Examen de Conocimientos Básicos de Estadística

Estudiante: Rut Lay Abad Rodríguez

I. Fundamentos de Estadística (20 puntos)

1. (4 p) **Definiciones**

   a. Define “población” y “muestra”.

   Población: Se refiere a todo el conjunto de personas, cosas o datos que queremos estudiar en general. Muestra: es una parte más pequeña de la población que se usa para hacer un estudio.

   Ej: Las mujeres cubanas que presentan lupus es una población con características específicas, pero como es dificil estudarlas a todas se toma una muestra de esta población como por ejemplo las mujeres cubanas con lupus atendidas en un hospital específico.

   b. Explica la diferencia entre “parámetro” y “estadístico”.

   Parámetro: Es un dato o medida que describe a toda la población. Sería, por ejemplo, el promedio de edad de todas las mujeres cubanas con lupus. Como incluye a toda la población, es un dato exacto, pero difícil de obtener.

   Estadístico: Es un dato o medida que describe a una muestra. Sería el promedio de edad de las mujeres con lupus que se atienden en ese hospital específico. Como es un cálculo hecho con la muestra, nos ayuda a estimar cómo es ese dato en toda la población.

   Estudiamos una muestra (mujeres atendidas en un hospital) para obtener estadísticos, y con eso tratamos de entender cómo es la población completa (todas las mujeres cubanas con lupus), que tiene parámetros.

2. (4 p) **Tipos de variables**

   a. ¿Cómo distingues una variable cuantitativa de una cualitativa?

   b. Da un ejemplo de cada una.

   Variable cuantitativa: Es la que se puede medir con números. Nos da una cantidad, como peso, edad o número de hijos. Se puede sumar o hacer cálculos coherentes con ella.

Variable cualitativa: Es la que describe una característica o una cualidad. No se mide con números, sino con palabras, como el color de ojos, el tipo de tratamiento o el grupo sanguíneo. Solo puedes clasificarla o describirla.

3. (6 p) **Medidas de tendencia central**

   Dado el conjunto de datos: {5, 7, 9, 10, 7, 12, 5, 9}

   a. Calcula la media, la mediana y la moda.

   b. ¿Qué te indica cada una sobre la distribución?

   Media: Nos da una idea del promedio general de los datos. Se calcula como la suma de todos los datos dividido ente la cantidad de datos:
   $(5 + 7 + 9 + 10 + 7 + 12 + 5 + 9)/8 = 64/8 = 8$

   Mediana: Nos dice el punto central de la distribución cuando los datos están ordenados. En este caso ordenamos: $5, 5, 7, 7, 9, 9, 10, 12$ Como la cantidad de datos es par, se escogen los dos valores del medio, y se calcula el promedio:
   $(7 + 9)/2 = 16/2 = 8$ Si la distribución dubiera un número de datos impar, solo sería escoger el valor central luego de que se hubiesen ordenado los datos.

   Moda: Nos muestra los valores más comunes (los que más se repiten). En esta distribución hay tres valores que se repiten con la misma frecuencia: 5,7 y 9 pol lo tanto esta distribución es multimodal (tiene múltiples modas)

4. (6 p) **Medidas de dispersión**

   Con los mismos datos del ítem anterior, calcula:

   a. La varianza muestral.

   La varianza se calcula de la siguiente manera:

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

   donde: s^2 es la varianza muestral, x_i son los datos, $\bar{x}$ es la media y n es el número de datos.

   Como anteriormente calculamos la media=8 y el número de datos= 8 también, el cálculo de la varinza será igual a:

   $$s^2 = \frac{1}{8-1}\sum_{i=1}^{8}(x_i - 8)^2$$

   Las diferencias al cuadrado para cada dato serán:

|x_i | $x_i - 8$ |$(x_i - 8)^2$| |-------|----------|------------| |5 | -3 | 9 | |5 | -3 | 9 | |7 | -1 | 1 | |7 | -1 | 1 | |9 | 1 | 1 | |9 | 1 | 1 | |10 | 2 | 4 | |12 | 4 | 16 |

$$s^2 = \frac{1}{7}(9 * 2 + 4 * 1 + 4 + 16) = \frac{42}{7} = 6$$

b. La desviación estándar muestral.

