TMA4255 Applied Statistics
May 2017
Solutions

1. A manufacturer of fertilizers studies the difference between the effect of three fertilizers A, B and C on the growth of plants. An experiment was conducted where plants were grown under identical conditions and one randomly assigned fertilizer: A to 7 plants, B to 7 plants, C to 7 plants. After a certain time the height of the plants were measured in cm. Data from the experiment (rounded to the nearest integer) are presented in the table below.

A	46	39	53	44	46	50	54	sum is 332	sum of squares is 15914
B	58	40	51	49	53	49	52	sum is 352	sum of squares is 17880
C	61	57	55	57	62	67	59	sum is 418	sum of squares is 25058

a) Assume that all measurements are independent and normally distributed with the same variance. A one-way analysis of variance model (ANOVA) was fitted to the data, and the MINITAB output is given below.

```
One-way ANOVA: A;B;C

Source   DF      SS      MS      F      P
Factor    ?       ?       ?      ?  0,001
Error     ?       ?    24.7
Total     ?  1023,2

S=4,970    R-Sq=56,55%   R-Sq(adj)=51,72%
```

Seven of the entries of this MINITAB printout are replaced by a question mark (?) Find numerical values for these seven missing entries. Show how you calculate these values.

Perform a hypothesis test that there is a difference between the effect of the three fertilizers. Write down the null hypothesis and the alternative. Base the test on what you have in the MINITAB printout above. Use significance level $\alpha = 0.05$.

Solution. If k is the number of samples and n is the total number of observations, then the number of degrees of freedom for "Factor", "Error" and "Total" are $k-1$, $n-k$ and $n-1$, respectivey (this is known from theory). In our case $k = 3$, $n = 21$, therefore the DF column is 2, 18, 20.

$$SSE = MSE \cdot DF = 24.7 \cdot 18 = 444.6 = 444.6.$$

$$SSF = SST - SSE = 1023.2 - 444.6 = 578.6$$

$MS = SS/DF$ therefore

$$MSF = 578.6/2 = 289.3.$$

Finally

$$F = MSF/MSE = 289.3/24.7 = 11.71.$$

The table is

```
One-way ANOVA: A;B;C

Source   DF      SS      MS      F      P
Factor    2   578,6   289,3  11,71  0,001
Error    18   444,6    24,7
Total    20  1023,2

S=4,970    R-Sq=56,55%   R-Sq(adj)=51,72%
```

Denote mean effects of the three feltilazers by μ_A, μ_B and μ_C. Then

$$H_0 : \mu_A = \mu_B = \mu_C, \quad H_1 : \text{not all are equal.}$$

Since

$$F = 11.71 > 3.55 = f_{0.05,2,18},$$

H_0 is rejected. Alternatively we can use the P-value. Since it is less than the significance level, the null hypothesis is rejected.

b) Perform a multiple comparison, using the Tukey test. Again $\alpha = 0.05$. Use that $q(0.05, 3, 18) = 3.61$, where $q(\alpha, m, n)$ is such a value that

$$P(Q_{m,n} \geq q(\alpha, m, n)) = \alpha,$$

where the random variable $Q_{m,n}$ has the studentized range distribution with m and n degrees of freedom.

Solution. Three hypothesis $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A \neq \mu_B$, $H_0 : \mu_A = \mu_C$ vs. $H_1 : \mu_A \neq \mu_C$ and $H_0 : \mu_B = \mu_C$ vs. $H_1 : \mu_B \neq \mu_C$ are tested simultaneously. The null hypothesis is rejected if 0 does not belong to the interval

$$I_{ij} = \left[\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - q(\alpha, k, k(n-1))\sqrt{\frac{MSE}{n}}, \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} + q(\alpha, k, k(n-1))\sqrt{\frac{MSE}{n}} \right],$$

where k is the number of samples and n is the size of each sample. In our case $k = 3$, $n = 7$, $MSE = 24.7$, $q(\alpha, k, k(n-1)) = 3.61$. The intervals are

$$I_{12} = [-9.62, 3.9]$$

$$I_{13} = [-19.04, -5.52]$$

$$I_{23} = [-16.18, -2.66]$$

Thus the first null hypothesis is not rejected while the two others are rejected.

c) Test the hypothesis that effects of fertilizers A and B are equal, using the two-sample t-test. Compare the result with the result obtained in b).

