UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ
ESTADÍSTICA APLICADA I: ESMA3101

REPASO EXAMEN FINAL
Diciembre de 2018.

Profesor: Didier A. Murillo Florez

I. Medidas estadísticas y gráficas:

1. Medidas de tendencia central:

La media: En estadística la media o media aritmética es una medida sobre la centralidad de un conjunto de datos. Sí n es una muestra aleatoria extraída de una población de tamaño N y X una variable entonces $\bar{X}$ se denota para la media muestral. Donde $\bar{X}$ se calcula de la siguiente manera:

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

Se espera que la media esté relacionada con el valor que más posibilidades tiene de ocurrir en una variable, por tanto si es así, la media es el centro de la distribución.

Para referirnos a la media o promedio poblacional usamos la letra griega μ.

La mediana: La mediana es una medida estadística de tendencia central que representa el 50 % del conjunto de datos.

La moda: En un conjunto de datos, la moda se refiere a la observación o datos que tiene mayor frecuencia, es decir; el dato que aparece más veces. La moda no es única y algunas ocasiones no existe moda.

Ejemplo: La siguiente muestra aleatoria representa la edad de 9 estudiantes de ESMA3101.
19 22 22 20 22 21 20 22 24
Calcular e interpretar la media, mediana y moda.
La media:

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{19 + 22 + 22 + 20 + 22 + 21 + 20 + 22 + 24}{9} = 21.33$$

En promedio los nueve estudiantes de ESMA3101 tienen 21.33 años.

La mediana: En este caso tenemos un conjunto de datos con 9 observaciones (impar). Entonces la mediana será el valor que divide el conjunto de datos ordenado en partes iguales.

Los datos ordenados son: **19 20 20 21 22 22 22 22 24** y la mediana el dato que está de color rojo. Dese cuenta que es la mitad del conjunto de datos.

El 50 % de los estudiantes muestreados tienen entre 19 y 22 años.

La moda: En este caso la moda es el dato 22.

La edad más frecuente entre los nueve estudiantes es 22 años.

Mediante una muestra aleatoria y los valores de la media, mediana y moda podemos tener alguna idea de la forma de la distribución de la población.
Distribuciones simétricas: La media, mediana y moda son iguales y representan el centro de la distribución.
Distribuciones sesgadas a derecha: La media es mayor que la mediana.
Distribuciones sesgadas a izquierda: La media es menor que la mediana.

2. Medidas de variabilidad:

   La Varianza: En estadística la varianza se refiere al cuadrado medio de la distancia entre cada una de las observaciones y el valor del promedio o media. Es una medida que nos indica que tanto se alejan los datos del centro de la distribución (media) o qué tan dispersos están los datos. La varianza se calcula de la siguiente manera y se denota por la letra S^2 si es calculada a partir de una muestra.

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

   La desviación estándar: La desviación estándar (raíz cuadrada de la varianza) es una medida de dispersión alternativa, expresada en las mismas unidades que los datos de la variable objeto de estudio. La desviación estándar S es la raíz cuadrada de la varianza

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

   Ejemplo: Continuando con la muestra aleatoria de la edad de 9 estudiantes de ESMA3101.
   19 22 22 20 22 21 20 22 24
   Calcular e interpretar la Varianza y desviación estándar.
   La varianza:

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{(19-21.33)^2 + (22-21.33)^2 + (22-21.33)^2 + ... + (24-21.33)^2}{9-1} = 2.25$$

   Entonces; la desviación media al cuadrado respecto a la media es de 2.25 años.
   La desviación estándar: Tan solo debemos calcular la raíz cuadrada de S^2. Entonces,

$$S = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} = \sqrt{2.25} = 1.5$$

   Entonces; la desviación media respecto a la media es de 1.5 años.

   Con los anteriores resultados nos podemos dar cuenta que los datos están muy cerca del centro o promedio de la distribución. Los datos no están dispersos o tiene poca dispersión. Tampoco se evidencia presencia de datos atípicos o outliers.

3. Medidas de posición.
   Los cuartiles: Los cuartiles son tres valores que dividen la muestra en cuatro partes iguales. El primer cuartil Q_1 representa el 25 % de los datos, el segundo cuartil Q_2 representa el 50 % y el tercer cuartil representa el 75 % de los datos. Al final obtenemos cuatro partes iguales y equivalentes al 25 %. El segundo cuartil siempre coincide con la mediana.
   Para calcular correctamente los cuartiles se debe ordenar de menor a mayor el conjunto de datos.

