

f. ¿Entre que valores queda ubicado el 40% central? Buscando el área 0.200 en la tabla encontramos para z el valor 0.52 = 0.1985 Y 0.53 = 0.2019, que interpolado resulta 0.524.

$$z = \frac{x_i - \mu}{\sigma} \quad \Rightarrow \quad x_i = z\sigma + \mu$$

x_i = **0.524 (25) +140 =13.10 + 140** x_i = **153.10** a la derecha

x_i = **-0.524 (25) +140 = -13.10 + 140** x_i = **126.90** a la izquierda



El 40% central queda ubicado entre 126.90 y 153.10

EJERCICIOS

1. Enumere las principales características de una distribución probabilística normal

Solución:

Las principales características son:

a. La curva normal tiene perfil de campana y presenta un solo pico en el centro exacto de la distribución.
b. La distribución probabilística normal es simétrica con respecto a su media.
c. La curva normal decrece uniformemente en ambas direcciones a partir del valor central

2. La media de una distribución probabilística normal es 60, y la desviación estándar es 5.

a. ¿Aproximadamente qué porcentaje de las observaciones se encuentra entre 55 y 65?
b. ¿Aproximadamente qué porcentaje de las observaciones se halla entre 50 y 70?
c. ¿Aproximadamente qué porcentaje de las observaciones se halla entre 45 y 75?

Solución:

Sea: μ=60 y σ=5

a. X: v.a. que indica porcentaje de las observaciones
   Sabemos lo siguiente:
   68% $\sim\mu \pm 1\sigma$, si probamos los valores tenemos:
   60-5<x<60+5
   55<x<65
   Por lo tanto: aproximadamente 68% de las observaciones está entre 55 y 65
b. Sabemos que :
   95%$\sim\mu \pm 2\sigma$,
   Remplazando μ y σ tenemos:
   60-2(5)<x<60+ 2(5)
   50<x<70, por lo tanto aproximadamente 95%de la observaciones están entre 50 y 70
c. Sabemos que 99.74%$\sim\mu \pm 3\sigma$
   Remplazando μ y σ tenemos
   60-3(5)<x<60+3(5)

3. Las cantidades de dinero en solicitudes de préstamo para causas sociales que recibe una institución, está aproximadamente distribuida en forma normal con una media de $70000 y una desviación estándar de $ 20 000. Una solicitud de préstamo se recibió esta mañana.

¿Cuál es la probabilidad de que

a. la cantidad solicitada sea de $ 80 000 o más?
b. la cantidad solicitada esté entre $ 65 000y80 000?
e. La cantidad solicitada sea de $ 65000 o más?
f. 20% de los préstamos sean mayores que cuál cantidad

4. La investigación sobre nuevos delincuentes juveniles que fueron puestos en libertad bajo palabra por un juez, reveló que 38% cometieron otro delito.

a. ¿Cuál es la probabilidad de que los últimos 100 nuevos delincuentes juveniles puestos en libertad bajo palabra, 30 o más delincan otra vez?
b. ¿Cuál es la probabilidad de que 40 o menos de los delincuentes cometan otro delito?
c. ¿Cuál es la probabilidad de que entre 30 y 40 de los delincuentes cometerán otro acto ilícito?

5. Pruebas realizadas en ampolletas eléctricas de la marca "X", indican que el período de duración se distribuye normalmente con media igual a 1.9 horas y desviación estándar igual a 65 horas. Estimar el porcentaje de ampolletas que se espera que duren:

a. Más de 2.2 horas
b. Menos de 1.8 horas.

6. El tiempo empleado, en minutos, en ir de un pueblo a una centro de salud por la ruta R se distribuye normalmente con media igual a 29 y desviación típica igual a 8; mientras que por la ruta M, la distribución es normal con media igual a 32 y desviación típica igual 3. ¿Qué ruta es conveniente usar si se dispone de:
   a. 30 minutos
   b. 35 minutos

7. En la clínica "X" , el salario medio del personal médico es de 20.0 n.s. por hora y la desviación típica es de 0.8 n.s.. Si se sabe que los salarios presentan una distribución normal ¿Que porcentajes del personal percibe salarios entre 15.0 y 18.0 por hora.

8. En una población de 2,548 adultos, la distribución de las estaturas es aproximadamente normal, con media 150 cms. y desviación estándar 27 centímetros. Calcule el número de dichas personas con estatura:
   a. Inferior a 95 centímetros
   b. Superior a 180 centímetros
   c. Comprendida entre 170 cms. Y 180 cms.

CAPITULO NUEVE

PRUEBAS DE HIPOTESIS

1. HIPOTESIS:

Son supuestos o enunciados que pueden o no ser verdaderas, relativas a una o más poblaciones y pueden ser:

(a) Hipótesis nula: Denotada con H_O, determina supuestos o conjeturas de la población o poblaciones bajo estudio, con el propósito de rechazar. En ella se indica que no hay cambios, que no hay diferencias o se propone un modelo teórico determinado. Por lo común es una afirmación de que el parámetro de población tiene un valor específico.

(b) **Hipótesis alternativa:** Denotado con H_1, determina supuestos o conjeturas de la población o poblaciones bajo estudio con el propósito de no rechazarla.

Afirmación o enunciado que se aceptará si los datos muestrales proporcionan amplia evidencia de que la hipótesis nula es falsa.

2. CLASES DE HIPOTESIS

Las hipótesis son:

(a) **Hipótesis simples**, es la hipótesis que da valores exactos para todos los parámetros desconocidos de la ley de probabilidad asumida.

(b) **Hipótesis compuesta**, es la hipótesis que no da valores exactos, sino tiene un conjunto de valores para todos los parámetros desconocidos de la ley de probabilidad asumida. Se refiere a regiones de valores.

Prueba de hipótesis: Es un procedimiento basado en la evidencia muestral y en la teoría de probabilidad que se emplea para determinar si la hipótesis es un enunciado razonable y no debe ser rechazada, o si es irrazonable y debe ser rechazada.

3. PROCEDIMIENTO DE CINCO PASOS PARA PROBAR UNA HIPOTESIS

Paso 1: Plantear Hipótesis nula y Alternativa

Paso 2: Seleccionar un Nivel de significación

Paso 3: Identificar el Valor estadístico de prueba

Paso 4: Formular una regla de decisión

Paso 5: Tomar una muestra y llegar a una decisión

Finalmente: Aceptar H_0, o bien rechazar H_0 y aceptar H_1

Nivel de significación: El riesgo que se asume acerca de rechazar la hipótesis nula cuando en realidad debe aceptarse por ser verdadera.

4. TIPOS DE ERROR

Error Tipo I: Se refiere a la probabilidad de rechazar la hipótesis nula, H_0, cuando en realidad es verdadera. Se busca minimizar este tipo de afirmación.
1- α: Se refiere a la probabilidad de no rechazar la hipótesis nula, H_0, cuando en realidad es verdadera. Se busca maximizar este tipo de error.
Error tipo II: Se refiere a la probabilidad de aceptar la hipótesis nula, H_0 cuando en realidad es falsa. Este tipo de error busca aceptar lo que espero que no se acepte.
1- β: Se refiere a la probabilidad de rechazar la hipótesis nula, H_0, cuando en realidad es falsa. No se busca maximizarlo por que nunca se va aceptar la H_0.
Lo ilustramos mejor en el siguiente cuadro:

Hipótesis Nula	**El investigador**	
	No Rechazar H_O	**Rechaza H_O**
Si H_O es verdadera	Decisión Correcta =$(1-\alpha)$	Error Tipo I = α **Nivel de significación**
Si H_O es falsa	Error Tipo II = β	Decisión Correcta = $(1-\beta)$ **Potencia**

Por tanto las probabilidades de error tipo I y tipo II están dadas por las siguientes proposiciones:

α= P (rechazar H_0 | H_0 es cierta) $\Rightarrow 1-\alpha$ = P (no rechazar H_o / H_0 verdadera)

β = P (no rechazar H_0 | h_0 es falsa) $\Rightarrow 1-\beta$ = P (rechazar H_0 / H_0 es falsa)

5. FUNCIÓN POTENCIA

Representa la probabilidad de rechazar la hipótesis nula cuando ésta es falsa, es decir, cuando el valor del parámetro de H_1 (hipótesis alternativa) es cierto. Esta definida mediante la siguiente función:

$$\mathbf{P(\theta) = 1 - \beta(\theta)}$$

Donde:
θ : Es el parámetro de interés
$\beta(\theta)$: Función característica del error tipo II

La potencia de una prueba es detectar que H_0 es, realmente falsa, de aquí el uso de la palabra **"potencia"**.

Valor estadístico de prueba: Es un valor determinado a partir de la información muestral, que se utiliza para aceptar o rechazar la hipótesis nula.
El valor estadístico de prueba que se utilizará ahora es el llamado valor Z
(o desvío normal),que se determina a partir de datos muestrales:

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Donde:

σ : Desviación estándar de la población
n : Número en la muestra
$\overline{X}$: Media muestral
μ : Media poblacional

Para tamaños de muestra n < 30, no se necesita un análisis preliminar del valor estadístico de la variable, pero para un tamaño de muestra n > 30, si se necesitaría un análisis preliminar de las variables.

6. TOMA DE UNA DECISION

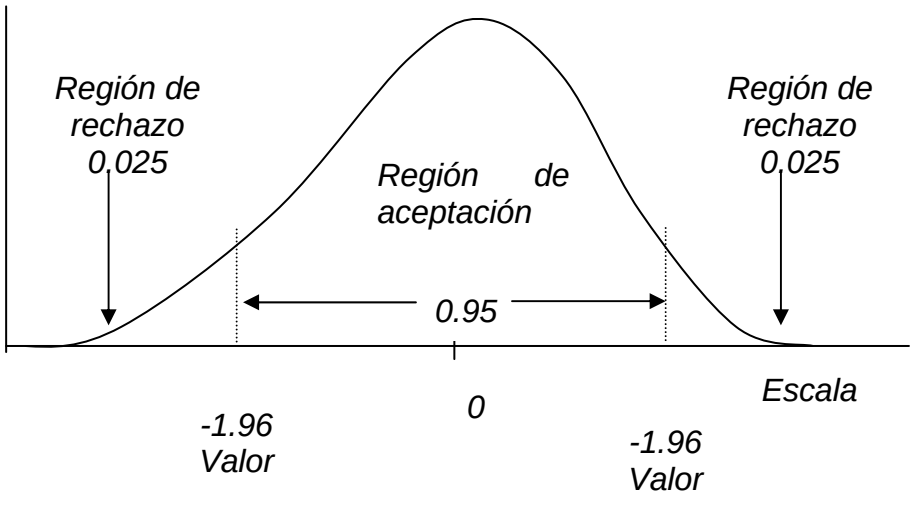
Es la de afirmar que no hay evidencias suficientes para "rechazar" la hipótesis nula. La hipótesis nula se rechaza en el nivel de significación 0.05. Se toma la decisión de rechazar $H_{0'}$ debido a que 2.34 se encuentra en la región de rechazo, es decir, más allá de 1.645, si hubiera sido calculado el valor igual a 1.645 o menor, la hipótesis nula sería aceptada.

7. PRUEBA DE DOS COLAS

Si no se especifica la dirección según la hipótesis alternativa, se aplica una prueba de dos colas o extremidades. Como ejemplo tenemos:

H_0: No hay diferencia entre las cantidades de asistencia técnica oferta y demanda.
H_1: Hay una diferencia entre las cantidades de oferta y demanda

Si se rechaza la hipótesis nula y se acepta H_1 podría la cantidad de asistencia técnica oferta ser mayor que la demanda o viceversa. Para dar cabida a estas dos posibilidades, el 5% que representa el área de rechazo se divide por igual en las dos colas de la distribución muestral.



