

Final exam Statistical Methods – Exemplary solutions

20 December 2022

1.

- a) Because of the Central Limit Theorem, \hat{P}_{1600} has approximately a normal distribution with mean p , which is the true population proportion of Dutch citizens who feel that Sinterklaas is an important part of Dutch tradition, and variance $\frac{p(1-p)}{1600}$.
- b) For the 95%-CI we use the formula
$$[\hat{P}_{1600} - \sqrt{\hat{P}_{1600}(1 - \hat{P}_{1600})/1600} \cdot z_{0.025}, \hat{P}_{1600} + \sqrt{\hat{P}_{1600}(1 - \hat{P}_{1600})/1600} \cdot z_{0.025}],$$

that is, $[\frac{1090}{1600} - \sqrt{\frac{1090}{1600}(1 - \frac{1090}{1600})} \cdot 1.96/40, \frac{1090}{1600} + \sqrt{\frac{1090}{1600}(1 - \frac{1090}{1600})} \cdot 1.96/40]$
 $= [0.681 - 0.023, 0.681 + 0.023] = [0.658, 0.704].$
- c) We are going to use the formula $n \geq z_{\alpha/2}^2 / (4 \cdot E_{max}^2)$. In our case, $E_{max} = 0.04/2 = 0.02$. Thus,
 $n \geq 1.96^2 / (4 \cdot 0.0004) = 2401.$

2.

- a) The mean weight of three independent and randomly samples patients is a random variable with a normal distribution and mean $\mu = 112$ and variance $\sigma^2 = 10^2/3 = 100/3 \approx 33.3$.
Thus, $P(\bar{X}_3 > 130) = P(Z > \frac{130-112}{10/\sqrt{3}})$
where $Z \sim N(0, 1) = 1 - P(Z \leq 3.12) = 1 - 0.9991 = 0.0009.$

- b) **Exemplary choice and motivation:**

Two independent samples t -test assuming equal variances (or standard deviations in both groups),
Motivation: The relative difference between the standard deviations is about 16% – 19% (depending on the point of view), so they are quite close and it can be assumed that the standard deviations are equal.

(Other choice could also be motivated - but motivation is always key.)

We need for the test assumptions either sufficiently large samples (both ≥ 30) or that the individual measurements are normally distributed.

Either condition is satisfied here.

- c) Parameter of interest: $\mu_1 - \mu_2$,
Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 < 0$.

The choice of test statistic has been made and motivated in b), so we here only give the formula and state it's distribution under H_0 :

$$T_2^{eq} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2/35 + S_p^2/35}}$$

which under H_0 has a t -distribution with $35 + 35 - 2 = 68$ degrees of freedom.

t -score of the test:

We first have to calculate $s_p^2 = \frac{(35-1)s_1^2 + (35-1)s_2^2}{35+35-2} = (7.4^2 + 6.2^2)/2 = 46.6$

$$t_2^{eq} = \frac{158-163}{\sqrt{46.6/35+46.6/35}} \approx \frac{-5}{1.632} \approx -3.06$$

We compare this to the critical value $-t_{68,0.01} \approx -t_{60,0.01} = -2.390$.

The t -score is less than the critical value.

Because this concerns a left-tailed test, we reject the null hypothesis in favour of the alternative hypothesis. Concluding, we have shown (at significance level $\alpha = 1\%$) that metformin is effective in reducing the weight.

- d) This is an unpaired testing problem because the individuals in both groups are not related at all. (They even have been randomized into both treatment groups.)

As for designing the experiment differently, there are many options possible; below you can find a selection of possible answers – others might be acceptable as well.

Exemplary answer about designing the study differently:

Yes, I would design it differently because one could take measurements before and after the treatment of metformin and thus naturally end up in a 2 dependent samples problem.

(“No could also be argued if done correctly.”)

3.

- a) We use the formula

$$[\hat{P}_1 - \hat{P}_2 - z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}, \hat{P}_1 - \hat{P}_2 + z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}]$$

for $\alpha = 10\%$ which yields

$$\begin{aligned} & \left[\frac{30}{40} - \frac{24}{36} - 1.645 \sqrt{\frac{30/40 \cdot (1 - 30/40)}{40} + \frac{24/36 \cdot (1 - 24/36)}{36}}, \right. \\ & \quad \left. \frac{30}{40} - \frac{24}{36} + 1.645 \sqrt{\frac{30/40 \cdot (1 - 30/40)}{40} + \frac{24/36 \cdot (1 - 24/36)}{36}} \right] \\ &= \left[\frac{3}{4} - \frac{2}{3} - 1.645 \sqrt{\frac{3}{4} \cdot \frac{1}{4} / 40 + \frac{2}{3} \cdot \frac{1}{3} / 36}, \frac{3}{4} - \frac{2}{3} + 1.645 \sqrt{\frac{3}{4} \cdot \frac{1}{4} / 40 + \frac{2}{3} \cdot \frac{1}{3} / 36} \right] \\ &\approx \left[\frac{1}{12} - 1.645 \cdot 0.104, \frac{1}{12} + 1.645 \cdot 0.104 \right] \\ &\approx [-0.088, 0.255]. \end{aligned}$$

- b) Parameter of interest: $p_1 - p_2$,
significance level: $\alpha = 5\%$,
 $H_0 : p_1 = p_2$ vs. $H_a : p_1 > p_2$.