   Ejemplo: Siguiendo con los 9 datos de la edad. Calcular los cuartiles Q_1, Q_2 y Q_3.

   Los datos ordenados son: **19 20 20 21 22 22 22 22 24**..

   Vamos a representar con color azul, rojo y verde a los datos que debemos tener presentes para calcular los cuartiles:

   19 20 20 21 22 22 22 22 24

   $Q_1 = \frac{20+20}{2} = 20$.Es decir, el 25 % de los estudiantes tiene 20 años

   Q_2= Mediana = 22. Es decir, el 50 % de los estudiantes tiene 22 años

   $Q_3 = \frac{22+22}{2} = 22$. Es decir, el 75 % de los estudiantes tiene 22 años

El rango Intercuartílico: Se refiere al 50 % central de los datos. Se puede usar como una medida de variabilidad en caso de presencia de outliers. Sa calcula como:

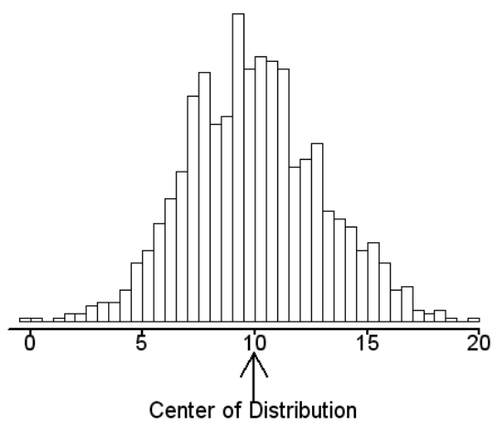
$$IQR = Q_3 - Q_1$$

Entonces el 50 % central de los datos o IQR para la edad de los 9 estudiantes es:

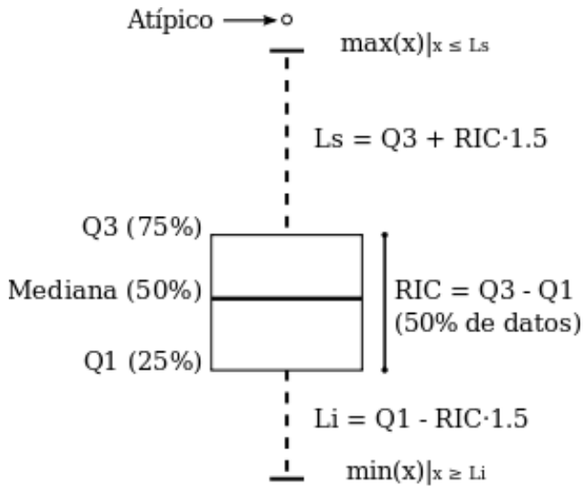
$$IQR = 22 - 20 = 2$$

4. Gráficas

El histograma. El histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. Sirven para obtener una "primera vista"general, o panorama, de la distribución de la población, o de la muestra, respecto a una característica, cuantitativa y continua (como la estatura o el peso)



Boxplot: El "Boxplot", al igual que el histograma permite tener una idea visual de la distribución de los datos. O sea, determinar si hay simetría, ver el grado de variabilidad existente y finalmente detectar "outliers". Pero además, el "Boxplot" es útil para comparar grupos.



Interpretación: La línea central de la caja representa la Mediana y los lados de la caja representan los cuartiles Q1 y Q3 respectivamente. Si la Mediana está bien al centro de la caja, entonces hay simetría. Si la Mediana está más cerca a Q3 que a Q1 entonces la distribución es sesgada a la izquierda, de lo contrario la distribución es sesgada a derecha. Si la caja no es muy alargada entonces se dice que no hay mucha variabilidad, es decir entre más larga la caja es mayor la variabilidad en los datos.

Si no hay "outliers" entonces las líneas laterales de la caja llegan hasta el valor mínimo por abajo, y hasta el valor máximo por arriba. Cuando hay "outliers" entonces éstos aparecen identificados en la figura y las líneas laterales llegan hasta los valores adyacentes a las fronteras interiores.

II. Probabilidad

Definición: La probabilidad es una forma de medir la incertidumbre asociada a un suceso o evento futuro. La probabilidad se expresa como un número entre 0 y 1.