La desviación estándar de la muestra se calcula como la raíz cuadrada de la varianza por lo que se calcula como:

$$s = \sqrt{s^2}$$

y para la distribución estudiada tendría un valor de:

$$s = \sqrt{6} \approx 2.45$$

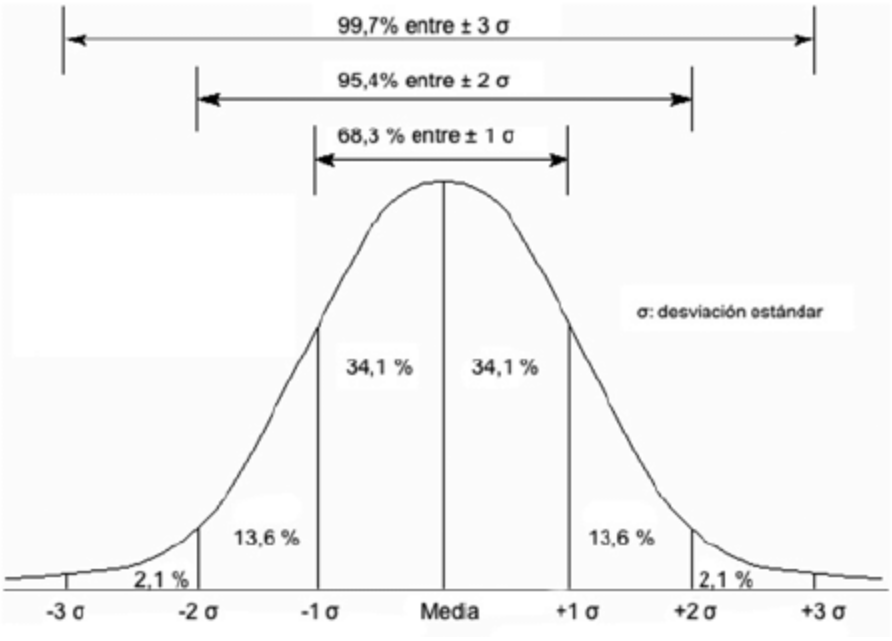
II. Probabilidad y Distribuciones (20 puntos)

5. (5 p) **Distribución normal**

   a. Describe sus principales propiedades.

   Dentro de las principales propiedades de una distribución normal se encuentran: Tiene una forma simétrica parecida a una campana, es simétrica respecto a la media pues la mitad de los datos se encuantran a la derecha de la media y la otra a la izquierda. En este tipo de distribución la media, mediana y moda suele coincidir en el centro de la misma. Además de la media que indica el centro de la distribución, es importante conocer en esta la desviación estándar que da una medida de qué tan dispersos están los datos. Por último, la curva normal nunca toca al eje de las x.

   b. ¿Qué porcentaje de datos cae dentro de ±1 σ de la media en una normal teórica?



[Pedro Romero-Aroca, Carlos García y Julio González-López. "Estadística Descriptiva e Inferencial". ISBN:978-84-90085-51-0.]

Como se observa en la figura, cerca del 68% de los datos cae en dentro de de ±1 σ de la media en una normal teórica.

F. (5 p) **Distribución binomial**

Un proceso tiene probabilidad de éxito $p = 0.2$ en cada ensayo. Si realizas $n = 10$ ensayos,

a. ¿Cuál es la probabilidad de obtener exactamente 3 éxitos?

Para la distribución binomial donde:

- (n) es el número de ensayos,
- (p) es la probabilidad de éxito en cada ensayo,
- (k) es el número exacto de éxitos.

