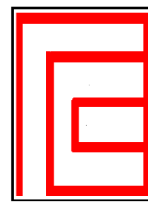




UNIVERSIDAD NACIONAL DE TUCUMAN

Facultad de Ciencias Económicas

Instituto de Investigaciones Estadísticas (INIE)



LICENCIATURA EN ADMINISTRACIÓN

LICENCIATURA EN ECONOMÍA

ESTADÍSTICA INFERENCIAL

AÑO 2025

RESUMEN

TEORIA PRIMER PARCIAL

CAPÍTULO 1: EXPERIMENTO ALEATORIO

Experimento aleatorio. Espacio muestral. Técnicas de conteo de puntos muestrales.

Objetivos:

El alumno debe ser capaz de:

- Encontrar el espacio muestral asociado a un experimento aleatorio dado.
- Aplicar fórmulas de conteo para determinar el número de elementos de algunos espacios muestrales.
- Reconocer bajo qué condiciones un experimento es aleatorio.

Resumen

D1. Experimento Aleatorio: Es un proceso que repetido bajo un mismo conjunto de condiciones controladas no conduce siempre a un mismo resultado.

D2. Espacio Muestral (\mathcal{S}): Es el conjunto de todos los posibles resultados de un experimento aleatorio.

D3. Acontecimiento, evento o suceso: Es un subconjunto del espacio muestral.

D4. Acontecimiento o suceso elemental: Se denomina así al conjunto formado por un único elemento del espacio muestral.

Conteo

T1. Principio de la multiplicación: En un experimento aleatorio que se realiza en dos partes o etapas, donde la primera parte tiene n_1 resultados posibles, y por cada resultado posible de la primera parte, la segunda parte tiene n_2 resultados posibles, el espacio muestral tendrá exactamente $n_1 \times n_2$ resultados.

T2. Principio de la adición: En un experimento aleatorio que se puede realizar de dos maneras mutuamente excluyente, donde la primera manera tiene n_1 resultados posibles, la segunda tiene n_2 resultados posibles, el espacio muestral tendrá exactamente $n_1 + n_2$ resultados posibles.

D5. Combinación: Se denomina combinación a una colección o grupo de elementos (el orden no es importante).

D6. Variación o permutación: Se denomina variación o permutación a cada secuencia u ordenación de elementos (el orden es importante).

Obs: Dos combinaciones son diferentes si difieren en el número de elementos o en al menos un elemento. Dos variaciones son diferentes si difieren en el número de elementos, en algún elemento, o en el orden de presentación de al menos un elemento.

T3. El número de variaciones posibles con n objetos distintos es $P_n = n!$.

T4. El número de variaciones posibles de n objetos distintos tomados de r a la vez es

$$V_{n,r} = \frac{n!}{(n-r)!}.$$

T5. El número de variaciones posibles con n objetos distintos ubicados en un círculo es $(n - 1)!$.

T6. El número de variaciones posibles con n objetos no todos distintos, de los cuales n_1 son de un tipo, n_2 son de un segundo tipo, ..., n_k son de un k -ésimo tipo es

$$\frac{n!}{n_1! n_2! \dots n_k!}.$$

T7. El número de formas posibles de partir un conjunto de n objetos en r celdas con n_1 elementos en la primera celda, n_2 en la segunda, ..., n_r elementos en la r -ésima celda es

$$\frac{n!}{n_1! n_2! \dots n_r!}.$$

T8. El número de combinaciones posibles de n objetos distintos tomados de r a la vez es

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

Ejemplos

Ejemplo 1: Indicar el espacio muestral más apropiado para cada uno de los siguientes experimentos aleatorios.

- Se arroja un dado y una moneda y se observa el número de la cara superior del dado y la cara de la moneda que se obtienen.
- Se observa el tiempo que demora el sistema informático de un banco en registrar el depósito realizado en efectivo en un cajero automático del interior del país y transferirlo a una cuenta corriente de una sucursal del Gran Bs. As.
- Se observa la cantidad de consultas médicas en la guardia del Hospital de Niños, en una semana del año elegida al azar, hasta que se detecta un caso con meningitis.
- Un supermercado de la ciudad de Yerba Buena ha inaugurado una nueva modalidad de compra por Internet que permite el pago con tarjeta de crédito. Cuando las facturas se realizan con errores se producen quejas de los clientes que demoran el pago por parte de la entidad que emitió la tarjeta de crédito. Se eligen diez facturas al azar y se registran los montos de los errores de facturación.

Resolución:

- Si se acuerda en observar primero el resultado del dado y luego el de la moneda, el resultado de dicha observación será un par ordenado, y el espacio muestral de todos los posibles pares ordenados será

$$\mathcal{S} = \{(x, y)/x = 1, 2, 3, 4, 5, 6; y = \text{cara, sello}\}.$$

- Al observar el tiempo que demora un sistema informático en registrar el depósito realizado en efectivo en un cajero automático del interior del país y transferirlo a una cuenta corriente de una sucursal del Gran Buenos Aires, podría ocurrir que por una falla del sistema el depósito y transferencia demoraran un plazo mayor al habitual. Por tal motivo se puede considerar no acotado el conjunto de resultados posibles.

El tiempo es continuo y puede ser igual a cualquier valor real positivo.

$$\mathcal{S} = \{x/x \in \mathbb{R}^+\} = (0, +\infty).$$

- Cuando se observa la cantidad de niños que realizan consultas en la guardia del Hospital de Niños, durante una semana, hasta que se detecta el primer caso de meningitis, puede suceder que el primero que concurra sea por meningitis, o el segundo, o el tercero, etc.

Al ser observaciones resultantes de una enumeración los resultados posibles son los números naturales.

$$\mathcal{S} = \{x/x \in \mathbb{N}\}.$$

- d) Al estudiar los montos de los errores de facturación de 10 facturas elegidas al azar se observa una 10-upla ordenada.

Cualquier factura seleccionada al azar puede no presentar ningún error o presentar algún error por exceso o por defecto, consecuentemente al observar los montos de los errores ellos pueden ser positivos, negativos o nulos.

Si se consideran errores de hasta centavos de peso, el espacio muestral no son todos los números reales, sino sólo aquellos que admiten una representación decimal con un máximo de dos decimales.

$$\mathcal{S} = \{(x_1, x_2, \dots, x_{10}) \in \mathbb{R}^{10} : x_i \text{ tiene una representación decimal con hasta dos decimales}\}.$$

Ejemplo 2: A un grupo de 200 estudiantes que comienzan a cursar una materia *Inferencia Estadística* se les pregunta si ya han aprobado *Estadística Descriptiva*, *Matemática I* y *Matemática II*. Los resultados fueron los siguientes: 150 ya aprobaron *Matemática I*, 130 *Estadística Descriptiva* y 80 *Matemática II*, 60 aprobaron las tres materias, 120 *Matemática I* y *Estadística Descriptiva*, 70 *Matemática I* y *Matemática II* y 65 *Estadística Descriptiva* y *Matemática II*. Representar en un diagrama de Venn la situación anteriormente descrita y responder cuántos de los alumnos que comenzaron a cursar *Estadística I*:

- Aprobaron al menos una de las tres materias.
- Aprobaron las materias *Estadística Descriptiva* y *Matemática I* pero no *Matemática II*.
- Aprobaron sólo la materia *Estadística Descriptiva*.
- No aprobaron ninguna de las tres materias.

Resolución:

Sean los conjuntos

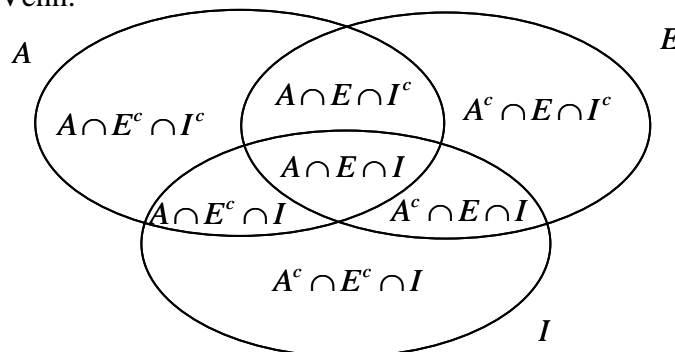
$$U = \{x/x \text{ es alumno que comienza a cursar Inferencia Estadística}\},$$

$$A = \{x/x \text{ alumno de } U \text{ que aprobó Matemática II}\},$$

$$E = \{x/x \text{ alumno de } U \text{ que aprobó Estadística Descriptiva}\},$$

$$I = \{x/x \text{ alumno de } U \text{ que aprobó Matemática I}\}.$$

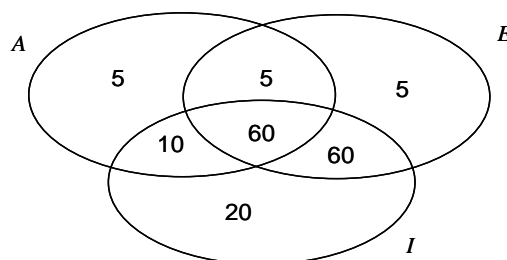
Para poder organizar la información suministrada en este ejercicio, se determinan los cardinales (número de elementos si el conjunto es finito) de cada conjunto, desde el conjunto más restrictivo hasta el menos restrictivo; para finalmente escribirlos en el diagrama de Venn.



Se denota con $|A|$ al cardinal del conjunto A .

- 60 estudiantes aprobaron las tres materias, consecuentemente $|A \cap E \cap I| = 60$
- Se sabe que $A \cap E = (A \cap E \cap I) \cup (A \cap E \cap I^c)$, por lo tanto $|A \cap E| = |A \cap E \cap I| + |A \cap E \cap I^c|$ debido a que los conjuntos $A \cap E \cap I$ y $A \cap E \cap I^c$ son mutuamente excluyentes. Consecuentemente
 $|A \cap E \cap I^c| = |A \cap E| - |A \cap E \cap I| = 65 - 60 = 5$, ya que 65 alumnos aprobaron *Estadística Descriptiva* y *Matemática II* y 60 aprobaron las tres materias.
- De manera análoga:
 $|A \cap E^c \cap I| = |A \cap I| - |A \cap E \cap I| = 70 - 60 = 10$, pues 70 aprobaron *Matemática I* y *Matemática II*, y
 $|A^c \cap E \cap I| = |E \cap I| - |A \cap E \cap I| = 120 - 60 = 60$ pues 120 aprobaron *Matemática I* y *Estadística Descriptiva*.
- $A = (A \cap E^c \cap I^c) \cup (A \cap E \cap I^c) \cup (A \cap E^c \cap I) \cup (A \cap E \cap I)$
 $|A| = |A \cap E^c \cap I^c| + |A \cap E \cap I^c| + |A \cap E^c \cap I| + |A \cap E \cap I|$ por ser los conjuntos $A \cap E^c \cap I^c$, $A \cap E \cap I^c$, $A \cap E^c \cap I$ y $A \cap E \cap I$ mutuamente excluyentes, consecuentemente,
 $|A \cap E^c \cap I^c| = |A| - |A \cap E \cap I^c| - |A \cap E^c \cap I| - |A \cap E \cap I| = 80 - 5 - 10 - 60 = 5$ pues 80 aprobaron *Matemática II*.
- De manera análoga:
 $|A^c \cap E \cap I^c| = |E| - |A \cap E \cap I^c| - |A^c \cap E \cap I| - |A \cap E \cap I| = 130 - 5 - 60 - 60 = 5$ y
 $|A^c \cap E^c \cap I| = |I| - |A \cap E^c \cap I| - |A^c \cap E \cap I| - |A \cap E \cap I| = 150 - 10 - 60 - 60 = 20$, teniendo en cuenta que 130 ya aprobaron *Estadística Descriptiva* y 150 *Matemática I*.

Si se escribe en cada conjunto su cardinal se obtiene:



- $A \cup E \cup I = \{x / x \text{ aprobó al menos una asignatura}\}$
 $|A \cup E \cup I| = 5 + 5 + 5 + 10 + 60 + 60 + 20 = 165$
- $A^c \cap E \cap I = \{x / x \text{ aprobó Matemática I y Estadística Descriptiva pero no Matemática II}\}$
 $|A^c \cap E \cap I| = 60$
- $A^c \cap E \cap I^c = \{x / x \text{ aprobó sólo Estadística}\}$ $|A^c \cap E \cap I^c| = 5$
- $(A \cup E \cup I)^c = \{x / x \text{ no aprobó ninguna de las tres materias}\}$
 $|(A \cup E \cup I)^c| = |U| - |A \cup E \cup I| = 200 - 165 = 35$

Ejemplo 3: Un empleado tiene facultad para escoger un curso de capacitación en finanzas o en administración de riesgos, cada uno de los cuales se ofrecen en tres horarios y con cuatro diferentes instructores. ¿Cuántas opciones se ofrecen al empleado?

Resolución:

Elegir un curso de capacitación, un horario y un instructor, sin tener a priori ninguna preferencia, puede pensarse como un experimento aleatorio que consta de tres etapas: la elección del curso, la elección del horario y la elección del instructor.

La elección del curso se puede hacer de $n_1 = 2$ maneras posibles, de Finanzas o de Administración de riesgos.

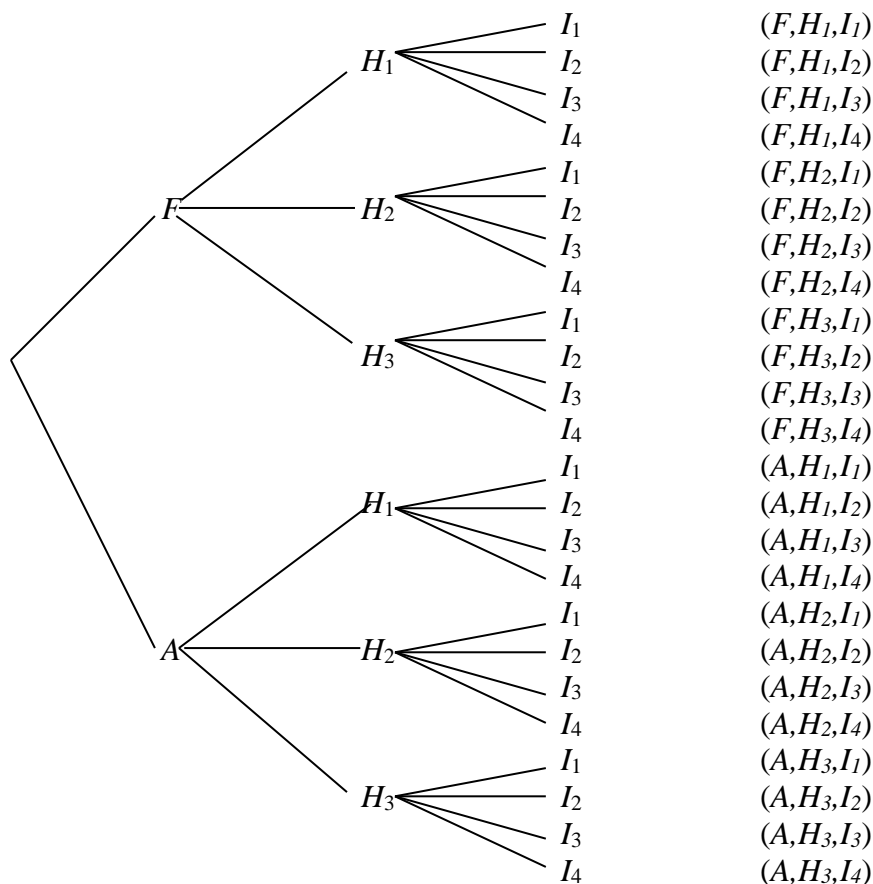
Por cada selección del curso, se pueden hacer $n_2 = 3$ elecciones posibles de horarios.

Por cada selección del curso y del horario, se pueden hacer $n_3 = 4$ elecciones posibles de instructores.

Entonces, por el Principio de la Multiplicación, el empleado tiene $n_1.n_2.n_3 = 2.3.4 = 24$ opciones distintas.

También se podrían contar las opciones enumerándolas en un diagrama de árbol.

Si se designa con F al curso de Finanzas, con A al curso de Administración de riesgos, con H_1, H_2 y H_3 los tres horarios y con I_1, I_2, I_3, I_4 los cuatro instructores.

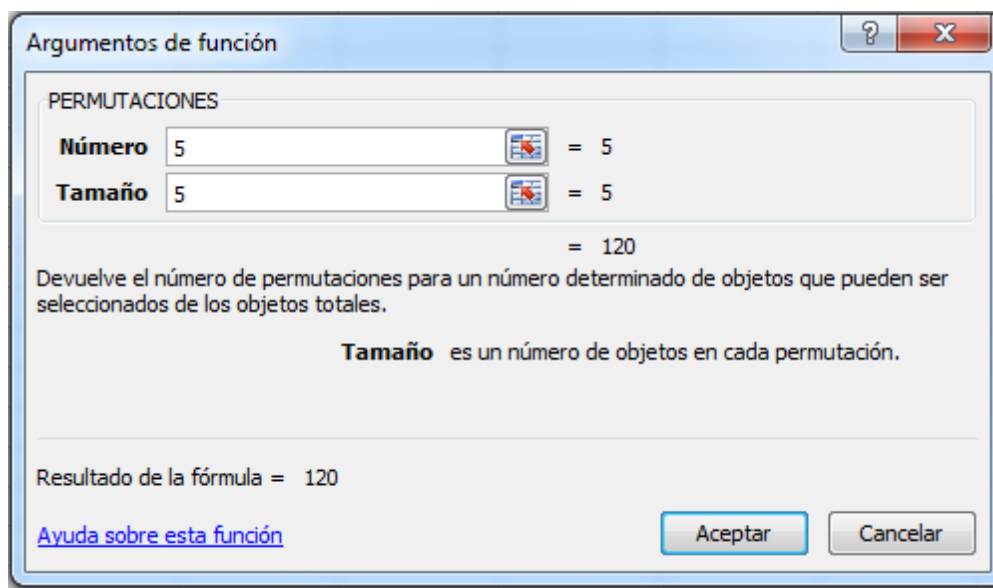


Ejemplo 4: Un viajante debe visitar cinco ciudades en un día para entregar mercadería. De acuerdo al orden en que visite las ciudades será el costo del viaje, pues los precios de los hoteles, restaurantes y la distancia a recorrer serán distintos en cada caso. ¿Los costos de cuántos viajes diferentes debe evaluar?

Resolución:

El orden en que se realizan las visitas a las ciudades es importante porque determina el costo del viaje. Entonces, la cantidad de recorridos por las cinco ciudades se puede contar con la cantidad de variaciones de las 5 ciudades, es decir, $P_5 = 5! = 120$ formas distintas

Nota: Este ejercicio se puede resolver con Microsoft Excel empleando el Menú Fórmulas, Insertar Función, PERMUTACIONES y, como se quiere saber los costos de cuántos viajes se deben calcular, se puede contar con la cantidad de variaciones de las 5 ciudades:



Ejemplo 5: Un inversionista desea integrar un portafolio con cinco acciones. Si hay 30 acciones en el mercado, ¿de cuántas maneras lo puede formar?

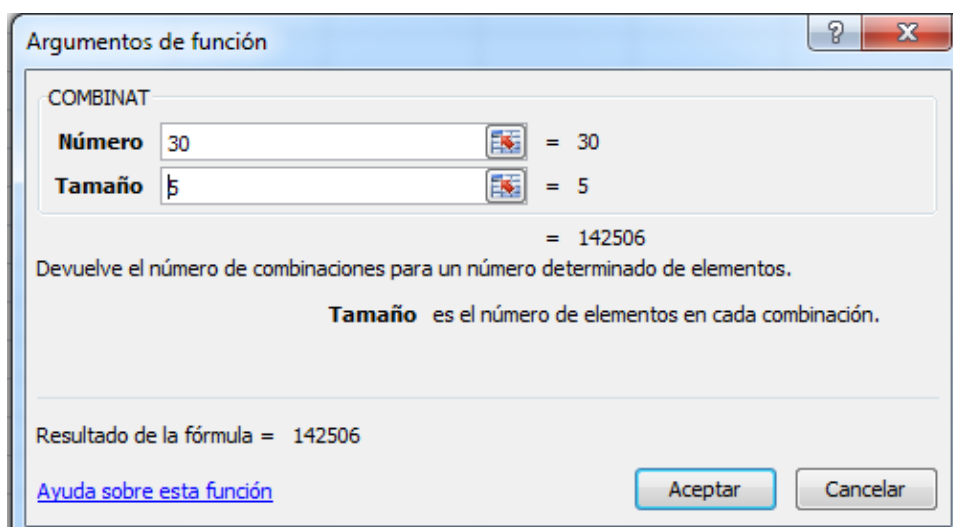
Resolución:

Al elegir un portafolio con 5 acciones no es importante el orden en la selección porque, por ejemplo, {acción 1, acción 3, acción 7, acción 10, acción 30} es el mismo portafolio que {acción 3, acción 10, acción 1, acción 30, acción 7}.

Al no importar el orden, se cuenta la cantidad de portafolios posibles con el número de combinaciones.

Esto es $\binom{30}{5} = \frac{30!}{5!(30-5)!} = \frac{30 \cdot 29 \cdot 28 \cdot 27 \cdot 26 \cdot 25!}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 25!} = 142.506$ elecciones posibles de portafolios.

Nota: En Microsoft Excel este ejercicio se resuelve con la fórmula COMBINAT que se encuentra en el Menú Fórmulas, Insertar Función, Todas y COMBINAT. Se cuenta la cantidad de portafolios posibles con el número de combinaciones:



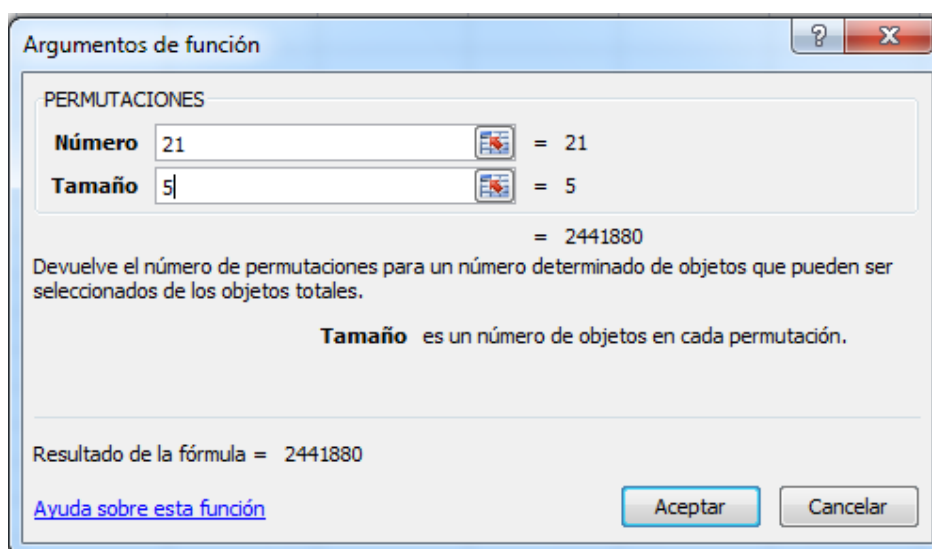
Ejemplo 6: El director técnico de la selección nacional viaja a un torneo de básquet con 21 jugadores. Suponiendo que todos los jugadores pueden jugar en cualquier posición, ¿cuántos equipos diferentes puede formar? (Nota: se considera que ordenamientos distintos de los mismos jugadores constituyen equipos diferentes)

Resolución:

Los ordenamientos distintos de los 5 jugadores seleccionados constituyen equipos diferentes, entonces es importante tener en cuenta el orden. Consecuentemente, el número de equipos que se pueden formar se cuenta con el número de variaciones de 21 jugadores tomados de 5 a la vez.

$$\text{Son } V_{21,5} = \frac{21!}{(21-5)!} = \frac{21 \cdot 20 \cdot 19 \cdot 18 \cdot 17 \cdot 16!}{16!} = 2.441.880 \text{ equipos diferentes.}$$

Nota: En Microsoft Excel este ejercicio se resuelve con la fórmula PERMUTACIONES que se encuentra en el Menú Fórmulas, Insertar Función, Todas y PERMUTACIONES. Por lo tanto, el número de equipos que se pueden formar con 21 jugadores tomados de 5 a la vez es:



CAPÍTULO 2: PROBABILIDAD

Definición de probabilidad. Asignación de probabilidad a un evento. Reglas aditivas. Probabilidad condicional. Independencia de sucesos. Reglas multiplicativas. Probabilidad total. Regla de Bayes.

Objetivos:

El alumno debe ser capaz de:

- Reconocer cuándo una función es una función de probabilidad.
- Comprender la relación entre los axiomas de probabilidad y las características de la frecuencia relativa.
- Identificar en qué situaciones es factible usar la definición clásica de probabilidad.
- Calcular probabilidades condicionales utilizando probabilidades no condicionales.
- Determinar si dos sucesos son independientes analizando si la ocurrencia de uno influye en la probabilidad de ocurrencia del otro.
- Calcular la probabilidad de que una “causa” haya originado una consecuencia cuando se conoce que ésta efectivamente ocurrió, usando el Teorema de Bayes.

Resumen

D1. Definición de Probabilidad de Laplace: Sea un espacio muestral \mathcal{S} y acéptese “a priori” que, por la simetría del experimento aleatorio, una persona razonable se sentiría indiferente con respecto a la preferencia de uno cualquiera de los acontecimientos elementales. Entonces la probabilidad de un acontecimiento es proporcional al número de acontecimientos elementales que lo forman.

D2. Definición Frecuentista de Probabilidad: Si cada vez que se realiza una serie suficientemente grande de repeticiones de un fenómeno aleatorio la razón (frecuencia relativa) del número de veces que un suceso A ocurre al número total de repeticiones, $f_n(A)$, es aproximadamente p , y si la frecuencia relativa $f_n(A)$ es habitualmente más próxima a p cuando mayor es el número de repeticiones, entonces se acepta de antemano en definir $P(A) = p$.

D3. Definición Axiomática de Probabilidad: Sea \mathcal{S} el espacio muestral de un experimento aleatorio. Una probabilidad P es una función que asigna a cada acontecimiento A de \mathcal{S} un número $P(A)$, llamado la probabilidad de A , que satisface los siguientes axiomas: (i) $P(A) \geq 0$, para todo acontecimiento A , (ii) $P(\mathcal{S}) = 1$, y (iii) para cualquier sucesión de acontecimientos mutuamente excluyentes A_1, A_2, \dots $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Obs: Las definiciones **D1** y **D2** pueden considerarse definiciones operacionales, pues indican la manera de calcular la probabilidad, en algunas situaciones particulares. La definición **D3** es conceptual, indica los axiomas que debe cumplir cualquier probabilidad, más no especifica cómo calcularlas.

Propiedades: Denotando por \mathcal{A} el conjunto de todos los acontecimientos de \mathcal{S} , se tiene que:

- | | |
|---|---|
| 1) $P(\emptyset) = 0$ | 4) $A, B \in \mathcal{A}, A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$ |
| 2) $A \in \mathcal{A}, P(A^c) = 1 - P(A)$ | 5) $A, B \in \mathcal{A}, P(B \cap A^c) = P(B) - P(A \cap B)$ |
| 3) $A \in \mathcal{A}, P(A) \leq 1$ | 6) $A, B \in \mathcal{A}, P(B \cup A) = P(A) + P(B) - P(A \cap B)$ |

Obs: Cuando \mathcal{S} es un conjunto finito, se puede considerar que \mathcal{A} es el conjunto de todos los subconjuntos de \mathcal{S} . Ese conjunto se denomina **conjunto partes de \mathcal{S}** y se denota con $\mathcal{P}(\mathcal{S})$.

D4. Probabilidad Condicional: Sean $A, B \in \mathcal{A}$, con $P(B) \neq 0$, la probabilidad condicional de A dado B , que se indica por $P(A|B)$, se define como

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

D5. Eventos independientes: A y B son independientes si $P(A \cap B) = P(A)P(B)$.

T1. Sean $A, B \in \mathcal{A}$, con $P(B) \neq 0$, entonces $P(A \cap B) = P(A|B)P(B)$.

T2. Sean $A, B \in \mathcal{A}$, con $P(B) \neq 0$, entonces A y B son independientes si $P(A|B) = P(A)$.

D6. Partición del Espacio Muestral: Los eventos C_1, C_2, \dots, C_k constituyen una partición del espacio muestral \mathcal{S} si (i) $\forall i, j$ con $i \neq j, C_i \cap C_j = \emptyset$ y (ii) $\bigcup_{i=1}^k C_i = \mathcal{S}$.

T3. Teorema de la Probabilidad Total: Si C_1, C_2, \dots, C_k constituye una partición del espacio muestral \mathcal{S} , con $P(C_i) \neq 0$ para todo $i = 1, \dots, k$, entonces para cualquier evento $A \in \mathcal{A}$

$$P(A) = \sum_{i=1}^k P(C_i)P(A|C_i).$$

T4. Teorema de Bayes: Si C_1, C_2, \dots, C_k constituye una partición del espacio muestral \mathcal{S} , con $P(C_i) \neq 0$ para todo $i = 1, \dots, k$, entonces para cualquier evento $A \in \mathcal{A}$, con $P(A) \neq 0$

$$P(C_r|A) = \frac{P(C_r)P(A|C_r)}{\sum_{i=1}^k P(C_i)P(A|C_i)}, \text{ para } r = 1, \dots, k.$$

Ejemplos

Ejemplo 1: Identificar qué definición de probabilidad se aplicaría para calcular la probabilidad de que al arrojar dos dados la suma de los resultados dé 7. Calcular dicha probabilidad.

Resolución:

Si se lanzan dos dados, se observa un par ordenado, entonces el conjunto de resultados posibles del experimento es $\mathcal{S} = \{(x, y) / x, y = 1, 2, 3, 4, 5, 6\}$.

El suceso de interés es $A = \{(x, y) / x + y = 7\} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$.

Si el dado no está cargado todas las caras del dado tienen la misma posibilidad de ocurrir, por lo que es posible aplicar la definición de probabilidad de Laplace (**D1**) para calcular la probabilidad del suceso A . De esta manera:

$$P(A) = \frac{|A|}{|\mathcal{S}|} = \frac{6}{36} = \frac{1}{6}.$$

Ejemplo 2: Dado $\mathcal{S} = \{1, 2, 3\}$, probar que la función $P: \mathcal{P}(\mathcal{S}) \rightarrow \mathbb{R}$ definida a continuación representa una función de probabilidad.

$$P(\mathcal{S}) = 1; P(\emptyset) = 0; P(\{1\}) = 0,3; P(\{2\}) = 0,5; P(\{3\}) = 0,2;$$

$$P(\{1, 2\}) = 0,8; P(\{1, 3\}) = 0,5; P(\{2, 3\}) = 0,7$$

Resolución:

Una función es una función de probabilidad cuando cumple con los axiomas (i), (ii) y (iii) de la definición **D3**.

El conjunto $\mathcal{P}(\mathcal{S})$ en este caso es el siguiente

$$\mathcal{P}(\mathcal{S}) = \{\mathcal{S}, \emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}$$

- Por la forma en que se definió P , se tiene que $P(A) \geq 0$, para todo acontecimiento $A \in \mathcal{P}(\mathcal{S})$, y por tanto se satisface el axioma (i)
- Por definición de P , $P(\mathcal{S}) = 1$ y consecuentemente el axioma (ii) se satisface.
- Por tratarse de un espacio muestral finito, el conjunto de acontecimientos es finito y consecuentemente cualquier sucesión de acontecimientos disjuntos se reduce a una unión finita de acontecimientos no vacíos en $\mathcal{P}(\mathcal{S})$. Las posibles uniones finitas de acontecimientos disjuntos no vacíos son las siguientes:

$$\{1\} \cup \{2\} = \{1, 2\}, P(\{1\}) + P(\{2\}) = 0,3 + 0,5 = 0,8 = P(\{1, 2\})$$

$$\{1\} \cup \{3\} = \{1, 3\}, P(\{1\}) + P(\{3\}) = 0,3 + 0,2 = 0,5 = P(\{1, 3\})$$

$$\begin{aligned}\{2\} \cup \{3\} &= \{2, 3\}, P(\{2\}) + P(\{3\}) = 0,5 + 0,2 = 0,7 = P(\{2, 3\}) \\ \{1\} \cup \{2, 3\} &= S, P(\{1\}) + P(\{2, 3\}) = 0,3 + 0,7 = 1 = P(S) \\ \{2\} \cup \{1, 3\} &= S, P(\{2\}) + P(\{1, 3\}) = 0,5 + 0,5 = 1 = P(S) \\ \{3\} \cup \{1, 2\} &= S, P(\{3\}) + P(\{1, 2\}) = 0,2 + 0,8 = 1 = P(S)\end{aligned}$$

Se puede observar que todas cumplen con el axioma (iii)

Consecuentemente, la función P define una función de probabilidad en $\mathcal{P}(S)$.

Ejemplo 3: Dado $S = \{1, 2, 3, 4\}$, indicar por qué las siguientes funciones $P: \mathcal{P}(S) \rightarrow \mathbb{R}$ no pueden definir una función de probabilidad.

- a) $P(\{1\}) = 0,3, P(\{2\}) = 0,4, P(\{3\}) = P(\{4\}) = 0,25$
b) $P(\{1\}) = 0,3, P(\{2\}) = 0,2, P(\{3\}) = P(\{4\}) = 0,25, P(\{1, 2\}) = 0,6$

Resolución:

Una función es una función de probabilidad cuando cumple con los axiomas i), ii) y iii) de la definición **D3**.

- a) Razonando por el absurdo, si se tratara de una probabilidad, por el tercer axioma se tendría que

$$P(S) = P(\{1\} \cup \{2\} \cup \{3\} \cup \{4\}) = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) = 1,2$$

resultado inconsistente con el axioma 2 que dice $P(S) = 1$. Consecuentemente P no es una probabilidad.

- b) Razonando por el absurdo, si se tratara de una probabilidad, por el tercer axioma se tendría que

$$P(\{1, 2\}) = P(\{1\}) + P(\{2\}) = 0,5 \neq 0,6.$$

Consecuentemente P no es una probabilidad.

Ejemplo 4:

Si $P(A) = \frac{1}{7}$, $P(B) = \frac{1}{5}$ y $P(A \cap B) = \frac{1}{8}$, calcule: a) $P(A^c)$; b) $P(A^c \cap B^c)$; c) $P(A^c \cap B)$ y d) $P(A^c \cup B^c)$.

Resolución:

- a) $P(A^c) = 1 - P(A) = 1 - \frac{1}{7} = \frac{6}{7}$, por propiedad 2 de probabilidad.

- b) $A^c \cap B^c = (A \cup B)^c$ por ley de De Morgan (el complemento de la unión es la intersección de los complementos). Luego

$$P(A^c \cap B^c) = P[(A \cup B)^c] = 1 - P(A \cup B)$$

$$\text{Por propiedad 6, } P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{7} + \frac{1}{5} - \frac{1}{8} = \frac{61}{280}.$$

$$\text{Entonces, } P(A^c \cap B^c) = 1 - \frac{61}{280} = \frac{219}{280}$$

- c) $P(A^c \cap B) = P(B) - P(A \cap B) = \frac{1}{5} - \frac{1}{8} = \frac{3}{40}$.

- d) $A^c \cup B^c = (A \cap B)^c$ por ley de De Morgan (el complemento de la intersección es la unión de los complementos). Luego

$$P(A^c \cup B^c) = P[(A \cap B)^c] = 1 - P(A \cap B) = 1 - \frac{1}{8} = \frac{7}{8}.$$

Ejemplo 5: El gerente de una fábrica de focos desea determinar cuál es la probabilidad de producir un foco defectuoso. Indicar cómo podría ser determinada esta probabilidad y qué definición se utilizaría.

Resolución:

Para conocer la probabilidad de producir un foco defectuoso, el gerente de la fábrica puede observar la condición de defectuoso o no defectuoso de una **gran cantidad** de focos producidos y calcular la razón (frecuencia relativa) del número de veces que observa un foco defectuoso entre la cantidad de focos observados. Si al aumentar el número de focos que se observan, esa razón se mantiene estable alrededor de un valor P , entonces se acepta en definir $P(A) = p$.

En este caso se está usando la definición frecuentista de probabilidad.

Ejemplo 6: Un experimento consiste en lanzar una moneda no convencional, para la cual la probabilidad de obtener una cara es 2 veces la de obtener sello. Escribir el espacio muestral asociado a este experimento y calcular la probabilidad de cada suceso elemental.

Resolución:

El espacio muestral asociado al experimento es: $\mathcal{S} = \{\text{cara}, \text{sello}\}$.

No es posible asignar probabilidad a cada suceso usando la definición clásica o de Laplace porque las probabilidades de los sucesos elementales no son iguales.

Se usará la definición axiomática que establece que la probabilidad del espacio muestral es igual a uno y que la probabilidad de la unión de sucesos elementales es igual a la suma de las probabilidades.

$$\begin{aligned}\mathcal{S} = \{\text{cara}\} \cup \{\text{sello}\} &\Rightarrow P(\mathcal{S}) = P(\{\text{cara}\}) + P(\{\text{sello}\}) \\ 1 &= 2P(\{\text{sello}\}) + P(\{\text{sello}\}) \\ 1 &= 3P(\{\text{sello}\}) \\ \frac{1}{3} &= P(\{\text{sello}\})\end{aligned}$$

y consecuentemente la $P(\{\text{cara}\}) = \frac{2}{3}$.

Ejemplo 7: El gobierno de la provincia de Tucumán invitó a las empresas constructoras registradas para presentarse a la licitación para la construcción de una nueva escuela pública. La empresa constructora A está considerando presentarse. En los últimos años su principal competidor B se ha presentado en el 70% de las licitaciones de edificios escolares. Si B no se presenta en la licitación, la probabilidad de que A obtenga el trabajo es de 0,5 y si B se presenta a la licitación, la probabilidad de que A gane la licitación es 0,25.

