CATEGORICAL DATA ANALYSIS, 5 credits (intermediate studies), 3.9.–22.10.2015. Literature: Alan Agresti. An Introduction to Categorical Data Analysis, 2. edition. Lecturer: Pekka Pere.

Suggested solutions for the 3rd set of exercises

1.

a) The number of successes $Y \in 0, 1, \ldots, n$ follows a Binomial distribution $\mathsf{Bin}(n, \pi)$. Each observed value y yields a different confidence interval.

b) Coverage probability is the probability that a confidence interval, determined by y, covers the true π. $C_n(\pi)$ sums the probabilities of such ks that yield a confidence interval which include π. The probabilities are determined by the binomial formula $\sum_{k=0}^{n} \binom{n}{k} \pi^k (1-\pi)^{n-k}$.

c) Let us alter π slightly. If the same confidence intervals cover the altered π and the original π then $C_n(\pi)$ changes slightly because the associated sum of binomial probabilities changes smoothly with π. If a confidence interval which originally covered π does not cover the altered π then the coverage probability changes abrubtly as a binomial probability drops out of the sum in $C_n(\pi)$. A peak in the opposite direction results results if a confidence interval covers the altered but not the original π resulting in an extra binomial probability entering $C_n(\pi)$.

2.

a) The upper limit of a one-sided Clopper–Pearson confidence interval is defined by π for which

$$\sum_{k=0}^{y} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \alpha$$

(Agresti 2013, 603). Setting $y = 0$ yields

$$\binom{n}{0} \pi^0 (1-\pi)^n = (1-\pi)^n = \alpha.$$

Hence

$$\begin{aligned} 1 - \pi &= \alpha^{1/n} \qquad \Leftrightarrow \\ \pi &= 1 - \alpha^{1/n}. \end{aligned}$$

The one-sided $100 \times (1 - \alpha)$ % Clopper–Pearson confidence interval is

$$[0, 1 - \alpha^{1/n}].$$

b) The lower limit of a one-sided Clopper–Pearson confidence interval is defined by π for which

$$\sum_{k=y}^{n} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \alpha$$

(Agresti 2013, 603). Now $y = n$ or

$$\binom{n}{n} \pi^n (1-\pi)^{n-n} = \pi^n = \alpha.$$

Thus

$$\pi = \alpha^{1/n}.$$

The one-sided $100 \times (1-\alpha)$ % Clopper–Pearson confidence interval is

$$[\alpha^{1/n}, 1].$$

3. Let $n^* = 1/n$. The first order Maclaurin approximation for the upper limit $1 - \alpha^{n^*}$ of the one-sided 95 % Clopper–Pearson confidence interval is

$$\begin{aligned} 1 - \alpha^{n^*} &\approx (1 - \alpha^{n^*})\big|_{n^*=0} + \left(-\alpha^{n^*} \log \alpha\right)\big|_{n^*=0} \times n^* \\ &= 1 - 1 - \log \alpha \times n^* \\ &= -\frac{\log \alpha}{n}. \end{aligned}$$

By substituting $\alpha = 0.05$ one obtains the upper limit of the rule of three:

$$-\frac{\log 0.05}{n} \approx -\frac{-2.995732}{n} \approx \frac{3}{n}.$$

Extra comments: The approximation is good for $n > 30$ according to Wikipedia. Jovanovic (2005) suggests that $3/(n+1)$ is a better approximation.[1]

4.

a) The 95 % Wald confidence interval

$$\hat{\pi} \pm z_{0,025}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

is $[0, 0]$ if $\hat{\pi} = 0$. The given R code returns of course the same interval.

b) Command `prop.test(0,34, correct=FALSE)` gives 95 % score confidence interval $[0, 0.1015]$. The same interval to four decimal places is obtained with the code given in the exercise.

c) Formula $1 - 0.025^{1/34}$ yields 0.1028. The 95 % Clopper–Pearson confidence interval is $[0, 0.1028]$.

Command `binom.test(0,34)` and the script of Agresti at `http://www.stat.ufl.edu/~aa/cda/R/one-sample/R1/index.html` produce both exactly the same 95 % Clopper–Pearson confidence interval $[0, 0.1028]$.

d) Typing `midPci(0,34,0.05)` in R returns mid-*p* Clopper–Pearson confidence interval $[0, 0.0843]$.

