UNIVERSIDAD DE GUANAJUATO
DIVISIÓN DE CIENCIAS E INGENIERÍAS

Estadística Avanzada
Tarea 2

Gil Estéfano Rodríguez Rivera

NUA: 390906 ge.rodriguezrivera@ugto.mx

22 de agosto de 2021

Esta tarea se hizo basada en el Capítulo 6 de la *Tenth edition* del libro BIOSTATISTICS: A FOUNDATION FOR ANALYSIS IN THE HEALTH SCIENCES de **Wayne W. Daniel** y **Chad L. Cross**.

El **repositorio de GitHub** se encuentra aquí: https://github.com/gilesitorr/Estadistica_Avanzada/tree/main/Tarea_2.

El **Notebook de Google Colab** se encuentra aquí: https://colab.research.google.com/drive/1iX1eyJvYuRue040j1_ALdq5dJlSdE20w?usp=sharing.

1. Primera parte

Instrucciones

1. **For each of the following exercises construct 90, 95, and 99 percent confidence intervals for the population mean, and state the practical and probabilistic interpretations of each.**

2. **Indicate which interpretation you think would be more appropriate to use when discussing confidence intervals with someone who has not had a course in statistics, and state the reason for your choice. Explain why the three intervals that you construct are not of equal width.**

3. **Indicate which of the three intervals you would prefer to use as an estimate of the population mean, and state the reason for your choice.**

En respuesta al **segundo inciso**: El intervalo de confianza es un rango de valores, este rango se define a partir de la renuencia a equivocarse al rechazar la hipótesis inicial. Entre más confianza (o fe) le tengas a la hipótesis inicial, menos vas a querer rechazarla y vas a requerir de información muy contundente para que cambies de opinión y rechaces esa hipótesis.

En respuesta al **tercer inciso**: El valor más usado en general es la confianza del 95 %, por ello, salvo que quiera estar muy seguro de mi afirmación y requiera de información más contundente (en tal caso usaría uno de 99 % o más).

Problema 6.2.5

1. **Some studies of Alzheimer's disease (AD) have shown an increase in 14CO2 production in patients with the disease. In one such study the following 14CO2 values were obtained from 16 neocortical biopsy samples from AD patients.**

   1009 1280 1180 1255 1547 2352 1956 1080

   1776 1767 1680 2050 1452 2857 3100 1621

   Assume that the population of such values is normally distributed with a standard deviation of 350

Se usará una distribución normal por indicación del problema.
Como referencia, el **promedio de la muestra** es de $1747{,}625$.

- El intervalo de **confianza del** $90\,\%$ es de $(1603{,}7003,\ 1891{,}5497)$.
- El intervalo de **confianza del** $95\,\%$ es de $(1576{,}1282,\ 1919{,}1218)$.
- El intervalo de **confianza del** $99\,\%$ es de $(1522{,}2399,\ 1973{,}0101)$.

Problema 6.3.5

1. **A sample of 16 ten-year-old girls had a mean weight of 71.5 and a standard deviation of 12 pounds, respectively. Assuming normality, find the 90, 95, and 99 percent confidence intervals for m.**

Se usará una distribución normal por indicación del problema.

- El intervalo de **confianza del** $90\,\%$ es de $(66{,}5654,\ 76{,}4346)$.
- El intervalo de **confianza del** $95\,\%$ es de $(65{,}6201,\ 77{,}3799)$.
- El intervalo de **confianza del** $99\,\%$ es de $(63{,}7725,\ 79{,}2275)$.

Problema 6.4.5

1. **Krantz et al. (A-12) investigated dose-related effects of methadone in subjects with torsade de pointes a polymorphic ventricular tachycardia.**

In the study of 17 subjects, nine were being treated with methadone for opiate dependency and eight for chronic pain.

The mean daily dose of methadone in the opiate dependency group was 541 mg/day with a standard deviation of 156, while the chronic pain group received a mean dose of 269 mg/day with a standard deviation of 316.

