Tentative solutions to TMA4255 Applied Statistics, August 8, 2012

Problem 1 Use of cannabis among youths in Norway

a) Hypotheses:
Let p_{2004}, p_{2005}, ..., p_{2008} be the probability that a randomly chosen youth from the surveyed population has used cannabis in the given year.

$$H_0 : p_{2004} = p_{2005} = \cdots = p_{2008} \text{ vs. } H_1 : \text{at least one pair differs}$$

We will use a χ^2-test for homogeniety, where the test statistic approximately follows a χ^2-distribution with $(c-1)\cdot(r-1)$ degrees of freedom. Here $c = 5$ and $r = 2$, yielding $4 \cdot 1 = 4$ degrees of freedom.

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

Result	2004	2005	2006	2007	2008	Total
Used	222 (201.5)	261 (212.4)	206 (191.4)	357 (395.7)	281 (325.9)	1327
Not used	1432 (1452.5)	1482 (1530.6)	1365 (1379.6)	2891 (2852.3)	2394 (2349.1)	9564
Total	1654	1743	1571	3248	2675	10891

Showing how the Used and 2004 cell expected value is calculated: $1327 \cdot 1654/10891 = 201.5$. The contribution from this cell to the test statistic is $\frac{(222-201.5)^2}{201.5} = 2.09$

The test statistic consists of 10 terms, and is given as

$$X^2 = \frac{(222-201.5)^2}{201.5} + \frac{(261-212.4)^2}{212.4} + \cdots + \frac{(2394-2349.1)^2}{2349.1} = 27.7$$

The null hypothesis is rejected if the test statistics is larger than $\chi^2_{0.05,4} = 9.49$. Clearly, the null hypothesis is rejected.

Conclusion:
There is reason to believe at at least two of the years have different probability of cannabis usage among the youths.

Problem 2 Process development

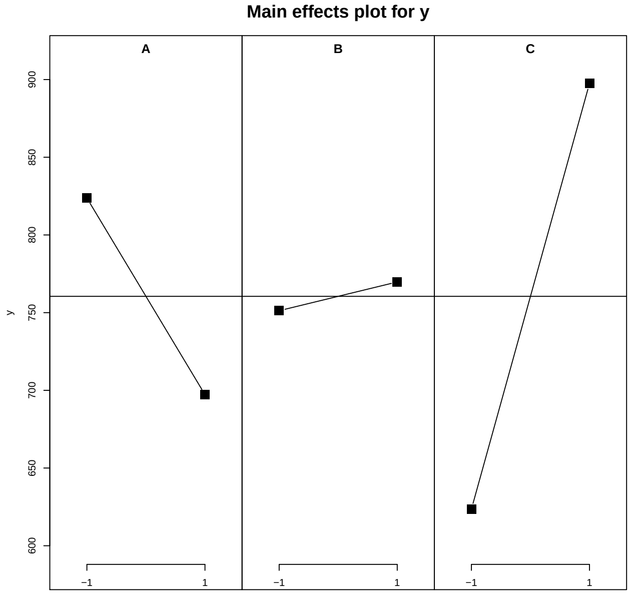
Run	A	B	C	Response
1	-1	-1	-1	550
2	1	-1	-1	669
3	-1	1	-1	633
4	1	1	-1	642
5	-1	-1	1	1037
6	1	-1	1	749
7	-1	1	1	1075
8	1	1	1	729

Intercept	A	B	C	AB	AC	BC	ABC
760.50	-126.5	*	274.0	-42.0	-190.5	-9.5	13.0

a) Let y_i be the response in run i.

$$\begin{aligned}\hat{B} &= \text{mean response with B is high} - \text{mean response when B is low}\\ &= (y_3 + y_4 + y_7 + y_8)/4 - (y_1 + y_2 + y_5 + y_6)/4\\ &= (633 + 642 + 1075 + 729)/4 - (550 + 669 + 1037 + 749)/4\\ &= 769.75 - 751.25 = 18.5\end{aligned}$$

The main effects plot for B shows that the mean B response at the low level is at 769.75, and going from the low to the high level the mean B response increases with 18.5 to 751.25. The increase from the low to the high mean level of B is the B main effect.



b) The “Std. Error” column gives the estimated standard deviation of the regression coefficients. Let s^2 be the estimated variance in the regression model (estimate for σ^2). Due to the orthogonality of the DOE design all estimated standard deviations are $s/\sqrt{n}$ where $n = 16$. From the printout we see that $47.46/4 = 11.865$ for all regression coefficients.

