Homework 7

Here we present solutions to the problems posted in the seventh homework assignment. The solutions and related commentary are put in italics. Remember that problems can have several different but correct ways of solving them.

Multiple choice questions

1. In statistics, categorical data

   A are analyzed using the same methods as numerical data;

   ***B** are analyzed by statistically examining relevant percentages or proportions;

   C are impossible to analyze in a coherent manner;

   D are often referred to as quantitative data as, in contrast, numerical data are referred to as qualitative data;

   E have to be transformed to numerical data and then analyzed by standard statistical methods for qualitative data.

   Solution: *Categorical data analysis is based on properly constructed contingency table that are designed to analyze percentages (or proportions) of data falling in various categories of interest.*

2. The difference between two sample percentages is statistically significant at 5% if

   ***A** it exceeds its two standard deviation;

   B it is 3% or less;

   C the percentages are based on a sample of at least 1,200;

   D the percentages are calculated to an accuracy of two significant figures;

   E there was a significant time lag between the two samples.

   Solution: *Under the assumption that the normal model applies, the chances that a difference of two percentages is between plus and minus two its standard deviation is 95% (even slightly more). Thus exceeding two standard deviation has about 5% chances. Therefore, in testing for significance of differences of two percentages, we consider it significantly different at level 5% if it exceed its two standard deviations in either direction.*

3. 2500 individuals randomly selected from the population were asked whether or not they would vote in the next general election. 50% said "yes". A 95% confidence interval for the percentage in the population who would have replied "yes" if asked is, approximately:

   A 45% to 55%;

   B 46% to 54%;

   C 47% to 53%;

   ***D** 48% to 52%;

E 49% to 51%.

Solution: *Estimate of standard deviation in sampling percentages is* $\sqrt{\hat{p}(1-\hat{p})/n}$. *Thus in our case it is* $\sqrt{0.5^2/2500} = 0.5/50 = 0.01$ *or* 1%. *According to the normal model, for the percentage being two standadrd deviation within its estimate occurs approximately 95% of time. Thus the 95% confidence interval is from 48% to 52%.*

4. Suppose the sample size in the previous question had been 10,000, and a confidence interval has been computed based on the same formula. Which one of the following is true?

   A the confidence level would be increased to 97.5%;

   B the confidence level would be increased to 98%;

   ***C** the confidence interval width would be halved;

   D the confidence interval width would be quartered;

   E there would be a 95% chance that the additional respondents would say "yes".

   Solution: *In our previous computation the value of standard deviation will be changed to* $\sqrt{0.5^2/10000} = 0.005$. *Thus standard deviation will be halved and the same is true to the confidence width.*

5. 36 out of 100 randomly selected individuals said that they preferred Brand X, when asked to say which of three brands they preferred. Thus, 36% is an estimate of the percentage in the population that prefer brand X, and its standard error is:

   A $36\% \pm 9.6\%$, that is, between 25% and 45% approximately;

   ***B** 4.8%;

   C $\pm 9.6\%$;

   D $36\% \pm 3\%$;

   E 0.6%.

   Solution: *According to previously used formula the standard deviation is* $\sqrt{0.36(1-0.36)/100} = 0.6 * 0.8/10 = 0.048$, *i.e. 4.8%.*

6. Suppose the sample size in the previous question had been 400, and 36% said they preferred Brand X. Which one of the following is true?

   A the standard error increases by a factor of 4;

   B the standard error increases by a factor of 2;

   C the standard error decreases by a factor of 4;

   ***D** the standard error decreases by a factor of 2;

   E none of the above is true.

   Solution: *The standard error (deviation) is proportional to* $1/\sqrt{n}$, *where n is sample size thus a four fold increase in n results in the two fold decrease in the standard error.*

7. In testing the homogeneity of several sample percentages patterns

A the null hypothesis is that none of the percentages are equal;

B a very large value of the chi-square test statistic suggests that the null hypothesis can not be rejected;

C the degrees of freedom associated with the chi-square test statistic is the sum of the number of rows and the number of columns in the contingency table involved in implementing the test by computer;

***D** the null hypothesis is that the row (or column) percentage patterns are the same;

E a very large value of the chi-square test statistic suggests that the wrong contingency table was used.

Solution: *In a test for homogeneity of the percentage patterns, we have two or more different populations that are examined for the same percentage patterns of a certain categorical variable that has two or more different levels.*

Problems

1. (a) A large banking organisation engaged a market research company to monitor the attitude of its staff towards proposals for changs in work practices. The market research company sampled 500 staff members at random and found 45% in favour of accepting the changes. Calculate a 95% confidence interval for the proportion of staff in favour.

   (b) Following six months of stalemate in negotiations, the bank commissioned another survey. This time, they required that the proportion in favour be estimated to within 2 percentage points, with 95% confidence. What sample size was required? (Consult lecture slides or Section 5.4, page 183-185.)

   (c) The bank agreed to a sample of 1000 for the second survey. 481 favourable responses were counted after the second survey. Calculate the 95% confidence interval for the proportion of those in favour. Compare with previously computed one and comment on it.

   (d) A month later, the staff union conducted a complete ballot of its member and found 44% in favour. How this complete count compares with previously conducted statistical analysis?

