Maestría en Minería de Datos y Descubrimiento del Conocimiento
Curso de nivelación en Estadística

Práctica 2 - Estadística

A) *Estimación*

1. ¿Cuál de las siguientes cantidades es una variable aleatoria?

   - a. La media poblacional
   - b. El tamaño N de la población
   - c. El tamaño n de la muestra
   - d. La media muestral
   - e. La varianza de la media muestral
   - f. El valor más grande de la unidades de la muestra
   - g. La varianza de la población

2. Se analizó una muestra de 12 piezas de pan blanco de cierta marca y se determinó el porcentaje de carbohidratos contenido en cada una de las piezas, obteniéndose los siguientes valores:
   76.93 76.88 77.07 76.68 76.39 75.09 77.67 76.88 78.15 76.50 77.16 76.42

   a) Estimar la media del porcentaje de carbohidratos contenido en las piezas de pan de esta marca.

   b) Estimar la mediana del porcentaje de carbohidratos.

   c) Estimar la proporción de piezas de pan de esta marca cuyo contenido de carbohidratos no excede el 76.5 %.

3. Consideremos muestras aleatorias $X_1, \ldots, X_n$ para cada una de las siguientes distribuciones:

   a) Normal de media θ y varianza σ_0^2 (con σ_0^2 conocido).

   b) Poisson de parámetro θ.

   c) Geométrica de parámetro θ.

   Encontrar en cada caso el estimador de momentos de θ. Para los dos primeros casos, decidir además si estos estimadores son insesgados y hallar su error cuadrático medio.

4. Suponga que se está evaluando la posibilidad de ayudar económicamente una población de bajos recursos. La ayuda se materializará en un subsidio por hogar a aquellos hogares donde el ingreso promedio familiar esté por debajo del nivel de probreza.

   Suponga que de una población de 10000 hogares se entrevistan a 40, resultando 16 ser pobres.

   a) Provea un estimador de la proporción poblacional de los de los hogares pobres y su valor estimado.

   b) Encuentre la varianza del estimador propuesto en (a), provea un estimador insesgado de dicha varianza y su valor estimado.

5. Se sabe que el tiempo de duración de una clase de lámparas tiene distribución uniforme de parámetro $(0, \theta)$. Se han probado 20 lámparas, obteniéndose los siguientes tiempos de duración (en días):
   45 53 50 61 39 40 45 47 38 53 54 60 34 46 34 50 42 60 62 50
   Hallar un estimador de θ y dar el valor estimador en este caso.

6. Durante 20 días se ha registrado el número de llamadas en una central telefónica, obteniéndose los siguientes valores:
   35 41 38 40 34 36 41 48 42 39 57 41 35 37 38 41 43 44 46 47
   Supongamos que el número de llamadas diarias sigue una distribución Poisson de parámetro λ. Hallar un estimador de λ y hallar el valor estimado en este caso.

7. a. (Este item es opcional, lo que hay que saber es el resultado. La demostración puede verse en cualquier libro) Demuestre que

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

   es un estimador insesgado de la varianza poblacional.

   b. Demuestre que S^2/n es un estimador insesgado de la varianza de la distribución de $\overline{X}$.

B) *Intervalos de confianza*

8. ¿Verdadero o falso?

   a) El centro del intervalo con 95 % de confianza para la media poblaciónal de una distribución normal es una variable aleatoria.

b) La probabilidad de que un intervalo de confianza de nivel 0.80 para la media poblacional contenga a la media poblacional es 0.80.

c) Un intervalo de confianza de nivel 0.90 para la media poblacional contiene al 90 % de los valores de la población.

d) Aproximadamente el 95 % de muchos intervalos de confianza de nivel 0.95 para la media poblacional, basados en repetidas muestras aleatorias de una misma población contendrán la media de la población.