8. PRUEBAS PARA LA MEDIA DE POBLACION

8.1. MUESTRA GRANDE Y SE CONOCE LA DESVIACION ESTANDAR DE LA POBLACION

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0.1)$$

H_0	H_1	R_c
$\mu=\mu_o$	$\mu=\mu_1$ ($<\mu_o$)	$Z<-z_{1-\alpha}$
	$\mu \neq \mu_1$ ($>\mu_o$)	$Z> z_{1-\alpha}$
	$\mu\neq \mu_o$	$Z<-z_{1-\alpha/2}$ ó $Z> z_{1-\alpha/2}$ $\lvert Z\rvert> z_{1-\alpha/2}$
$\mu\geq\mu_o$	$\mu<\mu_o$	$Z<-z_{1-\alpha}$
$\mu\leq\mu_o$	$\mu>\mu_o$	$Z> z_{1-\alpha}$

Ejemplo: Un comprador de ladrillos cree que la calidad de los ladrillos está disminuyendo. De experiencias anteriores, la resistencia media al desmoronamiento de tales ladrillos es 200 Kg, con una desviación típica de 10 Kg. Una muestra de 100 ladrillos arroja una media de 195 Kg. Probar la hipótesis. La calidad media no ha cambiado, contra la alternativa que ha disminuido.

Solución:

1. H_0 : μ = 200kg. Y H_1 : μ < 200 kg.

2. Escogemos el nivel de significación α = 0.05

3. La estadística de prueba es X . Desde que la muestra es grande n = 100, la distribución de $\overline{X}$ es:

$$N\left(200, \frac{10^2}{100}\right) = N(200, 10)$$

(Teorema central del límite). Luego,

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

4. R .C. =<-∞, $\bar{x}_C$ > donde $\bar{x}_c$ tal que $P\left[\bar{X} < \bar{x}_c / H_o\right] = \alpha$

ó $P\left[\dfrac{\bar{X} - \mu_O}{\sigma/\sqrt{n}} < \dfrac{\bar{x}_c - 200}{10/10}\right] = P\left[Z < \bar{x} - 200\right] = 0.05$

de donde $Z_\alpha = \bar{x}_c - 200 = -1.64$, luego $\bar{x}_c = 198.36$

R.C. = < - ∞, 198.36>.

5. Cálculo de media muestral: del enunciado una muestra de n = 100, da $\bar{x} = 195$.

6. Conclusión: Puesto que $\bar{x} = 195 \in$ R.C. = < - ∞, 198.36>. Rechazamos H_0.

8.2. MUESTRA GRANDE Y SE DESCONOCE LA DESVIACION ESTANDAR DE LA POBLACION

$$T = \frac{X - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$$

H_0	H_1	R_c
$\mu=\mu_o$	$\mu=\mu_1$ ($<\mu_o$)	$T<-t_{1-\alpha}$
	$\mu=\mu_1$ ($>\mu_o$)	$T> t_{1-\alpha}$
	$\mu\neq\mu_o$	$T<-t_{1-\alpha/2}$ o $T> t_{1-\alpha/2}$
$\mu\geq\mu_o$	$\mu<\mu_o$	$T<-t_{1-\alpha}$
$\mu\leq\mu_o$	$\mu>\mu_o$	$T> t_{1-\alpha}$

Ejemplo: Una máquina para enlatar conservas de pescado ha sido regulando para que el contenido de una lata sea 16 onzas. Usando α = 0.05, ¿Diría Ud. Que la máquina ha sido adecuadamente regulada, si una muestra de 20 latas dio un peso medio de 16.05 onzas y una desviación típica de 1.5 onzas?.

Solución:

1. $H_o : \mu = 16$ y $H_1 : \mu \neq 16$

2. α = 0.05

3. Puesto que n = 20 es pequeño y suponiendo que la población tiene distribución aproximadamente normal, usamos la variable aleatoria.

T = $\frac{\overline{X} - \mu_0}{s/\sqrt{n}}$ que tiene una distribución t con n – 1 = 19 grados de libertad como estadística de prueba.

4. Región Crítica: T < - $t_{\alpha/2}$ = - 2.093 ó T > $t_{\alpha/2}$ = 2.093, con $\alpha/2$ = 0.025, para buscar en la tabla tomamos
1 - $\alpha/2$ = 0.975 luego, R.A. = < - 2.093, 2.093>

5. De los datos $\overline{x}$ = 16.05, s = 1.5 para n = 20, entonces

t = $\frac{\overline{x} - \mu_0}{s/\sqrt{n}}$ = $\frac{16.05 - 16}{1.5/\sqrt{20}} \cong$ 0.148

6. Conclusión: desde que t = 0.148 $\in$ R.A, aceptamos H_o ; es decir se acepta que la maquina ha sido adecuadamente regulada.

9. PRUEBA DE HIPÓTESIS PARA UNA PROPORCION

Las pruebas de hipótesis con relación a proporciones son básicamente iguales a las relativas con medias. Para probar la hipótesis de la proporción se usa la siguiente estadística de prueba:

$$Z = \frac{\hat{P} - P_O}{\sqrt{\frac{p_o q_o}{n}}} \sim N(0.1)$$

Donde: $\hat{P}$ = Proporción Muestral.

P_0 = Proporción Poblacional.

$\sqrt{\frac{P_0 Q_0}{n}}$ = Error Estándar de la Proporción Poblacional.

H_0	H_1	Rc
$p = p_0$	$p = p_1$ ($< p_0$)	$Z < -z_{1-\alpha}$
	$p = p_1$ ($> p_0$)	$Z > z_{1-\alpha}$
	$p \neq p_0$	$Z < -z_{1-\alpha/2}$ o $Z > z_{1-\alpha/2}$
$p \geq p_0$	$p < p_0$	$Z < -z_{1-\alpha}$
$p \leq p_0$	$p > p_0$	$Z > z_{1-\alpha}$

Ejemplo: Un investigador afirma que el 80% de los niños de un PP.JJ trabajan. ¿Cree Ud. Que tal afirmación es cierta si, en una encuesta aplicada a ese PP. JJ resulta que el 88% de niños trabajan? (n = 1000).

Solución:

1.- Ho : p=0.8
H1 : p≠ 0.8

2.- $\propto$ =0.01

3.- La estadística para la prueba es:

$$Z = \frac{\hat{P} - P_O}{\sqrt{\frac{p_o q_o}{n}}}$$ ~ N(0.1) y tiene Rc:

zc <-2.58 y zc >2.58

4.- $\hat{P}$ =0.88 n=1000

$$Z = \frac{0.88 - 0.8}{\sqrt{\frac{(0.8)(0.2)}{1000}}} = 0.08$$

5.- Como Z = 0.08 < Zc = 2.58, entonces no se rechaza H_0, es decir, no hay pruebas suficientes para afirmar que el 80% de los niños en un PP.JJ. trabajan.

10. PRUEBA DE HIPOTESIS SOBRE LA DIFERENCIA ENTRE MEDIAS

Para muchos problemas prácticos es interesante determinar si existe o no una diferencia significativa entre las medias μ_x y μ_y de dos poblaciones de las variables X y Y. La prueba de hipótesis para dos medias, tiene la misma aplicación que la de una sola media, salvo que se necesitan dos muestras de cada población.

10.1. PRUEBA DE DIFERENCIA DE MEDIAS CON $\sigma_1 = \sigma_2$ PERO DESCONOCIDAS, EN MUESTRAS PEQUEÑAS

Si se quiere probar la hipótesis sobre la diferencia de medias, cuando los tamaños de las muestras son pequeños y las poblaciones tiene distribuciones normales, con varianzas iguales, se utiliza la siguiente estadística de prueba:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_o}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

Donde:

$$S^2{}_p = \frac{(n_1 - 1)(S^2{}_1) + (n_2 - 1)(S^2{}_2)}{n_1 + n_2 - 2}$$

Que tiene una distribución t con n_1+n_2 –2 grados de libertad.

La prueba de hipótesis que se desean probar son:

H_0	H_1	Rc
$\mu_1 - \mu_2 = d_o$	$\mu_1 - \mu_2 < d_o$	$T < -t_{1-\alpha}$
	$\mu_1 - \mu_2 > d_o$	$T > t_{1-\alpha}$
	$\mu_1 - \mu_2 \neq d_o$	$T < -t_{1-\alpha/2}$ o $T > t_{1-\alpha/2}$

Nota: do es una cantidad que toma valores positivos o cero y la cual representa la diferencia propuesta entre los valores desconocidos de las medias.

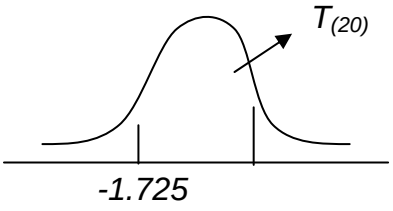
Ejemplo: En un estudio, se ha determinado que en la región 1 las familias gastan un promedio de 85 u.m. por consumo eléctrico, con una desviación estándar $S_1 = 4$ de una m.a de 12 familias. En la región 2 se ha tomado una m.a de 10 familias y se ha encontrado que las familias gastan un promedio de 81 u. m., y con una desviación estándar $S_2 = 5$, verifique si en las dos regiones las familias presentan el mismo consumo promedio. Suponga que las poblaciones son aproximadamente normales y tiene varianzas poblacionales iguales.
Solución:

1.- Ho : $\mu_1 = \mu_2$ ($\mu_1 - \mu_2 = 0$)
H1 : $\mu 1 \neq \mu 2$ ($\mu 1 - \mu 2 \neq 0$)

2.- $\alpha = 0.01$

3.- La estadística para la prueba es:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_o}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$ y tiene RC:



T- 1.725 y T> 1.725

4.- $\bar{X}_1$ =85 $\bar{X}_2$ =81

s_1 =4 s_2 = 5

n_1 =12 n_2 = 10

$$S_p = \sqrt{\frac{11(16)+9(25)}{12+10-2}} = 4.478$$

$$T = \frac{(85-81)-0}{4.478\sqrt{\frac{1}{12}+\frac{1}{10}}} = 2.07$$

5.- Se rechaza H_O y se concluye que las dos regiones no presentan el mismo consumo promedio de la electricidad.

10.2. PRUEBA DE DIFERENCIA DE MEDIAS, CON $\sigma_1 \neq \sigma_2$ Y DESCONOCIDA EN MUESTRAS PEQUEÑAS

Si se quiere probar la hipótesis sobre la diferencia de medias, cuando los tamaños de las muestras son pequeños y las poblaciones tienen distribuciones normales, con varianzas diferentes, se utiliza la siguiente estadística de prueba:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_o}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}} \sim t(k)$$

Que tiene una distribución t con K grados de libertad.

$$k = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[s_1^2/n_1\right]^2}{n_1+1} + \frac{\left[s_2^2/n_2\right]^2}{n_2+1}} - 2$$

Siendo:

Si $k \geq 30$, el estadístico indicado sigue aproximadamente una ley normal estándar y el procedimiento a seguir es como en el caso donde se conocen las varianzas.

