ANÁLISIS DE DATOS CUANTITATIVOS – ESTADÍSTICA INFERENCIAL

Estadística Inferencial

- Permite probar hipótesis para poder generalizar resultados de la muestra a la población.
- Se utiliza para:
  - Probar hipótesis
  - Estimar parámetros de la población.

Nivel de Significancia

- Se refiere al nivel de la probabilidad de equivocarse.
- Se fija antes de probar hipótesis inferenciales.
- También se conoce como nivel alfa (α)
- Se expresa en términos de probabilidad.

Nivel de Confianza

- Nivel de confianza es el complemento del respectivo nivel de significancia (1-α).
- Por ejemplo un 95% de intervalo de confianza refleja un nivel de significancia de 0.05.
- Con un nivel de significancia de 0.05:
  - Un investigador tiene 95% de seguridad para generalizar sin equivocarse y sólo 5% en contra (0.95 + 0.05 = 1.00).
- Con un nivel de significancia de 0.01:
  - Un investigador tiene 99% en su favor y 1% en contra (0.99 y 0.01 = 1.00) para generalizar.

Pruebas de hipótesis

- Al probar hipótesis se pueden cometer errores debido a que en estadística se trabaja con estimaciones y nunca estaremos completamente seguros.
- Los resultados al probar una hipótesis pueden ser:
  - Aceptar una hipótesis verdadera (decisión correcta).
  - Rechazar una hipótesis falsa (decisión correcta).
  - Aceptar una hipótesis falsa (conocido como error del Tipo II o error beta).
  - Rechazar una hipótesis verdadera (conocido como error del Tipo I o error alfa).

Para reducir errores en las pruebas de hipótesis

- Elegir muestras representativas probabilísticas.
- Tener un mayor conocimiento de la población.
- Inspeccionar cuidadosamente los datos.
- Seleccionar de las pruebas estadísticas apropiadas.
- El valor p (p value) representa la probabilidad de cometer un error tipo I (rechazar la hipótesis nula cuando es verdadera).
  - Mientras menor es el valor de p, menor es la probabilidad de rechazar la hipótesis nula incorrectamente.

Tipos de análisis estadísticos para probar hipótesis

- Paramétricos
- No paramétricos

Analisis estadístico paramétrico

- La distribución de la población de la VD es normal.
- Las variables son de tipo intervalo o radio.
- Cuando se estudian dos o más poblaciones, estas tienen una varianza homogénea (dispersión similar en las distribuciones).

Métodos de analisis estadístico paramétrico más usados

- Coeficiente de correlación de Pearson y regresión lineal.
- Prueba t.
- Análisis de varianza unidireccional (ANOVA en un sentido o oneway).
- Análisis de varianza factorial (ANOVA).

Coeficiente de correlación de Pearson

- Se relacionan las puntuaciones recolectadas de una variable con las puntuaciones obtenidas de la otra, con los mismos participantes o casos.
- El coeficiente de correlación es positivo o negativo.
- El coeficiente de correlación toma valores entre 0 y 1.
  - Mientras su valor está más cerca de 1, más fuerte la correlación.
  - Mientras su valor es más cercano a 0 menor es la correlación.
- El coeficiente de determinación r^2 explica el porcentaje de la variación de cierta variable debido a la variación de otra variable y viceversa.

Correlación de Pearson

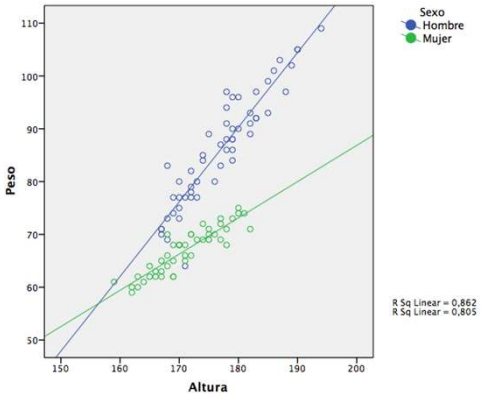
Se realizó una prueba de hipótesis, para comprobar la existencia de una correlación entre peso y estatura.

```
> cor.test(misdatos$Peso,misdatos$Estatura)

        Pearson's product-moment correlation

data:  misdatos$Peso and misdatos$Estatura
t = 11.8494, df = 18, p-value = 6.182e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8552787 0.9769678
sample estimates:
      cor
0.9414723
```

Los resultado de la prueba fueron los siguientes: Existe una correlación positiva entre peso y estatura (r=0.941) y es estadísticamente significativa (p<=0.05)

Regresión Lineal

- Es un modelo matemático para estimar el efecto de una variable sobre la otra.
- Asociado al coeficiente de correlación de Pearson **r**
  - Mientras mayor es r, mayor es la capacidad de predicción.
- Hipótesis a probar: correlacionales y causales.
- Variables: Una independiente y otra dependiente.
- Conociendo la recta de regresión y la tendencia, podemos predecir los valores de una variable conociendo los de la otra variable.



