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.

Examination 23.10.2015 – Solution sketches

1.

a) The associated rowwise probabilities are

		Y		
		y_1	y_2	Σ
X	x_1	π_1	$1-\pi_1$	1
	x_2	π_2	$1-\pi_2$	1

The odds measure how probable y_1 is compared to y_2 given $X = x_i$ ($i = 1, 2$). For $X = x_1$ the odds of y_1 relative to y_2 are $\pi_1/(1-\pi_1)$. Likewise for $X = x_2$ the odds are $\pi_2/(1-\pi_2)$. *E.g.* if $\pi_1/(1-\pi_1) = 3$ then y_1 is three times as probable as is y_2 given $X = x_1$.

The odds ratio θ is the probability of y_1 relative to the probability of y_2 if $X = x_1$ relative to the same ratio if $X = x_2$:

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

The odds ratio describes the relative difference of this ratio between the two classes of X. If $\theta = 1$ then there is no difference between the probabilities of y_1 happening in the two classes.

The odds ratio could be described in a similar manner using cell probabilities:

		Y		
		y_1	y_2	Σ
X	x_1	π_{11}	π_{12}	π_{1+}
	x_2	π_{21}	π_{22}	π_{2+}
	Σ	π_{+1}	π_{+2}	1

One or the other explanation suffices.

b) A homogeneous association applies if

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)}.$$

Here $\theta_{XY(k)}$, $k = 1, \ldots, K$, is the odds ratio between X and Y for the kth category of Z. In this case the odds ratio, and hence the odds of y_1, does not depend on Z.

Conditional independence of X and Y emerges if

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1.$$

In this circumstance probability of event y_1 does not depend on the value of X.

c) The logistic regression model is now of form

$$\log \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \alpha + \beta_1 x + \beta_2 z$$

where $\mathbf{x}' = [x\ z]$ and x and z are indicator variables with 2 and K categories, respectively. According to the model the odds ratio between X and Y is β_1 or the same regardless of the category of z. Hence a homogeneous association between X and Y applies. Conditional independence would mean that the odds ratio between X and Y equaled 1 in which case $\beta_1 = 0$. The model would then simplify to

$$\log \frac{\pi(z)}{1-\pi(z)} = \alpha + \beta_2 z.$$

2.

a) Two options for a confidence interval for a proportion are the ones based on the Wald and Rao test statistics

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \overset{as.}{\sim} \mathsf{N}(0,1)$$

and

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \overset{as.}{\sim} \mathsf{N}(0,1),$$

respectively. Here $\hat{\pi}$ is the MLE for π. Both statistics are derivatives of the likelihood function and follow asymptotically the Standard Normal distribution. The crucial difference is that in the former the variance of the MLE $\pi(1-\pi)/n$ is estimated by evaluating it at $\hat{\pi}$ but in the latter at the null value π_0.

The confidence intervals are derived by employing the duality between test statistics and confidence intervals: A $(1-\alpha) \times 100\%$ confidence interval is composed of such values of π which would not be rejected in (two-sided) testing at significance level α.

The Wald $(1-\alpha)\%$ confidence interval can be uncovered from the inequalities

$$-z_{\alpha/2} < \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} < z_{\alpha/2}$$

by manipulation:

$$\hat{\pi} - z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n} < \pi_0 < \hat{\pi} + z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}.$$

The confidence interval is

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}.$$

The confidence interval based on Rao's score test statistic is obtained correspondingly by squaring both sides of the equation(s) relating the value of the test statistic and the asymptotic critical values $\pm z_{\alpha/2}$:

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \pm z_{\alpha/2}.$$

The result is

$$\frac{(\hat{\pi}-\pi_0)^2}{-\pi_0(1-\pi_0)/n} = (\pm z_{\alpha/2})^2 \Leftrightarrow$$

$$(n+z_{\alpha/2}^2)\pi_0^2 - (2pn+z_{\alpha/2}^2)\pi_0 + np^2 = 0 \qquad || : n \Leftrightarrow$$

$$\left(1+\frac{z_{\alpha/2}^2}{n}\right)\pi_0^2 - \left(2\hat{\pi}+\frac{z_{\alpha/2}^2}{n}\right)\pi_0 + \hat{\pi}^2 = 0.$$

