Solution. Exam Stat 310 Dec 17 2013.

Please send comments to Thore.Egeland@umb.no

Exercise 1

a) (1)=<u>1</u>. Since there are two groups
(2)=7.4817/0.9463=<u>7.91</u>
(3)=22*0.9463=<u>20.82</u>

b) *Formulate hypotheses and test if there is a significant difference between the exposed workers and the control group. Formulate a conclusion.*
The hypotheses are

$$H_0 : \tau_1 = \tau_2,$$

$$H_1 : \tau_1 \neq \tau_2.$$

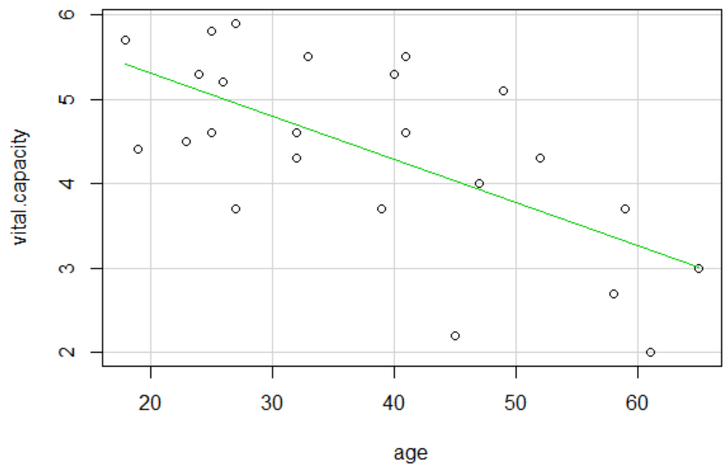
The p-value is 0.01 which is below 0.05. We therefore reject the null hypothesis and claim that there is a difference between the exposed and the control group.

c) *Comment on the assumptions of the model in 1a). Refer to Figure 1.*
The error terms should be (i) normally distributed, (ii) have constant variance and (iii) be independent. The assumption of normality seems OK as the points are close to the line. The upper left plot should show no systematic pattern for assumption (ii) to be OK and this may perhaps be dubious. [Different answers are accepted as long as answers refer to the plot]. The assumption of independence is judged based on the information given. One could comment that this assumption could be questioned as there could be correlation within factories.

d) The p-value = 0.0005 is below 0.05 and we therefore reject the null hypothesis and claim that there is a significant age effect. We find

$$\widehat{\beta}_0 = \underline{6.3} \text{ (intercept)}$$

$$\widehat{\beta}_1 = \underline{-0.05} \text{ (slope)}$$



e) *A statistician suggested that an analysis of covariance should be performed to study the effect of* ***group****,* ***age*** *and interaction between* ***group*** *and* ***age*** *on* ***vital.capacity.*** *Formulate the model. Include the assumptions in your description of the model.*

*Test if there is some **group** effect based on the output from Table 2 and Table 3.*

The model and assumptions are

$$y_i = \beta_0 + \beta_1 * group_i + \beta_2 * age_i + \beta_3 group_i * age_i + e_i$$

$e_i, i = 1, 2, ..., 24$ are assumed independent and normally distributed with constant variance, i.e., $e_i \sim NID(0, \sigma^2)$