The estimated effect for B is by definition twice the estimated coefficient for B.

The Estimate is the estimated regression coefficient, the Std.Error is the estimated standard deviation of the regression coefficient, the t-value is the value of the t-statistics (see below), the p-value is from the test described below.

The t-statistic: Estimate/Std.Error=3.688/11.865=0.311.
H_0: The coefficient for the covariate B is zero, H_1: different from zero. A p-value of 0.76 means that we do not reject H_0 at significance level 0.05 and assume that the B coefficient is zero - and can be removed from the model.

What are the significant covariates in the model? Significant covariates are A, C and AC (and the intercept).

c) Since we have an orthogonal design the presence of factors othogonal to A and C does not change the parameter estimates for the regression coefficients in the model. But, the regression model is important for the estimation of the error variance σ^2 and the Std.Error will then change with the change in the model.

Just looking at the estimated coefficients in the reduced model we see that the etching rate will increase with C and decrease with A. This would suggest to keep A at the low level and C at the high level. The interaction between A and C is negative, so with A at low level and C at high level the net effect is positive.

We may also calculate the estimated response (predictions) with the four combinations of A and C, which confirms that A low and C high is optimal.

A low and C low: $\hat{y} = 776.062 + 50.812 - 153.062 - 76.812 = 597$.
A low and C high: $\hat{y} = 776.062 + 50.812 - 153.062 + 76.812 = 1056.75$.
A high and C low: $\hat{y} = 776.062 - 50.812 - 153.062 + 76.812 = 649$.
A high and C high: $\hat{y} = 776.062 - 50.812 + 153.062 - 76.812 = 801.5$.
Calculate a 95% prediction interval for the etch rate based on your chosen levels for A and C. Since we have an othogonal design, the covariance matrix for the regression coefficients will be diagonal (all correlations are zero). The formula for the prediction interval with covariates $\boldsymbol{x_0}$ is

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

The covariate vector is $x_0 = (1, -1, 1, -1)$ for the intercept, A at low and C at high and thus AC at low level. $\boldsymbol{B}$ is the vector of regression coefficients for the intercept, A, C and AC, thus $(776.06, -50.81, 153.06, -76.81)$. The matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is a diagonal matrix with 1/16 on the diagonal. s is read off the printout as 41.96. The t critical value is $t_{16-3-1}(0.25) = t_{12}(0.25) = 2.18$.

$\boldsymbol{x}_0^T\boldsymbol{B} = y_0 = 1056.75$ and we add
$2.18\cdot\sqrt{1+(1,-1,1,-1)diag(1/16)(1,-1,1,-1)}\cdot 41.96 = 2.18\cdot\sqrt{1+4/16}\cdot 41.96 = 102.3$.
The interval is then $[954.45, 1159.05]$.

d) We now assume that in a pilot study with three factors only runs 1, 4, 6 and 7 from the table in the start of this problem were performed.

This is a half fraction of a 2^3 experiment, thus a 2^{3-1} experiment. The generator for the design is $AB = -C$, and the defining relation is thus $I = -ABC$. The alias stucture is: $A = -BC$, $B = -AC$, $C = -AB$. The defining relation has three letters, and thus this is a resolution III experiment.

