Norwegian University of Science and Technology
Department of Mathematical Sciences

Tentative solutions TMA4255 Applied Statistics 30 May, 2014

Problem 1 Darwins corn plants

a) First we inspect the data:
Do we have independent samples?
The two samples are paired by design, pairs of cross-fertilization (X_{1i}) and self-fertilization (X_{2i}) plants grown together under identical conditions. The observations of height of the cross-fertilization and self-fertilization plants are correlated. This means that we cannot use methods that are designed for independent samples, and we need to choose a method for paired samples. This can be done by forming differences, $D_i = X_{1i} - X_{2i}$ and using one-sample methods.

Do we have normality?
When X_{1i} and X_{2i} are assumed normally distributed, $X_{1i} \sim N(\mu_1, \sigma^2)$ and $X_{2i} \sim N(\mu_2, \sigma^2)$, $i = 1, ..., 15$, it follows that the difference $D_i = X_{1i} - X_{2i}$ also can be assumed normally distributed, $N(\mu, \sigma^2)$ and $\bar{D} = \frac{1}{15}(\sum x_{1i} - \sum x_{2i} \sim N(\mu_D, \frac{\sigma^2}{15})$. D_i is i.i.d (although $X_{1i} - X_{2i}$ are dependent). With these assumptions we can use a t-test for the difference, D_i (t-test for paired data).

We want to test if the mean height of the cross-fertilized plants are larger than the mean hight of the self-fertilized plants. The hypothesis can be written as

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

or define $\mu_D = \mu_1 - \mu_2$

$$H_0 : \mu_D = 0 \text{ vs. } H_1 : \mu_D > 0$$

We test this hypothesis by performing a t-test based on the test statistic

$$T = \frac{\bar{D}}{S_D/\sqrt{n}},$$

where $S_D = \sqrt{\frac{1}{14}\sum_{i=1}^{15}(D_i - \bar{D})^2}$. Under H_0 T $\sim t_{n-1}$. It is given that $\bar{d} = 21.53$, $s_d = 38.29$ and $n = 15$.

$$t_{obs} = \frac{21.53}{38.29/\sqrt{15}} = 2.177.$$

This is a one-sided test, and using significance level $\alpha = 0.05$ we reject the null hypothesis when $t_{obs} > t_{\alpha,n-1}$. From the tables we find that the critical values are $t_{0.05,14} = 1.761$.

Conclusion: We reject the null hypothesis and we have reason to believe that cross-fertilization plants are taller than self-fertilization plants.

We could have performed a non-parametric Wilcoxon signed-rank test for paired data in a), however, when data is normally distributed this test has lower power than the t-test.

b) When we assume that X_{1i} and X_{2i} are not normally distributed we can perform a non-parametric test. If we assume nothing about the data, we can perform a sign test for paired data. We test the hypothesis that the median of the differences ($\tilde{m_D} = \tilde{m_1} - \tilde{m_2}$) is larger than zero.

$$H_0 : \tilde{m}_1 = \tilde{m}_2 \text{ vs. } H_1 : \tilde{m}_1 > \tilde{m}_2$$

or written as

$$H_0 : \tilde{m}_D = 0 \text{ vs. } H_1 : \tilde{m}_D > 0$$

We test this hypothesis by performing a sign-test based on a binomial variable test statistic X with $n = 15$ trials and success probability $p = \frac{1}{2}$. We reject the null hypothesis in favor of the alternative hypothesis only if the proportion of plus signs is sufficiently greater than 1/2 (when x is large).

We compute the differences, d_i

$$58, -67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60, -48.$$

Replace each positive difference by a "+" symbol and each negative difference by a "-" symbol (discard any zero differences) we obtain the sequence

$$+, -, +, +, +, +, +, +, +, +, +, +, +, +, -,$$

let $x = 13$ and $n = 15$.

We can calculate the p-value of this test as $P = P(X \geq 13 \text{ when } p = 1/2) = 1 - \sum_{x=0}^{12} b(x; 15, \frac{1}{2}) = 1 - 0.9963 = 0.0037$. The p-value is therefore smaller than the significance level, we can reject the null hypothesis.

