Test Mathematical Statistics (M5-201800139), November 9, 2018, 8.45-11.45 h.

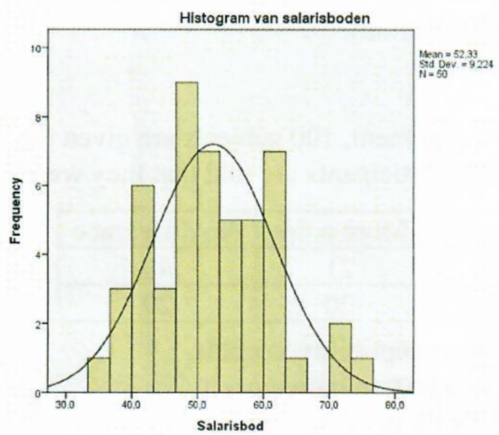
Lecturer Dick Meijer and module coordinator Pranab Mandal

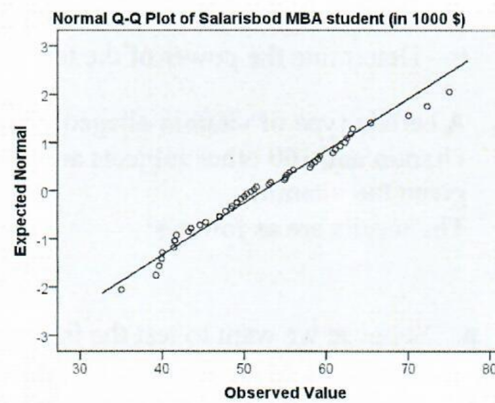
This test consists of 7 exercises. A formula sheet and the probability tables are added.
A regular scientific calculator is allowed, a programmable calculator (“GR”) is not.

1. The table below shows the highest salary offer (in thousands of dollars) that each member of a sample of 50 MBA students was offered who recently graduated from the *Graduate School of Management* of Rutgers, the state university of New Jersey. The offers are readily arranged in ascending order:

35.0	41.4	43.5	47.0	49.1	51.2	55.0	58.2	61.1	63.2
39.2	41.5	44.6	47.7	49.6	51.5	55.2	58.6	61.7	65.4
39.6	41.5	45.3	48.4	50.0	53.0	55.5	59.0	62.3	70.0
39.9	41.7	47.0	48.5	50.3	53.2	56.0	59.2	62.5	72.3
40.0	43.2	47.0	49.1	50.8	54.9	58.0	60.8	63.0	75.0

The mean and the standard deviation of the sample are 52.33 and 9.22, respectively and the skewness coefficient and the kurtosis of the sample are 0.378 and 3.342, respectively.





a. Are there any outliers according to the $1.5 \times IQR$-rule?

b. Does the normal distribution apply to the data? Comment, successively, on the observed numerical measures, the histogram and the normal Q-Q plot. Finally draw a total conclusion.

c. As an extra check we determined the value of Shapiro-Wilk's test statistic: $W = 0.976$. Give 1. the hypotheses, 2. the Rejection Region and 3. the conclusion of this test, for $\alpha = 0.10$.

d. Find a 99%-confidence interval for the expected salary offer of MBA-students and give an appropriate interpretation for the obtained interval.

2. $X_1, \dots, X_n$ is a random sample of X, which is uniformly distributed on $[\theta, 2\theta]$ for some $\theta > 0$.

a. Determine the value of $a \in \mathbb{R}$ such that $T_1 = a \cdot \sum_{i=1}^{n} X_i$ is an unbiased estimator.

b. Show that the estimator T_1 (with the value of a, determined in a.) is a consistent estimator.

c. Determine the value of $b \in \mathbb{R}$ such that T_2 is the best estimator of the shape $b \cdot \sum_{i=1}^{n} X_i$.

d. Determine the maximum likelihood estimator $\hat{\theta}$ of θ.

3. A film critic suspects that the duration of movies of a certain producer is larger than the duration of movies of another producer. “On average, movies of producer (1) are substantially longer than movies of producer (2).” To verify this statement, we looked up the durations of the movies of last year for both of the producers. The results are presented in the table below, including the mean duration and the standard deviation of the observed durations.

Producer 1	86	155	97	91	136	87	114	118	$\overline{x} = \mathbf{110.5}$ and $s_X = \mathbf{25.11}$
Producer 2	102	98	129	81	92				$\overline{y} = \mathbf{100.4}$ and $s_Y = \mathbf{17.84}$

Assume that both data sets can be modelled as random samples of the durations for both producers.

a. Do these observations confirm the suspicion of the critic? In order to answer this question, carry out a (complete) test at a 5% significance level, assuming normality and equal variances.

b. Carry out a non-parametric test as well, as an alternative for the test in a. (so without normality assumption). Give only the following parts of the testing procedure:

1. The hypotheses.
2. The observed value of the suitable test statistic
3. The p-value of the test, using a normal approximation.
4. Your conclusion in words (again for $\alpha = 5\%$).

4. For an unknown parameter $\theta > 0$ the density function of a variable X is defined as:

$$f(x) = \begin{cases} \dfrac{2\theta^2}{(\theta + x)^3}, & \text{if } x \geq 0 \\ 0\ , & \text{elsewhere} \end{cases}$$

Suppose we have only one observation x available.

a. Determine the moment estimator of θ.

b. Determine, with Neyman-Pearson's lemma, the most powerful test on $H_0: \theta = 1$ against $H_1: \theta = 2$ at a 25% level of significance (for $n = 1$ observation).

c. Determine the power of the test in b.

5. A certain type of vitamin allegedly prevents colds. In an experiment, 100 subjects are given the vitamin and 100 other subjects are given a placebo. All 200 participants are told that they were given the vitamin.

The results are as follows:

	Less colds	More colds	No difference
Control group	39	21	40
Treated group	51	20	29

a. Suppose we want to test the influence of the vitamin on susceptibility to colds. Should we apply a test on independence to these data or a test on homogeneity?

b. Apply the test selected in a. using the testing procedure with $\alpha = 5\%$.

c. Determine a 95%-confidence interval for the difference of the proportions "Less colds" in the treated and the control group.

6. Consider the **simple linear regression** model. Prove the equality $SS_{Total} = SS_{regr} + SS_{Error}$ by evaluating $SS_{Error} = \sum_i (y_i - \hat{y}_i)^2$, using $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \overline{y} + \hat{\beta}_1(x_i - \overline{x})$.
(Note that $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ and $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ are the least squares estimates.)

7. Consider the **multiple linear regression model**. Observations Y_i $(i = 1, 2, \ldots, n)$ depend on k predictor variables, as follows: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$,
where the values of the predictor variables are represented by x_{ij} and the disturbances ε_i are independent and all $N(0, \sigma^2)$-distributed. The predicted values are $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$.
In vector notation: $Y = X\beta + \varepsilon$, with least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$.

a. Prove that the vector of predicted values $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$ and the vector of residuals $y - \hat{y}$ are **orthogonal**, where, of course, $y = (y_1, y_2, \ldots, y_k)^T$.

b. Prove that $\sum_i (y_i - \hat{y}_i) = 0$: the sum of residuals always equals zero.

-- *END* --

Grade = $1 + \frac{\textit{number of points}}{60} \times 9$, rounded at 1 decimal.

1				2				3		4			5			6	7		Total
a	b	c	d	a	b	c	d	a	b	a	b	c	a	b	c		a	b	
2	2	3	3	2	2	4	4	6	4	2	4	2	1	6	3	4	3	3	60