The squared equation is a polynomial of second degree in π_0. The roots of it or the bounds of the confidence interval can be solved in the usual way with the quadratic formula. (The solution is not required here.)

b) The score confidence interval is recommendable. The Wald confidence interval may perform poorly unless π lies close to 0.5 or n is large. "Typically" the Wald interval is too short leading to a lower coverage probability than intended.

The applicability of the Wald interval is limited to introductory teaching purposes and sample size determination according to Newcombe (1998)[1]:

Even though, for large n and mesial p [$= \hat{\pi}$] *(for example, for 81/263 in Table I), methods 1 and 2* [two versions of the Wald interval] *approximate acceptably to the better methods, it is strongly recommended that intervals calculated by these methods should no longer be acceptable for the scientific literature; highly tractable alternatives are available which perform much better. Use of the simple asymptotic standard error of a proportion should be restricted to sample size planning (for which it is appropriate in any case) and introductory teaching purposes.*

In a similar tone Schilling and Doi (2014)[2] put the simulation results for the Wald interval in parentheses to indicate that they are out of the league compared to other confidence intervals.

c) The sample proportion of religious Finns is

$$\frac{1516}{4930} \approx 0.3075051.$$

The Wald 99% confidence interval for this proportion is

$$0.3075051 \pm 2.576\sqrt{\frac{0.3075051 \times (1-0.3075051)}{4930}}$$

or about

$$(0.2906, 0.3244).$$

Above 2.576 is the 99.5 percentile of the Standard Normal distribution.

[1] R.G. Newcombe (1998): Two-sided confidence intervals for the single proportion: Comparison of Seven Methods, *Statistics in Medicine*, 17, 857–872.

[2] M.F. Schilling and J.A. Doi (2014): A Coverage Probability Approach to Finding an Optimal Binomial Confidence Procedure, *American Statistician*, 68, 133–145. Table I.

Alternatively the 99% score (or Wilson) interval, say, could be calculated:

$$\begin{aligned}
& \hat{\pi}\frac{n}{n+z_{\alpha/2}^2}+\frac{1}{2}\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\pm\frac{z_{\alpha/2}}{n+z_{\alpha/2}^2}\sqrt{\hat{\pi}(1-\hat{\pi})n+\frac{1}{4}z_{\alpha/2}^2} \\
\approx\ & 0.3075051\times\frac{4930}{4930+2.576^2}+\frac{1}{2}\frac{2.576^2}{4930+2.576^2} \\
& \pm\frac{2.576}{4930+2.576^2}\sqrt{0.3075051(1-0.3075051)4930+\frac{1}{4}\times 2.576^2}.
\end{aligned}$$

The score confidence interval is

$$(0.2908, 0.3247).$$

Extra comments: The confidence intervals almost match. The obvious explanation is the large sample size.

d) A correctly calculated reasonable single one or two-sided confidence interval is an appropriate answer.

The rule of three is the simplest solution (a one-sided confidence interval). The upper bound of this confidence interval is

$$\frac{3}{300}=\frac{1}{100}=0.010.$$

The rule-of-three confidence interval is [0,0.010].[3]

Alternatively, the (two-sided) Wilson

$$\hat{\pi}\frac{n}{n+z_{\alpha/2}^2}+\frac{1}{2}\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\pm\frac{z_{\alpha/2}}{n+z_{\alpha/2}^2}\sqrt{\hat{\pi}(1-\hat{\pi})n+\frac{1}{4}z_{\alpha/2}^2},$$

Clopper–Pearson

$$1-(\alpha/2)^{1/300},$$

or mid-*p* Clopper–Pearson

$$1-\alpha^{1/300}$$

intervals could be calculated ($\alpha=0.05$). (The mid-*p* Clopper–Pearson formula follows from solving equation $\frac{1}{2}\binom{n}{0}\pi_0{}^0(1-\pi_0)^{n-0}=\alpha/2$.) They are, $[0, 0.013]$, $[0, 0.012]$, and $[0, 0.010]$, respectively.