The hypotheses and test statistics are

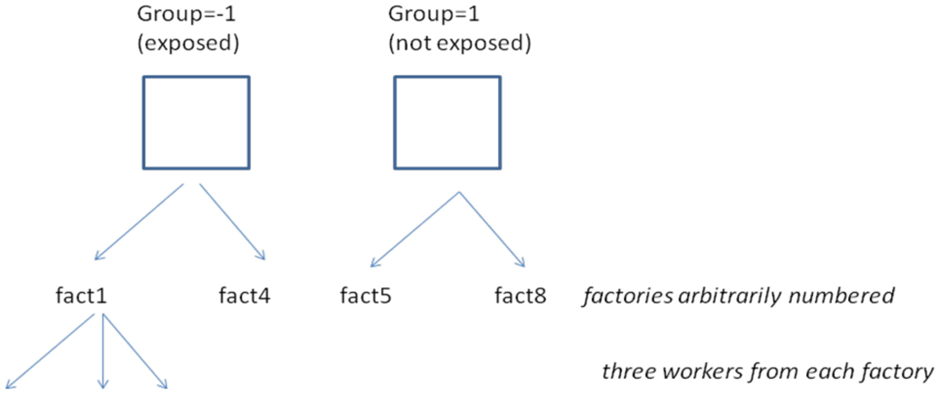
$$H_0 : \beta_1 = 0 \text{ and } \beta_3 = 0$$
$$H_1 : \beta_1 \neq 0 \text{ and}, \text{or } \beta_3 \neq 0$$
$$F_0 = \frac{[SS_E(red) - SS_E(full)]/r}{MS_E(full)} = \frac{[16.24 - 15.26]/2}{0.7632} = 0.64$$

From table (draw figure): $F_{0.05;2,20} = 3.49$

The critical value 3.49 is not exceeded and therefore the null is not rejected. There is no significant effect of **group.**

Exercise 2

a) *Explain why it is reasonable to consider **factory** to be a random effect and **group** a fixed effect. Explain why this is a nested design. Draw a figure.*
The factories are randomly chosen and so we are not specifically interested in these four factories and therefore this is a random effect.
We are specifically interested in the effect of exposure and therefore **group** is fixed.



b) *Formulate hypotheses and test if there is a significant difference between the exposed workers and the control group. Formulate a conclusion.*

$H_0 : \tau_1 = \tau_2,$
$H_1 : \tau_1 \neq \tau_2.$
The test statistic is

$$F_0 = \frac{7.48}{1.47} = 5.08$$

The critical value $F_{0.05,1,6} = 5.99$

is not exceeded and therefore we don't reject.

c) *Estimate σ^2 and find a 95% confidence interval for σ^2. Interpret the confidence interval. Estimate the correlation coefficient for* ***vital.capacity*** *for workers in the same factory. Comment on the answer.*
We find

$$\widehat{\sigma^2} = 0.75$$

95% confidence interval:

$$\left[\frac{11.99}{28.85}, \frac{11.99}{6.91}\right] = \underline{[0.42, 1.74]}.$$

We are 95% certain that this interval contains the true variance. The correlation coefficient is

$$\hat{\rho} = \frac{\widehat{\sigma_\beta^2}}{\widehat{\sigma_\beta^2} + \widehat{\sigma^2}}, \text{ where}$$

$$\widehat{\sigma^2} = 0.75$$

$$\widehat{\sigma_\beta^2} = \frac{1.47 - 0.75}{3} = 0.24$$

Therefore

$$\hat{\rho} = \frac{0.24}{0.24 + 0.75} = \underline{0.24}$$

Exercise 3

a) It was argued that differences in ***vital.capacity*** *can be explained by different smoking habits. We would like to design a new study where* ***vital.capacity*** *and* ***smoking.status*** *('Yes' or 'No') is recorded.* ***vital.capacity*** *is assumed to be normally distributed with known standard deviation 1.1. Explain what we need specify to calculate the required sample size.*
We need to specify the response variable, in this case **vital.capacity.** Furthermore the power, typically 0.8 and the significance level, typically 0.05, must be specified. Finally the smallest relevant difference, in this case difference in **vital.capacity** between smokers and non-smokers must be given. *Balanced* means that there should be equally many smokers and non-smokers.

b) With the specifications above we find

$z_\alpha = z_{0.05} = 1.64$ (comment: 1.645 or 1.65 is also OK) since the area to the right of 1.64 is 0.05 (equivalently: area to the left is 1-0.5=0.95), $z_{1-\beta} = z_{0.8} = -0.84$ since the area to the right of -0.84 is 0.8 (equivalently: area to the left of 0.84=0.8).

From this follows

$$n \geq \left[\frac{(1.64-(-0.84))1.1}{0.5}\right]^2 = 30$$

<u>and so a total of 30 (29 or 31 is OK) individuals is needed.</u>