Since $n > 10$ we may also use an normal approximation, since $np = nq > 5$, but this will be an approximation. We may use the normal approximation with

$$\mu = np = 15 \cdot 0.5 = 7.5$$

and

$$\sigma = \sqrt{npq} = \sqrt{15 \cdot 0.5 \cdot 0.5} = 1.936492,$$

we can find that

$$z = \frac{13 - 7.5}{1.936492} = 2.84$$

(we can also use 12.5 (instead of 13) to correct for that the binomial/discrete distribution is approximated by a continuous/normal distribution). We get the p-value $P = P(X \geq 13) \approx P(Z \geq 2.84) = 1 - P(Z \leq 2.84) = 1 - 0.9977 = 0.0023$, which leads to the rejection of the null hypothesis. We have reason to believe that the cross-fertilized plants are taller than self-fertilized plants.

The sign test are known to have lower power to detect any significance compared to the t-test. We see that the p-value is higher for the sign test (the binomial and normal approximation gave approximate the same p-values) than the t-test. The advantage of the sign test is that it can always be used as it is based on no assumptions.

Problem 2 House sparrows

a) Hypothesis:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

We test this hypothesis by performing a t-test based on the test statistic

$$T = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta}_1)}$$

From the MINITAB output we find that the estimated coefficient $\hat{\beta}_1 = 0.11584$ and estimated standard deviation of the regression coefficients β_1 is $\hat{SE}(\hat{\beta}_1) = 0.06609$, which give us the test statistic $t_{obs} = \frac{0.11584}{0.06609} = 1.75$.

This is a two-sided test, and using significance level $\alpha = 0.05$ we reject the null hypothesis when $t_{obs} > t_{\alpha/2,n-k-1}$ or when $t_{obs} > t_{1-\alpha/2,n-k-1}$, where $n = 901$ and $k = 3$ is the number of parameters estimated. From the tables we find that the critical values are $t_{0.025,897} = 1.96$ and $t_{0.0975,897} = -1.96$. The observed value are not more extreme than the critical values and we do not reject the null hypothesis.

Conclusion: β_1 is not significant, and x_1 do not explain a significant portion of the variation in the response.

R^2, is the coefficient of multiple determination and is interpreted as the amount of variability in the response that is accounted for by the regression and is defined as

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

where SSE is the sum-of-squares of error, SSR is the regression sum-of-squares, and SST is the total sum of squares. The R^2 found in the MINITAB output as $R - Sq = 24,1\%$. The regression is found to explains 24.1% of the variation in the data. A better goodness of fit measure is the adjusted R^2, when comparing models.

b) The estimated variance in the regression model, $s^2 = SSE/(n - k - 1)$ is an appropriate estimator for σ^2. The estimated standard deviation in the regression model (model error

variance) is found in the MINITAB output to be $S = 1,54791$. Estimated variance $s^2 = 1.54791^2 = 2.4$.

Find the error sum of squares, SSE, which is the variation in response that remains unexplained after taking into account the covariates.

$$s = \sqrt{SSE/(n-k-1)} \Rightarrow SSE = s^2 \cdot (n-k-1)$$

gives us

$$SSE = 1.54791^2 \cdot 897 = 2149.24.$$

Find the mean error sum of squares, MSE, $MSE = s^2$:

$$MSE = SSE/DF = 2149.24/897 = 2.4.$$

Find the total sum of squares, SST, the total variation in response:

$$SST = SSR + SSE = 682.54 + 2149.24 = 2831.78.$$

Find the F test statistics, F, used to test if the regression is significant

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_1 : \text{ at least one different from zero}$$

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-k-1)}} = \frac{\frac{SSR}{k}}{s^2} = \frac{\frac{682.54}{3}}{1.54791^2} = 94.95.$$

The p-value of the test is $P = 0.000$, which indicates that we can reject the null hypothesis and assume that at least one of the covariates are nonzero and that the regression is significant. This is seen in Figure 1, where both x_2 and x_3 is significant (p-value smaller than $\alpha = 0.05$).

