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 6th set of exercises

1.

R command `binom.test(4,6)` computes the 95 % Clopper–Pearson confidence interval and returns (0.2227781,0.9567281). This is the interval Agresti has computed.

Extra comments. For assurance, Wald, Wilson and mid-p Clopper–Pearson confidence intervals were calculated with the R codes in Exercise 3.4 (in descending order of width):

- $[0.289, 1.044]$ (Wald)
- $[0.223, 0.957]$ (Clopper–Pearson)
- $[0.262, 0.940]$ (mid-p Clopper–Pearson)
- $[0.300, 0.903]$ (Wilson or score).

Notes:

- The Wald interval is the widest. Being wider than the other intervals is "atypical" for a Wald interval (Agresti 2007, 9, cf. Exercise 3.4) but not a one-of-a-kind incident (Agresti and Coull 1998)[1].
- The Wald interval overshoots (the upper limit is larger than unity). The Wald interval can be unsensible even when the estimated probability (here 0.67) does not lie close to the edges of the parameter space [0,1] if the sample size is small.
- Mid-p Clopper–Pearson interval is not always the narrowest sensible interval.

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

2.

a) What is seeked for are the probability of the observed table

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

and of the more extreme table

		verdict		
		not guilty	guilty	Σ
gender of the judge	female	0	4	4
	male	7	1	8
	Σ	7	5	12

The probabilities are the hypergeometric ones

$$\frac{\binom{4}{1}\binom{8}{6}}{\binom{12}{7}} \approx 0.1414 \quad \text{and} \quad \frac{\binom{4}{0}\binom{8}{7}}{\binom{12}{7}} \approx 0.0101.$$

The one-sided *p*-value is $0.1414 + 0.0101 \approx 0.152$.

Check: R command `dhyper(0:1, 4, 8, 7, log = FALSE)` returns `0.01010101 0.14141414` or the probabilities above. R code

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

confirms the *p*-value above.

b) The mid *p*-value is 0.081:

$$0.5 \times 0.14141414 + 0.01010101 \approx 0.081.$$

c) The *p*-value 0.152 or the more meaningful mid *p*-value 0.081 are not small enough for rejection of the null hypothesis of an odds ratio of unity at conventional risk levels. The data, though askew, is not too unlikely under the null hypothesis. The claim in the article is not statistically sound.

Extra comments. The same inference results from the transposed tables

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

and

		gender of the judge		
		female	male	Σ
verdict	female	0	7	7
	male	4	1	5
	Σ	4	8	12

The probabilities are the same

$$\frac{\binom{7}{1}\binom{5}{3}}{\binom{12}{4}} \approx 0.1414 \quad \text{and} \quad \frac{\binom{7}{0}\binom{5}{4}}{\binom{12}{4}} \approx 0.0101,$$

as can be checked with the line `dhyper(0:1, 7, 5, 4, log = FALSE)`. Also

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

returns p equal to 0.152.

In general, the hypergeometric probability can be computed from the original table or the transpose of it. A few lines of algebra reveals that the probability equals

$$\frac{n_{1+}n_{2+}n_{+1}n_{+2}}{nn_{11}n_{12}n_{21}n_{22}}$$

(using well-established notation by now) or is the same for both tables (original or transposed).

3.

a) The slope of the curve

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

is

$$\begin{aligned}
\frac{\partial \pi(x)}{\partial x} &= \frac{[\exp(\alpha + \beta x)]\beta[1 + \exp(\alpha + \beta x)] - [\exp(\alpha + \beta x)]\beta\exp(\alpha + \beta x)}{[1 + \exp(\alpha + \beta x)]^2} \\
&= \beta\left[\frac{\exp(\alpha + \beta x)]}{1 + \exp(\alpha + \beta x)} - \frac{[\exp(\alpha + \beta x)]^2}{[1 + \exp(\alpha + \beta x)]^2}\right] \\
&= \beta\{\pi(x) - [\pi(x)]^2\} \\
&= \beta\{\pi(x)[1 - \pi(x)]\}.
\end{aligned}$$

b) The slope of the curve $\pi(x)$ or $\beta\{\pi(x)[1 - \pi(x)]\} = \beta\pi(x) - \beta[\pi(x)]^2$ ($\beta \neq 0$) is a second degree polynomial in $\pi(x)$. It opens downward if $\beta > 0$ or upward if $\beta < 0$. A second degree polynomial is twice differentiable so that the global maximum or minimum is located at the value at which the first derivative equals zero or at an extremum value of the argument. In the present circumstance the slope tends to zero as $\pi(x)$ tends to 0 or 1. The vertex is hence located at the root of the first derivative of the slope ($\beta \neq 0$):

$$\frac{\partial\{\beta\pi(x)[1 - \pi(x)]\}}{\partial\pi(x)} = \beta[1 - \pi(x) - \pi(x)] = \beta[1 - 2\pi(x)] = 0 \Leftrightarrow$$

$$1 - 2\pi(x) = 0 \Leftrightarrow \pi(x) = 1/2.$$

The slope peaks in absolute value or the curve $\pi(x)$ is at its steepest when $\pi(x) = 1/2$.

