Tentative solutions
TMA4255 Applied Statistics
August 2016

Problem 1 Treatments for stress reduction

a) Test the hypothesis

$$H_0 : \sigma_A = \sigma_B \text{ vs. } H_1 : \sigma_A \neq \sigma_B$$

by calculating a 95% confidence interval for $\frac{\sigma_A}{\sigma_B}$.

$$P(F_{1-\alpha/2,9,9} < F < F_{\alpha/2,9,9}) = 0.95$$
$$P(F_{0.975,9,9} < \frac{\sigma_A}{\sigma_B} \cdot \frac{s_B}{s_A} < F_{0.025,9,9}) = 0.95$$
$$P(\sqrt{F_{0.975,9,9}} \cdot \frac{s_A}{s_B} < \frac{\sigma_A}{\sigma_B} < \sqrt{F_{0.025,9,9}} \cdot \frac{s_A}{s_B}) = 0.95$$

A 95% confidence interval for $\frac{\sigma_A}{\sigma_B}$:

$$\frac{\sigma_A}{\sigma_B} \in [\sqrt{0.2484} \cdot \frac{0.972}{1.333}, \sqrt{4.0260} \cdot \frac{0.972}{1.333}] = [0.3434, 1.463]$$

We see that 1 lies within the confidence interval and we keep $H_0 : \sigma_A = \sigma_B$.

Based on the data, can we conclude that the stress reduction scores are different for the two treatment groups (group A and group B)?

We test the hypotheses:

$$H_0 : \mu_A = \mu_B \text{ vs. } H_1 : \mu_A \neq \mu_B$$

Perform a two-sample t-test with equal sample size and variance (from the hypothesis test above). With equal variance between the groups we can use the pooled estimate of the standard deviation:

$$S_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$
$$= \frac{(10-1) \cdot 0.972^2 + (10-1) \cdot 1.333^2}{10 + 10 - 2} = 1.36$$

$$T = \frac{\bar{X}_A - \bar{X}_B}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

Under H_0 $T \sim t_{n_A+n_B-2}$. It is given that

$$t_{obs} = \frac{3.5 - 3}{\sqrt{1.36}\sqrt{\frac{1}{10} + \frac{1}{10}}} = 0.958$$

This is a two-sided test, and using significance level α= 0.05 we reject the null hypothesis when $|t_{obs}| > t_{\alpha/2,n_A+n_B-2}$. From the tables we find that the critical value are $t_{0.025,18} = 2.101$.

Conclusion: We cannot reject the null hypothesis and we have reason to believe that the stress reduction scores are the same for the two treatment groups A and B.

Assumptions for the two-sample t-test:
Each sample comes from a normally distributed population and independent samples. From the hypothesis in the first part of this problem, we found that $\sigma_A = \sigma_B$, so assume equal variance in the two samples.

b) We test the hypotheses:

$$H_0 : \tilde{\mu}_A = \tilde{\mu}_B \text{ vs. } H_1 : \tilde{\mu}_A \neq \tilde{\mu}_B.$$

In the Wilcoxon rank-sum test we

- order the scores in increasing order without regard of which group they came from
- give them a rank
- sum the rank of the smallest group 1 (here choose group A)

Group	Order	Rank
B	1	1.5
B	1	1.5
B	2	4
A	2	4
A	2	4
A	3	8
A	3	8
B	3	8
B	3	8
B	3	8
A	4	14.5
A	4	14.5
A	4	14.5
A	4	14.5
A	4	14.5
B	4	14.5
B	4	14.5
B	4	14.5
A	5	19.5
B	5	19.5

Sum the ranks of group A, W_A =116.

From W_A we can calculate $W_B = \frac{(n_A+n_B)(n_A+n_B+1)}{2} - W_A = \frac{(20)(21)}{2} - 116 = 94$. The test statistics is then standardized,

$$U_A = 116 - \frac{10(10+1)}{2} = 61$$
$$U_B = 94 - \frac{10(10+1)}{2} = 39$$

This is a two-sided test and we take the $min(U_A, U_B)$ (here $= U_B = 39$). From the tables we can find a critical value (here 23). We reject H_0 if $U \leq$ critical value. We can also base the test on the p value, it is found in Table 1 to be 0.4272.

