Norwegian University of Science and Technology
Department of Mathematical Sciences

Tentative solutions
TMA4255 Applied Statistics
June 7, 2013

Problem 1 Body fat and foot treatment

a) Independent samples?
The two samples are paired *by design* (same person before and after). The two observations of the same person are correlated. This is also clearly seen in the scatter plot of the before vs the after measurements. 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 and using one-sample methods.

Normality?
From the normal plots we see that the before sample and the after samples are not normally distributed. However, from the normal plot we see that the differences are approximately normally distributed. (If you study the printout you also find a p-value for testing the null hypothesis that the data is a random sample from a normal population. A high p-value will not reject this null hypothesis.) These findings mean that when analysing the differences we may assume a normal distribution, and can use a t-test.

Then decide which of the four analyses you think fits the research question and the data the best.

According to the above assessment the t-test applied to the difference between the after and before values is preferred. The reason for not selected the Wilcoxon signed-rank test (analysis C) is that this test has lower power than the t-test when the data are normal.

Write down the null- and alternative hypotheses begin tested, and report the result of the hypothesis test.
Assume that $D_i = X_{2i} - X_{1i}$ and D_i i.i.d $N(\mu, \sigma^2)$, $i = 1, \ldots, 40$.

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

The p-value from the one-sample t-test for the differences is 0.095. We use significance level 0.05 and do not reject the null hypothesis. We do not have sufficient evidence to believe that there is a effect of foot treatment on the body fat percentage measurements.

Problem 2 Process control with resistors

Based on $k = 30$ samples, each based on three observations (thus a rational subgroup size of $n = 3$) assumed to be in control, we find $\bar{\bar{x}} = 5.095$ and $\bar{s} = 0.058$.

Remark: here the rational subgroup size is 3, not 30.

a) Construct a S-chart and a $\bar{X} - S$-chart (with 3 σ 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 $n = 3$, we have $c_4 = 0.8862$, $B_3 = 0$ and $B_4 = 2.568$. Thus, the S-chart has lower limit equal to 0 and upper limit equal to $2.568 \cdot 0.058 = 0.149$.

$\bar{X} - S$-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 $n = 3$, we have $c_4 = 0.8862$, $A_3 = 1.954$ Thus, the chart has lower limit equal to $5.095 - 1.954 \cdot 0.058 = 4.98$ and upper limit equal to $5.095 + 1.954 \cdot 0.058 = 5.21$.

A new sample is measured, with $\bar{x} = 5.15$ and $s = 0.10$. This is within the control limits both for the $\bar{X} - S$ chart and the S-chart.

What would be the advantange of using a Cusum–chart istead of a Shewhart–chart (as constructed here)?
Assume that a process is drifting out of control, that is, the change in a parameter of interest is slow and monotone. Using a Shewhart–chart we will spend a long time before the drift is detected, but with a Cusum–chart the cumulative sum is monitored, and this makes the Cusum–chart more suitable to quickly detect the drift.

Problem 3 We use the χ^2 goodness of fit test, based on calculated expected frequencies using the distribution under the null hypothesis.

We need to calculate the expected frequency for the number of bomb hits, which again is based on calculating the probability the number of bomb hits under the null hypothesis. We have 5 cells in our observed data table. Define:

- p_i expected probability for cell i (probability of the given number of bomb hits to a small area).
- o_i observed count in cell i (observed number of small areas with the given number of bomb hits).
- e_i expected count in cell i (expected number of small areas with the given number of bomb hits).

Let $X \sim Poisson(\hat{\lambda} = 0.93)$

$$
\begin{aligned}
P(X = 0) &= \frac{e^{-0.93}0.93^0}{0!} = \exp -0.93 = 0.39 \\
P(X = 1) &= \frac{e^{-0.93}0.93^1}{1!} = 0.37 \\
P(X = 2) &= \frac{e^{-0.93}0.93^2}{2!} = 0.17 \\
P(X = 3) &= \frac{e^{-0.93}0.93^3}{3!} = 0.05 \\
P(X \geq 4) &= 1 - P(X \leq 3) = 1 - 0.39 - 0.37 - 0.17 - 0.05 = 0.02
\end{aligned}
$$

NB: differences with your solution may be due to rounding - here two decimals are used.

The expected value for the number of bomb hits is found as $e_i = n \cdot p_i$, with $n = 576$. That is, $e_0 = 576 \cdot 0.39 = 227.64$, $e_1 = 576 \cdot 0.37 = 213.12$, and so on.

Table of probabilies, expected and observed frequencies.

Hits (k)	0	1	2	3	4 or more	Total
Observed	229	211	93	35	8	576
Expected	227.6	213.1	97.9	28.8	11.5	575.5

The test statistic is

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

which is approximately χ^2 distributed with $k-1-1$ degrees of freedom, where k is the number of groups used. We have used $k = 5$. The last 1 is deducted because we have estimated the parameter λ.

$$X^2 = (229 - 224.6)^2/224.6 + \cdots + (8 - 11.5)^2/11.5 = 2.75.$$

Critical value in the χ^2 distribution with 3 degrees of freedom is for $\alpha = 0.05$ equal to 7.815 (text book, page 740, row 3).