9. Una máquina de refrescos está ajustada de tal manera que la cantidad de líquido despachada tiene distribución normal con una desviación estandar igual a 0.15 decilitros. Halle un intervalo de confianza del 95 % para la media de los refrescos que sirve esta máquina si una muestra aleatoria de 36 refrescos tiene un contenido promedio de 2.25 decilitros.

10. Se registraron las siguientes mediciones del tiempo de secado, en horas, de una marca de pintura látex:

    1,3 1,5 1,7 2 2,1

    Suponiendo que las mediciones representan una muestra aleatoria de una población normal, encuentre un intervalo de confianza de 99 % para el tiempo medio de secado.

11. En una muestra aleatoria de 1000 casas en una determinada ciudad, se encuentra que 228 de ellas tienen calefacción de petróleo. Encuentre un intervalo de confianza de 99 % para la proporción de hogares en dicha ciudad que tiene este tipo de calefacción.

12. Un intervalo de confianza de nivel 0.90 para la media del número de menores en cada hogar de una población, dió como resultado el intervalo (0.7,2.1). ¿Podemos concluir de este intervalo que en el 90 % de los hogares de la población habitan entre 0.7 y 2.1 menores?

13. Un área de la Ciudad de Buenos Aires contiene 80000 hogares. Para estimar la cantidad de autos por hogar se toma una muestra de tamaño 100. Se estima la media poblacional (μ) y se ha calculado que $(1{,}44, 1{,}76)$ es un intervalo de confianza para μ de nivel 95 %. Como μ es algún número fijo (desconocido), o bien está dentro o bien está fuera del intervalo $(1{,}44, 1{,}76)$, por lo que no tiene sentido decir que $P\,(1{,}44 \leq \mu \leq 1{,}76) = 0{,}95$. ¿Qué es lo que queremos significar entonces, cuando decimos que el intervalo $(1{,}44, 1{,}76)$ es un intervalo de confianza de nivel 0.95?

14. Suponga que hemos tomado una muestra aleatoria de tamaño n grande de una población y hemos construido un intervalo con nivel aproximado de confianza de 90 % para la media poblacional con varianza conocida. Si quisiéramos reducir a la mitad el tamaño de un intervalo de confianza del mismo nivel aproximado (de 90 %), ¿en cuánto deberíamos aumentar el tamaño de nuestra muestra?

15. Un analista cuantifica su incertidumbre acerca de su estimación de la media poblacional reportando el intervalo $\left(\overline{X} - \frac{S}{\sqrt{n}}, \overline{X} + \frac{S}{\sqrt{n}}\right)$. Asumiendo que la distribución de la población es normal y que $n = 10$ ¿Cuál es el nivel de este intervalo? Si n aumenta, ¿qué le ocurre al nivel de confianza de este intervalo? ¿Disminuye o aumenta? Justifique su respuesta.

16. Suponga que un intervalo de confianza de nivel 95 % para una proporción p basado en una muestra aleatoria de tamaño 400 de una cierta población muy grande de tamaño N = 4,000,000 es igual a $(0{,}43, 0{,}47)$. Indique para cada una de las siguientes afirmaciones si son verdaderas o falsas. Justifique su respuesta.

    a) Si infinitas personas tomaran muestras aleatorias simples de tamaño 400 de la misma población y cada una calculara la proporción muestral, el 95 % de las proporciónes muestrales obtenidas caerían en el intervalo (0.43, 0.47).

    b) Si la población hubiese sido de tamaño 4N entonces la longitud del intervalo de confianza obtenido en base a la muestra de tamaño 400 se hubiese reducido a la mitad.

    c) Si la población hubiese sido de tamaño 4N entonces la longitud del intervalo de confianza obtenido en base a la muestra de tamaño 400 se hubiese duplicado.

    d) La probabilidad de que en una nueva muestra aleatoria de tamaño 400 la proporción muestral caiga dentro del intervalo (0.43,0.47) es igual a 95 %.

    e) De cada 1000 personas que obtengan muestras aletaorias independientes de tamaño 400, aproximadamente 950 de ellas calcularán intervalos de confianza para p iguales a $(0{,}43, 0{,}47)$.

    f) De cada 1000 personas que construyan intervalos de confianza de nivel 0.95 para p basándose cada una de ellas en muestras independientes de tamaño 400, aproximadamente 950 de ellas obtendrá intervalos que cubran al valor verdadero de p.