- ¿Cuál es la probabilidad de que A obtenga el trabajo?
- Si A no obtiene el trabajo, ¿cuál es la probabilidad de que B se haya presentado a la licitación?

Resolución:

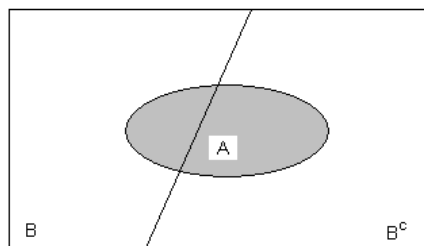
Sean los sucesos:

B : La empresa B se presenta a la licitación.

A : La empresa A consigue el trabajo.

Se sabe que $P(B) = 0,70$, luego $P(B^c) = 1 - P(B) = 0,3$. También se tiene que $P(A|B^c) = 0,5$ y $P(A|B) = 0,25$.

Se puede considerar que B y B^c determinan una partición del espacio muestral, y consecuentemente la situación presentada en este caso se puede visualizar gráficamente de la siguiente manera:



a) Por Teorema de Probabilidad Total:

$$\begin{aligned} P(A) &= P(A|B)P(B) \\ &\quad + P(A|B^c)P(B^c) \\ &= 0,25 \times 0,7 + 0,5 \times 0,3 = 0,325 \end{aligned}$$

b) $P(A^c|B) = 1 - P(A|B) = 1 - 0,25 = 0,75$

$$P(A^c) = 1 - P(A) = 1 - 0,325 = 0,675$$

$$P(B|A^c) = \frac{P(B) \times P(A^c|B)}{P(A^c)} = \frac{0,7 \times 0,75}{0,675} = 0,78 \text{ por Teorema de Bayes.}$$

Ejemplo 8: El 70% del personal del departamento de contabilidad de una empresa son graduados y 30% son pasantes. El 30% de los empleados elabora pólizas de nóminas. Si el hecho de ser o no graduado y el elaborar pólizas son independientes, calcule la probabilidad de que un empleado cualquiera sea pasante y elabore pólizas.

Resolución:

Sean los sucesos:

G : el trabajador es graduado $P(G) = 0,7$

P : el trabajador es pasante $P(P) = 0,3$

E : el trabajador elabora pólizas $P(E) = 0,3$

$$P(P \cap E) = P(P) \times P(E) = 0,3 \times 0,3 = 0,09 \text{ por ser independientes los sucesos } P \text{ y } E.$$

CAPÍTULO 3: VARIABLES ALEATORIAS

Variables aleatorias discretas y continuas. Función de masa y función de densidad de probabilidad. Función de distribución. Distribuciones marginales, conjuntas y condicionales. Variables aleatorias independientes.

Objetivos:

El alumno debe ser capaz de:

- Definir variable aleatoria.
- Definir variables aleatorias asociadas a un experimento aleatorio dado.
- Distinguir cuando una variable aleatoria es discreta o continua.
- Definir función de masa, función de densidad de probabilidad y función de distribución acumulada para variables aleatorias unidimensionales.
- Calcular valores de probabilidad para variables aleatorias unidimensionales utilizando la información de la función de masa, la función de densidad de probabilidad y la función de distribución acumulada.
- Definir función de masa, función de densidad de probabilidad y función de distribución acumulada para variables aleatorias bidimensionales.
- Calcular valores de probabilidad para variables aleatorias bidimensionales utilizando la información de la función de masa o la función de densidad de probabilidad.

Resumen

D1. Variable Aleatoria: Sea un espacio muestral \mathcal{S} de un experimento aleatorio. Una variable aleatoria X es una función cuyo dominio es \mathcal{S} y cuyo recorrido es un conjunto de números reales, es decir:

$$X: \mathcal{S} \rightarrow \mathbb{R}$$

$$s \mapsto X(s)$$

Obs: Las probabilidades estaban definidas en subconjuntos de \mathcal{S} . Una variable aleatoria induce ahora probabilidades en \mathbb{R} de manera tal que $P(A) = P(\{s \in \mathcal{S}: X(s) \in A\})$, donde A es un subconjunto de \mathbb{R} .

D2. Función de Distribución Acumulada (fda): Sea X una variable aleatoria. La función de distribución acumulada F_X es la función definida para todo número real x por $F_X(x) = P(X \leq x)$, es decir

$$F_X: \mathbb{R} \rightarrow \mathbb{R}$$

$$x \mapsto P(\{s \in \mathcal{S}: X(s) \leq x\})$$

Obs: Las variables aleatorias se denotan con letras de molde mayúsculas (X), y las letras de molde minúsculas (x) denotan valores de su recorrido.

Propiedades de la función de distribución acumulada:

- (1) F_X es una función no decreciente
- (2) $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- (3) $\lim_{x \rightarrow \infty} F_X(x) = 1$
- (4) $P(a < X \leq b) = F_X(b) - F_X(a)$, cuando $b \geq a$

D3. Variable Aleatoria Discreta: Se dice que una variable aleatoria X es discreta cuando el conjunto de valores posibles ($\text{Rec}(X)$) es numerable (finito o infinito semejante a los números naturales).

D4. Distribución de Probabilidad (para una variable aleatoria discreta): Sea X una variable aleatoria discreta con $\text{Rec}(X) = \{x_i: i \in I \subset \mathbb{N}\}$, entonces la función de probabilidad o función de masa se denota por p_X y es tal que

$$p_X: \mathbb{R} \rightarrow \mathbb{R}$$

$$x \mapsto p_X(x) = \begin{cases} P(X = x) & \text{si } x \in \text{Rec}(X) \\ 0 & \text{si } x \notin \text{Rec}(X). \end{cases}$$

Propiedades de la función de masa:

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

$$(2) \sum_{i \in I} p_X(x_i) = 1$$

D5. Variable Aleatoria Continua: Se dice que una variable aleatoria X es continua cuando el conjunto de valores posibles ($\text{Rec}(X)$) es no-numerable (semejantes a un intervalo en \mathbb{R}).

D6. Distribución de Probabilidad (para una variable aleatoria continua): Sea X una variable aleatoria continua. Si existe f_X tal que:

$$(1) f_X(x) \geq 0, \text{ para todo } x \text{ real,}$$

$$(2) \int_{-\infty}^{\infty} f_X(x) dx = 1 \text{ y}$$

$$(3) P(a < X < b) = \int_a^b f_X(x) dx,$$

entonces se dice que f_X es la función de probabilidad o función de densidad de X .

D7. Distribución de Probabilidad Conjunta (para variables aleatorias discretas): Sean X e Y variables aleatorias discretas, la función $f(x, y)$ es una función de distribución de probabilidad conjunta o función de masa de probabilidad de las variables aleatorias X e Y , si:

$$(1) f(x, y) \geq 0 \text{ para todo } (x, y),$$

$$(2) \sum_x \sum_y f(x, y) = 1,$$

$$(3) P(X = x, Y = y) = f(x, y).$$

Obs: Para cualquier región A del plano, se tiene que

$$P((X, Y) \in A) = \sum_{(x, y) \in A} f(x, y)$$

D8. Distribución de Probabilidad Conjunta (para variables aleatorias continuas): Sean X e Y variables aleatorias continuas, la función $f(x, y)$ es una función de distribución de probabilidad conjunta o función de densidad conjunta de las variables aleatorias X e Y , si:

$$(1) f(x, y) \geq 0 \text{ para todo } (x, y),$$

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

$$(3) P((X, Y) \in A) = \iint_A f(x, y) dx dy \text{ para cualquier región } A \text{ del plano.}$$

D9. Distribuciones Marginales: Sean X e Y variables aleatorias con función de distribución de probabilidad conjunta $f(x, y)$. Las distribuciones marginales de X e Y están dadas respectivamente por:

$$g(x) = \sum_y f(x, y) \text{ y } h(y) = \sum_x f(x, y) \text{ para el caso discreto, y}$$

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \text{ ... y } h(y) = \int_{-\infty}^{\infty} f(x, y) dx \text{ para el caso continuo.}$$

D10. Distribuciones Condicionales: Sean X e Y variables aleatorias con función de distribución de probabilidad conjunta $f(x, y)$. La distribución condicional de Y dado $X = x$ está dada por:

$$f(y|x) = \frac{f(x, y)}{g(x)}, g(x) > 0.$$

Similarmente, la distribución condicional de X dado $Y = y$ está dada por:

$$f(x|y) = \frac{f(x, y)}{h(y)}, h(y) > 0.$$

Donde $g(x)$ y $h(y)$ son las distribuciones marginales de X e Y respectivamente.

D11. Independencia Estadística: Sean X e Y variables aleatorias con función de distribución de probabilidad $f(x, y)$ y distribuciones marginales $g(x)$ y $h(y)$, respectivamente. Las variables X e Y se dicen estadísticamente independientes si y solo si

$$f(x, y) = g(x)h(y) \text{ para todo par } (x, y).$$

Ejemplos

Ejemplo 1: Sea X una variable aleatoria discreta con función de probabilidad dada por la siguiente tabla:

x	-5	-2	0	1	3	8
$p(x)$	0,1	0,2	0,1	0,2	a	0,1

- Determinar a .
- Encontrar la función de distribución acumulada de X .
- Usando la función de masa y la función de distribución acumulada, calcular las siguientes probabilidades:
 - $P(X \leq 1)$
 - $P(X > -2)$
 - $P(-2 < X \leq 3)$

Resolución:

- Para que p sea una función de probabilidad, debe satisfacer las propiedades de este tipo de funciones. Una de ellas es que las imágenes por p de todos los valores posibles de x sumen 1.

$$\sum_{i \in I} p(x_i) = 1$$

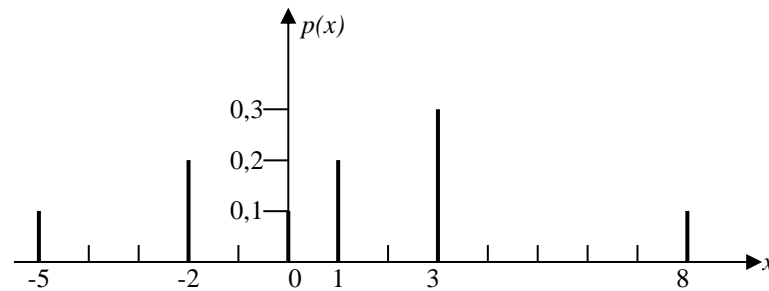
Consecuentemente,

$$0,1 + 0,2 + 0,1 + 0,2 + a + 0,1 = 1$$

$$0,7 + a = 1$$

$$a = 0,3.$$

La representación gráfica de $p(x)$ es la siguiente:



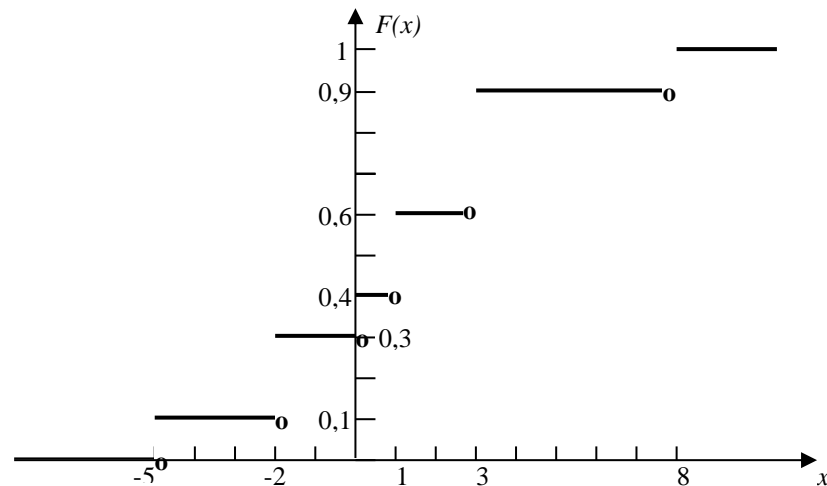
b) La función de distribución acumulada está definida por $F(x) = P(X \leq x)$. Es decir, F asigna a cada valor de x , la probabilidad que se ha acumulado desde $-\infty$ hasta el valor x inclusive.

- Si se toma x , tal que $x < -5$, desde $-\infty$ hasta el valor de x ningún punto tiene probabilidad positiva. Por lo tanto, $F(x) = 0$.
- Si se considera x , tal que $-5 \leq x < -2$, desde $-\infty$ hasta el valor de x el único punto que tiene probabilidad no nula es -5 . Por lo tanto, la única probabilidad acumulada hasta ese x es $P(X = -5)$. Entonces, $F(x) = 0,1$.
- Si se selecciona x , tal que $-2 \leq x < 0$, desde $-\infty$ hasta el valor de x los únicos puntos que tienen probabilidad no nula son -5 y -2 . En consecuencia, la probabilidad acumulada hasta ese x es $P(X = -5) + P(X = -2)$. Entonces, $F(x) = 0,3$.
- Si se selecciona x , tal que $0 \leq x < 1$, desde $-\infty$ hasta el valor de x , los únicos puntos que tienen probabilidad no nula son -5 , -2 y 0 . Por lo tanto, la probabilidad acumulada hasta ese x es $P(X = -5) + P(X = -2) + P(X = 0)$. Entonces, $F(x) = 0,4$.
- Si se selecciona x , tal que $1 \leq x < 3$, desde $-\infty$ hasta el valor de x los únicos puntos que tienen probabilidad no nula son -5 , -2 , 0 y 1 . En consecuencia, la probabilidad acumulada hasta ese x es $P(X = -5) + P(X = -2) + P(X = 0) + P(X = 1)$. Entonces, $F(x) = 0,6$.
- Si se selecciona x , tal que $3 \leq x < 8$, desde $-\infty$ hasta el valor de x los únicos puntos que tienen probabilidad no nula son -5 , -2 , 0 , 1 y 3 . Por lo tanto, la probabilidad acumulada hasta ese x es $P(X = -5) + P(X = -2) + P(X = 0) + P(X = 1) + P(X = 3)$. Entonces, $F(x) = 0,9$.
- Si se selecciona x , tal que $x \geq 8$, desde $-\infty$ hasta el valor de x los puntos que tienen probabilidad no nula son -5 , -2 , 0 , 1 , 3 y 8 . Por lo tanto, la probabilidad acumulada hasta ese x es $P(X = -5) + P(X = -2) + P(X = 0) + P(X = 1) + P(X = 3) + P(X = 8)$. Entonces, $F(x) = 1$.

Consecuentemente, la función de distribución acumulada viene dada por:

$$F(x) = \begin{cases} 0 & \text{si } x < -5 \\ 0,1 & \text{si } -5 \leq x < -2 \\ 0,3 & \text{si } -2 \leq x < 0 \\ 0,4 & \text{si } 0 \leq x < 1 \\ 0,6 & \text{si } 1 \leq x < 3 \\ 0,9 & \text{si } 3 \leq x < 8 \\ 1 & \text{si } x \geq 8 \end{cases}$$

La representación gráfica de la función F es la siguiente:



c) *Cálculo de probabilidades usando la función de masa:*

Para calcular probabilidades de eventos con un número finito de elementos, se suman las probabilidades sobre los valores de x que cumplen la condición dada por el evento.

- i)
$$P(X \leq 1) = P(X = -5) + P(X = -2) + P(X = 0) + P(X = 1)$$
$$= 0,1 + 0,2 + 0,1 + 0,2 = 0,6$$
- ii)
$$P(X > -2) = P(X = 0) + P(X = 1) + P(X = 3) + P(X = 8)$$
$$= 0,1 + 0,2 + 0,3 + 0,1 = 0,7$$
- iii)
$$P(-2 < X \leq 3) = P(X = 0) + P(X = 1) + P(X = 3)$$
$$= 0,1 + 0,2 + 0,3 = 0,6$$

Cálculo de probabilidades usando la función de distribución acumulada:

Para calcular probabilidades de un evento a partir de la función de distribución acumulada se intenta escribir el evento como la intersección, unión o complemento de intervalos del tipo $(-\infty, b]$ ó $(a, b]$, y luego se usa la definición de F , o bien la propiedad $P(a < X \leq b) = F(b) - F(a)$ para el cálculo de las respectivas probabilidades.

- i) $P(X \leq 1) = F(1) = 0,6$ valor que se obtiene directamente de la expresión de $F(x)$ al mirar cuál es la imagen de 1.
- ii) $P(X > -2) = 1 - P(X \leq -2) = 1 - F(-2) = 1 - 0,3 = 0,7$.
- iii) $P(-2 < X \leq 3) = F(3) - F(-2) = 0,9 - 0,3 = 0,6$.

Ejemplo 2: Una compañía tiene cinco solicitantes para dos puestos de trabajo: dos mujeres y tres hombres. Suponga que los cinco candidatos están igualmente calificados y que no existe

preferencia por ningún género al escoger. Se define la variable aleatoria X que cuenta el número de mujeres elegidas para cubrir los dos puestos.

- Determinar p_X .
- Representar gráficamente la función de masa de X .

Resolución:

- El experimento aleatorio es

ε : se eligen al azar dos solicitantes para ocupar los dos puestos de trabajo.

Si se denominan M_1 y M_2 a las dos mujeres solicitantes y H_1, H_2 y H_3 a los tres hombres solicitantes, y si se considera que no hay diferencia entre los dos puestos de trabajo, el espacio muestral del experimento sería

$$\mathcal{S} = \{\{M_1, M_2\}, \{M_1, H_1\}, \{M_1, H_2\}, \{M_1, H_3\}, \{M_2, H_1\}, \{M_2, H_2\}, \{M_2, H_3\}, \{H_1, H_2\}, \{H_1, H_3\}, \{H_2, H_3\}\}.$$

Al definir la variable aleatoria

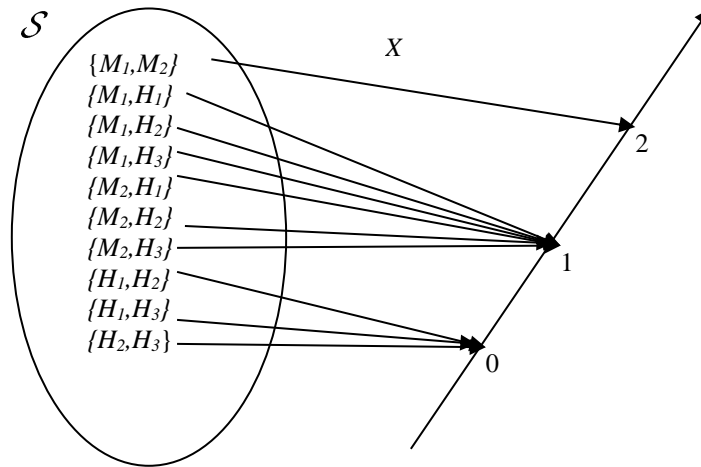
$$X: \mathcal{S} \rightarrow \mathbb{R} \\ s \mapsto \text{n}^\circ \text{de mujeres elegidas para cubrir los dos puestos}$$

el recorrido de X resulta

$$\text{Rec}(X) = \{0, 1, 2\},$$

por lo tanto una variable discreta, ya que su recorrido es un conjunto finito.

La representación gráfica de la variable aleatoria X , como función de \mathcal{S} en $[0,1]$, es la siguiente:



La función de masa de la variable X es una función

$$p_X: \mathbb{R} \rightarrow \mathbb{R} \\ x \mapsto p_X(x) = \begin{cases} P(X = x) & \text{si } x \in \text{Rec}(X) \\ 0 & \text{si } x \notin \text{Rec}(X) \end{cases}$$

Para determinar $p_X(x) = P(X = x)$, es necesario identificar el suceso correspondiente a $\{X = x\}$ en el espacio muestral; es decir,

$$\{s \in \mathcal{S}: X(s) = x\}, \text{ en cada } x \text{ del recorrido.}$$

- El suceso correspondiente a $\{X = 2\}$ en el espacio muestral es $\{M_1, M_2\}$.

En este caso, todos los puntos muestrales en \mathcal{S} son equiprobables, por lo tanto,

$$p_X(2) = P(X = 2) = P(\{M_1, M_2\}) = \frac{1}{10} = 0,1.$$

- El suceso correspondiente a $\{X = 1\}$ en el espacio muestral es $B = \{\{M_1, H_1\}, \{M_1, H_2\}, \{M_1, H_3\}, \{M_2, H_1\}, \{M_2, H_2\}, \{M_2, H_3\}\}$. Por lo tanto,

$$p_X(1) = P(X = 1) = P(B) = \frac{6}{10} = 0,6.$$

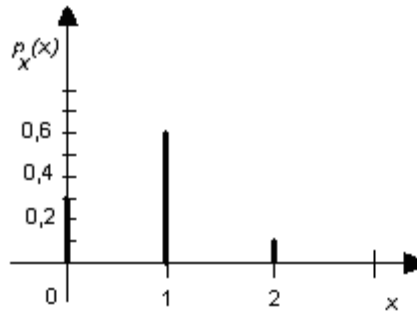
- El suceso relacionado a $\{X = 0\}$ en el espacio muestral es $C = \{\{H_1, H_2\}, \{H_1, H_3\}, \{H_2, H_3\}\}$. Por lo tanto,

$$p_X(0) = P(X = 0) = P(C) = \frac{3}{10} = 0,3.$$

Entonces, la función de masa está definida por la siguiente tabla.

x	0	1	2
$p_X(x)$	0,3	0,6	0,1

b) La representación gráfica de la función de masa de X es la siguiente:



Observación: Si en el enunciado de este ejercicio el número de mujeres y varones hubiera sido mayor, la enumeración de los elementos de \mathcal{S} discriminando los posibles pares de mujeres y varones sería más dificultosa e innecesaria. Sin embargo se podría razonar de la siguiente manera.

$$\mathcal{S} = \{\{V, V\}; \{V, M\}; \{M, M\}\}$$

donde V denota un varón y M una mujer. En este espacio muestral no se distingue entre las mujeres o entre los varones, consecuentemente, dependiendo del número de mujeres y del número de varones, los puntos muestrales tendrán probabilidades diferentes.

Si a denota el número de mujeres y b el número de varones, las probabilidades se calculan usando la teoría de conteo, para ello es conveniente observar que en este caso el orden no es importante.

- El número de grupos o combinaciones de dos varones que se pueden formar entre b varones es $\binom{b}{2}$, y el número de grupos de dos personas que se pueden formar entre las $a+b$ personas es $\binom{a+b}{2}$. Por lo tanto

$$P(\{V, V\}) = \frac{\binom{b}{2}}{\binom{a+b}{2}}.$$

- El número de grupos o combinaciones de un varón y una mujer que se pueden formar es $\binom{b}{1} \times \binom{a}{1}$, consecuentemente

$$P(\{V, M\}) = \frac{\binom{b}{1} \binom{a}{1}}{\binom{a+b}{2}}.$$

- El número de grupos o combinaciones de dos mujeres que se pueden formar es $\binom{a}{2}$, consecuentemente

$$P(\{M, M\}) = \frac{\binom{a}{2}}{\binom{a+b}{2}}.$$

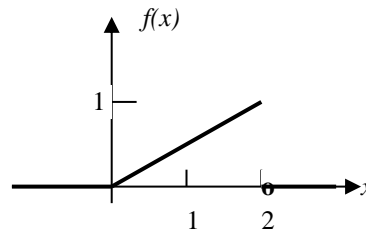
Ejemplo 3: Sea X una variable aleatoria continua con función de densidad de probabilidad

$$\text{dada por: } f(x) = \begin{cases} \frac{x}{2} & \text{si } 0 \leq x \leq 2 \\ 0 & \text{en otro caso} \end{cases}$$

- Determinar la función de distribución acumulada.
- Determinar, usando la función de densidad y la distribución acumulada, la probabilidad de que X sea mayor que 1 pero menor de $3/2$.

Resolución:

El gráfico de la función de densidad f es el siguiente:



- La definición de la función de distribución acumulada es

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt,$$

que representa la probabilidad acumulada desde $-\infty$ hasta el valor x seleccionado.

- Si $x < 0$, $F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt = 0$.
- Si $0 \leq x \leq 2$,

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_{-\infty}^x f(t) dt = \int_{-\infty}^0 0 dt + \int_0^x \frac{t}{2} dt = \frac{1}{2} \int_0^x t dt = \frac{1}{2} \frac{t^2}{2} \Big|_0^x = \frac{1}{4} (x^2 - 0^2) = \frac{1}{4} x^2. \end{aligned}$$

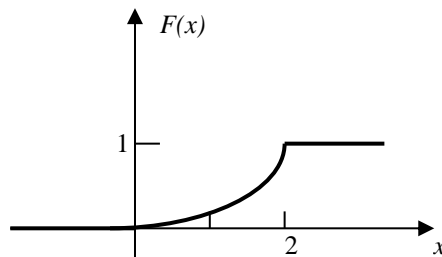
- Si $x > 2$,

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_{-\infty}^x f(t) dt = \int_{-\infty}^0 0 dt + \int_0^2 \frac{t}{2} dt + \int_2^x 0 dt = \frac{1}{2} \int_0^2 t dt = \frac{1}{2} \frac{t^2}{2} \Big|_0^2 = \frac{1}{4} (2^2 - 0^2) = 1. \end{aligned}$$

Consecuentemente, la función de distribución acumulada viene dada por:

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{4} x^2 & \text{si } 0 \leq x \leq 2, \\ 1 & \text{si } x > 2 \end{cases}$$

y su representación gráfica es la siguiente:

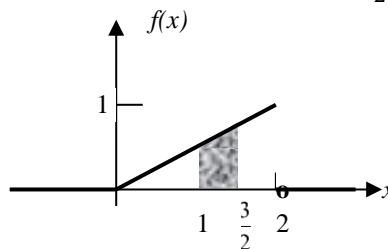


b) *Cálculo de probabilidades usando la función de densidad:*

Por definición de la función de densidad, la probabilidad de que una variable aleatoria asuma valores en un intervalo se calcula integrando su función de densidad entre los extremos del intervalo. De esta manera

$$P\left(1 < X < \frac{3}{2}\right) = \int_1^{\frac{3}{2}} f(x) dx = \int_1^{\frac{3}{2}} \frac{x}{2} dx = \frac{1}{2} \int_1^{\frac{3}{2}} x dx = \frac{1}{2} \frac{x^2}{2} \Big|_1^{\frac{3}{2}} = \frac{1}{4} \left(\left(\frac{3}{2}\right)^2 - 1^2 \right) = \frac{5}{16}.$$

Esta probabilidad se puede interpretar como el área comprendida entre la función de densidad, el eje x , y las rectas verticales $x = 1$ y $x = \frac{3}{2}$.



Cálculo de probabilidades usando la función de distribución acumulada:

Consiste en usar la definición de F , o la propiedad $P(a < X \leq b) = F(b) - F(a)$. Por tratarse de una **variable aleatoria continua** no tiene importancia que la probabilidad que se quiere calcular sea la de un intervalo semiabierto, cerrado o abierto, puesto que estos intervalos, como conjuntos, difieren ricamente en uno o dos puntos y la probabilidad en un punto es nula en este tipo de variables. Consecuentemente

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b).$$

De esta manera

$$P(1 < X < 3/2) = P(1 < X \leq 3/2) = F(3/2) - F(1) = \frac{1}{4} \left(\frac{3}{2}\right)^2 - \frac{1}{4} 1^2 = \frac{5}{16}.$$

Ejemplo 4: Se lanzan 3 monedas equilibradas y se observa si cada una de ellas cae en cara o en sello. Sea X la cantidad de caras obtenidas y sea Y la cantidad de caras menos la cantidad de sellos obtenidos. Encontrar (a) la distribución conjunta de la variable aleatoria bidimensional (X,Y) , (b) la distribución marginal de X y (c) la distribución marginal de Y .

Resolución:

El experimento aleatorio es

ε : se lanzan tres monedas equilibradas.

El espacio muestral es

$$\mathcal{S} = \{(c,c,c), (c,c,s), (c,s,c), (s,c,c), (c,s,s), (s,c,s), (s,s,c), (s,s,s)\},$$

donde c denota una cara y s un sello.

El recorrido de la variable X : *Cantidad de caras obtenidas* es $\text{Rec}(X) = \{0,1,2,3\}$, y por lo tanto X es discreta porque tiene un recorrido finito.

El recorrido de la variable Y : *Diferencia entre la cantidad de caras y de sellos* es $\text{Rec}(Y) = \{-3,-1,1,3\}$, y por lo tanto Y es discreta porque tiene un recorrido finito.

En la siguiente tabla se muestra la correspondencia entre los elementos del espacio muestral y sus respectivos valores de X , Y .

δ	$X(s)$	$Y(s)$
(c,c,c)	3	3
(c,c,s),(c,s,c),(s,c,c)	2	1
(c,s,s),(s,c,s),(s,s,c)	1	-1
(s,s,s)	0	-3

Los elementos en δ son equiprobables, con una probabilidad $1/8$ de ocurrir cualquiera de ellos al realizar el experimento. La tabla anterior permite determinar los valores del recorrido de (X,Y) y la probabilidad de ocurrencia de cada uno de esos pares ordenados. Por ejemplo: $f(0,-3) = P(X = 0, Y = -3) = P(\{(s,s,s)\}) = 1/8$, por otro lado $f(2,1) = P(X = 2, Y = 1) = P(\{(c,c,s), (c,s,c), (s,c,c)\}) = 3/8$.

Razonando de esta manera es posible construir la función de masa conjunta de (X,Y) , de la siguiente manera:

$f(x,y)$		y				$g(x)$
		-3	-1	1	3	
x	0	1/8	0	0	0	1/8
	1	0	3/8	0	0	3/8
	2	0	0	3/8	0	3/8
	3	0	0	0	1/8	1/8
$h(y)$		1/8	3/8	3/8	1/8	1

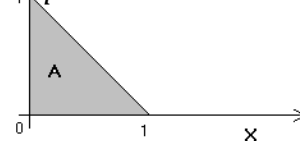
La función de distribución de probabilidad marginal de X , $g(x)$, se obtiene en cada x , sumando las probabilidades sobre cada fila, es decir $g(x) = \sum_y f(x,y)$.

La función de distribución de probabilidad marginal de Y , $h(y)$, se obtiene en cada y , sumando las probabilidades sobre cada columna, es decir $h(y) = \sum_x f(x,y)$.

Ejemplo 5: Sea la variable aleatoria bidimensional (X, Y) distribuida uniformemente en la región sombreada A, indicada en la figura. Por lo tanto

$$f(x,y) = \begin{cases} \frac{1}{\text{área}(A)}, & \text{si } (x,y) \in A \\ 0 & \text{en cualquier otro punto} \end{cases}$$

Encontrar las funciones de densidad marginales de X e Y .



Resolución:

La superficie sombreada es la superficie de un triángulo: $\frac{b \times h}{2} = \frac{1}{2}$.

Para poder escribir los límites de la región, es necesario determinar la ecuación de la recta que pasa por los puntos $(1,0)$ y $(0,1)$. La ecuación de esta recta es $y = 1-x$. Como los puntos de la región se encuentran sobre o por debajo de esta recta, cualquier punto de la región A verifica la inecuación $y \leq 1-x$.

Entonces, la función de densidad conjunta es

$$f(x,y) = \begin{cases} 2 & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1-x \\ 0 & \text{en otro caso} \end{cases}$$

La función de densidad marginal de X se obtiene integrando sobre y la función de densidad conjunta.

$$g(x) = \int_{-\infty}^{\infty} f(x,y) dy = \begin{cases} \int_0^{1-x} 2 dy = 2y \Big|_0^{1-x} = 2(1-x) & \text{si } 0 \leq x \leq 1 \\ \int_{-\infty}^{\infty} 0 dy & \text{en cualquier otro } x \end{cases}$$

Es preciso dividir el cálculo de la marginal, debido a la definición por ramas de la función de densidad conjunta de (X,Y) .

Se puede observar que la función de densidad conjunta también puede ser escrita de la siguiente manera

$$f(x,y) = \begin{cases} 2 & \text{si } 0 \leq y \leq 1, 0 \leq x \leq 1-y \\ 0 & \text{en otro caso} \end{cases}$$

consecuentemente, la función de densidad marginal de Y se obtiene integrando sobre x la función de densidad conjunta.

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \begin{cases} \int_0^{1-y} 2dx = 2x \Big|_0^{1-y} = 2(1-y) & \text{si } 0 \leq y \leq 1 \\ \int_{-\infty}^{\infty} 0dx & \text{en cualquier otro } y \end{cases}$$

CAPÍTULO 4: ESPERANZA MATEMÁTICA

Esperanza matemática. Media, varianza y covarianza de variables aleatorias. Media, varianza y covarianza de combinaciones lineales de variables aleatorias. Teorema de Chebyshev.

Objetivos:

El alumno debe ser capaz de:

- Calcular la esperanza y la varianza de una variable aleatoria.
- Dado un par de variables aleatorias calcular la covarianza entre las mismas.
- Calcular la esperanza y la varianza de funciones de variables aleatorias cuya función de probabilidad sea conocida.
- Calcular cotas de probabilidad aplicando el teorema de Chebyshev.

Resumen

D1. Esperanza Matemática: Sea X una variable aleatoria con distribución de probabilidad $f_X(x)$, la media (valor esperado o esperanza matemática) de X se denota con μ y es

$$\mu = E(X) = \sum_x x f_X(x), \text{ cuando } X \text{ es una variable discreta, ó}$$

$$\mu = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx, \text{ cuando } X \text{ es una variable continua.}$$

T1. Sea X v.a. con distribución de probabilidad $f_X(x)$. El valor esperado de la variable aleatoria $g(X)$ está dada por

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x) f_X(x), \text{ si } X \text{ es discreta, y}$$

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx, \text{ si } X \text{ es continua.}$$

D2. Sean X e Y v.s.as. con distribución de probabilidad conjunta $f(x, y)$. La media de la v.a. $g(X, Y)$ es

$$\mu_{g(X, Y)} = E[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y), \text{ si } X \text{ e } Y \text{ son discretas, y}$$

$$\mu_{g(X, Y)} = E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy, \text{ si } X \text{ e } Y \text{ son continuas.}$$

D3. Varianza: Sea X una v.a. con media μ . La varianza de X , $V(X)$, es

$$\sigma_X^2 = E[(X - \mu)^2].$$

T2. Sea X v.a. con distribución de probabilidad $f_X(x)$ y media μ . La varianza de X , $V(X)$, está dada por

$$\sigma_X^2 = \sum_x (x - \mu)^2 f_X(x), \text{ si } X \text{ es discreta, y}$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx, \text{ si } X \text{ es continua.}$$

D4. Desviación Estándar: La desviación estándar se denota por σ_X y es la raíz cuadrada positiva de σ_X^2 .

T3. Sea X v.a. con media μ , entonces $V(X) = E(X^2) - \mu^2$.

T4. Sea X v.a. con distribución de probabilidad $f_X(x)$. La varianza de la variable aleatoria $g(X)$ está dada por

$$V[g(X)] = E\left[\{g(X) - \mu_{g(X)}\}^2\right] = \sum_x \{g(x) - \mu_{g(X)}\}^2 f_X(x), \text{ si } X \text{ es discreta, y}$$

$$V[g(X)] = E\left[\{g(X) - \mu_{g(X)}\}^2\right] = \int_{-\infty}^{\infty} \{g(x) - \mu_{g(X)}\}^2 f_X(x) dx, \text{ si } X \text{ es continua.}$$

D5. Covarianza: Sean X e Y vs.as. con distribución de probabilidad conjunta $f(x, y)$ y medias μ_X y μ_Y , respectivamente. La covarianza de X e Y , $Cov(X, Y)$ ó σ_{XY} , es

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)].$$

T5. $\sigma_{XY} = E(XY) - \mu_X \mu_Y$.

Propiedades de las medias, varianzas y covarianzas de combinaciones lineales de vs.as.:

Sean X e Y vs.as. y sean a, b, c y d constantes

(1) $E(aX + b) = a E(X) + b$

(2) $E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)]$

(3) $E[g(X, Y) \pm h(X, Y)] = E[g(X, Y)] \pm E[h(X, Y)]$

(4) $E[g(X) \pm h(Y)] = E[g(X)] \pm E[h(Y)]$

(5) Si X e Y son independientes entonces $E(X \cdot Y) = E(X) \cdot E(Y)$

(6) $\sigma_{aX+b}^2 = a^2 \sigma_X^2$

(7) $\sigma_{aX+bY}^2 = a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab\sigma_{XY}$

(8) Si X e Y son independientes entonces $\sigma_{aX+bY}^2 = a^2 \sigma_X^2 + b^2 \sigma_Y^2$

(9) $Cov(aX + b, cY + d) = ac Cov(X, Y)$

(10) $Cov(\sum_{i=1}^A a_i X_i, \sum_{j=1}^B b_j Y_j) = \sum_{i=1}^A \sum_{j=1}^B a_i b_j Cov(X_i, Y_j)$, donde a_1, \dots, a_A y b_1, \dots, b_B son

constantes y X_1, \dots, X_A e Y_1, \dots, Y_B son vs.as.

T6. Teorema de Chebyshev: La probabilidad de que cualquier variable aleatoria X asuma un valor dentro de k desviaciones estándares de la media es al menos $1 - 1/k^2$, es decir

$$P(\mu - k\sigma_X < X < \mu + k\sigma_X) \geq 1 - 1/k^2.$$

Ejemplos

Ejemplo 1: En un concurso de un programa televisivo, un concursante lanza un dado y el anfitrión le paga tantos billetes de \$100 como puntos señale la cara que cae hacia arriba, excepto cuando sale 5 o 6, en cuyo caso es el concursante quien debe pagar al anfitrión tantos billetes de \$100 como puntos muestre la cara superior. ¿Quién de los dos tiene ventaja en el juego, el concursante o el anfitrión?

