Introducción a la Inferència Estadística

Yolanda Colom Torrens

Introducción

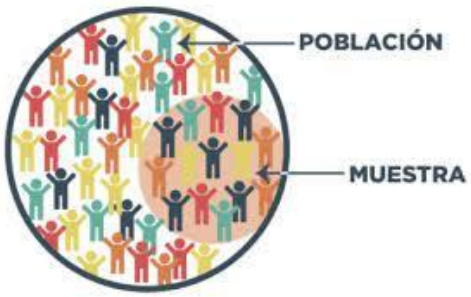
La inferencia estadística es una técnica esencial en el análisis y tratamiento de datos. En esta actividad, nos centraremos en resolver problemas matemáticos contextualizados para mejorar las competencias específicas y transversales relacionadas con la aplicación de los principios clave de la matemática, la estadística y la informática para el análisis y tratamiento de datos. La unidad tiene como objetivo identificar el método de muestreo y el tamaño de la muestra, implementar la estimación puntual y la estimación mediante intervalos de confianza, y al mismo tiempo interpretar correctamente los resultados.

Se buscará identificar el tipo y dimensión de la muestra, lograr aplicar de manera correcta estimación mediante intervalos de confianza y estimación puntual. Finalmente será importante la interpretación correcta de los resultados obtenidos.

Todas estas son habilidades necesarias para la obtención del Bachelor en data science, ser capaz de estructurar la información que estamos estudiando, la capacidad de seleccionar correctamente los datos más relevantes para implementar técnicas de modelización estadística, poder presentar y visualizar los datos para luego de un análisis de datos guiarnos en la toma de decisiones.

Los contenidos de la actividad se estructuran en cuatro secciones:

- Introducción general y conceptos básicos.
- Muestreo, estimación puntual.
- Estimación mediante.
- Intervalos de confianza.



Problema 1:

Si buscamos lograr una muestra representativa y realizar inferencias precisas sobre las medidas tomadas por las personas para enfrentar al cambio climático, debemos optar por un tipo particular de muestreo: el aleatorio simple. Esta técnica nos permite asegurar que cada individuo dentro de la población tenga igual oportunidad de ser seleccionado para formar parte de nuestra encuesta.
2.1. Para calcular el tamaño de muestra necesario para estimar la proporción de individuos que toman medidas para prevenir el cambio climático, podemos utilizar la fórmula:
n = z^2 * p * (1-p) / e^2.

Donde:
Cuando se trata de recopilar datos sobre las actitudes de las personas hacia los esfuerzos de mitigación del cambio climático, obtener una estimación precisa requiere una cuidadosa consideración y cálculo de varios factores, como nuestro nivel de confianza en nuestros resultados (95%), un valor asumido para las proporciones desconocidas entre los encuestados (en este caso, establecido de forma conservadora en un 50%) y un margen de error aceptable. Dados estos factores, podemos determinar el tamaño de muestra necesario mediante la siguiente fórmula:

n = (1.96)^2 * 0.5 * (1 - 0.5) / (0.02)^2 = 2401.64.

Esto significa que se necesita un tamaño de muestra de aproximadamente 2402 para obtener resultados confiables. Obtener una muestra verdaderamente representativa requiere una cuidadosa consideración del equilibrio de género dentro de cada grupo de edad y cálculos correspondientes en función del tamaño de muestra. Simplificando un poco las cosas, nuestro análisis presumirá una distribución uniforme entre hombres y mujeres en todos los rangos etarios.

15-19	24020.11	=	264.22	265
20-24	24020.125	=	300.25	301
25-29	24020.14	=	336.28	336
30-34	24020.205	=	492.01	492

	15-19	20-24	25-29	30-34	TOTAL
Hombres	265	301	336	492	1394
Mujeres	265	301	336	492	1394
Total	530	602	672	984	2788

Para determinar los márgenes de error relacionados con las proporciones estimadas en nuestra encuesta, podemos utilizar la siguiente ecuación después de obtener una muestra de aproximadamente 160 respuestas: **e = z * sqrt(p * (1 - p) / n).**

Así es como funciona: "z" se refiere a los valores críticos basados en nuestro nivel de confianza elegido. Dado que buscamos una tasa de confianza de aproximadamente el 95%, utilizar "z" = 1,96 debería ser suficiente. De manera similar, "p" implica las proporciones estimadas de individuos dentro de diferentes grupos de género/edad que están tomando medidas con respecto al cambio climático. Por último, "n" representa los tamaños de muestra individuales en juego aquí (asumiendo una distribución igual entre hombres y mujeres).
La proporción de personas en cada grupo de edad se puede calcular dividiendo la cantidad correspondiente en la tabla anterior

Problema 2:

Hipótesis nula

- H0: El peso de cada unidad =750 gr.
- H1: El peso de cada unidad = 750 gr.