(The unreasonable Wald interval $[0, 0]$ is not an acceptable answer.)

[3]This example is from Wikipedia (`https://en.wikipedia.org/wiki/Rule_of_three_(statistics)`; read 23.10.2015.

3.

a) Using the notation in the table below, the null hypothesis of marginal homogeneity would be $\pi_{+1} = \pi_{1+}$. In the context of the exercise it would mean that landlords were as likely to reply to a heterosexual as a homosexual couple.

		Y		
		y_1	y_2	Σ
X	x_1	π_{11}	π_{12}	π_{1+}
	x_2	π_{21}	π_{22}	π_{2+}
	Σ	π_{+1}	π_{+2}	1

b) Notation of the table

		Y		
		y_1	y_2	Σ
X	x_1	n_{11}	n_{12}	n_{1+}
	x_2	n_{21}	n_{22}	n_{2+}
	Σ	n_{+1}	n_{+2}	n

is used below. The test statistic is

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{4 - 50}{\sqrt{4 + 50}} \approx -6.260.$$

It follows the Standard Normal distribution in large samples if the null hypothesis is valid. The 2.5th percentile of the Standard Normal is -1.960. The null hypothesis is rejected at the 5% level because $-6.260 < -1.960$. Homosexual couples are less likely to receive answers to their contacts.

Extra comments: R command `2*pnorm(-6.260)` returns *p*-value 0.0000000009 of the test statistic.

c) McNemar's test is suitable because the data is composed of matched pairs (two observations per landlord).

Under marginal homogeneity it should hold that $n_{+1}/n \approx n_{1+}/n$ or $n_{+1} \approx n_{1+}$. Then

$$n_{11} + n_{12} \approx n_{11} + n_{21}$$

or

$$n_{12} \approx n_{21}.$$

Let $n^* = n_{12} + n_{21}$. Under the null hypothesis the observations in cells "12" and "21" are binomially distributed with equal probability 0.5. The independence assumption underlying the Binomial distribution is admissible because each observation in these cells is associated with a different landlord. The mean and variance of the frequencies ("N_{12}" and "N_{21}") are obtained from the Binomial distribution:

$$0.5 \times n^* = \frac{n_{12} + n_{21}}{2}$$

and

$$0.5 \times 0.5 \times n^* = \frac{n_{12} + n_{21}}{4}.$$

McNemar's test statistic

$$\frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{(n_{12} - n_{21})/2}{\sqrt{(n_{12} + n_{21})/4}} = \frac{n_{12} - (n_{12} + n_{21})/2}{\sqrt{(n_{12} + n_{21})/4}}$$

follows the Standard Normal distribution when the sample size is large enough.

If n_{12} differs greatly from n_{21} then the absolute value of the test statistic inflates and the null hypothesis is rejected by the test. In this case also n_{1+} and n_{+1} differ considerably. The data is then in conflict with the null hypothesis of marginal homogeneity.

Extra comments: The data could be arranged with a χ^2 test in mind:

	yes	no	Σ
answered the heterosexual couple	227	181	408
answered the homosexual couple	181	227	408

Also this table suggests that the homosexual couple has received fewer responses than the heterosexual couple. A χ^2 test is not applicable, though, because the observations are not independent: A response from each landlord is on both rows of the table. (It is a coincidence that the cell frequencies appear mirrorwise.)

4. a)–d) Pages 70–71, 99–101, and 104–105 of Agresti (2007)[4] and Exercise 6.3.

[4] A. Agresti (2007): *An Introduction to Categorical Data Analysis, 2. edition*. Wiley. Hoboken, NJ.