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.

6th exercise set (16.10.)

1. Agresti (2007, 108) writes:

The sample proportion estimate – – is $p = 4/6 = 0.67$ *– –. From converting small-sample tests using the binomial distribution, a 95 % confidence interval based on these six observations equals (0.22,0.96).*

What kind of (Wald, Wilson, Clopper–Pearson or mid-p Clopper–Pearson) confidence interval has Agresti calculated? (Hint: Try which one of the R commands in Exercise 3.4 returns the interval.)

2. Anneli Auer has been accused of murdering her husband 2006. She has been twice convicted from this murder and twice the conviction has been overturned. She has also been sentenced from sexual abuse of her children. Auer has plead to the European Court of Human Rights because of this verdict. The latest news (8.10.2015) is that careful analysis with modern software of a recording of the telephone call of Auer at about the time of death of her husband suggests that she had partly recorded the call beforehand or that the call is a fabrication of hers. The Supreme Court will ponder the claim.

Helsingin Sanomat's monthly supplement Kuukausiliite (3/2015, 17) pointed out that women judges have been more inclined to find Auer guilty than men judges (table). The caption of the article was "Nainen naiselle susi" or "A Woman is a wolf towards another woman". Is the argument statistically sound?

a) Calculate the p-value for Fisher's exact test statistic for the null hypothesis of an odds ratio equal to unity against the alternative hypothesis that the odds ratio is larger than unity.

b) As point a), but calculate the mid p-value.

c) What do you infer?

		verdict		
		not guilty	guilty	Σ
gender of the judge	female	1	3	4
	male	6	2	8
	Σ	7	5	12

(Hint: Compute the probabilities by yourself and check your results with R commands `dhyper(0:1, 4, 8, 7, log = FALSE)` and

```
x <- as.matrix(c(1,6,3,2))
dim(x) <- c(2,2)
x
fisher.test(x, alternative="less")
```

3. Logistic regression model

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x$$

($\beta \neq 0$) is explored. It holds that

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

a) Prove that the slope of the curve $\pi(x)$ is $\beta\pi(x)[1 - \pi(x)]$ (pp. 100–101 of the book).

b) Prove that the slope of the curve $\pi(x)$ is at its steepest when $\pi(x) = 0,50$.

c) Prove that the slope of the curve $\pi(x)$ is at its steepest when $x = -\alpha/\beta$.

4–5. Wikipedia[1]: *The Space Shuttle Challenger disaster occurred on January 28, 1986, when the NASA Space Shuttle orbiter Challenger – – broke apart 73 seconds into its flight, leading to the deaths of its seven crew members – – . – – Disintegration of the vehicle began after an O-ring seal in its right solid rocket booster (SRB) failed at liftoff. – – Approximately 17 percent of Americans witnessed the launch live because of the presence of Payload Specialist Christa McAuliffe, who would have been the first teacher in space. – – The temperature on the day of the launch was far lower than had been the case with previous launches: below freezing at 28 to 29 °F (-2.2 to -1.7 °C); previously, the coldest launch had been at 53 °F (12 °C).*

Connection between thermal distress of the primary O-rings (ring shaped seals) and temperature at the time of launch of the shuttle can be modeled using logistic regression. The response variable takes value 1 if at least one of the primary O-rings suffers from thermal distress and 0 if none do.[2]

The logistic regression, estimated by the method of maximum likelihood, is

$$\begin{array}{rcccc} \log \dfrac{\hat{\pi}(c)}{1-\hat{\pi}(c)} & = & 7{,}614 & - & 0{,}418c. \\ & & (3{,}933) & & (0{,}195) \end{array}$$

Above c is temperature, $\hat{\pi}(c)$ is the probability for thermal distress, potentially associated with temperature, and approximative standard errors of the estimates are in parentheses. The approximative covariance matrix of the constant and the regression coefficient, as reported by the R software, is below.

```
             (Intercept)           C
(Intercept)  15.4718002 -0.7587027
C            -0.7587027  0.0379570
```

[1] `Https://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster` (read 11.10.2015).

[2] The data and exercise are to a great extent are from the book A. Agresti (2013): *Categorical Data Analysis, 3. edition*. Wiley. Hoboken, New Jersey. (Exercise 5.6.) Fahrenheit units have been converted to degrees Celcius. The disaster is considered in more detail in Exercise 2.4 of the book C.R. Bilder and T.M. Loughin (2015): *Analysis of Categorical Data with R.* CRC Press, Boca Raton, Florida. Both exercises are based on the article S. Dalal, E. Fowlkes, and B. Hoadley (1989): Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. *Journal of the American Statistical Association*, 84, 945–957.

The fit of the model is depicted in the figure. The computations and the figure have been produced with R. The observations have been jittered for construction of the graph only (without jittering many of the observations would lie exactly on top of each other).

a) Conduct a two-sided test at risk level 5 % on whether temperature is associated with thermal distress or is not. Explain carefully each step of your test. How does the covariance matrix above relate to the test statistic you have calculated?

b) Interpret the estimated coefficient of the explanatory variable (the effect of temperature on thermal distress).

c) The average temperature has been around 20.9 °C at the time of launch of the shuttle. What is the estimated probability of thermal distress at this temperature? If the temperature changes slightly from the average, how much does the estimated probability change? Interpret the figure.

d) Calculate a 95 % confidence interval for the probability of thermal distress at the average temperature 20.9 °C.