Hay tres formas de calcular probabilidades:

1. **Probabilidad axiomática:** Consiste en definir tres axiomas de probabilidad y a partir de estos desarrollar una teoría que cumpla dichos axiomas.
   - Axioma 1: $P(S) = 1$
   - Axioma 2: $0 \leq P(A) \leq 1$
   - Axioma 3: $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$, donde los eventos A_i son mutuamente excluyentes.

   Algunas propiedades importantes:

   a) Sí A es un evento con probabilidad $P(A)$, entonces la probabilidad del complemento es $P(A^c) = 1 - P(A)$

   b) Dos o más eventos son mutuamente excluyentes si la intersección entre ellos es vacía. Por tanto la probabilidad es cero. Para el caso de dos eventos que son mutuamente excluyentes se cumple que $P(A \cap B) = 0$

   c) La propiedad de la adición; $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Dese cuenta que si A y B son eventos mutuamente excluyentes, entonces: $P(A \cup B) = P(A) + P(B)$

   d) Si dos eventos A y B son independientes entonces; $P(A \cap B) = P(A) * P(B)$

2. Método clásico o teórico:

   Definición: Si un experimento aleatorio tiene un espacio muestral equiprobable S que contiene N elementos y A es un evento de S que ocurre de $\#(A)$ maneras distintas entonces la probabilidad de ocurrencia de A está dada por:

$$P(A) = \frac{\#(A)}{N}$$

3. Método frecuentista:

   Definición: Si un experimento se repite n veces y $n(A)$ de esa veces ocurre el evento A, entonces la frecuencia relativa de A se define como:

$$\frac{n(A)}{n} \to P(A)$$

   La probabilidad es el valor en el cual se estabiliza la frecuencia relativa del evento después de haber repetido el experimento un número grande de veces. La existencia de este valor está garantizando por un resultado llamado La Ley de los Grandes números. Desde el punto de vista práctico se puede considerar que la frecuencia relativa de un evento es un estimado de la probabilidad de ocurrencia del evento.

III. Variables aleatorias

Definición: Una variable aleatoria es una función que toma los valores del espacio muestral y los transforma en números reales. La variable aleatoria es discreta si esta toma un número finito o un número contable de valores.

Ejemplo 1: Suponga el lanzamiento de tres monedas, sea la variable aleatoria $X :=$"número de caras". En este caso, el espacio muestral está definido por

$$S = \{CCC, CC+, C++, C+C, +C+, ++C, +CC, +++\}$$

Por ejemplo, cuando evaluamos a X en S para el elemento CCC encontramos que $X(CCC) = 3$, donde 3 es un número real.

Vemos que la variable aleatoria X puede tomar los siguientes valores $\{0, 1, 2, 3\}$. Estos valores corresponden al soporte de la variable aleatoria.

Variable aleatoria Binomial:

Definición: Una variable aleatoria Binomial cuenta el número de éxitos de un determinado experimento que es repetido un número finito de veces n. Si X es una variable aleatoria Binomial, se escribe $X \sim Binomial(n, p)$. p es la probabilidad de éxito.

La función de probabilidad para la variable aleatoria $Binomial(n, p)$ esta dada por:

$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, con $x = 0, 1, 2, ..., n$

donde;

- n es el número de ensayos o repeticiones del experimento.
- x es el número de éxitos.
- p es la probabilidad de éxito
- $1 - p$ es la probabilidad de fracaso.

Ejemplo: Si $X \sim Binomial(18, 0.4)$. Encuentre:

a) $P(X = 3)$
b) $P(X \geq 2)$

Solo debemos aplicar de función de probabilidad,

a)

$$P(X = 3) = \binom{18}{3} 0.4^3 (1 - 0.4)^{18-3} = 0.0245$$

b)

$$\begin{aligned} P(X \geq 2) = 1 - P(X < 2) &= 1 - \left[\binom{18}{0} 0.4^0 (1 - 0.4)^{18-0} + \binom{18}{1} 0.4^1 (1 - 0.4)^{18-1} \right] \\ &= 1 - \left[0.00010156 + 0.001218719 \right] \\ &= 1 - 0.001320279 \\ &= 0.9986 \end{aligned}$$

Variable aleatoria Poisson:

Definición: La distribución de Poisson es la distribución de probabilidad que modela el número de ocurrencias de eventos independientes en un intervalo.
La función de probabilidad para la variable aleatoria $Poisson(\lambda)$ esta dada por:

$P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$, con $x = 0, 1, 2, ...$
donde;

- λ es el promedio.
- x es el número de eventos independientes.

