Trial Exam "Applied Statistics" 2018

This test consists of 4 exercises. The formula sheet and tables are added separately.

1. We observe independent random variables $X_1, X_2, \ldots, X_n$ drawn from a $N(\mu, 1)$-distribution: they all have the following density function:

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} \quad \text{for some (unknown) parameter } \mu \in \mathbb{R}$$

a. Show that the maximum likelihood estimator of μ is given by $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$.

b. Show that $\bar{X}$ is a consistent estimator of μ.

c. Show that the likelihood ratio test for testing H_0: $\mu = 0$ against H_1: $\mu \neq 0$ rejects the null hypothesis for large positive or negative values of the sample mean.

d. Determine for the test in c. the rejection region and the power at $\mu = 1$ if $n = 16$ and $\alpha = 5\%$.

2. Estimation of the market value of houses is usually done by real estate experts, but recently computer programs have been developed to get a more objective tool for estimation, based on a set of characteristics of the house. Researchers investigated whether the results of estimation by the computer and by the experts are different, on average. For ten randomly chosen apartments the values were estimated by both the computer and the local expert. The results are shown in the table below.

apartment	1	2	3	4	5	6	7	8	9	10
computer	71 000	87 500	92 000	78 000	80 000	86 500	94 500	73 000	96 000	80 000
expert	70 000	86 000	90 000	78 500	81 000	85 000	94 000	74 500	94 000	79 500

a. Can you conclude from these observations that there is a structural difference in estimated values between computer and experts? Conduct a suitable parametric test with α = 5% and clearly state in step 1 of the testing procedure which assumptions are necessary for this test.

b. Determine a 95%-confidence interval for the difference of expected values (by computer and experts).

c. Give a correct interpretation of the interval that you determined in b.

d. Which test can be used as non-parametric alternative of the test in a. (if the normality assumption does not hold)? Give for this test (only!): the test statistic and its observed value and compute the p-value.

3. If we want to test H_0: $\sigma^2 = 10$ against H_1: $\sigma^2 \neq 10$, based on a random sample of $n = 20$ observations, drawn from a normal distribution with unknown parameters, give (only) the proper test statistic and determine the rejection region if $\alpha = 5\%$.

4. Pollution of drinking water is a major health risk. Arsenic is one of the poisonous chemical substances found in drinking water. In Arizona (US) the quantity of arsenic has been assessed as to investigate whether the quantity is larger in rural areas than in urban regions. For both types of areas 10 samples of drinking water were chosen at random and the quantities of arsenic in *ppb* (*parts per billion*) were determined, as the table shows.

Urban area: $\boldsymbol{x_1}$	3	7	25	10	15	6	12	25	15	7	$\bar{x}_1 = 12.5,\ s_1 = 7.63$
Rural area: $\boldsymbol{x_2}$	48	44	40	38	33	21	20	12	1	18	$\bar{x}_2 = 27.5,\ s_2 = 15.3$

a. Can we assume, despite of the observed difference in sample standard deviations, that the variances of the arsenic quantities are the same? Assume (approximately) normal distributions for the quantities and give for the appropriate test with significance level 5%:
   1. The hypotheses
   2. The value of the test statistic.
   3. The Rejection Region
   4. The conclusion that you can draw with respect to the equality of variances.

b. Is the expected arsenic quantity in drinking water in rural areas larger than in urban areas? Conduct an appropriate (parametric) test to answer this question.
Use a significance level of 1% and give all steps of the testing procedure.

c. Which test is a non-parametric alternative for the test in b.?
Give the formula of the test statistic and its (approximate) distribution under H_0.

Solutions Exercise 1

a. $L(\mu) = \prod_{i=1}^{n} f(x_i \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2} = (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}\sum(x_i-\mu)^2}$

$\ln L(\mu) = -\frac{1}{2}n \cdot \ln(2\pi) - \frac{1}{2}\sum(x_i-\mu)^2$, so: $\frac{d}{d\mu}\ln L(\mu) = 0 + \frac{1}{2}[\sum x_i - n\mu] = 0$ if $\mu = \frac{1}{n}\sum x_i = \overline{x}$

For this value of μ L attains its maximum since $\frac{d^2}{d\mu^2}\ln L(\mu) = -\frac{1}{2}n < 0$ (L is concave downward for all values of μ, hence for $\mu = \overline{x}$ too.) $\hat{\mu} = \frac{1}{n}\sum X_i = \overline{X}$ is the maximum likelihood estimator of μ.