Se usará una distribución t porque el tamaño de las muestras es muy pequeño (menor a 30).

En la prueba de hipótesis, se hará la suposición de que las **varianzas de las dos muestras son diferentes**. Esto es porque aunque ambos grupos son suministrados con la misma sustancia, sus condiciones y, por ende, sus posibles reacciones sean diferentes.

Como referencia, **la diferencia de los promedios** es de 272.

- El intervalo de **confianza del** $90\,\%$ es de $(39{,}2969,\ 504{,}7031)$.
- El intervalo de **confianza del** $95\,\%$ es de $(-18{,}1099,\ 562{,}1099)$.
- El intervalo de **confianza del** $99\,\%$ es de $(-156{,}0847,\ 700{,}0847)$.

Los intervalos de confianza de $95\,\%$ y $99\,\%$ contienen el 0, por lo que en ese caso, se puede inferir que los tratamientos son iguales.

Extra: Si se usa una prueba pero suponiendo **varianzas iguales**:

- El intervalo de **confianza del** $90\,\%$ es de $(64{,}9292,\ 479{,}0708)$.
- El intervalo de **confianza del** $95\,\%$ es de $(20{,}5683,\ 523{,}4317)$.
- El intervalo de **confianza del** $99\,\%$ es de $(-74{,}4198,\ 618{,}4198)$.

En este caso, se obtienen intervalos menos conservadores. Es decir, desde los intervalos de confianza del $95\,\%$ ya se infiere que las dosis son diferentes.

2. Segunda parte

Instrucciones

1. **For each of the following exercises state the practical and probabilistic interpretations of the interval that you construct.**
2. **Identify each component of the interval: point estimate, reliability coefficient, and standard error. Explain why the reliability coefficients are not the same for all exercises**

Problema 6.5.1

1. **Luna et al. (A-14) studied patients who were mechanically ventilated in the intensive care unit of six hospitals in Buenos Aires, Argentina.**

   The researchers found that of 472 mechanically ventilated patients, 63 had clinical evidence of ventilator-associated pneumonia (VAP).

   Construct a 95 percent confidence interval for the proportion of all mechanically ventilated patients at these hospitals who may be expected to develop VAP

Se usará una distribución binomial, pues se trata de un problema de lidiar con proporciones.

Como referencia, la **proporción de la muestra** es de $0{,}1335$. El intervalo de **confianza del** $95\,\%$ es de $(0{,}1028,\ 0{,}1642)$.

3. Tercera parte

Instrucciones

1. **For each of the following exercises state the practical and probabilistic interpretations of the interval that you construct.**

2. **Identify each component of the interval: point estimate, reliability coefficient, and standard error. Explain why the reliability coefficients are not the same for all exercises**

Problema 6.6.1

1. **Horwitz et al. (A-18) studied 637 persons who were identified by court records from 1967 to 1971 as having experienced abuse or neglect. For a control group, they located 510 subjects who as children attended the same elementary school and lived within a five-block radius of those in the abused/neglected group.**

   In the abused/neglected group, and control group, 114 and 57 subjects, respectively, had developed antisocial personality disorders over their lifetimes.

   Construct a 95 percent confidence interval for the difference between the proportions of subjects developing antisocial personality disorders one might expect to find in the populations of subjects from which the subjects of this study may be presumed to have been drawn.

Se usará una distribución binomial, pues se trata de un problema de lidiar con proporciones.

Como referencia, la **proporción de la muestra** es de $0{,}0672$. El intervalo de **confianza del** $95\,\%$ es de $(0{,}0268,\ 0{,}1076)$.

Problema 6.7.3

1. **A physician would like to know the mean fasting blood glucose value (miligrams per 100ml) of patients seen in a diabetes clinic over the past 10 years.**

   Determine the number of records the physician should examine in order to obtain a 90 percent confidence interval for μ if the desired width of the interval is 6 units and a pilot sample yields a variance of 60.