Regresión Lineal

- La ecuación de regresión lineal es: Y = a + bX.
  - Y es un valor de la variable dependiente
  - a es la intersección con el eje vertical
  - b la pendiente de inclinación.
  - X es el valor que fijamos de la variable independiente.
  - Por lo general los valores de a y b son proporcionados por los programas estadísticos.

Regresión lineal en R

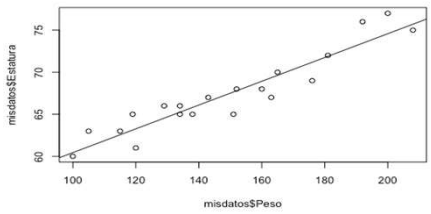
- > misdatos<-read.csv("Peso_Estatura.csv")
- > regl <- lm(misdatos$Estatura~misdatos$Peso)
- > par(cex=.8)
- > plot(misdatos$Peso, misdatos$Estatura)
- > abline(regl)

Dataset

Contenido de Peso_Estatura.csv

Peso,Estatura
100,60
120,61
105,63
115,63
119,65
134,65
129,66
143,67
151,65
163,67
160,68
176,69
165,70
181,72
192,76
208,75
200,77
152,68
134,66
138,65

Regresión lineal en R



Prueba t

- Es una prueba estadística para evaluar si dos grupos difieren entre si de manera significativa respecto a sus medias.
- Se simboliza por "t".
- Hipótesis a probar:
  - Hipótesis alternativa (Ha): Los grupos difieren significativamente entre sí.
  - Hipótesis nula (Ho): Los grupos no difieren significativamente.
- Función en R: t.test(x,y)

Ejercicio

En una academia de inglés se quiere probar si el uso de una aplicación de m-Learning puede complementar la enseñanza del profesor y mejorar el rendimiento de el examen que se toma al final del módulo de inglés básico. La academia de inglés tiene dos paralelos de inglés básico. El profesor del paralelo 1 complementó la enseñanza utilizando la herramienta de m-Learning, mientras que en el paralelo 2 el profesor usó los métodos convencionales de enseñanza. Se seleccionan al azar 10 estudiantes del paralelo 1 y 10 estudiantes del paralelo 2. El examen final tuvo 30 preguntas y el número de preguntas respondidas correctamente por los estudiantes se encuentran en la siguiente tabla.

Dataset

Paralelo 1	Paralelo 2
23	15
18	20
25	21
22	15
20	14
24	16
21	18
24	19
21	14
22	17

Ejercicio

- **Identifique una pregunta de investigación.**

¿Puede el uso de una aplicación de m-Learning complementar la enseñanza del profesor y mejorar el rendimiento del examen que se toma al final del módulo de inglés básico?

- **Escriba la hipótesis de investigación que respondería la pregunta.**

El uso de la herramienta de m-Learning mejora el rendimiento de los estudiantes en el examen final del módulo de inglés básico.

Ejercicio

- **Escriba la hipótesis estadística que utilizaría para representar la hipótesis**

Considerando que se nos ha dado un número de preguntas respondidas correctamente para dos muestras de estudiantes de cursos de inglés. Un grupo recibe la intervención del uso de la herramienta mLearning mientras que el otro grupo no. Asumiendo que ambas poblaciones son independientes y normalmente distribuidas. Si la población de estudiantes de inglés que usó la herramienta de mLearning tiene una media del número de preguntas respondidas correctamente en el examen final de μ_1 y la población de estudiantes de inglés que no usó la herramienta de mLearning tiene una media del número de preguntas respondidas correctamente en el examen final de μ_2, entonces podemos expresar las hipótesis nula y alternativa de la siguiente forma:

- Hipótesis Nula (H_0): $\mu_1 <= \mu_2$
- Hipótesis Alternativa (H_o): $\mu_1 > \mu_2$

Ejercicio

- **Identifique la prueba de hipótesis y el comando en R que usará para realizar la prueba. Justifique su elección.**

```
> paralelo1 = c(23, 18, 25, 22, 20, 24, 21, 24, 21, 22)
> paralelo2 = c(15, 20, 21, 15, 14, 16, 18, 19, 14, 17)
> t.test(paralelo1,paralelo2,alternative="greater")
```

Se usa una prueba de hipótesis t, debido a que se quiere hacer una comparación de medias para poder determinar si la herramienta de mLearning mejora o no el rendimiento de los estudiantes en un examen final de inglés.

Ejercicio

- **Describa los resultados obtenidos y llegue a una conclusión en relación a la hipótesis probada.**

Welch Two Sample t-test

data: paralelo1 and paralelo2

t = 4.9151, df = 17.469, p-value = 6.067e-05

alternative hypothesis: true difference in means is greater than 0 95 percent confidence interval:

3.297722 Inf

sample estimates: mean of x mean of y

22.0 16.9

El valor p es menor a 0.05 por tal motivo se rechaza la hipótesis nula. Existe evidencia de un incremento en la media del número de respuestas correctas en el examen final de Inglés cuando se usa la herramienta de mLearning.

ANOVA

- Es una prueba estadística para analizar si más de dos grupos difieren significativamente entre sí en cuanto a sus medias y varianzas.
- La prueba t es utilizada para dos grupos mientras que ANOVA se usa para tres o más grupos.
- Hipótesis a probar:
  - Hipótesis alternativa (Ha): hay diferencia entre dos o más grupos, los grupos difieren significativamente entre si.
  - Hipótesis nula (Ho): los grupos no difieren entre si.

Estadística no paramétrica

- Se usan para muestras con distribuciones no normales.
- Se usan para datos con escalas de medición nominales u ordinales.

Métodos o pruebas estadísticas no paramédicas

- Chi cuadrada
- Los coeficientes de correlación para rangos ordenados de Spearman y Kendall.

Chi cuadrada χ^2

- Evalua la hipótesis para saber si hay una relación entre dos variables categóricas.
- Hipótesis a probar: Si existe una correlación (si hay una relación entre dos variables).
- Variables involucradas: Dos
- Nivel de medición de las variables: Nominal u ordinal.
- La Chi cuadrada se calcula a través de:
  - Una tabla de contingencia o tabulación cruzada de dos dimensiones.
    - Cada dimensión representa una variable.

Ejercicio

Se ha observado en un gimnasio que aparentemente más mujeres tienden a hacer ejercicios juntas mientras que más hombres tienden a hacer ejercicios solos. Los siguientes datos fueron recolectados:

	Hombres	Mujeres
En grupo	12	24
Solos	22	10

Le han pedido a usted que determine si esta diferencia es significante.

Ejercicio

- Identifique la pregunta de investigación.
- Escriba la hipótesis estadística (hipótesis alternativa e hipótesis nula).
- Realice el test estadístico correspondiente, describa los resultados obtenidos y llegue a una conclusión en relación a la hipótesis probada.

Ejercicio

- **Identifique la pregunta de investigación.**

¿Es cierto que las mujeres tienden a hacer ejercicios en grupo y los hombres tienden a hacer ejercicios solos?

Ejercicio

- **Escriba la hipótesis estadística (hipótesis alternativa e hipótesis nula).**
- Hipótesis Nula (H_o): No hay diferencia en las preferencias de hacer ejercicios para mujeres y hombres.
- Hipótesis alternativa (H_a): Hay una diferencia entre las preferencias de hacer ejercicios para mujeres y hombres.

Ejercicio

La prueba chi cuadrado de independencia.

```
> matriz = matrix(c(24,12, 10, 22), ncol=2)
> chisq.test(matriz,correct=F)

        Pearson's Chi-squared test

data:  matriz
X-squared = 8.5, df = 1, p-value = 0.003551
```

Ejercicio

- **Describa los resultados obtenidos y llegue a una conclusión en relación a la hipótesis probada.**

El valor de p es menor a 0.05 por tal motivo se rechaza la hipótesis nula. Hay una diferencia significativa en las preferencias de hombres y mujeres para hacer ejercicio. Las mujeres prefieren ejercitar en grupo más que los hombres.