In this special case the upper endpoint of the confidence interval could be solved analytically explicitly (from ”Background theory”):

$$\begin{aligned} \tfrac{1}{2}(1-\pi_0)^n &= \alpha/2 && \Leftrightarrow \\ (1-\pi_0)^n &= \alpha && \Leftrightarrow \\ 1-\pi_0 &= \alpha^{1/n} && \Leftrightarrow \\ \pi_0 &= 1 - \alpha^{1/n}. \end{aligned}$$

[1] `Https://en.wikipedia.org/wiki/Rule_of_three_\%28statistics\%29` (read 19.9.2015). B.D. Jovanovic (2005): Confidence Intervals, Binomial, When no Events are Observed. *Encyclopedia of Biostatistics*. Wiley.

Evaluating the right hand side of the last equation at $\alpha = 0.05$ and $n = 34$ yields 0.0843. It is the same figure as produced by the command `midPci(0,34,0.05)`.

e) The rule of three confidence interval is

$$[0, 3/34] = [0, 0.0882].$$

f) The intervals are collected here in descending order:

- $[0, 0.1028]$ (Clopper–Pearson)
- $[0, 0.1015]$ (score)
- $[0, 0.0882]$ (rule of three)
- $[0, 0.0843]$ (mid-*p* Clopper–Pearson)
- $[0, 0]$ (Wald).

The Clopper–Pearson interval is the widest. It tends to be too wide in general.

The score interval is almost as wide as the Clopper–Pearson interval. The score interval is typically narrower than it or even the mid-*p* version of it (Table 5.2 in Newcombe 2013 and Table 1 in Schilling and Doi 2014)[2]. Here the score interval is only slightly narrower than the Clopper–Pearson interval and is wider than the mid-*p* Clopper–Pearson interval. Possible explanations are the extremity of the present data and that the Wilson interval has poor coverage probability for π close to 0 or 1 (e.g. Agresti 2013, 33).

The rule of three interval lies close to the mid-*p* Clopper–Pearson interval but belongs in a sense to another league: It is the sole interval calculated as a one-sided interval. The Clopper–Pearson interval, say, is calculated by inverting two one-sided tests at the level $\alpha = 0.05/2 = 0.025$. The rule of three interval is based on a single one-sided test at level $\alpha = 0.05$.

The mid-*p* Clopper–Pearson interval is the shortest of the comparable and meaningful intervals. Indeed, its width is taken as the reference point in the simulations of Newcombe (*op. cit.*, Table 5.2).

The Wald interval is senseless reflecting the problems it faces near the boundaries of the parameter space $[0, 1]$.

[2] R.G. Newcombe (2013): *Confidence Intervals for Proportions and Related Measures of Effect Size*. CRC. Boca Raton, FL. M.F. Schilling and J.A. Doi (2014): A Coverage Probability Approach to Finding an Optimal Binomial Confidence Procedure. *The American Statistician*, 68, 133–145.

5.

a) Let $\mu_1 = n\pi_{10} = n\pi_0$, $\mu_2 = n\pi_{20} = n(1-\pi_0)$, and $\hat{\pi} = n_1/n$. It is proved below that $X^2 = z_s^2$:

$$
\begin{aligned}
X^2 &= \textstyle\sum_{i=1}^{2} \frac{(n_i - \mu_i)^2}{\mu_i} \\
&= \frac{(n_1 - n\pi_{10})^2}{n\pi_{10}} + \frac{(n_2 - n\pi_{20})^2}{n\pi_{20}} \\
&= \frac{(n_1 - n\pi_0)^2(1-\pi_0) + [n - n_1 - n(1-\pi_0)]^2\pi_0}{n\pi_0(1-\pi_0)} \\
&= \frac{(n_1 - n\pi_0)^2(1-\pi_0) + (n_1 - n\pi_0)^2\pi_0}{n\pi_0(1-\pi_0)} \\
&= \frac{(n_1 - n\pi_0)^2}{n\pi_0(1-\pi_0)} \\
&= \frac{(n_1/n - \pi_0)^2}{\pi_0(1-\pi_0)/n} \\
&= \frac{(\hat{\pi} - \pi_0)^2}{\pi_0(1-\pi_0)/n} \\
&= z_s^2.
\end{aligned}
$$

b) Test statistic z_s follows asymptotically the Standard Normal distribution so the square of it follows asymptotically χ^2 distribution with 1 degrees of freedom ($\chi^2(1)$). In the present circumstance $X^2 = z_s^2$. Hence $\chi^2(1)$ must be the asymptotic distribution of X^2.