This means that we don't reject the null hypothesis, and stick with the Poission distribution.

Problem 4 Teaching reading

a) We would like to investigate if the expected reading score varies between the teaching methods.
Write down the null- and alternative hypothesis and perform one hypothesis test based on the summary statistics in the table above.

What are the assumptions you need to make to use this test?
What is the conclusion from the test? Hypotheses:
Let μ_A, μ_B and μ_c be the expected reading scores for each of the three methods.

$$H_0 : \mu_A = \mu_B = \mu_C \text{ vs. } H_1 : \text{at least one pair differs}$$

This hypothesis can be tested using one-way analysis of variance. We need to fill in the ANOVA table (SS, MS, df, F), which can be calculated from the summary statistics.

Let $\bar{x}_A$ denote the average and s_A the standard deviation of method A. Ditto for methods B and C. Let $\bar{x}$ denote the grand mean.

$$\begin{aligned} SSA &= n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2 + n_C(\bar{x}_C - \bar{x})^2 \\ &= 22 \cdot (41.05 - 44.02)^2 + 22 \cdot (46.73 - 44.02)^2 + 22 \cdot (44.27 - 44.02)^2 \\ &= 357.005 \\ SSE &= (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2 \\ &= 25511.712 \end{aligned}$$

Source	SS	df	MS	F
Method	357.005	2	178.5	4.47
Error	2511.712	63	39.9	
Total	2868.717	65		

The F statistic, here observed to be 4.47, should be compared with the critical value $f_{0.05,2,63}$. We find $f_{0.05,2,60} = 3.15$ in Table A.6, and we thus reject the null hypothesis. (We know that $f_{0.05,2,63} < f_{0.05,2,60}$.)

Assumptions:
The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

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

Conclusion:
There is reason to believe that the expected reading score is not the same for all the methods.

b) Let $\bar{X}_B$ be the mean of a random sample from using method A and $\bar{X}_C$ the mean of a random sample from using method C. A natural estimator for γ is

$$\hat{\gamma} = \frac{\bar{X}_B}{\bar{X}_C}$$

We turn to first order Taylor approximations with

$$\begin{aligned} h(\bar{X}_B, \bar{X}_C) &= \frac{\bar{X}_B}{\bar{X}_C} \\ \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_B} &= \frac{1}{\bar{X}_C} \\ \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_C} &= -\frac{\bar{X}_B}{\bar{X}_C^2} \end{aligned}$$

where the random variable $\bar{X}_B$ has $\mathrm{E}(\bar{X}_B) = \mu_B$ and $\mathrm{Var}(\bar{X}_B) = \sigma_B^2/n_B$, and $\bar{X}_C$ has $\mathrm{E}(\bar{X}_C) = \mu_C$ and $\mathrm{Var}(\bar{X}_C) = \sigma_C^2/n_C$.

Define

$$\begin{aligned} h'_B(\mu_B, \mu_C) &= \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_B} \mid_{\bar{X}_B = \mu_B, \bar{X}_C = \mu_C} = \frac{1}{\mu_C} \\ h'_C(\mu_B, \mu_C) &= \frac{\partial h(\bar{X}_B, \bar{X}_C)}{\partial \bar{X}_C} \mid_{\bar{X}_B = \mu_B, \bar{X}_C = \mu_C} = -\frac{\mu_B}{\mu_C^2} \end{aligned}$$

We assume that the two samples are independent. The first order Taylor approximation for two independent samples:

$$\begin{aligned}
\mathrm{E}(h(\bar{X}_B, \bar{X}_C)) &\approx h(\mu_B, \mu_C) = \frac{\mu_B}{\mu_C} \\
\mathrm{Var}(h(\bar{X}_B, \bar{X}_C)) &\approx (h'_B(\mu_B, \mu_C))^2 \,\mathrm{Var}(\bar{X}_B) + (h'_C(\mu_B, \mu_C))^2 \,\mathrm{Var}(\bar{X}_C) \\
&= \left(\frac{1}{\mu_C})\right)^2 \cdot \sigma_B^2/n_B + \left(-\frac{\mu_B}{\mu_C^2}\right)^2 \cdot \sigma_C^2/n_C
\end{aligned}$$

Estimates using numerical values $n_B = 22$, $n_C = 22$, $\hat{\mu}_B = \bar{x}_B = 46.73$, $\hat{\mu}_C = \bar{x}_C = 44.27$, $\hat{\sigma}_B^2 = s_B^2 = 7.388^2$, $\hat{\sigma}_C^2 = s_C^2 = 2^2$ are as follows.

$$\begin{aligned}
\hat{\gamma} &= \frac{46.73}{44.27} = 1.06 \\
\mathrm{E}(h(\bar{X}_B, \bar{X}_C)) &\approx \frac{46.73}{44.27} = 1.06 \\
\mathrm{Var}(h(\bar{X}_B, \bar{X}_C) &\approx (\frac{1}{44.27})^2 \cdot 7.388^2/22 + (\frac{46.73}{44.27^2})^2 \cdot 5.767^2/22 \\
&= 0.00127 + 0.00086 = 0.00212 \\
\mathrm{SD}(h(\bar{X}_B, \bar{X}_C)) &\approx \sqrt{0.00212} = 0.046
\end{aligned}$$

