Final exam Statistical Methods – Solutions

15 December 2015

1.

a) To construct a confidence interval for μ a t-distribution has to be used because the population standard deviation is not known and the sample standard deviation is used to construct confidence intervals.

More precisely, confidence intervals are based on the distribution of the statistic

$$T = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}},$$

which has approximately a t-distribution with $n-1$ degrees of freedom if the sample size n is large $n > 30$ or if the data come from a normal population.

b) Any confidence interval is of the form $[\hat{\mu}-E, \hat{\mu}+E]$, so we find $E = (78.599-68.081)/2 = 5.259$. On the other hand we know that $E = t_{n-1,\alpha/2} \cdot s/\sqrt{n}$ so in our case $E = t_{15,\alpha/2} \cdot 12.00/\sqrt{16.00} = t_{15,\alpha/2} \cdot 3$. Therefore, $t_{15,\alpha/2} = 5.259/3 = 1.753$. Table 3 shows that it corresponds to $\alpha = 0.1$, and the confidence level is 90%.

c) If for 100 samples the interval in part b) is computed, then approximately 90% of these intervals will contain the unknown value of μ.

2.

a) Point estimate $\hat{p} = \frac{77}{100} = 0.77$.

b) The margin of error is given by $E = z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ and since we are interested in the 90% confidence interval, we take $z_{0.05} = 1.645$ (or 1.64, both lead to the same answers after rounding):

$$E = 1.645 \cdot \sqrt{\frac{0.77 \cdot 0.23}{100}} \approx 0.069.$$

As a result, the 90% confidence interval is given by $[0.701, 0.839]$.

c) We should use our result from part a). We then get that the minimal sample size should be greater than

$$\left(\frac{1.96}{E_{\max}}\right)^2 \cdot \hat{p}(1-\hat{p}),$$

where $\hat{p}$ is the estimate from part a), and since $E_{\max}$ is the margin of error from part b), we get that n should be greater than

$$\left(\frac{1.96}{0.069}\right)^2 \cdot 0.77 \cdot 0.23 = 142.9001,$$

so the minimal number of students that should be surveyed to a obtain a 95% confidence interval with the same margin of error is 143.

3.

a) There is no relationship between the two groups of Facebook users, therefore both samples are independent.

b) Here are the usual steps of hypothesis testing:

Step 0: We are investigating the difference of population means (deactivated vs. normal Facebook users).

Step 1: H_0: $\mu_1 = \mu_2$, H_a: $\mu_1 \neq \mu_2$ (there is effect, but no direction of the effect is considered), significance level $\alpha = 0.1$.

Step 2: Samples are independent, we are using x_1, x_2, s_1, s_2.

Step 3: The sample standard deviations are quite different, so it is safe to assume that $\sigma_1 \neq \sigma_2$.

(NB. you can assume that standard deviations are equal, but you have to motivate it. In this question it will only change the number of degrees of freedom in the distribution of the test statistics under the null hypothesis, but not the value of the test statistic nor the final conclusion).

We therefore use the test statistic T_2 that under the null hypothesis has a t-distribution with $\tilde{n} = \min\{35, 35\} = 35$ degrees of freedom. The observed value of the test statistic is

$$t_2 = \frac{7.98 - 8.82}{\sqrt{1.07^2/36 + 2.20^2/36}} \approx -2.060.$$

Since the test is two-tailed, the critical region is based on $t_{\tilde{n},\alpha/2} = t_{35,0.05} = 1.690$ and contains the values to the left of -1.690 and to the right of 1.690.

Step 4: The observed value of the test statistic is in the critical region $-2.060 < -1.690$, so we reject the null hypothesis.

We have enough evidence to support the claim that social networking has an effect on cognitive skills.

c) Both samples have to be independent. Moreover, either both are from a normal population, or both are big ($n_1 > 30$ and $n_2 > 30$). In this case both requirements are met in this case.

---

4. Let $n_1 = 240, x_1 = 49, n_2 = 300, x_2 = 77$.

a) The usual point estimate for the difference between proportions is given by the difference of sample proportions:

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2} \approx 0.204 - 0.257 = -0.053.$$

b) Here are the usual steps of hypothesis testing:

Step 0: We are investigating the difference of population proportions.

Step 1: H_0: $p_1 = p_2$, H_a: $p_1 < p_2$ (the proportion in North Holland is supposedly higher), significance level $\alpha = 0.05$.

Step 2: We are using x_1, x_2, n_1, n_2, but also $\overline{p} = (x_1 + x_2)/(n_1 + n_2) \approx 0.233$.

