Examen Final de Estadística I, 22 de Junio de 2012.
Grados en ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER.

EXAM RULES: 1) Use separate booklets for each problem. 2) Perform the calculations with at least two significant decimal places. 3) You may not leave the exam during the first 30 minutes. 4) You are not allowed to leave the classroom without handing in the exam.

1. We are interested in the temperature of a thermostat of a certain model of a car at 100km/h. A simple random sample of 30 cars is taken and the temperature of the thermostat at 100km/h is measured. The results are as follows (in degrees Celsius):

65.8	69.4	69.4	69.7	71.5	72.2	74.1	75.4	75.8	76.3
77.2	77.6	77.6	77.9	78.3	78.7	78.9	81.2	81.2	81.7
82.3	82.3	82.4	84.5	84.7	85.2	85.4	88.2	102.5	105.5

(a) **(0.5 points)** Group the data in intervals of the same length starting with $[65, 69)$, and obtain the absolute and relative frequency distributions of the data.

Answer:

Interval	Abs. Frec.	Rel. Frec.
$[65, 69)$	1	1/30
$[69, 73)$	5	5/30
$[73, 77)$	4	4/30
$[77, 81)$	7	7/30
$[81, 85)$	8	8/30
$[85, 89)$	3	3/30
$[89, 93)$	0	0
$[93, 97)$	0	0
$[97, 101)$	0	0
$[101, 105)$	1	1/30
$[105, 109)$	1	1/30
	30	1

(b) **(0.25 points)** What percentage of the observations are between 78 and 81 degrees Celsius?

Answer:

There are three observations, 78.3, 78.7 y 78.9, of a total of 30. Then, 10%.

(c) **(1 point)** Calculate the three sample quartiles. Interpret them.

Answer:

The three sample quartiles are:

$$\begin{aligned} x_{\left(\frac{31}{4}\right)} &= x_{(8)} = 75.4 \\ x_{\left(\frac{31}{2}\right)} &= \frac{x_{(15)} + x_{(16)}}{2} = \frac{78.3 + 78.7}{2} = 78.5 \\ x_{\left(\frac{93}{4}\right)} &= x_{(23)} = 82.4 \end{aligned}$$

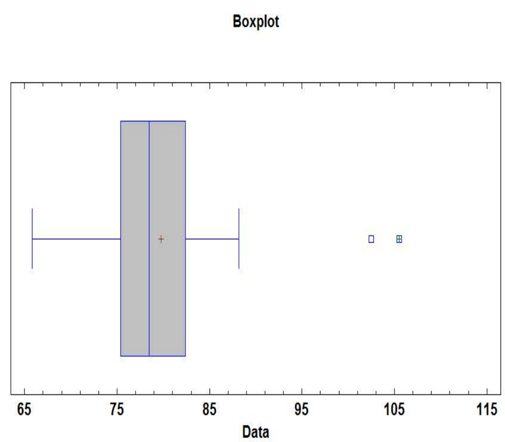
Therefore, the 25% of the observations are smaller than 75.4, the 50% of the observations are smaller than 78.5 and the 75% of the observations are smaller than 82.4. Consequently, the three sample quartiles split the sample in four subsamples that contains approximately the same number of observations.

OBS: The sample quartiels can be also estimated in some alternative ways. For instance:

$$\begin{aligned}
x_{\left(\frac{31}{4}\right)} &= 0.25x_{(7)} + 0.75x_{(8)} = 75.075 \\
x_{\left(\frac{31}{2}\right)} &= \frac{x_{(15)} + x_{(16)}}{2} = \frac{78.3 + 78.7}{2} = 78.5 \\
x_{\left(\frac{93}{4}\right)} &= 0.75x_{(23)} + 0.25x_{(24)} = 82.925
\end{aligned}$$

(d) **(0.75 points)** Draw a boxplot of the data and identify the outliers (if any). Justify your answer.

Answer:



There are two outliers because their values are larger than $Q_3 + 1.5 \times IQR$, where IQR is the interquartile range.

2. Let Y be a continuous random variable defined in the interval $[0, 1]$ with density function:

$$f_Y(y) = \begin{cases} 4y - 4y^3 & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) **(0.5 points)** Obtain the cumulative distribution function of Y.

Answer:

$$F_Y(y) = \begin{cases} 0 & 0 < y \\ \int_0^y \left(4u - 4u^3\right) du = 2y^2 - y^4 = y^2\left(2 - y^2\right) & 0 \le y \le 1 \\ 1 & y > 1 \end{cases}$$

(b) **(0.5 points)** Find the probabilities $P\left(0 < Y < \frac{1}{2}\right)$ and $P\left(\frac{1}{4} < Y < \frac{3}{4}\right)$.