IV. Distribución Normal

Definición: Si X es una variable aleatoria continua que se distribuye normalmente entonces se escribe $X \sim Normal(\mu, \sigma)$, donde μ es la media y σ es la desviación estándar.

V. Distribución Normal Estándar

Definición: A la distribución Normal que tiene media $\mu = 0$ y desviación estándar $\sigma = 1$ se le llama normal estándar y se representa por la letra Z. Se escribe $Z \sim Normal(0, 1)$. Las áreas bajo la curva normal representan las probabilidades estándar y se pueden encontrar en tablas o usando algún programa estadístico como `RStudio`.

Se puede pasar de de una distribución normal con media μ y desviación estándar σ a una distribución Normal estándar. El proceso se llama estandarización. Donde Z se calcula como:

$$Z = \frac{X - \mu}{\sigma}$$

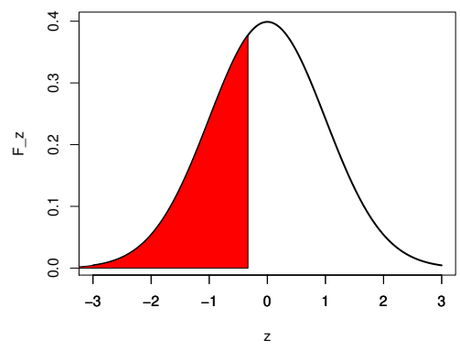
Donde X es el valor que toma la variable aleatoria $Normal(\mu, \sigma)$.

Ejemplos: Si $X \sim Normal(5, 6)$, mediante estandarización encontrar las siguientes probabilidades:

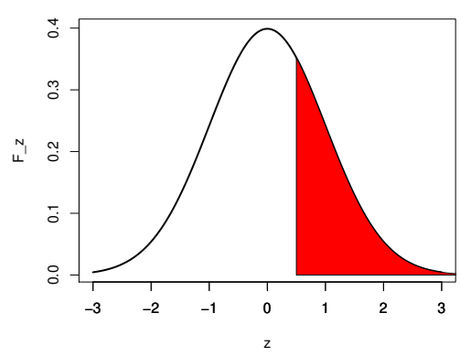
a) $P(X < 3)$
b) $P(X > 8)$

Entonces,

a) $P(X < 3) = P\left(Z < \dfrac{3-5}{6}\right) = P(Z < -0.333) = 0.3695$



b) $P(X > 8) = 1 - P(X < 8) = 1 - P\left(Z < \dfrac{8-5}{6}\right) = 1 - P(Z < 0.5) = 1 - 0.6914 = 0.3086$



VI. Distribuciones muestrales.

Teorema del Límite Central (TLC): Si se toman varias muestras de tamaño $n \geq 30$ de una población con media μ y desviación estándar σ, el TLC garantiza que la distribución del promedio de cada una de las muestras tiene distribución Normal con media $\mu = \mu_{\bar{X}}$ y desviación estándar $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$. En conclusión,

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Para describir la distribución del promedio muestral podemos tener dos casos:

a) Caso 1: Si se toman muestras de tamaño n de una población que tiene distribución $Normal(\mu, \sigma)$. Entonces la distribución del promedio muestral $\bar{X}$ es:

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

En este caso no es necesario usar el TLC. Porque no importa el tamaño de la muestra cuando la población es normal con parámetros conocidos.

b) Caso 2: Si se toman muestras de tamaño $n \geq 30$ de una población que tiene distribución desconocida. Entonces por el TLC la distribución del promedio muestral $\bar{X}$ es:

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

con $\mu = \mu_{\bar{X}}$ y $n \geq 30$.

VII. Estimación puntual.

En estadística siempre estamos interesados en aproximar o estimar el valor de ciertos parámetros de la población que son desconocidos. El camino clásico de hacer esto es siempre tomando una muestra aleatoria y usando algunos estadísticos en la muestra para estimar el valor de sus respectivos parámetros en la población.

El caso más frecuente es querer estimar el promedio poblacional. Para hacer esto usamos el estadístico de la media, es decir $\bar{X}$. El proceso siempre es; usar los datos de muestra aleatoria, calcular $\bar{X}$ y luego hacer inferencia sobre μ usando el valor de $\bar{X}$.