Resolución:

El experimento aleatorio es

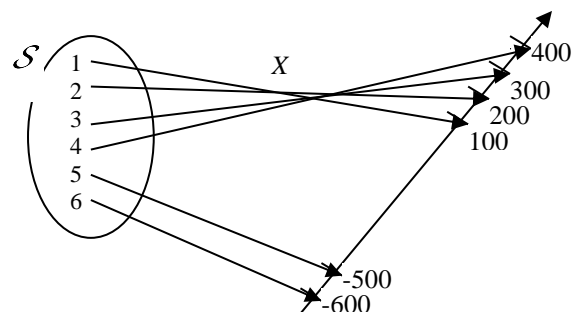
ε : se lanza un dado y se observa el resultado.

El espacio muestral es $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.

La variable aleatoria X que representa la ganancia del concursante, es una función que a cada elemento de \mathcal{S} le asigna un número real de la siguiente forma:

$$X: \mathcal{S} \rightarrow \mathbb{R}$$

$$s \rightarrow X(s) = \begin{cases} 100s & \text{si } s < 5 \\ -100s & \text{si } s \geq 5 \end{cases}$$



Para calcular $p(x) = P(X = x)$, se debe determinar el suceso equivalente en el espacio muestral. Por ejemplo, $p(100) = P(X = 100) = P(s = 1) = \frac{1}{6}$.

x	100	200	300	400	-500	-600
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Para determinar quién tiene ventaja en el juego, se calcula el valor esperado, que indica lo que el concursante gana o pierde *en promedio* cada vez que lanza el dado, o lo que se esperaría que sea la ganancia promedio si se jugara un número grande de veces.

$$\begin{aligned} E(X) &= \sum_{x=1}^6 x \cdot p(x) = \sum_{x=1}^6 x \cdot \frac{1}{6} = \frac{1}{6} \sum_{x=1}^6 x \\ &= \frac{1}{6} (100 + 200 + 300 + 400 - 500 - 600) = -100 \times \frac{1}{6} = -16,67 \end{aligned}$$

La interpretación de este valor esperado es que si el concursante jugara un número infinitamente grande de veces este juego, *en promedio* perdería \$16,67 en cada lanzamiento del dado.

Por lo tanto, el juego es favorable al anfitrión.

Ejemplo 2: El consumo de combustible de un cierto automóvil es una variable aleatoria medida en Km por litro. Se supone que la densidad de probabilidad de esta variable aleatoria es la siguiente

$$f(x) = \begin{cases} x - 10, & \text{si } 10 \leq x \leq 11 \\ 12 - x, & \text{si } 11 < x \leq 12 \\ 0, & \text{en cualquier otro punto.} \end{cases}$$

- Determinar la esperanza y varianza del consumo.
- Si el precio del combustible es \$4 por litro ¿Cuál es la esperanza del costo de un viaje de 100 Km en este automóvil?

Resolución:

a) En el caso continuo $E(x) = \int_{-\infty}^{\infty} xf(x)dx$.

$$\begin{aligned} E(X) &= \int_{-\infty}^{10} x \cdot 0 dx + \int_{10}^{11} x(x - 10) dx + \int_{11}^{12} x(12 - x) dx + \int_{12}^{\infty} x \cdot 0 dx \\ &= \int_{10}^{11} (x^2 - 10x) dx + \int_{11}^{12} (12x - x^2) dx = \left(\frac{x^3}{3} - 5x^2 \right) \Big|_{10}^{11} + \left(6x^2 - \frac{x^3}{3} \right) \Big|_{11}^{12} \\ E(X) &= \left(\frac{11^3}{3} - 5 \cdot 11^2 \right) - \left(\frac{10^3}{3} - 5 \cdot 10^2 \right) + \left(6 \cdot 12^2 - \frac{12^3}{3} \right) - \left(6 \cdot 11^2 - \frac{11^3}{3} \right) = 11 \end{aligned}$$

Para el cálculo de la varianza, se puede usar la expresión

$$V(X) = E(X^2) - [E(X)]^2.$$

$$\begin{aligned} E(X^2) &= \int_{10}^{11} x^2(x - 10) dx + \int_{11}^{12} x^2(12 - x) dx = \int_{10}^{11} (x^3 - 10x^2) dx + \int_{11}^{12} (12x^2 - x^3) dx \\ &= \left(\frac{x^4}{4} - 10 \frac{x^3}{3} \right) \Big|_{10}^{11} + \left(4x^3 - \frac{x^4}{4} \right) \Big|_{11}^{12} \end{aligned}$$

$$= \left(\frac{11^4}{4} - 10 \frac{11^3}{3} \right) - \left(\frac{10^4}{4} - 10 \frac{10^3}{3} \right) + \left(4 \cdot 12^3 - \frac{12^4}{4} \right) - \left(4 \cdot 11^3 - \frac{11^4}{4} \right)$$

$$= \frac{727}{6}$$

Por lo tanto,

$$V(X) = E(X^2) - [E(X)]^2 = \frac{727}{6} - 11^2 = \frac{727}{6} - 121 = \frac{1}{6}.$$

- b) X es la cantidad de kilómetros recorridos con un litro de combustible. Sea Z la cantidad necesaria de combustible para recorrer 100 km.

$$\begin{array}{lcl} x \text{ km} & \text{---} & 1 \text{ litro} \\ 100 \text{ km} & \text{---} & z \text{ litros} \end{array}$$

La relación entre Z y X viene dada por la expresión

$$Z = \frac{100}{x} \text{ litros.}$$

La densidad de X informa los valores de su recorrido, $\text{Rec}(X) = [10, 12]$, puesto que la probabilidad de que asuma valores fuera de este intervalo es nula. Para determinar los valores entre los cuales varía Z , se debe tener en cuenta los valores del recorrido de X , y así se puede determinar que $\text{Rec}(Z) = [8, 33; 10]$.

Si P denota el costo de un viaje de 100 km, se tiene que $P = 4 \cdot Z = \frac{400}{x}$, consecuentemente:

$$\begin{aligned} E(P) &= \int_{-\infty}^{\infty} \frac{400}{x} f(x) dx = 400 \left[\int_{10}^{11} \frac{x-10}{x} dx + \int_{11}^{12} \frac{12-x}{x} dx \right] \\ &= 400 \left[\int_{10}^{11} \left(1 - \frac{10}{x} \right) dx + \int_{11}^{12} \left(\frac{12}{x} - 1 \right) dx \right] \\ &= 400 [(x - 10 \ln x)|_{10}^{11} + (12 \ln x - x)|_{11}^{12}] = 36,41. \end{aligned}$$

El costo esperado de un viaje de 100 Km en este auto es de \$36,41.

Ejemplo 3: La demanda X , por semana, de cierto producto es una variable aleatoria con distribución de probabilidades $f(x) = \frac{1}{5}$, si $x = 1, 2, 3, 4, 5$ y cero en cualquier otro caso. Se supone que el costo de fabricación es \$1 por artículo, mientras que el fabricante vende cada artículo por \$9. Cualquier artículo que no se venda al cabo de la semana debe almacenarse con un costo de \$1 por artículo. Si en una semana se fabrican cuatro artículos, ¿cuál es la utilidad esperada en esta semana?

Resolución:

Sea $U(X)$ la utilidad de una semana.

Si la demanda de artículos no supera al número de artículos fabricados, $x \leq 4$

$$U(x) = 9x - 4 - 1(4 - x) = 10x - 8,$$

puesto que, la utilidad es el ingreso por ventas menos el costo de producción de las 4 unidades y menos el costo de almacenamiento de las unidades no vendidas.

Por otro lado, si la demanda es de 5 unidades, y en la semana se produjeron 4 artículos, sólo se venden 4, y así

$$U(x) = 9 \times 4 - 4 = 32 \text{ para } x = 5.$$

De esta manera,

x	1	2	3	4	5
$U(x)$	2	12	22	32	32

$p(x)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
--------	---------------	---------------	---------------	---------------	---------------

Consecuentemente, la utilidad esperada se calcula de la siguiente manera

$$E(U(X)) = \sum_{x=1}^5 U(x)f(x) = \frac{1}{5} \sum_{x=1}^5 U(x) = \frac{1}{5}(2 + 12 + 22 + 32 + 32) = 20$$

Por lo tanto se tiene que la utilidad esperada por el fabricante en una semana en que produce 4 artículos es de \$20.

Ejemplo 4: Las variables aleatorias X e Y tienen la siguiente distribución conjunta:

$f(x,y)$		x		
		1	2	3
y	1	0	1/6	1/12
	2	1/5	1/9	0
	3	2/15	1/4	1/18

- Calcular $E(X)$, $V(X)$, $E(X \times Y^2)$, $Cov(X,Y)$ y $V(X+Y)$
- ¿Son independientes las variables dadas?
- Calcular $P(1 \leq X \leq 2, Y > 1)$ y $P(1 \leq Y \leq 2 | X = 2)$.

Resolución:

a) -Cálculo de $E(X)$ y $V(X)$: La distribución marginal de X es:

x	1	2	3
$g(x)$	$\frac{1}{3}$	$\frac{19}{36}$	$\frac{5}{36}$

Por lo tanto, $E(X) = \sum_{x=1}^3 xg(x) = 1 \times \frac{1}{3} + 2 \times \frac{19}{36} + 3 \times \frac{5}{36} = \frac{65}{36}$.

Para el cálculo de la varianza de X , se observa que

$$E(X^2) = \sum x^2 g(x) = 1^2 \times \frac{1}{3} + 2^2 \times \frac{19}{36} + 3^2 \times \frac{5}{36} = \frac{133}{36}.$$

Por lo tanto,

$$V(X) = E(X^2) - E^2(X) = \frac{133}{36} - \left(\frac{65}{36}\right)^2 = \frac{563}{1296}.$$

- Cálculo de $E(X \times Y^2)$, $Cov(X,Y)$ y $V(X+Y)$: Se observa que $X \cdot Y^2 = g(X,Y)$, por lo tanto teniendo en cuenta la definición **D2** se tiene que

$$\begin{aligned} E(X \cdot Y^2) &= \sum x \cdot y^2 f(x,y) \\ &= 1 \times 1^2 \times 0 + 2 \times 1^2 \times \frac{1}{6} + 3 \times 1^2 \times \frac{1}{12} + 1 \times 2^2 \times \frac{1}{5} + 2 \times 2^2 \times \frac{1}{9} + 3 \times 2^2 \times 0 + \\ &\quad + 1 \times 3^2 \times \frac{2}{15} + 2 \times 3^2 \times \frac{1}{4} + 3 \times 3^2 \times \frac{1}{18} \\ &= \frac{341}{36}. \end{aligned}$$

Considerando que $Cov(X,Y) = E(XY) - E(X)E(Y)$. Se puede trabajar inicialmente con la distribución marginal de Y y después calcular su esperanza.

y	1	2	3
$h(y)$	$\frac{1}{4}$	$\frac{14}{45}$	$\frac{79}{180}$

$$\begin{aligned}
 E(Y) &= \sum_{y=1}^3 yh(y) = 1 \times \frac{1}{4} + 2 \times \frac{14}{45} + 3 \times \frac{79}{180} = \frac{197}{90} \\
 E(X \cdot Y) &= \sum x \cdot y \cdot f(x, y) \\
 &= 1 \times 1 \times 0 + 2 \times 1 \times \frac{1}{6} + 3 \times 1 \times \frac{1}{12} + 1 \times 2 \times \frac{1}{5} + 2 \times 2 \times \frac{1}{9} + 3 \times 2 \times 0 + 1 \times 3 \times \frac{2}{15} \\
 &\quad + 2 \times 3 \times \frac{1}{4} + 3 \times 3 \times \frac{1}{18} \\
 &= \frac{689}{180} \\
 Cov(X, Y) &= \frac{689}{180} - \frac{65}{36} \times \frac{197}{90} = -\frac{403}{3240}
 \end{aligned}$$

Para el cálculo de la $V(X+Y)$ se observa que $V(X+Y) = V(X) + V(Y) + 2Cov(X, Y)$.
Por lo tanto

$$\begin{aligned}
 E(Y^2) &= \sum_{y=1}^3 y^2 h(y) = 1^2 \times \frac{1}{4} + 2^2 \times \frac{14}{45} + 3^2 \times \frac{79}{180} = \frac{980}{180} \\
 V(Y) &= E(Y^2) - [E(Y)]^2 = \frac{980}{180} - \left(\frac{197}{90}\right)^2 = \frac{5291}{8100} \\
 V(X + Y) &= \frac{563}{1296} + \frac{5291}{8100} - 2 \times \frac{403}{3240} = \frac{27179}{32400}
 \end{aligned}$$

b) Las variables X e Y no son independientes puesto que no se cumple

$$f(x, y) = g(x)h(y),$$

donde g y h denotan las distribuciones marginales de X e Y respectivamente. En particular se puede observar que, por ejemplo:

$$0 = f(1, 1) \neq g(1)h(1) = 1/3 \times 1/4.$$

$$c) P(1 \leq X \leq 2, Y > 1) = \frac{1}{5} + \frac{1}{9} + \frac{2}{15} + \frac{1}{4} = \frac{25}{36}$$

$$P(1 \leq Y \leq 2 | X=2) = P(1 \leq Y \leq 2 | X=2) = \frac{P(1 \leq Y \leq 2, X=2)}{P(X=2)} = \frac{\frac{1+1}{9}}{\frac{19}{36}} = \frac{10}{19}.$$

CAPÍTULO 5: DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

Distribuciones discretas de probabilidad: distribución Bernoulli, distribución discreta uniforme, distribución binomial, distribución hipergeométrica, distribución geométrica y distribución de Poisson.

Objetivos:

El alumno debe ser capaz de:

- Conocer la expresión matemática de la función de masa de las siguientes variables aleatorias: uniforme, Bernoulli, binomial, hipergeométrica, geométrica y Poisson.
- Para una distribución dada de una variable aleatoria discreta calcular su esperanza y varianza.
- Dado un problema que involucre variables aleatorias discretas identificar la distribución de probabilidad que describa su comportamiento.
- Calcular probabilidades bajo las distribuciones binomial y Poisson y utilizar estos valores para la resolución de problemas.

Resumen

D1. Distribución Bernoulli: La variable aleatoria X se dice que tiene distribución Bernoulli con parámetro π , ($X \sim B(\pi)$) si y sólo si

$$f(x; \pi) = \pi^x (1 - \pi)^{1-x}, x = 0, 1.$$

D2. Distribución Uniforme Discreta: La variable aleatoria X se dice que tiene distribución Uniforme discreta si y sólo si

$$f_X(x; k) = \frac{1}{k}, x = x_1, x_2, \dots, x_k.$$

D3. Distribución Binomial: La variable aleatoria X se dice que tiene distribución Binomial con parámetros n y π , ($X \sim b(n, \pi)$) si y sólo si

$$f_X(x; n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, x = 0, 1, \dots, n.$$

D4. Distribución Geométrica: La variable aleatoria X se dice que tiene distribución Geométrica con parámetro π , ($X \sim \text{Geom}(\pi)$), si y sólo si

$$f_X(x; \pi) = \pi(1 - \pi)^{x-1}, x = 1, 2, \dots$$

D5. Proceso de Bernoulli: Se dice que un proceso es de Bernoulli cuando presenta las siguientes características:

- (1) El proceso consiste en repeticiones (intentos) de un mismo procedimiento.
- (2) Los resultados de cada uno de los intentos pueden clasificarse como un éxito o un fracaso.
- (3) La probabilidad de éxito, representada por π , permanece constante para todos los intentos.
- (4) Los intentos repetidos son independientes.

T1. En un proceso de Bernoulli con n intentos y probabilidad de éxito π , la variable aleatoria que cuenta el número de éxitos en las n repeticiones tiene distribución Binomial con parámetros n y π .

D6. Distribución Hipergeométrica: La variable aleatoria X se dice que tiene distribución Hipergeométrica con parámetros N , n y k ($X \sim \text{Hip}(N, n, k)$) si y sólo si

$$f_X(x; N, n, k) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, x = 0, 1, \dots, \min(n, k).$$

T2. La variable aleatoria que cuenta el número de éxitos en una muestra aleatoria de tamaño n seleccionada, sin reposición, de un conjunto constituido por N elementos, de los cuales k son considerados éxitos y $N-k$ fracasos, tiene distribución Hipergeométrica con parámetros N, n y k .

D7. Distribución Poisson: La variable aleatoria X se dice que tiene distribución Poisson con parámetro β ($X \sim P(\beta)$) si y sólo si

$$f_X(x, \beta) = e^{-\beta} \frac{\beta^x}{x!}, x \in N_0$$

D8. Proceso de Poisson: Se dice que un proceso es de Poisson cuando consiste en observar la ocurrencia de un fenómeno concreto en un proceso físico particular y presenta las siguientes características:

(1) Hipótesis de independencia: el número de ocurrencias en dos intervalos de tiempo o regiones disjuntos son independientes.

(2) Hipótesis de proporcionalidad y homogeneidad: la probabilidad de exactamente una ocurrencia en un intervalo de tiempo muy pequeño, o una región muy pequeña, es proporcional a la longitud del intervalo, o al tamaño de la región, y no depende del intervalo en particular.

(3) Hipótesis de regularidad: la probabilidad de tener más de una ocurrencia en un intervalo de tiempo particular muy pequeño, o región muy pequeña, es despreciable.

T3. En un proceso de Poisson, la variable aleatoria que cuenta el número de ocurrencias en un intervalo de tiempo fijo $[0, t]$ tiene una distribución Poisson con parámetro λt , donde λ es el número medio de ocurrencias por unidad de tiempo.

T4. Sea X una variable aleatoria Binomial con parámetros n y π . Cuando $n \rightarrow \infty, \pi \rightarrow 0$, y $n\pi$ permanece constante, la distribución Binomial se puede aproximar por una distribución Poisson con media $n\pi$, es decir

$$b(x; n, \pi) - p(x; n\pi) \xrightarrow{n \rightarrow \infty} 0,$$

donde $b(x; n, \pi)$ y $p(x; n\pi)$ denotan las funciones de masa de la Binomial y la Poisson respectivamente.

Observación: La aproximación de la Binomial a la Poisson se considera razonable en la práctica, para un n finito, cuando: $\pi < 0,1$ y $n\pi > 1$.

Esperanza y varianza de algunas distribuciones discretas

Distribución	Función de masa	Parámetros	Esperanza	Varianza
$B(\pi)$	$\pi^x(1-\pi)^{1-x}, x = 0, 1$	π	π	$\pi(1-\pi)$
$Unif(k)$	$\frac{1}{k}, x = 1 \dots k$	k	$\sum_{i=1}^k x_i/k$	$\frac{\sum_{i=1}^k x_i^2}{k} - \frac{(\sum_{i=1}^k x_i)^2}{k^2}$
$b(n, \pi)$	$\binom{n}{x} \pi^x(1-\pi)^{n-x}, x = 0 \dots n$	n, π	$n\pi$	$n\pi(1-\pi)$
$P(\beta)$	$\frac{e^{-\beta} \beta^x}{x!}, x = 0, 1, \dots$	β	β	β
$Hip(N, n, k)$	$\frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}$	N, n, k	$\frac{nk}{N}$	$\frac{N-n}{N-1} n \frac{k}{N} \left(1 - \frac{k}{N}\right)$
$Geom(\pi)$	$\pi(1-\pi)^{x-1}, x = 1, 2, \dots$	π	$\frac{1}{\pi}$	$\frac{1-\pi}{\pi^2}$

Ejemplos

Ejemplo 1: Una empresa vitivinícola produce vinos finos y ha solicitado catadores expertos capaces de discernir entre un vino fino y uno ordinario el 90% de las veces, con solo degustar un sorbo de cada tipo. Todos los aspirantes al cargo deben probar nueve tipos de vino y decidir si se trata de vino fino u ordinario. La empresa ha determinado que quienes acierten por lo menos en seis de los nueve ensayos serán contratados.

- Determinar la probabilidad de que un individuo que no conoce nada de vinos y sólo está adivinando logre pasar la prueba y ser contratado.
- Calcular la probabilidad de que un catador experto (que en efecto es capaz de acertar el 90% de las veces) no logre pasar la prueba.
- Calcular la media y el desvío estándar del número de aciertos para un catador experto y para otro que adivina.

Resolución:

El proceso que se realiza es

ε : degustar un sorbo de 9 marcas de vino y clasificarlas como vino fino y ordinario.

Este proceso es de Bernoulli porque:

- Consiste en repeticiones de un mismo experimento (degustación y clasificación).
 - Los resultados de cada intento se pueden clasificar en éxito o fracaso (acierta o no la calidad del vino).
 - La probabilidad de éxito permanece constante para todos los intentos (se puede pensar que la probabilidad de acertar es inherente a cada catador y que no depende del tipo de vino que se está experimentando).
 - Se puede suponer que los intentos son independientes (el hecho de que acierte al calificar un vino en una degustación no modifica su chance de calificar correctamente a otro vino).
- a) Si un individuo no conoce nada de vinos (novato) y está adivinando, la probabilidad de que acierte es $\pi = \frac{1}{2}$. Para que un novato sea contratado, debe acertar en la clasificación de por lo menos 6 vinos.

Se define la variable aleatoria

X : número de clasificaciones correctas en las 9 degustaciones.

Se puede suponer que la variable tiene una distribución binomial con parámetros $n = 9$ y $\pi = \frac{1}{2}$, puesto que cuenta el número de éxitos en 9 intentos de un proceso Bernoulli.

$$\begin{aligned} P(X \geq 6) &= \sum_{x=6}^9 f\left(x; 9, \frac{1}{2}\right) = 1 - P(X \leq 5) \\ &= 1 - \sum_{x=0}^5 f\left(x; 9, \frac{1}{2}\right) = 1 - 0,7461 = 0,2539 \end{aligned}$$

Por lo tanto la probabilidad de que un novato pase la prueba es de 0,2539.

- b) Si el catador es experto, $\pi = 0,9$, no logrará pasar la prueba si clasifica correctamente menos de 6 vinos.

$$P(X < 6) = \sum_{x=0}^5 f(x; 9, 0,9) = 0,0083.$$

- c) Para el catador experto:

$$\begin{aligned} E(X) &= n\pi = 9 \times 0,9 = 8,1 \\ V(X) &= n\pi(1 - \pi) = 9 \times 0,9 \times 0,1 = 0,81 \\ \sigma &= \sqrt{V(X)} = \sqrt{0,81} = 0,9. \end{aligned}$$

Para el catador que adivina:

$$\begin{aligned} E(X) &= n\pi = 9 \times 1/2 = 4,5 \\ V(X) &= n\pi(1 - \pi) = 9 \times 1/2 \times 1/2 = 2,25 \\ \sigma &= \sqrt{V(X)} = \sqrt{2,25} = 1,5. \end{aligned}$$

Ejemplo 2: De un lote de discos usados, que contienen trabajos de alumnos para revisión, se estima que aproximadamente el 60% contiene virus. Si los discos se someten al detector de virus de McAfee, uno por uno, y el detector es perfecto (detecta toda vez que el disco tenga un virus).

- a) Calcular la probabilidad de que al examinar el séptimo disco se detecte virus por primera vez.
b) ¿Tuvo que agregar algún supuesto para poder resolver el ejercicio?

Resolución:

- a) El experimento aleatorio es
 ε : se examina una serie de discos con el antivirus McAfee hasta que se detecta virus por primera vez.

Se considera éxito (E) si se detecta virus y fracaso (F) si no se lo detecta.

En el contexto de variable aleatoria, se considera éxito a lo que se está buscando observar, aunque esté asociado a algo no deseable, como lo es el caso de que un disco tenga virus. En consecuencia, no se debe confundir éxito, con lo deseable.

El espacio muestral de este experimento es

$$S = \{E, FE, FFE, FFFE, FFFFE, \dots\}.$$

La variable aleatoria X asocia a cada resultado el número de discos observados hasta que se detecta virus por primera vez, por lo tanto

$$\text{Rec}(X) = \{1, 2, 3, 4, \dots\}.$$

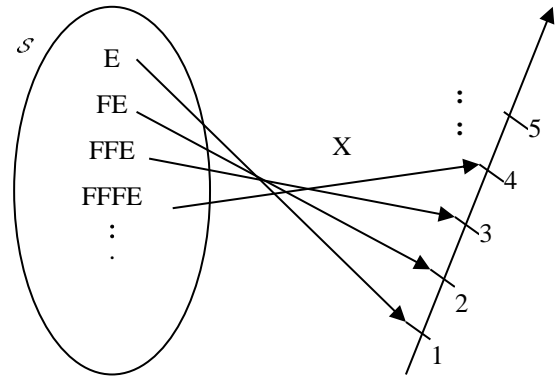
Consecuentemente, esta variable es discreta, porque tiene un recorrido infinito numerable.

Si los sucesivos ensayos son independientes, la distribución de probabilidad de esta variable se puede identificar con una distribución *geométrica* y su función de masa es

$$g(x; \pi) = (1 - \pi)^{x-1} \cdot \pi \quad x = 1, 2, 3, \dots$$

donde π denota la probabilidad de éxito. De esta manera la probabilidad de que al examinar el séptimo disco se detecte virus por primera vez se calcula de la siguiente manera:

$$P(X = 7) = g(7; 0,6) = (1 - 0,6)^{7-1} \times 0,6 = 0,00246.$$



- b) Se utilizó el supuesto de independencia de los ensayos para poder clasificar a la variable como geométrica.

La función de masa de la variable geométrica es $f(x) = (1 - \pi)^{x-1} \pi$, puesto que, si recién en el tiro x se obtuvo un éxito, se infiere que los primeros $x-1$ intentos fueron fracasos, consecuentemente, la probabilidad de tener $x-1$ fracasos y un éxito es $(1 - \pi)^{x-1} \pi$.

Ejemplo 3: El departamento de reservas de una aerolínea habilitó un centro de atención telefónica. Sea X la variable aleatoria que cuenta el número de llamadas que reciben en un intervalo de tiempo $[0, t]$.

- ¿Con qué distribución de probabilidad describiría Ud. la variable X ? ¿Por qué?
- Suponiendo que en promedio entran 2 llamadas cada 5 minutos, calcular la probabilidad de que al observar 3 períodos disjuntos de 10 minutos, en ninguno de ellos lleguen llamadas.
- Por cada diez minutos, el mantenimiento del centro de atención tiene un costo fijo de 2 pesos, y un costo adicional de 1 peso si se reciben más de 10 llamadas. Suponiendo que en media entran 2 llamadas cada 5 minutos, calcular el costo esperado por hora de funcionamiento del centro de atención.
- Resuelva los apartados b) y c) empleando Microsoft Office Excel y la aplicación Probability Distributions¹.

Resolución:

- a) X : número de llamadas telefónicas que entran a un conmutador en $[0, t]$.

X es una variable aleatoria Poisson y su función de masa es

$$f(x; \lambda) = e^{-\lambda t} (\lambda t)^x / x!, \quad x = 0, 1, 2, 3, \dots$$

donde λ representa el número medio de llamadas por unidad de tiempo.

¹ Aplicación desarrollada por Matthew Bognar de la universidad de Iowa, que calcula y grafica probabilidades de variables aleatorias con diferentes distribuciones, según los parámetros que se le indiquen. Disponible en forma gratuita en Google play.

Se afirma que sigue una distribución Poisson, puesto que puede considerarse que el proceso que consta en observar las llamadas a la central posee las siguientes características:

- (1) Hipótesis de independencia: el número de llamadas en un intervalo de tiempo $[t_1, t_2]$ es independiente del número de llamadas en otro intervalo de tiempo $[t_3, t_4]$, con $t_2 < t_3$.
 - (2) Hipótesis de proporcionalidad y homogeneidad: la probabilidad de que ocurra exactamente una llamada en un intervalo de tiempo muy pequeño es proporcional a la longitud del intervalo de tiempo y no depende del particular intervalo.
 - (3) Hipótesis de regularidad: la probabilidad de tener más de una llamada en un intervalo de tiempo muy pequeño es despreciable.
- b) Si se considera un lapso de tiempo de 10 minutos, $\lambda t = \frac{2}{5} \times 10 = 4$, y la probabilidad de que no lleguen llamadas en un período de 10 minutos es:

$$P(X = 0) = e^{-4} \times \frac{4^0}{0!} = 0,018.$$

Por la hipótesis de independencia del proceso de Poisson, el número de ocurrencias en dos intervalos de tiempos disjuntos son independientes, luego la probabilidad de seleccionar 3 períodos disjuntos de 10 minutos en los cuales no se registren llamadas es

$$P(X = 0) \times P(X = 0) \times P(X = 0) = 0,018^3 = 0,00000614.$$

- c) Sea X el costo del mantenimiento del centro por cada 10 minutos. La función costo, que se denota con C , puede considerarse de la siguiente manera:

$$C = \begin{cases} 2 & \text{si } X \leq 10 \\ 3 & \text{si } X > 10 \end{cases}$$

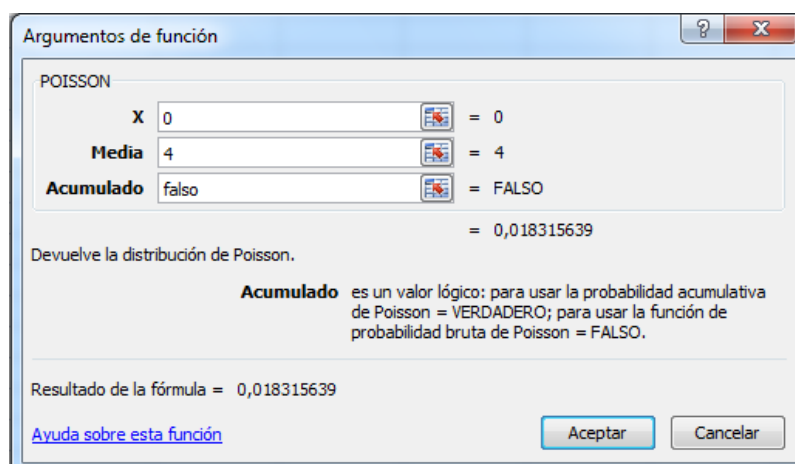
Por cada periodo de diez minutos el costo medio de mantenimiento es:

$$\begin{aligned} E(C) &= 2 \times P(X \leq 10) + 3 \times P(X > 10) \\ &= 2 \times 0,997 + 3 \times 0,028 = 2,003. \end{aligned}$$

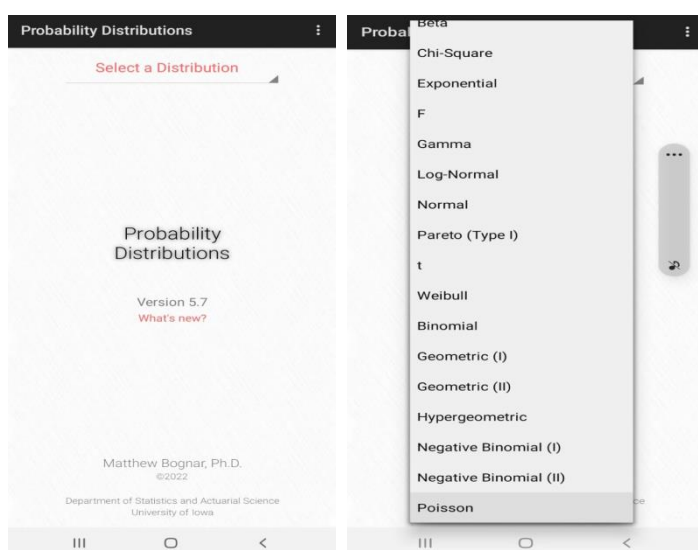
Luego el costo esperado por hora de funcionamiento es igual a $6 \times 2,003 = 12,02$. Puesto que es el costo esperado de los primeros diez minutos, más el de los segundos 10 minutos, y así sucesivamente hasta sumar el costo esperado de los últimos 10 minutos de la hora.

- d-b) Para encontrar la probabilidad de que en ninguno de los 3 períodos disjuntos de 10 minutos lleguen llamadas, primero se debe calcular el número medio de llamadas λ en un lapso de tiempo de 10 minutos, $\lambda t = \frac{2}{5} \times 10 = 4$ y, para encontrar la probabilidad, al usar Excel se selecciona la opción Fórmulas, Más Fórmulas, Estadísticas, POISSON.DIST y se procede a llenar el cuadro como se muestra en la imagen siguiente:

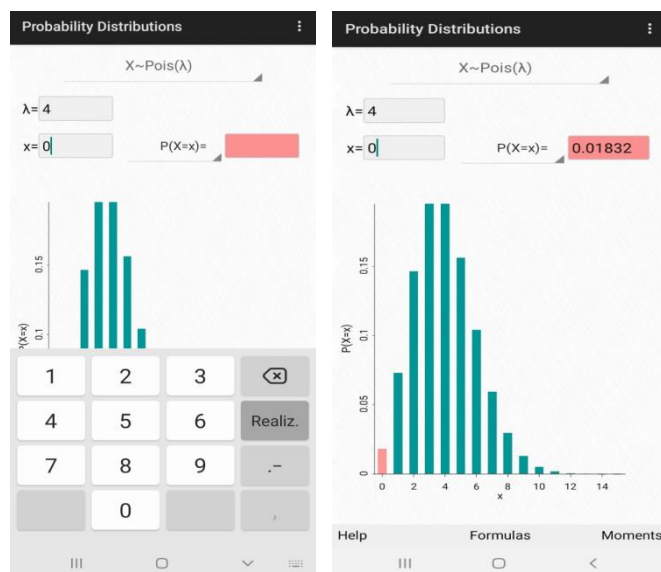
Por lo tanto, la probabilidad de que no lleguen llamadas en un lapso de tiempo de 10 minutos es $P(X=0)=0,018$.



También es posible calcular esta probabilidad utilizando la aplicación *Probability*



Distributions. Luego de instalar la aplicación seleccionamos la distribución bajo la cual queremos calcular la probabilidad, en este caso Poisson después especificamos su parámetro y el valor particular del recorrido para el cual deseamos calcular la probabilidad, como se indica en las imágenes, así obtenemos el resultado.



Luego la probabilidad de seleccionar 3 períodos disjuntos de 10 minutos en los cuales no se registren llamadas es:

$$P(X = 0) \times P(X = 0) \times P(X = 0) = 0.018^3$$

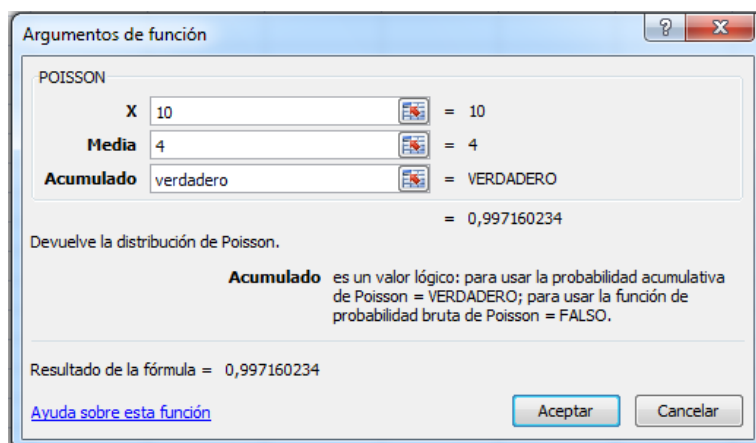
d-c) La función de costo es:

$$f(x) = \begin{cases} 2 & x \leq 10 \\ 3 & x > 10 \end{cases}$$

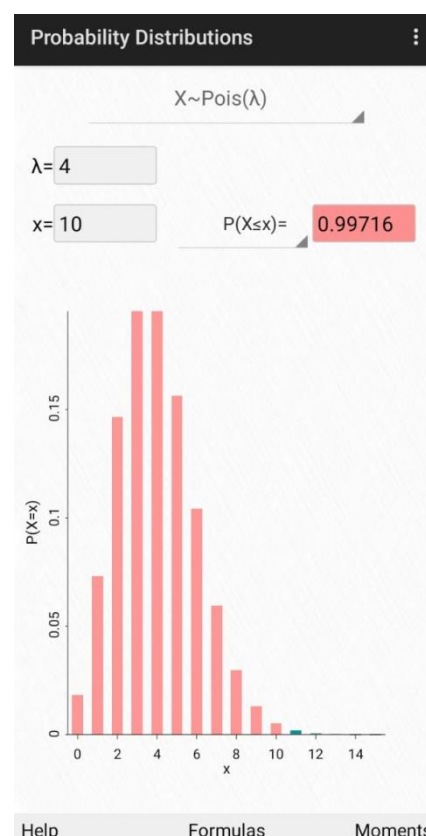
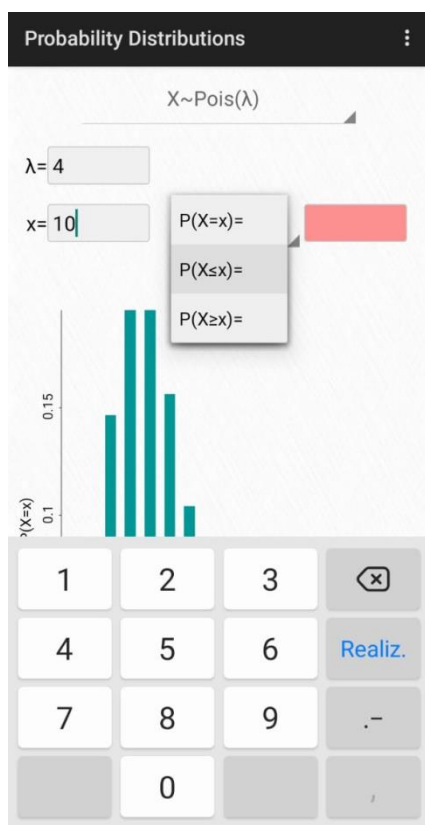
Por cada 10 minutos el costo de mantenimiento es:

$$E(C) = 2 \times P(X \leq 10) + 3 \times P(X > 10)$$

De manera que se emplea la fórmula POISSON.DIS en Excel para encontrar $P(X \leq 10)$:



También puede utilizar la aplicación Probability Distributions como se indica en la siguiente imagen:



Así la $P(X > 10) = 1 - P(X \leq 10) = 1 - 0,99716 = 0,028$ y reemplazando en la función de costos y tomando esperanza de la variable:

$$E(C) = 2 \times 0,997 + 3 \times 0,028 = 2,003.$$

Luego el costo esperado por hora de funcionamiento es $6 \times 2,003 = 12,02$.