11. PRUEBA DE HIPÓTESIS RELATIVA A LAS VARIANZAS DE DOS POBLACIONES

El procedimiento en la prueba de comparación de varianzas es el mismo que las pruebas de una sola varianza. Excepto que la estadística de prueba es la variable aleatoria y tiene la siguiente forma:

$$\boxed{F = \frac{s^2_1}{s^2_2}} \sim F_{(n_1-1,n_2-2)}$$

11.1. IDENTIDAD FUNDAMENTAL

Un estimador insesgado de la varianza de la población se obtiene combinando varias varianzas muestrales, las tres varianzas muestrales se pueden expresar por:

$$S_1^2 = \frac{\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.})^2}{n_1 - 1}, \qquad S_2^2 = \frac{\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.})^2}{n_2 - 1}, \qquad S_3^2 = \frac{\sum_{j=1}^{n_3}(x_{3j} - \bar{x}_{3.})^2}{n_3 - 1}$$

y la varianza combinada es:

$$\hat{\sigma}_c^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + (n_3 - 1)S_3^2}{n_1 + n_2 + n_3 - 3}$$

$$\hat{\sigma}_c^2 = \frac{\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.})^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.})^2 + \sum_{j=1}^{n_3}(x_{3j} - \bar{x}_{3.})^2}{n-3} = \frac{\sum_{i=1}^{S}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})}{n-3} \quad \text{.....(1)}$$

Una segunda manera de estimar la varianza σ_2 es mediante la relación:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

Que se convierte en:

$$\sigma^2 = n\sigma^2_{\bar{x}}$$

Por consiguiente, estimando $\sigma^2_{\bar{x}}$ podemos estimar σ_2.
Para la primera muestra tenemos:

$$\sigma^2 = n_1\sigma^2_{\bar{x}_1} = n_1(\bar{x}_{1.} - \bar{x})^2$$

Para la segunda y terceras muestras, tenemos:

$$\sigma^2 = n_2\sigma^2_{\bar{x}_2} = n_2(\bar{x}_{2.} - \bar{x})^2$$

$$\sigma^2 = n_3\sigma^2_{\bar{x}_3} = n_3(\bar{x}_{3.} - \bar{x})^2$$

Por tanto:

$$3\sigma^2 = \sum n_i(\bar{x}_{i.} - \bar{x})^2 \quad \Rightarrow \quad \sigma^2 = \frac{1}{3}\sum n_i(\bar{x}_{i.} - \bar{x})^2$$

Para una estimación insesgada, utilizamos los grados de libertad 3 -1 = 2 en vez de 3. Entonces la segunda forma de estimar σ_2 es:

$$\hat{\sigma}^2 = \frac{1}{3-1}\sum_{i=1}^{3} n_i(\bar{x}_{i.} - \bar{x})^2 \quad \text{.....(2)}$$

Dado H_0: $\mu_1=\mu_2=\mu_3$, hay una tercera forma de estimar la varianza de la población $\sigma 2$. Debido a esta hipótesis podemos considerar las 3 muestras juntas como una gran muestra de tamaño $n = n_1+n_2+n_3$. En este caso, un estimador insesgado de σ_2 es:

$$\hat{\sigma}^2 = \frac{\sum\sum(x_{ij} - \bar{x})^2}{n-1} \quad \text{.....(3)}$$

Esto nos lleva a que podamos establecer una relación entre los numeradores de los tres estimadores (1), (2) y (3) de $\sigma 2$, como sigue:

$$\underbrace{\sum\sum(x_{ij} - \bar{x})^2}_{\text{Total}} = \underbrace{\sum\sum(x_{ij} - \bar{x}_{i.})^2}_{\text{Dentro}} + \underbrace{\sum\sum(\bar{x}_{i.} - \bar{x})^2}_{\text{Entre}} = \sum\sum(x_{ij} - \bar{x}_{i.})^2 + \sum\sum n_i(\bar{x}_{i.} - \bar{x})^2 \ldots(4)$$

Donde:

$\sum\sum(x_{ij} - \bar{x})^2$: suma total de las desviaciones al cuadrado.

$\sum\sum(x_{ij} - \bar{x}_{i.})^2$: suma de cuadrados dentro de los grupos.

$\sum\sum n_i(\bar{x}_{i.} - \bar{x})^2$: suma de cuadrados entre los grupos.

La fórmula (4) indica que la suma total de desviaciones al cuadrado está dividida en dos partes. Expresada en términos del análisis de regresión, la suma total de desviaciones al cuadrado "entre" corresponde a las desviaciones "explicadas" y la suma "dentro" a las "no explicadas".

Obtención de la fórmula (4):

Construimos la siguiente identidad:

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x})$$

Elevando el cuadrado ambos miembros obtenemos:

$$(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_{i.})^2 + 2(x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}) + (\bar{x}_{i.} - \bar{x})^2$$

Sumando todos los valores, hallamos:

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 = \sum_i\sum_j(x_{ij}-\bar{x}_{i.})^2 + 2\sum_i\sum_j(x_{ij}-\bar{x}_{i.})(\bar{x}_{i.}-\bar{x}\ ) + \sum_i\sum_j(\bar{x}_{i.}-\bar{x}\ )^2 \quad ...5)$$

Total Dentro Entre

El término de los productos cruzados se puede calcular como sigue:

$$2\sum_i\sum_j(x_{ij}-\bar{x}_{i.})(\bar{x}_{i.}-\bar{x}\ ) = 2\sum_i\left[(\bar{x}_{i.}-\bar{x}\ )\sum_j(x_{ij}-\bar{x}_{i.})\right]$$

Pero $\sum_i^{n_i}(x_{ij}-\bar{x}_{i.})$ es la suma de las desviaciones respecto a la media dentro de un grupo, que es evidentemente cero. Así pues, el término del producto cruzado desaparece y la fórmula (5) se convierte en la relación presentada en la ecuación (4). Esta identidad se cumple independientemente si se cumple o no la hipótesis de $\mu_1=\mu_2=\mu_3$.

Ejemplo:
En la siguiente tabla se presenta el rango de notas de los alumnos de un curso divididos en tres grupos:

	Grupo 1	Grupo 2	Grupo 3
	3	4	7
	6	7	6
	5	7	7
	4	4	7
	7	8	8
Total	25	30	35
Media	5	6	7

Aplicar las diversas relaciones que se han deducido.

Solución:

1. la primera forma de estimar la varianza era por el método de la varianza combinada:

	x_{1j}	$x_{1j}-\bar{x}_1$	$(x_{1j}-\bar{x}_1)^2$	x_{1j}^2	x_{2j}	$x_{2j}-\bar{x}_2$	$(x_{2j}-\bar{x}_{12})^2$	x_{2j}^2	x_{3j}	$x_{3j}-\bar{x}_3$	$(x_{3j}-\bar{x}_3)^2$	x_{3j}^2
	3	-2	4	9	4	-2	4	16	7	0	0	49
	6	+1	1	36	7	1	1	49	6	-1	1	36
	5	0	0	25	7	1	1	49	7	0	0	49
	4	-1	1	16	4	-2	4	16	7	0	0	49
	7	2	4	49	8	2	4	64	8	1	1	64
Total	25		10	135	30		14	194	35		2	247
Media	5				6				7			

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2}{n-3} = \frac{1}{15-3}(10+14+2) = 2.17$$

2. la segunda forma de estimar la varianza σ2 era:

$$\hat{\sigma}^2 = \frac{1}{3-1}\sum_{i=1}^{3} n_i(\bar{x}_{i.}-\bar{x})^2$$, puesto que n1= n2= n3= 5, esto se convierte en:

$$\hat{\sigma}^2 = \frac{1}{3-1}(5)\left[(5-6)^2+(6-6)^2+(7-6)^2\right] = 5$$

3. la tercera manera de estimar la varianza σ2 era:

$$\hat{\sigma}^2 = \frac{\sum\sum(x_{ij}-\bar{x})^2}{n-1} = \frac{1}{15-1}\left[135+194+247-(15)(6)^2\right] = 2.57$$

Por lo tanto la identidad fundamental se cumple:

$$\sum\sum(x_{ij}-\bar{x})^2 = \sum\sum(x_{ij}-\bar{x}_{i.})^2 + \sum\sum n_i(\bar{x}_{i.}-\bar{x})^2$$

36 = 26 + 10

11.2. APLICACION DE LA DISTRIBUCION DE F

La relación de dos cantidades, u y v, que tienen distribuciones de χ^2 independientes, divididas por sus respectivos grados de libertad ϕ_1 y ϕ_2 se denominó razón de la varianza y la distribución de esta razón de la varianza se denominó distribución F.

Cuando aplicamos esto para demostrar la igualdad de las varianzas $\sigma_1^2 = \sigma_2^2$, encontramos que la razón de los estimadores insesgados de σ_1^2 y σ_2^2 satisfacía las condiciones para la razón de la varianza y presentaba distribución de F.
En función del presente problema, en el que queremos demostrar la igualdad de las medias, formaremos una relación con las sumas de cuadrados "dentro" y "entre" los grupos:

$$F = \frac{\frac{1}{3-1}\sum n_i(\bar{x}_{i.}-\bar{x})^2}{\frac{\sum\sum(x_{ij}-\bar{x}_{i.})^2}{n-3}}$$

Donde el numerador y denominador son estimadores insesgados de la varianza de la población σ2. Por consiguiente, esta razón también tiene una distribución de F. Esta razón se puede presentar esquemáticamente como:

$$F = \frac{\text{var}\,ianza\,estimada"entre"}{\text{var}\,ianza\ estimada"dentro"}$$

Recordemos que la varianza estimada "dentro", que es una varianza combinada, estima $\sigma 2$ independientemente de si es o no cierta la hipótesis nula $\mu_1 = \mu_2 = \mu_3$.

Pero la varianza estimada de la suma de cuadrados "entre" estima σ_2 solo cuando (a) las muestras proceden de la misma población, o (b) cuando las medias poblacionales de las diferentes poblaciones son iguales. Como vemos, la (b) es equivalente a (a). Cuando las medias de la población no son iguales, la varianza estimada de la suma de cuadrados "entre" será σ_2+c, donde c > 0 es una discrepancia debida a la desigualdad de las medias de la población y el resultado es que la razón de las varianza F será grande. Si invertimos el razonamiento, podemos decir que si F es significativamente grande hay razón para dudar de la igualdad de las medias, o de que las muestras provengan de la misma población.

Tabla de Análisis de la Varianza

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Fc
Entre	$S_A = \sum_{i=1}^{a} n_i(\overline{x}_{i.} - \overline{x})^2$	a – 1	$S'_A = \frac{S_A}{a-1}$	$FC = \frac{S'_A}{S'_e}$
Dentro	$S_e = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i.})^2$	n - a	$S'_e = \frac{S_e}{n-a}$	
Total	$S_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2$	n -1	$S'_T = \frac{S_T}{n-1}$	

Donde a = Número de Muestras.

$$n = \sum_{i=1}^{a} n_i \ .$$

EJERCICIOS

1.- Se dispone de la siguiente información
H0:μ=10
H1:μ >10
La media de la muestra es 12 para una de tamaño 36. La desviación estándar de la población es 3. Utilice el nivel de significación de 0.02
   a. ¿Es ésta una prueba de una o de dos colas?
   b. Enuncie la regla de decisión
   c. Calcule el valor del estadístico de prueba.
   d. ¿Cuál es su decisión respecto H0?
   e. Determine el valor p

2.- Una cadena de restaurantes afirma que el tiempo medio de espera de clientes por atender está distribuido normalmente con una media de 3 min. y una desviación estándar de 1 min. su departamento de aseguramiento de la calidad halló en una muestra de 50 clientes en un cierto establecimiento que el tiempo medio de espera era de 2.75 min. al nivel de significación de 0.05, ¿es dicho tiempo menor de 3 min.?
   a. Enuncie las hipótesis nula y alternativa
   b. Formule la regla de decisión.
   c. Calcule el valor estadístico de prueba.
   d. ¿Cuál es su decisión respecto de H0interprete el resultado. ¿Cuál es el valor p?

3.- Una muestra de 64 observaciones se selecciona de una población normal. La media muestral es de 215, y la desviación estándar de la muestra es 15. Realice la siguiente prueba de hipótesis utilizando el nivel de significación de 0.03.
H0:μ=220
H1:μ >220

   a. ¿Es ésta una prueba de una o de dos colas?
   b. Enuncie la regla de decisión.
   c. Calcule el valor estadístico de prueba.
   d. ¿Cuál es su decisión con respecto a H0?
   e. ¿Cuál es el valor p?

4.- Cuando fue contratada como servidora en un restaurante se dijo a Claudia Rojas: "puedes obtener, en promedio, más de $20 (dólares) al día en propinas". A los primeros 35 días de su trabajo en el restaurante, el importe medio diario de sus propinas fue $24.85, con una desviación estándar de $3.24. a nivel de significación de 0.01, ¿puede Claudia concluir que está ganando más de $20 en propinas?
   a. Exprese la hipótesis nula y la hipótesis alternativa.
   b. ¿Cuál es la regla de decisión?
   c. Evalúe el valor estadístico de prueba
   d. ¿Cuál es su decisión respecto a la hipótesis nula? Interprete el resultado.

5.- Una muestra de 65 observaciones se seleccionó de una población algo normal. La media de la muestra es 2.67, y la desviación estándar 0.75. Una muestra de 50 observaciones se selecciona de una segunda población algo normal. La media de la muestra es 2.59, y la desviación estándar 0.66. Efectúe la siguiente prueba de hipótesis utilizando el nivel de significación de 0.08.

H0:$\mu 1 = \mu 2$

H1:$\mu 1 > \mu 2$

a. ¿Es ésta una prueba de una o dos colas?
b. Exprese la regla de decisión
c. Calcule el valor estadístico de prueba.
d. ¿Cuál es su decisión respecto a HP?
e. ¿Cuál es el valor p?