b. Since $\overline{X}$ is an unbiased estimator of μ, with variance $\frac{\sigma^2}{n} = \frac{1}{n}$, the mean squared error $MSE(\overline{X}) = var(\overline{X}) = \frac{1}{n}$, so $\lim_{n\to\infty} MSE(\overline{X}) = 0$, proving that $\overline{X}$ is a consistent estimator.

c. $\Lambda(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} f(x_i \mid 0)}{\sup_{\mu\in\mathbb{R}} \prod_{i=1}^{n} f(x_i \mid \mu)} = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\overline{x})^2}} = \prod_{i=1}^{n} e^{-\overline{x}\cdot x_i + \frac{1}{2}\overline{x}^2} = e^{-\overline{x}\cdot\sum x_i + \frac{n}{2}\overline{x}^2}$

$= e^{-\overline{x}\cdot n\overline{x} + \frac{n}{2}\overline{x}^2} = e^{-\frac{n}{2}\cdot\overline{x}^2}$

Λ decreases for increasing $\overline{x}^2$: the likelihood ratio test rejects for small values of Λ, which is equivalent to rejecting for large negative and large positive values of $\overline{X}$:
using symmetry of the distribution under H_0 we will reject if $\overline{X} \le -c$ or if $\overline{X} \ge c$.

d. If $n = 100$ and $\alpha = 5\%$ we find from $P(\overline{X} \ge c \mid H_0) = \frac{\alpha}{2} =: 2.5\%$, $\frac{c-0}{\sqrt{\frac{1}{16}}} = 1.96$ or $c = \frac{1}{4}\cdot 1.96 = 0.49$.

The Rejection region is $(-\infty, -0.49] \cup [0.49, \infty)$

$$\beta(1) = P(\overline{X} \le -0.49 \text{ or } \overline{X} \ge 0.49 \mid \mu = 1) = P\left(Z \le \frac{-0.49-1}{\sqrt{\frac{1}{16}}}\right) + P\left(Z \ge \frac{0.49-1}{\sqrt{\frac{1}{16}}}\right)$$

$$= \Phi(-5.96) + \Phi(+2.04) = 0.0000 + 0.9788 = 97.88\%$$

Exercise 2

a. We should apply a method for paired samples here, since determination of the value of each apartment is done twice: by the computer and by an expert. The difference of the estimates (computer – expert) are: 1000, 1500, 2000, −500, −1000, 1500, 500, −1500, 2000, 500
Summary, using the calculator: $n = 10, \overline{x} = 600$ and $s \approx 1243$.

1. The differences $X_1, \ldots, X_{10}$ of the estimates are independent and all $N(\mu, \sigma^2)$-distributed, with an unknown expected difference μ and variance σ^2
2. We will test H_0: $\mu = 0$ against H_1: $\mu \neq 0$ with $\alpha = 5\%$
3. Test statistic $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 0}{S/\sqrt{10}}$
4. The distribution under H_0: $T \sim t_{10-1}$
5. Observed value: $t = \frac{\overline{X}}{S/\sqrt{10}} = \frac{600}{1243/\sqrt{10}} \approx 1.53$
6. The hypotheses imply a two-sided test: reject H_0, if $T \le -c$ or $T \ge c$.
   From the t_9-table it follows that $c = 2.262$, such that $P(T_9 \ge c) = \frac{\alpha}{2} = 0.025$
7. The observed $t = 1.53$ is not contained in the rejection region, so we cannot reject H_0.

8. At a 5% level of significance the estimates by the computer and the experts are not structurally different.

b. The 95%-confidence interval for the expected difference is $\left(\overline{X} - c\frac{S}{\sqrt{n}}, \overline{X} + c\frac{S}{\sqrt{n}}\right) \approx (-288, 1488)$, where we use the given n = 10, $\overline{x} = 600$ and $s \approx 1243$ (in a.) and c = 2.262 from the t_9-table.

c. At a confidence level 95% the expected difference in values lies between –€ 288 and €1488

d. The alternative for the parred sample t-test is the sign test on the median.