Problem 5 Concrete

a) Write down the estimated regression equation:

$$\hat{y} = -574.2 + 0.02670x_1 + 1.3612x_2 + 124.08x_3 + 23.33x_4$$

The estimated regression coefficient for x_3 is 124.08. If we look at two construction jobs, A and B, that have the same values for x_1, x_2 and x_4, but the value for x_3 is 1 unit (percent point) higher for job 1 than for job 2. Then the regression model estimates that job 1 will need 124.08 m^3 more concrete than job 2.

Is the effect of x_3, `waste`, significant in this model? The null hypothesis to be tested is $H_0 : \beta_3 = 0$ versus the alternative $H_1 : \beta_3 \neq 0$. To to that a regression t-test is used. The t-statistics is 2.51 with 23 degrees of freedom. The critical value in the t-distribution for a two-sided test with significance level 0.05 is $t_{0.025,23} = 2.069$. We thus reject the null hypothesis and we conclude that the effect of waste is signficant in this model.

R^2: coefficient of multiple determination is defined as

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

We have $R^2 = 6586535/7162388 = 0.9196$, which can be given in percentage as 91.96%. This can be interpreted as the proportion (percentage) of variability in the data that is explained by the Model A regression model.

Do you think model A is a good model for the data?

- Linearity: looking at the scatter plots we see a linear trend in all of the covariates vs. y. There are three observations with high value for y that deviates from the linear trend for x_3 and x_4. In the plot of the studentized 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.
- Covariates included in the model: The covariates x_2, x_3, x_4 are significant in the model. The x_1 covariate gives a p-value above 0.05 when testing each of the covariates. This might be due to x_1 being correlated with one or several of the other covariates. Pairwise scatterplots of the covariates would help us assess this.
- Normality of errors: looking at the normal plot for the studentized residuals the assumption of normality seems plausible.
- Explanatory powers: the model explains 91.96% of the variability of the data, which is a high number.

Conclusion: the model seem to be good.

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_2 and x_3, are different for Model A and Model B. The p-values for the coefficients in Model B are smaller than those for Model A. Model A explained 91.96% of the variablity in the data, while Model B explains 89.1%.

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 $n-k-1 = 28-4-1 = 23$ is the degrees of freedom for $SSE(modelA)$. Under the null hypothesis the test statistic F follows a Fisher distribution with $m = 2$ and $n - k - 1 = 23$ degrees of freedom.

$$F = \frac{\frac{SSR(modelA)-SSR(modelB)}{m}}{\frac{SSE(modelA)}{n-k-1}} = \frac{\frac{6586535-6381667}{2}}{25037} = 4.09$$

The critical value in the Fisher distribution is 3.42 at level 0.05 and the null hypothesis is rejected. This means that model A is preferred.

Based on the F-test I would prefer model A. Looking at the R^2-adjusted also gives the same conclusion (comparing 90.56% for Model A with 88.2% for Model B).

c) The best prediction for the quantity of concrete to be used in a new construction job with $x_2 = 300$ and $x_3 = 4.4$ is the fitted value with these covariates,

$$\hat{y} = -815.6 + 1.7575 \cdot 300 + 219.65 \cdot 4.4 = 678.1$$

Let $\boldsymbol{x}_0 = (1, 300, 4.4)$ be the new 3×1 covariate vector for our prediction.

From our textbook we find a $(1-\alpha)\cdot 100$ % prediction interval for new observation Y_0 to be

$$[\boldsymbol{x}_0^T\boldsymbol{B} \pm t_{n-k-1}(\frac{\alpha}{2})\sqrt{(1+\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)s^2}]$$

where $\boldsymbol{B} = (-815.6, 1.7575, 219.65)$ is our vector of estimated regression coefficients and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is the 3 ×3 matrix given in the printout from Model B, and $s^2 = 31229$ (MSE, or S^2) from Model B. The value for the the vector-matrix-vector-product $\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$ was given in the text. For those intereste - the value was calculated from

$$\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 = \begin{bmatrix} 1 & 300 & 4.4 \end{bmatrix} \begin{bmatrix} 0.970948 & -0.0000188 & -0.212061 \\ -0.000019 & 0.0000007 & -0.000046 \\ -0.212061 & -0.0000458 & 0.051464 \end{bmatrix} \begin{bmatrix} 1 \\ 300 \\ 4.4 \end{bmatrix} = 0.0357$$

Further we need $t_{28-2-1,0.025} = t_{25,0.025} = 2.060$.

$$678.1 \pm 2.060 \cdot \sqrt{(1+0.0357)\cdot 31229} = 678.1 \pm 2.060 \cdot 179.8 = [307.7, 1048.5]$$

Thus, the 95% prediction interval is given as [307.7,1048.5], and can be interpreted as an interval were a new value for the concrete quantity lies with probability 95%.