Norwegian University of Science and Technology
Department of Mathematical Sciences

Tentative solutions TMA4255 Applied Statistics 16 May, 2015

Problem 1 Manufacturer of fertilizers

a) Are these independent samples?
Yes, since the two fertilizer samples are not related (one fertilizers were given to n_1 plants and the other fertilizer to different n_2 plants, grown under identical conditions to be able to compare the two fertilizers).

Do we have normality?
This is a small sample and we need to check the normality assumption from both populations. However, normality is assumed in the problem.

Do the populations have equal variance? Yes, since s_{x1} and s_{x2} are not that different. A test could be performed for testing equal variance. However, equal variance is assumed in the problem.

With these assumptions we can use a two-sample t-test with equal variance (two-sample t-test with pooled variance).

We want to test if is a difference in the effect of new and old fertilizers on plant height? The hypothesis can be written as

H_0 : The mean growth heights of the plants given the two types (X_1 and X_2) of fertilizers are the same

vs.

H_1 : The mean growth heights of the plants given the two types (X_1 and X_2) of fertilizers are different

or

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2$$

or written as

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_1 : \mu_1 - \mu_2 \neq 0$$

We test this hypothesis by performing a two sample t-test with equal variance based on the test statistic

$$T = \frac{(\bar{x_1} - \bar{x_2}) - d_0}{s_{pooled}/\sqrt{1/n_1 + 1/n_2}},$$

where $d_0 = 0$ and $s_{pooled} = \sqrt{\frac{s_{x1}^2(n_1-1)+s_{x2}^2(n_2-1)}{n_1+n_2-2}} = \sqrt{\frac{(1.71^2)(5)+(1.91^2)(6)}{6+7-2}} = \sqrt{3.32}$.

Under H_0 T $\sim t_{n_1+n_2-2}$.

$$t_{obs} = \frac{54.25 - 53.07}{\sqrt{3.32}\sqrt{1/6 + 1/7}} = 1.164.$$

This is a two-sided test, and using significance level $\alpha = 0.05$ we reject the null hypothesis when $|t_{obs}| > t_{\alpha/2,n_1+n_2-2}$. From the tables we find that the critical values are $t_{0.025,11} = 2.201$.

Conclusion: We can not reject the null hypothesis and we have reason to believe that the mean heights of plants given the two types (X_1 and X_2) are not significantly different. If we look up in the table in "Tabeller og formler" we find that the p-value is larger than 0.15.

b) When we assume that X_{1i} and X_{2i} are not normally distributed we can perform a non-parametric test. When the normality assumption does not hold, a non-parametric alternative to the t-test can often have better statistical power. For example, for two independent samples when the data distributions are asymmetric (that is, the distributions are skewed) or the distributions have large tails, then the Wilcoxon rank-sum test (also known as the Mann–Whitney U test) can have three to four times higher power than the t-test.

If we cannot assume any underlying parametrical distributional, but continuous distribution and equal shape of the distribution for the two populations, the we can perform a Wilcoxon signed-rank test for the data. The Wilcoxon rank-sum test is used for two independent samples. The Wilcoxon rank-sum test is robust towards outliers since only ranks are considered.

We test the hypothesis that the median of the two samples are different.

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 \text{ vs. } H_1 : \tilde{\mu}_1 \neq \tilde{\mu}_2$$

or written as

$$H_0 : \tilde{\mu}_1 - \tilde{\mu}_2 = 0 \text{ vs. } H_1 : \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$$

The observations are arranged in ascending order: and ranked from 1 to 13, ranks marked * belong to x_1, the smallest sample:

Ordered data	Ranks
51.0	1.5
51.0	1.5
52.1	3.5
52.1	3.5*
52.9	5*
53.0	6
53.3	7
54.0	8.5*
54.0	8.5*
55.5	10
55.6	11
56.1	12*
56.4	13*

Since this is a two-sided test the test observator is the $\min(U_1, U_2)$, so we need to calculate both U_1 and U_2. We calculate the sum of ranks for the smaller sample, x_1 (as we have unequal sample size), for fertilizer X_1 are $w_1 = 3.5 + 5 + 8.5 + 8.5 + 12 + 13 = 50.5$. We then need to calculate $u_1 = w_1 - \frac{n_1(n_1+1)}{2} = 50.5 - \frac{(6)(7)}{2} = 29.5$.

