Resit exam Statistical Methods – Exemplary solutions

22 February 2023

1.

a) Define the events $A = \{$encountered sister replied with "yes" to the question: "Are you Ava?"$\}$ and $B = \{$Ava was truly encountered.$\}$. We are looking for $P(A)$. We know that $P(A|B) = 0.75$ and $P(A|\overline{B}) = 0.75$. Furthermore, $P(B) = P(\overline{B}) = 0.5$. By the law of total probability,
$P(A) = P(A|B) \cdot P(B) + P(A|\overline{B}) \cdot P(\overline{B}) = 0.75 \cdot 0.5 + 0.75 \cdot 0.5 = 0.75$.

b) We are looking for $P(B|A)$. By Bayes' theorem,

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\overline{B}) \cdot P(\overline{B})} = \frac{0.75 \cdot 0.5}{0.75 \cdot 0.5 + 0.75 \cdot 0.5} = \frac{1}{2}.$$

c) We are looking for $P(\overline{A} \cap \overline{B})$. This equals $P(\overline{A}|\overline{B}) \cdot P(\overline{B}) = (1 - P(A|\overline{B})) \cdot P(\overline{B}) = (1 - 0.75) \cdot 0.5 = 0.125$.

2.

a) The probability space consists of sample space plus probability measure.

Sample space: $\Omega = \{5, 10, 15, 20, 40\}$

Probability measure: $P(5) = 0.35 \cdot \frac{1}{2} = 0.175$, $P(10) = 0.5$, $P(15) = 0.3 \cdot \frac{1}{2} = 0.15$, $P(20) = 0.25 \cdot \frac{1}{2} = 0.125$, $P(40) = 0.1 \cdot \frac{1}{2} = 0.05$.

b) Let the random variable X model the random waiting time of a new customer.

$E(X) = \sum_{x \in \Omega} x \cdot P(x) = 5 \cdot 0.175 + 10 \cdot 0.5 + 15 \cdot 0.15 + 20 \cdot 0.125 + 40 \cdot 0.05 = 0.875 + 5 + 2.25 + 2.5 + 2 = 12.625$.

c) By the law of large numbers, because the sample size is large, $\overline{X}_{10000} \approx E(X) = 12.625$.

d) By the central limit theorem, because the sample size is large, $\overline{X}_{10000} \overset{\text{approx.}}{\sim} N(E(X), \sigma^2/10000)$, where $E(X) = 12.625$ and $\sigma^2/10000 \approx 0.00587$.

3.

a) For example:
Systematic sampling relates to sampling every kth theoretically available data point, if there is a systematic way of aligning them.
Convenience sampling relates to sampling only easily obtainable measurements.

b) A histogram is a plot of the (relative) frequencies of the data points in a sample. Therein, it is displayed how many data points fall into previously defined bins. It can illustrate numerical data of interval or ratio scale.

c) $P(X > -2.5) = P(Z > \frac{-2.5+3}{2}) = P(Z > 0.25) = 1 - P(Z \leq 0.25) = 1 - 0.5987 = 0.4013$. (here: $Z \sim N(0, 1)$.)

d) Chi-squared distributions assign probability mass only to intervals in the positive numbers whereas normal distributions assign probability mass to every real interval.

Chi-squared distributions are right-skewed whereas normal distributions are symmetric.

e) p-values are the probability that the test statistic exhibits under H_0 a more extreme value than the observed value (of the test statistic). They can be used to reach a test conclusion: reject H_0 if $p \leq \alpha$.

4.

a) Fisher's exact test, because the claim is directed.

b)

$$E_{i,j} = \frac{\text{(i-th row total)(j-th row total)}}{\text{grand total}}$$

expected frequencies:

		actual		
		milk	tea	
guessed	milk	6,5	6,5	
	tea	8,5	8,5	

c) All $E_{i,j} \geq 5$. Requirements are met.

d) $\alpha = 0.01$
Test statistic $X^2 = \sum_{(i,j)} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_1$ distributed under H_0.
Observed value: $X^2 = \frac{(10-6.5)^2}{6.5} + \frac{(3-6.5)^2}{6.5} + \frac{(5-8.5)^2}{8.5} + \frac{(12-8.5)^2}{8.5} \approx 6.6516$.
The critical value is $\chi^2_{1,0.01} = 6.635$.
$6.6516 > 6.635$ so H_0 is rejected.
There is enough evidence to confirm that the actual and guessed order are dependent.

