HomeWork Assignment 1 Applied Statistics 2018 *(includes the use of SPSS)*

Hand in your hand written solutions (a.–f., SPSS-output is not necessary) on **Friday 16/3 at 11.00** (Kanifing)

Part 1: Do not use software (SPSS) for this part: only use a (simple) calculator, with a statistics menu

A new interface for a smart (programmable) heating thermostat was designed by students: the aim was that users could intuitively program the weekly heating schedule without consulting the user`s guide. 34 potential users were asked to program a given schedule for the thermostat. In the table you find the (ordered) task completion times TCT, in minutes.

2.28	2.37	2.39	2.55	2.63	2.69	2.72
2.76	2.82	2.84	2.91	2.98	3.06	3.17
3.17	3.39	3.43	3.44	3.48	3.52	3.55
3.60	3.79	4.06	4.18	4.58	5.07	5.37
5.96	5.97	6.31	6.61	8.96	10.34	

a. Use a simple calculator to compute the sample mean and standard deviation (round in two decimals)

b. Compute the intervals $(\bar{\boldsymbol{x}} - \boldsymbol{k} \cdot \boldsymbol{s}, \bar{\boldsymbol{x}} + \boldsymbol{k} \cdot \boldsymbol{s})$ for $\boldsymbol{k} = \mathbf{1}, \mathbf{2}$ and 3, and compare the number of observed times in the intervals to the expected number according to the empirical rule: show the results in a table.

c. Determine the 5-number-summary of the observations and determine outliers, using the $\mathbf{1.5} \times \boldsymbol{IQR}$-rule.

d. Comparing the sample mean and median, what does the difference tell you about the data's distribution?

e. Suppose we want to estimate $\boldsymbol{E}(\boldsymbol{X}^2)$ for the task completion time $\boldsymbol{X}$.

1. Show that $\boldsymbol{T} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{X}_{\boldsymbol{i}}^{\mathbf{2}}$ is an unbiased estimator of $\boldsymbol{E}(\boldsymbol{X}^2)$.
2. How would you define an alternative estimator of $\boldsymbol{E}(\boldsymbol{X}^2)$, using the sample mean and the sample variance? And determine its value for the observed completion times (the estimate!)

Part 2: applying SPSS or other statistical software: any version of SPSS will do, such as provided version 14 on the flash drive. But you can use other software of even Excel. The directions below are for recent SPSS-version (alternatives for old versions as 14.0 are given as well) .

1. Enter the 38 service times above in an empty SPSS-file and use the "*Variable View*"-tab, left/down in the SPSS-screen, to give the variable Service time a proper *Name* containing your own first name (e.g. "Time"), choose a *Label* likewise (e.g. "service time") and set the number of *Decimals* to 2 (since the observations are given in two decimals).
2. Determine with SPSS "the **classical numerical summary**" (section 1.4) and check your results in a.
   Use the menu's: *Analyze → Descriptive Statistics → Descriptives:* click on *"options"* for the right choice
3. Make SPSS produce a box plot and a histogram of the task completion times:
   - **Box plot** via *Analyze → Descriptive Statistics → Explore.*
   - **Histogram** via *Graphs (→ Legacy Dialogs) → Histogram*: give the histogram a title "Histogram of …" and graph the *Normal Curve* in the histogram (choose this option in the dialog box).
4. Use SPSS to graph both the **exponential** and **normal Q-Q plot**: go to *Analyze→ Descr. Stat.→ Q-Q plots (in SPSS14.0: Graphs→ Q-Q plots*

----

f. If we would consider the option of a **normal** or an **exponential** model for these task completion times, what would you choose on the basis of: 1. The results in b., d. and e.

2. The numerical values of the skewness and the kurtosis (note that SPSS reports $\boldsymbol{kurtosis} - \mathbf{3}$),
3. The histogram and the box plot and
4. The two QQ-plots?