$w_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2} - w_1 = 91 - 50.5 = 40.5$. Then $u_2 = w_2 - \frac{n_2(n_2+1)}{2} = 40.5 - \frac{(7)(8)}{2} = 12.5$.
The $\min(u_1, u_2) = u_2 = 12.5$.

We can find the p-value for the test in "Wilcoxons to-utvalgstest" in "Tabeller og formler". We need to find $P(U_2 \leq 12) = 0.117$ and $P(U_2 \leq 13) = 0.147$, then $P(U_2 \leq 12.5) = 0.132$. This is a two-sided test so we get a p-value of $2 \cdot P(U_2 \leq 12.5) = 2 \cdot 0.132 = 0.264$.

Conclusion: We can not reject the null hypothesis. We have reason to believe that the mean heights of plants given the two types (X_1 and X_2) are not significantly different.

If we assume normality U_1 and U_2 will have test statistics,
$\mu = \frac{n_1 n_2}{2} = 6 \cdot 7/2 = 21$ and $\sigma = \frac{n_1 n_2 (n_1+n_2+1)}{12} = 588/12 = 49$. Two-sided test, use $\min(u_1, u_2) = u_2$
The standard normal test statistics is

$$Z = \frac{U_2 - \mu_{U_2}}{\sigma_{U_2}} = \frac{12.5 - 21}{\sqrt{49}} = -1.214286$$

We reject H_0 if ($Z < -1.960$ or if $Z > 1.960$ or) $|Z| > Z_{\alpha/2} = 1.96$.
Conclusion: We can not reject H_0.

P-value is approximately 0.23.

For the non-parametric Wilcoxon signed-rank test for the data in a), this test has lower power than the t-test when data is normally distributed. We see that the p-value is much larger in the non-parametric test than in the t-test.

Problem 2 Cement hydration

a) Write down the fitted regression model.

$$y = 62.41 + 1.5511x_1 + 0.5102x_2 + 0.1019x_3 - 0.1441x_4 \quad (1)$$

Hypothesis for x_3:

$$H_0 : \beta_3 = 0 \text{ vs. } H_1 : \beta_3 \neq 0$$

The p-value is found to be 0.896. Given that the truth is that $\beta_3 = 0$ there is a 0.896 probability to observe a test statistic T which is at least as extreme ($t_{obs} \leq -0.14$ or $t_{obs} \geq 0.14$) as what we have observed.

Is this a good model?

- Linearity: looking at the scatter plots and correlation we see a linear trend and correlation in x_2 vs. y, x_4 vs. y, x_4 vs. x_2 and x_3 vs. x_1. In the plot of the standardized residuals vs. fitted value we see no clear trend, and thus may assume that linearity in the parameters of the model may be an adequate assumption. No clear trend in the standardized residuals vs. fitted value plot also indicates equal variance of errors (homoscadasticity).
- Covariates included in the model: No covariate gave a p -value below 0.05 when testing each of the covariates (x_1 was almost significant, p-value of 0.071). All of the covariates seems to have a high correlation with y (x_3 has the lowest correlation with y). We may try to refit the model without the x_3 term, this will also change the estimated coefficients for the other covariates. All of the covariates and response seems to be highly correlated (not x_3 and x_4). In an overall level the regression is found to explain more than just the average yield level (p -value for the regression is 0.000).
- Normality of errors: looking at the qq-plot for the standardized residuals the assumption of normality seems plausible.
- Explanatory powers: the model explains 98.2% of the variability of the data, which is a high number.

Conclusion: the model seem to be adequate.

b) Model B only includes two covariates, while Model A has four. The estimated regression coefficients for the variables that are present in both models, x_1 and x_2 are different for Model A and Model B (because the covariates are correlated). The p-values for the coefficients in Model B are smaller than those for Model A. Model A explained (R^2_{adj}) 97.4% of the variability in the data, and Model B explains also 97.4%.

