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 2nd set of exercises

1. The quantities in the formulae are

$$l'(\pi) = \frac{y - n\pi}{\pi(1-\pi)}$$

and

$$i(\pi) = \frac{n}{\pi(1-\pi)}.$$

The Wald statistic is thus

$$\sqrt{i(\pi)}\Big|_{\pi=\hat{\pi}}(\hat{\pi}-\pi_0) = \sqrt{\frac{n}{\pi(1-\pi)}}\Bigg|_{\pi=\hat{\pi}}(\hat{\pi}-\pi_0) = \sqrt{\frac{n}{\hat{\pi}(1-\hat{\pi})}}(\hat{\pi}-\pi_0) = \frac{\hat{\pi}-\pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}.$$

Correspondingly, Rao's score statistic is

$$\frac{l'(\pi)}{\sqrt{i(\pi)}}\Bigg|_{\pi=\pi_0} = \frac{\dfrac{y-n\pi}{\pi(1-\pi)}}{\sqrt{\dfrac{n}{\pi(1-\pi)}}}\Bigg|_{\pi=\pi_0} = \frac{\dfrac{y-n\pi_0}{\pi_0(1-\pi_0)}}{\sqrt{\dfrac{n}{\pi_0(1-\pi_0)}}} = \frac{\dfrac{y/n-\pi_0}{\pi_0(1-\pi_0)/n}}{\sqrt{\dfrac{n}{\pi_0(1-\pi_0)}}} = \frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

2. The variance of the MLE $P = \sum_{i=1}^{n} Y_i/n$ is $\pi(1-\pi)/n$. The number of observations n is fixed. The variation in the variance originate hence from the product $\pi(1-\pi)$. The maximum of it can be found by differentiation:

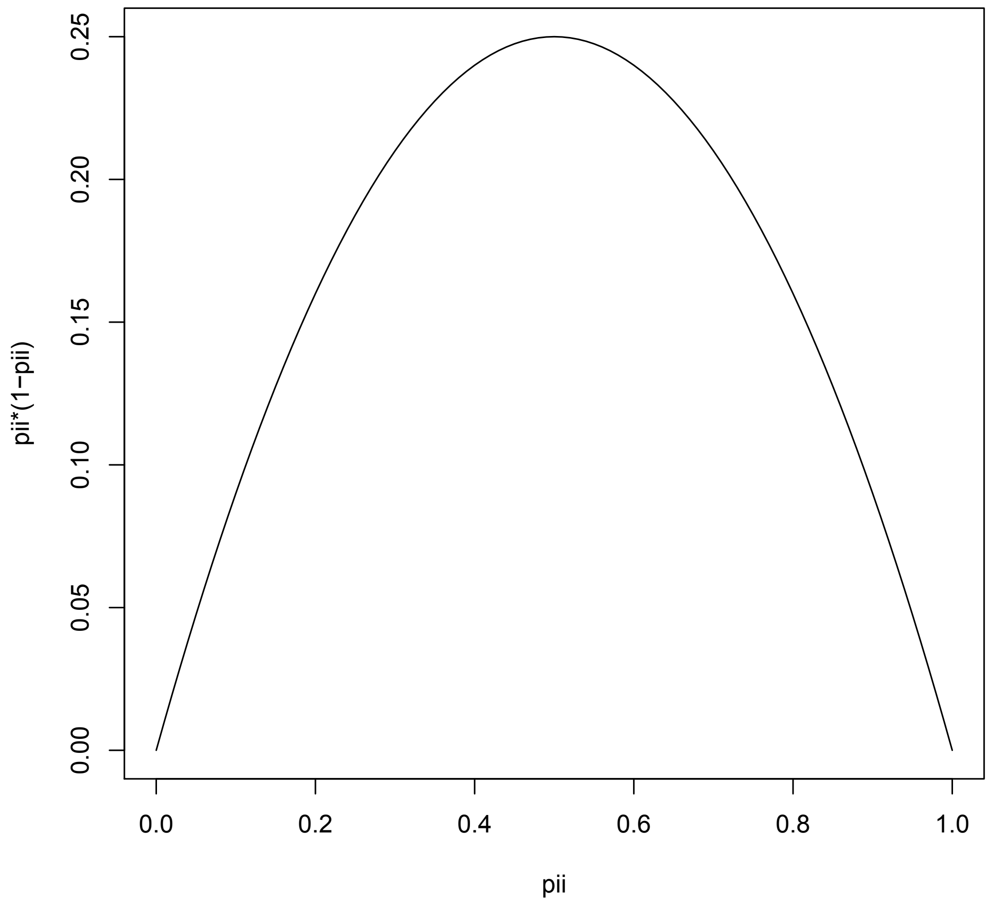
$$\frac{\partial}{\partial\pi}\pi(1-\pi) = 1 - 2\pi = 0 \Leftrightarrow \pi = 0,5.$$

