Second midterm exam
December 15th, 2021

**Time: 75 minutes**

- **You are not allowed to use any documentation apart from the formula sheet you have received, and the Z(0,1) table.**
- **Use 4 decimal digits in all calculations and results.**

1. (3 points) The following estimator, to be used on a simple random sample of size 4, is proposed to estimate the average of failures in a batch of computer equipment:

$$\hat{\mu} = \frac{X_1 + X_2 - 2X_3 - 2X_4}{12}$$

   a. (0,5 points) Check if the estimator is not unbiased

   We calculate the estimator expectation:

$$E(\hat{\mu}) = \frac{\mu + \mu - 2\mu - 2\mu}{12} = -\frac{\mu}{6}$$

   b. (1 point) Calculate its bias. Does the estimator overestimate or underestimate the parameter value?

   Bias: $-\frac{\mu}{6} - \mu = \frac{-7\mu}{6}$; since the bias is negative, the estimator underestimates the parameter value

   c. (0,5 points) Calculate its variance

$$Var(\hat{\mu}) = \frac{\sigma^2 + \sigma^2 + 4\sigma^2 + 4\sigma^2}{144} = \frac{10\sigma^2}{144} = 0{,}07\sigma^2$$

   d. (1 point) For which sample size its variance is smaller than that of the sample mean distribution.

$$0{,}07\sigma^2 < \frac{\sigma^2}{n} \qquad n < \frac{1}{0{,}07} = 14{,}28$$

   For n < 15 its variance is smaller than that of the sample mean.

2. (4 points) Randomly selected performance programs are periodically executed on a network to check its performance. The programs are selected from a collection which have essentially the same type of instructions and input-output calls. 51 measures are taken, obtaining an average response time of 2,7 seconds with a sample variance of 0,3721 $s^2$. Quality standards establish that averge response times should be smaller than 3 seconds.

   a. (2 points) Can it be stated that the network fulfills this condition at a significance level of 1,5%?

Contraste para la media

$H_0 : \mu \geq 3$

$H_1 : \mu < 3$

Rechazamos si: $\frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} < -Z_\alpha$

$\bar{x} = 2,7 \qquad \hat{s}^2 = 0,3721 \rightarrow \hat{s} = 0,61$

$P(Z > Z_\alpha) = \alpha = 0,015; \qquad P(Z < Z_\alpha) = 1 - 0,015 = 0,985; \qquad Z_\alpha = 2,17$

$$\frac{2,7 - 3}{0,61/\sqrt{51}} = -3,51 < -2,17$$

We reject the null hypothesis: This network fullfils the quality standard.

b. (2 points) When reviewing the work, data collection errors were detected in 2% of the simples. After threatening the responsible of the disaster with the dismissal, a sample of size 60 was taken, obtaining an error percentage of 1,8%. Has the percentage of errors been reduced to the same level of significance?

$H_0 : p_1 = p_2$

$H_1 : p_1 > p_2$

Rechazamos si : $Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 (\frac{1}{n_1} + \frac{1}{n_2})}} > Z_\alpha$

$$\hat{p_0} = \frac{n_1 \hat{p_1} + n_2 \hat{p_2}}{n_1 + n_2} = \frac{51 \cdot 0,02 + 60 \cdot 0,018}{51 + 60} = 0,0189$$

$\hat{q_0} = 0,9811$

$$Z_0 = \frac{0,02 - 0,018}{\sqrt{0,0189 \cdot 0,9811 (\frac{1}{51} + \frac{1}{60})}} = 0,077 < 2,17$$

Aceptamos $H_0$: el porcentaje de errores no ha disminuido.

3. (3 points) Data are available from a sample of students from the Polytechnic School. The information available is:
   - **Average**: Final grade in Advanced Maths
   - **IA**: Academic Index before following the subject
   - **Assistance**: Binary variable that shows whether the student attends class frequently. 1 = less than 5 absences, 0 = 5 or more absences.

   The following R outputs show the summary estatistics of the three variables:

```
summary(IA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.558   6.580   7.170   7.138   7.716   9.788

summary(Promedio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.281   6.856   7.355   7.344   7.815   9.511
```

```
table(Asistencia)
Asistencia
 0  1
21 79
```

A multiple regression model is built with these variables in order to explain variable Average. The result is summarized in the following R output:

```
> summary(lm(Promedio ~ IA + Asistencia))

Call:
lm(formula = Promedio ~ IA + Asistencia)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25934 -0.38176 -0.03849  0.39748  1.41526

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.38034    0.40747   8.296 6.29e-13 ***
IA           0.47299    0.05399   8.761 6.33e-14 ***
Asistencia   0.74319    0.13387   5.552 2.46e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5448 on 97 degrees of freedom
Multiple R-squared:  0.5168,Adjusted R-squared:  0.5068
F-statistic: 51.87 on 2 and 97 DF,  p-value: 4.801e-16
```

a. (1 point) Write the equation for the model and interpret its determination coefficient.

$$Promedio = 3.38 + 0.47IA + 0.74Asistencia$$

El modelo explica el 51,68 % de la variabilidad observada en los promedios.

b. (1,5 points) Interpret all model coefficients and analyze whether they are significant or not to explain variable Average.

Interpretación y análisis de los coeficientes:

- `IA`: 0.47. El coeficiente es significativo, por lo que se acepta que ejerce una influencia en el valor del promedio. Si IA aumenta en un punto, independientemente de si el alumno asiste o no a clase, la nota promedio aumenta en 0.47 puntos por término medio.
- `Asistencia`: 0.74. El coeficiente es significativo, por lo que se acepta que influye en el valor del promedio. Para un mismo IA, el asistir regularmente a clase (`Asistencia`=1) aumenta el promedio en 0.74 respecto de los que no lo hacen (`Asistencia`=0), por término medio.

c. (0,5 points) From the model results can you tell if it is interesting for students to attend class frequently? Justify your answer.

Si, pues independientemente de su IA, asistir a clase supone, por término medio, un aumento de 0.74 puntos en el promedio.