Statistics I Final Exam, 24 June 2014.
Degrees in 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. The following data set represents the number of sold cars in a certain month in 23 car concessionaires:

   9 8 14 23 23 27 18 18 8 21 20 18 14 20 29 18 33 19 23 24 24 14 19

   (a) (0.5 points) Is the monthly sales mean of this data set larger than 20? Is the monthly sales median of this data set larger than 20? Is it true that the most frequent monthly sale is 20? Justify your answers.

   (b) (0.5 points) Compute the sample quasi-variance and coefficient of variation of the data set.

   (c) (1 point) Draw a boxplot for the data set, obtaining the appropriate numerical measures for that. Is it true that the distribution shape is heavily positive skewed? Justify your answer.

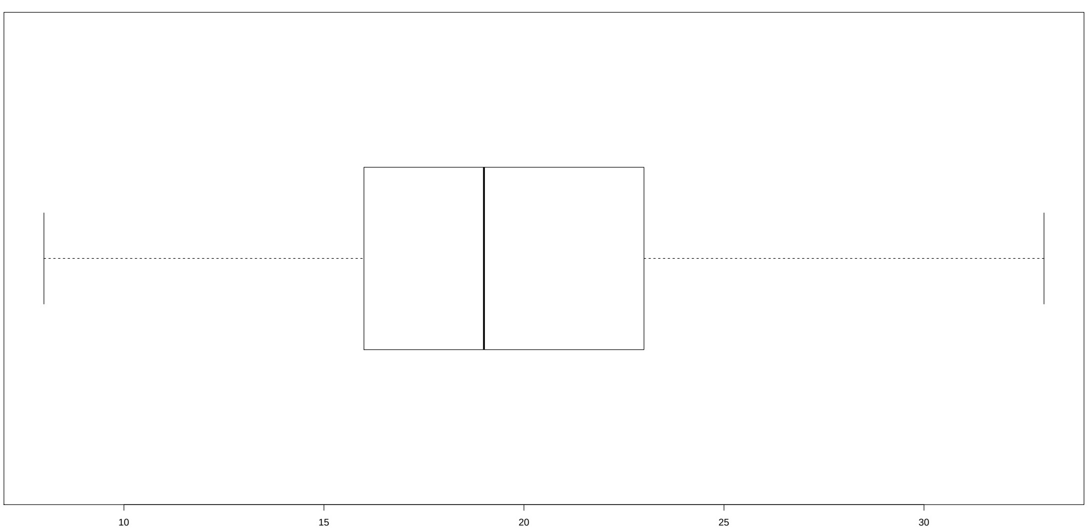
   (d) (0.5 points) Based on the boxplot drawn in the previous item, are there outliers in the data set? Justify your answer.

Solution.

(a) The sample mean is $\overline{x} = 19.30$. The sample median is $M = 19$. The sample mode is 18. Then, the answer is NO for all the questions.

(b) The sample quasi-variance is $\widehat{\sigma}^2 = 40.13043$. The sample coefficient of variation is 0.3281.

(c) The quartiles are $Q_1 = 14$, $Q_2 = 19$ and $Q_3 = 23$, respectively. Then, the sample IQR is $IQR = 23 - 14 = 9$. The boxplot is:



The distribution is slightly right skewed. Therefore, it is NOT heavily positive skewed.

(d) On the one hand, we have $Q_1 - 1.5IQR = 14 - 1.5 \times 9 = 0.5$. On the other hand, we have $Q_3 + 1.5IQR = 23 + 1.5 \times 9 = 36.5$. As the minimum and maximum of values in the sample are 8 and 33, respectively, and $0.5 < 8$ and $36.5 > 33$, we conclude that there are no outliers in the sample.

2. A young entrepreneur is starting a business in a city. Following certain information about similar businesses in the city, it is known that the 25% have annual income below 1 million euros (included), the 50% of them have annual income between 1 (not included) and 2 million euros (included), and the rest have annual income between 2 (not included) and 3 million euros (included). Given the previous information, and assuming that businesses are uniformly distributed among each of the considered anual income intervals, answer the following questions:

   (a) (0.75 points) Obtain the cumulative distribution function of the business annual income (in million euros).

   (b) (0.25 points) Obtain the probability that the business annual income is between 0.5 (not included) and 1.5 million euros (included).

   (c) (0.75 points) Compute the expected value and the standard deviation of the business annual income.

   (d) (0.75 points) If the business is successful, the young entrepreneur plans to start another business smaller than the previous one and with annual income equal to a quarter of the first business annual income. Which are the expected annual income (in million euros) of the second business? and the standard deviation?

Solution.