Step 3: All values of $x_1, x_2, n_1 - x_1, n_2 - x_2$ are greater than 5. We are using the test statistic Z_p that under the null hypothesis has approximately the standard normal distribution. The observed value of the test statistic is

$$z_p = \frac{0.204 - 0.257}{\sqrt{(0.233 \cdot 0.767)/240 + (0.233 \cdot 0.767)/300}} \approx -1.45.$$

The test is left-tailed, so the P-value equals $P(Z_p \leq -1.45) = 0.0735$.

Step 4: The calculated P-value is greater than the significance level, so we do not reject the null hypothesis.

We do not have enough evidence to support the claim that the proportion of app users whose daily average exceeds 5.56 km is higher in North Holland than the corresponding proportion in South Holland.

c) When performing a test, we commit 'Type I error' when we reject the null hypothesis when it is actually true, and we commit 'Type II error' when we fail to reject the null hypothesis when it is actually false. The significance level α is the probability of Type I error.

5.

a) Preference for *RStudio* and Windows were determined after 50 students were selected. Therefore, this is a test for independence. Here are the hypotheses:

$$H_0 : \text{preference for } \textit{RStudio} \text{ and preference for Windows are independent}$$
$$H_a : \text{preference for } \textit{RStudio} \text{ and preference for Windows are dependent}$$

b) Under the null hypothesis, we expect the following table

	Windows	other	total
pure R	10	10	20
RStudio	15	15	30
total	25	25	50

according to the formula

$$E_{ij} = \frac{(i\text{-th row total}) \cdot (j\text{-th column total})}{\text{grand total}}.$$

c) All entries of the expected frequencies table are bigger than 5. We therefore can use the chi-square test statistic given by

$$X^2 = \sum_{\text{cells}} \frac{(O-E)^2}{E} = \sum_{(i,j)} \frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

which in this case has under the null hypothesis approximately a chi-square distribution with 1 degree of freedom.

d) This is a one-tailed test (even though the alternative has no direction) so we reject the null hypothesis for high values of the test statistic, i.e., values above the critical value $\chi^2_{(r-1)(c-1),\alpha}$. We use Table 4 and find $\chi^2_{1,0.05} = 3.841$. Since the observed value of the test statistic $\chi^2 = 3.00$ is less than the critical value, we do not reject the null hypothesis. In other words, we do not have enough evidence to support the claim that preference for *RStudio* and preference for Windows are dependent.

6.

a) The regression equation is based on the least-squares estimates b_0 and b_1. Using the data characteristics we first find

$$b_1 = r \cdot \frac{s_y}{s_x} = -0.90 \cdot \frac{3.68}{1.60} = -2.07,$$

and then

$$b_0 = \overline{y} - b_1 \cdot \overline{x} = 43.37 + 2.07 \cdot 5.34 \approx 54.42.$$

Therefore, the regression equation is given by

$$\hat{y} = 54.42 - 2.07 \cdot x.$$

The predicted maximum speed for the windspeed of 5 Bft ($x = 5.00$) is therefore given by

$$\hat{y} = 54.42 - 2.07 \cdot 5.00 = 44.07.$$

b) The amount of variation in the response variable that can be approximately accounted for by the explanatory variable is given by the so-called coefficient of determination, in this setting equal to the sample linear correlation coefficient squared. Since $r = -0.90$, $r^2 = 0.81$, so approximately 81% of the variation is explained by the regression line.

c) Here are the usual steps of hypothesis testing:

Step 0: The test is about the population linear correlation coefficient ρ.

Step 1: H_0: $\rho = 0$, H_a: $\rho \neq 0$,
significance level $\alpha = 0.01$.

Step 2: We are using the sample linear correlation coefficient r and the sample size $n = 25$.

Step 3: We are using the test statistic T_ρ that under the null hypothesis has a t-distribution with $n - 2 = 23$ degrees of freedom. The observed value of the test statistic is

$$t_\rho = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{-0.90}{0.091} \approx -9.890.$$

Since the test is two-tailed, the critical region is based on $t_{n-2,\alpha/2} = t_{23,0.005} = 2.807$ and contains the values to the left of -2.807 and to the right of 2.807.

Step 4: The observed value of the test statistic is in the critical region $-9.890 < -2.807$, so we reject the null hypothesis.

We have enough evidence to reject the claim that linear correlation coefficient between the maximum speed and the windspeed is zero.

d) In order to test $\beta_1 = 0$ the errors have to be independent and come from a normal distribution with a fixed standard deviation. Normality can be checked with the normal QQ plot of residuals. In this case we see that the points follow a fairly straight line - therefore it is reasonable to assume that this assumption in met. The residual plot is used to verify whether the standard deviation of the errors is fixed. It does not show any specific pattern, which means that it is reasonable to assume that this assumption is also met.