Ejemplo: Suponga una muestra aleatoria del ingreso anual de 20 familias en Mayagüez. Los datos de la muestra son:

64355 67888 53964 68500 68775 75814 84733 64962 60401 60256 70695 58847 80081 88281 54878 79018 49674 53583 60434 85589

Estamos interesados en hacer inferencia sobre el ingreso anual de todas las familias en Mayagüez. Es simple, una estimación puntual del ingreso anual es $\bar{X}$, por lo tanto solo resta calcular $\bar{X}$.

$$\bar{X} = \sum_{i=1}^{20} \frac{x_i}{n} = \frac{64355 + 67888 + 53964 + ... + 85589}{20} = 67536.4$$

Entonces, una estimación para el ingresos anual promedio de todas las familias en Mayagüez está dado por 67536.4. Por supuesto que puede haber un error de estimación, quizá el ingreso anual promedio de todas las familias sea levemente mayor o menor. Pero la estimación puntual no considera el error!

También se puede hacer estimación puntual con otros estadísticos, tales como; $\hat{p}, S$,...

VIII. Estimación por intervalos de confianza.

Quizá no estamos muy conformes con la estimación puntual, pues sabemos que existe un error de estimación y lo queremos tener en cuenta. Una forma de tener presente ese error es construir intervalos con cierta confianza de que el parámetro que estamos estimando estará dentro del intervalo.

1. Si estamos estimando el promedio poblacional μ, podemos tener dos casos:

   1) **Caso 1 (σ es conocida):** Conocemos el valor de la desviación estándar poblacional σ, tenemos una muestra aleatoria de donde calculamos $\bar{X}$. Entonces un intervalo con el $(1-\alpha)*100\,\%$ de confianza esta dado por:

   $$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

   2) **Caso 2: (σ es desconocida):** Desconocemos el valor de la desviación estándar poblacional σ, tenemos una muestra aleatoria de donde calculamos S y $\bar{X}$. Entonces un intervalo con el $(1-\alpha)*100\,\%$ de confianza esta dado por:

   $$\bar{X} \pm t_{(\frac{\alpha}{2}, n-1)} \frac{S}{\sqrt{n}}$$

   Ejemplo: Sí asumimos que conocemos la desviación estándar de ingresos anual en todas las familias de Mayagüez, $\sigma = 12000$. Un intervalo con el 95 % de confianza para el ingreso anual promedio de todas las familias de Mayaguez está dado por:

   $$\Big(62277.26, 72795.54\Big)$$

   Interpretación: Estamos seguros con una confianza del 95 % de que el ingreso familiar promedio de todas las familias en Mayaguez esta entre 62277.26 y 72795.54 dolares.

   En este caso usamos el intervalo Z, porque conocíamos el valor de σ.

   Ejemplo: Sí desconocemos la desviación estándar del ingreso anual en todas las familias de Mayagüez. Con $S = 11532.7$ y $\bar{X} = 67536.4$ un intervalo con el 95 % de confianza para el ingreso anual promedio de todas las familias de Mayagüez está dado por:

   $$\Big(62138.93, 72933.87\Big)$$

Interpretación: Estamos seguros con una confianza del 95 % de que el ingreso familiar promedio de todas las familias en Mayaguez esta entre 62138.93 y 72933.87 dolares.

En este caso usamos el intervalo t-Student, porque desconocemos el valor de σ. En vez de σ usamos S.

En los dos intervalos anteriores usted debe verificar los resultados!

2. Intervalo de confianza para la proporción poblacional.

   Si X es el número total de éxitos en un total de n intentos, entonces una estimación puntual para p, está dada por $\hat{p} = \dfrac{X}{n}$.

   Ahora, si se cumple que:

   - $n\hat{p} \geq 10$
   - $n(1-\hat{p}) \geq 10$

   Entonces un intervalo de confianza para la proporción poblacional p, esta dado por:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

IX. Pruebas de Hipótesis

Formalismo en pruebas de hipótesis

Podemos organizar la metodología para realizar pruebas de hipótesis en unos simples pasos:

1. Definir el parámetro de interés y los datos del problema.
2. Plantear las hipótesis nula (H_0) y la hipótesis alterna (Ha)
3. Método de análisis y supuestos: Calcular el estadístico de prueba.
4. Encontrar el valor crítico de la distribución, luego especificar la región de rechazo y no rechazo.
5. Decisión y conclusión.

