Tentative solutions
TMA4255 Applied Statistics
9 August, 2014

Problem 1 Reiki treatment of Fibromyalgia

a) From this study can we conclude that the VAS score at enrollment (e.g. before treatment) differs between groups, A, B, C and D?

Write down the null hypothesis and the alternative hypothesis, perform one hypothesis test based on the descriptive measures above. Use significance level $\alpha = 0.05$.
Specify the assumptions you make and the conclusion of the test.

Hypothesis:

$$H_0 : \text{ the mean pain is equal across groups } \text{ vs. } H_1 : \text{ the groups are not equal}$$

or

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

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 in group A. Ditto for groups B, C and D. Let $\bar{x}$ denote the grand mean.

$$\begin{aligned} SSA &= n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2 + n_C(\bar{x}_C - \bar{x})^2 + n_D(\bar{x}_D - \bar{x}) && (1) \\ &= 0.25 + 0 + 4 + 2.25 = 6.5 && (2) \end{aligned}$$

$$\begin{aligned} SSE &= (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2 + (n_D - 1)s_D^2 && (3) \\ &= (25 - 1)2.2^2 + (25 - 1)2.6^2 + (25 - 1)2.1^2 + (25 - 1)2.4^2 = 522.48 && (4) \end{aligned}$$

	SS	df	MS	F
Group	6.5	3	2.17	0.39
Error	522.48	96	5.44	
Total	528.98	99		

Critical value: $f_{0.05,3,96} = 2.70$. ($f_{0.05,3,60} = 2.76$ and $f_{0.05,3,120} = 2.68$)

Assumptions:
The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{5}$$

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

Conclusion: we can not reject H_0, we have no reason to believe that there are difference in the VAS score before enrollment between the different groups.

b) Based on these data, do we have reason to believe that Reiki treatment has an larger effect on pain, measured as VAS score, than placebo treatment given by an actor? Write down the null hypothesis and the alternative hypothesis, choose a test statistics and perform a hypothesis test. Use significance level $\alpha = 0.05$.
Specify the assumptions you make.

Assumptions:

- Independent samples
- Each sample comes from a normally distributed population.
- Equal variance in the samples

Then we can use a unpaired t-test.

We are interested in finding out if Reiki treatment had an larger effect on pain than the placebo treatment. A negative value of the difference μ_X-μ_Y would indicate a lower VAS score for Reiki treatment than for the placebo treatment.

In hypotheses testing the alternative hypothesis tells us what predictions we made about the effect and the predicted direction of this effect. If the alternative hypothesis predicts direction of the effect we have a one-sided hypothesis.

A one-sided hypothesis in our problem can be written as

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_1 : \mu_X < \mu_Y.$$

However, it does not indicate the direction of the effect, positive or negative (that is, it is not given a specific value of d_0 ($\mu_X - \mu_Y < d_0$)).

It may be more useful here to test a two-sided hypothesis

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_1 : \mu_X \neq \mu_Y$$
$$H_0 : \mu_X - \mu_Y = d_0 \text{ vs. } H_1 : \mu_X - \mu_Y \neq d_0,$$

where $d_0 = 0$ and we do not make a choice over the direction that the effect takes (could be negative or positive). It is appropriate when we predict an effect, but we don't predict the direction of the effect.

The two-sided hypothesis is most appropriate to use here, however, the one-sided hypothesis is ok to use, since we here may not be so interested in a larger effect in the wrong direction.

With equal variance between the groups we can use the pooled estimate of the standard deviation:

$$T = \frac{\hat{D}}{SE(\hat{D})} = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \tag{6}$$

$$T = \frac{-0.4}{1.23\sqrt{1/25 + 1/25}} = -1.149767 \tag{7}$$

Two-sided hypothesis:
Reject H_0 if $|T| > t_{\alpha/2,n_X+n_Y-2}$, $t_{\alpha/2,48} \approx 2.021$ ($t_{\alpha/2,40} = 2.021$ $t_{\alpha/2,60} = 2.000$)

One-sided hypothesis:
Reject H_0 if $T > t_{\alpha,48} \approx t_{\alpha,40} = 1.684$.

Conclusion: we can not reject H_0, we have no reason to believe that Reiki treatment has an effect/larger effect on pain, measured in VAS score, than placebo treatment given by an actor. The treatments doesn't differ.

Problem 2 Cheddar cheese Taste

a) From the p-value for the Acetic coefficient, β_1, given in Figure 1, the acetic acid variable is significant at a 5% significance level.

What can you conclude from the given p-values in Figure 1, 2 and 3 about the three chemicals influence on taste?

P-values comes from testing each of the hypotheses:

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

$$H_0 : \beta_2 = 0 \text{ vs. } H_1 : \beta_2 \neq 0$$

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

We see that all the three variables has a significant effect on the taste of cheddar cheese, $p < 0.05$.

Comment on the values of R^2 in Figure 1, 2 and 3.

We observe that the sum of R^2 from the 3 different models are larger than 100%. The three models together explains more than 100% of the variability in the data. This is of course not possible, and there must be some common information in the three variables x_1, x_2 and x_3. This would mean that the three covariates x_1, x_2, x_3 are correlated.