Answer:

$$\begin{aligned}
P\left(0 < Y < \frac{1}{2}\right) &= F_Y\left(\frac{1}{2}\right) - F_Y(0) = \frac{1}{2^2}\left(2 - \frac{1}{2^2}\right) - 0 = \frac{7}{16}. \\
P\left(\frac{1}{4} < Y < \frac{3}{4}\right) &= F_Y\left(\frac{3}{4}\right) - F_Y\left(\frac{1}{4}\right) = \frac{3^2}{4^2}\left(2 - \frac{3^2}{4^2}\right) - \frac{1}{4^2}\left(2 - \frac{1}{4^2}\right) = \frac{11}{16}.
\end{aligned}$$

(c) **(1 point)** Find the expectation and the standard deviation of Y.

Answer:

$$E[Y] = \int_0^1 y\left(4y - 4y^3\right) dy = \left(\frac{4}{3}y^3 - \frac{4}{5}y^5\right)\Big|_{y=0}^{y=1} = \frac{8}{15}.$$

$$V[Y] = E\left[Y^2\right] - E[Y]^2 = \int_0^1 y^2\left(4y - 4y^3\right) dy - \left(\frac{8}{15}\right)^2 = \left(y^4 - \frac{2}{3}y^5\right)\Big|_{y=0}^{y=1} - \frac{64}{225} = \frac{11}{225}.$$

$$DT[Y] = \sqrt{\frac{11}{225}} \simeq 0.2211.$$

(d) **(0.5 points)** Find the expectation and the standard deviation of $2Y + 3$.

Answer:

$$E[2Y+3] = 2E[Y] + 3 = 2\frac{8}{15} + 3 = \frac{17}{5}$$

$$DT[2Y+3] = 2DT[Y] \simeq 0.4422.$$

3. A student takes a multiple-choice test, totally unprepared. The test consists of 40 questions, each with three possible options (only one is correct). To pass the exam, it is necessary to get 20 or more questions right. The student decides to answer each of the questions randomly, in such a way that the answer to one question is not affected by the answers to the other questions.

(a) **(1 point)** What is the probability that the student will answer the first twenty questions correctly and the remaining twenty questions incorrectly?

Answer:

Let:

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th answer is correct} \\ 0 & \text{if the } i\text{-th answer is not correct} \end{cases}$$

for $i = 1, \ldots, 40$. Each X_i has a $\mathrm{Ber}\left(\frac{1}{3}\right)$ distribution. Therefore, because of the independence of the variables,

$$P(X_1 = 1, \ldots, X_{20} = 1, X_{21} = 0, \ldots, X_{40} = 0) = P(X_1 = 1) \cdots P(X_{20} = 1) P(X_{21} = 0) \cdots P(X_{40} = 0) =$$

$$= \left(\frac{1}{3}\right)^{20} \left(\frac{2}{3}\right)^{20} = 8.62 \times 10^{-4}.$$

(b) **(1 point)** What is the exact probability that the student will answer more than 36 questions correctly.

Answer:

The random variable $Y =$"Number of correct answers" is $Y = \sum_{i=1}^{40} X_i$ and has a $\mathrm{Bin}\left(40, \frac{1}{3}\right)$

distribution. Therefore:

$$\begin{aligned} P(Y>36) &= P(Y=37)+P(Y=38)+P(Y=39)+P(Y=40) = \\ &= \binom{40}{37}\left(\frac{1}{3}\right)^{37}\left(\frac{2}{3}\right)^{3}+\binom{40}{38}\left(\frac{1}{3}\right)^{38}\left(\frac{2}{3}\right)^{2}+\binom{40}{39}\left(\frac{1}{3}\right)^{39}\left(\frac{2}{3}\right)+\binom{40}{40}\left(\frac{1}{3}\right)^{40} = \\ &= 6.5012\times10^{-15}+2.5662\times10^{-16}+6.5802\times10^{-18}+8.2252\times10^{-20} = \\ &= 6.7645\times10^{-15}. \end{aligned}$$

(c) **(0.5 points)** Based on the Central Limit Theorem, what is the approximate probability that the student will answer more than 20 questions correctly.