Formally: let $SSR(modelA)$ be the regression sums of squares for model A and $SSR(modelB)$ be the regression sums of squares for model B. Further, $SSE(modelA)$ is the error sums of squares for the full model A. The difference in number of parameters between model A and B is $m = 2$ and $n-k-1 = 13-4-1 = 8$ is the degrees of freedom for $SSE(modelA)$. Under the null hypothesis the test statistic F follows a Fisher distribution with $m = 2$ and $n-k-1 = 8$ degrees of freedom.

$$F = \frac{SSR(modelA) - SSR(modelB)}{m} / \frac{SSE(modelA)}{n-k-1} = \frac{2667.90 - 2657.9}{2} / \frac{47.86}{8} = 0.836. \tag{2}$$

The critical value in the Fisher distribution with m and (n-k-1) df is 4.46 at level 0.05 and the null hypothesis can not be rejected. This implies that β_3 and β_4 are simultaneously zero. This means that model B is preferred to model A.

Looking at the R^2_{adj} also gives the same conclusion, as they are approximately the same.

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.
- 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.

Looking first at C_p, these seems to be good models

- the model containing x_1 and x_2 contains 3 parameters, C_p value is 2.7. C_p indicats overfitting.
- the model containing x_1, x_2 and x_4 contains 4 parameters, C_p value is 3.0. C_p indicats overfitting.
- the model containing x_1, x_2 and x_3 contains 4 parameters, C_p value is 3.0. C_p indicats overfitting.

(and the full modell, C_p=5, p=5, and is assumed to be a good fit, and should not use C_p to evaluate model for the full model).

There are not much difference in these three models above when looking at C_p. Based on the adjusted R^2 and S I would recommend the model with the three variables x_1, x_2, and x_4. (The model with only x_1 and x_2 is also a good alternative). But, we also need to examine residual plots and model fit for this model in order to conclude.

Problem 3 Pain and hair colour

a) In the ANOVA output the sums of squares (SS) of the total variability in the data is decomposed into variability of pain score between hair colours types (SSA/"HairColour"), and variability within hair color types (SSE/"Error", the residuals), SST=SSA+SSE. The sum of squares give these decomposed numerical values. A degree of freedom (DF) is associated with each sum, reflecting the amount of information in the sum (- and technically associating the scaled sum with a χ^2-distribution with this number of degrees of freedom). The Mean squares (MS) are the Sum of squares (SS) divided by the degrees of freedom. The F-value is the ratio between the MS for the hair colours types ("HairColour") and the MS for the ERROR/residuals. The p-value related to the F-value and the F-distribution. The null hypotheses testes are wrt the parameters means, μ_i (or α_i) begin equal (or the same for the other parameters). S is the estimate for σ, and is the $\sqrt{MSE}$

Missing entries:
DF Error: number of observations - number of groups = 19-4=15, or DF Total - DF HairColour.
SS HairColour: MS HairColour= SS HairColour /DF HairColour, that gives MS HairColour

$\times$ DF HairColour =SS HairColour, $453.6 \times 3 = 1360.8$
SS Error: SS Total - SS HairColour=2362.5-1360.8=1001.7.
MS Error: SS Error/DF Error = 1001.7/15=66.78.
F : MS HairColour/MS Error=453.6/66.78=6.792.
S: In the one-way ANOVA we assume a common variance for the pain in all groups. This means that we may pool together the individual empirical variances to estimate a common variance for all groups. The weighing factor is the number of observations minus one (for the group mean) in each group. Here S_i^2 is the estimator for σ^2 based on the ith group.

$$S^2 = \frac{\sum_{i=1}^k (n_i - 1)S_i^2}{\sum_{i=1}^k (n_i - 1)} = \frac{\sum_{i=1}^k (n_i - 1)S_i^2}{n - k} = \frac{SSE}{n - k}.$$

Numerically

$$S = \sqrt{(5-1)\cdot 9.284^2 + (5-1)\cdot 8.325^2 + (5-1)\cdot 8.526^2 + (4-1)\cdot 5.447^2}/\sqrt{19-4} = 8.172$$

or

$$S = \sqrt{SSE/n-k} = \sqrt{MSE} = \sqrt{66.78} = 8.172$$

Assumptions :
The one-way ANOVA can be written as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{3}$$

where ϵ_{ij} is i.i.d. $\mathcal{N}(0, \sigma^2)$, that is the error terms are independent and normally distributed with the same variance across treatment groups.