Probabilidad del suceso ≤ 748, **$z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$**

- $\bar{x}$ media muestral.
- μ media poblacional.
- σ desviación típica poblacional.
- n tamaño de la muestra.

$z = (748 - 750) / (5 / \sqrt{100}) = -4$

Gracias a una tabla de distribución normal estándar, calculamos que z sea menor o igual a -4 es de aproximadamente 0.00003.
Como el resultado de la probabilidad es menor a 5%, podemos rechazar la hipótesis nula y concluir que la afirmación del gerente de que el peso de cada unidad es de 750 gr esta incorrecta.

b.- Para calcular los límites especificados del 95%, debemos buscar el v.c z para esta confianza. Con una distribución normal estándar, vemos que z = 1.96aprox.
Luego, con la fórmula del intervalo de confianza estudiaremos la media poblacional:

intervalo de confianza = $\bar{x} \pm z * (\sigma / \sqrt{n})$.

- $\bar{x}$ media muestral.
- σ desviación típica poblacional.
- n tamaño de la muestra.

$5 \pm 1.96 * (0.005 / \sqrt{64}) = (4.996, 5.004)$.

Resultando que los limites para este caso serán 4.996 cm y 5.004 cm.

Problema 3:

Con una distribución binomial podemos observar que n = número de reservas y p = la probabilidad de que n aparezca. **n = 260, p = 0.94**

Para que la empresa pueda acomodar a todos los pasajeros sabemos que debemos acertar en 260 ensayos, por lo tanto: **P(X=260) = (260 sobre 260) * (0.94)^260 * (0.06)^(0) = 0.0053**

Con estos resultados podemos analizar que es imposible acomodar a todos los pasajeros con reservas.

Problema 4:

a) Para el cálculo de μ usaremos la media muestral, **$\bar{x}$ = 150 gr.**
b) Al verificar si **$E(\bar{x}) = \mu$** podremos ver si x es un estimador correcto o no. Ya que el valor de X es 150gr este nos sirve.
   Para ver si $\bar{x}$ es un estimador consistente, verificamos si su varianza se aproxima a un valor nulo cada vez que n aumente. La varianza de **$\bar{x} = \sigma^2/n$**, donde σ es la desviación estándar poblacional. **σ = 10 gr n = 15**
   la varianza de $\bar{x}$ es de 6,67 gr^2. A medida que n aumenta, la varianza de $\bar{x}$ disminuye, lo que indica que $\bar{x}$ es un estimador consistente de μ.

c) Al revisar la varianza muestral s^2 = 100 gr^2 siendo este el mejor estimador puntual de σ^2.
d) Al igual que antes debemos revisar la veracidad de s^2, corroboramos si E(s^2) = σ^2.
   y como s^2 = 100 gr^2, podemos usarlo.

Para confirmar si s^2 es un buen estimador, necesitamos ver si la varianza de s^2 se acerca a cero a medida que aumenta el tamaño de la muestra. La varianza de s^2 se calcula como 2(σ^4/n), donde σ es la desviación estándar poblacional. En este caso, σ es de 10 gramos y n es de 15, lo que resulta en una varianza de s^2 de 1333,33 gramos al cuadrado. Pero a medida que aumenta el tamaño de la muestra, la varianza de s^2 disminuye, lo que significa que s^2 es un buen estimador consistente de σ^2. En otras palabras, podemos confiar en s^2 para estimar con precisión la varianza de la población a medida que aumenta el tamaño de la muestra.

Problema 5:

Con la fórmula del intervalo de confianza en donde podremos ver la diferencia de medias poblacionales con varianzas desconocidas pero iguales, revisando su nivel de confianza >= 95%.

$\bar{x}A - \bar{x}B \pm t(0.025, 88) * sqrt(((nA-1)*sA^2 + (nB-1)sB^2)/(nA+nB-2)(1/nA + 1/nB))$

- $\bar{x}$A media de la muestra A.
- $\bar{x}$B media de la muestra B.
- sA desviación estándar (A).
- sB desviación estándar (B).
- nA dimensión (A).
- nB dimensión (B).

418 - 402 ± t(0.025, 88) * sqrt(((40-1)*26^2 + (50-1)22^2)/(40+50-2)(1/40 + 1/50))

Intervalo de confianza = 16 ± 1.989 * 7.657 = **(0.3, 31.7)**

Con este análisis podemos afirmar que las duraciones de bombillas de las distintas marcas A y B se encuentra entre 0.3 y 31.7 horas.

5.2

p ± z(0.005) * sqrt(p*(1-p)/n)

- P proporción de clientes que compran carne una vez por semana.
- N dimension de la muestra.
- Z (0.005) valor crítico de la distribución normal estándar, confianza del 99%.

0.68 ± 2.576 * sqrt(0.68*0.32/300).

Intervalo de confianza = (0.613, 0.747).

Por En conclusión se observa que con un nivel de confianza >= al 99%, la proporción de demandantes que adquirirán la carne una vez a la semana esta entre [0.613, 0.747].