Answer:

As $Y \sim \text{Bin}\left(40, \frac{1}{3}\right)$, then:

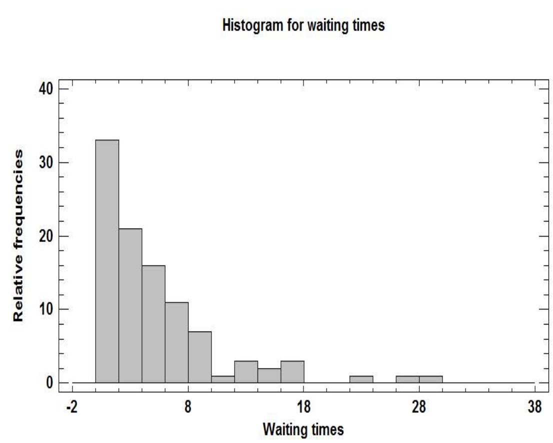
$$E[Y] = 40\times\frac{1}{3} = \frac{40}{3}$$
$$DT[Y] = \left(40\times\frac{1}{3}\times\frac{2}{3}\right)^{1/2} = \left(\frac{80}{9}\right)^{1/2} = \frac{80^{1/2}}{3}.$$

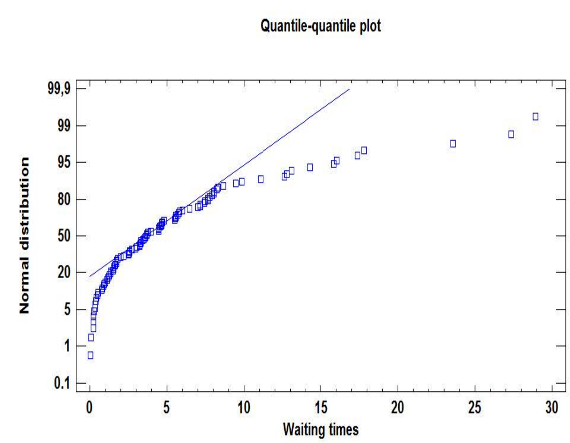
Therefore:

$$\begin{aligned} P(Y>20) &= P\left(\frac{Y-\frac{40}{3}}{\frac{80^{1/2}}{3}} > \frac{20-\frac{40}{3}}{\frac{80^{1/2}}{3}}\right) \simeq P\left(Z>\frac{20}{80^{1/2}}\right) = P(Z>2.2360) = \\ &= 1-P(Z\le 2.2360) = 1-0.9871 = 0.0129, \end{aligned}$$

where $Z \sim N(0,1)$.

4. The owners of a shopping center are particularly interested in customers' satisfaction. With this in mind, they carry out a survey in which, among other indicators, the data on the waiting times (in minutes) to access the center's elevators was collected. The data led to the following two graphs:





(a) **(0.75 points)** Is it reasonable to say that the waiting times follow a Normal distribution?

Answer:

The plots clearly show that the Normal distribution is not adequate to describe the data properly. First, the histogram is clearly skewed, while the qq-plot clearly shows that the sample quantiles are not distribution along a straight line when compared with the quantiles of a Normal distribution.

(b) **(1 point)** The owners of the shopping center claim that the waiting times have a mean of 6 minutes and a standard deviation of 5 minutes. If 50 people take the elevators independently, what is the probability that the sum of their waiting times will be between 5.5 and 6 hours.

Answer:

Let T be the random variable "Waiting time to access the elevators". We know that $E[T] = 6$ and $DT[T] = 5$. Then, from the CLT:

$$Z = \frac{\overline{T} - 6}{\frac{5}{\sqrt{50}}} \underset{aprox.}{\sim} N(0, 1).$$

Therefore,

$$P\left(330 < \sum_{i=1}^{50} T_i < 360\right) = P\left(\frac{330}{50} < \overline{T} < \frac{360}{50}\right) = P\left(6.6 < \overline{T} < 7.2\right) =$$
$$= P\left(\frac{6.6 - 6}{\frac{5}{\sqrt{50}}} < Z < \frac{7.2 - 6}{\frac{5}{\sqrt{50}}}\right) = P(0.8485 < Z < 1.6970) =$$
$$= P(Z < 1.6970) - P(Z < 0.8485) = 0.9545 - 0.7995 = 0.1550.$$

(c) **(0.75 points)** Assume that the waiting times have a mean of 6 minutes and a standard deviation of 5 minutes. Obtain the lower bound on the probability, that the total of the waiting times for 25 people will be between 2 and 3 hours (use the Chebyshev's inequality, i.e., for a random variable X with expectation μ and variance σ^2, then $P(|X - \mu| < k) \geq 1 - \frac{\sigma^2}{k^2}$, for any positive constant k).

Answer:

Using the Chebyshev's inequality, we have that the random variable $\sum_{i=1}^{25} T_i$ has expectation $25 \times 6 = 150$ and variance $25 \times 25 = 625$. Consequently:

$$P\left(120 < \sum_{i=1}^{25} T_i < 180\right) = P\left(150 - 30 < \sum_{i=1}^{25} T_i < 150 + 30\right) \geq 1 - \frac{625}{900} = 0.3055.$$