Statistics I Final Exam, 27 May 2015.
Degrees in ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER, TUR.

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. Vendors doing business with a particular company were sampled to determine the economic impact of company business on their gross sales. A sample of 15 firms that provide services to the company had the following percentages of total annual sales as a result of sales to the company:

   27 12 14.9 1.2 0.1 1 0.1 5.3 7.6 5 1 1 3.2 3 7

   (a) (0.5 points) Is the sample mean of the 15 percentages larger than the sample median? If true, what does this result suggest? Justify your answers.

   (b) (0.5 points) Calculate the three sample quartiles. Interpret them in terms of the percentages.

   (c) (0.5 points) Compute the sample quasi-variance and coefficient of variation of the 15 percentages.

   (d) (1 point) Draw a boxplot of the data and identify the outliers (if any). Justify your answer.

   Solution.

   (a) The sample mean of the 15 percentages is $\overline{x} = 5.96$, while the sample median is $M = 3.2$. Then, the sample mean is larger than the sample median. This result suggests that the data distribution is positively skewed, i.e., there are a few firms for which the percentages of total annual sales as a result of sales to the company are higher than others.

   (b) The sample quartiles are $Q_1 = x_{(4)} = 1$, $Q_2 = x_{(8)} = 3.2$ and $Q_3 = x_{(12)} = 7.6$, respectively. Therefore, the 25% of the percentages are smaller than 1%, the 50% of the observations are smaller than 3.2% and the 75% of the observations are smaller than 7.6%. Consequently, the three sample quartiles split the sample in four subsamples that contains approximately the same number of percentages. In general, the percentages of total annual sales as a result of sales to the company for most of the firms are less than the 7.6%.

   (c) The sample quasi-variance is $\widehat{\sigma}^2 = 53.26686$, while the sample coefficient of variation is $CV = 1.224566$.

   (d) To build the boxplot, we need the IQR which is given by $IQR = Q_3 - Q_1 = 7.6 - 1 = 6.6$. Moreover, to construct the boxplot fences and to detect outliers, if any, $Q_1 - 1.5IQR = 1 - 1.5 \times 6.6 = -8.9$ while $Q_3 + 1.5IQR = 7.6 + 1.5 \times 6.6 = 17.5$. Additionally, the minimum and maximum of values in the sample are 0.1 and 27, respectively. Therefore, there is a single outlier because $17.5 < 27$. The boxplot is in Figure 1:

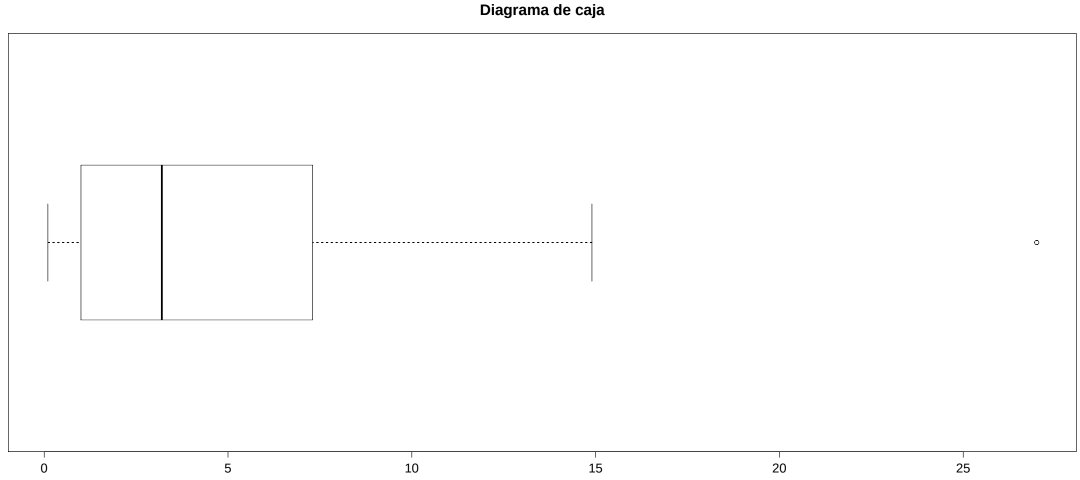
2. A sample of 500 persons is questioned regarding political affiliation and attitude toward a proposed national health care plan in the USA. The responses are cross-classified according to the political affiliation and opinion categories displayed in the following table:

		Attitude		
Affiliation	Favor	Indifferent	Oppose	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

   Given the previous information, answer the following questions:

   (a) (1 point) Find the joint (relative) frequency distribution of two variables and the two marginal (relative) frequency distributions.