CAPÍTULO 6: DISTRIBUCIONES CONTINUAS DE PROBABILIDAD

Distribuciones continuas de probabilidad: distribución Normal, distribución uniforme continua, distribución Exponencial. Aproximación de probabilidades binomiales usando la distribución Normal. Corrección por continuidad.

Objetivos:

El alumno debe ser capaz de:

- Reconocer la función de densidad y parámetros de las distribuciones estudiadas.
- Representar gráficamente las funciones de densidad de diversas variables aleatorias continuas, identificando los cambios que se producen en los gráficos al variar los parámetros.
- Calcular probabilidades, esperanza y varianza conociendo la función de densidad de la variable aleatoria continua involucrada.
- Calcular valores de probabilidad de variables aleatorias binomiales usando su aproximación a la distribución normal e identificando la corrección de continuidad que debe realizarse en cada caso.
- Resolver problemas de aplicación que involucren distintas distribuciones de probabilidad.
- Valorar la importancia del conocimiento de las distintas distribuciones de probabilidad en la medida que permiten modelar y resolver problemas de la vida real.

Resumen

D1. Distribución Uniforme Continua: La variable aleatoria X se dice que tiene distribución Uniforme Continua con parámetros α y β ($X \sim U(\alpha, \beta)$) si y sólo si

$$f_X(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{si } x \in [\alpha, \beta] \\ 0 & \text{caso contrario} \end{cases}$$

D2. Distribución Exponencial: La variable aleatoria X se dice que tiene distribución exponencial con parámetro β , ($X \sim \text{Exp}(\beta)$) si y sólo si

$$f_X(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & x > 0 \\ 0 & \text{otro caso} \end{cases}$$

D3. Distribución Normal Estándar: La variable aleatoria Z se dice que tiene distribución Normal Estándar ($Z \sim N(0, 1)$) si y sólo si

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, z \in \mathbb{R}.$$

D4. Distribución Normal: La variable aleatoria X se dice que tiene distribución Normal con parámetros μ y σ ($\sigma > 0$), ($X \sim N(\mu, \sigma^2)$) si y sólo si

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, x \in \mathbb{R}.$$

Propiedades de la distribución normal:

- (1) Si $X \sim N(\mu, \sigma^2)$, entonces $\frac{X-\mu}{\sigma} \sim N(0, 1)$.
- (2) Si $X \sim N(\mu, \sigma^2)$, entonces $E(X) = \mu$ y $V(X) = \sigma^2$.

Notación: * La función de distribución acumulada de una variable aleatoria Normal Estándar (Z) se denota con $\Phi(\cdot)$: $\Phi(z) = P(Z \leq z)$.

* z_α denota el valor del recorrido de la normal estándar que deja a su derecha una región de probabilidad α , es decir $\alpha = P(Z > z_\alpha)$.

T1. Si $X \sim b(n, \pi)$ entonces la distribución límite de $\frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$ cuando $n \rightarrow \infty$ es la distribución $N(0, 1)$.

T2. Si $X \sim b(n, \pi)$ entonces la proporción muestral ($\hat{\pi} = X/n$) es tal que la distribución límite de $\frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}}$ cuando $n \rightarrow \infty$ es la distribución $N(0, 1)$.

Esperanza y varianza de algunas distribuciones continuas

Distribución	Parámetros	Esperanza	Varianza
$U(\alpha, \beta)$	α, β	$\frac{\alpha + \beta}{2}$	$\frac{(\beta - \alpha)^2}{12}$
$Exp(\beta)$	β	β	β^2
$N(\mu, \sigma^2)$	μ, σ	μ	σ^2

Correcciones por continuidad: Se quiere calcular probabilidades relacionadas con X , una variable aleatoria discreta que asume únicamente valores enteros, y se sabe que una transformación de ella, por algún motivo, tiene una distribución que puede aproximarse con la distribución de una variable aleatoria Y continua. En el intento de mejorar dicha aproximación, antes del cálculo de probabilidades sobre X se realizan correcciones como las que se presentan a continuación.

$$(1) P(X = a) = P(a - 1/2 \leq X \leq a + 1/2) \quad (2) P(X < a) = P(X \leq a - 1/2)$$

$$(3) P(X \leq a) = P(X \leq a + 1/2) \quad (4) P(X > a) = P(X \geq a + 1/2)$$

$$(5) P(X \geq a) = P(X \geq a - 1/2)$$

Aunque, por tratarse de una variable aleatoria discreta, estas correcciones no modifican las probabilidades calculadas sobre X , si mejoran la aproximación al calcular la probabilidad basada en la distribución de Y .

Obs.: En situaciones donde la variable de interés es continua (Y) pero se mide por algún procedimiento que la discretiza, observando en realidad una variable X discreta, para calcular la probabilidad de observar algún evento (probabilidades sobre X) conociendo únicamente la distribución de Y , se debe corregir por continuidad siendo que la cantidad con la que se corrige (que en el caso anterior era $1/2$) ahora debe ser la mitad de la menor unidad que se puede observar (saltos en la variable X).

Ejemplos

Ejemplo 1: Una máquina produce esferas de metal, cuyos diámetros siguen una distribución normal con media $\mu = 5$ cm y desviación típica $\sigma = 0,2$ cm. Para los usos que tiene destinados, la esfera se considerará inservible si su diámetro cae fuera del intervalo $[4,8; 5,2]$ (en cm).

- ¿Qué porcentaje de esferas inservibles produce la máquina?
- ¿Cuál es la probabilidad de que entre 10 esferas elegidas al azar ninguna sea inservible?
- Resolver los apartados anteriores empleando Microsoft Excel y/o la aplicación Probability Distributions.

Resolución:

Sea X el diámetro de una esfera de metal producida por la máquina. Por hipótesis del problema, X tiene distribución Normal con parámetros $\mu = 5$ cm; $\sigma^2 = 0,04$ cm².

Si se realiza la transformación $Z = \frac{X-\mu}{\sigma}$ se obtiene una variable normal estándar, cuyos valores de probabilidad están tabulados.

- Se considera inservible la esfera, cuando su diámetro X es mayor que 5,2 cm, o menor que 4,8 cm,

$$\begin{aligned} P(X > 5,2 \text{ ó } X < 4,8) &= P(X < 4,8) + P(X > 5,2) = \\ &= P\left(\frac{X-5}{0,2} < \frac{4,8-5}{0,2}\right) + P\left(\frac{X-5}{0,2} > \frac{5,2-5}{0,2}\right) = P(Z < -1) + \\ &P(Z > 1) \end{aligned}$$

$$P(X > 5,2 \vee X < 4,8) = \Phi(-1) + [1 - \Phi(1)] = 2\Phi(-1) = 2 \cdot 0,15866 = 0,3173$$

Donde $\Phi(a)$ es el valor de la probabilidad acumulada hasta el valor a bajo la curva de la distribución Normal Estándar, recordemos que por la simetría de la distribución Normal Estándar alrededor de 0, $\Phi(-a) = 1 - \Phi(a)$. y que por ser la distribución Normal continua $P(Z \leq a) = P(Z < a) = \Phi(a)$. Por lo tanto, el 31,73% de las esferas son inservibles.

- Sea el experimento

ε : se seleccionan 10 esferas al azar y se observa si son o no inservibles.

Se puede suponer que el proceso de seleccionar aleatoriamente esferas y observar si son o no inservibles cumple las condiciones de un proceso de Bernoulli, puesto que (i) la selección es aleatoria, lo que asegura la independencia entre los resultados de una selección y otra, (ii) siempre se observa un éxito (la esfera es inservible) o fracaso (la esfera no es inservible), y (iii) la probabilidad de éxito es constante ($\pi = 0,3173$). Consecuentemente si se define la variable aleatoria

Y : Número de esferas inservibles entre las 10 observadas,

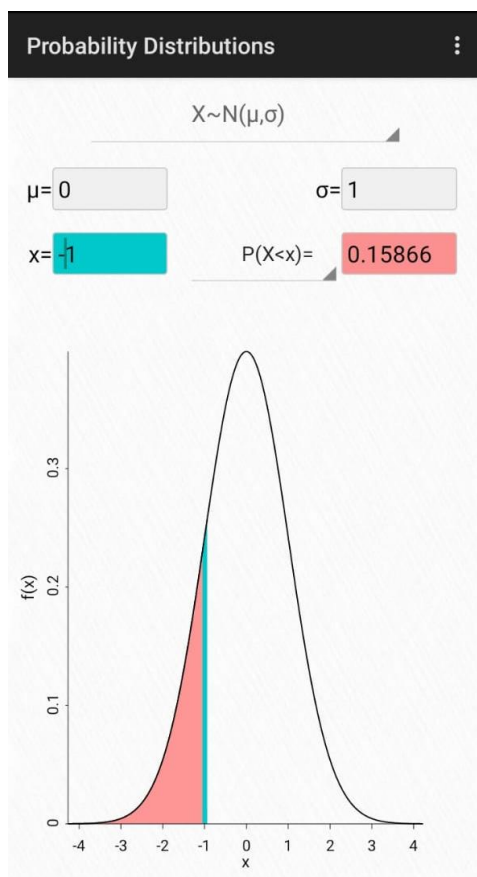
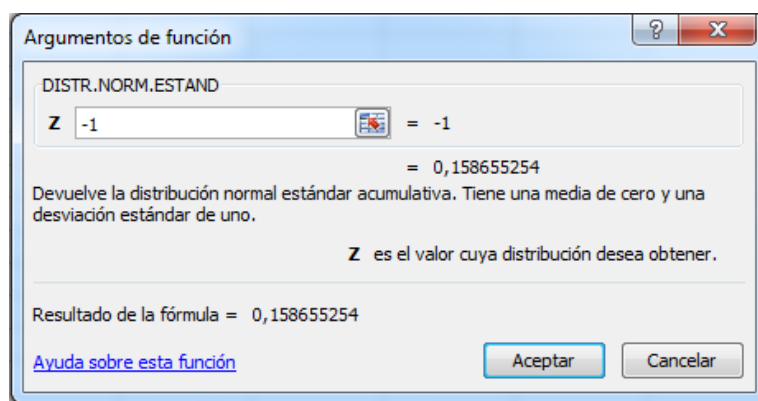
entonces Y tendrá distribución binomial con parámetros $n = 10$ y $\pi = 0,3173$ y así la probabilidad de que ninguna esfera sea inservible entre las 10 seleccionadas será la probabilidad de que $Y=0$,

$$P(Y = 0) = b(0; 10; 0,3173) = \binom{10}{0} 0,3173^0 \times 0,6827^{10} = 0,022.$$

- Para conocer cuál es el porcentaje de esferas inservibles que produce la máquina se procede a transformar X mediante un proceso de estandarización en una variable Normal Estándar Z :

$$P(X > 5,26 \text{ ó } X < 4,8) = 2 \cdot \Phi(-1)$$

Para conocer las probabilidades con Excel se recurre al Menú Fórmulas, Más Fórmulas y se usa la función DISTR.NORM.ESTAND.N:



Y utilizando la aplicación Probability Distributions:

Al conocer las probabilidades, se puede calcular el porcentaje de esferas inservibles:

$$\begin{aligned}
 P(X > 5,2 \vee X < 4,8) &= \\
 &= 2\Phi(-1) = 2 \cdot 0,15866 \\
 &= 0,3173
 \end{aligned}$$

c-b) Se define a Y como el número de esferas inservibles entre las 10 observadas, y empleando Excel, el Menú Fórmulas, Más Fórmulas, DISTR.BINOM.N se puede calcular la probabilidad de que ninguna de las 10 esferas elegidas al azar sea inservible al rellenar el cuadro de la siguiente forma:

Argumentos de función

DISTR.BINOM

Núm_éxito 0 = 0

Ensayos 10 = 10

Prob_éxito 0,3174 = 0,3174

Acumulado 0 = FALSO

= 0,021961542

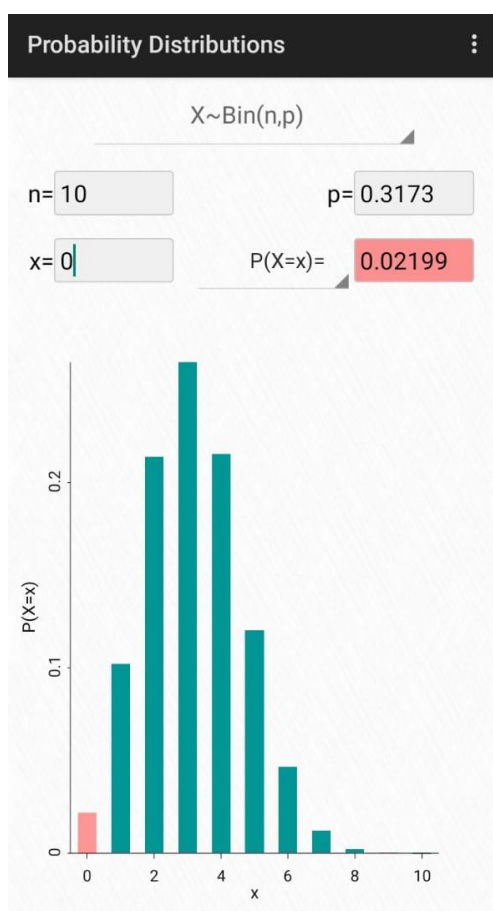
Devuelve la probabilidad de una variable aleatoria discreta siguiendo una distribución binomial.

Acumulado es un valor lógico: para usar la función de distribución acumulativa = VERDADERO; para usar la función de probabilidad bruta = FALSO.

Resultado de la fórmula = 0,021961542

[Ayuda sobre esta función](#)

Aceptar Cancelar



Con la aplicación

Por lo que $P(Y = 0) = b(0; 10; 0,3173) = 0,022$.

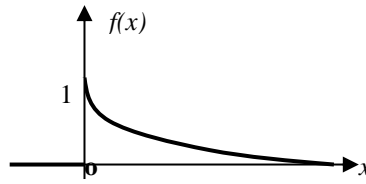
Ejemplo 2: Sea la variable aleatoria continua X , con distribución exponencial con parámetro $\beta = 1$. Representar gráficamente su función de densidad, calcular su mediana y la probabilidad de que asuma valores mayores que 1.

Resolución:

Si $\beta = 1$, la función de densidad de la variable será

$$f(x) = \begin{cases} e^{-x}, & \text{si } x > 0 \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

- El gráfico de f es:



- La mediana es el valor hasta el cual se acumula la mitad de la probabilidad.

$$P(X \geq Me) = P(X \leq Me) = \frac{1}{2}.$$

De esta manera,

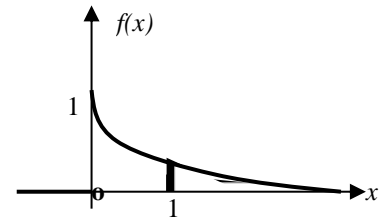
$$\begin{aligned} P(X \leq Me) &= \int_{-\infty}^{Me} f(x) dx = \int_0^{Me} e^{-x} dx = -e^{-x} \Big|_0^{Me} = -e^{-Me} + 1 = \frac{1}{2} \\ e^{-Me} &= \frac{1}{2} \\ -Me &= \ln \frac{1}{2} \\ Me &= -\ln \frac{1}{2} = 0,69. \end{aligned}$$

Por otro lado, $E(X) = \beta = 1 > Me = 0,69$. Con esto se observa que, a diferencia de lo que ocurre en la normal, no hay coincidencia entre la esperanza y la mediana porque la distribución exponencial es asimétrica positiva.

- Para calcular la probabilidad de que X asuma valores mayores que 1, se procede de la siguiente manera:

$$\begin{aligned} P(X > 1) &= \int_1^{\infty} f(x) dx = \int_1^{\infty} e^{-x} dx = \lim_{t \rightarrow \infty} \int_1^t e^{-x} dx \\ &= \lim_{t \rightarrow \infty} -e^{-x} \Big|_1^t = \lim_{t \rightarrow \infty} (-e^{-t} + e^{-1}) = e^{-1} = 0,37 \end{aligned}$$

Esta probabilidad representa el área comprendida entre el eje horizontal y la función de densidad, a la derecha de la recta $x=1$.



CAPÍTULO 7: DISTRIBUCIONES MUESTRALES

Muestreo aleatorio. Distribuciones muestrales. Teorema central del límite.

Objetivos:

El alumno debe ser capaz de:

- Definir población y muestra.
- Reconocer las condiciones para que una muestra particular sea aleatoria.
- Reconocer la diferencia entre estadístico, estimador y estimación.
- Identificar la distribución de algunos estadísticos.
- Analizar bajo qué condiciones es válido aplicar el Teorema Central del Límite.
- Conocer la importancia del Teorema Central del Límite en la Estadística.

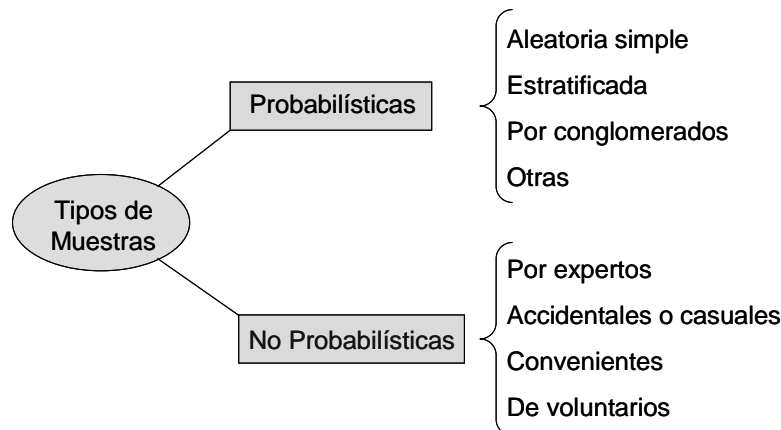
Resumen

D1. Población: Es el conjunto de todas las unidades/objetos en los cuales se quiere conocer el comportamiento de alguna característica de interés.

Observación 1: el conjunto de los valores asumidos por la característica de interés en la población se denomina *Población Estadística*, esta puede ser modelada por alguna distribución de probabilidad.

D2. Muestra: Es un subconjunto de la población.

Observación 2: Dependiendo el proceso por el cual se selecciona una muestra, las mismas se pueden clasificar según el siguiente esquema



D3. Muestra Aleatoria: Se dice que X_1, X_2, \dots, X_n constituye una muestra aleatoria (m.a.) de tamaño n de una población X , si X_1, X_2, \dots, X_n son variables aleatorias independientes e idénticamente distribuidas como X . Si $f(x)$ es la función de distribución de probabilidades de X , la distribución de probabilidades conjunta de la m.a. será:

$$f(x_1, x_2, \dots, x_n) = f_X(x_1)f_X(x_2)\dots f_X(x_n).$$

Observación 3: Cuando se habla de conocer alguna información sobre alguna característica de interés en una población, y si se supone que esa población (variable aleatoria) sigue alguna distribución, el problema se reduce a conocer los parámetros que caracterizan a esa distribución particular.

D4. Estadístico: Sea X_1, X_2, \dots, X_n m.a., un estadístico (T) es cualquier función de la m.a., es decir $T = g(X_1, X_2, \dots, X_n)$.

Notación: Generalmente los parámetros se denotan con letras griegas μ, θ, π, σ y los estadísticos con letras de imprenta mayúsculas o cuando ya están direccionados a estimar algún parámetro se los denota por $\hat{\mu}, \hat{\theta}, \hat{\pi}, \hat{\sigma}$.

Algunos estadísticos importantes

Media muestral: $\bar{X} = n^{-1} \sum_{i=1}^n X_i$

Mediana muestral: $Me = \begin{cases} X_{[(n+1)/2]} & \text{si } n \text{ es impar} \\ \frac{(X_{[n/2]} + X_{[n/2+1]})}{2} & \text{si } n \text{ es par} \end{cases}$, siendo $X_{[1]}, X_{[2]}, \dots, X_{[n]}$ la m.a. de

tamaño n acomodada en orden de magnitud creciente.

Moda: Es el valor de la muestra que ocurre con mayor frecuencia

Rango: $R = X_{[n]} - X_{[1]}$ siendo $X_{[1]}$ y $X_{[n]}$ las observaciones más pequeña y más grande de la muestra respectivamente.

Varianza Muestral: $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Desviación estándar muestral: es la raíz cuadrada positiva de la varianza muestral

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

T1. $S^2 = [n(n-1)]^{-1} \{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2\}$

D5. Distribución muestral: Es la distribución de probabilidad de un estadístico.

T2 Reproducibilidad de la distribución Normal: Si X_1, X_2, \dots, X_n son variables aleatorias independientes que tienen distribuciones normales con medias $\mu_1, \mu_2, \dots, \mu_n$ y varianzas $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$. Entonces: $Y = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$

tiene una distribución Normal con media $\mu_Y = a_1 \mu_1 + a_2 \mu_2 + \dots + a_n \mu_n$ y varianza $\sigma_Y^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2$.

T3. Teorema del límite central: Si \bar{X} es la media de una muestra aleatoria de tamaño n que se toma de una población con media μ y varianza finita σ^2 , entonces la forma límite de la distribución de

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

conforme $n \rightarrow \infty$, es la distribución normal estándar $n(z; 0, 1)$.

Observación 4: Para conocer las distribuciones muestrales de otros estadísticos que se utilizarán en esta materia consulte el Anexo I.

Ejemplos

Ejemplo 1: Se extrae una muestra aleatoria de la población $\{1, 2, 3\}$.

a) Determine la distribución muestral de \bar{X} en los siguientes casos:

- $n = 2$, muestreo con reposición.
- $n = 2$, muestreo sin reposición.
- $n = 3$, muestreo con reposición.

b) Responde las siguientes preguntas:

- ¿La distribución muestral depende del tamaño de la muestra?
- ¿La distribución muestral depende del tipo de muestreo realizado?

Resolución:

La **distribución muestral** es la distribución de probabilidad de un estadístico. Para conocerla, en situaciones donde la población es finita y el tamaño muestral es pequeño, es suficiente con enumerar todas las posibles muestras con sus respectivas probabilidades y el valor que asume el estadístico para cada muestra.

a) i) La siguiente tabla presenta la enumeración de todas las posibles muestras de tamaño $n=2$, con reposición, conjuntamente con el valor asumido por el estadístico al ser evaluado en cada muestra.

Muestra $\{s_1, s_2\}$	$\{1, 1\}$	$\{1, 2\}$	$\{1, 3\}$	$\{2, 1\}$	$\{2, 2\}$	$\{2, 3\}$	$\{3, 1\}$	$\{3, 2\}$	$\{3, 3\}$
\bar{x}	1	1,5	2	1,5	2	2,5	2	2,5	3

Se debe observar que en este caso, cada muestra tiene la misma probabilidad de aparecer. $P\{s_1, s_2\} = 1/9$ para cualquier $\{s_1, s_2\}$. Consecuentemente, la distribución de probabilidad del estadístico es la siguiente:

\bar{x}	1	1,5	2	2,5	3
$f(\bar{x})$	1/9	2/9	1/3	2/3	1/9

ii) La siguiente tabla presenta la enumeración de todas las posibles muestras de tamaño $n=2$, sin reposición, conjuntamente con el valor asumido por el estadístico al ser evaluado en cada muestra.

Muestra	$\{1, 2\}$	$\{1, 3\}$	$\{2, 1\}$	$\{2, 3\}$	$\{3, 1\}$	$\{3, 2\}$
\bar{x}	1,5	2	1,5	2,5	2	2,5

Considerando que cada muestra tiene probabilidad $1/6$ de aparecer, la distribución de probabilidad del estadístico es la siguiente:

\bar{x}	1,5	2	2,5
$f(\bar{x})$	1/3	1/3	1/3

iii) La siguiente tabla presenta la enumeración de todas las posibles muestras de tamaño $n=3$, con reposición, conjuntamente con el valor asumido por el estadístico al ser evaluado en cada muestra.

Muestra	\bar{x}
$\{1, 1, 1\}$	1
$\{1, 1, 2\}$	4/3
$\{1, 1, 3\}$	5/3
$\{1, 2, 1\}$	4/3
$\{1, 2, 2\}$	5/3
$\{1, 2, 3\}$	2
$\{1, 3, 1\}$	5/3
$\{1, 3, 2\}$	2
$\{1, 3, 3\}$	7/3
$\{2, 1, 1\}$	4/3
$\{2, 1, 2\}$	5/3
$\{2, 1, 3\}$	2
$\{2, 2, 1\}$	5/3
$\{2, 2, 2\}$	2

Muestra	\bar{x}
$\{2, 2, 3\}$	7/3
$\{2, 3, 1\}$	2
$\{2, 3, 2\}$	7/3
$\{2, 3, 3\}$	8/3
$\{3, 1, 1\}$	5/3
$\{3, 1, 2\}$	2
$\{3, 1, 3\}$	7/3
$\{3, 2, 1\}$	2
$\{3, 2, 2\}$	7/3
$\{3, 2, 3\}$	8/3
$\{3, 3, 1\}$	7/3
$\{3, 3, 2\}$	8/3
$\{3, 3, 3\}$	3

Considerando que cada muestra tiene probabilidad $1/27$ de aparecer, la distribución de probabilidad del estadístico es la siguiente:

\bar{x}	1	4/3	5/3	2	7/3	8/3	3
$f(\bar{x})$	1/27	1/9	2/9	7/27	2/9	1/9	1/27

- b) De los incisos i) y iii) anteriores se puede observar que si cambia el tamaño de la muestra puede cambiar la distribución del estadístico. Entonces, la distribución muestral sí depende del tamaño de muestra seleccionado.

De los incisos i) y ii) es evidente que si cambia el tipo de muestreo puede cambiar la distribución del estadístico. Entonces, la distribución muestral también depende del tipo de muestreo realizado.

Ejemplo 2: Se extrae una muestra aleatoria de tamaño 54 de una población con distribución de probabilidad Binomial con parámetros $n = 3$, $\pi = 1/3$.

- a) ¿Cuál es la probabilidad de que la media muestral supere a $10/9$? ¿Qué teorema permite asegurar este resultado?
b) ¿Se llegaría a un resultado diferente si la población fuera normal con $\mu = 1$, $\sigma^2 = 2/3$?
c) ¿Podría usar el mismo teorema utilizado en a) si la muestra fuera de tamaño 9?

Resolución:

- a) Para poder calcular una probabilidad necesitamos conocer previamente la distribución muestral del estadístico. Cuando el tamaño muestral (n') es grande, el Teorema Central del Límite asegura que la distribución de la media se aproxima a una normal. El tamaño de muestra $n' = 54$ es un tamaño razonablemente grande para usar tal aproximación.

$$E(\bar{X}) = E(X) = n\pi = 3 \times \frac{1}{3} = 1, V(\bar{X}) = \frac{V(X)}{n'} = \frac{n\pi(1-\pi)}{n'} = \frac{3 \times 1/3 \times 2/3}{54} = \frac{1}{81}$$

Para dejar en claro la notación empleada, n' denota el tamaño muestral, es decir el tamaño de la muestra tomada que en este caso es 54. Por otra parte, n denota el parámetro de la distribución Binomial que está siendo muestreada, en este caso $n=3$.

$$P\left(\bar{X} > \frac{10}{9}\right) = P\left(\frac{\bar{X} - 1}{\sqrt{1/81}} > \frac{10/9 - 1}{1/9}\right) \cong P(Z > 1) = 0,1587$$

- b) En el caso en que la población de origen fuera normal con $\mu = 1$, $\sigma^2 = \frac{2}{3}$, también se obtiene

$$E(\bar{X}) = E(X) = 1, Var(\bar{X}) = \frac{Var(X)}{n'} = \frac{\sigma^2}{n'} = \frac{2/3}{54} = \frac{1}{81},$$

y se procede a realizar la misma estandarización realizada en a), obteniéndose el mismo resultado 0,1587.

La diferencia en este caso es que, si la muestra proviene de una población normal, la media muestral tiene distribución normal, por lo tanto, la probabilidad antes calculada sería exactamente igual a 0,1587, no aproximadamente igual a ese valor.

- c) No se podría haber usado el Teorema Central del Límite para un tamaño muestral 9, que es muy pequeño como para considerar válida la aproximación de las probabilidades con la distribución asintótica.

Ejemplo 3: La primera tarea de un curso introductorio de programación consiste en codificar, editar y correr un breve programa. La experiencia indica que el tiempo que les insume esta tarea a los estudiantes es una variable aleatoria uniforme en el intervalo (24 minutos, 48 minutos).

- a) Genere con su computadora 1.000 muestras de tamaño 75 de esta población y construya la distribución de medias.

- b) Calcule la media y varianza de la distribución muestral construida. Compare estos resultados con los valores de los parámetros poblacionales.
c) ¿Qué observa de la forma de la distribución muestral que construyó?

Resolución:

Para resolver el ejercicio usando Microsoft Excel se requiere de la herramienta Análisis de Datos ubicada en el Menú Datos; en caso de no contar con la misma a continuación se detallan los pasos para su instalación:

1° Abrir Excel

2° En el Menú Archivo, seleccionar Opciones y luego Complementos (en la parte inferior del cuadro de dialogo),

3° En complemento de aplicaciones inactivas seleccionar Herramientas para análisis y Aceptar.

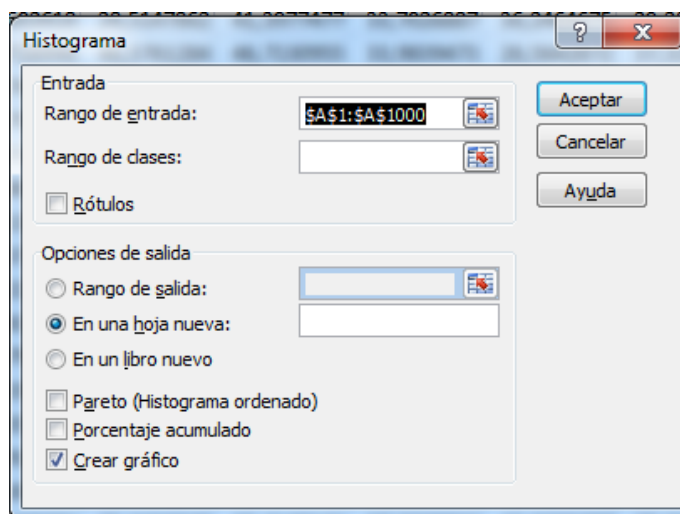
- a) Para generar las 1.000 muestras de tamaño 75 se selecciona en el Menú Datos, Análisis de Datos y luego Generación de Números Aleatorios. En la ventana de generación de números aleatorios se llena el cuadro con los datos del ejercicio:

El software genera las 1.000 muestras de tamaño 75 y las presenta en una hoja nueva, donde cada columna es una muestra.

Para construir la distribución muestral de la media se calcula la media de cada muestra, utilizando la función PROMEDIO. A continuación, se muestra la pantalla de Excel con las muestras generadas y las medias calculadas resaltadas:

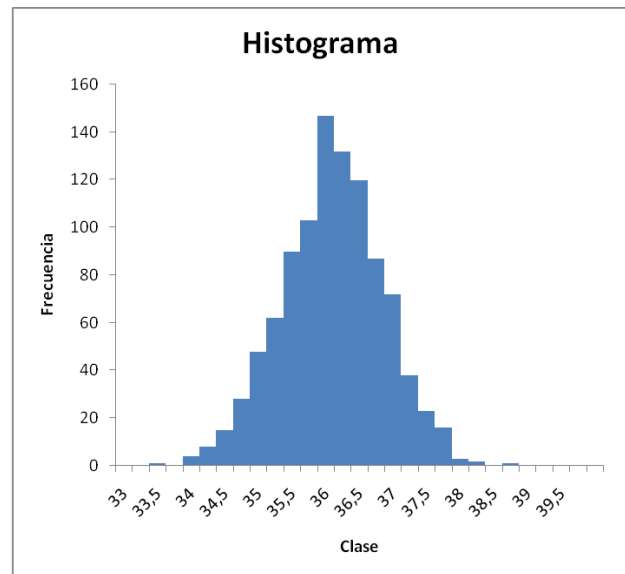
	PROMEDIO								
	=PROMEDIO(A1:A75)								
	A	B	C	D	E	F	G	H	I
57	38,83492538	43,7437666	31,111301	29,5760979	47,9362774	37,225013	27,3670461	41,5698721	25,092
58	39,54466384	43,1856441	28,5470138	24,5800958	33,2075564	37,8981292	25,8486892	26,7093112	37,63
59	39,15866573	34,4182867	28,0870388	35,8311716	37,0550859	47,8981903	32,1909238	27,7530442	43,012
60	47,4418775	27,6212043	24,7932371	25,3667409	43,7086093	47,5034028	42,6502274	28,6473586	27,232
61	36,13293863	42,044496	37,5048067	42,8838771	41,4197211	43,9246803	47,6857814	45,73455	26,11
62	41,4153264	38,6056703	47,0624714	27,0821253	42,4334239	35,4012268	43,6778466	44,5912046	38,765
63	41,38602863	47,4843593	45,5221412	42,0100711	45,9520859	26,2896207	35,6451308	27,8775597	28,05
64	40,52394177	34,7134617	28,3778191	36,4017457	40,1079134	45,1632435	30,9677419	30,1788995	32,063
65	25,15652944	35,7747734	25,7798395	38,0314341	42,9322184	45,9008148	24,09888	43,8477737	26,585
66	29,93353069	28,1917783	28,7513657	39,0466018	38,3192846	27,7816095	44,2923673	43,8485061	42,567
67	27,29160436	47,1239967	26,4068117	28,3192236	47,0009461	38,5851619	26,3108615	39,6025269	34,588
68	47,92455824	39,9592273	34,1751152	26,784753	44,5106357	36,6734825	32,1557665	33,8315989	28,328
69	28,24890896	35,9102756	41,013947	29,7899716	38,1574145	44,0946074	31,7778253	32,3996704	43,037
70	32,88015381	30,6059145	43,6287729	24,4321421	38,9880062	38,6730552	34,7998901	33,1592151	29,004
71	36,09338664	41,6863308	47,6953032	33,9070406	24,7610096	41,1018403	45,2335582	41,7925352	36,025
72	30,29755547	47,3950011	36,0736106	44,8504898	37,9149754	37,0397046	31,1977294	42,6582842	36,867
73	29,50065615	24,1743217	24,0300302	33,6814478	34,5318155	46,2018494	29,6420179	44,631489	29,748
74	39,78270821	35,2598651	33,4565874	43,9481185	36,3138524	26,4536882	26,0589007	33,9517197	35,795
75	44,75746941	34,2710654	34,0850246	26,6690268	33,5459456	44,9098178	41,2065798	26,0632954	39,517
76	=PROMEDIO(A1:A75)		35,3197986	34,8306723	36,4948247	36,6867153	36,0369884	35,1331718	35,307
77	PROMEDIO(número1; [número2]; ...)								

Para graficar la distribución muestral de la media se puede construir un Histograma. Excel permite realizarlo con la función Histograma en el menú análisis de datos.



En Rango de entrada se completa con las celdas de las 1.000 medias muestrales calculadas, ordenadas de menor a mayor, y en Rango de clases se colocan los límites superiores de los intervalos de clase que se desea, si no se completa esta ventana el programa los calculará automáticamente. De esta manera, en una hoja nueva se obtienen los intervalos y las frecuencias de los mismos junto con el Histograma como se muestra a continuación:

Intervalo de clases (minutos)	Frecuencia
más de 33 a 33,25	0
más de 33,25 a 33,5	1
más de 33,5 a 33,75	0
más de 33,75 a 34	4
más de 34 a 34,25	8
más de 34,25 a 34,5	15
más de 34,5 a 34,75	28
más de 34,75 a 35	48
más de 35 a 35,25	62
más de 35,25 a 35,5	90
más de 35,5 a 35,75	103
más de 35,75 a 36	147
más de 36 a 36,25	132
más de 36,25 a 36,5	120
más de 36,5 a 36,75	87
más de 36,75 a 37	72
más de 37 a 37,25	38
más de 37,25 a 37,5	23
más de 37,5 a 37,75	16
más de 37,75 a 38	3
más de 38 a 38,25	2
más de 38,25 a 38,5	0
más de 38,5 a 38,75	1
más de 38,75 a 39	0



Distribución muestral de la media de 1.000
muestras de una distribución $U(24, 48)$

- b) La media de la distribución de \bar{X} , se estima con el promedio de las medias $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_{1000}$, donde \bar{X}_j denota la media muestral de la j -ésima muestra, para $j=1, \dots, 1.000$. Este promedio se denota con $\bar{\bar{X}}$, medias de medias, y en el caso de las muestras seleccionadas, su valor observado resulta $\bar{\bar{x}} = 35,98$, que es la estimación del parámetro solicitado.

La varianza de \bar{X} se estima usando los valores $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{1000}$ observados, y su estimador se denota con $S_{\bar{X}}^2$ y en el caso de las muestras seleccionadas la estimación resultó $s_{\bar{x}}^2 = 0,576$.

Se observa que $\bar{\bar{x}}$ es muy próximo a 36, valor de la media de la distribución poblacional a partir de la cual se simularon las muestras y también coincide con la esperanza de la distribución muestral de \bar{X} . Por otro lado, se sabe que $V(\bar{X})$ es $\frac{V(X)}{n} = 0,64$, valor no muy distante de la estimación realizada con $s_{\bar{x}}^2$.