This is a maximum because $\partial^2\, \pi(1-\pi)/\partial\pi^2 = -2 < 0$. The second derivative is a negative constant or does not depend on π. The variance is a concave function, and the variance of the MLE is minimized at the edges of the parameter space $[0,1]$ at $\pi = 0$ and $\pi = 1$.

The variance of the MLE thus peaks at $\pi = 0,5$ and is minimized at $\pi \approx 0$ or $\pi \approx 1$. This explains why π is difficult to estimate accurately if $\pi \approx 0,5$.

The graph below illuminates. It is produced with R software with the commands

```
curve(x*(1-x), 0, 1, xlab="pii", ylab="pii*(1-pii)")
```



Extra comments: The distribution of the MLE $\hat{\pi}$ is skewed in small samples if $\pi \approx 0$ or $\pi \approx 1$. (An intuitive explanation: The distibution of the MLE kind of cuts at the edge of the parameter space.) It complicates statistical inference because the asymptotic Normal approximation does not apply. The true coverage probability of confidence intervals, say, is not then what is intended (95 %, say). Despite that accurate point estimation is possible if $\pi \approx 0$ or $\pi \approx 1$, accurate interval estimation may not be.

3.

a)

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

$$(n+z^2_{\alpha/2})\pi_0^2 - (2\hat{\pi}n + z^2_{\alpha/2})\pi_0 + n\hat{\pi}^2 = 0 \qquad || : n \Leftrightarrow$$

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

b) Let

$$a = 1 + \frac{z^2_{\alpha/2}}{n} > 0,$$

$$b = -(2\hat{\pi} + \frac{z^2_{\alpha/2}}{n})$$

and

$$c = \hat{\pi}^2.$$

The formula in a) can now be expressed as

$$ax^2 + bx + c = 0,$$

where $a \neq 0$. The roots of it are (the formula for the roots of a quadratic equation):

$$\frac{-b \pm \sqrt{b^2-4ac}}{2a}.$$

In the present circumstance the roots are

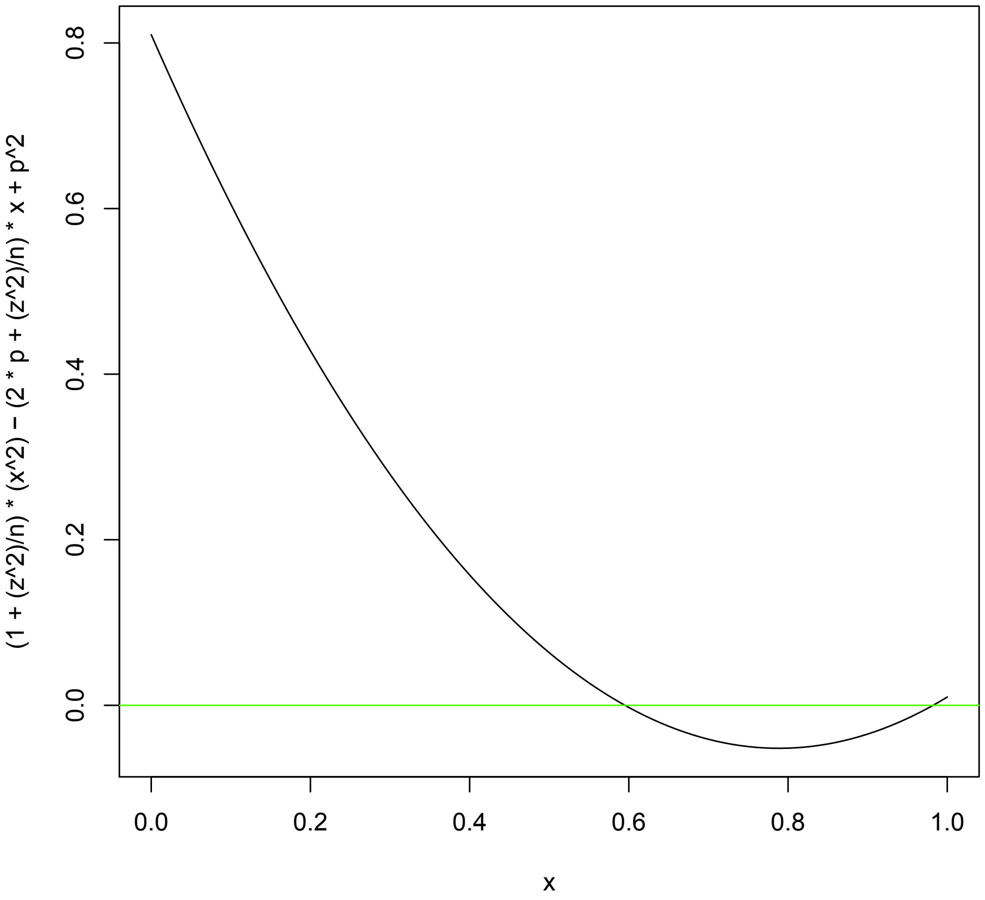
$$\frac{(2\hat{\pi}+\frac{z^2_{\alpha/2}}{n}) \pm \sqrt{(2\hat{\pi}+\frac{z^2_{\alpha/2}}{n})^2 - 4(1+\frac{z^2_{\alpha/2}}{n})\hat{\pi}^2}}{2(1+\frac{z^2_{\alpha/2}}{n})}.$$