1. Since there are 3 negative and 7 positive differences, we observe for X, the number of positive differences the value $X = 7$. (model: $X \sim B(10, p)$ with p = probability of a positive difference.)
2. We test $H_0: p = \frac{1}{2}$ against $H_1: p \neq \frac{1}{2}$ (or: $H_0: median = 0$ against $H_1: median \neq 0$).
3. The (2-sided) p-value is $2 \cdot P(X \geq 7|H_0) = 2 \cdot [1 - P(X \leq 6|p = 0.5)] = 2(1 - 0.828) = 34.4\%$
(*Hence we cannot reject H_0 for $\alpha = 5\%$ or any other $\alpha < 34.4\%$).*

Exercise 3

Test statistic S^2 and for this test the rejection region is two sided: reject H_0 if $S^2 \leq c_1$ or if $S^2 \geq c_2$, where $\frac{(20-1)S^2}{10} \sim \chi^2_{20-1}$ Hence $P(S^2 \leq c_1|H_0) = P\left(\chi^2_{19} \leq \frac{19}{10}c_1\right)$, so $\frac{19}{10}c_1 = 8.91$ and $c_1 = \frac{10}{19} \cdot 8.91 \approx 4.68$

Similarly $c_2 = \frac{10}{19} \cdot 32.85 \approx 17.3$

Exercise 4

a. We will apply the F-test:

1. Test $H_0: \sigma_1^2 = \sigma_2^2$ (*or:* $\sigma_1 = \sigma_2$) versus $H_1: \sigma_1^2 \neq \sigma_2^2$ with $\alpha = 5\%$
2. $F = \frac{s_X^2}{s_Y^2} = \frac{7.63^2}{15.3^2} \approx 0.249$
3. This is a two-tailed test: reject H_0 if $F \leq c_1$ or $\geq c_2$: $P(F_9^9 \geq c_2) = \frac{\alpha}{2} = 0.025$, so (according to the F_9^9-tabel) $c_2 =$ 4.03 and $P(F_9^9 \leq c_1) = P\left(F_9^9 \geq \frac{1}{c_1}\right) = \frac{\alpha}{2} = 0.025$, so $\frac{1}{c_1} = 4.03$, or $c_1 \approx 0.248$.
4. Since $F = 0.249$ is (just) not in the rejection region, we will not reject H_0, meaning that the variances of the arsenic quantities are not differing significantly at a 5% level of significance.

b. 1. We have two independent, random samples of arsenic quantities $X_1, \ldots, X_{10}$ and $Y_1, \ldots, Y_{10}$, drawn from the $N(\mu_1, \sigma^2)$-distribution for the towns and the $N(\mu_2, \sigma^2)$-distribution for rural areas (implicitly we assume equal σ's!)
(*formally in short*: $X_1, \ldots, X_{10}, Y_1, \ldots, Y_{10}$ are independent and $X_i \sim N(\mu_1, \sigma^2)$ and $Y_j \sim N(\mu_2, \sigma^2)$.)

2. Test H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 < \mu_2$ with α = 1%.
3. Test statistic $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2\left(\frac{1}{10} + \frac{1}{10}\right)}}$ with $S^2 = \frac{9S_1^2 + 9S_2^2}{10+10-2} = \frac{S_1^2 + S_2^2}{2}$.
4. T is under H_0 t-distributed with $df = n_1 + n_2 - 2 = 18$.
5. Observed: $s^2 = \frac{7.63^2 + 15.3^2}{2} \approx 146.15$ ($s \approx 12.09$) . Hence $t = \frac{12.5 - 27.5}{\sqrt{146.15\left(\frac{1}{10} + \frac{1}{10}\right)}} = -2.77$.
6. The test is left-sided: reject H_0 if $T \leq c$, where $c = -2.552$, since $P(T_{18} \geq 2.552) = \alpha = 1\%$.
7. t = -2.77 lies in the rejection region, so reject H_0.
8. The expected arsenic quantity in rural areas is greater than in the towns, at a 5% level of significance.

Alternative with p-value

Steps 6./7. The p-value for the observed $t = -2.77$, $P(T_{18} \leq -2.77) = P(T_{18} \geq 2.77)$, lies between 0.5% en 1%. Hence it is less than α, so reject H_0.

c. The Wilcoxon`s rank sum test with test statistic: $W = \sum_{i=1}^{n_1} R(X_i)$ $(n_1 = 10)$
W is under H_0 approximately normally distributed with

$$\mu = E(W) = \frac{1}{2}n_1(N+1) = \frac{1}{2} \cdot 10 \cdot 21 = 105$$

and $\sigma^2 = var(W) = \frac{1}{12}n_1n_2(N+1) = \frac{1}{12} \cdot 10 \cdot 10 \cdot 21 = 175$