- c) La distribución muestral construida es claramente simétrica, con forma aproximadamente de campana, que podría parecerse a una campana de Gauss (distribución Normal), forma que era de esperar en virtud del Teorema Central del Límite.

CAPÍTULO 8: ESTIMACIÓN PUNTUAL

Inferencia Estadística. Métodos clásicos de estimación. Estimación Puntual. Propiedades de los estimadores. Estimación de la media. Estimación de diferencias entre dos medias. Estimación de una proporción. Estimación de diferencia entre dos proporciones. Estimación de varianzas. Estimación de la razón entre dos varianzas. Estimadores de máxima verosimilitud.

Objetivos:

El alumno debe ser capaz de:

- Deducir conclusiones acerca de los parámetros de la población a partir de la información obtenida en una muestra aleatoria.
- Identificar las propiedades de los estimadores o estadísticos.
- Dada una muestra aleatoria de tamaño n , con distribución de probabilidad conocida con excepción de alguno de sus parámetros, calcular los estimadores máximo verosímiles de los mismos.

Resumen

D1. Parámetro: Las características de la población, en general desconocidas, sobre las cuales tenemos interés se denominan parámetros y usualmente se representan por letras griegas tales como μ, θ, π, σ , entre otras.

D2. Estimador: Un estimador es un estadístico cuya forma funcional no involucra ningún parámetro y que se construye con la intención de atribuir un valor que pudiera representar un parámetro de interés desconocido en la población.

D3. Estimación puntual: Es el valor asumido por un estimador cuando se evalúa en los valores obtenidos en una muestra particular.

D4. Estimador insesgado: Se dice que $\hat{\theta}$ es un estimador insesgado para el parámetro θ si $E(\hat{\theta}) = \theta$.

D5. Sesgo: $E(\hat{\theta}) - \theta$

Obs: Un estimador se dice *sesgado* cuando el sesgo es no nulo.

D6. Error Cuadrático Medio: $ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$

Estimadores para media, proporción y varianza

<u>Parámetro</u>	<u>Estimador</u>	<u>Propiedades</u>
μ	$\bar{X} = \frac{X_1 + \dots + X_n}{n}$	Inssegado
π	$\hat{\pi} = \frac{\text{Nro de éxitos en la muestra}}{n}$	Inssegado
σ^2	$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$	Inssegado
σ^2	$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$	Sesgado

D7. Eficiencia: Dados dos estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ insesgados para el parámetro θ , decimos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.

D8. Función de Verosimilitud: Sean X_1, X_2, \dots, X_n i.i.d. como X con función de distribución de probabilidad $f_X(x; \theta)$. Entonces la función de verosimilitud es

$$L(\theta) = f(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_X(x_i; \theta)$$

considerada como función de θ dados los valores de la muestra (x_1, x_2, \dots, x_n) .

Estimación por Máxima Verosimilitud: Dadas las observaciones x_1, x_2, \dots, x_n de una muestra aleatoria X_1, X_2, \dots, X_n de una población con f.d.p. $f_X(x; \theta)$, la estimación por máxima verosimilitud de θ , $\hat{\theta}$, es aquella función de los datos que maximiza la función verosimilitud

$$L(\theta) = f(\theta; x_1, x_2, \dots, x_n).$$

Obs: En la búsqueda de un estimador para un parámetro, la elección del estimador podrá basarse en sus propiedades o en el método de obtención del mismo.

Ejemplos

Ejemplo 1:

Se obtuvo una muestra al azar de una población exponencial con media β , resultando los valores 3, 5, 4, 3, 2, 4, 3 y 5. Obtenga la estimación de β por máxima verosimilitud.

Resolución:

Sean X_1, X_2, \dots, X_n i.i.d $X \sim \text{Exp}(\beta)$

Considerando que la función de densidad de la distribución exponencial es:

$$f(x) = \begin{cases} (1/\beta)e^{(-x/\beta)} & \text{si } x > 0 \\ 0 & \text{en otro caso,} \end{cases}$$

para obtener la estimación de β por máxima verosimilitud, se maximiza la función de verosimilitud

$$L(\beta; x_1, x_2, \dots, x_n) = (1/\beta)e^{(-x_1/\beta)} \dots (1/\beta)e^{(-x_n/\beta)} = \beta^{-n} \exp \sum_{i=1}^n -(x_i/\beta),$$

que es la distribución conjunta de las variables aleatorias de la muestra aleatoria vista como función del parámetro.

Puesto que el logaritmo es una función estrictamente creciente, el máximo de la función de verosimilitud coincide con el máximo de su logaritmo, y ya que en la mayoría de las funciones de probabilidad la búsqueda del máximo es más sencilla si se trabaja con el logaritmo, se recomienda aplicar el logaritmo natural cada vez que se busque maximizar la verosimilitud.

Aplicando logaritmo natural

$$\ln L(\beta) = -n \ln \beta - \sum_{i=1}^n x_i / \beta,$$

el máximo de la verosimilitud se busca derivando $\ln L(\beta)$ con respecto a β e igualando a cero:

$$\begin{aligned} \frac{\partial \ln[L(\beta)]}{\partial \beta} &= -\frac{n}{\beta} - (-1)\beta^{-2} \sum_{i=1}^n x_i = 0 \\ -\frac{n}{\beta} + \frac{\sum_{i=1}^n x_i}{\beta^2} &= 0 \\ \frac{-n\beta + \sum_{i=1}^n x_i}{\beta^2} &= 0 \\ -n\beta + \sum_{i=1}^n x_i &= 0. \end{aligned}$$

Luego el valor de β que maximiza la función de verosimilitud es $\frac{\sum_{i=1}^n x_i}{n}$.

Por lo tanto, el estimador máximo verosímil de β es $\hat{\beta} = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}$.

En este ejemplo particular la estimación por máxima verosimilitud de β es 3,25.

Se puede verificar que la segunda derivada de la función de verosimilitud evaluada en el punto crítico encontrado es menor que cero lo que confirma que se trata de un punto de máximo de $L(\beta)$.

Ejemplo 2:

Evalúe si el estimador $S_*^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ es un estimador insesgado del parámetro σ^2 .

Resolución:

El numerador de S_*^2 puede expresarse de la siguiente manera:

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - \sum_{i=1}^n 2(\bar{X} - \mu)(X_i - \mu) + \sum_{i=1}^n (\bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + n(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \end{aligned}$$

Aplicando esperanza,

$$\begin{aligned} E(S_*^2) &= E \left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^n E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \right] = \frac{1}{n} (\sum_{i=1}^n \sigma_{X_i}^2 - n\sigma_{\bar{X}}^2) \end{aligned}$$

Como, $\sigma_{X_i}^2 = \sigma^2$ para $i = 1, 2, \dots, n$ y $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$, se tiene que,

$$E(S_*^2) = \frac{1}{n} (n\sigma^2 - n\frac{\sigma^2}{n}) = \left[\frac{n-1}{n} \right] \sigma^2 \neq \sigma^2.$$

Luego el estimador S_*^2 no es insesgado.

Ejemplo 3:

Al finalizar un curso, los alumnos pueden promocionar la materia, quedar regulares o quedar libres. Estimar por máxima verosimilitud la probabilidad de promocionar, si se sabe que ésta es 1/3 de la probabilidad de regularizar la materia, y que en una muestra aleatoria de 40 alumnos se encontraron 7 que promocionaron, 17 de regularizaron y el resto quedaron libres.

Resolución:

Sea X la variable aleatoria que asume el valor 0 si el alumno queda libre, 1 si regulariza y 2 si promociona.

Debido a que la probabilidad de promocionares $1/3$ de la de regularizar, la función de masa de X resulta ser:

x	0	1	2
$p(x)$	$1 - 4\theta$	3θ	θ

donde θ es la probabilidad de promocionar la materia.

La verosimilitud en este caso puede escribirse como

$$L(\theta; x_1, \dots, x_n) = \theta^{n_2} (3\theta)^{n_1} (1 - 4\theta)^{n_0},$$

donde n_k es el número de elementos en la muestra que presentan el valor k , para $k=0,1$ y 2.

Para obtener el estimador por máxima verosimilitud, se escribe la log-verosimilitud y se deriva respecto del parámetro que se quiere estimar. Es decir:

$$\begin{aligned} \ln L(\theta; x_1, \dots, x_n) &= n_2 \ln \theta + n_1 (\ln 3 + \ln \theta) + n_0 \ln(1 - 4\theta) \\ &= (n_2 + n_1) \ln \theta + n_0 \ln(1 - 4\theta) + n_1 \ln 3 \\ \frac{d \ln L(\theta; x_1, \dots, x_n)}{d\theta} &= \frac{n_2 + n_1}{\theta} + \frac{n_0}{1 - 4\theta} (-4) \\ &= \frac{(1 - 4\theta)(n_2 + n_1) - 4\theta n_0}{\theta(1 - 4\theta)} \\ &= \frac{(n_2 + n_1) - 4\theta(n_2 + n_1 + n_0)}{\theta(1 - 4\theta)} \end{aligned}$$

Igualando a cero esta expresión se obtiene el estimador por MV de θ es $\hat{\theta} = \frac{n_2 + n_1}{4n}$, con $n = n_2 + n_1 + n_0$. La estimación por MV de θ es $24/160=0,15$.

Se puede verificar que la segunda derivada de la función de verosimilitud evaluada en el punto crítico encontrado es menor que cero.

Como en este caso se trata de una distribución discreta, la verosimilitud se puede interpretar como la probabilidad de obtener la muestra obtenida, y el valor estimado de θ es el valor que maximiza esa probabilidad.

CAPÍTULO 9: ESTIMACIÓN POR INTERVALOS

Estimación por intervalos: de la media, de diferencias entre dos medias, de una proporción, de diferencia entre dos proporciones, de varianzas y de la razón entre dos varianzas.

Objetivos:

El alumno debe ser capaz de:

- Deducir conclusiones acerca de los parámetros de la población a partir de la información obtenida en una muestra aleatoria.
- Valorar la distribución normal como una herramienta fundamental para la construcción de intervalos de confianza.
- Construir intervalos de confianza para diferentes parámetros, identificando las condiciones en las cuales pueden ser utilizados.

Resumen

Estimación por intervalos: La estimación por intervalos es un método de estimación que incorpora, a la estimación puntual, información con respecto a su variabilidad y se basa en la búsqueda de intervalos aleatorios que contengan al parámetro, que se intenta estimar, con una probabilidad prefijada. En la construcción de los intervalos aleatorios se tiene en cuenta un estimador o estadístico cuya distribución muestral posea información acerca del parámetro que se quiere estimar, y se escogen los extremos aleatorios $\hat{\theta}_I$ y $\hat{\theta}_S$ de tal manera que contengan ese parámetro con probabilidad $1 - \alpha$, es decir $P(\hat{\theta}_I < \theta < \hat{\theta}_S) = 1 - \alpha$. De esta forma, se espera que, de todas las posibles muestras, el $(1 - \alpha) \times 100\%$ produzca intervalos que contenga al verdadero parámetro. Los intervalos de confianza son los intervalos que se obtienen cuando se calculan los intervalos aleatorios con los datos de una particular muestra. Es en este sentido que se dice que se tiene una confianza del $(1 - \alpha) \times 100\%$ de que un particular intervalo de confianza calculado a partir de una muestra sea uno de los que contenga al verdadero parámetro.

Construcción de Intervalos para la media poblacional de una distribución normal

Siendo la varianza conocida:

$$IC_{100(1-\alpha)\%}(\mu) = [\bar{x} - z_{\alpha/2} \cdot \sigma/\sqrt{n}; \bar{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n}]$$

Siendo la varianza desconocida:

$$IC_{100(1-\alpha)\%}(\mu) = [\bar{x} - t_{\alpha/2}^{(n-1)} \cdot s/\sqrt{n}; \bar{x} + t_{\alpha/2}^{(n-1)} \cdot s/\sqrt{n}]$$

donde $t_{\alpha}^{(n-1)}$ denota el punto de la distribución t con $n-1$ grados de libertad que deja a la derecha un área de probabilidad α .

Obs: La hipótesis de normalidad puede ser relajada, bajo el cumplimiento de otros supuestos. (ANEXO II, sit. 1.2, 2.2 y 2.3)

Obs.: Para conocer los intervalos de confianza de diversos parámetros que se utilizarán en esta disciplina consulte el ANEXO II.

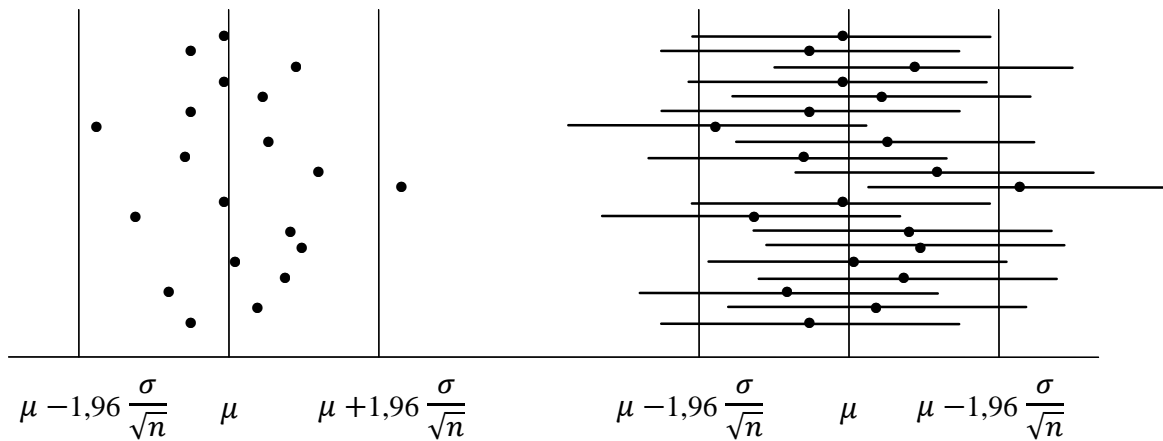


Figura: Intervalo probabilístico e intervalo del 95% de confianza para la media muestral con varianza conocida

En la Figura se presentan los intervalos de confianza para la media poblacional de 20 muestras seleccionadas aleatoriamente de una población normal con varianza conocida, calculados a partir de:

$$IC_{95\%}(\mu) = [\bar{x} - 1,96 \cdot \sigma/\sqrt{n}; \bar{x} + 1,96 \cdot \sigma/\sqrt{n}]$$

Obs: Cada punto en la figura indica un valor observado de \bar{X} .

D1. Amplitud del Intervalo: La amplitud del intervalo de confianza calculado usando la expresión $(\hat{\theta}_{I(Obs)}; \hat{\theta}_{S(Obs)})$, se define como $A = \hat{\theta}_{S(Obs)} - \hat{\theta}_{I(Obs)}$.

Obs: La amplitud de un Intervalo de Confianza, es inversamente proporcional al grado de precisión de la estimación puntual.

T1. El tamaño muestral necesario para estimar la media poblacional de una distribución normal con un error que no supere las L unidades, a un nivel de confianza del $100(1-\alpha)\%$, es el menor número natural que satisface $n \geq (z_{\alpha/2} \sigma / L)^2$.

T2. El tamaño muestral necesario para estimar la proporción poblacional con un error que no supere L , a un nivel de confianza del $100(1-\alpha)\%$, es el menor número natural que satisface $n \geq (z_{\alpha/2}^2 \pi(1 - \pi)) / L^2$.

Obs.: Si se utiliza $\hat{\pi}$ como una estimación de π se puede tener una confianza aproximada de $100(1-\alpha)\%$, de que el error sea menor o igual que L . En caso de no contar con información para estimar π se encuentra un límite superior de n cuando $\pi = 1/2$.

Ejemplos

Ejemplo 1: Un fabricante de papel para impresoras tiene un proceso de producción que opera en forma continua durante todo el turno de producción. Se espera que la hoja de papel tenga una longitud promedio de 11 pulgadas y se sabe que la desviación estándar es de 0,02 pulgadas. De manera periódica se seleccionan muestras para determinar si la longitud promedio de las hojas se mantiene en 11 pulgadas o si algo se ha desajustado en el proceso de producción y produjo un cambio en la longitud promedio del papel. Si esto ocurre, se necesita una acción correctiva. Suponga que se elige una muestra aleatoria de 100 hojas y que la longitud promedio es 10,998 pulgadas.

- Establecer una estimación mediante un intervalo del 95% de confianza para la longitud promedio del papel. ¿Es necesario realizar algún supuesto sobre la distribución de la población para poder determinar el intervalo de confianza?
- De acuerdo con el resultado obtenido, ¿es necesario realizar una acción correctiva al proceso de producción?

Resolución:

- Se construye un intervalo del 95% de confianza para la esperanza (μ) de las longitudes de las hojas. La población de las longitudes de las hojas tiene una distribución desconocida, pero se conoce su varianza $\sigma^2 = 0,02^2 = 0,0004$, además se cuenta con un tamaño muestral grande.

El intervalo de $(1-\alpha)100\%$ de confianza para μ tiene la forma

$$IC_{95\%}(\mu) = \left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] \Rightarrow \left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right].$$

Esta expresión es válida tanto en el caso en que la población de origen es Normal (independientemente del tamaño de la muestra) como cuando la población no es Normal pero se cuenta con un tamaño muestral grande, en virtud del Teorema Central del Límite. Ésta última es la situación en este ejercicio en particular, por lo que no es necesario hacer suposiciones sobre la forma de distribución de probabilidad de la población de las longitudes de las hojas.

En este caso se pide un nivel de confianza del 95%, luego $\alpha = 0,05$ y por lo tanto, $z_{\alpha/2} = z_{0,025} = 1,96$ es el valor z que deja a su derecha el 0,025 de probabilidad bajo la curva de la distribución Normal Estándar. Se conoce además que $\bar{x} = 10,998$. Por lo tanto, el intervalo de 95% de confianza resulta

$$IC_{95\%}(\mu) = \left[10,998 - 1,96 \frac{0,02}{\sqrt{100}} \right] \Rightarrow \left[10,998 - 1,96 \frac{0,02}{\sqrt{100}}, 10,998 + 1,96 \frac{0,02}{\sqrt{100}} \right]$$

$$IC_{95\%}(\mu) = [10,994 \Rightarrow 11,002].$$

La interpretación del IC obtenido es que se tiene una confianza del 95% de que el intervalo $[10,994; \Rightarrow 11,002]$ sea uno de los que contiene el valor del parámetro μ .

- Como el valor 11, que es el valor que indica que el proceso de producción funciona como debe, está incluido en el intervalo, no hay razón para pensar que el proceso necesite un ajuste.

Ejemplo 2: Se realizó un registro del tiempo, que invierten comprando, $n = 64$ clientes elegidos al azar en un supermercado. El promedio y la varianza de los 64 tiempos registrados fueron 33 y 256 minutos, respectivamente.

- Estimar por intervalos, el valor esperado del tiempo que cada cliente invierte comprando en este supermercado. Utilice un nivel de confianza de $1-\alpha = 0,90$. ¿Fue necesario realizar algún supuesto para determinar el intervalo de confianza?
- Interpretar el intervalo de confianza obtenido.

Resolución

- Se construye un intervalo de 90% de confianza para la esperanza μ de la población de tiempos invertidos en compras por los clientes en el supermercado. El intervalo de $(1-\alpha)100\%$ de confianza para μ resulta de la siguiente forma:

$$IC_{90\%}(\mu) = \left[\bar{x} - z_{0,05} \frac{s}{\sqrt{n}} \right] \Rightarrow \left[\bar{x} - z_{0,05} \frac{s}{\sqrt{n}}, \bar{x} + z_{0,05} \frac{s}{\sqrt{n}} \right].$$

puesto que se desconoce la varianza poblacional, pero se cuenta con un tamaño muestral grande. Esta expresión es válida cuando la muestra aleatoria se ha tomado de cualquier población, siempre y cuando tenga varianza finita. En este caso

particular, $\bar{x} = 33, s^2 = 256$ y $\alpha = 0,10$, lo que significa que $z_{0,05} = 1,645$, por lo que el intervalo de confianza resulta

$$IC_{90\%}(\mu) = \left[33 - 1,645 \frac{\sqrt{256}}{\sqrt{64}} \right] \Rightarrow \Rightarrow 33 + 1,645 \frac{\sqrt{256}}{\sqrt{64}},$$

$$IC_{90\%}(\mu) = [29,71; \Rightarrow 36,29].$$

Los supuestos necesarios para la utilización de la fórmula antes enunciada, son los del Teorema Central del Límite, es decir que los tiempos registrados sean provenientes de una muestra aleatoria de una población (tiempo que se invierte en realizar la compra dentro del supermercado) con varianza finita.

- b) Se tiene una confianza del 90% de que el intervalo $[29,71; 36,29]$ es uno de los intervalos que contiene el valor del parámetro μ .

Ejemplo 3: En una fábrica los obreros nuevos requieren alrededor de un mes de capacitación para alcanzar la máxima eficiencia en una operación de ensamblaje. Se sugiere un nuevo método de capacitación y se realiza una prueba para compararlo con el método tradicional. Para este fin se capacitan dos grupos de nueve obreros nuevos durante un período de tres semanas; en uno de los grupos se aplica el método nuevo y en el otro el método tradicional. Al final de las tres semanas de capacitación se midió el tiempo (en minutos) que le toma a cada empleado ensamblar el dispositivo. Las mediciones obtenidas fueron las siguientes:

Procedimiento	Mediciones (minutos)								
Tradicional	32	37	35	28	41	44	35	31	34
Nuevo	35	31	29	25	34	40	27	32	31

- a) Realizar una estimación por intervalo de la diferencia de las esperanzas de los tiempos de ensamblaje entre ambos grupos de obreros, utilizar un nivel de confianza del 95%. Suponer que los tiempos de ensamblaje tienen una distribución aproximadamente Normal y que las varianzas de las dos poblaciones son iguales.
- b) Teniendo en cuenta el IC calculado en (a), ¿es razonable decir que los tiempos medios de ensamblaje para los dos métodos son diferentes?
- c) Construir un intervalo del 90% de confianza para el cociente de varianzas de los tiempos de ensamblaje de los dos grupos. ¿Se justifica el supuesto de igualdad de varianzas entre las poblaciones, realizado en el apartado a)?

Resolución

- a) Sean

μ_1 : Tiempo medio de ensamblaje de obreros capacitados con el método tradicional

μ_2 : Tiempo medio de ensamblaje de obreros capacitados con el método nuevo

La construcción de un intervalo de confianza para $\mu_1 - \mu_2$, para poblaciones normales con varianzas desconocidas pero iguales, se construye usando la expresión

$$IC_{95\%}(\mu_1 - \mu_2) = \left[(\bar{x}_1 - \bar{x}_2) - t_{0,025}^{(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \Rightarrow \Rightarrow \right.$$

$$\left. (\bar{x}_1 - \bar{x}_2) + t_{0,025}^{(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

siendo $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, la estimación de la varianza común a las dos poblaciones.

De los datos se obtiene

$$\bar{y}_1 = 35,22; s_1^2 = 24,445; \bar{y}_2 = 31,56; s_2^2 = 20,027$$

$$s_p^2 = \frac{8 \cdot 24,445 + 8 \cdot 20,027}{16} = 22,236 \Rightarrow s_p = 4,716$$

Considerando $\nu = 9 + 9 - 2 = 16$ y un $\alpha = 0,05$ se tiene $t_{\alpha/2}^{(\nu)} = t_{0,025}^{16} = 2,12$.

El intervalo de confianza es:

$$IC_{95\%}(\mu_1 - \mu_2) = [(35,22 - 31,56) - 2,12 \cdot 4,716\sqrt{1/9 + 1/9} \Rightarrow;$$

$$(35,22 - 31,56) + 2,12 \cdot 4,716\sqrt{1/9 + 1/9}],$$

$$IC_{95\%}(\mu_1 - \mu_2) = [-1,05 \Rightarrow; \Rightarrow 8,37].$$

- b) El valor cero pertenece al intervalo de confianza, indicando que las medias poblacionales pueden ser iguales, por esta razón no se puede afirmar que los tiempos medios de ensamble sean diferentes para obreros entrenados con uno u otro método.
- c) Debido al supuesto de normalidad en los tiempos de ensamble, la forma del intervalo de confianza del 90% de confianza para el cociente de varianzas es:

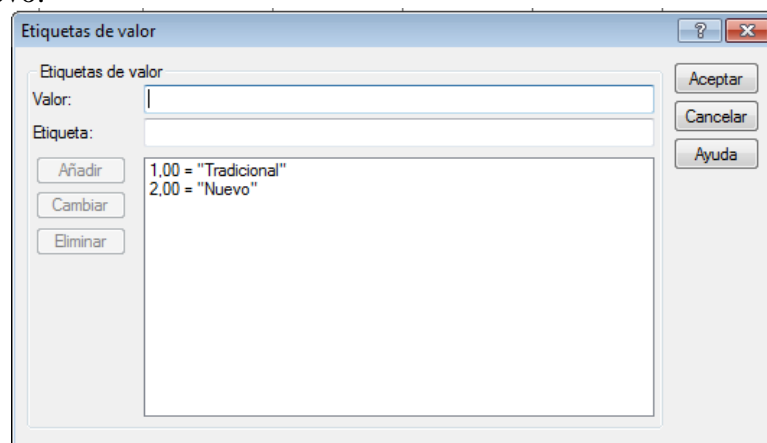
$$IC_{90\%}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left[\frac{s_1^2}{s_2^2} \frac{1}{f_{0.05}(\nu_1, \nu_2)}; \frac{s_1^2}{s_2^2} f_{0.05}(\nu_2, \nu_1)\right],$$

$$IC_{90\%}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = [0,322 \Rightarrow; \Rightarrow 4,626].$$

El valor 1 se encuentra contenido en el intervalo, indicando que las varianzas de los tiempos de ensamble de los dos grupos no son significativamente diferentes, por lo que es razonable suponer varianzas iguales en la construcción del intervalo en el apartado a).

Resolución del apartado a) del Ejemplo 3 con SPSS:

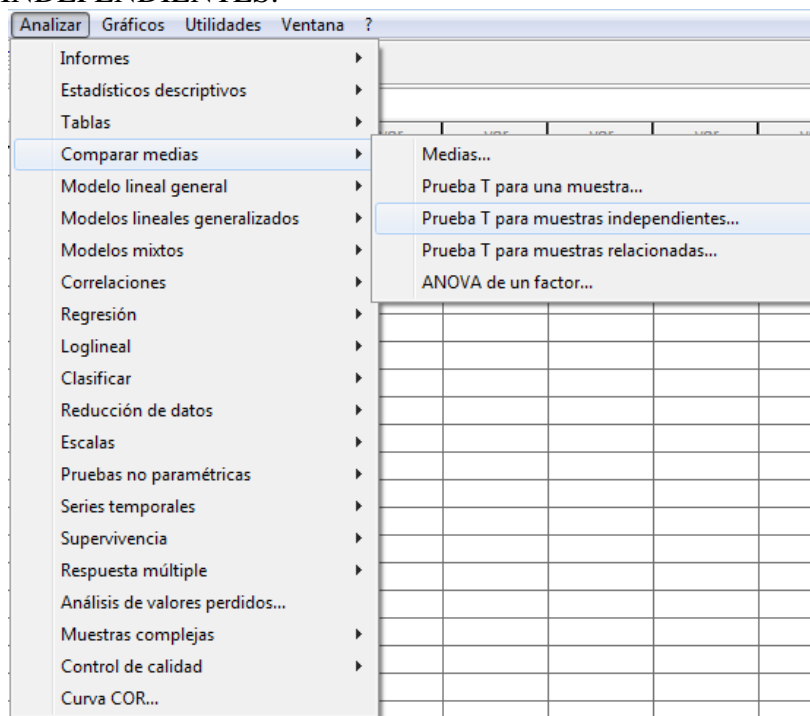
1° Para resolver el punto a) primero hay que crear la variable Procedimiento utilizando la solapa Vista de Variables y en la columna de Nombre se escribe el nombre de la variable y en la columna de valores se otorga el valor 1 al método Tradicional y 2 al método Nuevo:



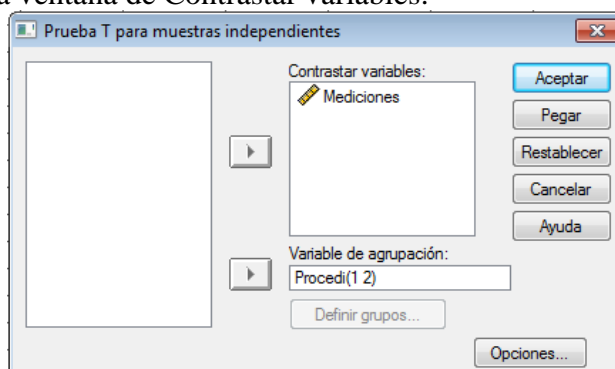
2° En lo que respecta a la variable Mediciones, en la columna Nombre se crea la misma y los valores se importan con la solapa Vista de Datos simplemente escribiéndolos de acuerdo al método que corresponden:

	Procedi	Mediciones
1	1,00	32,00
2	1,00	37,00
3	1,00	35,00
4	1,00	28,00
5	1,00	41,00
6	1,00	44,00
7	1,00	35,00
8	1,00	31,00
9	1,00	34,00
10	2,00	35,00
11	2,00	31,00
12	2,00	29,00
13	2,00	25,00
14	2,00	34,00
15	2,00	40,00
16	2,00	27,00
17	2,00	32,00
18	2,00	31,00

3° Una vez cargados los datos, menú ANALIZAR, COMPARAR MEDIAS, PRUEBA T PARA MUESTRAS INDEPENDIENTES:



4° Aceptar. A continuación se elige como Variable de agrupación a Procedimiento y a Mediciones en la ventana de Contrastar variables.



Luego, se procede a definir los grupos con la opción Definir grupos, el Grupo 1 es el método Tradicional y el Grupo 2 es el Nuevo:

5° Al aceptar se obtienen los resultados:

Estadísticos de grupo

Procedi		N	Media	Desviación típ.	Error típ. de la media
Mediciones	Tradicional	9	35,2222	4,94413	1,64804
	Nuevo	9	31,5556	4,47524	1,49175

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
									Inferior	Superior
Mediciones	Se han asumido varianzas iguales	,061	,807	1,649	16	,119	3,66667	2,22292	-1,04571	8,37904
	No se han asumido varianzas iguales			1,649	15,844	,119	3,66667	2,22292	-1,04949	8,38282

CAPÍTULO 10: PRUEBAS DE HIPÓTESIS I

Hipótesis Estadística. Hipótesis nula y alternativa. Hipótesis simples y compuestas. Prueba de una hipótesis estadística. Pruebas de una y dos colas. Tipos de errores que se pueden cometer al decidir un test y sus probabilidades. Valor P. Relación entre test de hipótesis e intervalo de confianza.

Objetivos:

El alumno debe ser capaz:

- Especificar la hipótesis nula a contrastar en una Prueba de Hipótesis.
- Dada una hipótesis nula identificar el test adecuado para probarla.
- Deducir conclusiones acerca de los parámetros de la población a partir de una muestra aleatoria.
- Conocer la relación entre intervalo de confianza y Prueba de Hipótesis.
- Dada una hipótesis alternativa específica, calcular la potencia de la Prueba.

Resumen

D1. Prueba de Hipótesis Estadística: Es un procedimiento estadístico que permite validar o rechazar una hipótesis estadística a través de los resultados de una muestra.

D2. Hipótesis estadística: Es una afirmación o conjetura acerca de una o más poblaciones.

D3. Hipótesis Nula (H_0): Es la hipótesis cuya aceptación o rechazo se quiere decidir, es decir, es la hipótesis que se quiere testar o poner a prueba.

D4. Hipótesis Alternativa (H_1): Es la hipótesis competidora de H_0 .

D5. Error Tipo I: Rechazar la hipótesis nula cuando es verdadera.

D6. Error Tipo II: No rechazar la hipótesis nula cuando es falsa.

Situaciones posibles al probar una hipótesis estadística

	<u>H_0 es verdadera</u>	<u>H_0 es falsa</u>
<u>no se rechaza H_0</u>	decisión correcta	error tipo II
<u>se rechaza H_0</u>	error tipo I	decisión correcta

Obs: La probabilidad de error Tipo I se denota con α y la del error Tipo II con β .

Lógica en el procedimiento de una prueba de hipótesis estadística

Con la hipótesis nula en mente, se escoge un estadístico (E) cuya distribución puede determinarse cuando la hipótesis nula es verdadera. En esta situación, teniendo en cuenta la distribución del estadístico bajo la hipótesis nula, se determina una región de rechazo (RC) que corresponde a valores posibles del estadístico que ocurren con muy baja probabilidad cuando la hipótesis nula es verdadera, en otras palabras: $P(E \in RC | H_0 \text{ verdadera}) = P(\text{Error tipo I}) = \alpha$, para α pequeño.

Se procede a observar la información relevante en los datos provenientes de una muestra aleatoria, y se obtiene el valor observado del estadístico para la muestra seleccionada.

Si el valor observado del estadístico cae en la región de rechazo, se deduce que la probabilidad de observar éste o un dato más extremo, bajo el supuesto que la hipótesis nula es verdadera, es muy baja. Consecuentemente, con la muestra se ha obtenido un dato que ocurre con baja probabilidad cuando la hipótesis nula es verdadera, refutando la idea de que realmente la hipótesis nula sea verdadera. Entonces se dice que los datos se juzgan como

refutantes de la hipótesis, y en consecuencia es rechazada a favor de la hipótesis alternativa dado que los datos son más probables de haber ocurrido bajo el supuesto de esta alternativa.

D7. Hipótesis Simple: Una hipótesis es simple, cuando caracteriza completamente la distribución del estadístico de prueba (ej. $\theta = \theta_0$).

D8. Hipótesis Compuesta: Una hipótesis es compuesta, cuando no caracteriza completamente la distribución del estadístico de prueba (ej. $\theta \neq \theta_0$ ó $\theta < \theta_0$).

D9. Nivel de significación de una prueba de hipótesis: Cuando la hipótesis nula es una hipótesis simple, el nivel de significación (α) es directamente la probabilidad de error tipo I, $\alpha = P(\text{Error tipo I})$. Cuando la hipótesis nula es una hipótesis compuesta el nivel de significación es la máxima probabilidad de error tipo I, $\alpha = \max_{H_0} P(\text{Error tipo I})$.

Obs. Cuando la hipótesis alternativa es una hipótesis compuesta no es posible calcular “la” probabilidad de error tipo II, $\beta = P(\text{Error tipo II})$, puesto que ésta dependerá de cada particular hipótesis alternativa.

D10. Poder de una prueba de hipótesis: Es la probabilidad de rechazar correctamente la hipótesis nula para una hipótesis alternativa específica: $1 - \beta = 1 - P(\text{Error tipo II}) = P_\theta(RC|\theta \in H_1)$.

Obs: Como el poder se calcula bajo una hipótesis alternativa específica, consecuentemente $Poder(\theta) = P_\theta(RC|\theta \in H_1)$. Por ejemplo, cuando $H_0: \mu \geq \mu_0$ vs. $H_1: \mu < \mu_0$, la probabilidad de rechazar correctamente la hipótesis nula cambia con los diversos $\mu < \mu_0$.

D11. Función de Potencia de una prueba de hipótesis: Es la probabilidad de la región de rechazo. Específicamente

$$\begin{aligned} \text{Potencia: } H_0 \cup H_1 &\rightarrow [0, 1] \\ \theta &\mapsto P_\theta(RC) \end{aligned}$$

Obs: La función potencia resulta

$$P(\text{Región de rechazo}) = \begin{cases} 1 - \beta & H_1 \text{ es verdadera (Decisión correcta)} \\ P(\text{Error Tipo I}) & H_0 \text{ es verdadera (Decisión incorrecta)} \end{cases}$$

Pasos para la construcción de una prueba de hipótesis:

Paso 1: Determinar H_0 y H_1 .

Paso 2: Prefijar α .

Paso 3: Escoger el estadístico de prueba para evaluar H_0 .

Paso 4: Especificar la distribución del estadístico de prueba bajo H_0 y construir la región de rechazo con el α prefijado.

Paso 5: Con la información proveniente de la muestra aleatoria encontrar el valor observado del estadístico.

Paso 6: Concluir que se tiene evidencia para rechazar H_0 si el valor observado del estadístico cae en la región de rechazo y que no se tiene evidencia para rechazar H_0 cuando cae fuera de la región de rechazo.

D12. Prueba de una cola: Una prueba de cualquier hipótesis estadística donde la alternativa es unilateral (por ejemplo $\theta < \theta_0$) recibe el nombre de prueba de una cola.

D13. Prueba de dos colas: Una prueba donde la alternativa es bilateral (por ejemplo $\theta \neq \theta_0$) recibe el nombre de prueba de dos colas.

D14. Valor P: Es el nivel más bajo de significancia en el cual el valor observado del estadístico de prueba es significativo.

Obs: El valor P se calcula a partir de la muestra. Se compara su valor con el nivel de significación α para decidir si debe rechazar la hipótesis nula:

Si el valor P es menor o igual a α : se tiene evidencia para rechazar H_0 .

Si el valor P es mayor a α : no se tiene evidencia para rechazar H_0 .