Solution. We test $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A \neq \mu_B$. The t-test is based on the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where n and m are sample sizes, $\bar{X}$ and $\bar{Y}$ are sample means and

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

(S_X^2 and S_Y^2 are sample variances of the two samples). T has (under H_0) the t-distribution with $n+m-2$ degrees of freedom. H_0 is rejected if

$$|T| \geq t_{\alpha/2,n+m-2}.$$

We have

$$\sum X_i = 332,\ \sum Y_i = 352,\ \sum X_i^2 = 15914,\ \sum Y_i^2 = 17880,$$

therefore

$$(n-1)S_X^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - \frac{1}{n}(\sum X_i)^2 = 15914 - 332^2/7 = 168$$

$$(m-1)S_Y^2 = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{1}{m}(\sum Y_i)^2 = 17880 - 352^2/7 = 180$$

$$S_p^2 = 29$$

$$T = \frac{47.43 - 50.29}{\sqrt{29 \cdot 2/7}} = -0.99$$

$$t_{0.025,12} = 2.179$$

H_0 is not rejected. The conclusion coincides with the corresponding conclusion of b).

d) Now we do not assume that the observations are normally distributed. Solve the hypothesis testing problem in a) using a suitable nonparametric test.

Solution. A suitable nonparametric test is the Kruskal-Wallis test. The table of ranks is

A	4.5	1	11.5	3	4.5	8	13
B	17	2	9	6.5	11.5	6.5	10
C	19	15.5	14	15.5	20	21	18

Sums of ranks in the rows are $R_1 = 45.5$, $R_2 = 62.5$, $R_3 = 123$. The test statistic of the Kruskal-Wallis test is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

where k is the number of samples (in our case $k = 3$), n_i are sample sizes (in our case $n_1 = n_2 = n_3 = 7$), n is the total number of observations ($n = 21$). The null hypothesis is rejected if $H \geq \chi^2_{\alpha,k-1}$. In our case $H = 12.09$, $\chi^2_{0.05,2} = 5.991$. H_0 is rejected.

2. 77 goals were scored on the 2012 European Football Championship. The table shows the number of games where 0,1,2 etc. goals were scored.

The number of goals in a game	The number of games
0	2
1	6
2	10
3	6
4	3
5	3
6	1
7+	0

a) Test the hypothesis that the number of goals, scored in a game, has the Poisson distribution with parameter $\lambda = 1$. The significance level is 0.05.

Solution. Poisson probabilities ($\lambda = 1$) are

$$p_k = \frac{e^{-1}}{k!},\ k = 0, 1, ...,$$

i.e.

$p_0 = 0.37$
$p_1 = 0.37$
$p_2 = 0.18$
$p_3 = 0.06$
$p_4 = 0.015$
$p_{5+} = 0.005$

To apply the Pearson goodness-of-fit test, we use the partition (of nonnegative integers)

$$\{0\},\ \{1\},\ \{2, 3, ...\}$$

with probabilities (under H_0)

$$p_{10} = 0.37,\ p_{20} = 0.37,\ p_{30} = 0.26.$$

This partition is chosen because the condition $np_{i0} \geq 5$ must be satisfied. Then

$$k_1 = 2,\ k_2 = 6,\ k_3 = 23;$$

$$np_{10} = 11.47,\ np_{20} = 11.47,\ np_{30} = 8.06.$$

The observed value of the test statistic is

$$d = \sum_{i=1}^{3} \frac{(k_i - np_{i0})^2}{np_{i0}} = 38.1 > 5.991 = \chi^2_{0.05,2}$$

The hypothesis is rejected.

b) Now we test the hypothesis that the number of goals, scored in a game, has a Poisson distribution, where the parameter is not specified. This is done using MINITAB. The MINITAB output is given below (except P-Value). Explain how the numbers in the column "Probability" have been obtained. Why the value of

"DF" is 4? How is the value of "Chi-Sq" obtained from columns "Observed" and "Expected"? What is the conclusion now? Is the missing P-Value in the MINITAB output greater or less than 0.05?

```
Poisson mean for C1 = 2,48387
```