Expanding and arranging terms suitably yields the requested expressions for the roots:

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

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

$$\hat{\pi}\frac{n}{n+z^2_{\alpha/2}} + \frac{1}{2}\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}} \pm z_{\alpha/2}\sqrt{\frac{1}{n+z^2_{\alpha/2}}\left[\hat{\pi}(1-\hat{\pi})\frac{n}{n+z^2_{\alpha/2}} + \frac{1}{2}\frac{1}{2}\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right]}.$$



The second last expression is due to the equality $y = n\hat{\pi}$.

Extra comments: Wilson (1927) was the first to explore the score confidence interval.[1] Agresti and Coull (1998) suggest that the above method of calculating a confidence interval would be called Wilson's method.[2] They propose (*op. cit.*, p. 120) that it could be used for all sample sizes and values of π. Newcombe's (1998) evaluation is that it it

[1] E.B. Wilson (1927): Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22, 209–212.

[2] A. Agresti ja B.A. Coull (1998): Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *American Statistician*, 52, 119–126.

the sole method which is easily calculated and works.[3]

c) The formula below for the roots gives the center of the confidence interval as the weighted mean of $\hat{\pi}$ and $1/2$:

$$\hat{\pi}\frac{n}{n+z_{\alpha/2}^2}+\frac{1}{2}\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}.$$

It lies (almost) always closer to $1/2$ than the center or MLE $\hat{\pi}$ of the Wald confidence interval.(If $\hat{\pi}=1/2$ then the centers are the same.) The center of the score interval approaches $\hat{\pi}$ as n tends to infinity. MLE $\hat{\pi}$ is consistent so the center converges to π. The lower and upper limits of the score confidence interval determined by the terms

$$\pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}^2}\left[\hat{\pi}(1-\hat{\pi})\frac{n}{n+z_{\alpha/2}^2}+\frac{1}{2}\frac{1}{2}\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right]}$$

converge toward zero as n tends toward infinity. The lower and upper limits of the score confidence interval approach then $\hat{\pi}$ and π.

4. The R code of the exercise yields exactly confidence interval $(0{,}596; 0{,}982)$. The confidence interval has been calculated correctly in the book.

5.

a) Altogether 4 observations are added to the data when the plus four method is used: 2 observations are added to the successes and 2 to the failures. After these revisions the proportion of successes is

$$p^*=\frac{y+2}{n+4}.$$

The center of the plus four confidence interval is p^*.

The center of the score confidence interval is

$$\left.\frac{y+z_{\alpha/2}^2/2}{n+z_{\alpha/2}^2}\right|_{z_{\alpha/2}^2=1{,}960}\approx\left.\frac{y+z_{\alpha/2}^2/2}{n+z_{\alpha/2}^2}\right|_{z_{\alpha/2}^2=2}=\frac{y+2^2/2}{n+2^2}=\frac{y+2}{n+4}=p^*.$$

The centers almost match.

b) The width of the plus four confidence interval is

$$2z_{0{,}025}\sqrt{p^*(1-p^*)/n^*}\overset{z_{0{,}025}\approx 2}{\approx}4\sqrt{p^*(1-p^*)/(n+4)}.$$

[3] R.G. Newcombe (1998): Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine*, 17, 857–872.