6.- Un estudio se realiza comparando el costo de alquiler o rentar un departamento de una recámara en Cincinati mostró que un valor promedio de las rentas era de $370, con una desviación estándar de $30. Una muestra de 40 departamentos en Pittsburgs señaló que la renta media es de $380, con una desviación estándar de $26. A nivel de significación de 0.05, ¿Existe una diferencia en las rentas medias entre Cincinati y Pittsburgs? Aplique el procedimiento de pruebas de hipótesis de cinco pasos.

CAPITULO DIEZ

DISEÑO DE LA MUESTRA

En este capítulo se analizara un aspecto de la inferencia estadística, la estimación. Esta comprende el evaluar, o predecir, el valor de un parámetro de población desconocida, por ejemplo, la media o la proporción poblacional, con base en información estadística obtenida de un grupo pequeño (muestra)

1. PAPEL DEL MUESTREO EN LA TEORIA Y METODOS ESTADISTICOS

En un sentido general, se puede considerar la teoría del muestreo como coexistente con los métodos estadísticos modernos. Casi todos los adelantos estadísticos modernos se refieren a inferencias que se pueden efectuar respecto a la población, cuando se dispone de información sólo de una muestra de elementos de la población. A continuación se menciona algunas de las formas en que ésta se refleja en los programas estadísticos.

a. Trabajos de investigación

En la mayoría de los trabajos de investigación, la población se compone de todas las personas (o establecimientos industriales, granjas, etc.) en una ciudad u otras áreas. Se obtiene o se desea información de una muestra de la población, pero se requieren inferencias sobre las características de toda la población.

b. Diseño y análisis de experimentación

En el diseño y análisis de experimentación, la población representa todas las posibles aplicaciones de varias técnicas alternativas que puedan usarse.

Por ejemplo, el experimento puede ser agrícola, en el cual se investiga el efecto de varios fertilizantes. La población es infinita, debido a que representa el uso de fertilizantes en todas las posibles granjas en cualquier época. El problema consiste en diseñar experimentos de modo que se disponga del máximo de información para realizar inferencias respecto a la población total, a base de una muestra de tamaño limitado.

c. Control de calidad

Para la aplicación de los métodos de control de calidad en un establecimiento industrial, por ejemplo la población es todo el producto que sale de una máquina. Se necesita inferencias sobre la forma cómo los productos cumplen con las especificaciones. El término "control de calidad" se aplica también a una verificación sobre la calidad del trabajo de campo efectuado en una encuesta por muestreo. La verificación por muestreo se ejecuta después que se ha completado la muestra. Las operaciones de oficina como edición, codificación y perforación, también están sujetas al control de la calidad, se verifican una muestra del trabajo para determinar si cumple con estándares aceptables.

2. RAZONES PARA MUESTREAR UNA POBLACION

- La naturaleza destructiva de ciertas pruebas, es decir, no se puede evaluar a toda la población. Por ejemplo: Si los catadores de vino tuvieran que beberse todo el vino para evaluarlo, consumirían toda la producción y no quedaría producto para vender.
- Puede ser imposible revisar o localizar a todos los elementos de la población. Por ejemplo: las poblaciones de peces, aves, etc.
- Puede ser prohibitivo el costo de estudiar a todos los elementos de la población.
- Los resultados de una muestra pueden ser una estimación adecuada del parámetro poblacional, ahorrando por tanto, tiempo y dinero.
- Puede necesitarse demasiado tiempo para estar en contacto (o entrevistar) a todos los elementos de la población.

3. EJEMPLOS DE MUESTRAS

a.- Fondos limitados

Es muy conocido el uso de encuestas por muestreo cuando existen fondos limitados para recoger información. El muestreo también se puede usar para ahorrar dinero en la tabulación. Por ejemplo, en el censo de 1981 en el Perú se combinó Censo con muestra, la mayoría de los datos fueron obtenidos con una base del ciento por ciento.

b.- Ahorro de tiempo

Otros ejemplos del Censo de 1981 en Perú ilustran cómo pueden usarse las muestras para ahorrar tiempo. El empadronamiento censal se realizó en abril de 1981. El tiempo requerido para procesar los resultados era tal que se esperaba que la publicación de los resultados comience en 1981 y continuaría durante 1982. La muestra considerada en el Censo, se proceso, publicándose los resultados preliminares antes que los resultados completos del Censo.

c.- Concentración en casos especiales

Algunas encuestas requieren entrevistas tan intensas y prolongadas, que es imposible considerar salvo mediante una base muestral. Más aún, el uso de muestreo permite prestar atención especial a un número limitado de casos. Ejemplos de ello los constituyen los estudios de presupuestos familiares y encuestas amplias sobre condiciones de salud.

d.- Muestreo para series de tiempo

Puede necesitarse información para series de tiempo, cuando sólo se dispone de datos para periodos especiales y los resultados se precisan con rapidez. La serie puede referirse a la actividad económica del país, de la cual sólo se dispone de las cifras anuales o mensuales o puede ser para elaborar una curva de experimentación en la cual únicamente se pueden hacer ensayos ocasionales.

e.- Control de los errores no muestrales

Un ejemplo interesante surgió en el citado censo de 1981: se trataba de un censo donde la relación entre los errores no muestrales y los muestrales, hacía preferibles los resultados de la muestra a los provenientes de un censo completo. En el Perú se realiza desde 1940 una encuesta muestral sobre la fuerza de trabajo.

En 1950, ella se basaba en una muestra de 20.000 hogares. La información obtenida en el censo de 1981 también incluía la situación del empleo dentro de la fuerza de trabajo. Cuando se dispuso de los resultados del censo, pareció evidente que las cifras tanto para personas desocupadas como para las empleadas, eran completamente diferentes de las estimadas mediante la encuesta por muestreo de la fuerza del trabajo; las diferencias eran muy superiores a la que podía esperarse debido a los errores muestrales.

El problema de información en el censo introdujo un error mucho mayor que el error de muestreo de la encuesta mensual (este mayor error era causado por la intervención de empadronadores que, en su mayor parte, no tenían experiencia en entrevistar). Los usuarios de los datos del censo fueron aconsejados por lo tanto, para que usaran los resultados muestrales como estadísticas nacionales más fidedignas sobre la fuerza del trabajo.

4. LIMITACIONES DEL MUESTREO

En ciertas condiciones, la utilidad del muestreo es dudosa. Se pueden mencionar tres puntos principales:

a.- Si se necesitan datos para áreas muy pequeñas, se requieren muestras desproporcionadamente grandes, pues la precisión de una muestra depende fuertemente del tamaño y no de la tasa del muestreo. En este caso, el muestreo puede ser casi tan costoso como un censo completo.

b.- Si los datos se necesitan a intervalos regulares y es importante medir los cambios muy pequeños de un periodo a otro, se necesitarán muestras muy grandes.

c.- Si existen costos generales fijos desusadamente grandes ligados a la encuesta por muestreo, debido al trabajo necesario para la selección de la muestra, control, etc, el muestreo puede ser poco práctico. Por ejemplo, en un país con muchas poblaciones pequeñas, puede ser más económico enumerar todos los hogares en la muestra de poblaciones, que enumerar una muestra de hogares dentro de las poblaciones muestreadas. Para la elaboración en la oficina, sin embargo, se puede usar una muestra de los hogares enumerados para reducir el trabajo y el costo de las tabulaciones.

5. TIPOS DE MUESTRAS

5.1. MUESTREO PROBABILISTICO

Es una muestra que se selecciona de modo que cada integrante de la población en estudio tenga una probabilidad conocida (no igual a cero) de ser incluido en la muestra.

Los **métodos de muestreo probabilísticos** tienen un objetivo, que es permitir que el azar determine los integrantes que se incluirán en la muestra. Existen varios métodos de muestreo de probabilidad entre ellos tenemos:

5.2. MUESTREO SIMPLE AL AZAR O MUESTREO ALEATORIO SIMPLE

Es el método más sencillo de muestreo. Se puede decir que si n es el tamaño de la muestra, cada una de las posibles combinaciones de tamaño n que se pueden formar con las N unidades de análisis que forman la población tiene la misma probabilidad de ser incluida en la muestra, asimismo, cada elemento o unidad de análisis de la población tiene la misma probabilidad de ser seleccionado que el de cualquier otro elemento.

Métodos de selección
La selección puede llevarse a cabo en dos formas distintas:
- Con reposición
- Sin reposición

Supongamos que cada elemento poblacional está identificado en una ficha. De las N fichas se extrae una al azar y luego se repone. Es posible, por lo tanto, que esta misma ficha pueda extraerse de nuevo.

Esta selección se llama **con reposición** y el número de muestras posibles de tamaño n de una población de tamaño N está dado por la siguiente expresión:

$$L = N^n$$

donde:

L = Número de muestras posibles con reposición de tamaño n.
N =Tamaño de la población
n = Tamaño de la muestra

Por otra parte, la selección se puede hacer de las n fichas simultáneamente o de n fichas sin responderlas. Este es el muestreo **sin reposición** el número de muestras posibles está dado por lo siguiente fórmula:

$L = \dfrac{N!}{n!(N-n)!} = \binom{N}{n}$, que es la combinación de n elementos de N elementos, siendo:

L = Número de muestras sin reposición, de tamaño n.
N = Tamaño de la población
n = tamaño de la muestra

Ejemplo: Sea una población hipotética de 12 personas cuyo ingreso medio se estima mediante una muestra, la población total se presenta en el **cuadro 1**.

Cuadro 1. Ingreso percápita de una población hipotética de 12 personas

Individuo	Ingreso $
A	1.300
B	6.300
C	3.100
D	2.000
E	3.600
F	2.200
G	1.800
H	2.700
I	1.500
J	900
K	4.800
L	1.900
Ingreso total	32.100
Ingreso medio	2.675

Supóngase que se estima mediante una muestra de dos individuos. Existen varias formas de seleccionar la muestra.

Por ejemplo, pueden usarse 12 fichas de igual dimensión con una de las doce letras A,B,C,...,L, inscrita, sin que dos tengan la misma letra. Se colocan luego las fichas en una urna, se mezclan cuidadosamente y luego se eligen aleatoriamente dos fichas que representarán a las personas escogidas. Este tipo de selección puede efectuarse en dos formas.

Hay otras formas de seleccionar dos personas al azar. Para muestreo sin reposición pueden considerarse todos los posibles pares de personas. AB, AD ,...,BC, BD,..., Etc. Podría escribirse una par de letras para cada uno de estos 66 pares en una ficha, luego se selecciona solo una ficha. Las posibles muestras y las oportunidades de selección son las mismas que antes.

En la práctica, no se usan fichas para seleccionar individualmente o en pares. El método usual es emplear una tabla de números aleatorios y elegir en la tabla dos números del 1 al 12. Los dos números representan a dos individuos. El efecto de utilizar tablas de números aleatorios es exactamente el mismo que si se usan las fichas.

Cualquiera que sea el método usado, se satisface el criterio para una muestra aceptable. En cada uno de los tres métodos, cada persona tiene una oportunidad de selección, se conocen las probabilidades y pueden ser calculadas. Para el pequeño universo en consideración, cualquiera de los tres métodos es práctico. (En situaciones más realistas, únicamente será práctico el procedimiento de la tabla de números aleatorios). Finalmente, los tres satisfacen las condiciones para una muestra aleatoria simple, ya que todas las posibles combinaciones de dos personas son igualmente probables.

6. DEFINICIONES Y RELACIONES QUE SE DERIVAN DE LA TEORÍA DE MUESTREO:

6.1. Desviación Estándar

Se demostrará que existe una medida de la variabilidad en la población original, que puede estimar mediante las observaciones en una sola muestra y con la cual es posible estimar el error esperado en la media de la muestra. La medida de la variabilidad en la población se llama desviación estándar, su cuadrado se llama varianza y se designa por el símbolo σ^2.

6.2. La Varianza

Se define como la media de los cuadrados de lo desvíos de las observaciones individuales respecto a su valor medio. Por lo tanto se calcula el siguiente procedimiento, si se puede observar todos los valores del universo.

$$\sigma^2 = \frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + ... + (Y_N - \bar{Y})^2}{N} = \frac{\sum_{i=1}^{N} (Y_N - \bar{Y})^2}{N}$$

donde las y_i con subíndice representan observaciones individuales y $\bar{Y}$ es la media de N observaciones para los N elementos del universo.