c) The slope peaks in absolute value if $\pi(x) = 1/2$. Substituting it to the formula of the logistic regression model and solving for x yields

$$\begin{aligned}
\log\frac{1/2}{1 - 1/2} &= \alpha + \beta x \quad \Leftrightarrow \\
0 &= \alpha + \beta x \quad \Leftrightarrow \\
x &= -\frac{\alpha}{\beta}.
\end{aligned}$$

The maximum absolute value of the slope takes place at $x = -\alpha/\beta$.

The proof can be carried out mirrorwise as well:

$$\begin{aligned}
\pi(-\alpha/\beta) &= \frac{\exp[\alpha + \beta(-\alpha/\beta)]}{1 + \exp[\alpha + \beta(-\alpha/\beta)]} \\
&= \frac{\exp(0)}{1 + \exp(0)} \\
&= \frac{1}{2}.
\end{aligned}$$

This is the value of $\pi(x)$ at which the curve is at its steepest according to point b).

4–5.

a) The Wald test statistic is the estimated coefficient of temperature devided by the estimated standard error of it:

$$\frac{-0{,}418}{0{,}1948} \approx -2{,}145.$$

(The calculations are carried out using more decimal places than given in the exercise.) The test statistic follows asymptotically a Standard Normal distribution. The 2.5th percentile of it is -1.960. The observed value of the test statistic is smaller($-2.145 < 1.960$) so the null hypothesis is rejected at the 5 %:n risk level. Temperature affects the probability of thermal distress.

The diagonal values of the covariance matrix are approximative variances (valid for large samples) of the estimators. For example the estimated variance of the estimated coefficient of temperature is the (2,2) element of the matrix or 0.0379570. The square root of it, 0.1948256, is the standard error of the estimated coefficient of temperature reported in the exercise. It is the denominator of the Wald test statistic.

Extra comments: The *p*-value of the test statistic is about 0.032 (calculated with the R command `2*pnorm(-2.145)`).

b) In the case of a logistic regression with a single explanatory variable

$$\begin{aligned} \log \frac{\pi(c)}{1-\pi(c)} &= \alpha + \beta c \quad \Leftrightarrow \\ \frac{\pi(c)}{1-\pi(c)} &= e^{\alpha+\beta c}. \end{aligned}$$

According to the latter equation the odds of thermal distress are multiplied by $\exp(\beta)$ if the explanatory variable c is changed by a unit:

$$\begin{aligned} \frac{\pi(c+1)}{1-\pi(c+1)} &= e^{\alpha+\beta(c+1)} \\ &= e^{\alpha+\beta c} e^{\beta}. \end{aligned}$$

In the case under study, the MLE of β is –0.418 so the odds are estimated to vary by $\exp(-0.418) \approx 0.66$ (the odds decrease).

c) The formula

$$\pi(c) = \frac{e^{\alpha+\beta c}}{1+e^{\alpha+\beta c}}$$

determines probability $\pi(c)$. Substitution of the MLEs of the parameters and $c = 20.9$ to the formula gives

$$\frac{e^{7.614-0.418\times 20.9}}{1+e^{7.614-0.418\times 20.9}} \approx 0.246.$$

Probability of thermal distress at the average temperature of 20.9 is estimated to be about 0.25.