La fórmula para la probabilidad de obtener exactamente (k) éxitos:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Donde:

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

Para el inciso a. tenemos:

$$\binom{10}{3} = \frac{10!}{3!(10-3)!} = 120$$

Por lo que:

$$P(X = 3) = \binom{10}{3} \cdot 0.2^3 \cdot (1 - 0.2)^{10-3} = 120 * 0.008 * 0.2097 = 0.2013$$

R/ La probabilidad de obtener exactamente 3 éxitos con una probabilidad de 0.2 y 10 ensayos es de 0.2013 o 20.13%.

b. ¿Cuál es la probabilidad de obtener al menos 1 éxito? Para calcular la probabilidad de tener al menos un éxito tenemos que calcular la probabilidad de tener 0 éxitos y se la restamos a 1.

$$P(X \geq 1) = 1 - P(X = 0)$$

$$\binom{10}{0} = \frac{10!}{0!(10-0)!} = 1$$

$$P(X = 0) = \binom{10}{0} \cdot 0.2^0 \cdot (1 - 0.2)^{10-0} = 1 * 1 * 0.8^{10} \approx 0.1074$$

R/ La probabilidad de que ocurra al menos un éxito es de 1-0.1074 $\approx$ 0.8926 o 89.26%.

6. (10 p) **Teorema central del límite**

a. Enuncia el teorema central del límite.

Si tomamos muchas muestras grandes, la distribución de las medias muestrales se vuelve aproximadamente normal, aunque los datos originales no sean normales.

b. ¿Por qué es importante para la inferencia estadística?

Este teorema y su aplicación nos permite trabajar con promedios de muestras usando la distribución normal, incluso si los datos originales no tienen forma de campana (no son normales). Nos permite utilizar las distribuciones normales para hacer inferencias aunque los datos no sean normales, siempre que la muestra sea lo suficientemente grande.

c. Describe un ejemplo práctico en el que lo aplicas para construir un intervalo de confianza de la media.

Ejemplo práctico: intervalo de confianza para la media empleando TLC

Supongamos que se estudia el tiempo promedio que tardan las mujeres cubanas con lupus en recibir un diagnóstico desde que aparecen los primeros síntomas.

Suponiendo que la distribución de los tiempos no es normal, pero tenemos una muestra aleatoria de:

- (n = 40) mujeres
- Media muestral: $\bar{x} = 6.5$ meses

- Desviación estándar muestral: (s = 1.2)

Gracias al **Teorema Central del Límite**, se puede usar la distribución normal para construir un intervalo de confianza del 95% para la media poblacional mediante la fórmula:

$$IC_{95\%} = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$$

Usando valor crítico (z = 1.96) para un 95% de confianza:

$$IC_{95\%} = 6.5 \pm 1.96 \cdot \frac{1.2}{\sqrt{40}} \approx 6.5 \pm 0.37$$

Entonces, con un 95% de confianza, el tiempo promedio real está entre **6.13 y 6.87 meses**.

III. Estimación e Intervalos de Confianza (20 puntos)

8. (8 p) **Intervalo de confianza para la media*

   Una muestra de tamaño *n* = 25 arroja una media muestral de 80 y desviación estándar muestral de 10.

   a. Calcula el intervalo de confianza al 95 % para la media poblacional (asume distribución normal).

   b. Interpreta el resultado.

a. Cálculo del intervalo de confianza al 95 %

- Tamaño de la muestra: (n = 25)
- Media muestral: ($\bar{x}$ = 80)
- Desviación estándar muestral: (s = 10)

Como (n < 30), asumimos que la variable a analizar es cuantitativa y la distribución es "paramétrica" usamos la **distribución t de Student** con (df = n - 1 = 24).
Un nivel de confianza del 95 %, implica un 5 % de error total (α=0.05). Como el intervalo es simétrico, ese error se reparte en ambos extremos de la distribución: α/2=0.025 a cada lado. Así se obtiene el valor crítico t_{α/2} para calcular el intervalo. Apoyándonos de una tabla, el valor crítico (t_{0.025, 24} $\approx$ 2.064).

