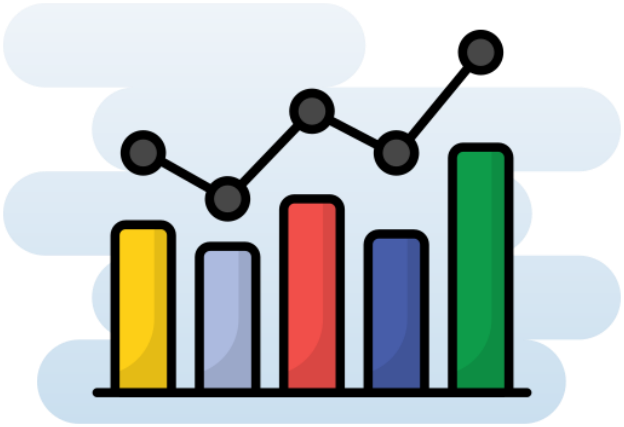
HERRAMIENTAS DE ESTADISTÍCA

PORTAFOLIO 1



Rubén Gaona Pérez

ÍNDICE

PORTAFOLIO 1

Usando las columnas 2 y 3 del excel DATOS_INFERENCIA obtenemos que la media y varianza poblacionales son:

μ = 84.32, σ^2 = 124.5

Y las muestrales son:

- Columna 2: $\bar{x}_2$= **107.7504267,** s_2^2 **= 589.2393091**
- Columna 3: $\bar{x}_3$= **106.48462,** s_3^2 **= 647.7452291**

Ejercicio 1:

En una carrera de 10 km se registraron los tiempos de llegada en minutos de 30 participantes en dos años distintos (2024 y 2025), los tiempos vienen dados por las columnas 2 y 3, respectivamente, del archivo DATOS_INFERENCIA. Con α=0.1 calcular:

1. Contrastar test de hipótesis (comparación de medias) con una de las columnas; varianza poblacional conocida. (Los parámetros se indican anteriormente al escoger las columnas. μ, σ2)

Usando la columna 3 tendremos:

$$H_0: \mu_3 = 84.32 \qquad H_1: \mu_3 > 84.32$$

Con α=0,1 calculamos el estadístico de contraste:

$$Z_0 = \frac{(\bar{x}_3 - \mu_3)\sqrt{n_3}}{\sigma_3} = 10.88$$

Por último, obtenemos la región de rechazo:

$$Z_{0,9} = 1.285$$

H_a	Aceptar H_0 si	Rechazar H_0 si
$\mu > \mu_0$	$K^0_{obs} \leq z_{1-\varepsilon}$	$K^0_{obs} > z_{1-\varepsilon}$

Como la prueba es unilateral y $Z_0 > Z_{1-\alpha}$, hay evidencia estadística suficiente para rechazar la hipótesis nula, con lo cual el tiempo medio de la carrera no es igual a 84.32.

2. Contrastar test de hipótesis para homogeneidad de varianzas (Ejemplo: datos de columna 1 y datos columna 2)

Comparamos las varianzas de ambas columnas:

$$H_0: \sigma_2^2 = \sigma_3^2 \quad H_1: \sigma_2^2 \neq \sigma_3^2$$

Con α=0,1 calculamos el estadístico de contraste:

$$f_0 = \frac{s_2^2}{s_3^2} = 0.90968$$

Por último, obtenemos la región de rechazo:

$$f_{29,\,29,\,0.05} = 1.85$$

H_a	Aceptar H_0 si	Rechazar H_0 si
$\sigma_X^2 \neq \sigma_Y^2$	$f_{n_X-1,n_Y-1,\frac{\varepsilon}{2}} \leq K_{\text{obs}}^0 \leq f_{n_X-1,n_Y-1,1-\frac{\varepsilon}{2}}$	$K_{\text{obs}}^0 < f_{n_X-1,n_Y-1,\frac{\varepsilon}{2}}$ o $K_{\text{obs}}^0 > f_{n_X-1,n_Y-1,1-\frac{\varepsilon}{2}}$

Como la prueba es bilateral y $f_0 < |f_{1-\alpha/2}|$, no hay evidencia estadística para rechazar la hipótesis nula, con lo cual consideramos varianzas homogéneas.

3. Contrastar test de hipótesis para comparación(diferencia) de dos medias (Ejemplo: datos de columna 1 y datos columna 2) de acuerdo al apartado (2.)

Usando las columnas 2 y 3 y considerando, según el apartado 2, que las varianzas son homogéneas, tendremos:

$$H_0: \mu_2 = \mu_3 \quad H_1: \mu_2 \neq \mu_3$$

Con α=0,1 calculamos el estadístico de contraste:

$$t_0 = \frac{\bar{x}_2 - \bar{x}_3}{s_p\sqrt{\frac{1}{n_2}+\frac{1}{n_3}}} = 0.19713$$

$$s_p = \sqrt{\frac{(n_2-1)s_2^2+(n_3-1)s_3^2}{n_2+n_3-2}} = 24.8695$$

Por último, obtenemos la región de rechazo:

$$t_{0,95} = 1.672$$

H_a	Aceptar H_0 si	Rechazar H_0 si
$\mu_X \neq \mu_Y$	$\lvert K_{\text{obs}}^0 \rvert \leq z_{1-\frac{\varepsilon}{2}}$	$\lvert K_{\text{obs}}^0 \rvert > z_{1-\frac{\varepsilon}{2}}$

Como la prueba es bilateral y $t_0 < |t_{n2+n3-2,\ 1-\alpha/2}|$, no hay evidencia estadística para rechazar la hipótesis nula, con lo cual consideramos ambas medias como iguales u homogéneas.