17. Supongamos que se observaron las realizaciones $x_1, ..., x_{10}$ de $X_i \sim N(\mu, \sigma^2)$ i.i.d.,$i = 1, .., 10$, y se obtuvo $\sum_{i=1}^{10} x_i = 100$ y $\sum_{i=1}^{10} x_i^2 = 1027$. Si se calculó un intervalo de confianza exacto para σ^2 asumiendo que μ es desconocida y se obtuvo $(a, 10)$, calcule el valor a de la cota inferior del intervalo y el nivel de confianza del intervalo.

18. Suponga que Ud. quiere estimar la proporción p de personas que está a favor del aborto en Argentina. Para ello tomará una muestra aleatoria simple de n habitantes de la Argentina. Defina $Y_1, ..., Y_n$ las respuestas que se obtendrán de los n entrevistados, siendo $Y_i = 1$ si el i-ésimo entrevistado fuera a responder estar a favor del aborto y $Y_i = 0$ de lo contrario. En vista de que la población de Argentina es muy grande en relación al tamaño de la muestra, es apropiado asumir que $Y_1, ..., Y_n$ son independientes e igualmente distribuidas con distribución Bernoulli(p).

a) ¿Cuál deberá ser n si se desea construir un intervalo de confianza para p de nivel aproximado 0.95 cuya máxima longitud posible sea 0.2?

b) Suponga que se obtiene una muestra de tamaño $m = 4n$, donde n es el valor encontrado en el inciso (a). ¿Cuál es la longitud máxima posible del intervalo para p de nivel 0.95 basado en una muestra de tamaño m?

c) Responda si es verdadero a falso (JUSTIFIQUE SU RESPUESTA):

c.1) Suponga que $n = 50$ y el total de personas en su muestra que declaró estar a favor del aborto es 7. Entonces, la proporción de personas en la Argentina que está a favor del aborto es 0.14.

c.2) Si la población de Argentina fuera 10 veces más grande se debería tomar una muestra 10 veces más grande (es decir, de tamaño $10n$) para que el estimador $\overline{Y}$ de p tuviera la misma varianza que el estimador que obtuvo con la muestra de tamaño n.

c.3) La varianza de p es igual a la varianza de $\overline{Y}$.

C) *Tests de hipótesis*

19. Una empresa vende dos variedades de soja. La variedad 1 tiene un rendimiento por ha. que puede considerarse una variable aleatoria con distribución $N(37, 25)$, y la variedad 2 tiene un rendimiento por ha. que puede considerarse $N(40, 25)$. Un cliente realizó una compra de semillas de la variedad 2 y antes de continuar comprando a esta empresa, quiere asegurarse de que las semillas que le enviaron realmente pertenecen a esa variedad y no a la variedad 1. Con ese fin, cultiva 10 parcelas de 1 ha. y obtiene los siguientes rendimientos:
37 - 39.50 - 41.70 - 42 - 40 - 41.25 - 43 - 44.05 - 38 - 38.50
El cliente quiere que la probabilidad de seguir comprando a esta empresa cuando las semillas no son de la variedad pedida sea 0.05.

a) Explicar por qué las hipótesis para este problema son

$$H_0 : \mu = 37, \qquad H_1 : \mu = 40$$

b) Encontrar un test para estas hipótesis y decir qué decisión se toma en base a la muestra dada. Calcular la probabilidad del error de tipo II.

c) Encontrar el test para el caso en que se cultiven n parcelas (con igual nivel).

d) Determinar el níumero n de parcelas a cultivar para que la probabilidad del error de tipo II sea menor o igual que 0.05.