Fórmula del intervalo de confianza:

$$IC = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

Sustituyendo los valores:

$$IC = 80 \pm 2.064 \cdot \frac{10}{\sqrt{25}} = 80 \pm 2.064 \cdot 2 = 80 \pm 4.128$$

Resultado:

$$IC_{95\%} = [75.872,\ 84.128]$$

b. Interpretación del resultado

Podemos decir que, **con un 95 % de confianza**, la media real de la población (por ejemplo, un parámetro como la puntuación promedio de calidad de vida o nivel de anticuerpos en mujeres cubanas con lupus) se encuentra **entre 75.87 y 84.13**.

Esto significa que **si repitiéramos el estudio muchas veces**, en el 95 % de los casos el intervalo calculado **incluiría la media poblacional real**.

9. (6 p) **Intervalo de confianza para proporción**

   De 200 pacientes, 50 presentaron un efecto adverso.

   a. Calcula el intervalo de confianza al 95 % para la proporción de efectos adversos.

   - Tamaño de la muestra: (n = 200)
   - Éxitos (efectos adversos): (x = 50)
   - Proporción muestral:

     $$\hat{p} = \frac{x}{n} = \frac{50}{200} = 0.25$$

     Para un nivel de confianza del 95 %, se utiliza el valor crítico Z_0.025 $\approx$ 1.96 porque el 5 % de error se reparte simétricamente en ambas colas de la distribución normal (2.5 % a cada lado). Este valor asegura que el 95 % central de los posibles valores de la estimación estén dentro del intervalo de confianza.

Fórmula del intervalo de confianza para proporción:

$$IC = \hat{p} \pm Z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Cálculo:

$$IC = 0.25 \pm 1.96 \cdot \sqrt{\frac{0.25 \cdot 0.75}{200}} = 0.25 \pm 1.96 \cdot 0.0306 = 0.25 \pm 0.06$$

$$IC_{95\%} = [0.19,\ 0.31]$$

Con 95% de confianza, se estima que entre el 19% y el 31% de todos los pacientes podrían tener efectos adversos.

b. ¿Qué sucede con la amplitud del intervalo si la muestra fuera de 400 pacientes, manteniendo la misma proporción?

Si mantenemos la proporción (0.25) pero usamos una muestra de 400 pacientes:

$$SE = \sqrt{\frac{0.25 \cdot 0.75}{400}} = 0.02165$$

$$IC = 0.25 \pm 1.96 \cdot 0.02165 = 0.25 \pm 0.0424 \Rightarrow [0.2076,\ 0.2924]$$

El intervalo es más estrecho por lo que hay mayor precisión. Aumentar la muestra reduce la incertidumbre.

10. (6 p) **Errores Tipo I y Tipo II**

a. Define error tipo I y error tipo II.

- **Error Tipo I (α):** Rechazar la hipótesis nula cuando **es verdadera**. Es un falso positivo. Ejemplo: Concluir que un tratamiento para lupus causa efectos adversos cuando en realidad **no los causa**.
- **Error Tipo II (β):** No rechazar la hipótesis nula cuando **es falsa**. Es un falso negativo. Ejemplo: Decir que un tratamiento **no tiene efectos adversos** cuando en realidad **sí los tiene**.

b. Explica la relación entre tamaño de muestra, nivel de significancia y potencia.

- **Potencia estadística** = (1 - β), es la capacidad de detectar un efecto si realmente existe.
- La potencia de un estudio aumenta con el tamaño de muestra. Cuantas más observaciones se recojan, mayor será la capacidad para detectar diferencias reales y menor el riesgo de cometer un error tipo II (no detectar un efecto que sí existe).
- Cuando se escode menor α (menor nivel de significancia) (más exigente para rechazar H_0) $\rightarrow$ menor riesgo de error tipo I, pero puede aumentar el riesgo de error tipo II. Cuanto más pequeño sea α (más estricto), más grande debe ser la muestra para mantener la misma potencia y detectar el mismo efecto.
- Hay un **balance** entre α, β y n. Elegirlos bien es clave en ensayos clínicos.

En estudios estadísticos, α es la probabilidad de cometer un error tipo I (falso positivo), β es la de cometer un error tipo II (falso negativo), y n es el tamaño de la muestra. Aumentar el tamaño muestral mejora la precisión y reduce la probabilidad de errores, especialmente el tipo II.