Find a 90% confidence interval for β_1.

CI for β_1:

$$[\text{coefficient} \pm t_{\alpha/2,(n-2)}\text{SE(coefficient)}],$$

$t_{0.05,28=1.701}$

$$[15.648 \pm 1.701 \cdot 4.496] = [8.000304, 23.2957]$$

What is the predicted taste score for a Acetic acid value of $x_1^0 = 7$?

$$y = -61.5 + 15.6 \cdot x_1$$
$$y = -61.5 + 15.648 \cdot 7 = 48.036$$

b) Find an appropriate estimate for σ, and calculate a 90% confidence interval for σ in the regression model in Equation (1) (Hint: use that SSE/σ^2 is chi-square distributed).

An appropriate estimate for σ is the estimated standard deviation in the regression model, $s = 13.8212$. Here $s = \sqrt{\frac{1}{n-2}SSE}$.

Use that $\mathrm{SSE}/\sigma^2 \sim \chi^2_{n-2}$ for a simple linear regression. Confidence interval for σ^2:

$$P(\chi^2_{\alpha/2,28} < \frac{SSE}{\sigma^2} < \chi^2_{1-\alpha/2,28}) = 0.90$$

$$P(41.337 < \frac{SSE}{\sigma^2} < 16.928) = 0.90$$

$$P(\frac{SSE}{41.337} < \sigma^2 < \frac{SSE}{16.928}) = 0.90$$

$$\begin{aligned} s^2 &= 13.82^2 = 190.9924 \\ s^2 &= SSE/(n-2) \\ SSE &= (n-2) \cdot s^2 = 5348.716 \end{aligned}$$

i.e. CI for σ^2:

$$(129.3929, 315.9686)$$

and to find CI for σ we take the square root of the interval limits.

CI for σ:

$$(11.3751, 17.7755)$$

How can we use this confidence interval to test the null hypothesis $H_0 : \sigma = 1$? Write down the alternative hypothesis, give the conclusion of this test and the significance level.

Hypothesis:

$$H_0 : \sigma = 1 \text{ vs. } H_1 : \sigma \neq 1$$

is tested with significance level 0.10 (i.e. 0.90). Rejecting when 1 is not in the 90% CI for σ. So we reject H_0, σ is not equal to 1.

c) Explain the term multicollinearity. Could this be a problem in the regression model in Equation (4)?

If one covariate is correlated with another covariate then we have collinearity. (Not linearity - but a tendency of linear dependence.). With several correlated covariates we call this multicollinearity.

We have that $\mathbf{B} = (\mathbf{X^T X})^{-1}\mathbf{X}^T Y$, $Cov(\mathbf{B}) = \sigma^2(\mathbf{X^T X})^{-1}$. When we have multicollinearity $\mathbf{X^T X}$ may have large diagonal elements. The covariance of $\mathbf{B}$ may be large since $\mathbf{X^T X}$ may be nearly singular.

This will make it difficult to know which variable to include in the model (several variables give much of the same information). The estimate of β_1 in a model with only x_1 will change if x_2 is also included into the model. This will also make prediction difficult since the prediction error will increase rapidly.

Comment on the correlation matrix in Figure 5 and pairwise scatter plot in Figure 6.

We see from Figure 5 that the correlation between each pair of chemicals is high, exceeding 0.6.

From the three-variable regression models in Figure 4 we see that acetic acid, X_1 is not significant, but in Figure 1 (Equation (1)) X_1 was significant. What may be the reason behind this? Justify your answer.

Since we have correlation between each pair of the variables in the model, multicollinearity may be a problem. So X_1 was significant in the model in Equation (1), but as we added more variables/regressors in the regression model (Equation (4)), the X_1 variables changed as these variables/regressors are dependent on each other (correlated), $\mathbf{x}_j^T\mathbf{x}_i \neq 0$.

This is an observational study. Would it be possible to design an experiment (design of experiment) to investigate the problem under study? Elaborate.

Design of experiment (DOE) creates orthogonality, where regressors are independent of each other, and $\mathbf{x}_j^T\mathbf{x}_i = 0$. Here, the estimate of the coefficient are not changed if we change the model. However, it may be difficult (or impossible) to design a study here, making cheeses with given x_1, x_2 and x_3 values.

Problem 3 Toy plastic bricks

Day i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of defects	3	3	1	3	4	6	4	6	4	4	5	2	3	1

a) Make the appropriate control chart to control the probability of defects, p. Use all the data in the table above to calculate control limits (use 3σ limits).
Can we assume that the number of defects in each sample is approximately normally distributed?

This is a p-chart (with 3σ limits): defect or not defect.