Conclusion: we can not reject H_0. This is the same conclusion as in a).

Assumptions:

- two continuous distributions, non-normal
- same shape and spread
- independent samples

We see from Figure 2 that the data are not normal (AD test gives p-value of 0.015), but very close to being normal. Wilcoxon rank-sum test will give best power of the test if very non-normal data. The t-test could give the best power, as it reasonably robust to departure from normality.

c) We test the hypotheses:

$$H_0 : \mu_A = \mu_B = \mu_C \text{ vs. } H_1 : \text{ at least two means are not equal.}$$

To perform a one-way ANOVA we need to find the treatment sums of squares (SSA) and the error sums of squares (SSE). Let $\bar{x}_A$ denote the average and s_A^2 the standard deviation of method A. Ditto for methods B and C. Let $\bar{x}$ denote the grand mean.

$$SSA = n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2 + n_C(\bar{x}_C - \bar{x})^2 =$$
$$10(3.5 - 2.83)^2 + 10(3 - 2.83)^2 + 10(2 - 2.82)^2 = 11.667$$

$$SSE = (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2 =$$
$$9 \cdot 0.972^2 + 9 \cdot 1.333^2 + 9 \cdot 0.816^2 = 30.48776.$$

We have that $n_A = n_B = n_C = 10$, k=3. The F statistic is found to be

$$F = \frac{\frac{11.667}{k-1}}{\frac{30.48776}{k(n-1)}} = \frac{\frac{11.667}{2}}{\frac{30.48776}{3(10-1)}} = \frac{5.8335}{1.129} = 5.166962,$$

and should be compared with the critical value $F_{0.05,2,27} = 3.3541$, and we thus reject the null hypothesis.

Assumptions:
The one-way ANOVA model is:

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j},$$

where the error terms are independent and normally distributed with the same variance across treatment groups.

Conclusion: at least two of the group means differ. To find out which of the groups differ, we need multiple testing.

d) Bonferroni:
$\mu_i - \mu_j$ is significant different from 0 if

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq t_{\alpha/2 \cdot m, N-k} \cdot S_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

m=3 (number of tests), N=30, k=3, $S_p = \sqrt{SSE/N - k}$. The critical value $t_{\alpha/2 \cdot m, N-k} = t_{0.025 \cdot 3,27} = t_{0.0083,27}$, we can not find this exact value in the tables, we find $t_{0.005,27} = 3.690$.

For group A and B, test if

$$H_0 : \mu_A - \mu_B = 0 \text{ vs. } H_1 : \mu_A - \mu_B \neq 0$$

$$|\bar{y}_A - \bar{y}_B| \geq 3.69 \cdot \sqrt{1.129}\sqrt{\frac{1}{10} + \frac{1}{10}}.$$

$$|3.5 - 3| \geq 1.75$$

Conclusion: not significant different from 0, and group A and group B have equal means.

For group A and C, test if

$$H_0 : \mu_A - \mu_C = 0 \text{ vs. } H_1 : \mu_A - \mu_C \neq 0$$

$$|\bar{y}_A - \bar{y}_C| \geq 3.69 \cdot \sqrt{1.129}\sqrt{\frac{1}{10} + \frac{1}{10}}.$$

$$|3.5 - 2.0| \geq 1.75$$

Conclusion: not significant different from 0, and group A and group C have equal means.

For group B and C, test if

$$H_0 : \mu_B - \mu_C = 0 \text{ vs. } H_1 : \mu_B - \mu_C \neq 0$$

$$|\bar{y}_B - \bar{y}_C| \geq 3.69 \cdot \sqrt{1.129}\sqrt{\frac{1}{10} + \frac{1}{10}}.$$

$$|3.0 - 2.0| \geq 1.75$$

Conclusion: not significant different from 0, and group B and group C have equal means.