Discuss all aspects separately and draw a total conclusion with respect to the choice of the model.

a	b	c	d	e	f	Total
1	1	2	1	2	3	10

Solutions:

a. $\overline{x} \approx 4.028 \approx 4.03$ and $s \approx 1.8608 \approx 1.86$ (so $s^2 \approx 3.4624 \approx 3.46$)

b. See the table:

	Interval	Observed number	Expected number
$k = 1$	(2.17, 5.89)	28	$0.68 \times 34 \approx 23.1 \approx 23$
$k = 2$	(0.31, 7.75)	32	$0.95 \times 34 \approx 32.3 \approx 32$
$k = 3$	(-1.55, 9.61)	33	$0.997 \times 34 \approx 33.9 \approx 34$

c. The 5-number-summary is:

1. minimum 2.28,

2. 25% of 34 is 8.5, so $Q_1 = x_{(9)} = 2.82$,

3. The median $m = \frac{x_{(17)}+x_{(18)}}{2} = \frac{3.43+3.44}{2} = 3.435$ ($n = 34$ is even)

4. $Q_3 = x_{(26)} = 4.58$ and

5. The maximum is 10.34

Outliers according to the "box-plot-method": $IQR = Q_3 - Q_1 = 4.58 - 2.82 = 1.76$. Outliers are outside the interval $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR) = (0.18, 7.22)$: 8.96 and 10.34.

d. The median m = 3.435 is smaller than the sample mean $\overline{x} = 4.03$, due to a skewness to the right of the distribution with two large outliers (as observed in c.)

e. 1. $T = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ is an unbiased estimator of $E(X^2)$, since

$$E(T) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) = \frac{1}{n} \cdot n \cdot E(X^2) = E(X^2)$$

2. Since $var(X) = E(X^2) - (EX)^2$, we have $E(X^2) = var(X) + (EX)^2$.

Using $\overline{X}$ as an estimator of EX and S^2 for $var(X)$, we find $S^2 + \left(\overline{X}\right)^2$as an estimator of $E(X^2)$.
Then the estimate of $E(X^2)$ is: $s^2 + (\overline{x})^2 = 3.46 + 4.03^2 \approx 19.70$

f. 1. From b., c. and d. we can conclude that the distribution of the observations is not symmetrical, since there are large outliers and $\overline{x} \geq$ median: the data are skewed to the right, probably caused by two large outliers. Only the first interval gives a larger deviation from the proportion of contained observations according to the empirical rule.

2. The numerical values of the skewness coefficient 1.912 (> 0, skewed to the right) is close to the reference value 2 of the exponential distribution, but the kurtosis 6.785 is between the reference values 3 and 9 of the normal respectively the exponential distribution. Considering the skewness we would prefer the exponential distribution.

3. The histogram shows some deviations from the normal distribution: it is not symmetrical (skewed to the right). We would prefer the exponential distribution.

4. Both the normal and the exponential Q-Q plot show a deviating pattern from the line $y = x$. Neither distribution seems to fit.

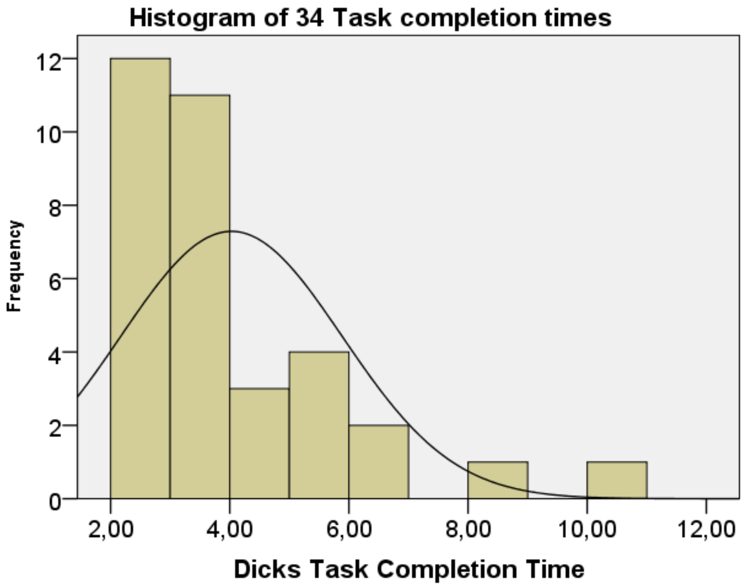
Overall Conclusion: when choosing between the normal and the exponential distribution most indicators (1., 2. and 3.) point to the exponential distribution, but the QQ-plot contradicts the choice of an exponential model for the service times. The **normal nor the exponential distribution seems an appropriate model for the service times**.