Below is the values for the question marks filled in

```
Analysis of Variance

Source           DF       SS      MS      F      P
Regression        3   682,54  227,51  94,95  0,000
Residual Error  897  2149,24    2,40
Total           900  2831,78
```

Figure 1: Printout from statistical analysis of the house sparrow data set.

c) The estimated regression model is found from the output in Figure 1

$$\hat{y} = 4.97 + 0.116 \cdot x_1 - 0.663 \cdot x_2 + 0.715 \cdot x_3$$

A point estimate $\hat{y}_0$ can be found for observed tarsus length $x_1^0 = 20$, bill depth $x_2^0 = 8$ and total badge size $x_3^0 = 19$.

$$\hat{y_0} = 4.97 + 0.116 \cdot 20 - 0.663 \cdot 8 + 0.715 \cdot 19 = 15.571.$$

It is given that the estimated standard deviation of $\hat{y}_0$ is 0.097.

A 95% confidence interval for the expected value of $\hat{y}_0$ ($E(\mathbf{Y}|\mathbf{X_0}) = \mu_{\mathbf{Y}}|\mathbf{x_0}$) can be calculated as

$$\hat{y_0} \pm t_{0.025,897} \cdot s\sqrt{\mathbf{x_0^T}(\mathbf{x^T x})^{-1}\mathbf{x_0}},$$

where $\hat{SD}(\hat{y_0}) = s\sqrt{\mathbf{x_0^T}(\mathbf{x^T x})^{-1}\mathbf{x_0}}$. The critical value is $t_{0.025,897} = 1.96$, we have

$$s\sqrt{\mathbf{x_0^T}(\mathbf{x^T x})^{-1}\mathbf{x_0}} = 0.097.$$

A 95% confidence interval for $\hat{y}_0$:

$$(15.571 \pm 1.96 \cdot 0.097) = (15.38088, 15.76112).$$

Calculate a 95% prediction interval for y_0:

$$\hat{y_0} \pm t_{0.025,897} \cdot s\sqrt{1 + \mathbf{x_0^T}(\mathbf{x^T x})^{-1}\mathbf{x_0}}.$$

We have that

$$\mathbf{x_0^T}(\mathbf{x^T x})^{-1}\mathbf{x_0} = (\frac{0.097}{1.54791})^2 = 0.003927.$$

A 95% prediction interval for y_0:

$$15.571 \pm 1.96 \cdot 1.54791\sqrt{1 + 0.003927} = (12.53115, 18.61085).$$

What is the difference in interpretation of the two intervals?

We are interested in finding a 95% confidence interval on the mean response, $\mu_{\mathbf{Y}}|\mathbf{x_0} = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \beta_3 x_3^0$ for a set of values $\mathbf{x_0}$. That is an interval within which we can say with 95% certainty that the mean response will fall.

$\hat{y}$ is an unbiased estimate for the mean response, which we are interested in finding a confidence interval for. $\hat{y_0}$ is therefore a estimate for mean response for a set of values $\mathbf{x_0}$.

The 95% prediction interval is an interval within which we can say with 95% probability that a new observed response, y_0, for a set of values $\mathbf{x_0}$ will fall.

The confidence interval for the mean response gives us the quality of predicted response, y_0. Prediction intervals account for the variability around the mean response inherent in any prediction, making the prediction interval wider than the confidence interval.

d) There are four principal assumptions made in the linear regression model in Equation (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.

We can not observe this random error term, and we therefore use the residuals (standardized) ($e_i = y_i - \hat{y}_i$) to test the assumptions from the fitted regression.

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 3 upper right panel we see that there seems to be a trend in the residuals not all symmetrical around 0. We have to look further at plots of residuals vs covariate to determine further the linearity of the residuals (not included in the problem).
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 3 lower right panel we see that there seems to be some small trend in the residuals, a curve in the residuals indicating a correlation over the observations.
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 3 upper right panel we see that there is some 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 3 upper left panel we see that some of the points lies outside the 95% confidence interval, and a S-shaped pattern i evident. A p-value of < 0.05 of the Anderson Darling test, indicates that the residuals are not normally distributed. Also the histogram lower left panel indicates a skewed distribution of the residuals.

How do you think this effects the results given in the outputs?

