Examen #1

Métodos Cuantitativos Aplicados, CIDE A.C.

Instrucciones

El examen tiene cuatro secciones. La primera consta de cinco preguntas sobre algunos conceptos básicos que hemos estudiado en clase. En general, no debes necesitar más de tres o cuatro oraciones para contesarlas. La segunda parte tiene de dos problemas de probablidad y la tercera dos preguntas de inferencia estadística. La última sección contiene un ejercicio de extra-crédito. Tienes dos horas para contestar este examen. Te recomiendo leer las preguntas con cuidado y tratar de no dejar respuestas en blanco. Si no tienes la respuesta final, siempre es mejor mostrar el procedimiento que tu crees que te llevaría a dicha respuesta.

Al terminar, deberás enviar el archivo con tus respuestas a mi dirección de correo electrónico (`sergio.bejar@cide.edu`). ¡Buena suerte!

Parte 1: Respuestas Rápidas (15 puntos).

1. ¿Qué significa el *p-value*? (3 pts.)

   Es la probabilidad de haber obtenido nuestro resultado por mera coincidencia.

2. ¿Qué es el error de muestreo aleatorio? (3 pts.)

   Es la diferencia que existe entre nuestros estadísticos muestrales y los parámetros poblacionales. En otras palabras, el EMA es que tan bien nuestra muestra aleatoria representa a la población.

3. ¿Por qué es importante hacer un muestreo aleatorio de calidad? (3 pts.)

   Porque un buen muestreo aleatorio reduce sesgos potenciales que afectan la calidad de nuestros resultados.

4. ¿Qué es la inferencia estadística? (3 pts.)

   Es un conjunto de métodos y técnicas que nos permiten hacer afirmaciones sobre la población por medio de una muestra.

5. Describe la importancia de incrementar el tamaño de una muestra (n). (3 pts.)

   Al incrementar n se reduce el E.M.A y por tanto la confiabilidad de nuestros resultados.

Parte 2: Probabilidad (30 puntos)

1. Una bolsa contiene bolas numeradas del 1 al 20, de manera que todas tienen la misma probabilidad de ser escogidas. (15 puntos).

   a) ¿Cuál es la probabilidad de que al sacar una bola, el número sea divisible por 3? (7.5 pts.)

   RESPUESTA:

   Evento A: Pr. de ser divisible por 3 = {3, 6, 9, 12, 15, 18}

   Evento B: Pr. de ser divisible por 5 = {5, 10, 15, 20}

   Evento A y B: Pr. de ser dividible por 3 y por 5 = {15}

   P(A) = $\frac{6}{20} = \frac{3}{10}$

   P(B) = $\frac{4}{20} = \frac{1}{5}$

   P(A y B) = $\frac{1}{20}$

   Por lo tanto:

   P(A) = $\frac{3}{10}$

   b) ¿Cuál es la probabilidad de que sea divisible por 3 o por 5? (7.5 pts.)

   RESPUESTA:

   Tenemos que calcular P(A) + P(B) - P(A y B). Usando los valores presentados en el inciso (a):

P(A) + P(B) - P(A y B) = $\frac{3}{10} + \frac{1}{5} - \frac{1}{20} = \frac{9}{20}$

2. Científicos en España han diseñado una prueba para detectar la presencia de la misteriosa enfermedad *cerebralitis*. Entre los que tienen la enfermedad, la probabilidad de que esta sea detectada por la prueba es del 86%. Sin embargo, la probabilidad de que la prueba indique erroneamente la presencia de la enfermedad en quienes no la tienen es del 8%. Se estima que el 16% de la población que se haga la prueba tiene la enfermedad. Si la prueba administrada a un individuo es positiva, ¿Cuál es la probabilidad de que esa persona en realidad tenga la enfermedad? (15 puntos).

RESPUESTA:

Denotemos los siguientes eventos:

Evento A: Tener la enfermedad.

Evento B: El examen sale positivo.

Usando la información dada tenemos lo siguiente:

P(A) = .16

P(no A) = .84

P(B|A) = .86

P(B|no A) = .08

Ahora aplicamos la fórmula de Bayes:

$\text{P(A|B)} = \frac{P(A)*P(B|A)}{(P(A)*P(B|A))+P(B|noA)*P(noA))} = \frac{.16*.86}{(.16*.86)+(.08*.84)} = .67$

Parte 3: Inferencia Estadística (55 puntos)

1. El 5 y 6 de octubre del 2023, la revista *Newsweek* organizó una encuesta que incluyó una muestra aleatoria de 1004 personas. Los individuos en la muestra contestaron la siguiente pregunta para medir cómo evalúan el trabajo del President Biden:

¿Usted aprueba o desaprueba la forma en la que el Presidente Biden está haciendo su trabajo como presidente?

En octubre del 2023, 33% de la muestra aprobaba el trabajo del Presidente Biden y el 67% desaprobaba su labor. Nótese que la variable aleatoria x puede tomar valores de 1 (si el individuo aprueba) y 0 (si el individuo desaprueba).

a) Calcula la media de x *(i.e.* $\bar{x} = \frac{1}{n}\sum_{x's} n_x * x$*)*. (5 pts.)