6.3. Error Estándar

En forma similar, se pueden computar las medias de todas las posibles muestras de tamaño n. Al elevar al cuadrado sus desviaciones de la media verdadera y promediar la suma de esos cuadrados, resulta la varianza de las medias muestrales. La raíz cuadrada de este número es desviación estándar de las medias, o como generalmente se llama, el **error estándar** para medias de muestras de tamaño "n". El error estándar varía con el tamaño de la muestra.

Se calcula el error estándar para todos los tamaños posibles en el ejemplo anterior y se obtiene el **cuadro 2**.

Cuadro 2. Error estándar de estimaciones del ingreso para varios tamaños de muestra

Tamaño de muestra	Error estándar ($\sigma_{\bar{y}}$)
1.....	S/. 1.505
2.....	1.015
3.....	786
4.....	642
5.....	537
6.....	454
7.....	383

El error estándar de muestra de tamaño 1 es igual a la **desviación estándar** de la población. Una vez conocida la desviación estándar, es fácil, obtener el error estándar para muestras de cualquier tamaño, sin calcular las numerosas estimaciones muestrales posibles. También se pueden estimar la desviación estándar y el error estándar con una muestra única.

6.4. Muestreo Aleatorio Sistemático

Si la técnica de muestreo donde los integrantes de la población se ordenan alfabéticamente en un archivo según la fecha en que se reciben, o por algún otro método. Se selecciona un punto de inicio y después se elige cada K-ésimo elemento de la población para la muestra.

N° de muestra U. en la muestra	1	2	3	i	k
1ª	y_1	y_2	y_3	y_i	y_k
2ª	y_{k+1}	y_{k+2}	y_{k+3}	y_{k+i}	y_{2k}
3ª	y_{2k+1}	y_{2k+2}	y_{2k+3}	y_{2k+i}	y_{3k}
j-ésima	$y_{(j-1)k+1}$	$y_{(j-1)k+2}$	$y_{(j-1)k+3}$	$y_{(j-1)k+i}$	y_{jk}
.	.	.	.	.	.
.	.				
n-ésima	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	$y_{(n-1)k+3}$	$y_{(n-1)k+i}$	y_{nk}

6.4.1. Ventaja del muestreo sistemático

1. El marco muestral no necesariamente se puede conocer, se puede ir generando paralelamente. En el campo se puede ir obteniendo el Marco Muestral.
2. El arranque es aleatorio, pero el problema es que cada 4 viviendas se selecciona n/N = 0.25).

6.4.2. Razones del Uso del Muestreo Sistemático

1. La muestra se distribuye a lo largo de toda la población.
2. Es posible captar el efecto de la estratificación.
3. La muestra se puede generar en el campo.

6.4.3. Limitaciones

1. Se podría estar aumentando sustancialmente la varianza del estimador si la población bajo estudio tiene una variación cíclica.
   Porque están captando los puntos: máximo y mínimo, luego los estimadores se estarán sobrestimando o subestimando.

2. Para el calculo de la V($\overline{x}_{Sist.}$) es muy difícil que una sola muestra. La fórmula que existen para estimar V($\overline{x}_{Sist.}$) a partir de una sola muestra son aproximaciones

   $P(\mu_1, \mu_{1+k}, \mu_{1+2k},..., \mu_{1+(n-1)k}) = 1/k = n/N \quad n = nK$
   .
   .
   .
   $P(\mu_k, \mu_{2k},..., \mu_{nk}) = 1/k$

   porque N = nk
   La P indica la probabilidad de selección en mas con reposición.

7. ESTIMADORES

a. Del total del poblacional

$$\hat{X} = N\overline{x}_j \text{ donde } \quad \overline{X}_j = \frac{\sum_{i=1}^{n} X_{ji}}{n}$$

De la media poblacional

$$\overline{X}_{j\,Sist.} = \frac{1}{n}\sum_{i=1}^{n} X_{ji}$$

$\overline{X}_{Sist.}$ = promedio de la muestra

b. Varianza de los estimadores en función de los componentes de la varianza

Cuando la composición interna de cada muestra disponible es heterogénea, entonces la varianza por muestreo sistemático es menor y es aconsejable aplicar este método de muestreo.
Es decir un buen muestreo sistemático es cuando la varianza del interior de cada muestra sistemática es grande (los datos son heterogéneos).
El problema ahora es determinar cuando esta varianza es grande; es decir determinar los criterios a utilizar para decir que la muestra es heterogénea.
Se puede pasar directamente de la varianza de una media muestral a la V(total) poblacional multiplicada por N^2.

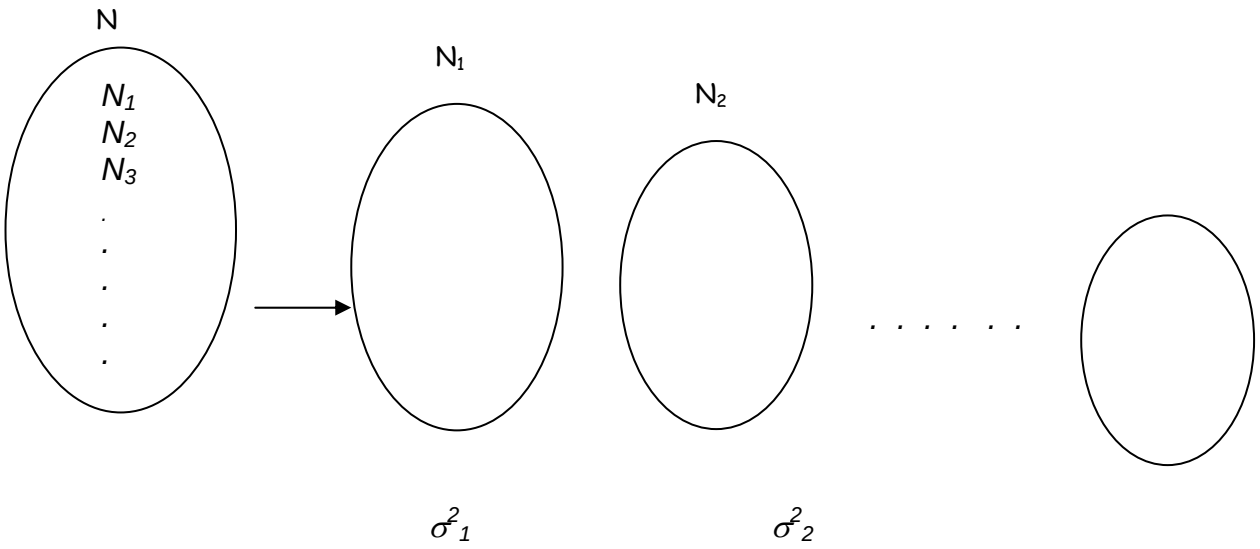
c. La varianza de los estimadores en función del coeficiente de corrección intraclásica

Esta fórmula significa que ρ = 0 las muestras son extremadamente heterogéneas si la muestra se comprende con la del mas.

Cuando ρ = 1 la varianza asume su máximo valor y lo que interesa es la varianza mínima para obtener estimadores preciso. El muestreo sistemático es aplicado universalmente.

d. Muestreo Aleatorio Estratificado

La estratificación de la población consiste en agrupar sus unidades en un cierto número de clases disjuntas denominadas estratos, constituidas por unidades similares: con esto se persigue que la varianza en cada uno de ellos sea pequeña.



σ^2_i pequeña (i=1,2,3,.....k)

- El muestreo estratificado consiste en seleccionar K muestras independientes, provenientes cada una de un estrato (N_i).
- En este tipo de muestras la eficiencia es mayor cuanto mayor sea la homogeneidad de cada estrato.

8. RAZONES DE SU USO

1.- Lograr una disminución de la varianza de los estimadores

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \quad \text{o} \quad \sigma^2 = \frac{\sum\left(X_i - X\right)^2}{N-1}$$

Se trabaja con S^2 porque el $\hat{S}^2$ es un estimador insesgado $E\left(\hat{S}^2\right) = \sigma^2$

$$\hat{S}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \qquad \hat{\sigma}^2 = \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2$$

conviene desagregar

$$S^2 = S^2{}_a + S^2{}_b$$

S^2_a varianza dentro de los estratos
S^2_b varianza entre estratos

Se busca homogeneidad al interior de cada estrato y heterogeneidad entre cada estrato.
Luego $S^2_a < S^2_b$ y a lo mas $S^2_a = S^2b$
Si se da $S^2_a > S^2_b$ el diseño muestral estará mal realizado
Dentro de cada estrato se fija un tamaño de muestra n_h

2.-Tener la posibilidad de aplicar diferentes métodos muestrales, en cada uno de los estratos. En el área rural no se ha podido hacer un registro completo de las unidades censales; las variables en el área urbana son más heterogéneas que en el área rural.

En esta área por ejemplo, el 80% de la PEA se dedica a la agricultura luego se seleccionan muestras de viviendas en forma aleatoria simple. Como no se tiene un registro completo de las unidades censales se aplica el muestreo sistemático
En el área urbana interesa aplicar la estratificación por la heterogeneidad de las variables socioeconómicas.

3.- Permite obtener información para cada dominio de estudio. (Ej. División política administrativa), satisface además el interés de análisis de los usuarios.
Ejemplo, estudios por estratos socioeconómicos de los hogares, o estratos económicos de empresas.

4.- Lograr una disminución de costos. Por ejemplo, si en la encuesta nacional de hogares no se estratificara es posible que el tamaño de la muestra sea mayor porque la varianza sería mayor sino se estratificara. Al estratificar una encuesta en cada estrato su varianza es muy pequeña, que incide en la varianza total.

En muchos casos cuando no se dispone de información de la variable de mayor importancia en la investigación, se necesita una información auxiliar necesariamente para estratificar. Por ejemplo, para seleccionar establecimientos podría ser el número de personas ocupadas por establecimiento, significaría tener información para cada establecimiento (unidad muestral) de dicha variable en toda la población lo que significaría mayores costos.

9. PROCEDIMIENTO PARA ESTRATIFICAR

a. Dalenius ha realizado una técnica para fijar los límites de estratos. Ejemplo se tiene la alternativa de formar 5 estratos o diez estratos. En esta última la varianza sería menor, pero Dalenius dice que llega un momento en que la varianza se estabiliza y de nada sirve aumentar el número de estratos, porque su costo es mayor que su utilidad marginal. Es decir si se escogió una buena variable de estratificación, se puede seleccionar la muestra hasta en dos estratos.

Para la formación de estratos, si es necesario, se debe disponer una variable auxiliar perfectamente correlacionada con la variable más importante motivo de la investigación

b. Al aumentar el número de estratos se disminuye la varianza de los estimadores.

10. SELECCION DE LA MUESTRA

La muestra de cada estrato es independiente. Es decir si seleccionamos un tamaño de la muestra n_1 en un estrato,(E_1) es independiente del tamaño de la muestra n_h en un estrato cualquiera (E_h).
En consecuencia se puede aplicar un tipo de muestreo diferente en cada estrato

10.1. MUESTREO POR CONGLOMERADO

Este método se emplea a menudo para reducir el costo de muestrear una población dispersa en un área geográfica grande.

10.2. MUESTREO NO PROBABILISTICO

Es el método de muestreo en el cual una muestra es seleccionada de tal manera que no todos los integrantes de la población tienen la probabilidad de ser incluidos en la muestra. Es decir, la inclusión en la muestra se basa en juicios de la persona que realiza el muestreo. Las muestras no probabilísticas pueden llevar a resultados con sesgo.

11. CRITERIOS PARA LA ACEPTABILIDAD DE UN METODO DE MUESTREO

Los métodos de muestreo pueden proporcionar datos de confiabilidad conocida con eficiencia y economía.
Para aceptar una muestra es necesario que ésta represente a la población, que tenga una confiabilidad que se pueda medir y que responda a un plan práctico y eficaz.

Los criterios más comúnmente usados para aceptar un método de muestreo son:

11.1.- Probabilidad de selección para cada unidad

La muestra debe seleccionarse de modo que represente adecuadamente a la población. Cada unidad debe tener una probabilidad conocida de ser elegida y esta probabilidad debe ser siempre distinta de cero, osea, $0<p\leq 1$.