Problem 2 Heat flux

a) Estimated regression equation

$$\hat{y} = 325.44 + 0.06753x_1 + 2.552x_2 + 3.800x_3 - 22.949x_4 + 2.417x_5.$$

Is x_5 significant?
(Estimated regression coefficient for x_5 is 2.417).

$$H_0 : \beta_5 = 0 \text{ vs. } H_1 : \beta_5 \neq 0$$

using a t-test. The t statistics is 1.24 with n-k-1 degrees of freedom, df=23. Critical value $t_{0.025,23} = 2.069$.

Conclusion: we accept H_0, and x_5, the time of day, is not significant (p-value 0.194).

Calculate the R^2 and explain how you can interpret this value.

R^2 is defined and calculated as

$$\frac{SSR}{SST} = \frac{MSR \cdot k}{SST} = \frac{2639.1 \cdot 5}{14681.9} = 89.9\%.$$

R^2 can be interpreted as the proportion of variability in the data that is explained by the regression model, Model A.

An appropriate estimate for σ is the estimated standard deviation in the regression model, $s = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE} = \sqrt{64.6} = 8.037413$.

Is a significant amount of variation explained by the model?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \text{ vs. } H_1 : \text{ at least one not equal to 0}$$

To test this we use the F statistic

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = 40.81,$$

we see from Figure 5 that a p-value approximately 0 is given for this F-test.

Conclusion: Model A explain a significant amount of variation.

b) There are four principal assumptions made in the linear regression model in Eqn. 1.

1. Linearity of the relationship between response and covariates.
2. Independence of the errors, ϵ_i (no serial correlation).
3. Homoscedasticity (constant variance) of the errors, ϵ_i. This means constant variance versus time and versus fitted value (or covariates).
4. Normality of the error,ϵ_i, distribution.

That is $\epsilon_i \sim N(0, \sigma^2)$ iid. We can not observe this random error term, and we therefore use the residuals (standardized) to test the assumptions from the fitted regression.

A residual is defined as $e_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is the estimated fitted values.

1. Detecting nonlinearity is usually most evident in a plot of the residuals versus fitted values. If linear, the points should be symmetrically distributed around a horizontal line. A curve-like pattern may indicate that the model makes systematic errors making unusually large or small predictions. Also residuals versus covariates can indicate nonlinearity. In Figure 6 upper right panel we see that there are no clear pattern, but not entirely random with some large values for small values of y. We have to look further at plots of residuals vs covariate to determine further the linearity of the residuals (not included in the problem). We also see from Figure 4 that we have a high correlation between y and x_1, x_4, x_5 (and also between some of the covariates).
2. By looking at the standardized residuals versus observation order we can detect correlation in the residuals. If there are no correlation, the residuals should be scattered randomly around 0, and there should be no trend. In Figure 6 lower right panel we see that there are no clear trend, although not entirely random.
3. Non constant variance can be detected by looking at for instance standardized residuals versus fitted values. Non constant variance is indicated by a trend, most often a "fan" shape. Also residuals versus covariates can indicate non constant variance. In Figure 6 upper right panel we see that there are no clear pattern, however maybe an indication of a "fan" shape, where the residual are larger for larger values of the fitted values.
4. A normal probability plot of the residuals (QQ-plot) is the best test for normally distributed errors. If the error distribution are normal the residual points should fall close to the diagonal line. An S-shaped pattern of deviations from the diagonal line indicates too many/too few large errors and a bow-shaped pattern indicates that the residuals are not symmetrically distributed (see also histogram of residuals). Violations of normality can be due to two reasons that the distributions of the response and/or covariates are significantly non-normal, and/or the linearity assumption is violated. A Anderson-Darling normality test can also be used to test for normality. In Figure 6 upper left panel we see that all of the points lies inside the 95% confidence interval (one point on the line), and we have a normal distribution. Although the points have a indication of a S-shaped curve. A p-value of > 0.05 of the Anderson Darling test, indicates that the residuals are normally distributed. Also the histogram lower left panel indicates a somewhat normal distribution of the residuals.

