Final exam Statistical Methods – Solutions

20 December 2016

1.

a) Point estimate $\hat{p} = \frac{24}{96} = 0.25$.

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 95% confidence interval, we take $z_{0.025} = 1.96$:

$$E = 1.96 \cdot \sqrt{\frac{0.25 \cdot 0.75}{96}} \approx 0.087.$$

As a result, the 95% confidence interval is given by $(0.163, 0.337)$.

c) If we were to select 100 different samples of the same size and construct the corresponding confidence intervals, then approximately 95 of these intervals would contain the unknown value of p.

2.

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

b) The margin of error is given by $E = t_{n-1,\alpha/2} s_n/\sqrt{n}$ and since we are interested in the 90% confidence interval and $n = 36$, we take $t_{35,0.05} = 1.69$:

$$E = 1.69 \cdot \frac{237.50}{\sqrt{36}} \approx 66.896.$$

As a result, the 90% confidence interval is given by $(625.964, 759.756)$.

c) The length of a confidence interval is $2E$. If one were to construct the 95% confidence interval, they would use $t_{35,0.025} = 2.030$ instead of $t_{35,0.05} = 1.690$, and it would be the only change in E. Since $t_{35,0.05} < t_{35,0.025}$, the new confidence interval would be bigger.

3.

a) Both samples are independent, there is no pairing between households in Amsterdam and Staphorst.

b) Here are the usual steps of hypothesis testing:

Step 0: We are investigating the difference of population means (Amsterdam vs. Staphorst households).

Step 1: H_0: $\mu_1 = \mu_2$, H_a: $\mu_1 < \mu_2$, significance level $\alpha = 0.01$.

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

Step 3: We 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 $\tilde{n} = \min\{40, 40\} = 40$ degrees of freedom. The observed value of the test statistic is

$$t_2 = \frac{0.366 - 0.927}{\sqrt{0.623^2/41 + 0.818^2/41}} \approx -3.494.$$

Since the test is left-tailed, the critical region is based on $t_{\tilde{n},\alpha} = t_{40,0.01} = 2.423$ and contains the values to the left of -2.423.

Step 4: The observed value of the test statistic is in the critical region $-3.494 < -2.423$, so we reject the null hypothesis.

We have enough evidence to support the claim that people living in large cities on average own fewer cars than people living in the rural area.

c) The P-value is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis is true.

d) Since we rejected the null hypothesis in b), the P-value must be less than or equal to the significance level.

---

4. Let $n_1 = 100, x_1 = 36, n_2 = 100, x_2 = 29$.

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

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2} = 0.36 - 0.29 = 0.07.$$

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: $p_1 \neq p_2$, significance level $\alpha = 0.05$.

Step 2: We are using x_1, x_2, n_1, n_2, but also $\overline{p} = (x_1 + x_2)/(n_1 + n_2) = 0.325$.

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

$$z_p = \frac{0.36 - 0.29}{\sqrt{(0.325 \cdot 0.675)/100 + (0.325 \cdot 0.675)/100}} \approx 1.06.$$

The test is two-tailed and the observed value is positive, so the P-value equals

$$2 \cdot P(Z_p \geq 1.06) = 2 \cdot (1 - P(Z_p \leq 1.06)) = 0.2892.$$

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 reject the claim that no matter which method is used, the proportion of users who switch off the ad blocking plug-in after seeing the unreadable article is the same.

c) The data is then presented in a table with 3 rows (corresponding to different methods) and 2 columns (switched off/did not switch off) and the test of homogeneity can be performed. The hypotheses are

H_0: proportions of users switching off the ad blocking plug-in are equal for all three methods

H_a: proportions of users switching off the ad blocking plug-in are not equal for all three methods.

We 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 2 degrees of freedom, where O_{ij} and E_{ij} are observed and expected frequencies, respectively.

5.

a) We need to combine observations corresponding to 2, 3, or 4 chocolate chips per cookie and to 12, 13, 14, or 15 chocolate chips per cookie. The resulting table of observed frequencies is:

Chocolate chips per cookie	4 or less	5	6	7	8	9	10	11	12 or more
Number of cookies	12	9	9	6	10	10	5	4	5

b) The hypotheses are

H_0: frequency counts agree with the claimed distribution

H_a: frequency counts do not agree with the claimed distribution.

Under the null hypothesis, we expect the following frequencies

4 or less	5	6	7	8	9	10	11	12 or more
7.00	6.44	8.54	9.80	9.80	8.68	6.93	5.04	7.77

c) All expected frequencies are bigger than 5.

d) This is a right-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 $\chi^2_{k-1,\alpha}$. Here $k = 9$. We use Table 4 and find $\chi^2_{8,0.1} = 13.362$. Since the observed value of the test statistic $\chi^2 = 8.032$ is less than the critical value, we do not reject the null hypothesis. In other words, we do not have enough evidence to reject the producer's claim.

6.

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

$$b_1 = r \cdot \frac{s_y}{s_x} = -0.64 \cdot \frac{12.05}{1.82} \approx -4.23,$$

and then

$$b_0 = \overline{y} - b_1 \cdot \overline{x} = 34.56 + 4.24 \cdot 8.93 \approx 72.42.$$

Therefore, the regression equation is given by

$$\hat{y} = 72.42 - 4.23 \cdot x.$$

b) Here are the usual steps of hypothesis testing:

Step 0: The test is about the slope of the regression equation β_1.

Step 1: H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$, significance level $\alpha = 0.05$.

Step 2: We are using the estimate b_1, its standard error s_{b_1} and the sample size $n = 20$.

Step 3: We are using the test statistic T_β that under the null hypothesis has a t-distribution with $n - 2 = 18$ degrees of freedom. The observed value of the test statistic is

$$t_\beta = \frac{b_1}{s_{b_1}} = \frac{-4.24}{1.20} \approx -3.533.$$

Since the test is two-tailed, the critical region is based on $t_{n-2,\alpha/2} = t_{18,0.025} = 2.101$ and contains the values to the left of -2.101 and to the right of 2.101.

Step 4: The observed value of the test statistic is in the critical region $-3.533 < -2.101$, so we reject the null hypothesis.

We have enough evidence to reject the claim that the slope is zero.

c) 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.

d) The outlier is the point corresponding to the size of incoming data of approximately 15 GB. If that point were removed, the sample linear correlation coefficient could get closer to -1, and would still indicate negative correlation.