(a) Let X be the business annual income (in million euros). Then, the probability density function is given by:

$$f(x) = \begin{cases} \frac{1}{4} & x \leq 1 \\ \frac{1}{2} & 1 < x \leq 2 \\ \frac{1}{4} & 2 < x \leq 3 \\ 0 & \text{en otro caso} \end{cases}$$

Therefore, the cumulative distribution function can be computed as follows:

- For $x \leq 0$,

$$F(x) = 0$$

- For $0 < x \leq 1$,

$$F(x) = \int_0^x \frac{1}{4} dx = \frac{x}{4}$$

- For $1 < x \leq 2$,

$$F(x) = \int_0^1 \frac{1}{4} dx + \int_1^x \frac{1}{2} dx = \frac{1}{4} + \frac{1}{2}(x-1) = \frac{1}{2}\left(x - \frac{1}{2}\right)$$

- For $2 < x \leq 3$,

$$F(x) = \int_0^1 \frac{1}{4} dx + \int_1^2 \frac{1}{2} dx + \int_2^x \frac{1}{4} dx = \\ = \frac{1}{4} + \frac{1}{2} + \frac{1}{4}(x-2) = \frac{1}{4}(x+1)$$

- For $3 < x$,

$$F(x) = 1$$

In summary,

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{4} & 0 < x \leq 1 \\ \frac{1}{2}\left(x - \frac{1}{2}\right) & 1 < x \leq 2 \\ \frac{1}{4}(x+1) & 2 < x \leq 3 \\ 1 & 3 < x \end{cases}$$

(b) The probability that the business annual income is between 0.5 (not included) and 1.5 million euros (included) is given by:

$$\Pr(0.5 < x \leq 1.5) = F(1.5) - F(0.5) = \frac{1}{2} - \frac{1}{8} = \frac{3}{8} = 0.375$$

(c) The expectation of X is given by:

$$E[X] = \int_0^3 x f(x)\, dx = \int_0^1 \frac{x}{4} dx + \int_1^2 \frac{x}{2} dx + \int_2^3 \frac{x}{4} dx =$$
$$= \frac{x^2}{8}{}_{|x=0}^{|x=1} + \frac{x^2}{4}{}_{|x=1}^{|x=2} + \frac{x^2}{8}{}_{|x=2}^{|x=3} = \frac{1}{8} + \frac{3}{4} + \frac{5}{8} = \frac{3}{2} = 1.5 \text{ million euros}$$

The variance of X is given by:

$$V[X] = E\left[X^2\right] - E[X]^2$$

where:

$$E\left[X^2\right] = \int_0^3 x^2 f(x)\, dx = \int_0^1 \frac{x^2}{4} dx + \int_1^2 \frac{x^2}{2} dx + \int_2^3 \frac{x^2}{4} dx =$$
$$= \frac{x^3}{12}{}_{|x=0}^{|x=1} + \frac{x^3}{6}{}_{|x=1}^{|x=2} + \frac{x^3}{12}{}_{|x=2}^{|x=3} = \frac{1}{12} + \frac{7}{6} + \frac{19}{12} = \frac{17}{6}$$

Therefore,

$$V[X] = \frac{17}{6} - \frac{9}{4} = \frac{7}{12} \simeq 0.5833$$

and the standard deviation is $\sqrt{\frac{7}{12}} \simeq 0.7637$ million euros.

(d) The expected annual income of the second business and its standard deviation are the quarter of those the first business. Therefore, the expected value is $\frac{3}{8} = 0.375$ million euros while the standard deviation is 0.1909.

3. Each workday a sales agent calls 25 independent households to try sell them a new product, earning a commission of 45 euros per sale. Based on experience, the agent makes, on average, a sale on one out of 6 calls when calling a household with school-age children, and a sale on one out of 10 calls when calling a household without school-age children. The agent does not know in advance the type of household for each call, but knows that 30% of households have school-age children. Obtain:

   (a) (0.5 points) The probability that a call results in a sale.

   (b) (0.5 points) The probability that the number of sales in a day is 3 or less.

   (c) (0.5 points) The mean, the variance and the standard deviation of the total sales commission earned in a day.

   (d) (1 point) The probability (or an approximation) that the agent earns in a workweek (5 workdays) more than 1000 euros in sales commissions.

Solution.

(a) By the Law of Total Probability, writing "school-age children" as SAE, the probability that a call results in a sale is:

$$\begin{aligned} \Pr(\text{Sale}) &= \Pr(\text{Sale}|\text{SAE}) \Pr(\text{SAE}) + \Pr(\text{Sale}|\text{No SAE}) \Pr(\text{No SAE}) = \\ &= \frac{1}{6} \times \frac{3}{10} + \frac{1}{10} \times \frac{7}{10} = 0.12 \end{aligned}$$