The width of the 95 % score confidence interval is

$$2z_{0,025}\sqrt{\frac{1}{n+z_{0,025}^2}\left[p(1-p)\frac{n}{n+z_{0,025}^2}+\frac{1}{2}\frac{1}{2}\frac{z_{0,025}^2}{n+z_{0,025}^2}\right]}$$
$$\overset{z_{0,025}^2\approx 2}{\approx} 4\sqrt{\frac{1}{n+4}\left[p(1-p)\frac{n}{n+4}+\frac{1}{2}\frac{1}{2}\frac{4}{n+4}\right]}.$$

The widths can be compared with the help of expressions $p^*(1-p^*)$ and

$$p(1-p)\frac{n}{n+4}+\frac{1}{2}\frac{1}{2}\frac{4}{n+4}.$$

Idea: Let random variate P take value p with probability $n/(n+4)$ and value $1/2$ with probability $4/(n+4)$. (The probabilities sum to 1.) The expected value of P is

$$\mathsf{E}(P)=p\frac{n}{n+4}+\frac{1}{2}\frac{4}{n+4}=\frac{np+2}{n+4}=\frac{y+2}{n+4}=p^*.$$

Let $g(P)=P(1-P)$. It is a concave function ($\partial^2 P(1-P)/\partial^2 P=-2<0$.) By Jensen's inequality $g[\mathsf{E}(P)]\geq \mathsf{E}[g(P)]$ for a concave $g(P)$. It must hence be the case that

$$\begin{aligned} g[\mathsf{E}(P)] &= g[\mathsf{E}(P)]\{1-[\mathsf{E}(P)]\} \\ &= p^*(1-p^*) \\ &\geq \mathsf{E}[g(P)] \\ &= p(1-p)\frac{n}{n+4}+\frac{1}{2}(1-\frac{1}{2})\frac{4}{n+4} \end{aligned}$$

or

$$p^*(1-p^*)\geq p(1-p)\frac{n}{n+4}+\frac{1}{2}\frac{1}{2}\frac{4}{n+4}.$$

Plus four confidence interval is wider than the 95 % score confidence interval.

Extra comments:

- The latter confidence interval contains the former because the centers of them (almost) match (part a)). Plus four confidence interval is in general conservative or too wide (Agresti and Coull 1998). An obvious reason is the widerness of it compared to the score confidence interval.
- The arguments above are based on the approximation $z_{0,025}\approx 1,960\approx 2$.
- The center of the plus four confidence interval can be taken as an estimator of π. Agresti and Coull (*op. cit.*) name it the Wilson estimator.
- Agresti and Coull (*op. cit.*) think that the plus four confidence interval is ”dramatically” better than the Wald confidence interval.

- The plus four confidence interval works reasonably well already when $n = 20$ and π is not close to 0 or 1. If it is then the interval is too wide.

c) The plus four approximation does not apply for a 99 % confidence interval. The approximation is based on the use of the 97.5th percentile of the Standard Normal distribution in the approximation $z_{\alpha/2} \approx 2$. The approximation does not apply for other percentiles.

The 99.5th percentile of the Standar Normal distribution is 2,576. If the number of successes and failures were both increased by $2,576^2/2 \approx 3,318$ and the number of observations by $2,576^2 \approx 6,636$ then the confidence interval

$$p^* \pm z_{0,005}\sqrt{p^*(1-p^*)/n^*}$$

could be used to approximate the 99 % score confidence interval.

Extra comments: In general the score confidence interval can be approximated by calculating a $1-\alpha$ level confidence interval with the formula

$$p^* \pm z_{\alpha/2}\sqrt{p^*(1-p^*)/n^*},$$

where the data is modified by adding $z^2_{\alpha/2}/2$ observations to the successes and failures of the original data. The new number of observations is then $n^* = n + z^2_{\alpha/2}$. A confidence interval composed this way is called an Agresti-Coull confidence interval. A special case of it is the plus four confidence interval.

The exercise is essentially exercise 1.25 from Alan Agresti's book (2013) *Categorical Data Analysis, 3rd edition*, CUP.

