**Statistics I Final Exam, 25 June 2013.**
**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. Consider the following variables taken from 47 renowned Business Schools of the United States in 2004: mean starting salary, expressed in dollars, percentage of new graduates employees at the end of their studies (% employees), score given by external evaluators to the university studies (score) and percentage of accepted applicants in the Business School (% acceptance)

   (a) (0.5 points) The following table contains a statistical summary of the four variables:

| | Mean Starting Salary (in $) | % employees | % acceptance | score |
|---|---|---|---|---|
| Mininum | 66340 | 42 | 9.2 | 3 |
| Maximum | 107320 | 77.60 | 56.6 | 4.6 |
| Percentile 10% | 70245 | 50.08 | 15.12 | 3.2 |
| Percentile 25% | 75482 | 55.7 | 22.8 | 3.4 |
| Percentile 50% | 82357 | 62.2 | 31 | 3.6 |
| Percentile 75% | 92657 | 68 | 38.25 | 4.1 |
| Percentile 90% | 98611 | 70.74 | 46.3 | 4.42 |

   Obtain the interquartile range for all the variables.

   (b) (0.75 points) Draw boxplots for all the variables. Based on them, which are the most important features of the variables in terms of shape? Are there outliers? Justify your answer.

   (c) (0.75 points) The following table contains the absolute frequencies of the variable "mean starting salary" at certain classes:

| mean starting salary | Absolute frequency |
|---|---|
| $[65000, 71000)$ | 6 |
| $[71000, 77100)$ | 9 |
| $[77100, 83200)$ | 10 |
| $[83200, 89300)$ | 6 |
| $[89300, 95400)$ | 6 |
| $[95400, 101500)$ | 8 |
| $[101500, 107600)$ | 2 |

   Draw an histogram of the variable using the classes defined in the previous table. Describe the main features of the histogram.

   (d) (0.5 points) The dataset also contains a last categorical variable that classify each university into one of three categories "Normal", "High" and "Very high" in terms of the general quality of the university. We select a university in each category and observe the percentage of accepted applicants in the Business School. The obtained values are 18.2 for "Very high", 28 for "High" and 36.9 for "Normal". Decide and justify which of the universities is larger in relation with the percentage of accepted applicants in universities in the same category considering the summary of the variable "accepted applicants" in the following table:

| | Normal | High | Very High |
|---|---|---|---|
| Mean | 37.67 | 26.93 | 17.3 |
| Median | 36.9 | 27.65 | 17.05 |
| Standard deviation | 8.87 | 8.21 | 5.31 |
| Minimum | 24.3 | 12.8 | 9.2 |
| Maximum | 56.6 | 40 | 25.3 |
| Range | 32.3 | 27.2 | 16.1 |

**Solution.**

(a) The interquartile ranges are given by:

$$\begin{aligned} RIC &= Q_3 - Q_1 = 92657 - 75482 = 17175 \\ RIC &= Q_3 - Q_1 = 68 - 55.7 = 12.3 \\ RIC &= Q_3 - Q_1 = 38.25 - 22.8 = 15.45 \\ RIC &= Q_3 - Q_1 = 4.1 - 3.4 = 0.7 \end{aligned}$$

respectively.

(b) The five points of the boxplot for mean starting salary (in $) are:

$$\begin{aligned} x_{(\min)} &= 66340 \\ Q_1 &= 75482 \\ Q_2 &= 82357 \\ Q_3 &= 92657 \\ x_{(\max)} &= 107320 \end{aligned}$$

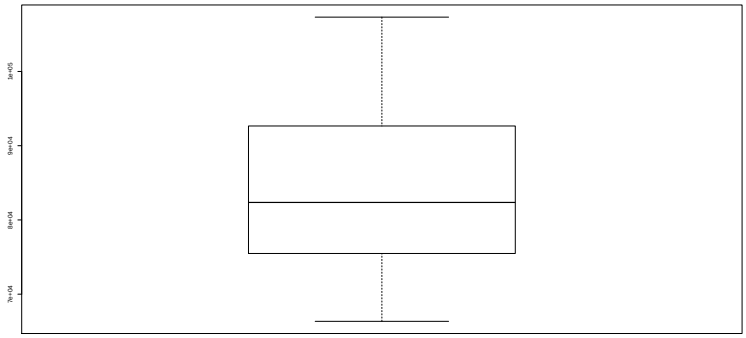


Figure 1: Boxplot for mean starting salary (in $)