1. If the data are not linearly related and you fit a linear model, your predications are likely to have serious errors, especially when you extrapolate beyond the range of the sample data.

2. Violation of nonrandom residuals indicates that the value of one observation is not completely independent of another observation. This often indicates a mis-specified model. This may lead to biased standard errors and hence significance tests may not be valid.
3. If the variance of error is increasing over the fitted values, the confidence intervals may be to narrow when we predict for values beyond the range of the data. Non constance variance may lead to biased standard errors and hence significance tests may not be valid.
4. Non normal residuals may effect the estimation of coefficients and calculation of confidence intervals. If there are some large outliers (the error distribution is skewed) these outliers can exert a disproportionate influence on the estimation of coefficients (that is based on minimization of squared error). The calculation of confidence intervals and various significance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

Conclusion: Both the standard error, and estimated coefficient may be biased and therefore the conclusion of the hypothesis tests may be wrong. The prediction and confidence interval may also not be trustworthy since we do not have normally distributed errors.

A transformation may correct for the non-normal errors (transformation may often help both non-normality and non-constant variation of the random errors) or a new specification of the model.

All models are wrong, some are useful.

Problem 3 Forging of piston rings

Based on 21 samples, each based on five observations (thus a rational subgroup size of $n = 5$) assumed to be in control, we find $\bar{\bar{x}} = 74.001$ and $\bar{s} = 0.00995$.

a) Construct a S-chart and a $\bar{X}$-S-chart (with 3σ limits).

S-chart has limits

$$\bar{S} \pm 3\frac{\bar{S}}{c_4}\sqrt{1-c_4^2}$$
$$[B_3\bar{S}, B_4\bar{S}]$$

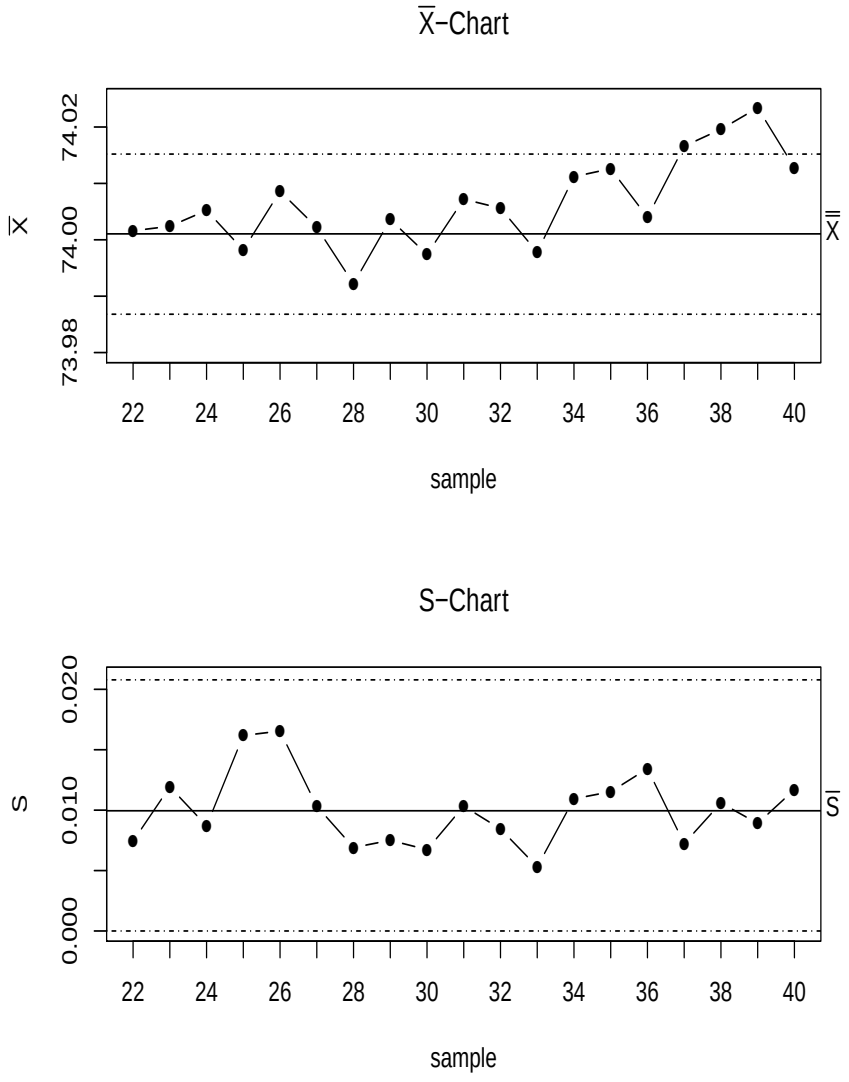
According to table A22 (page 766), and for rational subgroup size $n = 5$, we have $c_4 = 0.94$, $B_3 = 0$ and $B_4 = 2.089$. Thus, the S-chart has lower limit equal to 0 and upper limit equal to $2.089 \cdot 0.00995 = 0.020785$.