---

5. Let $n_1 = 150, x_1 = 29, n_2 = 200, x_2 = 64$.

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

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2} \approx 0.193 - 0.320 = -0.127.$$

Here give if the answer is correct. Give if the answer is missing or incorrect, but it is clear that the student knows the estimate is the difference of sample proportions.

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 $\alpha = 0.01$.

Step 2: We are using x_1, x_2, n_1, n_2, but also $\overline{p} = (x_1 + x_2)/(n_1 + n_2) \approx 0.266$.

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

$$z_p = \frac{0.193 - 0.320}{\sqrt{(0.266 \cdot 0.734)/150 + (0.266 \cdot 0.734)/200}} \approx -2.66.$$

The test is two-tailed, therefore the critical regions are to the right of $z_{\alpha/2} = z_{0.005} \approx 2.575$ and to the left of -2.575.

Alternatively, using P-values:

Let Z be standard normally distributed. The test is two-tailed, and the observed value of the test statistic is negative, therefore the P-value is given by $2P(Z \leq -2.66) = 2 \cdot 0.0039 = 0.0078$.

Step 4: The observed value of the test statistic is negative and in the left critical region: $-2.66 < -2.575$, so we reject the null hypothesis.

Alternatively, using P-values:

The P-value 0.0078 is smaller than the significance level $\alpha = 1\%$, therefore we reject the null hypothesis.

We have enough evidence to reject the claim that Germans and Dutch people have the same proportions of people that know the brand NicePhone.

c) First, when testing proportions, the single observations are coded as 0/1 (i.e. nominal level of measurement) whereas tests for means require (at least) interval scaled levels of measurement.
Second, when testing proportions, the claimed value of proportion differences in the null hypothesis is always 0, whereas the claimed difference of expected values in tests for means could be different.
Third, when testing proportions, always the pooled proportion estimator should be used for estimating the standard deviation, whereas, when testing means, the pooled variance estimator can only be used if both samples have the same variances.
Fourth, when testing proportions, quantiles from the standard normal distribution are used, whereas, when testing means, usually the t-quantiles are used because of (normally) unknown variances.

6.

a) Using the data characteristics we first find

$$b_1 = r \cdot \frac{s_y}{s_x} = 0.392 \cdot \frac{11.700}{26.047} \approx 0.176,$$

and then

$$b_0 = \overline{y} - b_1 \cdot \overline{x} = 39.325 - 0.176 \cdot 136.701 \approx 15.266.$$

b) Here are the usual steps of hypothesis testing:

Step 0: The test is about the slope β_1.

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

Step 2: We use the estimate b_1 and its standard error s_{b_1} and $n = 30$.

Step 3: We use the test statistic T_β that under the null hypothesis has a t-distribution with $n - 2 = 28$ degrees of freedom. The observed value of the test statistic is

$$t_\beta = \frac{b_1}{s_{b_1}} = \frac{0.176}{0.078} \approx 2.256.$$

Since the test is two-tailed, the critical region is based on $t_{n-2,\alpha/2} = t_{28,0.025} = 2.048$ and contains the values to the left of -2.048 and to the right of 2.048.

Step 4: The observed value of the test statistic is in the critical region $2.048 < 2.256$, so we reject the null hypothesis.

We have sufficient evidence to reject the claim that the slope coefficient β_1 is zero.

c) A normal QQ plot is a plot of the ordered sample against the suitably chosen theoretical quantiles of the standard normal distribution. It may be used to conclude whether a normal distribution could be suitable to describe the data distribution. (Especially by checking the tails.)

d) The normal QQ plot of residuals is used to verify normality of the errors. In this case we see that the points follow a fairly straight line – therefore it is reasonable to assume that the errors come from a normal distribution.