4. Usando el apartado (3.), encontrar un intervalo de confianza para la diferencia de medias poblacionales

Dado que en el apartado 2 hemos determinado que las varianzas eran homogéneas, calcularemos el intervalo de confianza:

$$(\bar{x}_2 - \bar{x}_3) \pm t_{n2+n3-2,1-\frac{\alpha}{2}} \sqrt{(\frac{1}{n_2} + \frac{1}{n_3})(\frac{(n_2-1)s_2^2 + (n_3-1)s_3^2}{n_2+n_3-2})} = 1.2658067 \pm 10.736377$$

Con $t_{n2+n3-2,\ 1-\alpha/2} = 1.672$

Intervalo: (-9.4705703, 12.0021837)

5. Encontrar un intervalo de confianza al para la desviación típica poblacional asumiendo que esta es desconocida (Columna usada en 1.)

El intervalo de la desviación típica poblacional asumiendo que esta es desconocida es el siguiente:

$$\left[\sqrt{\frac{(n_3-1)s^2}{x^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n_3-1)s^2}{x^2_{n-1,\alpha/2}}}\right] = [21.0094, 32.5699]$$

Con:

$X_{n-1,\ 1-\alpha/2} = 17{,}708$

$X_{n-1,\ \alpha/2} = 42{,}557$

Ejercicio 2.

Redactar un problema donde se relacionen como variables las dos columnas asignadas anteriormente realizando la estimación de un modelo de regresión lineal simple encontrando:

Un equipo de fútbol está investigando si el tiempo total de entrenamiento de alta intensidad (X o C2, en minutos acumulados en los 2 entrenamientos anteriores) influye sobre la frecuencia cardíaca media registrada en competición (C3, en latidos por minuto)

1. Recta de regresión de Y sobre X. Interpretar

Modelo de regresión simple

$$y_n = \beta_0 + \beta_1 * x_n + \varepsilon_n$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathrm{E} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_2 \end{pmatrix}$$

Si $\beta = (\mathrm{X}' * \mathrm{X})^{-1} * X' * Y$, tendremos

$$\beta = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} * \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{pmatrix} = \begin{pmatrix} 30 & 3232.51 \\ 3232.51 & 365392{,}573 \end{pmatrix}^{-1} * \begin{pmatrix} 3194{,}539 \\ 341406{,}936 \end{pmatrix},$$ obteniendo $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 124{,}1758 \\ -0{,}1642 \end{pmatrix}$

Con lo que la recta de regresión simple será:

$$\hat{y} = 123{,}1758 - 0{,}1642 * x$$

Dada la pendiente negativa, podemos deducir que el entrenamiento intenso realizado en las 2 sesiones previas a la competición se asocia con una disminución de la frecuencia cardíaca media en la competición. Cada minuto adicional de entrenamiento previo supone unos 0,164 latidos por minuto menos en la frecuencia cardíaca media.

El intercepto, con valor de 123,1758, será la frecuencia cardiaca media en la competición dada para un entrenamiento previo a la competición nulo, es decir, de 0 minutos.

2. Recta de regresión de X sobre Y . Interpretar

Mediante el mismo proceso, calculamos la recta de regresión de X sobre Y, obteniendo:

$$\beta = \begin{pmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} y_i^2 \end{pmatrix}^{-1} * \begin{pmatrix} \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i x_i \end{pmatrix}$$

$$= \begin{pmatrix} 30 & 3194{,}5386 \\ 3194{,}5386 & 358953{,}8405 \end{pmatrix}^{-1} * \begin{pmatrix} 3232{,}5128 \\ 341406{,}936 \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 123{,}6566 \\ -0{,}1494 \end{pmatrix}$$

La recta de regresión será:

$$\hat{x} = 123{,}6566 - 0{,}1494 * y$$

Nos indica una recta parecida a la de Y sobre X, obteniendo una pendiente negativa, asociando que una frecuencia cardiaca media menor en la competición con un entrenamiento previo mayor. Cada latido adicional menos en la frecuencia cardíaca media equivale a 0,1494 minutos más de entrenamiento previo.

El intercepto en este caso es de 123,6566, que indica el trabajo previo a la competición realizado para una frecuencia cardíaca de 0, parámetro que solo nos sirve para ajustar la recta.

3. Coeficiente de correlación. Interpretar

El coeficiente de correlación es $P = \pm\sqrt{R^2}$

$$R^2 = 1 - \frac{SCR}{SCT}$$

Donde la suma de los cuadrados de los residuos es: $SCR = Y' * Y - \left(X * \hat{\beta}\right)' * Y = (358953{,}8405) - (340630{,}0279) = 18323.8126$

Y la suma de los cuadrados totales es: $SCT = \sum_{i=1}^{n}(Y_i - \bar{Y}) = 18784.6116$

$$R^2 = 0.02453$$

El valor de R^2 nos indica que la variabilidad de la variable dependiente (Y o C3) se explica por la variable independiente (X o C2) solamente en un 2,45%, lo que quiere decir, que el modelo lineal no describe bien los datos, ya que el cambio de la variable dependiente dependerá de otros factores.

$P = -0.15662$ (usando el signo de β_1)

El valor de P nos hace saber que hay una relación negativa muy débil entre las variables, el valor es casi nulo, tenemos una baja asociación lineal.

Podemos concluir que no existe evidencia sólida que nos indique que un entrenamiento previo mayor tenga un impacto lineal grande sobre la frecuencia cardíaca media, por lo que no es un modelo útil y habría que tener en cuenta otras variables como: posición, edad, esfuerzo, etcétera.