IV. Pruebas de Hipótesis Generales (20 puntos)

11. (6 p) **Prueba *t* de una muestra** Plantea las hipótesis nula y alternativa para verificar si la media poblacional difiere de 100. Describe brevemente cómo se calcula la estadística *t* y cómo decides rechazar o no H₀.

1. Definimos la hipótesis nula y la alternativa y calculamos el valor de t.

H₀: μ = 100 (la media poblacional es igual a 100)

H₁: μ ≠ 100 (la media poblacional es diferente de 100)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Donde:

- ($\bar{x}$): media muestral
- (μ_0): es el valor de la media definido en la hipótesis nula
- (s): desviación estándar muestral
- (n): tamaño de la muestra

2. Buscar el valor crítico t en la tabla de *t* de Student:

- Para un nivel de confianza del 95 % en una prueba bilateral:
- (α = 0.05)
- ($\alpha/2$ = 0.025)
- Usa los grados de libertad (df = n - 1)
- Busca el valor ($t_{\alpha/2,df}$) en la tabla o con una función estadística.

3. Regla de decisión:

- Si ($|t_{\text{calculado}}| > t_{\text{crítico}}$), **rechazas** la hipótesis nula (H_0).
- Si ($|t_{\text{calculado}}| \leq t_{\text{crítico}}$), **no rechazas** (H_0).

12. (6 p) ***Chi*-cuadrado de independencia** En una tabla de contingencia 3×2, explica el proceso para: a. Calcular el estadístico x². b. Determinar los grados de libertad y el valor crítico al 5 %. c. Interpretar el resultado.

Objetivo: Verificar si **dos variables categóricas están asociadas**, como tipo de tratamiento recibido y presencia de efectos adversos en mujeres cubanas con lupus.

a. Cálculo del estadístico x²

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Donde:

- (O_{ij}): frecuencia observada en la celda (i, j)

- (E_{ij}): frecuencia esperada en la celda (i, j)

$$E_{ij} = \frac{(\text{Total fila i}) \cdot (\text{Total columna j})}{\text{Total general}}$$

Grados de libertad y el valor crítico al 5 %.

$$df = (r-1)(c-1)$$

Para una tabla 3×2:

$$df = (3-1)(2-1) = 2$$

El valor crítico de χ^2 al 5 % teniendo en cuenta 2 como grado de libertad tenemos: $\chi^2_{0.05,2}$= 5,9915.

Decisión

- Si ($\chi^2_{\text{calculado}} > \chi^2_{\text{crítico}}$), **rechazamos H_0**.

13. (8 p) **Prueba exacta de Fisher** a. ¿En qué situaciones es preferible usar Fisher en lugar de x²? b. ¿Cómo se interpreta el *p*-valor obtenido?

¿Cuándo se usa?

- Tablas de contingencia **2×2**
- Tamaños de muestra pequeños
- Frecuencias esperadas menores de 5

Porque la χ^2 es una aproximación que funciona bien con muestras grandes, pero no es precisa con datos pequeños. En cambio, la prueba de Fisher calcula la probabilidad exacta, por eso es más confiable en esos casos.

Interpretación del valor *p*

- Representa la probabilidad de observar una distribución **igual o más extrema** que la actual si las variables fueran independientes.
- Si (p < α), **rechazamos H_0** y concluimos que **hay asociación** entre las variables.

V. Pruebas de Hipótesis sobre Odds Ratio (OR) (20 puntos)

14. (4 p) **Definición de OR** a. Define el odds ratio.

El odds ratio (OR) es una medida estadística que se usa para comparar las probabilidades (odds) de un evento entre dos grupos.

Se calcula a partir de una tabla de contingencia 2×2:

$$\text{OR} = \frac{a \cdot d}{b \cdot c}$$

Donde:

	Evento Sí	Evento No	Total
Grupo Expuesto	a	b	a + b
Grupo No Expuesto	c	d	c + d

b. Explica qué significa un OR = 1, OR > 1 y OR < 1.