The distance between the median and the first quartile is smaller than the distance between the median and the third quartile. The variable is positively skewed. Moreover, as $x_{(\min)} = 66340$ and $x_{(\max)} = 107320$, then, there are no outliers in mean starting salary (in $) because:

$$\begin{aligned} Q_1 - 1.5 \times RIC &= 49719.5 \\ Q_3 + 1.5 \times RIC &= 118419.5 \end{aligned}$$

The five points of the boxplot for % employees are:

$$\begin{aligned} x_{(\min)} &= 42 \\ Q_1 &= 55.7 \\ Q_2 &= 62.2 \\ Q_3 &= 68 \\ x_{(\max)} &= 77.6 \end{aligned}$$

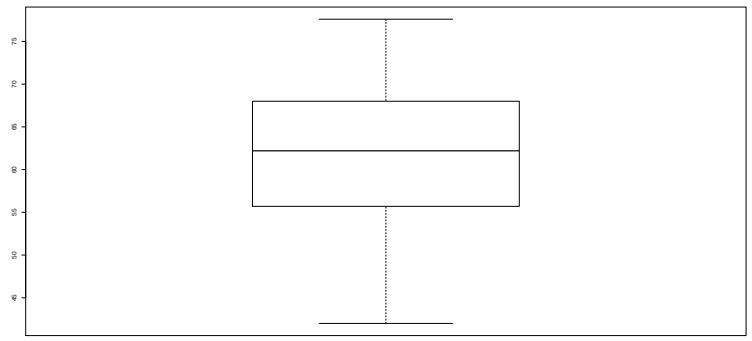


Figure 2: Boxplot for % employees

The distance between the median and the first quartile is larger than the distance between the median and the third quartile. The variable is slightly negatively skewed. Moreover, as $x_{(\min)} = 42$ and $x_{(\max)} = 77.6$, then, there are no outliers in % employees because:

$$\begin{aligned} Q_1 - 1.5 \times RIC &= 37.25 \\ Q_3 + 1.5 \times RIC &= 86.45 \end{aligned}$$

The five points of the boxplot for % acceptance are:

$$\begin{aligned} x_{(\min)} &= 9.2 \\ Q_1 &= 22.8 \\ Q_2 &= 31 \\ Q_3 &= 38.25 \\ x_{(\max)} &= 56.6 \end{aligned}$$

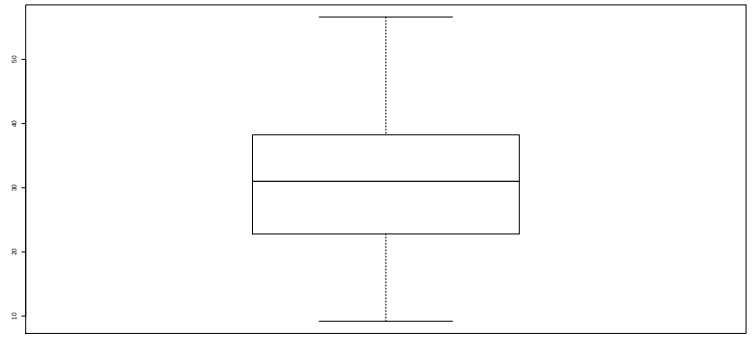


Figure 3: Boxplot for % acceptance

The distance between the median and the first quartile is slightly larger than the distance between the median and the third quartile. The variable is slightly positively skewed. Moreover, as $x_{(\min)} = 9.2$ and $x_{(\max)} = 56.6$, then, there are no outliers in % acceptance because:

$$\begin{aligned} Q_1 - 1.5 \times RIC &= -0.375 \\ Q_3 + 1.5 \times RIC &= 61.425 \end{aligned}$$

The five points of the boxplot for score are:

$$\begin{aligned} x_{(\min)} &= 3 \\ Q_1 &= 3.4 \\ Q_2 &= 3.6 \\ Q_3 &= 4.1 \\ x_{(\max)} &= 4.6 \end{aligned}$$