6. The confidence intervals, widths and center points are (from p. 10)

- $(0,714; 1,086)$, $0,372$ and $0,900$ (Wald)
- $(0,596; 0,982)$, $0,386$, $(0.596 + 0.982)/2 \approx 0.789$ and (score)
- $0,786 \pm 1,960 \times 0.110$ or $(0,570; 1,002)$, $0,432$ and $0,786$ (plus four).

It can be seen that

- the centers of the score and plus four confidence intervals almost match.
- the plus four confidence interval is wider and encompasses the score confidence interval.
- the Wald confidence interval is the shortest.

The results are in accordance with the derived general theoretical results and the statement in the book that the Wald confidence interval (for a proportion) tends to be too narrow.

7.

a) The two one-sided mid p-values are

$$\frac{1}{2}\mathsf{P}(T=t_i)+\mathsf{P}(T>t_i)$$

and

$$\frac{1}{2}\mathsf{P}(T=t_i)+\mathsf{P}(T<t_i).$$

They sum to 1:

$$\begin{aligned} & \frac{1}{2}\mathsf{P}(T=t_i)+\mathsf{P}(T>t_i)+\frac{1}{2}\mathsf{P}(T=t_i)+\mathsf{P}(T<t_i) \\ = \quad & \mathsf{P}(T<t_i)+\mathsf{P}(T=t_i)+\mathsf{P}(T>t_i) \\ = \quad & 1. \end{aligned}$$

b) The corresponding p-values are

$$\mathsf{P}(T=t_i)+\mathsf{P}(T>t_i)$$

and

$$\mathsf{P}(T=t_i)+\mathsf{P}(T<t_i).$$

The sum of them is larger than 1:

$$\begin{aligned} & \mathsf{P}(T=t_i)+\mathsf{P}(T>t_i)+\mathsf{P}(T=t_i)+\mathsf{P}(T<t_i) \\ = \quad & \mathsf{P}(T<t_i)+\mathsf{P}(T=t_i)+\mathsf{P}(T>t_i)+\mathsf{P}(T=t_i) \\ = \quad & 1+\mathsf{P}(T=t_i) \\ = \quad & 1+\pi_i \\ > \quad & 1. \end{aligned}$$

Probability $\mathsf{P}(T=t_i)=\pi_i$ is counted twice in the sum.

c) A one-sided mid p-value is $\frac{1}{2}\mathsf{P}(T=t_i)+\mathsf{P}(T>t_i)=\pi_i/2+\pi_{i+1}+\cdots+\pi_I$. The expexted value of it is $1/2$:

$$\begin{aligned} & \textstyle\sum_{i=1}^{I}\pi_i(\pi_i/2+\pi_{i+1}+\cdots+\pi_I) \\ = \quad & \textstyle\sum_{i=1}^{I}\frac{\pi_i^2}{2}+\sum_{i=1}^{I}\pi_i(\pi_{i+1}+\cdots+\pi_I) \\ = \quad & \frac{1}{2}\textstyle\sum_{i=1}^{I}\pi_i^2+\sum_{i=1}^{I}\pi_i\sum_{j=i+1}^{I}\pi_j \\ = \quad & \frac{1}{2}\textstyle\sum_{i=1}^{I}\pi_i^2+\sum_{i=1}^{I}\sum_{j=i+1}^{I}\pi_i\pi_j \\ = \quad & \frac{1}{2}(\textstyle\sum_{i=1}^{I}\pi_i)^2 \\ = \quad & \frac{1}{2}. \end{aligned}$$

The fourth equality follows from the formula $(\sum_{i=1}^{I} \pi_i)^2 = \sum_{i=1}^{I} \pi_i^2 + 2\sum_{i=1}^{I} \sum_{j=i+1}^{I} \pi_i \pi_j$. The fifth equality is obtained by noting that $\sum_{i=1}^{I} \pi_i = 1$.

d) The expected value of the p-value is larger than $1/2$ because of the previous point:

$$\begin{aligned} & \textstyle\sum_{i=1}^{I} \pi_i(\pi_i + \pi_{i+1} + \cdots + \pi_I) \\ > & \textstyle\sum_{i=1}^{I} \pi_i(\pi_i/2 + \pi_{i+1} + \cdots + \pi_I) \\ \overset{c)}{=} & \frac{1}{2}. \end{aligned}$$

Point c) is exercise 1.27 in Alan Agresti's book (2013) *Categorical Data Analysis, 3rd edition*, CUP.