Problem 3 Gene activity

a) The two-way ANOVA model can be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where α_i are the cell population levels, β_j are the gene levels and γ_{ij} the cell population and gene interaction. In the ANOVA output the sums of squares of the total variability in the data is decomposed into variability between the cell population (A), between the genes (B) and due to interaction between cell population and gene (AB), and the remaining variability is called the residual variability, SStot=SSA+SSB+SSAB+SSError. The sum of squares give these decomposed numerical values. A degree of freedom is associated with each sum, reflecting the amount of information in the sum - and technially associating the scaled sum with a χ^2 distribution with this number of degrees of freedom. The Mean squares are the Sum of squares divided by the degrees of freedom. The F-value is the ratio between the Mean Sq for the given factor (A, B or AB) and the Mean Sq for the residual. The p-value related to each of three tests based on the F-value and the F-distribution. The null hypotheses testes are wrt the parameters α_i begin equal (or the same for the other parameters).

Missing entries:
Df for Total: number of observation -1=45-1=44, or as the sum of the other dfs.
Sum Sq for AB: Mean Sq for AB $\cdot$ df for AB=$30.645 \cdot 4 = 122.98$.
Mean Sq for A: Sum Sq for A / df for A=$3.240/2 = 1.62$.
p-value for A: Tail in the Fisher distr with 2 and 36 df for F=5.935. Fom the table for 0.05 this critical value is around 3.3-3.3, which means that the p-value is below 0.05. For the 0.01 table the critical value is arond 5.4, which means that the p-value is below 0.01. Computer software (which is not available at the exam) would give 0.006 as the p-value.

The interaction term has a p-value below $2 \cdot 10^{-16}$ and is thus significant. We could proceed to look at differences between genes for specific cell populations or vice versa.

b) Assumptions:
We assume that the data are normally distributed, that is, $X_i \sim N(\mu_2, \sigma^2)$ for A_1 and B_1 and $Y_j \sim N(\mu_2, \tau^2)$ for A_1 and B_2, $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$, and that the two samples are independent.

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

Equal variances:
Assuming data to be normally distributed we may test the equality of variance by performing an F-test. The null and alternative hypothesis:

$$H_0 : \sigma^2/\tau^2 = 1 \text{ vs. } H_1 : \sigma^2/\tau^2 \neq 1$$

Let S_1^2 and S_2^2 be the variances of two independent random samples of size n_1 and n_2 taken from normal populations with variances σ^2 and τ^2, respectively, then

$$F = \frac{S_1^2/\sigma^2}{S_2^2/\tau^2}$$

has an F-distribution with n_1-1 and n_2-1 degrees of freedom. Under the null $F = S_1^2/S_2^2$ and since we have a two-sided test we reject the null when $f_{obs} > f_{\alpha/2,n_1-1,n_2-1}$ or when $f_{obs} < f_{1-\alpha/2,n_1-1,n_2-1}$. We only have tables for small values for α, so we need to use the relationship

$$f_{1-\alpha,\nu_1,\nu_2} = \frac{1}{f_{\alpha,\nu_2,\nu_1}}$$

From our data we have $f_{obs} = 0.34^2/0.42^2 = 0.66$ and with $\alpha = 0.02$ (why: only tables for 0.05 and 0.01 in the textbook) we find from the tables that the critical values are $f_{0.01,3,9} = 6.99$ and $f_{0.99,2,9} = 1/f_{0.01,3,9} = 1/6.99 = 0.14$. Thus we do not reject the null and conclude that we may assume that the variances are equal. Comment: using $\alpha = 0.05$ (not in the tables of the textbook) would give a cut-off of 5.08 and 0.20 - and thus the same conclusion as for $\alpha = 0.02$. Comment: in real life, where at computer is available the preferred strategy would be to not perform an hypothsis test for equal variance, but instead use the t-test with estimated variances. To test for equal variances or not is a highly debated topic in statistics.