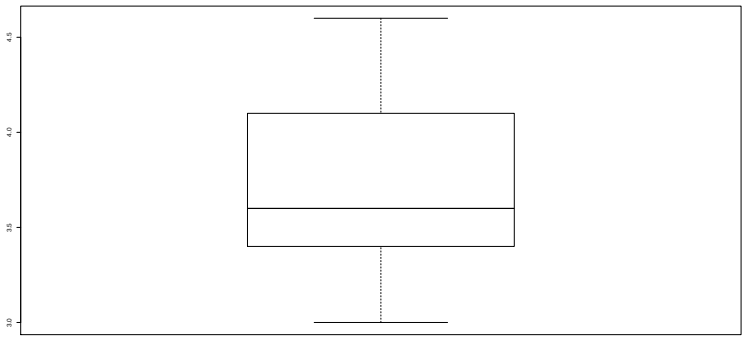



Figure 4: Boxplot for score

The distance between the median and the first quartile is smaller than the distance between the median and the third quartile. The variable is positively skewed. Moreover, as $x_{(\min)} = 3$ and $x_{(\max)} = 4.6$, then, there are no outliers in score because:

$$\begin{aligned} Q_1 - 1.5 \times RIC &= 2.35 \\ Q_3 + 1.5 \times RIC &= 5.15 \end{aligned}$$

(c) The histogram shows a clear bimodality suggesting the presence of two different populations. Indeed, this is the case, "Normal" and "High" qualifications on one hand and "Very High" qualifications on the other hand.

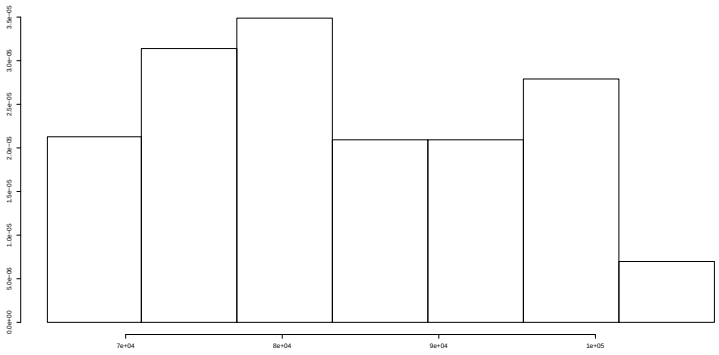



Figure 5: Histogram for mean starting salary

(d) First, we standardize the obtained rates as follows:

$$z_{Normal} = \frac{36.9 - 37.67}{8.87} = -0.086,\ z_{High} = \frac{28 - 26.93}{8.21} = 0.131,\ z_{VeryHigh} = \frac{18.2 - 17.3}{5.31} = 0.169$$

The one with larger acceptance rate is the category "Very High", with a standardized acceptance rate of 0.169.

2. Consider a random variable $X$ with cumulative distribution function:

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.25 & 0 \leq x < 2 \\ 0.75 & 2 \leq x < 7 \\ 1 & x \geq 7 \end{cases}$$

(a) (0.5 points) Obtain the probability function of the random variable $X$.

(b) (0.5 points) Obtain $P(0 < X < 3)$ and $P(3 \leq X < 7)$.

(c) (0.5 points) Obtain the mean and variance of the random variable $X$.

(d) (0.5 points) Say TRUE or FALSE for the following statement and justify your answer:

$$P(X \leq 2.5) = P(X < 2.5).$$

(e) (0.5 points) Let $Y = 2X - 1$ be a new variable. Obtain the mean and variance of the new random variable.

**Solution.**

(a) The probability function of the random variable $X$ is given by:

$$p(X) = \begin{cases} 0.25 & x = 0 \\ 0.5 & x = 2 \\ 0.25 & x = 7 \end{cases}$$

(b) On one hand

$$P(0 < X < 3) = P(X = 2) = 0.5$$

On the other hand,

$$P(3 \leq X < 7) = 0$$

(c) On one hand,

$$E[X] = 0 \times 0.25 + 2 \times 0.5 + 7 \times 0.25 = 2.75$$

On the other hand,

$$Var[X] = E\left[X^2\right] - E[X]^2 = 0^2 \times 0.25 + 2^2 \times 0.5 + 7^2 \times 0.25 - 2.75^2 = 14.25 - 7.5625 = 6.6875.$$

(d) The statement is TRUE because $X$ is a discrete random variable and has no weight at $x = 2.5$.

(e) On one hand:

$$E[Y] = E[2X - 1] = 2E[X] - 1 = 2 \times 2.75 - 1 = 4.5$$

On the other hand:

$$V[Y] = V[2X - 1] = 2^2 V[X] = 4 \times 6.6875 = 26.75$$

3. A certain company distributes mobile phones that are manufactured in two plants, A and B. This company packs the devices in packages with 5 mobile phones to distribute them at different stores.