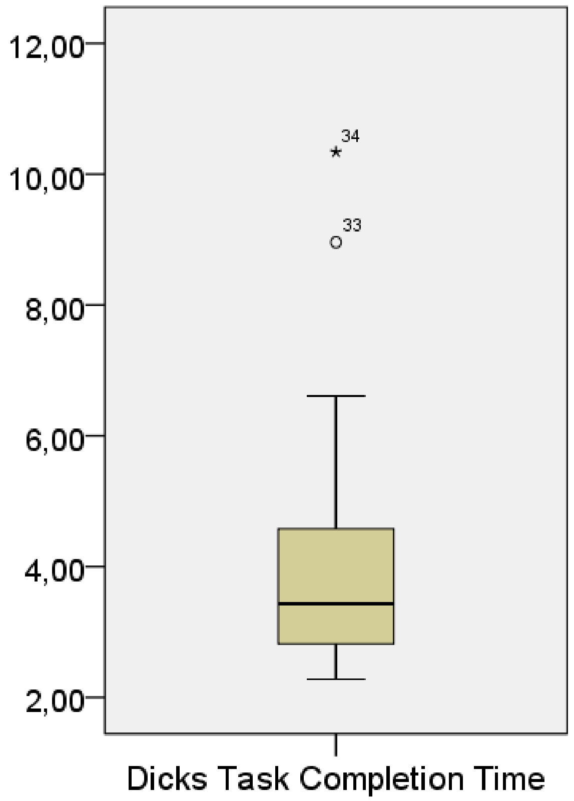
Relevant SPSS-output:

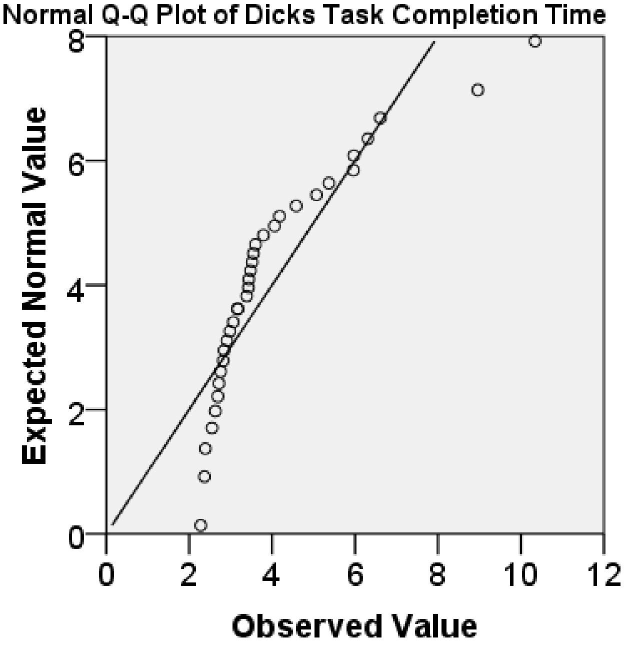
Descriptive Statistics

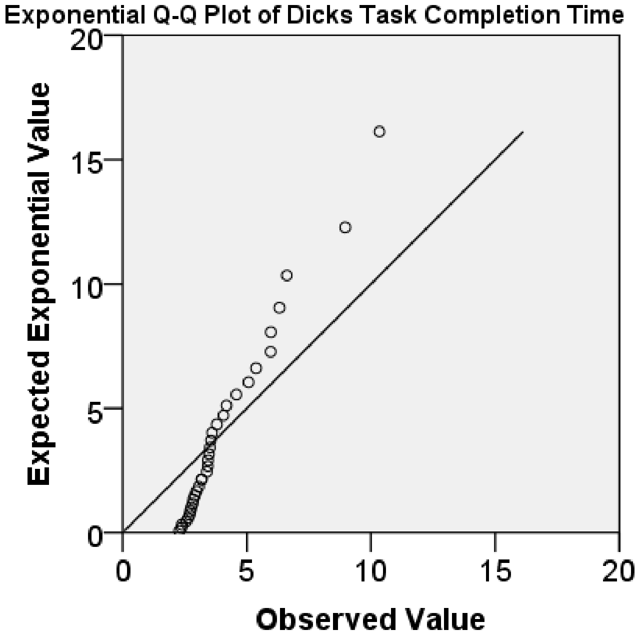
	N	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Dicks Task Completion Time	34	4,0279	1,86076	3,462	1,912	,403	3,785	,788
Valid N (listwise)	34							

(Note: the results for the mean and the standard deviation are the same as we obtained in a., using the calculator)


Histogram of 34 Task completion times




Normal Q-Q Plot of Dicks Task Completion Time


Exponential Q-Q Plot of Dicks Task Completion Time