Dado que se nos proporciona un tamaño deseado para el intervalo de confianza, de éste se puede inferir el error estándar que se desea.

La definición del intervalo de confianza:

intervalo = estadístico $\pm$ (Z score)(error estándar)*

Por lo que el tamaño del intervalo de confianza es de:

R = 2 (Z score)(error estándar)*

Al despejar, se obtiene que el error estándar dado el tamaño del intervalo es de:

error estándar = R / (2(Z score))*

Con el error deseado, se puede calcular la d de Cohen y luego calcular el tamaño de la muestra con la varianza dada. El **tamaño de la muestra** es de 18 personas.

Problema 6.8.3

1. **A hospital administrator wishes to know what proportion of discharged patients is unhappy with the care received during hospitalization. How large a sample should be drawn if we let d=0.05, the confidence coefficient is .95, and no other information is available?**

2. **How large should the sample be if p is approximated by .25?**

No es posible determinar el tamaño de la distribución si no se conoce la varianza de la muestra (o alguna cantidad relacionada a ésta). Sin embargo, para una distribución binomial (de proporciones), el máximo se da cuando p=0.5. ¿Cuál es el valor del

tamaño de la muestra en ese caso? En el caso máximo (p=0.5), se requiere de una muestra de 384 personas.

Si p=0.25, entonces el tamaño estimado de la muestra es de 288 personas.

Problema 6.9.7

1. **Twenty air samples taken at the same site over a period of 6 months showed the following amounts of suspendes particulate matter (micrograms per cubic meter of air):**

   68 22 36 32

   42 24 28 38

   30 44 28 27

   28 43 45 50

   79 74 57 21

   Consider these measurements to be a random sample from a population of normally distributed measurements, and construct a 95 percent confidence interval for the population variance.

Como se está estudiando la desviación de la distribución, se usará una pruba χ^2.

Como referencia, la **varianza de la muestra** es de $295{,}6421$. El intervalo de **confianza del** $95\,\%$ es de $(170{,}9833,\ 630{,}6843)$.

Problema 6.10.7

1. **Measurements of gastric secretion of hydrochloric acid (milliequivalents per hour) in 16 normal subjects and 10 subjects with duodenal ulcer yielded the following results:**

   Normal subjects: 6.3, 2.0, 2.3, 0.5, 1.9, 3.2, 4.1, 4.0, 6.2, 6.1, 3.5, 1.3, 1.7, 4.5, 6.3, 6.2

   Ulcer subjects: 13.7, 20.6, 15.9, 28.4, 29.4, 18.4, 21.1, 3.0, 26.2, 13.0

   Construct a 95 percent confidence interval for the ratio of the two population variances. What assumptions must be met for this procedure to be valid?

Como se estudia la proporción de las desviaciones de dos distribuciones, se usará una distribución F.

Como referencia, la **proporción de las varianzas de las muestras** es de $15{,}9942$. El intervalo de **confianza del** $95\,\%$ es de $(5{,}1219,\ 60{,}2878)$.

4. Review Questions and Exercises

4.1. What is statistical inference?

Es el proceso de deducir las características de una población a partir del estudio de una muestra (que es un subconjunto de tal población).

4.2. Why is estimation an important type of inference?

Porque es un método que permite obtener información relevante sobre una población sin tener que estudiar la totalidad de la misma (proceso que puede ser incluso imposible en el mundo real).

4.3. What is a point estimate?

Es un valor que se obtiene de la muestra y que se usa para estimar un parámetro de la población.

4.4. Explain the meaning of unbiasedness

Un valor estimador no está sesgado cuando, luego de medir muchas veces el mismo, se obtiene que el promedio de los estimadores es similar al parámetro real de la población que se pretendía estimar.