t-test:
We choose to use a pooled estimate for the variance, $S_p^2 = \frac{(n_1-1)\cdot S_1^2+(n_2-1)\cdot S_2^2}{n_1-n_2-2}$ and get $s_p = \sqrt{\frac{2\cdot 0.34^2+8\cdot 0.42^2}{3+9-2}} = \sqrt{0.16} = 0.4$. The t-test is based on the t-statistic

$$T = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}S_p}}$$

which we calculate to be $t_{obs} = \frac{-29.1+33.7}{\sqrt{\frac{1}{3}+\frac{1}{9}\cdot 0.16}} = \frac{4.6}{0.27} = 17.0$. This is a two-sided test, and using significance level $\alpha = 0.05$ we reject the null hypothesis when $t_{obs} > t_{\alpha/2,n_1+n_2-2}$ or when $t_{obs} < t_{1-\alpha/2,n_1+n_2-2}$. From the tables we find that the critical values are $t_{0.025,10} = 2.23$ and $t_{0.975,10} = -2.23$. We have observed a value more extreme than the critical values and we reject the null hypothesis.

Conclusion:
We have reason to believe that there is a difference in the gene activity for gene B_1 and B_2 in cell population A_1.

c) Let $\bar{X}$ be the mean of group 1 and $\bar{Y}$ the mean of a group 2. A natural estimator for γ is

$$\hat{\gamma} = \frac{2^{\bar{X}}}{2^{\bar{Y}}}$$

We turn to first order Taylor approximations with

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

where the random variable $\bar{X}$ has $\mathrm{E}(\bar{X}) = \mu_1$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n_1$, and $\bar{Y}$ has $\mathrm{E}(\bar{Y}) = \mu_2$ and $\mathrm{Var}(\bar{Y}) = \tau^2/n_2$.

Define

$$\begin{aligned}
h'_X(\mu_1,\mu_2) &= \frac{\partial h(\bar{X},\bar{Y})}{\partial \bar{X}} \mid_{\bar{X}=\mu_1,\bar{Y}=\mu_2} = \ln 2 \frac{2^{\mu_1}}{2^{\mu_2}} \\
h'_Y(\mu_1,\mu_2) &= \frac{\partial h(\bar{X},\bar{Y})}{\partial \bar{Y}} \mid_{\bar{X}=\mu_1,\bar{Y}=\mu_2} = -\ln 2 \frac{2^{\mu_1}}{2^{\mu_2}}
\end{aligned}$$

The first order Taylor approximation for two independent samples:

$$\begin{aligned}
\mathrm{E}(h(\bar{X},\bar{Y})) &\approx h(\mu_1,\mu_2) = \frac{2^{\mu_1}}{2^{\mu_2}} \\
\mathrm{Var}(h(\bar{X},\bar{Y})) &\approx (h'_X(\mu_1,\mu_2))^2 \mathrm{Var}(\bar{X}) + (h'_Y(\mu_1,\mu_2))^2 \mathrm{Var}(\bar{Y}) \\
&= (\ln 2 \frac{2^{\mu_1}}{2^{\mu_2}})^2 \sigma^2/n_1 + (-\ln 2 \frac{2^{\mu_1}}{2^{\mu_2}})^2 \tau^2/n_2
\end{aligned}$$

Estimates using numerical values $n_1 = 3$, $n_2 = 9$, $\hat{\mu}_1 = \bar{x} = -29.1$, $\hat{\mu}_2 = \bar{y} = -33.7$, $\hat{\sigma}^2 = s_1^2 = 0.34^2$, $\hat{\tau}^2 = s_2^2 = 0.42^2$ are as follows.

$$\begin{aligned}
\hat{\mathrm{E}}(h(\bar{X},\bar{Y})) &\approx \frac{2^{-29.1}}{2^{-33.7}} = 24.25 \\
\hat{\mathrm{Var}}(h(\bar{X},\bar{Y})) &\approx (\ln 2 \frac{2^{-29.1}}{2^{-33.7}})^2 \cdot (\frac{0.34^2}{3} + \frac{0.42^2}{9}) = 282.57 \cdot 0.058 = 16.42
\end{aligned}$$