c)

- R^2: defined as

$$SSR/SST = 1 - SSE/SST,$$

and indicates the proportion of variation explained by the regression model. This can only increase as more variables are added to the model. R^2 should not be used comparing models with different number of covariates (however, if adding more variables to the model yields a very small increase in R^2 this indicates that this is not worthwhile). R^2 can be used to compare models with the same number of parameters.

- $R^2_{adjusted}$: defined as

$$1 - SSE/(n-k-1)/SST(n-1),$$

and makes a penalty for adding more predictors to the model. The best regression model is the one with the largest adjusted R^2-value.

- R^2_{pred}: defined as
$$1 - \frac{PRESS}{SST} = 1 - \frac{\sum y_i - \hat{y}_{i,-i}}{SST}$$
, reflecting prediction performance. The best regression model is the one with the largest predicted R^2-value.
- S: is the square root of
$$MSE = SSE/DF$$
and quantifies how far away our predicted responses are from our observed responses. We want this distance to be small and the best regression model is the one with the smallest MSE. As S is the square root of MSE, the best model is also the one with the smallest S.
- Mallows C_p (from textbook):
$$C_p = p + (s^2 - \hat{\sigma^2})(n-p)/\hat{\sigma^2}$$
where p=number of parameters estimated, n=number of observations, s^2= estimated variance (MSE) of model under investigation, $\hat{\sigma^2}$=estimated variance of the most complete model (Model A). We are in general looking for a small value for C_p. A rule of thumb is that we would like a model where $C_p \approx p$ A too high C_p may indicate a model that is underfitted (not explaining variability), and a too low C_p may indicate a model that is overfitting the data. By default $C_p = p$ for the model we use as the most complete model (Model A).

Compare models: The model with the largest adjusted R^2-value (97.6) and the smallest S (2.3087) is the model with the three variables x_1, x_2, and x_4, (whereas based on the R^2 criterion, the "best" model is with x_1 and x_2 R^2 = 97.9).

The $Vars$ column tells us the number of predictors (p-1) that are in the model (because intercept is in the model). But, we need to compare C_p to the number of parameters (p). We should add one to the numbers in $Vars$ to compare to C_p.

- Full model: has the highest adjusted R^2 (87.7%), a low Mallows' C_p value (6.0), and the lowest S value (8.0390) (should not use C_p to evaluate model for the full model).
- Best model with p=4 variables: the model containing x_1, x_2,x_3 and x_4 contains 5 parameters, has a lower C_p value (5.8), although S is slightly higher (8.16) and adjusted R^2 is slightly lower (87.3%).
- Best model with p=3 variables: the model containing x_2,x_3 and x_4 contains 4 parameters, has a slightly higher C_p value (7.6) and a lower adjusted R^2 (85.9%).
- Best model with p=2 variables: the model containing x_3 and x_4 contains 3 parameters, C_p=9.1, lower $R^2_{adjusted}$ (84.8%).
- One predictor model have very high C_p and low $R^2_{adjusted}$.

In this example, it isn't obvious which model fits the data best. The best two predictor model includes x_3 and x_4 and is tied for having the highest predicted R^2 (81.4%). This fact suggests that the models that include additional predictors may be overfitting the data. Overfit models appear to explain the relationship between the predictor and response variables for the data set used for model calculation but fail to provide valid predictions for new observations. If you are mainly interested in predictions for new observations, this two predictor model may be the best model and you will only need to measure data for two predictors.