4.5. Define the following:

- Reliability coefficient
- Confidence coefficient
- Precision
- Standard error
- Estimator
- Margin of error

Reliability coefficient (Coeficiente de confiabilidad):
Es el valor de la distribución estandarizada (ya sea la z -o normal estandarizada-, la t, la chi cuadrada, la F, etc) que se asocia al coeficiente de confianza que se desea.

Confidence coefficient (Coeficiente de confianza):

En el sentido práctico, es el valor entre 1 y 0 que determina cuánta certeza hay de que el parámetro real que se está estimando se encuentre dentro del intervalo de confianza. El intervalo se define como:

estimador $\pm$ *(coeficiente de confiabilidad)x(error estándar)*

Precision (Precisión):

Es la distancia del estimador hacia uno de los extremos del intervalo de confianza:

precisión = (coeficiente de confiabilidad)x(error estándar de la media)

Standard error (Error estándar):

Es la desviación estándar de un estimador. Es decir, es una medida de la dispersión de la muestra en torno a ese estadístico.

Estimator (Estimador):

Es una regla que indica cómo estimar un parámetro. Puede haber diversos estimadores para calcular un mismo parámetro.

Margin of error (Margen de error):

Es un sinónimo de la precisión.

4.6. Give the general formula for a confidence interval

estimador $\pm$ *(coeficiente de confiabilidad)x(error estándar)*

4.7. State the probabilistic and practical interpretations of a confidence interval.

Probabilística:

Luego de múltiples muestreos, la proporción del número de muestras correspondiente (usualmente el $95\,\%$ y, en general, el porcentaje que se desee) va a tener intervalos de confianza que contengan el parámetro real.

Práctica:

La proporción que se escoja (usualmente el $95\,\%$ y, en general, el porcentaje que se desee) indica la probabilidad de que el intervalo de confianza de la muestra de estudio contenga el parámetro real.

4.8. Of what use is the central limit theorem in estimation?

El Teorema del Límite Central indica que, dadas n variables aleatorias con distribuciones de probabilidad con varianzas no nulas y finitas y medias conocidas, entonces la suma de variables tiende a una distribución normal.

Este teorema es útil porque permite hacer estimaciones sobre los parámetros de la población (con un valor promedio y un error estándar) con sólo hacer diferentes muestreos.

4.9. Describe the t distribution.

Es una distribución simétrica con media 0. Tiene forma de campana y tiene como dominio los números reales. Depende de un parámetro conocido como grados de libertad (que es igual al tamaño de la muestra de estudio menos uno). Tiende a una distribución normal para muestras grandes.

4.10. What are the assumptions underlying the use of the t distribution in estimating a single population mean?

Lo primero a considerar es el tamaño de la muestra (la distribución t se usa cuando la muestra tiene menos de 30 elementos).

Se tiene que hacer la suposición de que se conoce la varianza de la población. También se debe hacer lasuposición de que la población sigue una distribución normal.

4.11. What is the finite population correction? When can it be ignored?

El cálculo del tamaño de población más elemental parte de suponer que se tienen poblaciones grandes (o infinitas, en el sentido práctico). En éstas, se supone que los muestreos sin reemplazo no suponen un traslape de datos en distintas muestras. Es decir, es un método que funciona cuando se tiene certeza de que un mismo elemento no va a aparecer en dos muestras diferentes. Esta fórmula no depende del tamaño de la población (pues éste no tiene efecto al ser infinito).

Cuando no se puede asegurar lo último, es decir, cuando la muestra es suficientemente pequeña (o finita), se utiliza una corrección para el cálculo de la muestra. Esa corrección depende del tamaño de la población.

4.12. What are the assumptions underlying the use of the t distribution in estimating the difference between two population means?

Se supone que ambas muestras de estudio vienen de poblaciones normalmente distribuidas. Asimismo, se puede suponer que las varianzas de ambas muestras son

iguales o distintas: lo importante es que sean conocidas.

5. Preguntas extra

Las siguientes preguntas no se encontraban en la asignación de la tarea pero se mencionaron en clase.

5.1. ¿Por qué la media de estatura de la gente de Japón cambió en una generación?