Pruebas de hipótesis para la media poblacional μ

Prueba a cola derecha	**Prueba a cola izquierda**	**Prueba a dos colas**
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu \neq \mu_0$

Al realizar pruebas de hipótesis para la media poblacional podemos encontrar dos casos:

a) **Caso 1 (σ es conocida):** Conocemos el valor de la desviación estándar poblacional σ, tenemos una muestra aleatoria de donde calculamos $\bar{X}$ o conocemos su valor de antemano. Entonces el estadístico de prueba bajo el supuesto que H_0 es cierta, esta dado por:

$$Z^* = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

b) **Caso 2: (σ es desconocida):** Desconocemos el valor de la desviación estándar poblacional σ, tenemos una muestra aleatoria de donde calculamos S y $\bar{X}$ o conocemos sus valores de antemano. Entonces el estadístico de prueba bajo el supuesto que H_0 es cierta, esta dado por:

$$t^* = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

Pruebas de hipótesis para la proporción poblacional p

Prueba a cola derecha	**Prueba a cola izquierda**	**Prueba a dos colas**
$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
$H_a : p > p_0$	$H_a : p < p_0$	$H_a : p \neq p_0$

El estadístico de prueba en una prueba de hipótesis para una proporción está dado por:

$$Z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Para ejemplos sobre pruebas de hipótesis puede consultar las notas de la clase y hacer los ejercicios de repaso!

El p-value: El p-value es la probabilidad de observar un valor mayor al estadístico de prueba bajo el supuesto de que la hipótesis nula H_0 es cierta.

X. Correlación

La ídea de la correlación es medir de alguna manera la relación que existe entre dos variables cuantitativas.

Para medir esa relación en terminos numéricos, se usa el coeficiente de correlación, el cual es una medida estadística similar a la media, mediana, desviación estándar, Q1, etc. El coeficiente de correlación:

Es una estadístico cuando se encuentra a partir de una muestra

Es un parámetro cuando pertenece a una población.

En el primer caso usamos el símbolo r, en el segundo caso usamos ρ.

Propiedades del coeficiente de correlación:

1. Siempre esta entre $-1 < r < 1$
2. r cercano a 0 significa muy poca o incluso ninguna correlación (relación)
3. r cercano a ± 1 significa una correlación muy fuerte $r = -1$ o $r = 1$ significa una correlación lineal perfecta (es decir, en el diagrama de dispersión los puntos forman una línea recta)
4. $r < 0$ significa una relación negativa (a medida que x se hace más grande y se hace más pequeño)
5. $r > 0$ significa una relación positiva (a medida que x aumenta, y aumenta)
6. r trata simétricamente x e y, es decir $cor(x, y) = cor(y, x)$

El coeficiente de correlación de Pearson solo mide las relaciones lineales, no funciona si una relación no es lineal y si se lesuma la presencia de outliers.

XI. Regresión lineal simple

En Estadística, un modelo de gresión lineal simple es una ecuación que intenta describir la relación lineal que existe entre una variable respuesta (dependiente) y una variable predictora (independiente):

$$y = \alpha + \beta x$$

Donde:

1. y es la variable respuesta. (variable dependiente)
2. x es la variable predictora. (variable independiente)
3. α es el $y-$ intercepto.
4. β es la pendiente de la recta.

La lógica aquí es esta: si sabemos el valor de x, podemos calcular el correspondiente valor de y. Lamentablemente, siempre hay errores en este cálculo, por lo que la respuesta y varía incluso para la misma x. Si el error es considerado, entonces el modelo se escribe de la siguiente manera:

$$y = \alpha + \beta x + \epsilon$$

Donde ϵ denota los errores. Para estos errores se supone son independientes idénticamente distribuidos Normal con media 0 y varianza σ^2. Es decir

$$\epsilon_i \sim Normal(0, \sigma^2)$$

Respecto a la línea de regresión;

Digamos que $\bar{X}$ es la media del vector x, y que y $\bar{Y}$ es la media del vector y, entonces $(\bar{X}, \bar{Y})$ es siempre un punto en la línea. Es decir la línea de regresión siempre pasa por el punto $(\bar{X}, \bar{Y})$.