C1	Observed	Poisson Probability	Expected	Contribution to Chi-Sq
0	2	0,083420	2,58601	0,132795
1	6	0,207204	6,42332	0,027898
2	10	0,257334	7,97734	0,512845
3	6	0,213061	6,60490	0,055398
4	3	0,132304	4,10143	0,295786
>=5	4	0,106678	3,30700	0,145220

N	N*	DF	Chi-Sq	P-Value
31	0	4	1,16994	

```
3 cell(s)  (50,00%)  with expected value(s) less than 5.
```

Solution. Numbers in the column "Probability" are probabilities of the Poisson distribution with parameter $\hat{\lambda} = 2.48387$ (estimate of λ), i.e.

$$0.083420 = \frac{e^{-2.48387}2.48387^0}{0!},$$

$$0.207204 = \frac{e^{-2.48387}2.48387^1}{1!},$$

etc.

The number of degrees of freedom is equal to $k - 1 - r$, where k is the number of sets in partition (in our case $k = 6$) and r is the number of estimated parameters ($r = 1$), so $k - 1 - r = 4$.

$$\text{Chi} - \text{Sq} = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

Conclusion:the null hypothesis is not rejected, because

$$1.16994 < 9.448 = \chi^2_{0.05,4}.$$

Since the null hypothesis is not rejected, the P-Value is greater than 0.05 (in fact it is 0.883).

3. The amounts (in grams) of a chemical compound Y that dissolved in 100 grams of water at various temperatures x, were recorded as follows:

For these data

$$\sum_{i=1}^{15} x_i = 660,\ \sum_{i=1}^{15} Y_i = 540,\ \sum_{i=1}^{15} x_i^2 = 33520,\ \sum_{i=1}^{15} x_i Y_i = 27048.$$

$x(^{o}\text{C})$	16	20	24	28	32	36	40	44	48	52	56	60	64	68	72
Y	13	18	24	26	27	30	31	35	42	43	44	47	49	53	58

We would like to fit a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, ..., 15,$$

where the ϵ_i are independent and normally distributed random variable with zero expectation and (unknown) variance σ^2.

a) Find the equation of the regression line. Estimate the amount of chemical that will dissolve in 100 grams of water at 70°C.

Solution. To find the equation we need estimates of β_0 and β_1. We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{27048 - 44 \cdot 540}{33520 - 44 \cdot 660} = 0.73393,$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{x} = 3.707.$$

Therefore the equation is

$$y = 3.71 + 0.734x.$$

The prediction value at $x = 70$ is

$$\hat{Y}(70) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 70 = 55.082.$$

b) A part of MINITAB output for these data is given below.

```
S = 1,72856   R-Sq = 98,4%    R-Sq(adj) = 98,3%

Analysis of variance

Source           DF       SS       MS        F
Regression        1   2413,2   2413,2   807,64
Residual Error   13     38,8      3,0
Total            14   2452,0
```

Using this output find the estimated standard deviation of the estimator of the slope. Find a 95% confidence interval for the slope.

How is the R-Sq value obtained from the numbers given in the Analysis of variance table?

The observed value 807.64 of some test statistic F is given in the Analysis of variance table. Which hypothesis is tested? What is the conclusion if the significance level is $\alpha = 0.01$?

Solution. It follows from the MINITAB output that the estimate of σ is 1.72856. Since

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{15} (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_{i=1}^{15} x_i^2 - n\bar{x}^2},$$

the estimated standard deviation of $\hat{\beta}_1$ is

$$\sqrt{\frac{S^2}{\sum_{i=1}^{15} x_i^2 - n\bar{x}^2}} = \sqrt{\frac{1.72856^2}{33520 - 660 \cdot 44}} = 0.026.$$

The interval is

$$\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{S^2}{\sum_{i=1}^{15} x_i^2 - n\bar{x}^2}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{S^2}{\sum_{i=1}^{15} x_i^2 - n\bar{x}^2}}\right] =$$

$$= [0.734 - 2.16 \cdot 0.026, 0.734 + 2.16 \cdot 0.026] = [0.678, 0.79].$$

$$R^2 = \frac{SSR}{SST} = \frac{2413.2}{2452.0} = 0.984.$$

The hypothesis $H_0 : \beta_1 = 0$ is tested versus the alternative $H_1 : \beta_1 \neq 0$. Since $f_{0.01,1,13} = 9.07 < 807.64$, the null hypothesis is rejected.