Problem 3 Time to tie show laces

a) We use the χ^2 goodness of fit test, bases on calculated frequencies using the distribution under the null hypothesis Hypothesis:

H_0 : the total time to tie shoe laces is normally distributed with mean 35 and standard deviation 10

$$\text{vs. } H_1 : \text{ not so}$$

We need to calculate the expected frequency for each time interval, which again is based on calculating the probability for each time interval under the null hypothesis. Define:

- p_i expected probability for class i
- o_i observed count in class i
- e_i expected count in class i

Let $X \sim N(35, 10)$.

$$\begin{aligned}
P(X \leq 10) &= P(Z \leq \frac{10-35}{10}) = \Phi(-2.5) = 0.0062 \\
P(X \leq 20) &= P(Z \leq \frac{20-35}{10}) = \Phi(-1.5) = 0.0668 \\
\mathbf{P(X \leq 30)} &= \mathbf{P(Z \leq \frac{30-35}{10}) = \Phi(-0.5) = 0.3085} \\
\mathbf{P(X \leq 40)} &= \mathbf{P(Z \leq \frac{40-35}{10}) = \Phi(0.5) = 0.6915} \\
P(X \leq 50) &= P(Z \leq \frac{50-35}{10}) = \Phi(1.5) = 0.9332 \\
P(X \leq 60) &= P(Z \leq \frac{60-35}{10}) = \Phi(2.5) = 0.9938
\end{aligned}$$

$$
\begin{aligned}
&p_1 = P(X \leq 10) = 0.0062\\
&p_2 = P(X \leq 20) - P(X \leq 10) = 0.0606\\
&p_3 = P(X \leq 30) - P(X \leq 20) = 0.2417\\
&\mathbf{p_4 = P(X \leq 40) - P(X \leq 30) = 0.3829}\\
&p_5 = P(X \leq 50) - P(X \leq 40) = 0.2417\\
&p_6 = P(X \leq 60) - P(X \leq 50) = 0.0606\\
&p_7 = P(X \geq\geq 60) = 0.0062
\end{aligned}
$$

The probability, that the total time to tie shoe laces for a random child, lies in the interval $(30, 40]$ is found in bold.

The expected value for each time interval is found as $e_i = n \cdot p_i$ with $n = 250$. Table of probabilities, expected and observed frequencies.

	1	2	3	4	5	6	7
Time	(∞,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,∞]
Probability under null	0.0062	0.0606	0.2417	0.3829	0.2417	0.0606	0.0062
Expected	1.55	15.15	60.43	95.73	60.43	15.15	1.55
Observed	9	16	51	72	79	11	12

The expected is lower than 5 for cell 1 and 7. Need to merge four of the cells to be make the approximation valid a χ^2 distribution.

	1	2	3	4	5
Time	(∞,20]	(20,30]	(30,40]	(40,50]	(50∞]
Probability under null	0.0668	0.2417	0.3829	0.2417	0.0668
Expected	16.7	60.43	95.73	60.43	16.7
Observed	25	51	72	79	23

The test statistics is

$$x^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

approximately χ^2 distributed with k-1=5-1= 4 degrees of freedom.

$$X^2 = \sum_{i=1}^{5} \frac{(25-16.7)^2}{16.7} + \frac{(51-60.43)^2}{60.43} + \frac{(72-95.73)^2}{95.73} + \frac{(79-60.43)^2}{60.43} + \frac{(23-16.7)^2}{16.7} = 4.12515 + 1.471536 + 5.88230$$

.

The null hypothesis is rejected if the test statistics is larger than $\chi^2_{0.05,4} = 9.488$.

Conclusion: clearly we can reject the null hypothesis and we have reason to believe that the total time to tie shoe laces is not normally distributed with mean 35 and standard deviation 10.

How could you instead test the null hypothesis that the total time to tie shoe laces is normally distributed, $N(\mu, \sigma)$, when μ and σ are unspecified?

We could instead test the hypothesis:

$$H_0 : \text{the total time to tie show laces is normally distributed} H_1 : \text{not so.}$$

Using the estimated mean and standard deviation of the data as estimate of μ and σ and perform the same procedure as above.