11.2.- Confiabilidad medible

Los valores de la muestra deben proporcionar medidas de la confiabilidad de las estimaciones que con ellos se calculan y de la precisión que se espera o desea que tengan.

11.3.- Viabilidad o factibilidad

El plan de muestreo adoptado debe ser práctico y permitir que sea realizado en la forma proyectada.

11.4.- Economía y eficiencia

El diseño muestral debe ser eficiente o sea, que sea capaz de proporcionar la mayor cantidad de información al menor costo, para lo cual debe hacerse el uso más efectivo posible de los recursos disponibles.

12. DEFINICION O DESCRIPCION DE ALGUNOS TERMINOS

12.1.- Unidad de análisis

Es la unidad o elemento para el cual se desea obtener información estadística sobre determinadas características o variables.

12.2.- Población o universo

Es el conjunto de todas las unidades de análisis cuyas características se desean investigar, es decir, de las N unidades o elementos que conforman la población.

12.3.- Marco de muestreo

La totalidad de la unidades de muestreo de donde se extrae la muestra constituye el marco muestral o de muestreo. El marco de muestreo puede ser una lista de personas o de unidades de vivienda, un archivo de registros o un conjunto de tarjetas perforadas; puede ser un mapa subdividido, o puede ser un directorio de nombres y direcciones en una cinta magnética para computadora.

12.4.- Unidades de muestreo

Es la unidad seleccionada del marco de muestreo. Puede o no coincidir con la unidad de análisis. Por ejemplo, para obtener información sobre personas, se puede usar una lista completa de un censo, o un directorio de personas y seleccionar una muestra de personas directamente. Sin embargo, también se podrá seleccionar una muestra de hogares e incluir en la encuesta a todas las personas de los hogares seleccionadas.

12.5.- Probabilidad de selección

La oportunidad que tiene una unidad de la población de ser incluida en la muestra, se denomina su probabilidad de selección. Los valores de la probabilidad van de 0 a 1.

12.6.- Estadística

Una estadística o estadígrafo es una cantidad que se calcula con las observaciones muestrales correspondientes a una característica determinada con el objeto de efectuar inferencias acerca de la población.

12.7.- Información independiente

Son datos conocidos antes de realizar las encuestas o investigación estadística que sirven para su diseño, estratificación o establecer probabilidades de selección.

12.8.- Fórmulas de estimación o estimadores

Son aquellos que utilizan los resultados de la muestra para producir una estimación sobre valores poblacionales o parámetros, por ejemplo:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ que se utiliza para estimar:

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

12.9.- Intervalos de Confianza

Es el intervalo alrededor del cual se espera que esté el verdadero valor poblacional con el fin de producir una estimación sobre valores poblacionales o parámetros.

12.10.- Error de Muestreo

Es la diferencia entre parámetro poblacional y la estadística muestral

12.11.- Distribución Muestral de Medias

Es una distribución probabilística que señala todas las medias muestrales posibles y sus probabilidades de ocurrencia.

12.12.- Teorema Central del Limite

En este caso si la población es normal, entonces la distribución de medias también manifiesta normalidad. Si la población no es normal, la distribución de muestreo de las medias se aproxima a la normal a medida que aumenta el tamaño de la muestra.

13. ESTIMACIONES PUNTUALES Y DE INTERVALO

13.1. ESTIMACION PUNTUAL

Es un valor único que se utiliza para estimar un valor de la población.

13.2. ESTIMACION POR INTERVALO

Es un conjunto de valores dentro de las cuales se espera que ocurra el parámetro de la población. Por lo general el intervalo se denomina intervalo de confianza.

Los factores que constituyen un intervalo de confianza para una media son:

1. El número de observaciones en la muestra (n)
2. La variabilidad de la población, que generalmente se estima mediante la desviación estándar (s)
3. El nivel de confianza. Se representa mediante el desvío normal o valor z Un intervalo de confianza de 95% para la media se obtiene usando la siguiente fórmula:

$$X \pm 1.96\frac{s}{\sqrt{n}}$$

Los factores que constituyen un intervalo de confianza para una proporción son:

1. El número de observaciones en la muestra.

2. El valor de $\overline{p}$, que se obtiene dividiendo el número de éxitos en la muestra (X) entre el número de observaciones en la misma (n).
3. El nivel de confianza. Está representado por el valor z.

14. INTERVALO DE CONFIANZA PARA UNA PROPORCION DE LA POBLACION

Un intervalo de confianza para una proporción se determina usando la siguiente fórmula:

$$\overline{p} \pm z\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

Donde:

$\overline{p}$: es la proporción muestral
z : es el valor z del grado de confianza seleccionado.
n : es el tamaño de la muestra.

14.1. SELECCION DEL TAMAÑO DE LA MUESTRA

El tamaño de la muestra necesario puede determinarse tanto para las medias como para las proporciones.
Los factores que determinan el tamaño de muestra para una proporción son:
1. El nivel de confianza deseado (z)
2. El máximo error permisible (E)
3. La variación de la población (que por lo general se estima con s)

14.2. VARIACION EN LA POBLACION

Para tener el tamaño de la muestra se necesita estimar la variación en la población.
La fórmula para el tamaño de muestra para una media es:

$$n = \left(\frac{z * s}{E}\right)^2$$

Donde:
E : es el error permisible
z :es el desvío normal asociado al grado de confianza seleccionado
s : es la desviación estándar de la muestra del estudio piloto.
Los factores que determinan el tamaño de muestra para una proporción son:
1. El nivel de confianza deseado (z)
2. El máximo error permisible (E)
3. La estimación de la proporción de la población. Sino se cuenta con una estimación, entonces se usa 0.50

15. TAMAÑO DE MUESTRA PARA PROPORCIONES

Para determinar el tamaño de muestra se deben especificar tres importantes aspectos:
1. El investigador debe decidir qué nivel de confianza emplear, por lo general 0.95 ó 0.99.
2. Debe indicar qué tan precisa debe ser la estimación de la proporción de la población.
3. La proporción de la población, $\overline{p}$, se debe aproximar con base en la experiencia o en un estudio piloto pequeño.
4. La fórmula para el tamaño de muestra para una proporción es:

$$n = \overline{p}(1-\overline{p})\left(\frac{z}{E}\right)^2$$

Factor de corrección de población finita se aplica si n / N es mayor que 0.50 tal factor de corrección es:

$$\sqrt{\frac{N-n}{N-1}}$$

EJERCICIOS

1. Una población consta de los cinco valores siguiente: 2,2,4,4,8
   a. ¿Cuántas muestras de tamaño 2 son posibles?
   b. Enliste todas las muestras posibles de tamaño 2, y calcule la media de cada muestra
   c. Determine la media de las medias muestrales y la media de la población. Compare los dos valores.
   d. Compare la dispersión poblacional con la media de las muestrales.

2. Una muestra de 81 observaciones se toma de una población normal. La media muestral es 40, y la desviación estándar de la muestra es 5. Determine el intervalo de confianza de 95% para la media de la población.

   Solución:

   De los datos del problema:
   n = 81 $\bar{x}$ (media muestral) = 40 $S_x = 5$

   El intervalo de confianza será: $\bar{x} \pm ZS_x$

   (30.2 ; 59.208)

3. Hay 300 enfermeras empleadas en un hospital. Una muestra de 30 reveló que 18 se graduaron en una escuela especial. Establezca un intervalo de confianza de 95% para la proporción de enfermeras graduadas en dicha escuela.
   N = 300 f = 0.1 Z =1.96
   n = 30
   Se graduaron en una escuela especial = 18
   P = 0.60
   Sp =0.085

   El intervalo de confianza será: (0.434 ; 0.766)

4. Explique qué significa error de muestreo

5. Un estudio de los establecimientos de Salud en un área Metropolitana mostró que existen 10. El Ministerio de Salud estudia el número de camas en cada establecimiento de salud. Los resultados son los siguientes: 90, 72, 75, 60, 75, 72, 84, 72, 88, 74, 105, 115, 68, 74, 80, 64, 104, 82, 48, 58, 60, 80, 48, 58 y 108.
   a. Utilice la tabla de números aleatorios y seleccione una muestra aleatoria de tamaño cinco a partir de esta población.
   b. Obtenga una muestra sistemática seleccionado un punto de inicio aleatorio entre los cinco primeros establecimientos y después selecciones cada quinta institución.

6. Las edades de seis médicos de una clínica son:

Nombre	Edad
Sr. Perez	54
Sra. Salas	50
Sr. Lara	52
Sra. Ruiz	48
Sr. Luna	50
Sr. Soto	52

a. ¿Cuántas muestra de tamaño dos son posibles?
b. Seleccione todas las muestras posibles de tamaño dos de la población de ejecutivos y calcule las medias.
c. Organice las medias en una distribución muestral.
d. ¿Cuál es la media de la población? ¿De la media muestral?
e. ¿Qué forma tiene la población?
f. ¿Qué forma tiene la distribución muestral?

7. En cierta región, los salarios diarios de los médicos se distribuyen normalmente con una desviación típica de 7.0. ¿De qué tamaño debe ser la muestra aleatoria que se tome, si se desea tener una seguridad del 96% de que la media muestral difiera de la media poblacional en menos de 3?

8. Si la desviación estándar de las estaturas de un grupo de niños de un distrito es de 5 cm. ¿Cuál es la probabilidad de que la estatura promedio de una muestra al azar de 100 de dichos niños:

a. Exceda en más de 1.5 cm. A al estatura promedio de todos los niños.
b. Sea menor en más de 1.5 cm. Con relación a la estura promedio de todos los niños?

CAPITULO ONCE

ANALISIS DE REGRESION

La teoría de la regresión pretende hacer un análisis sobre la relación que existe entre las variables dependientes e independientes (explicativas) para un conjunto de valores observados sobre estas variables.

1. NATURALEZA DEL ANALISIS DE REGRESION

El análisis de regresión está relacionado con el estudio de la dependencia de una variable, la variable dependiente, está en función de una o más variables explicativas con la perspectiva de estimar y/o predecir el valor (poblacional) medio o promedio de la primera en términos de valores conocidos o fijos (en muestreos repetidos) de las segundas.

El objetivo es determinar una ecuación de regresión que permita pronosticar el valor de una variable (denotado por Y y denominado variable dependiente) con base en otra variable (denotada por X y llamada variable independiente).

Ejemplo:
Se efectúa una encuesta de ingresos y gastos a 60 familias, que viven en un centro poblado. La información se presenta en el cuadro adjunto.

Y = Consumo de la familia.
X = Ingreso de la familia.

Ingreso de las Familias

f(y)/ x	80	100	120	140	160	180	200	220	240	260
Gasto de consumo Familiar por semana y $	55 60 65 70 75	65 70 74 80 85 88	79 84 90 94 98	80 93 95 103 108 113 115	102 107 110 116 118 125	110 115 120 130 136 140	120 136 140 144 145	135 137 140 152 157 160 162	137 145 155 165 175 180	150 152 165 168 180 185 191 173
E(y/x)	65	77	89	100	113	125	137	149	161	173

$$E(y) = \frac{\sum y}{n} \quad ; \quad E_{j(y)} = \left[\sum_{i=1}^{n_j} \frac{y_i}{n_j} \right] \quad ; \quad i = 1,2,3,\ldots,n_j \qquad j = 1,2,3,\ldots 10$$

Esperanza promedio: E(y/x=80) = 65
El consumo promedio de las familias que ganan 80 es 65.
E(y/x=260) = 173
El consumo promedio de las familias que ganan 260 es 173.

2. FUNCION DE REGRESION POBLACIONAL (FRP)

Para la construcción de la función de regresión poblacional la curva de regresión debe expresar todos los valores promedios de la variable dependiente para todos los valores fijos de la variable explicativa.
La regresión poblacional nos muestra cómo el valor promedio de Y varía en relación a los valores de la variable X.

El ejemplo anterior se trata de los valores promedios de consumo en cada valor fijo del ingreso.

$$\boxed{FRP \rightarrow E(y/x) = \beta_1 + \beta_2 x}$$

Donde:

β_1 , β_2 son parámetros desconocidos pero fijos que se denominan coeficiente de regresión, también llamados intercepto y coeficiente de la pendiente de la recta formada respectivamente.