(a) (1 point) It is known that the 2% of the phones that comes from plant A and the 1% of the phones that comes from plant B are defective. Given that the 20% of the phones comes from A and the 80% of the phones comes from B, obtain the probability that a mobile phone is defective.

(b) (0.5 points) Let $X$ be the number of defective mobile phones in a single package. Obtain the distribution of $X$ and identify the parameters of the distribution.

(c) (0.5 points) A certain store ask the company for 100 packages. Based on paragraph a), obtain the (approximated) probability that more than 10 mobile phones in the 100 packages are defective.

(d) (0.5 points) Obtain the (exact) probability that more than one of the 500 mobile phones sent to the store are defective. Approximate the same probability using the Central Limit Theorem. Is this a good approximation? Justify your answer.

**Solution.**

(a) Using the Total Probability Law,

$$p = 0.02 \times 0.2 + 0.01 \times 0.8 = 0.012$$

(b) We have $X \sim Bin\,(5, 0.012)$.

(c) Let $Y$ be the number of defective mobile phones in 100 packages. Then, $Y \sim Bin\,(500, 0.012)$ and

$$\Pr\,(Y > 10) = \Pr\left(\frac{Y-6}{2.4347} > \frac{10-6}{2.4347}\right) = \Pr\,(Z > 1.64) \simeq 0.05,$$

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

(d) Let $Y$ be the number of defective mobile phones in 100 packages. Then, $Y \sim Bin\,(500, 0.012)$ and

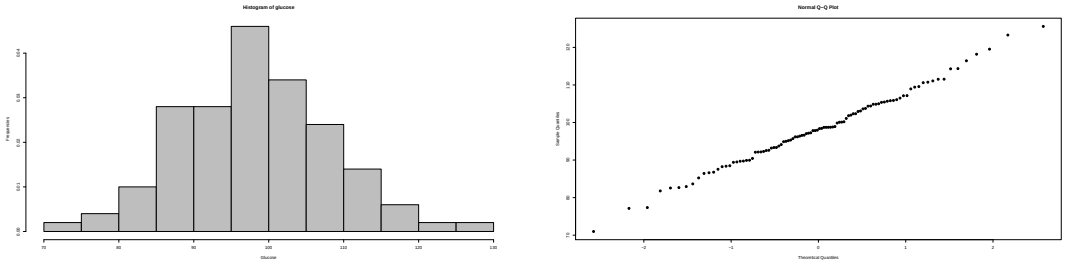
$$\Pr\,(Y > 1) = 1 - \Pr\,(Y \leq 1) = 1 - \Pr\,(Y = 0) - \Pr\,(Y = 1) =$$
$$= 1 - \binom{500}{0} 0.012^0\,(1-0.012)^{500} - \binom{500}{1} 0.012^1\,(1-0.012)^{499} =$$
$$= 1 - 0.0024 - 0.0169 = 0.9807.$$

On the other hand, this probability can be approximated by the CLT as done in c). Then:

$$\Pr\,(Y > 1) = \Pr\left(\frac{Y-6}{2.4347} > \frac{1-6}{2.4347}\right) = \Pr\,(Z > -2.05) \simeq 0.9798.$$

The approximation is good enough because $n$ is large and because $500 \times 0.012 \geq 5$ and $500 \times 0.988 \geq 5$.

4. A simple random sample of 100 people is selected and their glucose in blood (in mg/dL) was measured.




(a) (0.5 points) Can the glucose be described by a Normal distribution? Justify your answer.

(b) (1 point) The collected sample has the sample mean of 98.36 and variance of 100.288. Calculate the 95% confidence interval for the mean glucose.

(c) (0.5 points) Reason without doing any calculations if the width of a 90% confidence interval will be smaller, equal or greater than the one obtained with a 95% confidence level.

(d) (0.5 points) According to the experts the glucose in blood has the mean 98 and standard deviation of 10. Obtain the probability that the sum of the glucose in blood of the 100 people is larger than 9900 mg/dL.

**Solution:**

(a) Yes. The histogram is approximately symmetric and may fit well the normal density. The sample quantiles of the qq-plot are also fitting the line approximately.

(b) We have $z_{.975} = 1.96$. Hence the confidence interval equals $(96.39, 100.32)$.

(c) The 90% confidence interval will be narrower, as the confidence level is smaller.

(d) If $X_i$ is the glucose in blood for person $i$, then, we have:

$$\Pr\left(\sum_{i=1}^{100} X_i > 9900\right) = \Pr\left(\overline{X} > 99\right) = \Pr\,(Z > 1) \simeq .8413$$

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