De acuerdo a Hermanussen et. al (2015), tanto la estatura como la morfología del esqueleto tienen una relación con el estilo de vida de la población: esto viene de cambios como el ambiente, los aspectos psicosociales y emocionales. Por estos motivos, hubo un cambio en el último siglo en la altura de la población japonesa.

De acuerdo con Ohyama et al. (1987), la estatura de la población de Japón creció desde el final de la Segunda Guerra Mundial. Por ejemplo, entre los años 1961 y 1962, la estatura promedio era de $167{,}1 \pm 4{,}7\ cm$. Veinte años después, entre 1980 y 1981 fue de $170{,}4 \pm 5{,}4\ cm$. De hecho, el mismo artículo indica que luego del período de guerra hubo un aumento en la altura promedio de varios países. Esto tiene cabida dentro la conclusión de Hermanussen et al. (2015) porque luego de una guerra se esperaría un aumento en la esperanza y el estilo de vida de algunos países.

Fuentes:

- **Hermanussen, M., Scheffler, C., Groth, D. & Aßmann, C.** (2015). HEIGHT AND SKELETAL MORPHOLOGY IN RELATION TO MODERN LIFE STYLE. Journal of Physiological Anthropology.
- **Ohyama, S., Hisanaga, A., Inamasu, T., Yamamoto, A., Hirata, M., & Ishinishi, N**. (1987). SOME SECULAR CHANGES IN BODY HEIGHT AND PROPORTION OF JAPANESE MEDICAL STUDENTS. American journal of physical anthropology, 73(2), 179–183.

5.2. Describir los primeros cuatro momentos estadísticos respecto de la media

El **primer momento** respecto al origen es el promedio:

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i := \overline{x}$$

Es una medida que indica la tendencia central de la muestra. Como mención, note que el primer momento respecto a la media (o momento central) es cero:

$$E(X - E(X)) = \frac{1}{n}\sum_{i=1}^{n}\Big(x_i - E(X)\Big) = E(X) - \frac{n}{n}E(X) = 0$$

El **segundo momento central** se conoce como **varianza**:

$$E((X - E(X))^2) = \frac{1}{n}\sum_{i=1}^{n}\Big(x_i - E(X)\Big)^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(x_i^2 + E^2(X) - 2x_iE(X)\Big)$$

$$= E(X^2) + E^2(X) - 2E(X)E(X) = E(X^2) - E^2(X)$$

Es una medida de la dispersión de la muestra.

Estos momentos se pueden estandarizar al dividir los momentos centrales entre una potencia de la desviación estándar σ igual al orden del momento. Es decir, el N-ésimo momento estandarizado se define:

$$\frac{E((X - E(X))^N)}{\sigma^N}$$

El **tercer momento estandarizado** es el **sesgo**. Es una medida de la simetría de la muestra. Es decir, un sesgo negativo indica que la distribución está *cargada* a la derecha: es decir, que la mediana se encuentra a la derecha de la media. Por su parte, un sesgo positivo indica una distribución con una mediana a la izquierda de la media.

El **cuarto momento estandarizado** es la **curtosis**. Es una medida del *peso* que tienen los *outliers* en la distribución. En otras palabras, la *concentración* de la distribución en torno a la media. En ese sentido, una campana más estrecha que la campana normal tendrá una curtosis menor que la misma, pues los outliers contribuyen menos que los valores centrales.

Fuentes:

- **United States Naval Academy**. (s.f.). MOMENT STATISTICS. Recuperado el 21 de agosto del 2021 de https://www.usna.edu/Users/oceano/pguth/md_help/html/moment_stats_2.htm
- **Wolfram Language and System Documentation Center**. (s.f.). STATISTICAL MOMENTS AND GENERATING FUNCTIONS. Recuperado el 21 de agosto del 2021 de https://reference.wolfram.com/language/guide/StatisticalMomentsAndGeneratingFunctions.html