Test statistic:

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}},$$

where $\bar{P} = \frac{X_1 + X_2}{n_1 + n_2}$;

Z has under H_0 approximately an $N(0, 1)$ -distribution if certain requirements are met.

And indeed, the requirements are met: $x_1 = 30, n_1 - x_1 = 10, x_2 = 24, n_2 = 36 - 24 = 12$ are all bigger than 5.

$$\text{Test score: } z = \frac{\frac{3}{4} - \frac{2}{3}}{\sqrt{0.711 \cdot (1 - 0.711) / 40 + 0.711 \cdot (1 - 0.711) / 36}} \approx 0.8$$

since $\bar{p} = 54/76 \approx 0.711$.

We compare this to the critical value $z_\alpha = 1.645$, and we see that the score is smaller, so we cannot reject the null hypothesis based on this right-tailed test.

Concluding, there is not enough evidence to support the claim that CS students like programming better than AI students.

- c) Exemplary justification:

These two approaches in general yield the same conclusion:

The critical value method gives us a significant test result if the test score exceeds the 95%-quantile of $N(0, 1)$.

The p -value method gives us a significant outcome if the area to the right of the test score z w.r.t. $N(0, 1)$ is smaller than 5%. So this is exactly the case if the score exceeds the 95%.

4.

- a) This concerns a test of homogeneity because the data were sampled at two different times, and also because we are primarily interested in the equality/difference of proportions of active users.

Hypotheses:

$H_0 : p_{11} = p_{21} \text{ and } p_{12} = p_{22} \text{ vs. } H_a : p_{11} \neq p_{21} \text{ or } p_{12} \neq p_{22}.$

Alternatively, phrased in words:

H_0 : The “before take-over” population and the “after take-over” population are homogeneous with respect to their proportion of active/inactive users.

vs. H_a : The “before take-over” population and the “after take-over” population are heterogeneous with respect to their proportion of active/inactive users.

- b) The expected frequencies under the null are given as the product of the marginal totals over the grand total. Thus,

$$\begin{aligned} E_{11} &= \frac{N_{1.} \cdot N_{.1}}{n} = \frac{1063 \cdot 922}{1963} = 499.28, \\ E_{12} &= \frac{900 \cdot 922}{1963} = 422.72, \\ E_{21} &= \frac{1063 \cdot 1041}{1963} = 563.72, \\ E_{22} &= \frac{900 \cdot 1041}{1963} = 477.28. \end{aligned}$$

These expected numbers are all bigger than 5, so the requirements of the test are met.

- c) We use the test statistic

$$X^2 = \sum_{\text{cell } (i,j)} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

which has under H_0 approximately a chi-squared distribution with $(r-1) \cdot (c-1) = (2-1) \cdot (2-1) = 1$ degree of freedom.

(from b) we know that the requirements are met).

The test score is $\chi^2 = (526 - 499.28)^2/499.28 + (396 - 422.72)^2/422.72 + (537 - 563.72)^2/563.72 + (504 - 477.28)^2/477.28$
 $= 5.88$

We compare this to the critical value $\chi^2_{1,0.01} = 6.635$.

The score is smaller and the test is right-tailed which is why we cannot reject the null hypothesis. Concluding, there is not enough evidence to support the claim that the proportion of active Twitter users has changed after the take-over.

5.

- a) The explainable variation is given by $r^2 = 0.393^2 \approx 0.154$.

- b) Coefficient estimates:

$$b_1 = r \frac{s_y}{s_x} = 0.393 \cdot \frac{1.011}{196.544} \approx 0.002$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 7.673 - 3991.751 \cdot 0.002 \approx -0.311$$

- c) In general:

Residual plots can be used to check the fixed error variance assumption, i.e. whether or not the spread of the residuals is about constant for all values of x or whether there is some trend or some other strange pattern.

QQ-plots based on the residuals can be used to check the normality assumption for the error terms.

In our present data example:

No clear pattern or trend is visible in the residual plot, so we consider the fixed error variance assumption to be met.

Furthermore, the QQ-plot depicts points which are arranged almost perfectly on a straight line, which could be used as an indication that the residuals (and thus the error terms) follow a normal distribution.

- d) **This subquestion is granted full points for everybody because s_{b_1} was missing on the exam. It has been announced but some students already left at that time.**

Parameter of interest: β_1 , the slope parameter in the linear regression model.

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

Test statistic: $\frac{B_1}{S_{B_1}}$

which under H_0 (requirements are met, see c)) has approximately a t -distribution with $n - 2 = 38 - 2 = 36$ degrees of freedom.

t -score: $t = \frac{0.002}{0.00079} \approx 2.532$.

We want to compare this score to the critical values $\pm t_{36, 0.025\%} = \pm 2.028$.

So we see that the test score belongs to the critical region on the right, thus we can reject H_0 .

Consequently, we conclude that the Quiz scores do have an influence on the exam grades.