A small change in temperature alters the probability by $\beta\pi(c)[1-\pi(c)]$. In the present case the probability changes in the opposite direction by about 0.08 at the average temperature 20.9:

$$-0.418 \times 0.246(1-0.246) \approx -0.077.$$

The dependency of the probability of thermal distress on temperature is illuminated in the figure. The logistic transformation ensures that the estimated probability lies in (0,1). The estimated coefficient of temperature is negative so the probability of thermal distress increases when the temperature decreases. The probability is almost 0.93 at the coldest observed temperature 12°C).

d) Approximative 95 % confidence interval for the logit or $\alpha + \beta c$ at c degrees is

$$\hat{\alpha} + \hat{\beta}c \pm 1.960 \times \mathsf{SE}.$$

Here SE is the square root of the variance

$$\mathsf{V}(\hat{\alpha} + \hat{\beta}c) = \mathsf{V}(\hat{\alpha}) + c^2\mathsf{V}\hat{\beta}) + 2c\mathsf{Cov}(\hat{\alpha}, \hat{\beta})$$

of the sum $\hat{\alpha} + \hat{\beta}c$. Above $\mathsf{Cov}(\hat{\alpha}, \hat{\beta})$ is the covariance of the estimators. The estimates of the variances and the covariance were stated in the assignment. Substitution of them into the formula above reveals the estimated variance and SE:

$$15.4718002 + (20.9)^2 \times 0.0379570 + 2 \times 20.9 \times (-0.7587027) \approx 0.338.$$

and

$$\mathsf{SE} \approx \sqrt{0.338} \approx 0{,}584.$$

Approximative 95 % confidence interval for the logit transformation at 20.9°C) is

$$7.614 - 0.418 \times 20.9 \pm 1.960 \times 0{,}584 \approx \left\{ \begin{array}{l} -2.259952 \\ 0.01913154. \end{array} \right.$$

Logistic function

$$\frac{e^c}{1+e^c} = \frac{1}{1+e^{-c}}$$

is monotonically increasing. Approximative 95 % confidence interval for the probability of thermal distress is hence obtained by evaluating the probability function

$$\pi(c) = \frac{e^{\alpha+\beta c}}{1+e^{\alpha+\beta c}}$$

at the lower and upper limits of $(\alpha + \beta c)$:

$$\left\{ \begin{array}{l} \dfrac{e^{-2.259952}}{1+e^{-2.259952}} \approx 0.094 \\ \dfrac{e^{0.01913154}}{1+e^{0.01913154}} \approx 0.505. \end{array} \right.$$

The requested confidence interval is (0.094, 0.505).

Extra comments.

- Statistical models should not be employed outside the observed range of variables. The probability is extrapolated into dataless zones in the figure. Such extrapolated evaluations are to be treated with caution. The extrapolated probability is essentially 1 at 0°C. On the other hand, the probability tends to 0 as the temperature increases without limit. This aspect of the model is not to be taken literally either.

- The problems with the O-rings had been pointed out beforehand by engineer Roger Boisjoly. He tried to cancel the flight on 28th of January 1986 because of the cool temperature. His superiors ignored his warnings. Later his colleagues and superiors started shunning him and he resigned. Boisjoly was 1988 awarded the Award for Scientific Freedom and Responsibility by the American Association for the Advancement of Science.[2]

- Many books have been written about the disaster.

- Related accidents still take place:

  - USA Today 29.10.2014[3]: *Unmanned Antares rocket explodes on launch. An unmanned commercial rocket headed for the International Space Station to deliver supplies exploded just after launching Tuesday, filling the sky with a massive fireball. The Antares rocket supplied by contractor Orbital Sciences blew up moments after liftoff at NASA's space launch facility on the Eastern Shore of Virginia – – .*

  - The Guardian 31.10.2014[4]: *One pilot dead as Virgin Galactic's SpaceShipTwo rocket plane crashes. Virgin Galactic says plane designed for commercial space travel was undertaking test flight in California when "in-flight anomaly" occurred. – – The California Highway Patrol confirmed that one person was dead and another had suffered "major injuries" in the accident – – .*

  - The New York Times 28.6.2015[5]: *An unmanned cargo ship destined for the International Space Station disintegrated minutes after being launched from the Cape Canaveral Air Force Station in Florida on Sunday morning, NASA said, raising questions about how the agency and its partners will continue keeping the station supplied. It was the third loss of a cargo ship headed to the space station in the past eight months.*

---

[2] `Https://en.wikipedia.org/wiki/Roger_Boisjoly` (read 13.10.2015).

[3] `Http://www.usatoday.com/story/tech/2014/10/28/nasa-rocket-explodes-wallops-island/18080871/` (read 13.10,.2015).

[4] `Http://www.theguardian.com/science/2014/oct/31/spaceshiptwo-richard-branson-virgin-crash-mojave` (read 13.10.2015)

[5] `Http://www.nytimes.com/2015/06/29/science/space/spacex-rocket-explodes-du\textbf{âĂć}ring-launch.html` (read 13.10.2015).