Hay al menos dos formas para resolver este problema. Una es aplicando las fórmulas dadas (i.e. el método largo). La otra es creando un `data.frame` que contega la información que hemos dado en el problema. Acá va la aplicación de las fórmulas:

```
aprueba <- 331  ## 331 aprueban el trabajo, es decir = 1
no_aprueba <- 673 ## 673 no aprueban, es decir = 0
n <- 1004  ## número de participantes

## Aplicamos la fórmula dada

media <- (aprueba*1 + no_aprueba*0)/n
```

```
media
```

```
[1] 0.3296813
```

b) Calcula la desviación estándar de x en la muestra (esto es, $s^2 = \sqrt{\frac{1}{n-1}} * \sqrt{\sum_{x=1}^{n}(x_i - \bar{x})^2}$. (5 pts.)

```
## Teniendo calculada la media podemos calcular d.e. muestral:

var <- ((aprueba*(1-media)^2) + (no_aprueba*(0-media)^2)) / (n - 1)

var  ## la varianza
```

```
[1] 0.2212119
```

```
dem <- sqrt(var)

dem ## la desviación estándar de la muestra
```

```
[1] 0.4703317
```

c) Calcula el intervalo de confianza para la variable x con un 95% de confiabilidad. (5 pts.)

Los datos que tenemos nos dan información sobre parámetros muestrales y los más que podemos hacer (usando el Teorema del Límite Central) es asumir que tanto la media muestral (.33) es muy parecida a la media poblacional. Para calcular la d.e. muestral:

```
sigma <- dem / sqrt(n)  ## fórmula de D.E. muestral

sigma
```

```
[1] 0.01484354
```

```
ic <- c(media - (1.96*sigma), media + (1.96*sigma)) ## fórmula I.C.

ic
```

```
[1] 0.3005879 0.3587746
```

d) Interpreta los resultados del intervalo de confianza que obtuviste en el inciso (c). (5 pts.)

El intervalo de confianza obtenido con la muestra de *Newsweek* nos indica que la aprobación de Biden va a estar entre el 30% y el 36% con un 95% de probabilidad.

2. En esta pregunta utilizaremos la base de datos `births` que viene el el paquete `openintro`. Dicha base contiene información sobre 150 nacimientos junto con información sobre las madres (la base la puedes cargar usando el código mostrado abajo). Nos interesa saber si hay diferencias significativas en el peso de los bebés cuyas madres fuman (f) versus el peso de los bebés cuyas madres no fuman. (NOTAS: (i) Para que el código funcione primero debes instalar el paquete `openintro`; (ii) lo columna `weight` tiene los datos sobre el `peso` del bebé; y (iii) la columna `smoke` tiene los datos de si la madre es fumadora (f) o no.)

```
```{r}
library(openintro)
library(tidyverse)
library(dplyr)
data(births)
head(births, 2)
```
```

a) Establece las hipótesis nula y alternativa. (5 pts.)

RESPUESTA:

H_0: $\overline{X}_{fumadoras} = \overline{X}_{nofumadoras}$

H_a: $\overline{X}_{fumadoras} \neq \overline{X}_{nofumadoras}$

b) Calcula la diferencia entre las medias muestrales del peso (`weight`) de los bebés de madres fumadoras (f) y el peso de los bebés de madres no fumadoras (nf). (10 pts.)

RESPUESTA:

```
## filtro fumadoras y no fumadoras. Luego uso `pull` para transformar columna en vector

smoker  <- births %>% filter(smoke == "smoker") %>% pull(weight)
nonsmoker <- births %>% filter(smoke == "nonsmoker") %>% pull(weight)

mean(nonsmoker) - mean(smoker)
```

```
[1] 0.4005
```

c) Calcula el valor de p. (5 pts.)

RESPUESTA:

```r
## Usamos `t.test` para simplificar

t.test(
  x           = smoker,
  y           = nonsmoker,
  alternative = "two.sided",
  mu          = 0,
  var.equal   = TRUE,
  conf.level  = 0.95
)
```

```
    Two Sample t-test

data:  smoker and nonsmoker
t = -1.5517, df = 148, p-value = 0.1229
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9105531  0.1095531
sample estimates:
mean of x mean of y
   6.7790    7.1795
```

El valor de p es 0.1229

d) Obtén el intervalo de confianza. (5 pts.)

RESPUESTA:

El intervalo de confianza es (-.91, .10) (Obtenido en el inciso anterior al usar `t.test`)

e) ¿Existe evidencia estadística a favor de la hipótesis nula? ¿Por qué? (10 pts.)

RESPUESTA:

No. El valor de p es mayor a .05.

Parte 4: Extra-Crédito (10 puntos)

El uso de termómetros para evaluar figuras políticas es muy común en estudios de opinión. En ese termómetro, valores más altos indican que los individuos tienen una opinión más favorable sobre la figura política. Y para asegurarnos que los datos son número enteros que se encuentran contenidos en el intervalo [0, 100] usamos la distribución *beta* (que es una aproximación a la normal). El código mostrado a continuación simula una muestra de 250,000 individuos con una $\mu = 40.015$ y $\sigma = 40.24$.

```
# library(tidyverse)
# library(stevemisc)

# Poblacion <- rbnorm(250000, mean = 40.01578, sd = 40.24403,
#                      lowerbound = 0,
#                      upperbound = 100,
#                      round = TRUE,
#                      seed = 8675309)
#
# Poblacion %>% as_tibble() %>%. ## pueden usar tibble or data.frame
#   mutate(uid = 1:n()) %>%
#   rename(term = value) %>%
#   select(uid, term) -> Poblacion
```

a) Simula 100,000 muestras aleatorias de 10 individuos cada una. ¿Cuál es la media de todas las medias muestrales? ¿Esta cerca de μ?

b) ¿Qué explica el resultado del inciso (a)?