   (b) (0.5 points) Find the conditional distribution of Attitude given Affiliation. Which attitude is more frequent among Republicans? Justify your answers.



(c) (0.5 points) Find the conditional distribution of Affiliation given Attitude. Which political affiliation has larger favor attitude toward the proposed national health care plan? Justify your answers.

(d) (0.5 points) Is it correct to say that among the people that do not oppose to the proposed national health care plan, the relative frequency of Democrat voters is larger than the frequency of Republican voters? Justify your answer.

Solution.

(a) The joint (relative) frequency distribution of two variables and the two marginal (relative) frequency distributions are given in the following table:

		Attitude		
Affiliation	Favor	Indifferent	Oppose	Total
Democrat	0.276	0.166	0.128	0.57
Republican	0.128	0.134	0.168	0.43
Total	0.404	0.3	0.296	1

(b) The conditional distributions of Attitude given Affiliation is given by:

Attitude|Affiliation	Favor	Indifferent	Oppose
$f_{\cdot\|Democrat}$	0.4842105	0.2912281	0.2245614
$f_{\cdot\|Republican}$	0.2976744	0.3116279	0.3906977

Then, the attitude more frequent among Republicans is Oppose.

(c) The conditional distributions of Affiliation given Attitude is given by:

Affiliation|Attitude	$f_{\cdot\|Favor}$	$f_{\cdot\|Indifferent}$	$f_{\cdot\|Oppose}$
Democrat	0.6831683	0.5533333	0.4324324
Republican	0.3168317	0.4466667	0.5675676

Then, the political affiliation with larger favor attitude toward the proposed national health care plan is Democrat.

(d) The answer is Yes. Note that,

Affiliation|Attitude	$f_{\cdot\|NoOppose}$	$f_{\cdot\|Oppose}$
Democrat	0.6278409	0.4324324
Republican	0.3721591	0.5675676

Thus, among the people that do not oppose to the proposed national health care plan, the relative frequency of Democrat voters (0.6278409) is larger than the frequency of Republican voters (0.3721591).

3. The number of overnight emergency calls to the answering service of a heating and air conditioning firm has the probabilities 0.05, 0.1, 0.15, 0.35, 0.2 and 0.15 for 0, 1, 2, 3, 4, and 5 calls per night, respectively.

   (a) (0.5 points) Obtain the probability function of the overnight emergency calls in one night. Which is the probability that the number of overnight emergency calls in one night is larger than 3?

   (b) (0.5 points) Which is the mean number of overnight emergency calls per night? and the variance?

   (c) (0.5 points) Assuming that the the numbers of overnight emergency calls in different days are independent, which is the mean number of overnight emergency calls per week? and the variance?

   (d) (1 point) Obtain the probability (or an approximation) that the number of overnight emergency calls per night in a year (365 days) is larger than 1300.

Solution.

(a) The probability function of the number of overnight emergency calls, X, is given by:

$$\Pr(X = x) = \begin{cases} 0.05 & x = 0 \\ 0.1 & x = 1 \\ 0.15 & x = 2 \\ 0.35 & x = 3 \\ 0.2 & x = 4 \\ 0.15 & x = 5 \end{cases}$$

Therefore, $\Pr(X > 3) = \Pr(X = 4) + \Pr(X = 5) = 0.2 + 0.15 = 0.35$.

(b) The mean is given by:

$$E[X] = 0 \times 0.05 + 1 \times 0.1 + 2 \times 0.15 + 3 \times 0.35 + 4 \times 0.2 + 5 \times 0.15 = 3$$

while the variance is given by:

$$\begin{aligned} V[X] &= (0-3)^2 \times 0.05 + (1-3)^2 \times 0.1 + (2-3)^2 \times 0.15 + \\ &+ (3-3)^2 \times 0.35 + (4-3)^2 \times 0.2 + (5-3)^2 \times 0.15 = \\ &= 9 \times 0.05 + 4 \times 0.1 + 1 \times 0.15 + 0 \times 0.35 + 1 \times 0.2 + 4 \times 0.15 = 1.8 \end{aligned}$$