(b) The number of sales in a day is $N \sim Bin(25, 0.12)$. Hence, the probability that the number of sales in a day is 3 or less is given by:

$$\Pr(N \leq 3) = \Pr(N=0) + \Pr(N=1) + \Pr(N=2) + \Pr(N=3) =$$
$$= \binom{25}{0} 0.12^0 (1-0.12)^{25} + \binom{25}{1} 0.12^1 (1-0.12)^{24} +$$
$$+ \binom{25}{2} 0.12^2 (1-0.12)^{23} + \binom{25}{3} 0.12^3 (1-0.12)^{22} =$$
$$= 1 \times 0.0409 + 25 \times 0.0056 + 300 \times 0.0008 + 2300 \times 0.0001 =$$
$$\simeq 0.6475$$

(c) The mean, variance and standard deviation of N is given by:

$$\begin{aligned} E[N] &= 25 \times 0.12 = 3 \\ V[N] &= 25 \times 0.12 \times 0.88 = 2.64 \\ DT[N] &= \sqrt{2.64} = 1.6248 \end{aligned}$$

Therefore, the mean, variance and standard deviation of the total sales commission earned in a day are given by:

$$\begin{aligned} E[45N] &= 45 \times E[N] = 45 \times 3 = 135 \text{ euros} \\ V[45N] &= 45^2 \times V[N] = 5346 \text{ euros}^2 \\ DT[45N] &= 45 \times DT[N] = 45 \times 1.6248 = 73.116 \text{ euros} \end{aligned}$$

(d) The number of sales in a workweek is $M \sim Bin(125, 0.12)$. Therefore, the total commission earned in a workweek is $W = 45M$. The mean, variance and standard deviation of M is given by:

$$\begin{aligned} E[M] &= 125 \times 0.12 = 15 \\ V[M] &= 125 \times 0.12 \times 0.88 = 13.2 \\ DT[M] &= \sqrt{13.2} = 3.6331 \end{aligned}$$

Using the CLT, we obtain:

$$\begin{aligned} &\Pr(W > 1000) = \Pr(45M > 1000) = \Pr(M > 22.22) = \\ &= \Pr\left(\frac{M - 15}{3.6331} > \frac{22.22 - 15}{3.6331}\right) \simeq \Pr(Z > 1.9872) = \\ &= 1 - \Pr(Z < 1.9872) = 1 - 0.9761 = 0.0239 \end{aligned}$$

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

4. The clients of a certain bank may deposit or withdraw money. Let X be a transaction made by a client. Then, X is positive, if the money is deposited, or negative, if the money is withdrawn. The bank analysts assume that the expectation and the standard deviation of X are 0 and 100 euros, respectively.

(a) (1 point) Compute the probability that the average transaction made by 100 independent clients is between -3 and 3 euros.

(b) (0.5 points) Without doing any additional calculations, decide if the probability in (a) will increase or decrease if we increase the sample size. Justify your answer.

(c) (0.5 points) Let assume that exactly 100 clients will visit the bank in one day. Compute the probability that the sum of transactions at the end of the day is positive.

(d) (0.5 points) Even though the bank assumed that the true expectation of X equals 0, the bank analysts suspect that lately the clients are mostly depositing and not withdrawing money. To check this suspicion, the analysts collect a sample of 25 transactions made by independent clients and obtain a sample mean of 50 euros. Obtain a 90% confidence interval for the mean assuming that X follows a normal distribution with standard deviation 100. How confident are you that the true mean is in fact larger than zero?

Solution:

(a) Since $n \geq 30$, by CLT we have that $\overline{X} \sim N\left(0, \frac{100}{\sqrt{n}}\right) = N(0, 10)$. Then, the required probability is given by:

$$\begin{aligned} &\Pr\left(-3 < \overline{X} < 3\right) = \Pr\left(\frac{-3}{10} < \frac{\overline{X}}{10} < \frac{3}{10}\right) \simeq \Pr(-0.3 < Z < 0.3) = \\ &= \Pr(Z < 0.3) - \Pr(Z < -0.3) = 0.6179 - 0.3821 = 0.2358 \end{aligned}$$

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

(b) If the sample size increases, the variance of $\overline{X}$ decreases. Therefore, the probability for $\overline{X}$ to be in the interval $(-3, 3)$ will increase.

(c) We want to compute the following probability:

$$\Pr\left(\sum_{i=1}^{n} X_i > 0\right)$$

This is similar to:

$$\Pr\left(\overline{X} > 0\right)$$

Then, as in a), by CLT we have that $\overline{X} \sim N\left(0, \frac{100}{\sqrt{n}}\right) = N(0, 10)$. Therefore, the required probability is given by:

$$\Pr\left(\sum_{i=1}^{n} X_i > 0\right) = \Pr\left(\overline{X} > 0\right) = \Pr\left(\frac{\overline{X}}{10} > 0\right) \simeq \Pr(Z > 0) = 0.5$$

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

(d) The confidence interval for the mean is given by:

$$\left(50 - 1.65\frac{100}{\sqrt{25}}, 50 + 1.65\frac{100}{\sqrt{25}}\right) = (17, 83)$$

Consequently, we are more than 90% confident that the true mean is positive.