   Solution:

   (a) The estimated proportion is $\hat{p} = 0.45$*. Thus the estimated standard deviation of it is* $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.45 * 0.55/500} = 0.022$*. The confidence interval at 95% confidence level and based on the normal model is* $(0.45 - 1.96 * 0.022, 0.45 + 1.96 * 0.022) = (0.406, 0.493)$*, i.e.* $(40.6\%, 49.3\%)$*.*

   (b) The sample size can be determined by requiring that $1.96 * 0.5/\sqrt{n} \leq 0.02$*. Solving from this for* n*, we obtain that* $n \geq (1.96 * 0.5/0.02)^2 = 2401$*, so the sample of at least 2401 members of the staff is required.*

   (c) The confidence interval is given by $0.481 \pm 1.96 * \sqrt{0.481 * (1 - 0.481)/1000} = 0.481 \pm 0.031$*, i.e. the confidence interval is given by* $(45.0\%, 51.2\%)$*. We see that this time the confidence interval has shifted toward left although due to a larger sample size, it is now narrower.*

(d) *The value 44% is not in line what has been reported in the above study based on sample 1000. This value is not covered by 95% confidence interval. It appears that the survey overestimated the support, possibly by setting some pressure on staff.*

2. Of 294 urban residents selected for interview in a market research sample survey, 29% refused to participate. Of 1,015 rural residents selected for the same survey, 17% refused. Are urban and rural response rates different? Conduct a statistical test at significance level 5%. Report p-value. Formulate final conclusion.

   Solution: *We carry out a sample test for difference in two proportions. The test is based on test statistic*

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

   that has been discussed in the lecture (consult lecture slides) and also is given in the textbook, page 230. We reject the null hypothesis if the above statistic exceeds in absolute value the critical value of 1.96 that is taken from the table of normal distribution and corresponds to the 5% significance level of the test. The value of the test statistic is

$$(0.29 - 0.17)/\sqrt{0.29 * (1 - 0.29)/294 + 0.17 * (1 - 0.17)/1015} = 4.14$$

   which exceeds the critical value of 1.96. The p-value, i.e. the observed significance level is for all practical reasons equal to zero. There is a very strong statistical evidence that there is difference in willingness to respond between residents in rural areas and those in urban areas.

3. In a study of 400 trials relating to 'white collar' crime, the numbers jailed or not were recorded for each of three categories of crime, as follows:

	Crime			
Jail	E	F	Fy	Sum
N	57	146	25	228
Y	22	130	20	172
Sum	79	276	45	400

   where 'E' stands for 'Embezzlement', 'F' for 'Fraud', and 'Fy' for 'Forgery'.

   - Construct the percentage table normalized along columns (i.e. use percentages within each column with respect to their totals).
   - Carry a statistical test whether jailing pattern varied between the crime categories.
   - Formulate the conclusion of your test.

   Solution:

   - *The percentage table normalized along columns is obtained by expressing the percentage of each entry with respect the column total and is presented below:*

	Crime		
Jail	E	F	Fy
N	72	53	56
Y	28	47	44
Total	100	100	100

- *We will use the chi-square test of homogeneity of jailing patterns between the crime categories. The null hypothesis that we want to test is that the model proportions are in each column the same:*

$$(p_{11}, p_{21}) = (p_{12}, p_{22}) = (p_{13}, p_{23}),$$

where $p_{i,j}$ are the within cell column proportions. The general chi-square statistic for testing proportions or percentages based on the contingency tables has the form

$$\chi^2 = \sum_{i,j=1}^{2,3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

In the above O_{ij} represent counts in cells of contingency table, while E_{ij} represent the expected values of such counts in the model under the assumption that the null hypothesis is true. The expected counts E_{ij} can be either exactly computed from the assumptions on the model or maybe estimated from the data. Under our model, the exact values of E_{ij} are $O_{.j} * p_{ij}$, where $O_{.j}$ is the total in the jth column. We note that the total number of independent parameters is $2 * 3 - 3 = 3$. Under the null hypothesis, the estimated values E_{ij} can be obtained by using estmators $\hat{p}_{ij} = (O_{i1} + O_{i2} + O_{i3})/n = O_{i.}/n$ and taking estimated $E_{ij} = O_{.j} * \hat{p}_{ij} = O_{i.}O_{.j}/n$. We note that there is only one independent parameter that has to be estimated $\hat{p}_{1j} = 228/400 = 0.57$, the other being $\hat{p}_{2j} = 1 - \hat{p}_{1j}$. Here are the actual values of E_{ij} listed in the table below

	Crime			
Jail	E	F	Fy	Tot
N	45	157	26	228
Y	34	119	19	172
Sum	79	276	45	400

From that we evaluate the chi-square statistic as follows

$$\begin{aligned} \chi^2 = & \frac{(57-45)^2}{45} + \frac{(146-157)^2}{157} + \frac{(25-26)^2}{26} \\ & + \frac{(22-34)^2}{34} + \frac{(130-119)^2}{119} + \frac{(20-19)^2}{19} = \\ & = 9.31 \end{aligned} \tag{1}$$

We use the chi-square distribution with the number degrees of freedom equal to the total number of independent parameters minus the number of estimated parameters, i.e. 3-1=2. From the table of this distribution we get the critical value for the test at level 5% to be 6.0 and we note the evaluated value is 9.31.

- *In the conclusion, we make the claim that based on the collected data summarized in the contingency table we can observe a different jailing patterns for the three types of crime considered, i.e. there is no homogeneity in the jailing patterns.*