(c) The mean and variance of the number of overnight emergency calls per week, Y, is the same as the mean of $7X$ and seven times the variance of X. Then,

$$E[Y] = 7 \times E[X] = 7 \times 3 = 21$$

while

$$V[Y] = 7 \times V[X] = 7 \times 1.8 = 12.6$$

(d) The mean and variance of the number of overnight emergency calls per year, M, is the same as the mean of $365X$ and 365 times the variance of X. Then,

$$E[M] = 365 \times E[X] = 365 \times 3 = 1095$$

while

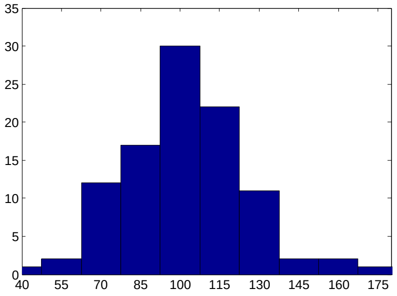
$$V[M] = 365 \times V[X] = 365 \times 1.8 = 657$$

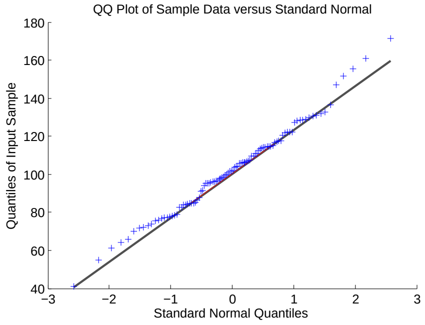
Using the TCL, we get:

$$\Pr(M > 1300) = \Pr\left(\frac{M - 1095}{25.63201} > \frac{1300 - 1095}{25.63201}\right) = \Pr\left(\frac{M - 1095}{25.63201} > 7.997812\right) \simeq$$
$$\simeq \Pr(Z > 7.997812) = 1 - \Pr(Z \leq 7.997812) = 1 - 1 = 0,$$

4. The number of downloads per hour of a specific mobile app can be modelled as a random variable with mean 100 and standard deviation 20.

   (a) (0.5 points) From a sample of size 100 the following graphs are obtained for the number of downloads per hour:





   Reason if the number of downloads per hour can follow a Normal law.

   (b) (0.75 points) Calculate the probability (not approximate) that in a given day (24 hours) the mean of downloads per hour takes values between 90 and 110, assuming that the number of downloads per hour is a simple random sample. Which other hypothesis do you need to assume?

   (c) (0.75 points) Calculate the probability that in a given week (168 hours) the total of downloads per hour is greater than 17000, assuming that the number of downloads per hour is a simple random sample. Which theorem do you apply in order to obtain the result?

   (d) (0.5 points) The number of downloads of this app was under control during 40 hours and a sample mean of 99.5 was obtained. Compute the 95% confidence interval for the mean of downloads per hour (assume $\sigma = 20$ and that the number of downloads per hour is a simple random sample.).

Solución:

(a) Atendiendo al histograma y al qq-plot parece que los datos no se alejan demasiado de la ley Normal.

(b) Si se supone que la v.a. $X =$"número de descargas por hora" sigue una ley $N(100, 20^2)$, entonces dadas $n = 24$ v.a. independientes y con la misma ley que X, la media muestral $\bar{X}_n \sim N(100, 20^2/24)$. Por tanto,

$$P(90 < \bar{X}_n < 110) = P(-2.45 < Z < 2.45) = 2\,P(Z < 2.45) - 1 = 2 \cdot 0.9929 - 1 = 0.9858,$$

donde $Z \sim N(0,1)$. Para poder calcular esta probabilidad es necesario asumir que X tiene ley Normal.

(c) Dadas $X_1, \ldots, X_n$, $n = 168$ v.a. independientes y con la misma ley que $X =$"número de descargas por hora", consideramos la v.a. $\sum_{i=1}^n X_i = n\,\bar{X}_n$. Puesto que $n \geq 30$, el Teorema Central del Límite asegura que $\bar{X}_n \sim N(100, 20^2/168)$. Entonces,

$$P(\sum_{i=1}^{n} X_i > 17000) = P(n\,\bar{X}_n > 17000) = P(\bar{X}_n > 17000/168) \approx P(Z > 0.77) = 0.2206,$$

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

(d) A partir de las descargas observadas en 40 horas ($X_1, \ldots, X_n$ v.a. i.i.d. con $n = 40$) se obtuvo una media muestral de $\bar{x}_n = 99.5$. Suponiendo $\sigma = 20$, el intervalo de confianza al 95% para la media de descargas es

$$\bar{x}_n \mp z_{1-\alpha/2}\,\sigma/\sqrt{n} = 99.5 \mp 1.96 \cdot 20/\sqrt{40} = 99.5 \mp 6.20 = (93.3, 105.7).$$