Pruebas de hipótesis sobre una media poblacional:

Parámetro de Interés: μ Estimador: \bar{X}

Estadístico de prueba: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ ó $\frac{\bar{X}-\mu}{S/\sqrt{n}}$

Hipótesis a Contrastar Región Crítica Supuestos (A.I)

$H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$ $\{\bar{x}: |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$ Sit. 1.1 y 1.2

$\{\bar{x}: |\bar{x} - \mu_0| > t_{\alpha/2}^{(n-1)}s/\sqrt{n}\}$ Sit. 2.1 y 2.2

$\{\bar{x}: |\bar{x} - \mu_0| > z_{\alpha/2}s/\sqrt{n}\}$ Sit. 2.3

$H_0: \mu = \mu_0$ vs. $H_1: \mu < \mu_0$ $\{\bar{x}: \bar{x} < \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}\}$ Sit. 1.1 y 1.2

$\{\bar{x}: \bar{x} < \mu_0 + t_{1-\alpha}^{(n-1)}s/\sqrt{n}\}$ Sit. 2.1 y 2.2

$\{\bar{x}: \bar{x} < \mu_0 + z_{1-\alpha}s/\sqrt{n}\}$ Sit. 2.3

$H_0: \mu = \mu_0$ vs. $H_1: \mu > \mu_0$ $\{\bar{x}: \bar{x} > \mu_0 + z_{\alpha}\sigma/\sqrt{n}\}$ Sit. 1.1 y 1.2

$\{\bar{x}: \bar{x} > \mu_0 + t_{\alpha}^{(n-1)}s/\sqrt{n}\}$ Sit. 2.1 y 2.2

$\{\bar{x}: \bar{x} > \mu_0 + z_{\alpha}s/\sqrt{n}\}$ Sit. 2.3

Selección del tamaño de una muestra

* Para una prueba sobre una media con potencia de $1 - \beta$, un nivel de significación α y una alternativa específica $\mu = \mu_0 + \delta$, cuando se conoce la varianza σ^2

- de una cola: $n = \frac{(z_{\alpha} + z_{\beta})^2 \sigma^2}{\delta^2}$

- de dos colas: $n \cong \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2}$

* La Tabla A.8 (Walpole, pag. 746 9ª Ed., pag 742 4ª Ed) proporciona los tamaños muestrales necesarios para pruebas sobre una media, con potencia de $1 - \beta$, un nivel de significación α y una alternativa específica $\mu = \mu_0 + \delta$, cuando se desconoce la varianza, pero se tiene información acerca de $\Delta = \frac{|\delta|}{\sigma}$.

Ejemplos

Ejemplo 1: La empresa Cumbre envasa café en latas grandes afirmando que su contenido promedio es de 1.500g. La Comisión Federal de Comercio (CFC) lleva a cabo estudios periódicos para comprobar las afirmaciones que hacen los fabricantes acerca de sus productos y desea verificar si la lata grande de café Cumbre cumple con lo afirmado en la etiqueta. De pruebas anteriores se sabe que el contenido de las latas de café presenta una desviación estándar de 100g. La CFC no quiere llevar a cabo una acción que perjudique a la empresa cuando ésta esté cumpliendo con sus especificaciones de peso, ya que podría enfrentar una demanda por parte de la empresa, consecuentemente decide fijar en 0,01 la probabilidad de que esto ocurra. Debido a la imposibilidad de verificar el contenido de todas las latas, la CFC seleccionó una muestra aleatoria de 36 latas de café y pesó su contenido neto, siendo la media de los pesos igual a 1.470 g. Teniendo en cuenta la información de la muestra seleccionada ¿la CFC puede afirmar que la empresa no está cumpliendo con las especificaciones sobre el contenido?

Resolución:

Sean las hipótesis: $H_0: \mu = 1.500$ vs. $H_1: \mu < 1.500$.

En este caso se opta por realizar una prueba de una cola, puesto que contenidos por debajo de lo especificado serían los problemáticos.

La CFC prefijó en 0,01 la probabilidad de rechazar que la empresa cumple con las especificaciones cuando verdaderamente las cumple, es decir que

$$P(\text{rechazar } H_0 | H_0 \text{ es verdadera}) = 0,01 = \alpha.$$

Si bien cada vez que se realiza una prueba de hipótesis es posible cometer dos tipos de errores, el error tipo I es habitualmente el más costoso, y por ello se fija en un valor pequeño la probabilidad de cometerlo. En este caso, si se comete el error de tipo I, la CFC se puede enfrentar una demanda por parte de la empresa si la sanciona injustamente.

Aunque no se conoce la distribución de la que provienen los datos, dado que el tamaño muestral es grande, y $\sigma = 100 < \infty$ se puede asumir por el Teorema Central del Límite que la distribución del estadístico $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ es aproximadamente normal

estándar. El estadístico de la prueba de hipótesis en este caso será $Z = \frac{\bar{X} - 1.500}{100/\sqrt{36}}$, que sigue una distribución aproximadamente Normal Estándar bajo H_0 .

Por tratarse de una prueba de una cola, el valor crítico bajo la distribución normal estándar que deja en la cola izquierda una probabilidad 0,01 es $z_{1-\alpha} = z_{0,99} = -2,33 = -z_\alpha$. Con este valor crítico se determina la región crítica o de rechazo de H_0 ,

$$\text{Región crítica: } z < -2,33.$$

Al calcular el valor del estadístico para la muestra seleccionada se obtiene

$$z_{obs} = \frac{1.470 - 1.500}{100/\sqrt{36}} = -1,8.$$

Finalmente, se compara el valor obtenido del estadístico del test con el valor crítico, y como $z_{obs} = -1,8 > -2,33 = z_{0,01}$, no hay razones suficientes para rechazar la hipótesis nula H_0 . Consecuentemente, no se tiene evidencia suficiente para rechazar que el contenido medio de las latas de café sea 1.500g.

Ejemplo 2: En referencia al Ejemplo 1, suponga que a la CFC le preocupa que los clientes estén recibiendo en promedio una cantidad menor a la que afirma el fabricante y que esto no sea detectado en sus estudios. Considera inadmisibles que reciban en promedio 1.470g o menos y que esto no sea captado por sus analistas.

- ¿Cuál es la probabilidad que esto ocurra?
- ¿Cuál es la potencia de la prueba de hipótesis realizada en el Ejemplo 1? Explicar su significado.

Resolución:

- Se pretende calcular cuál es la probabilidad de que el peso medio de los paquetes que está envasando la empresa sea 1.470g y la CFC no lo esté detectando. Es decir que se quiere calcular la probabilidad que en realidad el contenido medio de las latas de café sea de 1.470g y al realizar la prueba no se rechace que el contenido medio de las latas de café sea 1.500g, por lo tanto se quiere calcular la probabilidad de cometer un error tipo II, ante una alternativa $\mu = 1.470$.

$$\beta = P[\text{error tipo II}] = P[\text{aceptar } H_0 | \mu = 1.470] \quad (1)$$

En el Ejemplo 1 se calculó el z crítico $z_{1-\alpha} = -2,33$ que separa la zona de aceptación de la zona de rechazo bajo la hipótesis nula. En términos de los valores de \bar{X} , ese z crítico se corresponde con $\bar{x}_c = 1.500 - 2,33 \frac{100}{6} = 1.461,16 \cong 1.461$; consecuentemente, se rechaza H_0 cuando $\bar{x} \leq 1.461$ y no se rechaza en caso contrario. Luego β se calcula:

$$\beta = P[\bar{X} > 1.461 | \mu = 1.470]$$

$$= P \left[\frac{(\bar{X} - 1.470)}{100/6} > \frac{(1.461 - 1.470)}{100/6} \right]$$

$$\cong P[Z > -0.54] = 0.7054.$$

El cálculo de la probabilidad del error tipo II se calcula bajo H_1 por lo que hay que estandarizar considerando que la verdadera media es 1.470, y no 1.500.

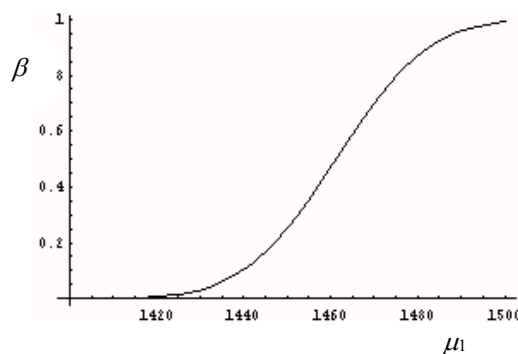
La interpretación del resultado obtenido es que si el peso medio de las latas de café fuese 1.470g, en aproximadamente el 70% de las veces la CFC no rechazará la hipótesis que la media es 1.500g. Con esto se observa que, aunque se prefijó en un valor pequeño para la probabilidad de cometer el error tipo I, la probabilidad de cometer el error tipo II es grande. Generalmente, a medida que se hace más pequeño uno de los errores, el otro crece. Se observa que aumentando el tamaño de la muestra es posible disminuir las probabilidades de cometer ambos errores.

También resulta de interés averiguar cuál es la probabilidad de que la empresa esté envasando una cantidad μ_1 , menor que 1.470g y la CFC no lo detecte. En este caso

$$\beta = P[\text{error tipo II}] = P[\text{aceptar } H_0 \mid \mu = \mu_1]$$

$$\beta = P[\bar{X} > 1461 \mid \mu = \mu_1] = 1 - \Phi \left[\frac{1461 - \mu_1}{100/6} \right]$$

En la representación gráfica de esta función $\beta(\mu_1)$, se observa que a medida que el valor de μ_1 se aproxima a 1.500, aumenta la probabilidad de cometer error tipo II.

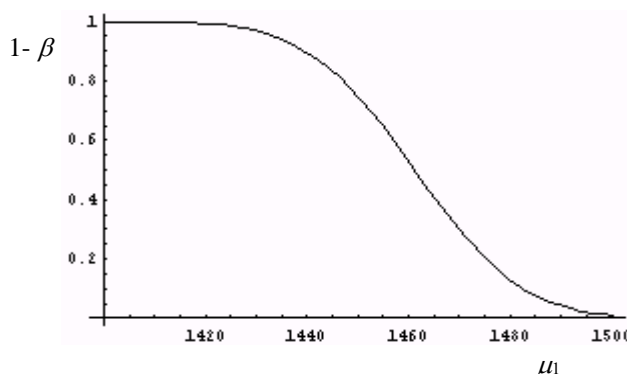


En este ejemplo, a la CFC le preocupa no poder advertir que $\mu \leq 1.470$. Como el valor más grande de β ocurre cuando $\mu = 1.500$ (porque la función es creciente), para los valores de $\mu \leq 1.470$, se tendrá que $\beta \leq 0.7019$.

b) Se define la potencia de una prueba de hipótesis como la probabilidad de rechazar H_0 cuando es falsa. Por lo tanto, la *Potencia* = $1 - \beta$ tendrá un valor distinto para cada valor diferente de la hipótesis alternativa.

En el caso en que $\mu_1 = 1.470$, la potencia es *Potencia* = $1 - 0.7019 = 0.2981$. Esto significa que aproximadamente el 30% de las veces que los datos provienen de una población con $\mu_1 = 1.470$, la prueba es capaz de captar que los datos provienen de una población con media menor que 1.500.

También se puede calcular la potencia para cada μ_1 particular, usando la expresión calculada en el ejercicio anterior.



$$\text{Potencia}(\mu_1) = 1 - \beta(\mu_1).$$

En el gráfico de esta función *Potencia* se observa que a medida que el valor de μ_1 se aproxima a 1.500, el test pierde capacidad de discriminar que los datos provienen de una población con media menor.

Ejemplo 3:

- En el mismo ejemplo precedente, ¿la CFC debería multar a la empresa Cumbre por no cumplir con las especificaciones del envase si la media de una muestra de tamaño 36 es de 1.460 g?
- Supóngase que el fabricante de café Cumbre luego de ser multado decide ajustar las máquinas que envasan para que su producto cumpla con las especificaciones del envase, pero sin envasar de más porque esto le ocasionaría pérdidas. Luego toma una muestra de tamaño 36 obteniendo una media de 1.540 g. ¿Puede estar conforme con el ajuste que le hizo a la máquina al 1% de significación?
- Por su parte, un grupo de consumidores de café Cumbre sabe que el productor ha incrementado el contenido de la lata y están satisfechos porque creen que ahora la lata tiene un contenido mayor al especificado en la etiqueta. Cuentan con una muestra semejante en tamaño y un valor de media muestral igual a la obtenida por la empresa. ¿A qué conclusión llegan los consumidores al 1% de significación?
- ¿Son contradictorios los resultados de b) y c)? Justifique.

Resolución:

- a) El razonamiento es semejante al del Ejemplo 1. Siendo las hipótesis a contrastar:

$$H_0: \mu = 1.500 \text{ vs. } H_1: \mu < 1.500.$$

Se puede considerar que en este caso la probabilidad del error tipo I que se está dispuesto a tolerar es

$$P(\text{rechazar } H_0 | H_0 \text{ es verdadera}) = 0,01 = \alpha.$$

Aunque no se conoce la distribución de la que provienen los datos, dado que el tamaño muestral es grande y $\sigma = 100 < \infty$, se puede asumir, por el Teorema

Central del Límite, que la distribución del estadístico $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ es aproximadamente

normal estándar. Consecuentemente el estadístico de prueba $Z = \frac{\bar{X} - 1.500}{100/\sqrt{36}}$ sigue una

distribución aproximadamente Normal Estándar bajo H_0 . Se determina que el valor crítico bajo la distribución normal estándar es $z_{1-\alpha} = -2,33$, y consecuentemente la región crítica o de rechazo de H_0 queda determinada por:

$$\text{Región crítica: } z < -2,33.$$

Calculando el valor del estadístico observado para la particular muestra, se tiene que

$$z_{obs} = \frac{1.460 - 1.500}{100/\sqrt{36}} = -2,4.$$

Finalmente, comparando el valor observado del estadístico del test con el valor crítico, como $z_{obs} = -2,4 < -2,33 = z_{0,01}$, se rechaza H_0 . Consecuentemente, a un nivel de significación del 1%, la CFC rechaza la afirmación de la empresa Cumbre de que el contenido medio de sus latas de café sea de 1.500g y debería multar a la empresa porque su producto no cumple con las especificaciones del envase.

- b) En este caso las hipótesis a contrastar son:

$$H_0: \mu = 1.500 \text{ vs. } H_1: \mu \neq 1.500.$$

El estadístico de prueba coincide con el empleado en el apartado a). Por lo tanto, considerando $P(\text{rechazar } H_0 | H_0 \text{ es verdadera}) = 0,01 = \alpha$, se determina el valor crítico bajo la distribución normal estándar, como el valor z que deja en ambas colas de la distribución una probabilidad 0,005. Este valor es $z_{\alpha/2} = z_{0,005} = 2,575$. Con este valor crítico, se determina la siguiente región crítica o de rechazo de H_0

$$\text{Región crítica: } z < -2,575 \text{ ó } z > 2,575.$$

Calculando el valor observado del estadístico de prueba se obtiene

$$z_{obs} = \frac{1.540 - 1.500}{100/\sqrt{36}} = 2,4.$$

Finalmente, comparando el valor observado del estadístico de prueba con el valor crítico, se tiene que $-2,575 < 2,4 < 2,575$, y consecuentemente no hay evidencia suficiente para rechazar H_0 . El productor está satisfecho con el ajuste de las máquinas porque su producto cumple con las especificaciones del envase y no tiene pérdidas por envasar más cantidad de la estipulada.

c) En este caso las hipótesis a contrastar son:

$$H_0: \mu = 1.500 \text{ vs. } H_1: \mu > 1.500.$$

El estadístico de prueba para este caso coincide con el empleado en el apartado a). Por lo tanto, considerando $P(\text{rechazar } H_0 | H_0 \text{ es verdadera}) = 0,01 = \alpha$, se determina el valor crítico bajo la distribución normal estándar, como el valor z que deja a su derecha una probabilidad 0,01. Este valor es $z_\alpha = z_{0,01} = 2,33$, y la región crítica o de rechazo resulta

$$\text{Región crítica: } z > 2,33.$$

Calculando el valor observado del estadístico de prueba se obtiene

$$z_{obs} = \frac{1.540 - 1.500}{100/\sqrt{36}} = 2,4.$$

Finalmente, comparando el valor obtenido del estadístico de prueba con el valor crítico, $2,4 > 2,33$, se rechaza H_0 . Los consumidores deberían estar satisfechos de que ahora el envase tiene un contenido mayor al especificado en el envase.

d) Se observa que con el mismo tamaño de muestra y la misma media muestral se pueden obtener conclusiones diferentes. Por un lado, el fabricante está convencido de que está envasando 1.500 g, pero los consumidores creen que la lata contiene en promedio más de 1.500 g. ¿Cuál es el resultado más confiable?

Si se repite el procedimiento llevado a cabo en el Ejemplo 2, y se calcula la potencia de una prueba de dos colas y otro de una cola a derecha, se llega a los siguientes resultados:

- Prueba Unilateral, cola derecha:

$$\text{Potencia}(\mu_1) = 1 - \beta(\mu_1) = 1 - \Phi\left(2,33 + (1.500 - \mu_1)\frac{6}{100}\right)$$

- Prueba Bilateral:

$$1 - \beta(\mu_1) = 1 - \Phi\left(2,575 + (1.500 - \mu_1)\frac{6}{100}\right) + \Phi\left(-2,575 + (1.500 - \mu_1)\frac{6}{100}\right)$$

Si se representa en un mismo gráfico ambas Potencias, se obtiene lo siguiente:

La función Potencia de la prueba unilateral se encuentra por arriba de la función Potencia de la prueba bilateral. Por lo tanto, para valores de $\mu_1 > 1.500$, la prueba unilateral tiene más capacidad para discriminar entre H_0 y H_1 . Por eso se llega a distintos resultados. Lo que sucede en este caso es que las máquinas están envasando en promedio más de 1.500 g, lo que es detectado por la prueba unilateral de cola derecha, pero no por la prueba bilateral que tiene menor potencia.

CAPÍTULO 11: PRUEBAS DE HIPÓTESIS II

Cálculo de tamaños muestrales. Pruebas de hipótesis sobre medias de dos poblaciones, proporciones de una o dos poblaciones y varianzas de una o dos poblaciones

Objetivos:

El alumno debe ser capaz:

- Conocer la relación entre intervalo de confianza y prueba de hipótesis.
- Calcular el tamaño de muestra necesario para que una prueba de hipótesis tenga la potencia deseada cuando se quiere poner a prueba una hipótesis nula.
- Contrastar hipótesis sobre proporciones, para muestras pequeñas y grandes.
- Contrastar hipótesis sobre varianzas.

Resumen

Relación entre la estimación por intervalos y la prueba de hipótesis

La prueba de $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$ a un nivel de significación α es equivalente a calcular un intervalo de confianza del $100(1 - \alpha)\%$ de μ , y rechazar H_0 si μ_0 está fuera del IC.

Pruebas de hipótesis sobre diversos parámetros:

Parámetro de Interés	Estimador	Estadístico de prueba
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ con } \sigma_1 = \sigma_2$ $\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ con } \sigma_1 \neq \sigma_2$
π	$\hat{\pi} = \frac{\text{nº de éxitos}}{n}$	$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$
$\pi_1 - \pi_2$	$\hat{\pi}_1 - \hat{\pi}_2$	$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$
σ^2	S^2	$\frac{(n-1)S^2}{\sigma^2}$
σ_1^2 / σ_2^2	S_1^2 / S_2^2	$\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$

Selección del tamaño de una muestra

* Para una prueba de una cola sobre dos medias con potencia de $1 - \beta$, un nivel de significación α y una alternativa específica $\mu_1 - \mu_2 = d_0 + \delta$, cuando se conocen las varianzas σ_1^2 y σ_2^2 .

$$n = n_1 = n_2 = \frac{(z_\alpha + z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{\delta^2}$$

* La Tabla A.9 (Walpole, pg. 742 y 743) proporciona los tamaños muestrales necesarios para pruebas sobre una media o dos medias respectivamente, con potencia de $1 - \beta$, un nivel de significación α y una alternativa específica $\mu_1 - \mu_2 = d_0 + \delta$, cuando se desconoce la varianza pero se tiene información acerca de $\Delta = \frac{|\delta|}{\sigma}$.

Ejemplos

Ejemplo 1: El gerente de personal de una fábrica está pensando en cambiar el horario de trabajo en la misma. Para asegurarse que los empleados en su mayoría están a favor del cambio de horario y puesto que no puede entrevistar a los 500 empleados en un tiempo razonable decide tomar una muestra.

- Plantear la hipótesis nula y alternativa de una prueba a realizar que le ayude al gerente de personal a tomar la decisión.
- Proponer un estadístico de prueba e indicar su distribución bajo la hipótesis nula.
- Indicar cuál sería el resultado de una prueba de hipótesis que contraste las hipótesis planteadas en a), si 72 de 120 empleados seleccionados aleatoriamente están a favor del cambio de horario, a un nivel de significación del 1%.
- Explicar, en el contexto de esta aplicación, en qué consiste el error tipo I y el error tipo II.
- Comparar el estadístico propuesto en b) con el que se usa cuando se calcula un intervalo de confianza para una proporción.

Resolución:

- El gerente realizará el cambio de horario si la mayoría está a su favor. Para seleccionar las hipótesis nula y alternativa, se toma como nula aquella que denota ausencia de cambio. En este caso, denotando con π a la proporción de empleados a favor del cambio de horario, las hipótesis a contrastar serían:

$$H_0: \pi \leq 0,5 \text{ vs. } H_1: \pi > 0,5.$$

- El estadístico de prueba para contrastar una hipótesis sobre la proporción poblacional es

$$\frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}}.$$

Este estadístico tiene distribución aproximadamente normal cuando n es suficientemente grande, $\hat{\pi}$ es la proporción muestral (número de éxitos sobre el tamaño muestral) y π es la proporción de éxitos en la población.

Bajo la hipótesis nula, $H_0: \pi \leq 0,5$, el estadístico de prueba que se usa es $\frac{\hat{\pi} - 0,5}{\sqrt{0,5(1-0,5)/n}}$, que se sabe tendrá distribución aproximadamente normal cuando n es suficientemente grande, y $\pi = 0,5$.

En este caso se resuelve el problema para la hipótesis simple $\pi = 0,5$, puesto que la hipótesis nula es una hipótesis compuesta. También, es conveniente recordar en este caso que, la probabilidad del error tipo I será a lo sumo el valor del α prefijado.

- c) Por tratarse de una prueba de una cola, la región de rechazo será

$$RC = \{z: z > z_{\alpha}\} = \{z: z > 2,327\}$$

Calculando el valor observado del estadístico de prueba,

$$z_{obs} = \frac{72/120 - 0,5}{\sqrt{0,5(1-0,5)/120}} \cong \frac{0,6 - 0,5}{0,5/10,95} \cong 2,19,$$

se obtiene un valor fuera de la región de rechazo, y consecuentemente no se puede rechazar la hipótesis nula y se concluye que: A un nivel de significación del 1% no se tiene evidencia suficiente para afirmar que la proporción de empleados a favor del cambio de horario sea superior al 50%.

Con lo cual el gerente no tomaría la decisión de realizar el cambio de horario.

- d) El error tipo I por definición es rechazar H_0 cuando es verdadera. En el contexto de este ejercicio, cometer un error tipo I sería afirmar que la mayoría está a favor del cambio de horario cuando en realidad no es así.

El error tipo II, por definición, es no rechazar H_0 cuando es falsa. En el contexto de este ejercicio, cometer un error tipo II sería afirmar que la mayoría no está a favor del cambio de horario cuando en realidad si lo está.

- e) El estadístico de prueba para contrastar una hipótesis sobre la proporción poblacional es

$$\frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}}, \quad (1)$$

mientras que el estadístico usado en la construcción de un Intervalo de Confianza es

$$\frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}, \quad (2)$$

puesto que la amplitud del intervalo no puede depender del parámetro que se quiere estimar. De esta manera se resalta que el estadístico (2) se puede obtener de (1) reemplazando el error estándar de la proporción muestral por su estimador.

Con esto debería quedar claro que si se realiza una prueba de dos colas, la conclusión no tiene que coincidir necesariamente con la que se tome a partir del IC, puesto que los estadísticos sobre los que se basan las inferencias son diferentes.

Ejemplo 2: Se realiza una investigación sobre las causas de los accidentes automovilísticos, características de los conductores, lugares, etc. Con este objetivo, se seleccionó una muestra aleatoria de 50 personas solteras y se observó que 20 de ellos tuvieron al menos 1 accidente de tránsito en 3 años. Del mismo modo se seleccionó una muestra de 80 personas casadas y 19 de ellas tuvieron al menos un accidente de tránsito en el mismo período de tiempo.

- Plantear la hipótesis nula y alternativa de una prueba de hipótesis que ayude a decidir la proporción de personas que tuvieron al menos 1 accidente es la misma o no entre solteros y casados.
- Realizar una prueba de hipótesis que le permita contrastar las hipótesis planteadas en a), usando un nivel de significación del 5%.
- ¿Se puede evaluar si la proporción de personas que tuvieron al menos 1 accidente es la misma o no entre solteros y casados utilizando un Intervalo de Confianza? En caso afirmativo, indicar la conclusión usando un nivel de confianza del 95% y comparar con lo obtenido en b).

Resolución:

- a) Se quiere evaluar si la proporción de personas que tuvieron al menos 1 accidente en los últimos tres años es la misma o no entre solteros y casados. Se denota con π_1 la proporción de personas con al menos un accidente en los últimos tres años entre las personas solteras, y con π_2 a la misma proporción entre las casadas.

Las hipótesis a contrastar son:

$$H_0: \pi_1 = \pi_2 \text{ vs. } H_1: \pi_1 \neq \pi_2.$$

- b) Un estadístico direccionado a la diferencia de proporciones es

$$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2}}.$$

Este estadístico tiene distribución aproximadamente normal cuando n_1 y n_2 son suficientemente grandes, $\hat{\pi}_1$ y $\hat{\pi}_2$ son las proporciones muestrales de dos muestras independientes y π_1 y π_2 son las proporciones de éxitos en las poblaciones de las cuales se tomaron las muestras.

Bajo la hipótesis nula, $H_0: \pi_1 = \pi_2$, las muestras pueden ser consideradas como provenientes de una misma población y consecuentemente, el error estándar de $\hat{\pi}_1 - \hat{\pi}_2$ resulta $\sqrt{\pi^*(1-\pi^*)(1/n_1 + 1/n_2)}$, donde $\pi^* = \pi_1 = \pi_2$.

Bajo la hipótesis nula, el valor de π^* todavía queda indeterminado, por lo tanto para desarrollar la prueba de hipótesis el estadístico que se usa para encontrar la región de rechazo es

$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}^*(1-\hat{\pi}^*)(1/n_1 + 1/n_2)}},$$

donde $\hat{\pi}^* = (n_1\hat{\pi}_1 + n_2\hat{\pi}_2)/(n_1 + n_2)$, y su distribución asintótica es la normal estándar. El valor observado de este estadístico es

$$\begin{aligned} z_{obs} &= \frac{20/50 - 19/80}{\sqrt{39/130(1 - 39/130)(1/50 + 1/80)}} \\ &\cong \frac{0,40 - 0,2375}{\sqrt{0,3(1 - 0,7)0,0325}} \cong 1,967. \end{aligned}$$

Por tratarse de una prueba de dos colas, la región de rechazo será

$$RC = \{z: |z| > z_{\alpha/2}\} = \{z: |z| > 1,96\},$$

y se procede a rechazar la hipótesis nula, concluyéndose que: A un nivel de significación del 5% se tiene evidencia suficiente para afirmar que la proporción de personas que tuvieron al menos 1 accidente en los últimos 3 años difiere significativamente entre solteros y casados.

- c) También se puede evaluar si la proporción de personas que tuvieron al menos 1 accidente es la misma o no entre solteros y casados utilizando un Intervalo de Confianza, el estadístico usado para su construcción es

$$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2}}$$

ya que no se puede suponer que $\pi_1 = \pi_2$; y la expresión del IC resulta

$$\begin{aligned} &(\hat{\pi}_{1obs} - \hat{\pi}_{2obs} - z_{\alpha/2}\sqrt{\hat{\pi}_{1obs}(1-\hat{\pi}_{1obs})/n_1 + \hat{\pi}_{2obs}(1-\hat{\pi}_{2obs})/n_2}; \\ &\hat{\pi}_{1obs} - \hat{\pi}_{2obs} + z_{\alpha/2}\sqrt{\hat{\pi}_{1obs}(1-\hat{\pi}_{1obs})/n_1 + \hat{\pi}_{2obs}(1-\hat{\pi}_{2obs})/n_2}), \end{aligned}$$

teniendo en cuenta la información obtenida en ambas muestras,

$$IC_{95\%}(\pi_1 - \pi_2) = (-0,0022; 0,327).$$

Evalutando si la proporción de personas que tuvieron al menos 1 accidente es la misma o no entre solteros y casados a partir del IC, al observar que el cero está incluido en el intervalo, se concluye que no hay evidencia para afirmar que sean diferentes.

Esta conclusión difiere de la dada en (b) puesto que la inferencia se basa en un estadístico diferente (el estadístico en (b) usa el supuesto de igualdad de proporciones y éste no).

Ejemplo 3:

Un gerente de control de calidad de una compañía estaba convencido de que la balanza que usa para pesar materia prima está bien calibrada y es bastante precisa, siendo que su error de medición puede ser considerado una variable aleatoria con media 0 y desvío estándar 2. A fin de probar el equipo se pesó una carga de 107 kg y se registraron las mediciones 104,1; 105,2 y 110,2.

- Mediante el uso de una prueba de hipótesis evaluar si es verdadera la afirmación del gerente acerca de la precisión del error de medición. Usar un nivel de significación del 5%, y especificar claramente los supuestos utilizados.
- Calcular la potencia de la prueba de hipótesis realizada en a) cuando en realidad el desvío estándar es 5,076.

Resolución:

- La variable sobre la que se tiene interés es el error de medición, la diferencia entre el valor observado y la carga real (107kg). Consecuentemente, los errores observados son: 2,9; 1,8 y 3,2. Se supone que estos valores son los valores observados de una muestra aleatoria de tamaño 3 (E_1, E_2, E_3), de una variable aleatoria normal con media 0 y desvío estándar 2 cuando la balanza está bien calibrada. La distribución normal es clásicamente utilizada para modelar los errores de medición. Entonces, el estadístico $\frac{1}{2^2} \sum_{i=1}^3 E_i^2$ tiene distribución χ^2 con 2 grados de libertad.

Por tratarse de un error de medición, y como se quiere evaluar la afirmación del gerente acerca de la precisión del error, resulta de interés contrastar si el error estándar excede o no el valor especificado por él, es decir

$$H_0: \sigma \leq 2 \text{ vs. } H_1: \sigma > 2.$$

El nivel de significación que se usará para contrastar estas hipótesis es del 5%.

El estadístico de prueba en este caso es

$$\frac{1}{\sigma^2} \sum_{i=1}^3 E_i^2 \sim \chi_{(3)}^2, \text{ siempre que } E_1, E_2, E_3 \text{ i. i. } dN(0, \sigma^2).$$

Consecuentemente, bajo la hipótesis nula $\frac{1}{2^2} \sum_{i=1}^3 E_i^2 \sim \chi_{(3)}^2$, y la región de rechazo será $\{c/c > \chi_{(2)}^2(\alpha)\} = \{c/c > 5,992\}$.

El valor observado del estadístico es $\frac{1}{2^2} \sum_{i=1}^3 e_i^2 = \frac{1}{2^2} 21,89 = 5,4725$, cayendo fuera de la región de rechazo, lo que lleva a concluir que no se tiene evidencia suficiente para afirmar que el desvío estándar del error de medición sea superior a 2 a un nivel de significación del 5%.

- La potencia de una prueba es la probabilidad de rechazar correctamente la hipótesis nula. En el contexto de la prueba de hipótesis planteada en a) sería:

$$\begin{aligned} P_{\sigma=5,076} \left(\frac{1}{2^2} \sum_{i=1}^3 E_i^2 > 5,995 \right) &= P \left(\frac{1}{5,076^2} \sum_{i=1}^3 E_i^2 > \frac{5,992 \times 4}{5,076^2} \right) \\ &= P(\chi_{(3)}^2 > 0,9302) \\ &= 0,6281. \end{aligned}$$

CAPÍTULO 12: Pruebas χ^2

Prueba de bondad del ajuste. Prueba de homogeneidad, independencia e igualdad de varias proporciones.

Objetivos:

El alumno debe ser capaz de:

- Identificar las situaciones donde es conveniente aplicar cada una de las pruebas presentadas.
- Enunciar la hipótesis nula asociada a cada una de las pruebas.
- Verificar las condiciones para la aplicación de cada prueba.
- Calcular el valor del estadístico χ^2 e interpretar los resultados obtenidos.

Resumen

D1. Prueba de Bondad de Ajuste: Es un procedimiento estadístico que permite validar o rechazar la hipótesis de que alguna distribución muestral sigue una distribución teórica, es decir

$H_0: X$ tiene una distribución determinada vs. $H_1: X$ no tiene esa distribución.

Obs: Una prueba de bondad de ajuste permite evaluar si un modelo probabilístico es adecuado o no para describir la población de donde proviene la muestra.

D2. Distribución Muestral del Estadístico de Prueba en una Prueba de Bondad de Ajuste: Sea X_1, X_2, \dots, X_n i.i.d. como X . Considere los valores del recorrido de X divididos en k categorías mutuamente excluyentes. Si se define la variable aleatoria O_i como el número de observaciones de la muestra que pertenecen a la categoría i ($i=1, \dots, k$), entonces el estadístico

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

sigue una distribución χ^2 con $\nu = k - 1$ grados de libertad, cuando n tiende a infinito. Siendo e_i la frecuencia esperada en la categoría i bajo la distribución de X , es decir

$P(X \in \text{Categoría}_i) \times n$.

Obs 1: Para el cálculo de las frecuencias esperadas es necesario el conocimiento completo de la distribución teórica, tanto su forma cuanto sus parámetros.

Obs 2: Cuando una variable sigue una determinada distribución, pero se desconoce uno o más de sus parámetros, las frecuencias esperadas pueden ser calculadas usando como parámetros de la distribución teórica estimaciones obtenidas a partir de la información de la muestra aleatoria. En esta situación el estadístico de prueba definido en **D2** todavía tiene distribución χ^2 , pero los grados de libertad disminuyen tantas unidades como parámetros fueron estimados, es decir $\nu = k - 1 - c$ con c representando el número de parámetros que fueron estimados.

D3. Prueba de Independencia: Es un procedimiento estadístico que permite validar o rechazar la hipótesis de independencia entre dos variables categóricas.

$H_0: X$ e Y son independientes vs. $H_1: X$ e Y no son independientes.

Obs: La independencia entre dos variables categóricas, por ejemplo, X tal que $\text{Rec}(X) = \{x_1, \dots, x_a\}$ e Y tal que $\text{Rec}(Y) = \{y_1, \dots, y_b\}$, se satisface si y sólo si

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j), \quad \text{para } i=1, \dots, a; j=1, \dots, b.$$

Obs: Como a priori no se tiene información sobre la distribución de probabilidad conjunta, dada la muestra de tamaño n , $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, de (X, Y) , las frecuencias observadas serán

$X \backslash Y$	y_1	y_2	\dots	y_b	<i>Total</i>
x_1	n_{11}	n_{12}	\dots	n_{1b}	n_{1+}
x_2	n_{21}	n_{22}	\dots	n_{2b}	n_{2+}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_a	n_{a1}	n_{a2}	\dots	n_{ab}	n_{a+}
<i>Total</i>	n_{+1}	n_{+2}	\dots	n_{+b}	$n_{++} = n$

donde n_{ij} denota el número de observaciones de la muestra donde el valor de X observado es x_i y el valor de Y observado es y_j . La frecuencia esperada bajo independencia sería

$$e_{ij} = \frac{n_{i+} \times n_{+j}}{n_{++}},$$

donde $n_{i+} = \sum_{j=1}^b n_{ij}$, $n_{+j} = \sum_{i=1}^a n_{ij}$ y $n_{++} = \sum_{i=1}^a \sum_{j=1}^b n_{ij}$, que coincide con el tamaño muestral.

D4. Distribución Muestral del Estadístico de Prueba en una Prueba de Independencia: Sea $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ i.i.d. como (X, Y) . Cada elemento de la muestra resultará en algún (x_i, y_j) , con $x_i \in \text{Rec}(X)$ e $y_j \in \text{Rec}(Y)$. Si se define la variable aleatoria O_{ij} como el número de observaciones de la muestra donde el valor de X observado fue x_i y el valor de Y observado fue y_j , entonces, cuando X e Y son independientes, el estadístico

$$\sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

sigue una distribución χ^2 con $\nu = (a - 1) \times (b - 1)$ grados de libertad cuando n tiende a infinito, donde $e_{ij} = \frac{n_{i+} \times n_{+j}}{n_{++}}$.

D5. Prueba de Homogeneidad: Es un procedimiento estadístico que permite validar o rechazar la hipótesis de que una variable aleatoria categórica tiene la misma distribución (es homogénea) en varias subpoblaciones.

H_0 : la dn. de X es la misma para todas las subpoblaciones
vs. H_1 : la dn. de X cambia en al menos una subpoblación.

Obs: Sea X una v.a. con $\text{Rec}(X) = \{x_1, \dots, x_a\}$, y sea $P_k(X = x_i)$ la probabilidad de que X asuma el valor x_i en la subpoblación k , para $i=1 \dots a$ y $k=1 \dots K$. Se dice que la distribución de X es homogénea en las K subpoblaciones cuando

$$P_k(X = x_i) = P(X = x_i), \text{ para } k=1, \dots, K,$$

es decir que las probabilidades no cambian de subpoblación en subpoblación.