E(y/x = 80) = 65. Valor promedio de y para x =80

La diferencia entre el valor promedio obtenido y cada valor observado se debe al termino de perturbación (μ_1).

$Y_i = E(Y/x) + \mu_i$

$Y_i = \beta_1 + \beta_2 x + \mu_i$, reemplazando para c/u de los valores del consumo cuando el ingreso es 80, nos da las siguientes expresiones:

$Y_1 = 55 = \beta_1 + \beta_2 x + \mu_1$
$Y_2 = 60 = \beta_1 + \beta_2 x + \mu_2$
$Y_3 = 65 = \beta_1 + \beta_2 x + \mu_3$
$Y_4 = 70 = \beta_1 + \beta_2 x + \mu_4$
$Y_5 = 75 = \beta_1 + \beta_2 x + \mu_5$

3. FUNCION DE REGRESION MUESTRAL (FRM)

Es la que se obtiene a partir de una muestra de observaciones y nos permite estimar los parámetros de una función de la regresión poblacional, a partir de la información proporcionada por la muestra. Su forma estocástica es la siguiente:

$$FRM \rightarrow Y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + e_i$$

La diferencia con la FRP está dada en que en este último caso los valores de los parámetros son de los datos poblacionales (β_i). Asimismo el término de perturbación (μ_i) está referido a la diferencia de los valores promedios poblacionales respecto a cada uno de los valores mencionados.

Podemos afirmar lo siguiente:

$\hat{\beta}_1$ es un estimador de β_1
$\hat{\beta}_2$ es un estimador de β_2
e_i es un estimador de μ_i

4. SIGNIFICADO DEL TERMINO DE PERTURBACION (μ_i)

Se tiene un modelo más general, de la siguiente forma:

$$Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_2 + ... + \mu_i$$

Donde los valores de los parámetros (β) son referidos a la población. Suponiendo que alguien nos diera los valores de los β nos faltaría encontrar el valor del término de perturbación (u_i).

El u_i se simboliza como una bolsa donde están las otras variables respectivas del modelo y que no están incluidas en el mismo. Asimismo representa efectos aleatorios de la misma naturaleza de las u_i

En el caso del consumo por ejemplo u_i estaría representando el efecto de otras variables siguientes: riqueza, tamaño de la familia.
El u_i siempre está a partir de los residuales.

Sea el modelo: $Y = \beta_1 + \beta_2 x_2$

β_1 = 10; β_2=2 $u_i \sim N(0, 25)$

X_2	(Y_i)	$e_i = \hat{\mu}$	Y_i
2	14	-2	12
5	20	5	25
4	18	0	18
6	22	3	19

y teórico **y Empírico**

Valor promedio

$$\beta_1 + \beta_2 x_2 = Y$$

10+2(2)=14 2 = 12
10+2(5)=20 5 = 25
10+2(4)=18 0 = 18
10+2(6)=22 -3 = 19

E(y/x) u_i Y_i

MODELO LINEAL GENERAL

HIPÓTESIS

Supongamos que existe una relación lineal entre una variable Yi , K-1 variables explicativas X_2, X_3,..... X_K y un término de perturbación u podemos escribir:

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3\ x_{i3} + ... + \beta_k x_{ik} + \mu_i$$

(i=1,2,....n)

Para efectos del cálculo matricial tenemos los siguientes:

1. La ecuación tradicional es:

$$Y_1 = \beta_1 + \beta_2 X_{12} + \beta_3\ X_{13} + ... + \beta_k X_{1k} + \mu_i$$
$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3\ X_{23} + ... + \beta\ X_{2k} + \mu_2$$
$$Y_3 = \beta_1 + \beta_2 X_{32} + \beta_3\ X_{33} + ... + \beta_k X_{3k} + \mu_3$$
$$.$$
$$.$$
$$Y_n = \beta_1 + \beta_2 X_{n2} + \beta_3 X_{n3} + ... + \beta_k X_{nk} + \mu_n$$

$$\mathbf{Y_{n1} = X_{nk}\beta_{k1} + \mu_{n1}}$$

La ecuación matricial se escribe de la siguiente forma:

i = 1
i = 2
i = 3
.
.
.
.
i = n

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & .x_{n1} \\ 1 & x_{22} & x_{32} & .x_{n2} \\ 1 & x_{23} & x_{33} & .x_{n3} \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{2k} & x_{3k} & .x_{nk} \end{bmatrix} X \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ . \\ . \\ \hat{\beta}_n \end{bmatrix} + \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \\ . \\ . \\ \hat{\mu}_n \end{bmatrix}$$

o simplemente:

$$Y = X B + U$$

2. Supuestos del modelo

   a. relativas a las perturbaciones
      i. $E(u_i) = 0$
      ii. $Var(u_i)$ = constantes = σ^2_u (HOMOCEDASTICIDAD)
      iii. $Cov(u_i u_j) = o$ (INDEPENDENCIA) $i \neq j$
      iv. Es decir, $(u_i) = N(0,\sigma^2_u)$

b. relativas a las variables

i. Las variables x_2, x_3 ..... x_k son variables no aleatorias.

ii. La variable explicada y_i es aleatoria con *media*:
$E(y_i) = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik}$
O también: E(Y)=XB
Varianza: $E\left[(Y_i - E(Y_i))\right]^2 =$
$E\left[(\beta_1 + \beta_2 X_{i2} + ... + \beta_{ik} X_k + \mu_i) - \beta_1 - \beta_2 X_{i2} - ... - \beta_k X_{ik}\right]^2 = E(\mu_i)^2 = \sigma_u^2$

iii. la variable y_i (explicada) y x_2, x_3.....x_k (explicativas) no tienen errores de observación.

iv. Entre las variables: x_2, x_3.....x_k no debe haber relación lineal, es decir, $Cov(x_i x_j)=0 \quad i \neq j$
Lo anterior significa que el rango de la matriz X debe ser K; por consiguiente ninguna columna debe ser linealmente dependiente de otra columna.

v. Para poder estimar el modelo se requiere tomar una muestra de n elementos, tal que n>k.

1. ESTIMACIÓN DE LOS PARÁMETROS

El principio básico para estimar los parámetros es que la suma de los residuales de cada valor observado respecto al estimado sea lo más pequeña pero $\sum e_i = \sum (Y_i - \hat{Y}_i) = 0$ porque la recta estimada corta a los residuales por encima y debajo de manera que se compensa. En consecuencia se debe de minimizar la suma de los cuadrados de cada uno de los residuales.

$$e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \qquad \text{... (1)}$$

$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \quad ; \text{ Por propiedad: } (AB)' = B'A' \Rightarrow Y'X\hat{\beta} = \hat{\beta}'X'Y$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

Derivando respecto a $\hat{\beta}$ e igualando a cero.

$$\frac{de'e}{d\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y \qquad \text{... (2)}$$

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \text{... (3)}$$

Para el caso de 2 variables

$$(X'X) = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} \text{ y } X'Y = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

Sustituyendo en (1) tenemos:

$$\sum Y = \quad n\hat{\beta}_1 \quad + \hat{\beta}_2 \sum X$$

$$\sum XY = \hat{\beta}_1 \sum X \; + \hat{\beta}_2 \sum X^2$$

Para obtener la media y la varianza de $\hat{\beta}$ sustituimos en (3)

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \mu) \qquad \text{... (4)}$$

2. LA VARIANZA DEL TÉRMINO DE PERTURBACION

VARIANZA: $\quad Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

$$S^2 = \hat{\sigma}_\mu^2 = \frac{e'e}{n-k} = \frac{Y'Y - \hat{\beta}x'Y}{n-k}$$

$$S^2 = \frac{\sum (Y_i - \hat{Y}_i)}{n-k} = \frac{\sum e_i^2}{n-k}$$

3. INTERVALO DE CONFIANZA PARA LOS PARAMETROS

A fin de establecer los intervalos de confianza para los coeficientes de regresión (β_i) y teniendo la varianza poblacional desconocida se construye un intervalo asumiendo que esta variable tiene una distribución estadística "t" a partir de las estimaciones de los parámetros y sus varianzas por ejemplo: para β_1

$$Prob(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha$$

$$Prob(-t_{\alpha/2} \le \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \le t_{\alpha/2}) = 1 - \alpha$$

Despejando β_1

$$Prob(\hat{\sigma}_{\beta_1} t_{\alpha/2} \geq -\hat{\beta}_1 + \beta_1 \geq -t_{\alpha/2}\hat{\sigma}_{\beta 1}) = 1-\alpha$$

$$\beta_1 \in \left[\hat{\beta}_1 - \hat{\sigma}_{\beta 1} t_{\alpha/2}, \hat{\beta}_1 + \hat{\sigma}_{\beta 1} t_{\alpha/2}\right] \text{con un nivel de significancia } \alpha$$

Ejemplo:

Número de familia	Ingreso X	Consumo Y
1	80	70
2	100	65
3	120	90
4	140	95
5	160	110
6	180	115
7	200	120
8	220	140
9	240	155
10	260	150

$\mathbf{Y = X \beta + \mu}$

Donde: Y: Consumo
X: Ingreso

$$\begin{bmatrix} 70 \\ 65 \\ 90 \\ . \\ . \\ . \\ 150 \end{bmatrix} = \begin{bmatrix} 1 & 80 \\ 1 & 100 \\ 1 & 120 \\ . & . \\ . & . \\ . & . \\ 1 & 260 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ . \\ . \\ . \\ \mu_{10} \end{bmatrix}$$

$$\hat{\beta} = (x'x)^{-1}x'y$$

$$(x'x)^{-1} = \left[\begin{pmatrix} 1 & 1 & 1.....1 \\ 80 & 100 & 120.....260 \end{pmatrix} \begin{pmatrix} 1 & 80 \\ 1 & 100 \\ 1 & 120 \\ . & . \\ . & . \\ . & . \\ 1 & 260 \end{pmatrix} \right]^{-1} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} = \begin{pmatrix} 10 & 1700 \\ 1700 & 322000 \end{pmatrix}^{-1} =$$

$$\frac{1}{330000} \begin{pmatrix} 322000 & -1700 \\ -1700 & 10 \end{pmatrix} = \begin{pmatrix} 0.975757 & -0.005152 \\ -0.005152 & 0.0000303 \end{pmatrix}$$

$$x'y = \begin{pmatrix} 1.....1 \\ 80....260 \end{pmatrix} \begin{pmatrix} 70 \\ 65 \\ . \\ . \\ . \\ 150 \end{pmatrix} = \begin{pmatrix} 1110 \\ 205500 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} 0.975757 & -0.005152 \\ -0.005152 & 0.0000303 \end{pmatrix} \begin{pmatrix} 1110 \\ 205500 \end{pmatrix} = \begin{pmatrix} 24.4545 \\ 0.50909 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

4. ESTIMACION DE LA VARIANZA DEL TÉRMINO DE PERTURBACION

$$\sigma_\mu^2 = \frac{e'e}{(n-k)} = \frac{Y'Y - \beta'X'Y}{n-k}$$

$$y'y = \begin{pmatrix} 70 & 65.....150 \end{pmatrix} \begin{pmatrix} 70 \\ 65 \\ . \\ . \\ . \\ 150 \end{pmatrix} = 132100$$

$$\hat{\beta}'x'y = \begin{pmatrix} 24.4545 & 0.50909 \end{pmatrix} \begin{pmatrix} 1110 \\ 205500 \end{pmatrix} = 131762.49$$

Reemplazando en la fórmula tenemos:

$$\sigma_{\mu}^2 = \frac{132100 - 131762.49}{10 - 2} = \frac{337.51}{8} = 42.18875$$

Calculando Varianza de los parámetros de regresión:

$$\text{Var}\,\beta_i = (42.18875)\begin{pmatrix} 0.975757 & -0.005152 \\ -0.005152 & 0.0000303 \end{pmatrix}$$

$$\sigma_{\beta 1}^2 = 42.18875(0.975757) = 41.16596813$$
$$\sigma_{\beta 2}^2 = 42.18875(0.0000303) = 0.001278319125$$

la desviación estandarserá :

$$\hat{\sigma}_{\beta 1} = 6.416071082$$
$$\hat{\sigma}_{\beta 2} = 0.0357533$$

5. CONSTRUCCION DE INTERVALOS PARA β_i

$$\beta_i \in \left[\hat{\beta}_i - \hat{\sigma}_{\beta_i}\, t_{\alpha/2}\,,\ \hat{\beta}_i + \hat{\sigma}_{\beta_i}\, t_{\alpha/2}\right]$$

Para un nivel de significación del 5% (α = 0.05) observando en la tabla "t" de student tenemos: $t_{(n-k),\alpha/2} = t_{(10-2)\,0.05/2} = t_{(8)\,0.025} = 2.306$

$$\beta_2 \in \left[0.50909 - 0.0357(2.306), 0.50909 + 0.0357(2.306)\right]$$

$$\beta_2 \in \left[0.4268, 0.5919\right] \text{ con } \alpha = 0.05$$

Otra forma de expresarlo con prob.

$$P(0.4267 \leq \beta_2 \leq 0.5912) = 1 - 0.05 = 0.95$$

Dado un coeficiente de confianza del 95% en el I.p si se construye sin intervalos repetidos con los límites siguientes 0.4268 y 0.919, el 95% de ellos estarían verdadero parámetro poblacional.