$\bar{X}$-S-chart has limits

$$\bar{\bar{X}} \pm 3\frac{\bar{S}}{c_4\sqrt{n}} = \bar{\bar{X}} \pm A_3\bar{S}$$

According to table A22 (page 766), and for rational subgroup size $n = 5$, we have $c_4 = 0.94$ and $A_3 = 1.427$. Thus, the chart has lower limit equal to $74.001 - 1.427 \cdot 0.00995 = 73.9868$ and upper limit equal to $74.001 + 1.427 \cdot 0.00995 = 74.0152$

The figure below is the new samples with the control limits drawn in



The new samples are within the control limits for the S-chart (variance is in control), but the new samples are outside the control limits for the $\bar{X}$-S-chart. The process mean is out of control, not stable for sample (37)38-40. .

Problem 4 Good and bad husbands and wives

a) Hypothesis:

H_0 : the temperament of the husband and wife are independent vs. H_1 : not so

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

We need to calculate the expected frequency for each entry of the contingency table, which again is based on calculating the probability for each grade under the null hypothesis.

P(good husband $\cap$ good wife)= P(good husband) $\cdot$ P(good wife)
P(bad husband $\cap$ good wife)= P(bad husband) $\cdot$ P(good wife)
P(good husband $\cap$ bas wife)= P(good husband) $\cdot$ P(bad wife)
P(bad husband $\cap$ bad wife)= P(bad husband) $\cdot$ P(bad wife)

$\widehat{\text{P(good wife)}} = \frac{24+34}{111} = 0.523.$
$\widehat{\text{P(bad wife)}} = 1 - \widehat{\text{P(good wife)}} = 0.477.$
$\widehat{\text{P(good husband)}} = \frac{24+27}{111} = 0.459.$
$\widehat{\text{P(bad husband)}} = 1 - \widehat{\text{P(good husband)}} = 0.5405.$

The expected value for each table entry is found as $e_i = p_i \cdot n$, with $n = 111$.

$e_{11} = \widehat{\text{P(good husband)}} \cdot \widehat{\text{P(good wife)}} \cdot 111 = 26.65$
$e_{21} = \widehat{\text{P(bad husband)}} \cdot \widehat{\text{P(good wife)}} \cdot 111 = 31.35$
$e_{12} = \widehat{\text{P(good husband)}} \cdot \widehat{\text{P(bad wife)}} \cdot 111 = 24.35$
$e_{22} = \widehat{\text{P(bad husband)}} \cdot \widehat{\text{P(bad wife)}} \cdot 111 = 28.65$

Expected frequencies can also be calculated as (columns totals)$\cdot$(row totals)/(grand total). The table of observed and expected values (e) are as follows:

	Good wife	Bad wife	Total
Good husband	24 (26.65)	27 (24.35)	51
Bad husband	34 (31.35)	26 (28.65)	60
Total	58	53	111

The test statistics consists of 4 terms and is given as

$X^2 = \frac{(24-26.65)^2}{26.65} + ... + \frac{(26-28.65)^2}{28.65} = 1.020$

The null hypothesis is rejected if the test statistics is larger than $\chi^2_{0.05,1} = 3.841$.

Conclusion: clearly we can not reject the null hypothesis and we have no reason to believe that the temperament of husband and wife are dependent of each other.