Obs: A priori no se tiene información sobre la distribución de X en cada subpoblación, pero se sabe el tamaño de la muestra en ella, n_{+k} . Las frecuencias observadas serán

X	Subpoblación 1	Subpoblación 2	\dots	Subpoblación k
x_1	n_{11}	n_{12}	\dots	n_{1K}
x_2	n_{21}	n_{22}	\dots	n_{2K}
\vdots	\vdots	\vdots	\ddots	\vdots
x_a	n_{a1}	n_{a2}	\dots	n_{aK}
<i>Total</i>	n_{+1}	n_{+2}	\dots	n_{+K}

y la frecuencia esperada bajo homogeneidad sería

$$e_{ik} = n_{+k} \times \frac{n_{i+}}{n_{++}},$$

donde $n_{i+} = \sum_{k=1}^K n_{ik}$ y $n_{++} = \sum_{i=1}^a \sum_{k=1}^K n_{ik}$ que coincide con el tamaño muestral.

Obs: Bajo homogeneidad, pensar en la distribución de X en cada subpoblación es redundante y se puede pensar en que simplemente se ha observado una única población con $n_{1+}, n_{2+}, \dots, n_{a+}$ unidades en cada categoría. Así la proporción de unidades en cada categoría, bajo homogeneidad, se calcula como $\frac{n_{i+}}{n_{++}}$, y la frecuencia esperada correspondiente en cada subpoblación con $n_{+k} \times \frac{n_{i+}}{n_{++}}$.

D6. Distribución Muestral del Estadístico de Prueba en una Prueba de Homogeneidad: Sea $X_1^{(k)}, X_2^{(k)}, \dots, X_{n+k}^{(k)}$ una muestra aleatoria de X en la subpoblación k , para $k=1, \dots, K$. Si se define la variable aleatoria O_{ik} como el número de observaciones de la muestra de la subpoblación k , donde el valor de X observado es x_i , entonces si la distribución de X es la misma en todas las subpoblaciones, el estadístico

$$\sum_{k=1}^K \sum_{i=1}^a \frac{(O_{ik} - e_{ik})^2}{e_{ik}}$$

sigue una distribución χ^2 con $\nu = (a - 1) \times (K - 1)$ grados de libertad, cuando n tiende a infinito, donde $e_{ik} = n_{+k} \times \frac{n_{i+}}{n_{++}}$.

Obs: Cuando X tiene distribución Binomial, la prueba de homogeneidad permite evaluar si las proporciones de éxitos en cada subpoblación coinciden, consecuentemente la prueba de homogeneidad permite probar igualdad de proporciones.

Obs: En cualquiera de las pruebas antes descriptas, la utilización de la distribución asintótica χ^2 se recomienda para n finito, siempre que las frecuencias esperadas no sean inferiores a 5.

Ejemplos

Ejemplo 1: Durante la Segunda Guerra mundial se dividió el mapa de Londres en 576 cuadrículas de $1/4\text{km}^2$ y se contó el número de bombas caídas en cada cuadrícula durante un bombardeo alemán. Los resultados fueron los siguientes

Cantidad de impactos	0	1	2	3	4	5 o más
Frecuencia	229	211	93	35	7	1

- Contrastar la hipótesis de que la cantidad de bombas caídas por cuadrículas sigue una distribución de Poisson con parámetro 1. Indicar claramente los supuestos que se necesitan para contrastar esta hipótesis.
- ¿Se podría haber realizado el contraste del punto (a) si se sabe que los bombardeos estaban dirigidos a determinados objetivos?

Resolución:

- Si X_i denota el número de bombas en la cuadrícula i , para $i=1, \dots, 576$, y suponiendo que:

- El número de bombas que caen en la cuadrícula i es independiente del número de bombas que caen en cualquier otra cuadrícula $i' \neq i$.
- La distribución del número de bombas que caen en cada cuadrícula es la misma para todas las cuadrículas.

Entonces se puede afirmar que X_1, X_2, \dots, X_n son i.i.d. como X , y el objetivo de esta prueba de hipótesis es evaluar si X tiene distribución Poisson con parámetro 1. Es decir

$$H_0: X \sim P(1) \text{ vs. } H_1: X \not\sim P(1).$$

Como se está en presencia de una Prueba de Bondad de Ajuste, el estadístico de prueba es

$$\sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i},$$

donde O_i es la frecuencia observada para la celda i y e_i es la frecuencia esperada para la celda i bajo la distribución de probabilidad de la hipótesis nula. Bajo H_0 , este estadístico tiene una distribución χ^2 con $k - 1$ grados de libertad cuando n tiende a infinito, donde k es el número de celdas o categorías prefijadas en la distribución teórica. En una particular aplicación, donde n es un número fijo, se puede utilizar la distribución χ^2 como distribución aproximada siempre y cuando las frecuencias esperadas no sean inferiores a 5. De hecho, en el proceso de determinar las categorías se debe tener en cuenta esta limitación.

Considerando, inicialmente, las siguientes categorías en el recorrido de X

$$\{X = 0\}, \{X = 1\}, \{X = 2\}, \{X = 3\}, \{X = 4\}, \{X \geq 5\},$$

para obtener el valor de las frecuencias esperadas en cada celda, se multiplica la probabilidad de cada categoría bajo la distribución de la hipótesis nula por el total de observaciones realizadas. En este caso, por ejemplo, se tiene

$$e_i = P(X \in \text{Categoría } i) \times 576 \text{ donde } X \sim P(1).$$

En la tabla siguiente se observan las distintas frecuencias observadas y esperadas:

Cantidad de impactos	Frecuencias observadas	Frecuencias esperadas
0	229	$0,3679 \times 576=211,9$
1	211	$0,3679 \times 576=211,9$
2	93	$0,1839 \times 576=105,9$
3	35	$0,0613 \times 576=35,3$
4	7	$0,0153 \times 576=8,8$
5 o más	1	$0,0037 \times 576=2,1$
Total	576	

Como la frecuencia esperada en la última celda es menor a 5, se debe reconsiderar la división de las categorías en el recorrido de X . Una opción es combinar las dos últimas celdas, de modo que se tendrán en total 5 celdas, donde la última celda corresponderá al suceso “4 o más impactos” cuya frecuencia observada será 8 y la esperada 9,9.

Los grados de libertad de la distribución de probabilidad del estadístico de prueba son 4 ($k - 1$). Si se considera un nivel de significación del 5%, la región crítica es:

$$\chi^2 > 9,49, \text{ ya que } \chi_{0,05}^2 = 9,49.$$

El valor del estadístico de prueba es 3,73 y se encuentra en la región de aceptación de la H_0 . Por lo tanto, no se tiene evidencia suficiente para rechazar que el número de impactos por cuadrícula sigue una distribución de Poisson con media 1.

- b) Si se supiera que los bombardeos estaban dirigidos a determinados objetivos, el promedio de impactos en cada cuadrícula no sería el mismo, pues sería mayor en aquellas con objetivos estratégicos. Consecuentemente no se puede suponer que el número de impactos en cada una de las 576 cuadrículas constituya una muestra aleatoria de una distribución particular (puesto que ellas no serán igualmente distribuidas) y por lo tanto, no puede realizarse la prueba de la forma empleada en a).

Ejemplo 2: Una fábrica de automóviles quiere averiguar si la preferencia de modelo está relacionada al sexo de sus posibles clientes. Para ello, se tomó una muestra aleatoria de 2.000 posibles clientes y se obtuvieron los siguientes resultados:

Sexo	Modelo		
	I	II	III
Varón	402	311	437
Mujer	289	340	221

Contrastar la hipótesis de que la preferencia de modelo no está relacionada al sexo de los posibles clientes a un nivel de significación de 0,01.

Resolución:

Sean las hipótesis estadísticas

H_0 : El modelo de auto elegido es independiente del sexo del cliente.

H_1 : El modelo de auto elegido no es independiente del sexo del cliente.

En este ejemplo se está en presencia de dos variables aleatorias categóricas, X : sexo, con $\text{Rec}(X)=\{\text{varón}, \text{mujer}\}$ e Y : modelo de auto elegido, con $\text{Rec}(Y)=\{I, II, III\}$ y se pretende evaluar la independencia entre las mismas.

El estadístico de prueba es

$$\sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - e_{ij})^2}{e_{ij}},$$

donde O_{ij} es la frecuencia observada y e_{ij} es la frecuencia esperada bajo independencia para la celda de la fila i y columna j , para $i = 1, 2$ y $j = 1, 2$ y 3 . Bajo H_0 , este estadístico tiene una distribución χ^2 con $(a-1)(b-1)$ grados de libertad cuando el tamaño muestral tiende a infinito, donde a es el número de filas, cantidad de elementos del $\text{Rec}(X)$ y b es el número de columnas, cantidad de elementos del $\text{Rec}(Y)$. En este caso particular en que $n=2.000$, la distribución del estadístico de prueba puede considerarse como χ^2 siempre que las frecuencias esperadas no sean inferiores a 5.


Para obtener el valor de las frecuencias esperadas en cada celda ij , se multiplica el total de las observaciones de la fila i por el total de observaciones de la columna j y al resultado se lo divide en el total de observaciones. En la tabla siguiente se observan las frecuencias observadas y entre paréntesis las frecuencias esperadas para este caso.

Sexo	Modelo		
	I	II	III
Varón	402 (397,3)	311 (374,3)	437 (378,4)
Mujer	289 (293,7)	340 (276,7)	221 (279,7)

Como ninguna frecuencia esperada es inferior a 5, se puede afirmar que el estadístico de prueba posee una distribución aproximadamente χ^2 con 2 grados de libertad. La región crítica de esta prueba de hipótesis es $\chi^2 > 9,21$, para un nivel de significancia del 1%. El valor observado del estadístico de prueba para la particular muestra presentada en este ejemplo es 46,71. El valor observado del estadístico de prueba se encuentra en la región crítica, por lo tanto se puede afirmar que existe evidencia suficiente para rechazar la hipótesis de que el modelo de auto elegido es independiente del sexo del cliente.

Resolución del Ejemplo 2 con SPSS:

1° El Programa SPSS tiene dos solapas: Vista de variables y Vista de datos. Las variables Modelo, Sexo y Frecuencia se generan en Vista de variables donde se define el nombre de la variable, su tipo que en este caso es numérica, la anchura de la fila, los decimales a utilizar, el ancho de la columna, la alineación, la medida y, por último, los valores donde se crean los datos. En la columna Nombre y Etiqueta se escriben los nombres de las variables:

Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores
Modelo	Numérico	8	2	Modelo	Ninguna 
Sexo	Numérico	8	2	Sexo	Ninguna
Frecuencia	Numérico	8	2	Frecuencia	Ninguna

Para generar los datos de la variable Modelo y para indicarle al programa que existen tres modelos distintos, hacer clic en los puntos suspensivos de la columna Valores, en este caso de la fila que corresponde a Modelo, de esta manera se despliega la ventana de Etiquetas de valor donde se debe escribir en Valor el número del modelo, por

ejemplo 1, y la Etiqueta que en este caso sería Modelo I y hacer clic en añadir, se realiza la misma operación para cada modelo y, una vez añadidos los tres modelos en la misma venta de Etiquetas de valor, se acepta. Este mismo procedimiento se lleva a cabo para la variable Sexo.

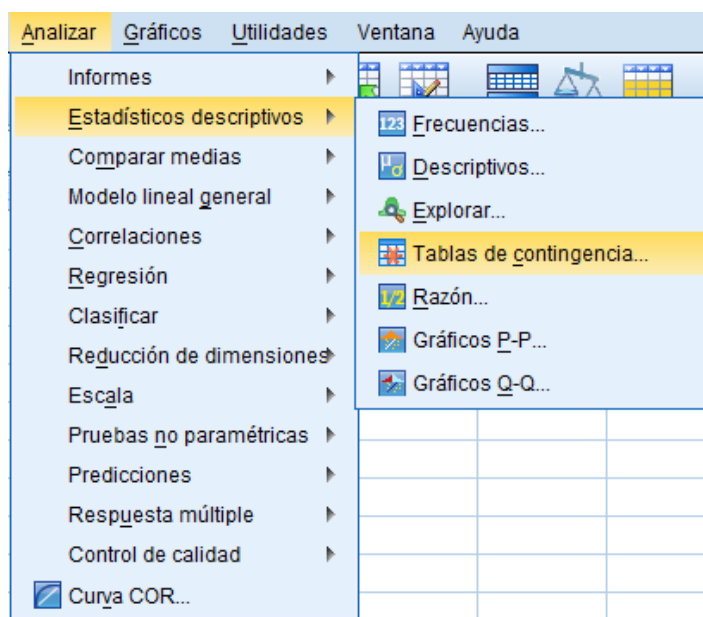
The image shows three sequential screenshots of the 'Etiquetas de valor' (Value Labels) dialog box in SPSS, illustrating the process of adding labels for a variable.

- First Screenshot:** The 'Valor' (Value) field contains '1' and the 'Etiqueta' (Label) field contains 'Modelo I'. The 'Añadir' (Add) button is highlighted.
- Second Screenshot:** The 'Valor' field is empty, and the 'Etiqueta' field is empty. The list of labels shows '1,00 = "Modelo I"', '2,00 = "Modelo II"', and '3,00 = "Modelo III"'. The 'Añadir' button is highlighted.
- Third Screenshot:** The 'Valor' field is empty, and the 'Etiqueta' field is empty. The list of labels shows '1,00 = "Varón"' and '2,00 = "Mujer"'. The 'Añadir' button is highlighted.

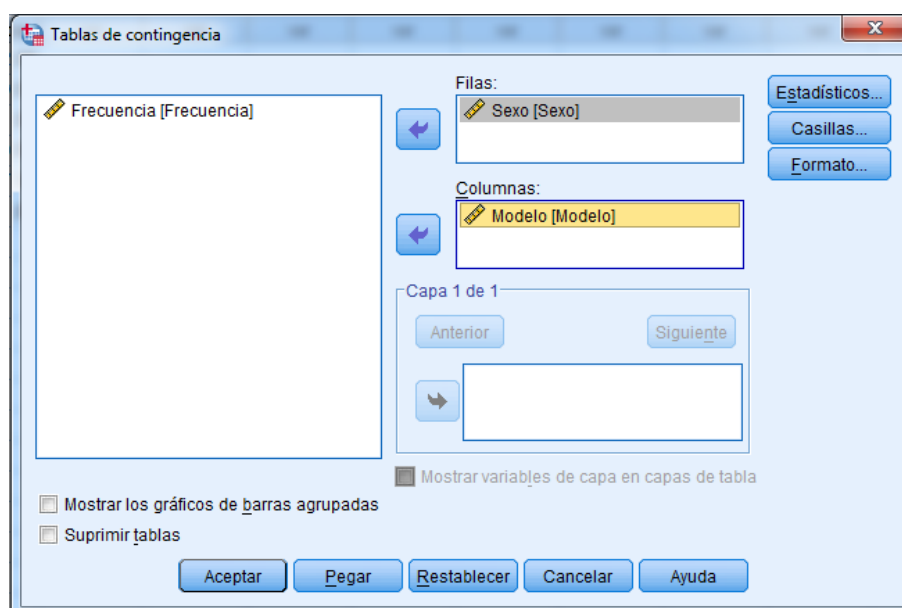
2° Completar las variables con los valores correspondientes en la solapa Vista de datos donde aparecen las columnas con los nombres respectivos:

	Modelo	Sexo	Frecuencia
1	1,00	1,00	402,00
2	1,00	2,00	289,00
3	2,00	1,00	311,00
4	2,00	2,00	340,00
5	3,00	1,00	437,00
6	3,00	2,00	221,00

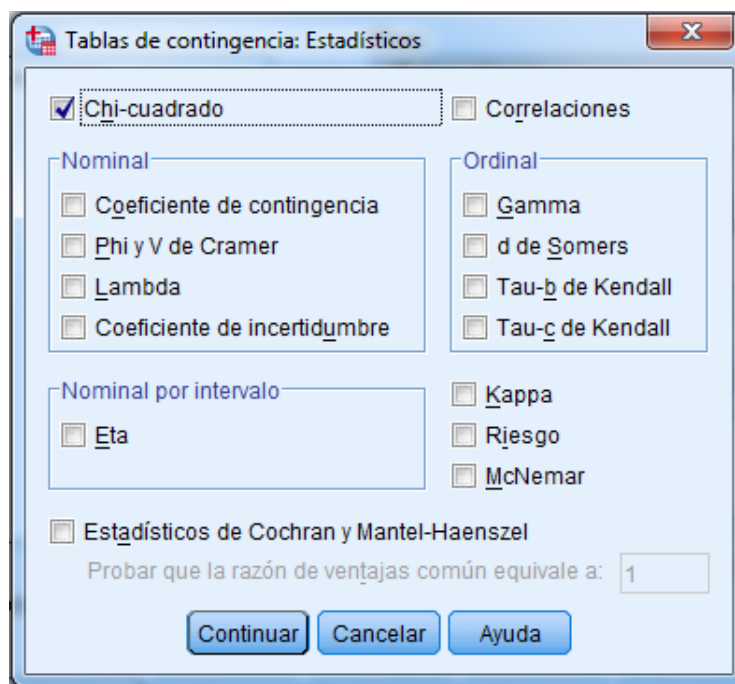
3° Para realizar la prueba, seleccionar el Menú ANALIZAR, ESTADÍSTICOS DESCRIPTIVOS, TABLAS DE CONTINGENCIA.



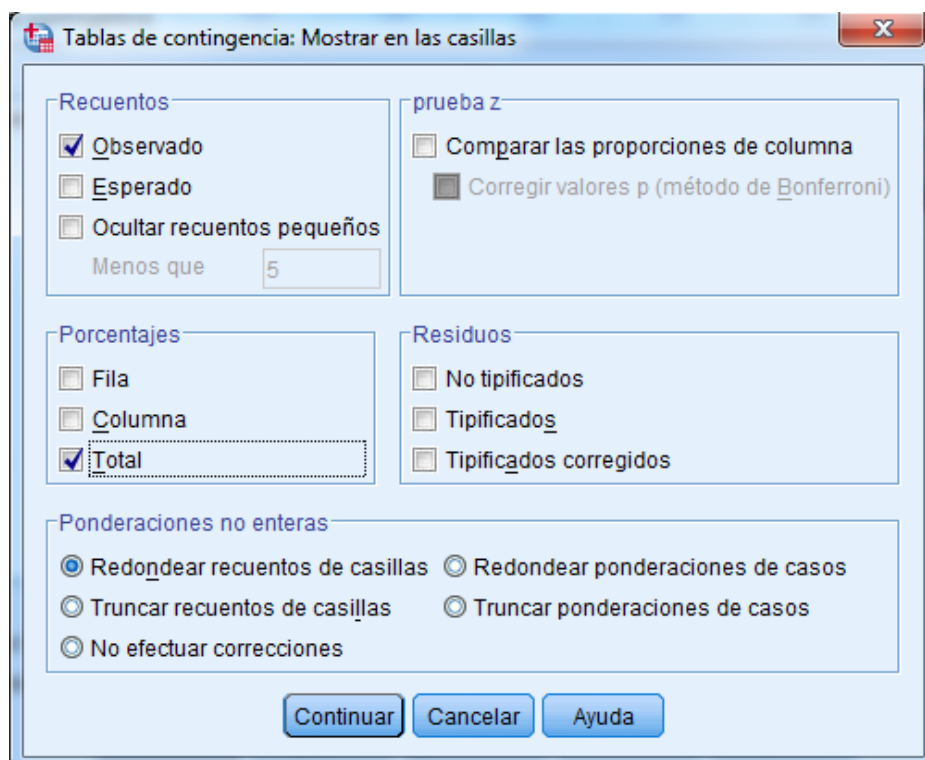
En la ventana de Tablas de contingencia, elegir Sexo en las Filas y Modelo en las Columnas:



Luego, hacer clic en ESTADÍSTICOS para tildar la opción Chi-Cuadrado y continuar:



En la ventana de Tablas de contingencia, seleccionar CASILLAS y tildar las opciones Total en Porcentajes y Observado en Recuentos y continuar:



4° Al aceptar se obtienen los resultados:

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Sexo * Modelo	2000	100,0%	0	,0%	2000	100,0%

Tabla de contingencia Sexo * Modelo

			Modelo			Total
			Modelo I	Modelo II	Modelo III	
Sexo	Varon	Recuento	402	311	437	1150
		% del total	20,1%	15,6%	21,9%	57,5%
	Mujer	Recuento	289	340	221	850
		% del total	14,5%	17,0%	11,1%	42,5%
Total		Recuento	691	651	658	2000
		% del total	34,6%	32,6%	32,9%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	46,728 ^a	2	,000
Razón de verosimilitudes	46,922	2	,000
Asociación lineal por lineal	8,836	1	,003
N de casos válidos	2000		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5.
La frecuencia mínima esperada es 276,68.

Ejemplo 3: Se hizo un estudio para evaluar la evolución del estándar de vida de las familias durante el año 2023 en el conglomerado urbano del Gran San Miguel de Tucumán. Para ello se recogieron muestras independientes de familias al fin de cada trimestre y se las clasificó según su estándar de vida. Los datos recogidos se muestran en la siguiente tabla:

Período	Estándar de vida			Total
	Bueno	Regular	Malo	
Marzo 2023	72	144	84	300
Junio 2023	63	135	102	300
Septiembre 2023	47	100	53	200
Diciembre 2023	40	105	55	200

Contrastar la hipótesis, con un nivel de significación de 0,05, de que las proporciones de familias dentro de cada categoría de estándar de vida son las mismas para cada uno de los cuatro períodos de tiempo.

Resolución:

Sean las hipótesis estadísticas

H_0 : Las proporciones de familias con estándar de vida bueno, regular y malo se mantienen constantes en los cuatro trimestres del año 2023.

H_1 : Las proporciones de familias con estándar de vida bueno, regular y malo no se mantienen constantes en los cuatro trimestres del año 2023.

En este ejemplo la variable X : estándar de vida, tiene $\text{Rec}(X)=\{\text{bueno}, \text{regular}, \text{malo}\}$ y se evalúa en cuatro períodos distintos. Se puede considerar que en cada período de tiempo se observa una subpoblación distinta:

subpoblación 1 (1° Trimestre), subpoblación 2 (2° Trimestre), subpoblación 3 (3° Trimestre) y subpoblación 4 (4° Trimestre).

El estadístico de prueba es:

$$\sum_{i=1}^a \sum_{k=1}^K \frac{(O_{ik} - e_{ik})^2}{e_{ik}}$$

donde O_{ik} es el número de observaciones de la muestra de la subpoblación k , donde el valor de X observado pertenece a la categoría i , y e_{ik} es la frecuencia esperada para la categoría i en la subpoblación k bajo la hipótesis de homogeneidad de las proporciones. Este estadístico, bajo H_0 , tiene una distribución χ^2 con $(a-1)(K-1)$ grados de libertad cuando el tamaño muestra tiende a infinito, donde a es el número de categorías en X , cantidad de elementos del $\text{Rec}(X)$ y K es el número de subpoblaciones. En este caso particular en que $n=300+300+200+200$, $a=3$ y $K=4$, y bajo el supuesto de que la distribución de X es la misma en todas las subpoblaciones (H_0), la distribución del estadístico de prueba puede considerarse como χ^2_6 siempre que las frecuencias esperadas en cada celda no sean inferiores a 5.

Para obtener el valor de las frecuencias esperadas en cada celda ik , se multiplica el total de las observaciones de la categoría i por el total de observaciones en la subpoblación k y al resultado se lo divide en el total de observaciones. En la tabla siguiente se observan las frecuencias observadas y entre paréntesis las frecuencias esperadas para este caso.

Período	Estándar de vida			Total
	Bueno	Regular	Malo	
Marzo 2003	72 (68,1)	144 (145,2)	84 (86,7)	300
Junio 2003	63 (68,1)	135 (145,2)	102 (86,7)	300
Septiembre 2003	47 (45,4)	100 (96,8)	53 (57,8)	200
Diciembre 2003	45 (45,4)	105 (96,8)	50 (57,8)	200

Como ninguna frecuencia esperada es inferior a 5, se puede decir que, bajo H_0 , el estadístico tiene una distribución aproximada χ^2 con $(4-1)(3-1)$ grados de libertad. Consecuentemente, en este caso se tiene 6 grados de libertad y la región crítica de del test es $\chi^2 > 12,59$, para un nivel de significación del 5%.

El valor del estadístico de prueba observado en la muestra provista en el ejemplo es 6,43. Por encontrarse el valor observado del estadístico de prueba en la región de aceptación de H_0 , no existe evidencia suficiente para rechazar que las proporciones de estándar de vida bueno, regular y malo se mantuvieran constantes en los cuatro trimestres del año 2.023.

CAPÍTULO 13: REGRESIÓN LINEAL SIMPLE

Regresión lineal simple. El método de mínimos cuadrados. Estimación de los parámetros de la regresión lineal. Propiedades de los estimadores por mínimos cuadrados. Intervalos de confianza y prueba de hipótesis para los parámetros de la regresión. Intervalo de predicción. Coeficiente de correlación. Prueba de hipótesis sobre el coeficiente de correlación.

Objetivos:

El alumno debe ser capaz

- ✓ Reconocer si el modelo de regresión lineal simple es un modelo adecuado para describir la relación entre dos variables.
- ✓ Determinar la mejor estimación de la relación lineal entre las variables basado en el criterio de mínimos cuadrados.
- ✓ Interpretar el significado de los parámetros estimados.
- ✓ Predecir valores de una de las variables para distintos niveles de la otra.
- ✓ Realizar inferencia acerca de la pendiente y la ordenada al origen de la regresión.
- ✓ Construir intervalos de confianza para los parámetros de la regresión.
- ✓ Calcular el coeficiente de correlación.

Resumen

Análisis de Regresión: En términos generales, en un análisis de regresión se relaciona el valor de una o más variables de interés, llamadas variables dependientes o respuestas, con un conjunto de variables independientes o regresoras, con el objetivo de:

- (i) Encontrar e interpretar constantes desconocidas en una relación conocida.
- (ii) Entender las razones de una relación estadística o esquemas de asociación.
- (iii) Predecir las variables respuestas dados ciertos valores de las regresoras.

Si se evalúa la regresión de una variable de interés Y , respecto de una única variable regresora X , se denomina *Análisis de Regresión Simple*. En cambio, cuando se evalúa la regresión de una variable de interés Y , respecto de un conjunto de variables regresoras X_1, X_2, \dots, X_k se presenta un *Análisis de Regresión Múltiple*.

D1. Modelo de Regresión Lineal Simple: Dados los valores fijos x_1, x_2, \dots, x_n , el modelo de regresión lineal simple establece la siguiente relación

$$Y_i = \alpha + \beta x_i + E_i, \text{ para } i=1, \dots, n,$$

donde: - x_1, x_2, \dots, x_n son cantidades fijas (no aleatorias).

- E_1, E_2, \dots, E_n son vs.as. con $E(E_i) = 0$ y $Cov(E_i, E_j) = \begin{cases} \sigma_E^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$.

- α, β y σ_E^2 son los parámetros del modelo (constantes desconocidas).

Obs: - El modelo **D1.** es compatible con un esquema de muestreo donde Y_i es la variable dependiente que se observará en un individuo seleccionado aleatoriamente de una subpoblación que presenta a x_i como valor de la variable regresora.

- α, β son los parámetros del modelo en que se tiene mayor interés. El parámetro α es el valor esperado para la variable dependiente Y cuando la variable regresora asume el valor 0. El parámetro β representa la variación esperada en la variable dependiente cuando la variable regresora aumenta una unidad.

Método de Mínimos Cuadrados para estimar α, β : Con este método se establece como estimadores para α, β , aquellos $\hat{\alpha}, \hat{\beta}$ que minimizan la suma de cuadrados de los residuos $e_i = y_i - \hat{\alpha} - \hat{\beta}x_i$, o errores en el ajuste del modelo. Es decir:

$$\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \leq \sum_{i=1}^n (y_i - a - bx_i)^2 \text{ para cualquier } a \text{ y } b.$$

Obs: El método de mínimos cuadrados no se basa en ningún supuesto sobre las distribuciones de probabilidad de los errores E_1, E_2, \dots, E_n .

D2. Estimadores de Mínimos Cuadrados para $\alpha y \beta$:

$$\hat{\beta} = \frac{n \sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n Y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad \hat{\alpha} = \frac{\sum_{i=1}^n Y_i - \hat{\beta} \sum_{i=1}^n x_i}{n}$$

Obs: - Al obtener $\hat{\alpha} y \hat{\beta}$, el valor de la variable dependiente se predice con el correspondiente valor de la recta estimada, es decir $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$.

- $\hat{\alpha} y \hat{\beta}$ son vs.as. puesto que son combinaciones lineales de las variables aleatorias Y_1, Y_2, \dots, Y_n .

- Los valores observados de $\hat{\alpha} y \hat{\beta}$ se denotan con a y b .

- Si se denota a $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ y $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$, se pueden expresar $\hat{\alpha} y \hat{\beta}$ como: $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ y $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}$.

T1. Propiedades de los Estimadores de Mínimos Cuadrados: Bajo el modelo de regresión lineal se tiene que:

$$- E(\hat{\beta}) = \beta y E(\hat{\alpha}) = \alpha$$

$$- V(\hat{\beta}) = \frac{\sigma_E^2}{S_{xx}} y V(\hat{\alpha}) = \sigma_E^2 \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \right)^2$$

T2. Un estimador insesgado para σ_E^2 es $S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n-2}$. Una estimación se denota por $s^2 = \frac{sse}{n-2} = \frac{\sum_{i=1}^n (y_i - a - bx_i)^2}{n-2}$.

T3. Bajo el modelo de regresión lineal juntamente con el supuesto de distribución normal para los errores E_1, E_2, \dots, E_n , se tiene que

$$\frac{\hat{\beta} - \beta}{S / \sqrt{S_{xx}}}$$

sigue una distribución t con $n-2$ grados de libertad. Un intervalo del $(1 - \gamma) \times 100\%$ de confianza para el parámetro β en el modelo de regresión lineal es:

$$IC_{(1-\gamma)100\%}(\beta) = \left(b - \frac{t_{\gamma/2} s}{\sqrt{S_{xx}}}, b + \frac{t_{\gamma/2} s}{\sqrt{S_{xx}}} \right),$$

Donde $t_{\gamma/2}$ es el punto crítico de la distribución t con $n-2$ grados de libertad, que deja a derecha una probabilidad $\gamma/2$, b es el valor observado de $\hat{\beta}$ en la muestra seleccionada (estimación), donde s es la raíz cuadrada positiva del valor observado de S^2 para la muestra seleccionada.

T4. Bajo el modelo de regresión lineal, juntamente con el supuesto de distribución normal para los errores E_1, E_2, \dots, E_n , se tiene que el estadístico

$$\frac{\hat{\alpha} - \alpha}{S \sqrt{\frac{\sum_{i=1}^n x_i^2}{n S_{xx}}}}$$

sigue una distribución t con $n-2$ grados de libertad. Un intervalo del $(1 - \gamma) \times 100\%$ de confianza para el parámetro α en el modelo de regresión lineal es:

$$IC_{(1-\gamma)100\%}(\alpha) = \left(a - \frac{t_{\gamma/2} s \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}, a + \frac{t_{\gamma/2} s \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}} \right),$$

Donde $a, t_{\gamma/2}, s$ y S_{xx} son los mismos ya definidos en D2 y T3.

Predicción: El objetivo en muchas investigaciones es establecer relaciones que permitan predecir valores de la variable dependiente. Para un determinado valor x_0 de la variable regresora se puede predecir el *valor promedio de la variable dependiente*, $E(Y|x_0) = \mu_{Y|x_0}$, o “el” *valor de la variable dependiente*, $Y_0 = \mu_{Y|x_0} + E_0$. En el primer caso se intenta estimar el valor de un parámetro, en cambio en el segundo se intenta predecir el valor de una variable aleatoria. La siguiente tabla presenta las diferencias entre los errores cometidos en ambas predicciones.

Tabla N° 1: Predicción para valores futuros de x_0

	Predecir un valor	
	Promedio	Individual
Característica a predecir	$\mu_{Y x_0}$	$\mu_{Y x_0} + E_0$
Predictor	$\hat{\mu}_{Y x_0} = \hat{\alpha} + \hat{\beta}x_0$	$\hat{\mu}_{Y x_0} = \hat{\alpha} + \hat{\beta}x_0$
Error en la Predicción	$\hat{\mu}_{Y x_0} - \mu_{Y x_0}$	$\hat{\mu}_{Y x_0} - \mu_{Y x_0} - E_0$
Varianza del Error	$V(\bar{Y}) + V(\hat{\beta})(x_0 - \bar{x})^2$	$V(\bar{Y}) + V(\hat{\beta})(x_0 - \bar{x})^2 + V(E_0)$
$IC_{\gamma \times 100\%}$	$a + bx_0 \pm t_{\gamma/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$	$a + bx_0 \pm t_{\gamma/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

T5. Descomposición de la suma de cuadrados: $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

Obs:

- Para una muestra particular $\sum_{i=1}^n (y_i - \bar{y})^2$ es una forma de medir la variabilidad de la muestra y $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$ se conoce como la suma de cuadrados totales.
- $(\hat{Y}_i - \bar{Y})$ es la diferencia entre predecir Y_i con el modelo de regresión lineal o con una media general, consecuentemente $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ se conoce como la suma de cuadrados explicados por la regresión. Debido que $\hat{Y}_i = \bar{Y} - \hat{\beta}\bar{x} + \hat{\beta}x_i$, se puede deducir que $SSR = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2$.
- $(y_i - \hat{Y}_i)$ es la diferencia entre el valor observado y el valor predicho con el modelo de regresión lineal, consecuentemente $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ se conoce como la suma de cuadrados del error del ajuste.

T6. Cuando $\beta = 0$, $\frac{SSR}{\sigma_E^2}$ y $\frac{SSE}{\sigma_E^2}$ son vs. as. χ^2 independientes con 1 y $n-2$ grados de libertad respectivamente, y por lo tanto $\frac{SSR}{SSE/(n-2)} \sim F(1, n-2)$.

Obs: El estadístico $\frac{SSR}{SSE/(n-2)}$, puede desempeñar el papel de estadístico de una prueba de hipótesis para $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$.

Obs: En la literatura estadística los resultados presentados en **T2**, **T5** y **T6** se resumen en la siguiente tabla de Análisis de Varianza para un modelo de Regresión Lineal Simple.

Fuente de Variabilidad	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F
Regresión	SSR	1	$SSR/1$	$\frac{SSR/1}{SSE/(n-2)}$
Error	SSE	$n-2$	$SSE/(n-2)$	
Total	SST	$n-1$	$SST/(n-1)$	

En la tabla de ANOVA:

- El estadístico F es el estadístico de prueba para contrastar $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$ en el modelo de regresión lineal simple (**T6**).
- El cuadrado medio correspondiente al error observado en una muestra particular es la estimación de σ_E^2 en el modelo de regresión lineal simple (**T2**).

D5. Coeficiente de Correlación Poblacional: Dadas dos variables aleatorias X e Y , el coeficiente de correlación poblacional se define como $\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$.

D6. Coeficiente de Correlación Muestral: Dadas $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ i.i.d. como (X, Y) , el coeficiente de correlación muestral o de Pearson se define como

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

T7. Cuando $Y = \alpha + \beta X + E$, se tiene que $\rho = \beta \sigma_X / \sigma_Y$.

T8. $r = b \sqrt{\frac{S_{xx}}{SST}}$.

D7. $R^2 = \frac{SSR}{SST}$ es la proporción de la variabilidad total explicada por la regresión, y se denomina *coeficiente de determinación* al valor observado de R^2, r^2 , en una muestra particular.

T9. Cuando $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ i.i.d. (X, Y) con distribución normal bivariada y $\rho_{XY} = 0$, entonces el estadístico

$$r \times \sqrt{\frac{n-2}{1-r^2}} \sim t^{(n-2)},$$

puede usarse para evaluar $H_0: \rho_{XY} = 0$ vs. $H_1: \rho_{XY} \neq 0$.

Ejemplos

Ejemplo 1: Existe la discusión sobre si una mayor inversión en educación se ve reflejada en una mayor calidad en el aprendizaje de los alumnos. La siguiente tabla muestra la inversión media anual por alumno en quince localidades y el resultado promedio alcanzado por sus alumnos en un Operativo de Evaluación de la Calidad.

- Presentar los datos en un gráfico de dispersión.
- De acuerdo al gráfico, ¿es razonable aplicar un análisis de regresión lineal para describir la relación entre ambas variables?
- ¿Cuál es la variable independiente y cuál la dependiente? ¿Por qué?
- Indicar la estimación por mínimos cuadrados de los coeficientes del modelo, α y β . Estimar σ^2 .
- Escribir la recta de regresión estimada y dibujarla en el gráfico de dispersión.
- Interpretar las estimaciones encontradas de la pendiente y ordenada al origen de la recta de regresión estimada.

Localidad	Gasto por alumno (en cientos de pesos)	Resultado promedio en Matemática
1	41	58
2	34	47
3	49	58
4	55	63
5	43	55
6	38	43
7	47	61
8	49	61
9	41	55
10	40	53
11	80	72
12	38	61
13	40	52
14	52	62
15	61	75

Resolución:

- Se representa en el eje X el gasto por alumno en cada localidad y en el eje Y el rendimiento promedio en Matemática de los alumnos.

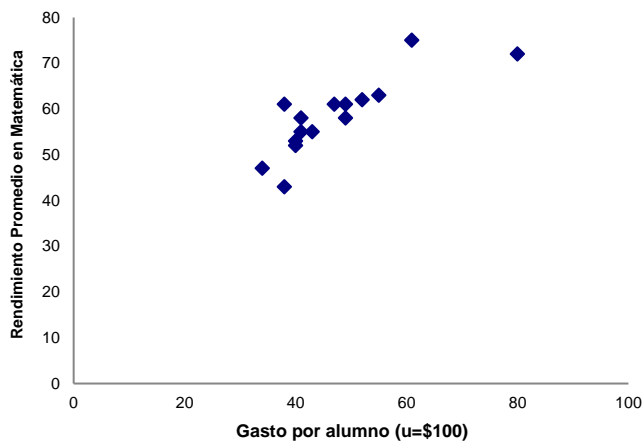


Diagrama de dispersión del gasto en el sistema educativo por alumno y rendimiento promedio en Matemáticas en un operativo de calidad Educativa, en 15 localidades

Nota: Para realizar el gráfico de dispersión en Excel primero se debe copiar la tabla con los datos en una hoja y seleccionarlos. En el menú "Insertar" ingresar a "Gráfico" y seleccionar tipo de gráfico "Dispersión" luego aceptar.