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0 \text{ vs. } H_1 : \text{ not all are equal}$$

This can be tested using the F-statistics, if there are significantly variation between the hair colour types than variation within the different types. $F = 6.792$ and we got a p-value for this test at 0.004, which is lower than our significance level and we can reject the null hypothesis, and we have reason to believe that the pain sensitivity is not the same for all hair colours.

b)

$$H_0 : \mu_{LightBlonds} - \mu_{DarkBrunette} \text{ vs. } H_1 : \text{not equal}$$

$$t = \frac{\bar{y}_{LightBlonds} - \bar{y}_{DarkBrunette}}{S\sqrt{\frac{1}{n_{LightBlonds}} + \frac{1}{n_{DarkBrunette}}}}. \tag{4}$$

t=$(59.200 - 37.400)/8.172\sqrt{1/5 + 1/5}$= 4.217918

Degrees of freedom to S found in ANOVA table (S_{pooled}) equals to (n-k) 15, for a two-sided test the critical value is $t_\alpha/2, 15$=2.131.

Conclusion: reject the null hypothesis. Light blonds are significantly different from Dark brunette.

Compute the 95% confidence interval on $\mu_{LightBlonds} - \mu_{DarkBrunette}$ as

$$y_{LightBlonds} - y_{DarkBrunette} \pm t_{\alpha/2} s \sqrt{\frac{1}{n_{LightBlonds}} + \frac{1}{n_{DarkBrunette}}},$$

where $t_{\alpha/2,15}$= 2.131 (df from S). Numerically

$$(59.200 - 37.400) \pm 2.131 \cdot 8.172 \sqrt{\frac{1}{5} + \frac{1}{5}} = 21.8 \pm 11.01392 = (10.79, 32.81).$$

All values in the interval are considered plausible values for the parameter being estimated. If the value of the parameter specified by the null hypothesis is contained in the 95% interval then the null hypothesis cannot be rejected at the 0.05 level. We observe that the interval do not cover zero, we can reject H_0, the difference in pain between Light blonds and dark brunettes is significant. Same result as with the t-test.

Problem 4 Satisfaction of custumers

a) Hypothesis:

$$H_0 : \text{column probabilities are the same for each row vs. } H_1 : \text{not so}$$

$$H_0 : p_{South-west} = p_{South-East} = p_{Middle} = p_{North} H_1 : \text{at least one differ}$$

We will use a χ^2-test for homogeneity, where the test statistics approximately follows a χ^2-distribution with $(c-1)(r-1)$ degrees of freedom. Here $c = 3$ and $r = 4$, yielding $2 \cdot 3 = 6$ degrees of freedom.

Expected frequencies are calculated as (columns totals)·(row totals)/(grand total). The table of observed and expected values (e) are as follows:

Region/Number of costumers that are:	Satisfied	Don't know	Discontent	Total
South-west	235(224,70)	74(64,24)	89(109,06)	398
South-East	654 (658,29)	203 (188,21)	309 (319,50)	1166
Middle	366 (388,99)	79(111,22)	244 (188,80)	689
North	179 (162,03)	54 (46,33)	54(78,64)	287
Total	1434	410	696	2540

Showing how the Satisfied and South-west cell expected value is calculated: $398 \times 1434/2540 = 224.70$.

The contribution from this cell to the test statistic is $\frac{(235-224.70)^2}{224.70} = 0.472$. The test statistic consists of 12 terms, and is given as $X^2 = \frac{(235-224.70)^2}{224.70} + ... + \frac{(54-78.64)^2}{78.64} = 44.780$.

The null hypothesis is rejected if the test statistics is larger than $\chi^2_{0.05,6} = 12.592$.

Conclusion: clearly we can reject the null hypothesis and we have reason to believe that the proportions of degree of satisfaction of the chain-store customers differ between the regions (South-west, South-East, Middle and North).