c) The two test statistics and associated tests yield the same inferences in two-sided testing. χ^2 test is two-sided, but test statistics z_s can be used for one-sided testing as well. It matters which test statistic one chooses if one wants to carry out a one-sided test.

6.

a) First the squared difference in the numerator of the X^2 test statistic is calculated:

$$\begin{aligned} n_{11} - \hat{\mu}_{11} &= n_{11} - \frac{n_{1+}n_{+1}}{n} \\ &= \frac{n_{11}n - (n_{11}+n_{12})(n_{11}+n_{21})}{n} \\ &= \frac{n_{11}(n_{11}+n_{12}+n_{21}+n_{22}) - n_{11}^2 - n_{11}n_{21} - n_{11}n_{12} - n_{12}n_{21}}{n} \\ &= \frac{n_{11}n_{22} - n_{12}n_{21}}{n}. \end{aligned}$$

Accordingly

$$\begin{aligned} n_{12} - \hat{\mu}_{12} &= \frac{n_{12}n_{21} - n_{11}n_{22}}{n}, \\ n_{21} - \hat{\mu}_{21} &= \frac{n_{12}n_{21} - n_{11}n_{22}}{n} \quad \text{and} \\ n_{22} - \hat{\mu}_{22} &= \frac{n_{11}n_{22} - n_{12}n_{21}}{n}. \end{aligned}$$

It is the case that

$$(n_{ij} - \hat{\mu}_{ij})^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n^2}.$$

Indices *i* or *j* do not affect the value of $(n_{ij} - \hat{\mu}_{ij})^2$! Intuition: The estimated expected frequencies sum to the marginal frequencies (check it as well!). The departure of the observed from the estimated expected frequency $n_{ij} - \hat{\mu}_{ij}$ must cancel in the neighbouring cell in the 2×2 table. The squared differences thus equal.

b)

$$\begin{aligned} \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{\hat{\mu}_{ij}} &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{n}{n_{i+}n_{+j}} \\ &= n\left(\frac{1}{n_{1+}n_{+1}} + \frac{1}{n_{1+}n_{+2}} + \frac{1}{n_{2+}n_{+1}} + \frac{1}{n_{2+}n_{+2}}\right) \\ &= n\frac{n_{2+}n_{+2} + n_{2+}n_{+1} + n_{1+}n_{+2} + n_{1+}n_{+1}}{n_{1+}n_{2+}n_{+1}n_{+2}} \\ &= n\frac{n_{2+}(n_{+1}+n_{+2}) + n_{1+}(n_{+1}+n_{+2})}{n_{1+}n_{2+}n_{+1}n_{+2}} \\ &= n\frac{n_{2+}n + n_{1+}n}{n_{1+}n_{2+}n_{+1}n_{+2}} \\ &= \frac{n^2(n_{1+}+n_{2+})}{n_{1+}n_{2+}n_{+1}n_{+2}} \\ &= \frac{n^3}{n_{1+}n_{2+}n_{+1}n_{+2}}. \end{aligned}$$

c) The numerator is fixed from the point of view of the *i,j* indexing, so the test statistic X^2 can be calculated this way because of the points a) and b):

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \\
&\overset{\text{a)}}{=} \frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n^2}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{\hat{\mu}_{ij}} \\
&\overset{\text{b)}}{=} \frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n^2}\frac{n^3}{n_{1+}n_{2+}n_{+1}n_{+2}} \\
&= \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}.
\end{aligned}
$$

The requested formulation is obtained.[3]

[3] The proof is from J. Wu and S. Coggeshall (2012): *Foundations of Predictive Analytics*. CRC Press. Boca Raton, FL. (Page 23.)