Test Mathematical Statistics (M5-201800139), October 26, 2018, 8.45-11.45 h.

Lecturer Dick Meijer and module coordinator Pranab Mandal

This test consists of 6 exercises. A formula sheet and the probability tables are added.
A regular scientific calculator is allowed, a programmable calculator ("GR") is not.

1. 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 (in 1000 Euro's) are shown in the table below.

apartment	1	2	3	4	5	6	7	8	9	10
computer	71	87.5	92	78	80	86.5	94.5	73	96	80
expert	70	86	90	78.5	81	85	94	74.5	94	79.5

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%. Apply the testing procedure and clearly state in step 1 which assumptions are necessary for this test.

b. Determine a 95%-confidence interval for the difference of expected values (by computer and experts) and give a correct **interpretation** of this interval in words.

c. 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, its observed value, the p-value and your conclusion if $\alpha_0 = 5\%$.

2. 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 $\alpha = 5\%$.

3. 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 quantity of arsenic 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.

4. An event occurs at a rate of p $(0 < p < 1)$: we define the variable X such that $X = 1$ if the event occurs and $X = 0$ otherwise: $P(X = 1) = p$ and $P(X = 0) = 1 - p$.
$x_1, \dots, x_n$ is a realization of a random sample $X_1, \dots, X_n$ of X.

a. Show that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

b. Show that $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a consistent estimator of p.

c. Show that $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is also the maximum likelihood estimator of p.

d. Assume that $n = 100$. Derive the most powerful test on $H_0: p = 0.2$ against $H_1: p = 0.3$ for $\alpha_0 = 5\%$ (no randomization required).

5. Interarrival times of customers in a system are modelled as exponentially distributed random variables with an unknown parameter λ. For the test on H_0: $\lambda = 1$ against H_1: $\lambda < 1$ we want to use the sample mean $\overline{X}$ of a random sample $X_1, \dots, X_n$ (with $n > 25$) as the test statistic.

a. Show that, for the given hypotheses and for a given α_0, the likelihood ratio test is equivalent to a test that rejects H_0 for large values of the sample mean. You can use that $\hat{\lambda} = {}^{1}\!/_{\overline{X}}$ is the *mle*.

b. Determine the sample size n such that - the level of significance $\alpha_0 = 5\%$ and
- the power of the test for $\lambda = \frac{1}{2}$ is at least 99%.

6. If in the simple linear regression model, with observed points $(x_1, y_1), \dots, (x_n, y_n)$, the regression constant β_0 is 0, the model reduces to:
$Y_i = \beta_1 x_i + \varepsilon_i$, where $\varepsilon_1, \dots, \varepsilon_n$ are independent and all $N(0, \sigma^2)$-distributed (unknown β_1 and σ^2).

a. Show that the least squares estimator of β_1 is, in this case, given by $\hat{\beta}_1 = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}$.

b. Determine the distribution of $\hat{\beta}_1$.

c. For the case that $\sigma^2$is known (e.g. $\sigma^2 = 10$), construct a 95%-confidence interval of β_1.

d. Determine the maximum likelihood estimator of the pair (β_1, σ^2).

(the normal density function is given by: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$)

-- *END* --

Grade = 1 + $\frac{\textit{number of points}}{57} \times 9$,
rounded at 1 decimal

1			2	3			4				5		6				Total
a	b	c		a	b	c	a	b	c	d	a	b	a	b	c	d	
6	3	4	2	4	6	2	2	2	3	5	3	4	3	2	2	4	57

