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.

## 2nd exercise set (18.9.)

1. Let the null hypothesis be that the parameter of the Binomial distribution is $\pi = \pi_0$. Denote the sample size by $n$ and the MLE of $\pi$ by $\hat{\pi}$. Show that the Wald and Rao's score test statistics are

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

and

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

respectively.

2. Parameter $\pi$ of the Binomial distribution is easier to (point) estimate (by the method of ML, say) accurately if $\pi$ is close to 0 or 1 compared to the case of $\pi$ lying close to 0,5. Explain why. (Hint: Consider the variance of the MLE $\hat{\pi} = \sum_{i=1}^{n} Y_i/n$.)

3. A way to derive a confidence interval for a proportion ($\pi$) is to derive it by squaring both sides of the equation relating Rao's score test statistic and its asymptotic critical values ($\pm z_{\alpha/2}$):

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

(p. 10 of the book). Above $\hat{\pi} = \sum_{i=1}^{n} y_i/n = y/n$ is the observed proportion, $\pi_0$ is the hypothetised value of $\pi$, $n$ is the sample size, $\alpha$ is the size of the test (e.g. $0,05$), and $z_{\alpha/2}$ is the $100 \times (1-\alpha/2)$. (the $97,5$th, say) percentile of the Standard normal distribution. The roots of the equation are the lower and upper limits of the confidence interval.

a) Show that the squared equation can be written equivalently as

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

b) Show that the roots of the equation above are

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

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

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

c) What happens to the center of the confidence interval as $n$ tends to infinity? What happens to the roots or the lower and upper bounds of the confidence interval as $n$ tends to infinity? (Hint: Examine the last formula for the roots.)

4. Let $n = 10$ and $p = 0.9$. The 95 % score confidence interval for $\pi$ is then $(0.596; 0.982)$ (p. 10 of the book). Check that the interval is correctly calculated, or calculate the roots in this case. You may want to employ the R code below in your checkup.

```
n <- 10
p <- 0.9
z <- 1.960
Wl <- (p*n+(z^2)/2)/(n+z^2)-(z/(n+z^2))*sqrt(p*(1-p)*n+0.25*(z^2))
Wu <- (p*n+(z^2)/2)/(n+z^2)+(z/(n+z^2))*sqrt(p*(1-p)*n+0.25*(z^2))
Wl
Wu
```

5. Let the confidence coefficient be 0,95. The score confidence interval can then be approximated by the following method (p. 10 of the book): Add 2 observations both to the successes and failures, and calculate the Wald confidence interval

$$p^* \pm z_{0,025}\sqrt{p^*(1-p^*)/n^*}$$

from the modified data. Above $p^*$ is the proportion of successes in the modified data and $n^* = n + 4$. Such a confidence interval is called an Agresti–Coull confidence interval in Agresti's 2007-book. Agresti calls it the plus four confidence interval in his 2013-book (p. 33) and refers to a related adjusted confidence interval as an Agresti–Coull confidence interval. The plus four confidence interval is a special case of this more generally applicable confidence interval.

Let us explore why the plus four confidence interval mimics the 95 % score confidence interval. (Hint: $z_{0,025} \approx 1,960 \approx 2$.)

a) Let $y$ be the number of successes in the original data. Prove that the center $p^*$ of the plus four confidence interval is

$$\frac{y+2}{n+4}.$$

Compare it with the center of the 95 % score confidence interval

$$\frac{y + z^2_{\alpha/2}/2}{n + z^2_{\alpha/2}}$$

(exercise 3).

b) Prove that the plus four confidence interval is wider than the 95 % score confidence interval (at least when the comparison is based on the approximation $z_{0,025} \approx 1,960 \approx 2$). (Hints: The longest expression for the lower and upper bounds of the 95 % score confidence interval in exercise 3. Jensen's inequality: Let $g(P)$ be a concave function. Then $g[\mathsf{E}(P)] \geq \mathsf{E}[g(P)]$. Set $g(P) = P(1-P)$. Note that $n/(n + z^2_{\alpha/2}) +$

$z^2_{\alpha/2}/(n + z^2_{\alpha/2}) = 1$. Compose a random variable $P$ associated with the previous weights so that it can take two values $p$ or 1/2 only. The comparison can be carried out between $p^*(1-p^*)$:n and the square bracket term (an expected value) in the aforementioned longest expression formulation.)

c) Is the plus four approximation valid for 99 % confidence intervals as well? If not, can you fabricate a corresponding adjustment to approximate a 99 % score confidence interval?

6. A toy example of 10 independent Bernoulli experiments with 9 successes is considered on p. 10 of the book. Agresti reports the Wald and score confidence intervals and the plus four confidence interval approximating the latter. Make a table of the confidence intervals, calculate the midpoints and widths of them, and compare your results to the previous theoretical results and to the statement in the book that the Wald confidence interval for a proportion is in general too wide (p. 9). Are your numerical results in accordance with these theoretical results?

7. Let statistic $T$ follow a discrete distribution $\mathsf{P}(T = t_i) = \pi_i > 0, i = 1, 2, \ldots, I$.

a) Calculate both the left-tail and right-tail $p$-values for the realized value $t_i$ of the statistic. Prove that the corresponding mid $p$-values sum to 1.

b) Prove that the sum of the corresponding $p$-values is larger than 1.

c) Prove that the expected value of a one-sided mid $p$-value is 0.5. (Hint — which you should justify: $\sum_{i=1}^{I} \pi_i(\pi_i/2 + \pi_{i+1} + \cdots + \pi_I) = (\sum_{i=1}^{I} \pi_i)^2/2$.)

d) Prove that the expected value of a one-sided $p$-value is larger than 0.5.