e) Explicar por qué el test planteado en c) sirve también para las hip?tesis

$$H_0 : \mu = 37 \qquad H_1 : \mu > 37.$$

20. Según una estadística realizada por el Centro de Estudiantes la cantidad de minutos por semana que un alumno de Computación dedica a practicar alguna actividad deportiva es una variable aleatoria $N(45, 144)$. Se quiere saber si durante el receso de verano los estudiantes dedican más tiempo al deporte.

    a) En base a una muestra aleatoria de 100 datos, plantear el test de hipótesis adecuado y dar las regiones de rechazo de nivel $\alpha_1 = 0,01$ y de nivel $\alpha_2 = 0,05$

    b) Se sabe que, para una muestra de 100 datos, se rechazó la hipótesis nula a nivel 0,05, pero no a nivel 0,01. Indicar qué valores puede tomar el promedio muestral.

    c) En un test de nivel 0.05 basado en una muestra de 100 datos, ¿cuál es la probabilidad de equivocarse al sacar la conclusión si en realidad $\mu = 48$ minutos?

    d) Calcular el tamaño de la muestra para el cual la probabilidad de rechazar H_0 cuando la media poblacional es 47 sea menor a 0,01.

21. En la construcción de un edificio debe usarse un concreto que tenga una tensión media mayor a 300 psi. Para verificar si el concreto preparado a partir del cemento Loma Blanca cumple con este requerimiento, se realizan 15 mediciones en forma independiente de la tensión de este concreto. Se observa una media muestral de 304 psi y un desvío estándar muestral de 10 psi. El constructor está dispuesto a correr un riesgo del 5 % de comprar cemento Loma Blanca cuando éste produce un concreto que no cumple con las especificaciones. Suponiendo que los datos están normalmente distribuidos:

    a) Plantear el test correspondiente. ¿Qué decisión se toma?

    b) Acotar el valor p si usa una tabla.

22. Se diseñó un nuevo sistema de riego de manera tal que el desvío del tiempo de activación sea menor que 6 segundos. Se lo prueba 11 veces, obteniéndose los siguientes tiempos de activación:

    $$27 - 41 - 22 - 27 - 23 - 35 - 30 - 24 - 27 - 28 - 22$$

    Suponiendo que el tiempo de activación (en segundos) es una v.a. con distribución normal:

    a) ¿Usted decidiría, a nivel 0.05, que el sistema cumple la especificación?

    b) Acotar el valor p usando una tabla.

23. Se supone que 1 de cada 10 fumadores prefiere la marca A. Después de una campaña publicitaria en cierta región de ventas, se entrevistó a 200 fumadores para determinar la efectividad de la campaña. El resultado de esta encuesta mostró que 26 personas preferían la marca A.

    a) ¿Indican estos datos, a nivel aproximado 0.05, un aumento en la preferencia por la marca A?

    b) Calcular el valor p aproximado.

    c) ¿Cuál es la probabilidad aproximada de decidir que la campaña publicitaria no fue efectiva, cuando en realidad la proporción de preferencia por la marca A después de la campaña es 0.15?

    d) ¿Qué tamaño de muestra debería tomarse para que la probabilidad de c) fuese a lo sumo 0.05?

24. En cada caso indique si la afirmación es verdadera o falsa y justifique:

    a) El nivel de significación de un test es igual a la probabilidad de que la hipótesis nula sea cierta.

    b) Un Error de tipo II es más grave que un Error de tipo I.

    c) Si el p-valor es 0.3, el test correspondiente rechazaría al nivel 0.01.

    d) Si un test rechaza al nivel de significación 0.06, entonces el p-valor es menor o igual a 0.06.

    e) Si un intervalo de confianza de nivel 0.99 para la media μ de una distribución normal calculado a partir de una muestra da como resultado $[-2{,}0, 3{,}0]$, entonces el test para las hipótesis

    $$H_0 : \mu = -3 \qquad H_1 : \mu \neq -3$$

    basado en los mismos datos rechazara la hipótesis nula al nivel 0.01.