Estimate for p: $\bar{p} = \frac{1}{m}\sum_{i=1}^{m}\hat{p}_i$, where $\hat{p}_i$ is the proportion of defects in the sample from day i, X_i/n, $i = 1, ...14$. m is the number of timepoints in control, X_i is the number of defects in sample i, assume $X_i \sim bin(n,p)$, and n is the rational subgroup/trials is $n = 250$.

$$\frac{1}{14}\sum_{i=1}^{14}(\frac{3}{250} + \frac{3}{250} + \frac{1}{250} + ... + \frac{1}{250}) = 0.014. \tag{8}$$

Control limits are estimated as

$$[LCL, UCL] = [\bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}]$$

$$[0.014 \pm 3\sqrt{\frac{0.014(1-0.014)}{250}}] = [-0.0083, 0.0363]$$

We see that LCL is negative. Since the p value can never be negative, the LCL should never be negative, and the LCL will be set to 0. See Walpole, Myers, Myers and Ye: “Probability and Statistics for Engineers and Scientists” chapter 17.

The control limits are then set to $[0, 0.0363]$.

Is the number of defects in a sample approximately normally distributed?

See Walpole, Myers, Myers and Ye: “Probability and Statistics for Engineers and Scientists” chapter 6.5, 8.4. A binomial random variable can be approximated by the normal distribution if n is large, with expectation $\mu = np$ and variance $\sigma^2 = np(1-p)$. We get the largest value of σ^2 by setting $p = 0.5$. Then always $\sigma^2 \leq n/4$. Demanding that $\sigma^2 \geq 5$ (or $np \geq 5$ and $n(1-p) \geq 5$) we automatically secured that $n \geq 20$ which given the central limit theorem gives us an good approximation.

We have that $\widehat{E(X_i)} = n\bar{p} = 3.5 < 5$, so we may not assume that the number of defects are approximately normally distributed.

b) How many observations, n, in each rational subgroup is needed to detect a change from $p = 0.2$ to $p_1 = 0.21$?

1. The problem doesn't specify the probability needed to detect a change. We will use probability 0.5, but other choices are of cause possible.
2. We will assume that $\bar{p} = p$ is known so that the upper limit of the control chart is UCL $= p + 3\sqrt{\frac{p(1-p)}{n}}$.
3. When the true probability of defect is p_1 we have $X_i \sim bin(n, p_1)$ and $\hat{p}_i = \frac{X_i}{n}$ has $E(\hat{p}_i) = p_1$, $Var(\hat{p}_i) = \frac{p_1(1-p_1)}{n}$.
4. We want to find n so that $P(\hat{p}_i > UCL) = 0.5$. (Actually we want $P(\hat{p}_i > UCL) + P(\hat{p}_i < LCL) = 0.5$, but the latter probability will be very small when the true p_1 is larger than p.)
5. $P(\hat{p}_i > UCL) = 0.5$

$$\begin{aligned} P(\hat{p}_i > UCL) &= P(\frac{\hat{p}_i - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} > \frac{UCL - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}) \\ &\approx 1 - \Phi(\frac{UCL - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}), \end{aligned}$$

since we here may assume normality when $np > 5$ & $n(1-p) > 5$.

$$1 - \Phi(\frac{UCL - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} = 0.5$$

$$UCL - p_1 = 0 \Leftrightarrow UCL = p_1$$

When we have probability 0.5 to detect a shift from p to p_1 UCL need to be set at p_1.

6. What does this mean wrt n?

$$UCL = p + 3\sqrt{\frac{p(1-p)}{n}}$$

and

$$UCL = p_1$$

$$p + 3\sqrt{\frac{p(1-p)}{n}} = p_1 \tag{9}$$
$$\frac{p(1-p)}{n} = \frac{(p_1-p)^2}{3^2} \tag{10}$$
$$n = \frac{9p(1-p)}{(p_1-p)^2}. \tag{11}$$

7. When $p = 0.2$ and $p_1 = 0.21$ we have

$$n = \frac{9 \cdot 0.2(1-0.2)}{(0.21-0.20)^2} = 14400. \tag{12}$$

If a probability of detection different from 0.5 is chosen the calculations become much more difficult.

Problem 4 Obesity and alcohol intake

a) Hypothesis:

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

$$H_0 : p_{\text{Low}} = p_{\text{Average}} = p_{\text{High}} \text{ vs. } H_1 : \text{at least one differ.}$$

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 $c = 4$ and $r = 3$, yielding $3 \cdot 2 = 6$ degrees of freedom.

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

	Alcohol Intake				
Obesity	0	1-2	3-5	6+	Total
Low	45(39,32)	45(38,31)	41(45,03)	34(42,34)	165
Average	39(38,36)	32(37,38)	46(43,94)	44(41,32)	161
High	33(39,32)	37(38,31)	47(45,03)	48(42,34)	165
Total	117	114	134	126	491

Showing how the Low and 0 cell expected value is calculated: $117 \cdot 165/491 = 39.32$. The contribution from this cell to the test statistic is $\frac{(45-39.32)^2}{39.32} = 0.82$. The test statistic consists of 12 terms, and is given as $X^2 = \frac{(45-39.32)^2}{39.32} + ... + \frac{(48-42.34)^2}{42.34} = 6.952$

The null hypothesis is rejected if the test statistics is larger than $\chi^2_{0.05,6} = 12.592$.
Conclusion: clearly we can not reject the null hypothesis and we have no reason to believe that the proportions of Low, Average and High obesity differ with respect to alcohol intake.