OR = 1: No hay asociación entre la exposición y el evento. Las probabilidades son iguales en ambos grupos.

OR > 1:Hay mayor probabilidad del evento en el grupo expuesto. Ejemplo: si OR = 2, el evento es el doble de probable en ese grupo.

OR < 1:Hay menor probabilidad del evento en el grupo expuesto. Ejemplo: si OR = 0.5, el evento es la mitad de probable en ese grupo.

15. (8 p) **Cálculo de OR** Dada la siguiente tabla 2×2 de exposición (E) y desenlace (D):

    | | D = Sí | D = No | | --- | --- | --- | | E = Expuesto | 40 | 20 | | E = No expuesto | 10 | 30 |

    a. Calcula el odds ratio. b. Interpreta el resultado en términos de riesgo relativo de desenlace.

$$\text{OR} = \frac{a \cdot d}{b \cdot c} =$$

donde: a=40: Expuestos con desenlace b=20: Expuestos sin desenlace c=10: No expuestos con desenlace d=30: No expuestos sin desenlace

$$\text{OR} = \frac{40 \cdot 30}{20 \cdot 10} = \frac{1200}{200} = 6$$

El OR = 6 indica que: los pacientes expuestos tienen 6 veces más probabilidades de presentar el desenlace (por ejemplo, un efecto adverso o complicación) que los pacientes no expuestos.

16. (8 p) **Intervalo de confianza y prueba de hipótesis sobre OR**

    a. Describe cómo se construye el intervalo de confianza al 95 % para el log(OR).

    b. Con los datos del ejercicio anterior, calcula (aproximadamente) el IC al 95 % para el OR y di si es estadísticamente significativo (suponer aproximación normal).

    c. Explica qué conclusión sacas si $1 \notin$ IC.

Para construir el IC del logaritmo del odds ratio (log(OR)), se usan los siguientes pasos:

$$\log(\text{OR}) \pm Z_{\alpha/2} \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Donde: a,b,c,d son las frecuencias de la tabla 2×2, $Z_{\alpha/2}$=1.96 para un nivel de confianza del 95 %.

Después, se exponen los límites para obtener el intervalo de confianza del OR:

$$\text{IC } 95\,\% = \left[e^{\text{Límite inferior}},\ e^{\text{Límite superior}}\right]$$

b. Cálculo del IC al 95 % para el OR del ejercicio anterior Recordamos la tabla:

	D = Sí	D = No
E = Expuesto	40	20
NE = No expuesto	10	30

Paso 1: Calcular OR

$$\text{OR} = \frac{40 \cdot 30}{20 \cdot 10} = 6$$

Paso 2: Calcular log(OR)

$$\log(6) \approx 1.79176$$

$$SE = \sqrt{\frac{1}{40} + \frac{1}{20} + \frac{1}{10} + \frac{1}{30}} \approx \sqrt{0.025 + 0.05 + 0.1 + 0.0333} \approx \sqrt{0.2083} \approx 0.4564$$

Paso 3: Calcular el intervalo para log(OR)

$$\text{Límite inferior} = 1.79176 - 1.96 \cdot 0.4564 \approx 0.897$$

$$\text{Límite superior} = 1.79176 + 1.96 \cdot 0.4564 \approx 2.686$$

Paso 4: Exponenciar para obtener el IC del OR

$$\text{IC } 95\,\% = \left[e^{0.897},\ e^{2.686}\right] \approx [2.45,\ 14.68]$$

El resultado sí, es estadísticamente significativo, porque el intervalo de confianza del OR al 95 % es: [2.45, 14.68] Este intervalo no incluye el valor 1, lo que indica que existe una asociación estadísticamente significativa entre la exposición y el desenlace. En otras palabras, la probabilidad de que el resultado se deba al azar es muy baja.

c. ¿Qué pasa si el 1 NO está dentro del IC? Si 1 no está dentro del intervalo de confianza del OR, significa que el OR es significativamente diferente de 1. Por tanto, hay evidencia estadística de una asociación entre la exposición y el desenlace.

¡Gracias Profe!