- Puesto que la nube de puntos parece rodear una línea recta ascendente, es razonable ajustar un modelo de regresión lineal simple.
- La variable independiente o explicativa es el gasto por alumno y la variable dependiente es el rendimiento promedio en Matemática, ya que con una mayor inversión en educación se espera un mejor rendimiento promedio en Matemática, o sea se espera que la inversión influya en el rendimiento.
- Los estimadores por mínimos cuadrados de α y β son los siguientes:

$$\hat{\beta} = \frac{n \sum_{i=1}^n x_i Y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, y \hat{\alpha} = \frac{\sum_{i=1}^n Y_i - \hat{\beta} \sum_{i=1}^n x_i}{n},$$

y el estimador de σ^2 es

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n-2}.$$

Para encontrar los estimadores se puede recurrir a un software estadístico, usar la planilla Excel o construir convenientemente una tabla con los valores necesarios para reemplazar en las fórmulas. Esto último ayuda en la interpretación de las fórmulas usadas en el proceso de estimación. Para ello se construye la siguiente tabla hasta la columna 5:

1	2	3	4	5	6	7	8
Localidad	Gasto por alumno (en cientos de pesos)	Res. Promedio en Matemática Y	X^2	XY	\hat{Y}	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1	41	58	1681	2378	54,7274	3,33	11,09
2	34	47	1156	1598	50,5876	-3,54	12,53
3	49	58	2401	2842	59,4586	-1,39	1,93
4	55	63	3025	3465	63,007	0,07	0,00
5	43	55	1849	2365	55,9102	-0,85	0,72
6	38	43	1444	1634	52,9532	-9,90	98,01
7	47	61	2209	2867	58,2758	2,79	7,78
8	49	61	2401	2989	59,4586	1,61	2,59
9	41	55	1681	2255	54,7274	0,33	0,11
10	40	53	1600	2120	54,136	-1,08	1,17
11	80	72	6400	5760	77,792	-5,68	32,26
12	38	61	1444	2318	52,9532	8,10	65,61
13	40	52	1600	2080	54,136	-2,08	4,33
14	52	62	2704	3224	61,2328	0,84	0,71
15	61	75	3721	4575	66,5554	8,53	72,76
Suma	708	876	35316	42470			311,61

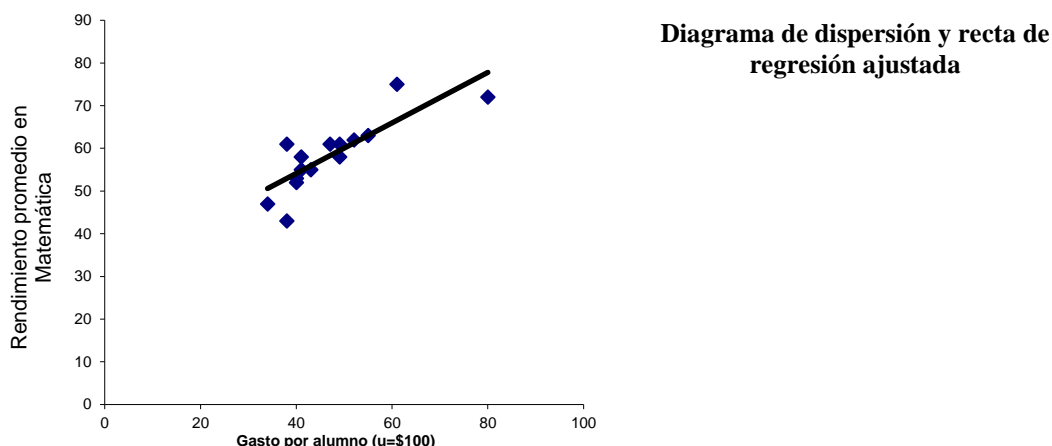
Primero se encuentra la estimación de β

$$b = \frac{15 \times 42470 - 708 \times 876}{15 \times 35316 - 708^2} \cong 0,59,$$

Luego se estima α y se puede construir la columna 6, 7 y 8 de la tabla de cálculos y encontrar la estimación de σ^2 .

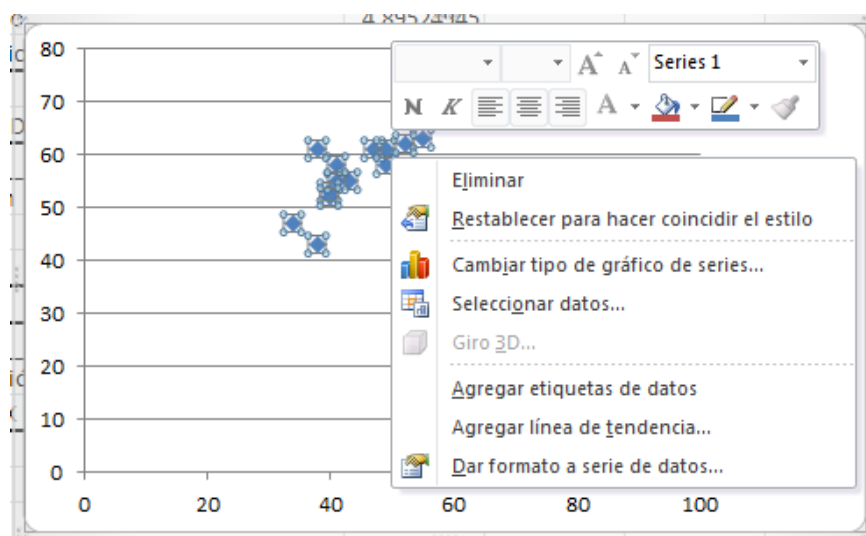
$$a = \frac{876 - 0,59 \times 708}{15} \cong 30,48 \text{ y } s^2 = \hat{\sigma}_E^2 = \frac{311,61}{13} = 23,96.$$

e) La recta de regresión estimada es $\hat{y} = 30,48 + 0,59x$.



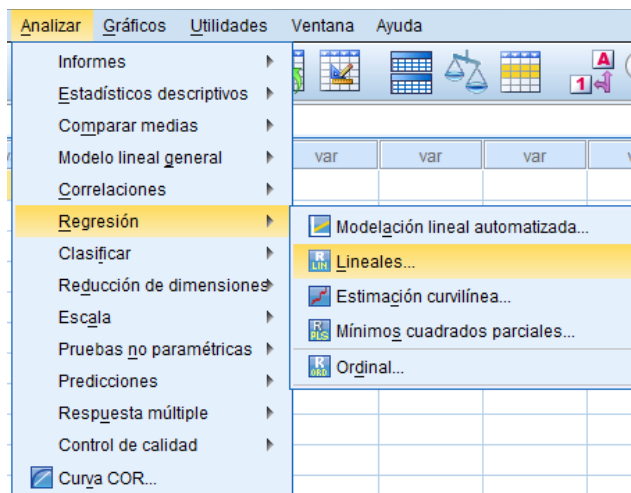
f) 0,59 puntos es la diferencia en el rendimiento medio de los alumnos de dos localidades distintas cuyos gastos en educación difieren en \$100, siendo mayor el puntaje en aquella localidad con mayor gasto en educación.

Nota: Para obtener la recta de regresión estimada junto con el gráfico, simplemente en el gráfico antes realizado se hace clic derecho sobre uno de los puntos que representa un dato; y se selecciona la opción “Agregar línea de tendencia...” en el cuadro que se despliega.

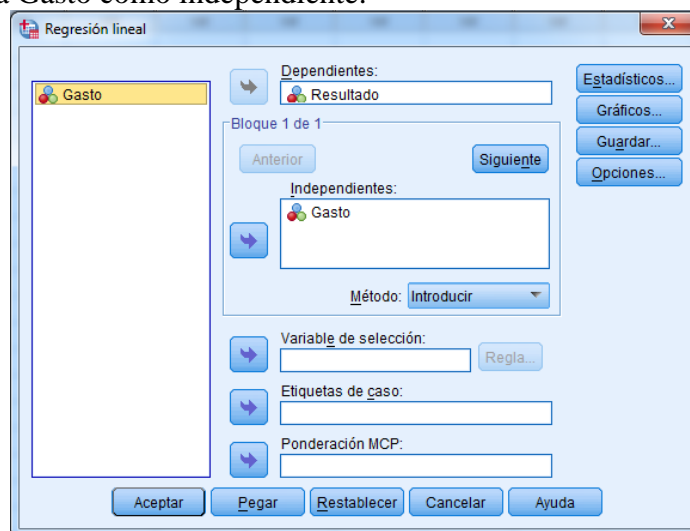


Resolución del Ejemplo 1 con SPSS:

1° Una vez creadas las variables en la solapa de Vista de variables y cargados los valores en Vista de datos, proceder a seleccionar el Menú ANALIZAR, REGRESIÓN, LINEAL como se muestra a continuación:



2° Luego, en la ventana de Regresión lineal, elegir a Resultado como variable dependiente y a Gasto como independiente:



3° Al aceptar se obtienen los resultados:

Estadísticos descriptivos

	Media	Desviación típ.	N
Resultado	58,4000	8,34780	15
Gasto_	47,2000	11,64474	15

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,825 ^a	,681	,656	4,89525

a. Variables predictoras: (Constante), Gasto_

ANOVA^b

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	664,075	1	664,075	27,712	,000 ^a
	Residual	311,525	13	23,963		
	Total	975,600	14			

a. Variables predictoras: (Constante), Gasto_

b. Variable dependiente: Resultado

Coefficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	30,484	5,452		5,592	,000	18,706	42,261
	Gasto_	,591	,112	,825	5,264	,000	,349	,834

a. Variable dependiente: Resultado

Ejemplo 2: Continuando con el Ejemplo 1:

- Construir un intervalo del 95% de confianza para α y para β . Realizar los supuestos necesarios.
- ¿Se puede rechazar la hipótesis de que la pendiente del modelo es nula?
- Predecir el rendimiento esperado en matemática en una localidad donde el gasto por alumno es \$5.800.
- Calcular un intervalo de confianza para la media del rendimiento promedio de los alumnos en matemática en una localidad cuya inversión es \$5.800.
- Luego de realizar el análisis, ¿qué conclusiones se puede tener en relación a si un mayor gasto en educación se ve reflejado en una mayor calidad en el aprendizaje de los alumnos?

Resolución:

- Suponiendo que el resultado promedio en la prueba de Matemática tiene una distribución Normal, el intervalo de confianza del $(1-\gamma) \times 100\%$ de confianza para el parámetro α se calcula con la siguiente expresión:

$$IC_{(1-\gamma).100\%}(\alpha) = \left(a - \frac{t_{\gamma/2} s \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}, a + \frac{t_{\gamma/2} s \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}} \right)$$

donde s es la raíz cuadrada de la estimación de σ^2 encontrada en el ejemplo 1, $s=4,89$, $t_{\gamma/2}$ es el valor de t que deja $\gamma/2$ de probabilidad a la derecha de la distribución con $\nu=15-2=13$ grados de libertad, en este caso $t_{\gamma/2}=2,16$ y

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 1898,4.$$

Reemplazando los valores en la expresión anterior se tiene que:

$$IC_{95\%}(\alpha) = \left(30,48 - \frac{2,16 \times 4,89 \times 187,93}{168,75}; 30,48 + \frac{2,16 \times 4,89 \times 187,93}{168,75} \right) \\ \cong (18,72; 42,24).$$

La expresión para el intervalo de confianza para el parámetro β es:

$$IC_{(1-\gamma).100\%}(\beta) = \left(b - \frac{t_{\gamma/2} s}{\sqrt{S_{xx}}}, b + \frac{t_{\gamma/2} s}{\sqrt{S_{xx}}} \right)$$

luego, reemplazando los valores se obtiene:

$$IC_{95\%}(\beta) = \left(0,59 - \frac{2,16 \times 4,89}{43,57}; 0,59 + \frac{2,16 \times 4,89}{43,57} \right) = (0,35; 0,83).$$

- b) Las hipótesis a contrastar son $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$.

El intervalo de confianza del 95% de confianza de β no contiene al valor cero por lo que se considera que hay razón suficiente para rechazar la hipótesis nula $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$.

Otra forma de contrastar la hipótesis planteada es usando el siguiente estadístico que tiene una distribución t con $(n-2)$ grados de libertad:

$$\frac{\hat{\beta} - \beta_0}{S/\sqrt{S_{xx}}}$$

Considerando un nivel de significación $\gamma = 5\%$, la región crítica de esta prueba es $t < t_{(1-\gamma/2)} = -2,16$ ó $t > t_{\gamma/2} = 2,16$.

En este caso el valor del estadístico observado es:

$$\frac{0,59}{0,113} \cong 5,21.$$

Como el valor observado del estadístico está en la región crítica del test se rechaza la hipótesis $H_0: \beta = 0$. Otra posibilidad para sacar una conclusión de este test es utilizar el valor P que en este caso es 0,00015, como este valor es inferior al valor de γ , prefijado en el 5%, se rechaza la hipótesis nula.

- c) Para calcular el valor esperado de Y cuando x es igual a \$5.800 se reemplaza este valor en la recta de regresión ajustada y se obtiene:

$$\hat{y} = 30,48 + 0,59 \times 58 \cong 64,7.$$

- d) La expresión del intervalo para la media del rendimiento promedio es:

$$IC_{(1-\gamma)100\%}(\mu_{y/x_0}) = \left(\hat{y}_0 - t_{\gamma/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \Rightarrow \hat{y}_0 \pm t_{\gamma/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right).$$

Reemplazando los valores correspondientes

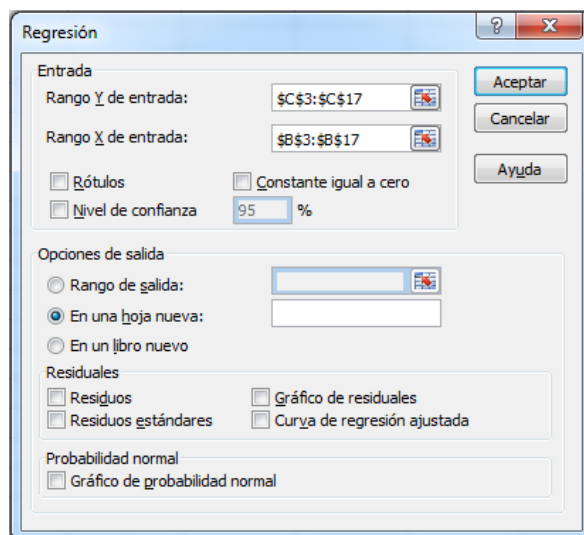
$$\left(64,7 - 2,16 \times 4,89 \sqrt{\frac{1}{15} + \frac{(58-47,2)^2}{1898,4}} \Rightarrow 64,7 \pm 2,16 \times 4,89 \sqrt{\frac{1}{15} + \frac{(58-47,2)^2}{1898,4}} \right),$$

el intervalo de confianza para la media del rendimiento promedio de los alumnos en matemática en una localidad cuya inversión es 5.800 pesos es:

$$IC_{(1-\gamma)100\%}(\mu_{y/x_0=58}) = (60,92 ; 68,48).$$

- e) Luego de realizar el análisis se puede decir que una mayor inversión en educación se ve acompañada por un mayor rendimiento promedio en matemática, ya que el signo del coeficiente β es positivo y su valor es significativamente distinto de cero.

Nota: Para estimar el modelo de regresión lineal simple con Excel, se emplea en el Menú Datos, en Análisis de datos se selecciona Regresión, y se completa el cuadro siguiente:



Algunos de los cálculos que se realizan aplicando las fórmulas ya vistas se encuentran en la salida obtenida que se transcribe a continuación:

Estadísticas de la regresión

Coefficiente de correlación múltiple	0,83
Coefficiente de determinación R^2	0,68
R^2 ajustado	0,66
Error típico (s)	4,90
Observaciones	15

ANÁLISIS DE VARIANZA

	<i>Grados de libertad</i>	<i>Suma de cuadrados</i>	<i>Promedio de cuadrados</i>	<i>F</i>	<i>Valor crítico de F</i>
Regresión	1	664,07	664,07	27,71	0,000153
Residuos	13	311,53	23,96		
Total	14	975,60			

	<i>Coefficientes</i>	<i>Error típico</i>	<i>Estadístico t</i>	<i>Probabilidad</i>	<i>Inferior 95%</i>	<i>Superior 95%</i>
Intercepción	30,48 ⁽¹⁾	5,4528 ⁽³⁾	5,59 ⁽⁵⁾	8,74585E-05 ⁽⁷⁾	18,71 ⁽⁹⁾	42,26 ⁽¹⁰⁾
Variable X 1	0,59 ⁽²⁾	0,112 ⁽⁴⁾	5,26 ⁽⁶⁾	0,000153047 ⁽⁸⁾	0,35 ⁽¹¹⁾	0,83 ⁽¹²⁾

⁽¹⁾Estimación puntual por mínimos cuadrados de α .

⁽²⁾Estimación puntual por mínimos cuadrados de β .

⁽³⁾Estimación del error estándar de $\hat{\alpha}$.

⁽⁴⁾Estimación del error estándar de $\hat{\beta}$.

⁽⁵⁾Valor del estadístico de prueba del test t con $H_0: \alpha = 0$ versus $H_1: \alpha \neq 0$.

⁽⁶⁾Valor del estadístico de prueba del test t con $H_0: \beta = 0$ versus $H_1: \beta \neq 0$.

⁽⁷⁾Valor P de test t con $H_0: \alpha = 0$ versus $H_1: \alpha \neq 0$

⁽⁸⁾Valor P de test t con $H_0: \beta = 0$ versus $H_1: \beta \neq 0$

⁽⁹⁾ y ⁽¹⁰⁾ Límite inferior y límite superior del intervalo del 95% confianza para α .

⁽¹¹⁾ y ⁽¹²⁾ Límite inferior y límite superior del intervalo del 95% confianza para β .

Las diferencias observadas entre los valores de la salida de Excel y los cálculos anteriores se deben a errores de redondeo.

ANEXOS

ANEXO I

DISTRIBUCIONES MUESTRALES DE ALGUNOS ESTADÍSTICOS

Estadístico	Situación	Supuestos	Distribución del estadístico
$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$	1.1	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ Para cualquier valor de $n \in \mathbb{N}$ 	$N(0, 1)$
	1.2	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. X, con $E(X) = \mu$ y $V(X) = \sigma^2 < \infty$ Para n grande, ($n \geq 30$) 	$N(0, 1)$ aprox
$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$	2.1	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ $n \in \mathbb{N}$ mayor que 1 	$t(n - 1)$
	2.2	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. X, con fdp en forma de campana $n \in \mathbb{N}$ mayor que 1 	$t(n - 1)$ aprox
	2.3	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. X, con $E(X) = \mu$ y $V(X) = \sigma^2 < \infty$ Para n grande, ($n \geq 30$) 	$N(0, 1)$ aprox
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	3.1	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes Para todos los valores de $n_1, n_2 \in \mathbb{N}$ 	$N(0, 1)$
	3.2	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. X_1 con $E(X_1) = \mu_1$ y $V(X_1) = \sigma_1^2 < \infty$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. X_2 con $E(X_2) = \mu_2$ y $V(X_2) = \sigma_2^2 < \infty$ Muestras aleatorias independientes $n_1, n_2 \in \mathbb{N}$ grandes ($n_1 \geq 30$ y $n_2 \geq 30$) 	$N(0, 1)$ aprox
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	4.1	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes $\sigma_1^2 = \sigma_2^2$ $n_1, n_2 \in \mathbb{N}$ mayores que 1. 	$t(n_1 + n_2 - 2)$

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

ANEXO I

DISTRIBUCIONES MUESTRALES DE ALGUNOS ESTADÍSTICOS

Estadístico	Situación	Supuestos	Distribución del estadístico
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	4.2	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. aprox $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. aprox $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes $\sigma_1^2 = \sigma_2^2$ $n_1, n_2 \in \mathbb{N}$ mayores que 1 	$t(n_1 + n_2 - 2)$ Aprox.
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	4.3	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. aprox $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. aprox $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes $n_1, n_2 \in \mathbb{N}$ mayores que 1 	$t(\nu)$ ³ aprox.
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma_D / \sqrt{n}}$	5.1	<ul style="list-style-type: none"> $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$ i.i.d (X_1, X_2) con $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_D^2)$ Para cualquier valor de $n \in \mathbb{N}$ 	$N(0, 1)$
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma_D / \sqrt{n}}$	5.2	<ul style="list-style-type: none"> $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$ i.i.d (X_1, X_2) con $V(X_1 - X_2) = \sigma_D^2 < \infty$ y $E(X_1 - X_2) = \mu_1 - \mu_2$ Para n grande, ($n \geq 30$) 	$N(0, 1)$ aprox
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_D / \sqrt{n}}$	5.3	<ul style="list-style-type: none"> $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$ i.i.d (X_1, X_2) con $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_D^2)$ $n \in \mathbb{N}$ mayor que 1 	$t(n-1)$

³ $\nu = (s_1^2 / n_1 + s_2^2 / n_2)^2 / [(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)]$

ANEXO I DISTRIBUCIONES MUESTRALES DE ALGUNOS ESTADÍSTICOS

Estadístico	Situación	Supuestos	Distribución del estadístico
$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_D / \sqrt{n}}$	5.4	<ul style="list-style-type: none"> $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$ i.i.d (X_1, X_2) con $X_1 - X_2$ en forma de campana, con media $\mu_1 - \mu_2$ $n \in \mathbb{N}$ mayor que 1 	$t(n-1)$ aprox
$\frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$	6.1	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $B(\pi)$ $\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i$ es la proporción muestral Para n grande, ($n \geq 30$), $n\hat{\pi} > 5$ y $n(1-\hat{\pi}) > 5$ 	$N(0,1)$ aprox
$\frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}}$	6.2	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $B(\pi)$ $\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i$ es la proporción muestral Para n grande, ($n \geq 30$), $n\pi > 5$ y $n(1-\pi) > 5$ 	$N(0,1)$ aprox
$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$	7.1	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $B(\pi_1)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $B(\pi_2)$ Muestras aleatorias independientes $\hat{\pi}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}$ y $\hat{\pi}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$ son las proporciones muestrales en cada muestra $n_1, n_2 \in \mathbb{N}$ grandes ($n_1 \geq 30$ y $n_2 \geq 30$) 	$N(0,1)$ aprox
$\frac{(n-1)S^2}{\sigma^2}$	8.1	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ $n \in \mathbb{N}$ mayor que 1 	$\chi^2_{(n-1)}$
$\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$	9.1	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes Para $n_1, n_2 \in \mathbb{N}$ mayores que 1 	$F(n_1-1, n_2-1)$

ANEXO II

INTERVALOS DE CONFIANZA

Parámetro	Estimador puntual	Estadístico que relaciona el estimador puntual con el parámetro	Situación	Distribución del estadístico	Supuestos	Intervalo del $(1 - \alpha) \times 100\%$ de confianza
μ	\bar{X}	$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$	1.1	$N(0,1)$	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ σ conocido Para cualquier valor de $n \in \mathbb{N}$ 	$IC(\mu) = \left(\bar{x} - z_{\alpha/2} \sigma / \sqrt{n}, \bar{x} + z_{\alpha/2} \sigma / \sqrt{n} \right)$
			1.2	$N(0,1)$ aprox	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. X, con $E(X) = \mu$ y $V(X) = \sigma^2 < \infty$ σ conocido Para n grande, $(n \geq 30)$ 	$IC(\mu) = \left(\bar{x} - z_{\alpha/2} \sigma / \sqrt{n}, \bar{x} + z_{\alpha/2} \sigma / \sqrt{n} \right)$
		$\frac{\bar{X} - \mu}{S / \sqrt{n}}$	2.1	$t(n-1)$	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ σ desconocido $n \in \mathbb{N}$ mayor que 1 	$IC(\mu) = \left(\bar{x} - t_{\alpha/2}^{(n-1)} s / \sqrt{n}, \bar{x} + t_{\alpha/2}^{(n-1)} s / \sqrt{n} \right)$
			2.2	$t(n-1)$ aprox	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. X, con fdp en forma de campana σ desconocido $n \in \mathbb{N}$ mayor que 1 	$IC(\mu) = \left(\bar{x} - t_{\alpha/2}^{(n-1)} s / \sqrt{n}, \bar{x} + t_{\alpha/2}^{(n-1)} s / \sqrt{n} \right)$
			2.3	$N(0,1)$ aprox	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. X, con $E(X) = \mu$ y $V(X) = \sigma^2 < \infty$ σ desconocido Para n grande, $(n \geq 30)$ 	$IC(\mu) = \left(\bar{x} - z_{\alpha/2} s / \sqrt{n}, \bar{x} + z_{\alpha/2} s / \sqrt{n} \right)$

ANEXO II

INTERVALOS DE CONFIANZA

Parámetro	Estimador puntual	Estadístico que relaciona el estimador puntual con el parámetro	Sit.	Distribución del estadístico	Supuestos	Intervalo del $(1 - \alpha) \times 100\%$ de confianza
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	3.1	$N(0,1)$	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes σ_1 y σ_2 conocidas Para todos los valores de $n_1, n_2 \in \mathbb{N}$ mayores a 1 	$IC(\mu_1 - \mu_2) = \left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \right. \\ \left. (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$
			3.2	$N(0,1)$ aprox	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. X_1 con $E(X_1) = \mu_1$ y $V(X_1) = \sigma_1^2 < \infty$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. X_2 con $E(X_2) = \mu_2$ y $V(X_2) = \sigma_2^2 < \infty$ Muestras aleatorias independientes σ_1 y σ_2 conocidas $n_1, n_2 \in \mathbb{N}$ grandes ($n_1 \geq 30$ y $n_2 \geq 30$) 	$IC(\mu_1 - \mu_2) = \left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \right. \\ \left. (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$
		$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	4.1	$t(n_1 + n_2 - 2)$	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes $\sigma_1^2 = \sigma_2^2$ desconocidas $n_1, n_2 \in \mathbb{N}$ mayores que 1 	$IC(\mu_1 - \mu_2) = \left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}^{(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right. \\ \left. (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}^{(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$

ANEXO II

INTERVALOS DE CONFIANZA

Parámetro	Estimador puntual	Estadístico que relaciona el estimador puntual con el parámetro	Sit.	Distribución del estadístico	Supuestos	Intervalo del $(1 - \alpha) \times 100\%$ de confianza
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	4.2	$t(n_1 + n_2 - 2)$ aprox	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. aprox $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. aprox $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes $\sigma_1^2 = \sigma_2^2$ desconocidas $n_1, n_2 \in \mathbb{N}$ mayores que 1 	$IC(\mu_1 - \mu_2) = \left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}^{(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right. \\ \left. (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}^{(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$
		$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	4.3	$t(\nu)$ aprox	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. aprox $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. aprox $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes σ_1 y σ_2 desconocidas $n_1, n_2 \in \mathbb{N}$ mayores que 1 	$IC(\mu_1 - \mu_2) = \left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}^{(\nu)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \right. \\ \left. (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}^{(\nu)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$
		$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_D / \sqrt{n}}$	5.1	$t(n-1)$	<ul style="list-style-type: none"> $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$ i.i.d (X_1, X_2) con $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_D^2)$ σ_D desconocidas Para cualquier valor de $n \in \mathbb{N}$ 	$IC(\mu_1 - \mu_2) = \left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}^{(n-1)} S_D / \sqrt{n}, \right. \\ \left. (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}^{(n-1)} S_D / \sqrt{n} \right)$

ANEXO II

INTERVALOS DE CONFIANZA

Parámetro	Estimador puntual	Estadístico que relaciona el estimador puntual con el parámetro	Sit.	Distribución del estadístico	Supuestos	Intervalo del $(1 - \alpha) \times 100\%$ de confianza
π	$\hat{\pi}$ prop. Muestral	$\frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1 - \hat{\pi})/n}}$	6.1	$N(0,1)$ aprox	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $B(\pi)$ Para n grande, ($n \geq 30$) 	$IC(\pi) = \left(\hat{\pi}_{obs} - z_{\alpha/2} \sqrt{\hat{\pi}_{obs}(1 - \hat{\pi}_{obs})/n}, \hat{\pi}_{obs} + z_{\alpha/2} \sqrt{\hat{\pi}_{obs}(1 - \hat{\pi}_{obs})/n} \right)$
$\pi_1 - \pi_2$	$\hat{\pi}_1 - \hat{\pi}_2$	$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}}$	7.1	$N(0,1)$ aprox	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $B(\pi_1)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $B(\pi_2)$ Muestras aleatorias independientes $n_1, n_2 \in \mathbb{N}$ grandes ($n_1 \geq 30$ y $n_2 \geq 30$) 	$IC(\pi_1 - \pi_2) = \left(\hat{\pi}_{1obs} - \hat{\pi}_{2obs} - z_{\alpha/2} s(\hat{\pi}_1 - \hat{\pi}_2), \hat{\pi}_{1obs} - \hat{\pi}_{2obs} + z_{\alpha/2} s(\hat{\pi}_1 - \hat{\pi}_2) \right)$ con $s(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\hat{\pi}_{1obs}(1 - \hat{\pi}_{1obs})}{n_1} + \frac{\hat{\pi}_{2obs}(1 - \hat{\pi}_{2obs})}{n_2}}$
σ^2	S^2	$\frac{(n-1)S^2}{\sigma^2}$	8.1	$\chi^2_{(n-1)}$	<ul style="list-style-type: none"> X_1, X_2, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ $n \in \mathbb{N}$ mayor que 1 	$IC(\sigma^2) = \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right)$
σ_1^2 / σ_2^2	S_1^2 / S_2^2	$\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$	9.1	$F(n_1 - 1, n_2 - 1)$	<ul style="list-style-type: none"> $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$ $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$ Muestras aleatorias independientes Para $n_1, n_2 \in \mathbb{N}$ mayores que 1 	$IC\left(\sigma_1^2 / \sigma_2^2\right) = \left(\frac{s_1^2}{s_2^2} \frac{1}{f_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2}{s_2^2} f_{\alpha/2}(n_2 - 1, n_1 - 1) \right)$

ANEXO III

PROBABILIDAD DEL ERROR TIPO II Y FUNCIÓN DE POTENCIA DE LA PRUEBA DE HIPÓTESIS

Hipótesis a Contrastar:

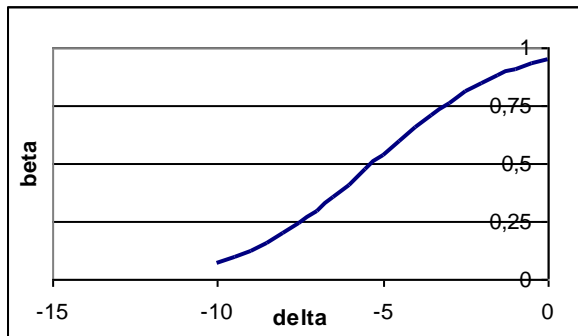
$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu < \mu_0$$

Regla de Decisión

Rechace si $\bar{x} < \mu_0 + z_{1-\alpha} \sigma / \sqrt{n}$

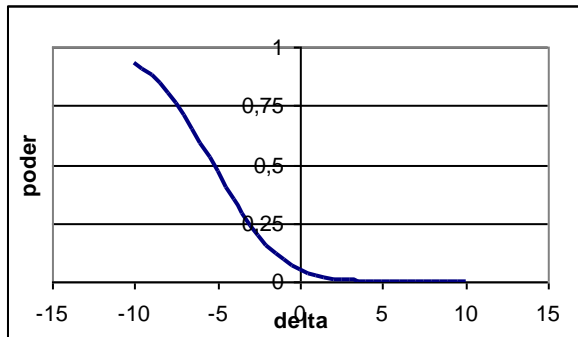
Beta

$$\beta = 1 - \Phi(z_{1-\alpha} - \delta / (\sigma / \sqrt{n}))$$



Función Potencia

$$FP(\delta) = \Phi(z_{1-\alpha} - \delta / (\sigma / \sqrt{n}))$$



Hipótesis a Contrastar:

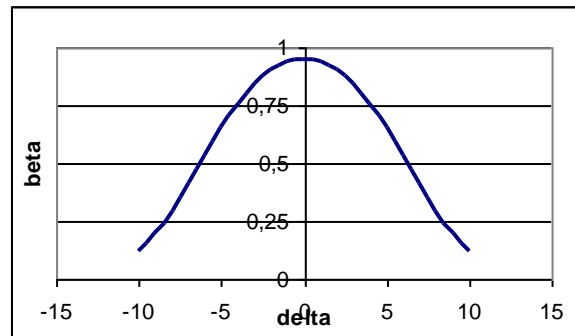
$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0$$

Regla de Decisión

Rechace si $|\bar{x} - \mu_0| > z_{\alpha/2} \sigma / \sqrt{n}$

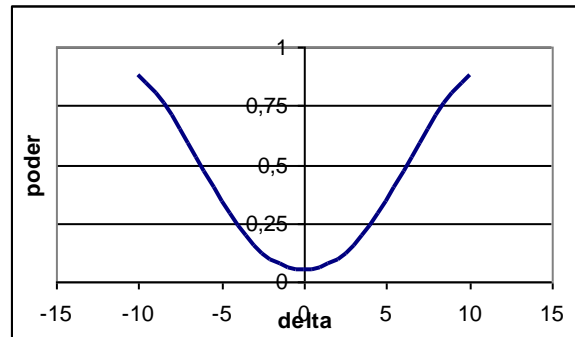
Beta

$$\beta = \Phi(z_{\alpha/2} - \delta / (\sigma / \sqrt{n})) - \Phi(-z_{\alpha/2} - \delta / (\sigma / \sqrt{n}))$$



Función Potencia

$$FP(\delta) = 1 + \Phi(-z_{\alpha/2} - \delta / (\sigma / \sqrt{n})) - \Phi(z_{\alpha/2} - \delta / (\sigma / \sqrt{n}))$$



Hipótesis a Contrastar:

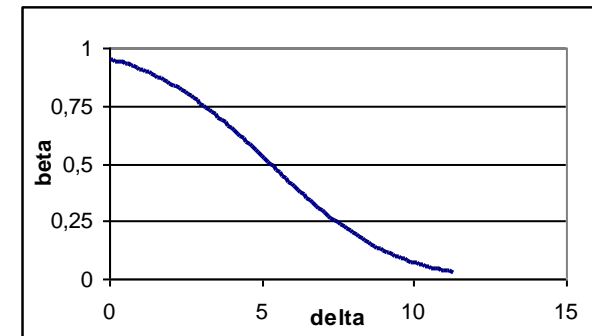
$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

Regla de Decisión

Rechace si $\bar{x} > \mu_0 + z_{\alpha} \sigma / \sqrt{n}$

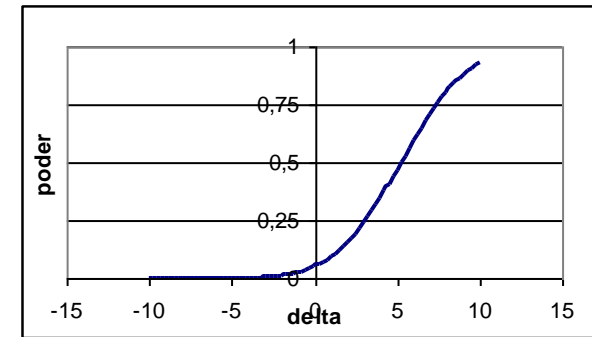
Beta

$$\beta = \Phi(z_{\alpha} - \delta / (\sigma / \sqrt{n}))$$



Función Potencia

$$FP(\delta) = 1 - \Phi(z_{\alpha} - \delta / (\sigma / \sqrt{n}))$$



Observaciones: * $\delta = \mu - \mu_0$ * $\alpha = 0,05$, $\sigma = 16$ y $n = 25$.

ANEXO IV

REGRESIÓN EN EXCEL

año	Vehículos Matriculados (0,000) - x	Accidentes en Carretera (000) - y	x ²	xy	yc	y-yc	(y-yc) ²
1	352	166	123.904	58.432	166	0	0
2	373	153	139.129	57.069	172	-19	369
3	411	177	168.921	72.747	184	-7	50
4	441	201	194.481	88.641	193	8	57
5	462	216	213.444	99.792	200	16	257
6	490	208	240.100	101.920	209	-1	1
7	529	227	279.841	120.083	221	6	37
8	577	238	332.929	137.326	236	2	5
9	641	268	410.881	171.788	256	12	148
10	692	268	478.864	185.456	272	-4	14
11	743	274	552.049	203.582	288	-14	186
Sumas	5.711	2.396	3.134.543	1.296.836		0	1.124.168
						s ²	125
						s	11.176

Resumen

Estadísticas de la regresión	
Coefficiente de correlación múltiple	0.967573241
Coefficiente de determinación R ²	0.936197977
R ² ajustado	0.929108863
Error típico	11.17620711
Observaciones	11

ANÁLISIS DE VARIANZA

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	16495.46792	16495.46792	132.0613574	1.11223E-06
Residuos	9	1124.168448	124.9076054		
Total	10	17619.63636			

	Coefficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95,0%	Superior 95,0%
			Coeficiente / error típico					
Intercepción a	55.85267521	14.4912535	3.854233536	0.003881115	23.07115733	88.63419309	23.07115733	88.63419309
Vehículos Matriculados (0,000) - x	0.311962979	0.027146584	11.49179522	1.11223E-06	0.250553094	0.373372864	0.250553094	0.373372864