Formula Sheet Mathematical Statistics

Probability Theory

$E(X+Y)=E(X)+E(Y)$ $E(X-Y)=E(X)-E(Y)$ $E(aX+b)=aE(X)+b$

$var(X)=E(X^2)-(EX)^2$ $var(aX+b)=a^2var(X)$

If X and Y are independent: $var(X+Y)=var(X)+var(Y)$, $var(X-Y)=var(X)+var(Y)$

$var(T)=E(var(T|V))+var(E(T|V))$

Distribution	**Probability/Density function**	**Range**	$\boldsymbol{E(X)}$	$\boldsymbol{var(X)}$
Binomial (n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$0,1,2,\dots,n$	np	$np(1-p)$
Poisson (μ)	$e^{-\mu}\mu^x/x!$	$0,1,2,\dots$	μ	μ
Uniform on (a,b)	$1/(b-a)$	$a<x<b$	$(a+b)/2$	$(b-a)^2/12$
Exponential (λ)	$\lambda\exp(-\lambda x)$	$x\geq 0$	$1/\lambda$	$1/\lambda^2$
Gamma (α,β)	$x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)/(\Gamma(\alpha)\beta^\alpha)$	$x>0$	$\alpha\times\beta$	$\alpha\times\beta^2$
Chi-square (χ_f^2)	is the Gamma distribution with $\alpha=f/2$ and $\beta=2$			

Testing procedure in 8 steps

1. Give a probability model of the observed values (the statistical assumptions).
2. State the null hypothesis and the alternative hypothesis, using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic if H_0 is true.
5. Compute (give) the observed value of the test statistic.
6. State the test and **a.** Determine the rejection region or
   b. Compute the p-value.
7. State your statistical conclusion: reject or fail to reject H_0 at the given significance level.
8. Draw the conclusion in words.

Bounds for Confidence Intervals:

* $\hat{p}\pm c\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

* $\overline{X}\pm c\dfrac{S}{\sqrt{n}}$ and $\left(\dfrac{(n-1)S^2}{c_2},\dfrac{(n-1)S^2}{c_1}\right)$

* $\overline{X}-\overline{Y}\pm c\sqrt{S^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}$, with $S^2=\dfrac{n_1-1}{n_1+n_2-2}S_X^2+\dfrac{n_2-1}{n_1+n_2-2}S_Y^2$ or: $\overline{X}-\overline{Y}\pm c\sqrt{\dfrac{S_X^2}{n_1}+\dfrac{S_Y^2}{n_2}}$

* $\hat{p}_1-\hat{p}_2\pm c\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

* (regression) $\hat{\beta}_i\pm c\times se(\hat{\beta}_i)$ and

$$\hat{\beta}_0+\hat{\beta}_1x_0\pm cS\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}},\text{ with } S^2=\frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

Prediction intervals: $\overline{X} \pm c\sqrt{S^2\left(1+\frac{1}{n}\right)}$

(regression) $\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm cS\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}$

Test statistics

- X (number of successes for a binomial situation)
- $T = \dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$ and S^2
- $T = \dfrac{\left(\overline{X}-\overline{Y}\right)-\Delta_0}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, with $S^2 = \dfrac{n_1-1}{n_1+n_2-2}S_X^2 + \dfrac{n_2-1}{n_1+n_2-2}S_Y^2$ or: $Z = \dfrac{\left(\overline{X}-\overline{Y}\right)-\Delta_0}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}}$
- $F = \dfrac{S_X^2}{S_Y^2}$
- $Z = \dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, with $\hat{p} = \dfrac{X_1+X_2}{n_1+n_2}$
- (regression) $T = \hat{\beta}_i / se\left(\hat{\beta}_i\right)$ and $F = \dfrac{SS_{Regr}/k}{SS_{Error}/(n-k-1)}$

Adjusted coefficient of determination: $R_{adj}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SS_{Error}}{SS_{Total}}$

Analysis of categorical variables

- 1 row and k columns: $\chi^2 = \displaystyle\sum_{i=1}^{k} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$ $(df = k-1)$
- $r \times c$-cross table: $\chi^2 = \displaystyle\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(N_{ij}-\hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$, with $\hat{E}_0 N_{ij} = \dfrac{\text{row total} \times \text{column total}}{n}$

  and $df = (r-1)(c-1)$.

Non-parametric tests

- Sign test: $X \sim B\left(n, \frac{1}{2}\right)$ under H_0
- Wilcoxon`s Rank sum test: $W = \displaystyle\sum_{i=1}^{n_1} R(X_i)$,

  under H_0 with: $E(W) = \dfrac{1}{2}n_1(N+1)$ and $var(W) = \dfrac{1}{12}n_1 n_2 (N+1)$

Test on the normal distribution

- Shapiro – Wilk`s test statistic: $W = \dfrac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$