6. DOCIMAS DE HIPOTESIS

Se refiere a una distribución de frecuencias, y se plantea con el fin de comprobar si se cumple una relación.

1.- Hipótesis propuesta o sometida a análisis.
H_0: $\beta_i = C$ hipótesis nula
H_1: $\beta_i \neq C$ hipótesis alternativas

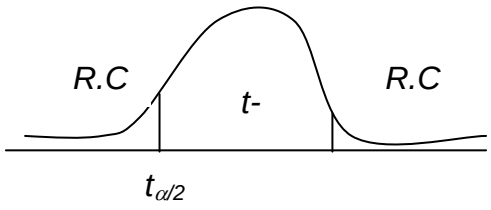
2.- Para comprobar si viene de una misma población con distribución "N" o "t" un caso particular de la docima es:
H_0: $\beta_i = 0$ hipótesis nula
H_1: $\beta_i \neq 0$ hipótesis alternativas

A fin de verificar o comprobar si en la ecuación $Y = \beta_0 + \beta_1 x_i$ β_1 proviene de una población con distribución normal. Si se cumple H_0, X_e son independientes, por lo tanto un cambio en "x" no afectaría "y".

Existen dos reglas o enfoques complementarios para decidir si se rechaza o no la hipótesis nula. Ambos enfoques pretenden que la variable que se considera tiene una distribución de probabilidad y que las pruebas de hipótesis encierran afirmaciones sobre los valores de los parámetros de dicha distribución.

	Verdadero	**Falso**
Aceptar	$1-\beta$	Error tipo II β
Rechazar	Error tipo I α	$1-\alpha$

3.-Establecer una región critica



$(1-\alpha)$ zona de aceptación de hipótesis nula $\beta_1 = 0$

$$C = \{t_c / -t_{\alpha/2} < t_c < t_{\alpha/2}\} = \{t_c / |t_c| \geq t_{n-k,\alpha/2}\}$$

4.- Obtención del estadístico "t"

$$t = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\beta i}} = \frac{\beta_i - 0}{\sigma_{\beta i}} = \frac{\hat{\beta}_i}{\hat{\sigma}_{\beta i}}$$

5.- Comparar el t calculado con el t de tabla con un nivel de significación α

Regla de decisión

Si $/T_C/ > t_{(N-K/\alpha/2)}$ ⟶ se rechaza H_0

Si el t calculado es mayor que el t de tabla, rechaza la hipótesis nula y al rechazar $\beta_i \neq 0$ en consecuencia X_I si explica el comportamiento de la variable dependiente.

El enfoque terminado es el **enfoque de la prueba de significancia.**

7. ENFOQUE DEL INTERVALO DE CONFIANZA

Consiste en hacer un intervalo, en construirlo

$$P(\hat{\beta}_i - \hat{\sigma}_{\beta i}\, t_{n-k,\,\alpha/2} < \beta < \hat{\beta}_i + \sigma_{\beta i} t_{n-k,\,\alpha/2}) = 1 - \alpha$$

El siguiente paso es comparar el β de la hipótesis nula con el intervalo establecido.

Regla de decisión:
Si el β de la H.N está dentro del intervalo se acepta la hipótesis nula, contrariamente si el β está fuera del intervalo rechaza hipótesis.

Ejemplo

- Por la prueba de significancia

$$\hat{\beta}_i = \begin{pmatrix} 24.4545 \\ 0.50909 \end{pmatrix} \qquad \hat{\sigma}_{\beta i} = 6.4160$$

$$\hat{\sigma}_{\beta 2} = 0.0357$$

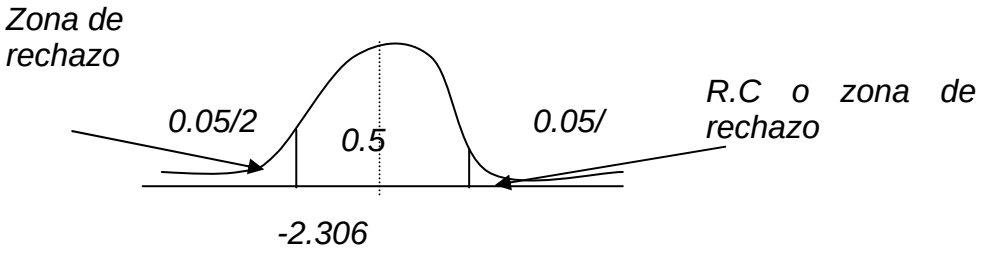
1. H_0: $\beta_2 = 0$
   H_1: $\beta_2 \neq 0$

2.- $\alpha = 0.05$

3.- $RC = \{/ t_c / > t_{10-2,0.05/2}\}$

4.- $t_c = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\beta i}} = \frac{0.50909 - 0}{0.0357} = 14.3$, $t_{10-2,0.05/2} = 2.306$

Zona de



5.- $|t_c| > t_{8,0.025}$

$14.3 > 2.306$

- Por el Enfoque del Intervalo de Confianza:

$$P(\hat{\beta}_i - \hat{\sigma}_{\beta i}, t_{n-k/\alpha/2} < \beta < \hat{\beta}_i + \sigma_{\beta i} t_{n-k,\alpha/2}) = 1 - \alpha$$

P(0.5091-0.0357(}2.306)<β<0.5091+0.0357(2.306))=0.95

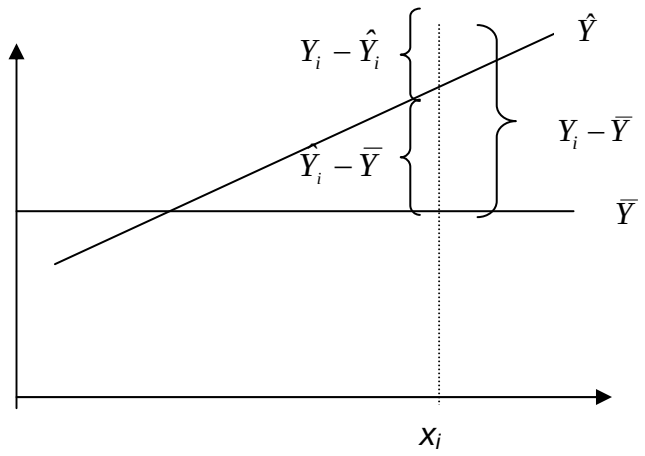
P(0.4268<β<0.5914)

P=0.95

8. COEFICIENTE DE DETERMINACION (R^2)

Es un indicador de la bondad de ajuste de la línea de regresión que mide la proporción de la variación total en la variable dependiente Y, que "se explica" o "se debe a" la variación de la variable independiente X.

El rango de R^2es el siguiente: $0 \leq R^2 \leq 1$.



Planteada la relación inicial, la misma se mantiene cuando se establece relaciones a partir de las sumatorias de sus desviaciones cuadráticas.
Por un proceso matemático particular se da:

$$\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \overline{Y})^2$$

SCT=SCR+SCE

SCT: variación total del Y_i observado con respecto a su media muestral.
La suma total de los cuadrados.

$$\sum (Y_i - \overline{Y})^2 = Y'Y - n\overline{Y}^2$$

SCR: Variación residual o no explicada de los valores de Y respecto a la línea de regresión.
Suma de los cuadrados residuales.

$$\sum (Y_i - \hat{Y}_i)^2 = Y'Y - \hat{\beta}'X'Y$$

SCE: Variación de los valores estimados $\hat{Y}_i$ con respecto a su media.
Suma de los cuadrados Explicados.

$$\sum (\hat{Y}_i - \overline{Y})^2 = \hat{\beta}'X'Y - n\overline{Y}^2$$

Resumiendo tenemos:

$$\left(Y'Y - \hat{\beta}'X'Y\right) + \left(\hat{\beta}'X'Y - n\overline{Y}^2\right) = \left(Y'Y - n\overline{Y}^2\right)$$

$$\boxed{R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \qquad 0 \le R^2 \le 1}$$

PROPIEDADES

1. Es una cantidad no negativa

2. Sus límites son $0 \le R^2 \le 1$

Es decir que R varía entre cero y uno

R^2=1 cuando el ajuste es perfecto, es decir los valores observados coinciden perfectamente con la recta estimada

$R^2 \approx 0$ es decir que no hay relación entre la variable dependiente y los variables explicativas.

Este R^2 no mide el grado de asociación entre x e y, para lo cual se acude a otro indicador

9. COEFICIENTE DE CORRELACION

Es una medida de asociación lineal entre dos variables

Poblacional **Muestral**

$$r = \frac{Cov(x,y)}{\sqrt{\sigma^2_x \sigma^2_y}} = \frac{\sum (x_i - \bar{x})(x_i - \bar{y})}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1} \frac{\sum (y - \hat{y})^2}{n-1}}}$$

$$r = \sqrt{R^2} \qquad 0 \le r \le 1$$

9.1. PROPIEDADES

Sus límites están entre: $0 \le r \le 1$

Es de naturaleza simétrica, es decir el coeficiente de correlación entre X y Y (r_{xy}) es igual al coeficiente de correlación entre Y y X (r_{yx})

Si X, Y son estadísticamente independientes, el coeficiente de correlación es cero; pero si r=0 no implica necesariamente independencia.

Es una medida de asociación lineal, es decir mide la asociación lineal entre dos variables. .

9.2. COEFICIENTE DE DETERMINACION MULTIPLE CORREGIDO

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \qquad 0 \le R^2 \le 1$$

En la medida que el número de variables independientes se incrementa, el R^2 tiende a incrementarse por tener el numerador de la fracción en mayor valor.
Una alternativa, es cambiar esta expresión dividiendo a cada uno de la sumatorias cuadráticas entre sus grados de libertad, obteniendo finalmente un cociente de varianza.

$$\boxed{\bar{R}^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2 / n-k}{\sum (Y_i - \bar{Y})^2 / n-1}}$$

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2_\mu}{\hat{\sigma}^2_y} \qquad \hat{\sigma}_\mu{}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-k} \quad , \quad \hat{\sigma}_y{}^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

Ejemplo:

En la práctica se calcula de la siguiente manera:

$$R^2 = \frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{\hat{\beta}' X' Y - n\overline{Y}^2}{Y' Y - n\overline{Y}^2} \quad \text{.....(1)}$$

$$\overline{R}^2 = 1 - \frac{\hat{\sigma}^2_{\mu}}{\hat{\sigma}^2_{y}} \quad \text{.....(2)}$$

$$\hat{\sigma}^2_{\mu} = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-k} = \frac{Y'Y - \hat{\beta}'(X'Y)}{n-k}$$

$$\sigma^2_{y} = \frac{\sum\left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{Y'Y - n\overline{Y}^2}{n-1}$$

En (1):

$$\frac{131764.5 - 10(111)^2}{132100 - 10(111)^2} = \frac{1317645 - 123210}{123210} = 0.96$$

En (2)

$$\sigma^2_{\mu} = \frac{132100 - 131762.49}{10-2} = \frac{337.51}{8} = 42.18$$

$$\sigma^2_{\mu} = \frac{132100 - 10(111)^2}{10-1} = \frac{8890}{9} = 987.8$$

Reemplazando tenemos:

$$\overline{R}^2 = 1 - \frac{42.18}{987.8} = 0.99$$