Exercise Sheet 6

Exercise 6.1 - Maximum Likelihood Estimation, Fisher Information
You are given an i.i.d. sample $X = (X_1, \dots, X_n)$ of continuous random variables X_i with the following density:

$$f(x_i;\theta) = \frac{1}{\sqrt{2\pi}}\theta^{-\frac{3}{2}}\, x_i^2 \exp\left(-\frac{x_i^2}{2\,\theta}\right), \quad \theta > 0, \quad i = 1, \dots, n. \tag{1}$$

Further, $\mathbb{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 3\theta$ hold.

(a) Determine the maximum likelihood estimator for θ.
Note: You don't have to show that the MLE is indeed a maximum.

$$\begin{aligned}
f(x;\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\
&= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}}\theta^{-\frac{3}{2}}\, x_i^2 \exp\left(-\frac{x_i^2}{2\,\theta}\right)\right) \\
&= (2\pi)^{-n/2}\theta^{-3n/2} \prod_{i=1}^{n} \left(x_i^2 \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\theta}\right)\right) = L(\theta;x) \\
l(\theta;x) &= -\frac{n}{2}\log 2\pi - \frac{3n}{2}\log\theta + 2\sum_{i=1}^{n}\log x_i - \frac{\sum_{i=1}^n x_i^2}{2\theta} \\
s(\theta;x) &= -\frac{3n}{2\theta} + \frac{\sum_{i=1}^n x_i^2}{2\theta^2} \\
0 &\overset{!}{=} -\frac{3n}{2\theta} + \frac{\sum_{i=1}^n x_i^2}{2\theta^2} \\
3n &= \frac{\sum_{i=1}^n x_i^2}{\theta} \\
\Rightarrow \qquad \hat{\theta}_{ML} &= \frac{\sum_{i=1}^n x_i^2}{3n}
\end{aligned}$$

(b) The provided distribution fulfils the criteria for Fisher regularity. Show that the expected Fisher information is $I_X(\theta) = \dfrac{1.5\,n}{\theta^2}$.

$$\begin{aligned}
I_X(\theta) &= -\mathbb{E}\left(\frac{\partial}{\partial\theta} s(\theta;X)\right) \\
&= -\mathbb{E}\left(\frac{3n}{2\theta^2} - \frac{\sum_{i=1}^n X_i^2}{\theta^3}\right) \\
&= -\frac{3n}{2\theta^2} + \frac{n}{\theta^3}\underbrace{\mathbb{E}(X_1^2)}_{Var(X_1)=3\theta} \\
&= \frac{3n}{2\theta^2}
\end{aligned}$$

(c) Determine the approximate distribution of the following statistic:

$$T(X) = \frac{\sum_{i=1}^{n} X_i^2}{2\theta^2} - \frac{3n}{2\theta}.$$

The statistic $T(X) = \frac{\sum_{i=1}^{n} X_i^2}{2\theta^2} - \frac{3n}{2\theta}$ is exactly the score function and thus

$$T(X) \xrightarrow{D} N(0, I_X(\theta)) = N\left(0, \frac{3n}{2\theta^2}\right).$$

(d) A standard approach to the construction of the $(1-\alpha)$ confidence interval for a parameter is the Wald interval. In case of the MLE, we make use of its asymptotic variance so that the interval looks like this:

$$[\widehat{\theta}_{ML} \pm z_{1-\alpha/2}\sqrt{I_X^{-1}(\widehat{\theta}_{ML}))}]$$

Calculate an approximate $(1-\alpha)$ confidence interval ($\alpha = 0.05$) for θ using the following sample:

$$x = (-2, 1, 0, -0.5, 2, 1, 0, 0.5, 2, -1, 0, 0.5, 2, 1, 0, -0.5).$$

$n = 16, \quad \sum_{i=1}^{16} x_i^2 = 4(2^2 + 1^2 + 0^2 + 0.5^2) = 4 \times 5.25 = 21$

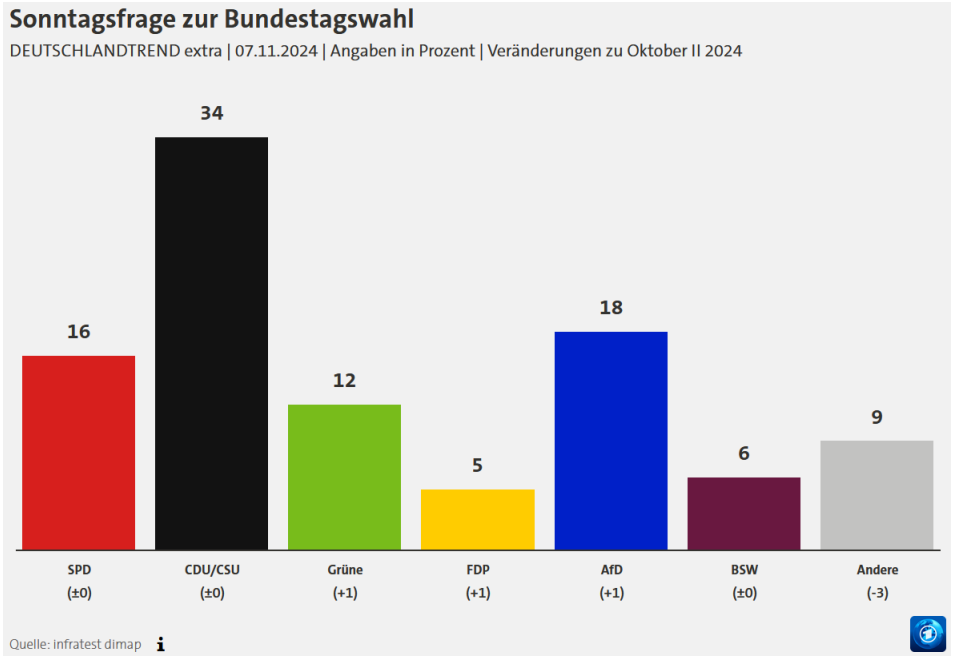
$$\hat{\theta}_{ML} = \frac{21}{3 \cdot 16} = 0.437 \qquad I_X(\hat{\theta}_{ML}) = \frac{3 \cdot 16}{2\hat{\theta}_{ML}^2} = 125.3878 \Rightarrow \sqrt{\frac{1}{I_X(\hat{\theta}_{ML})}} \approx 0.0893$$

$z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution, so $\Phi^{-}1(0.975) = 1.96$. With this, we have all components to calculate the Wald-CI for θ:

$$[0.4375 \pm 1.96 \times 0.0893] \approx [0.2624, 0.6125]$$

Exercise 6.2 - Probability and Confidence Intervals

The "Sonntagsfrage" in the ARD Deutschlandtrend in November gave the following results:



Source: infratest dimap

The numbers above the bars in the chart indicate the estimated percentage of votes a party would receive if federal elections were held next Sunday. The number in brackets below each bar indicate the change in comparison to the results the week before. The following information is attached to the results:

- Population: Eligible voters in Germany
- Data collection method: Randomized telephone and online survey
- Data collection period: 7. November 2024
- Sample size: 1.065 respondents (Mixture of telephone and internet)
- Margin of error: 2 percentage points at a proportion of 10 percent, 3 percentage points at a proportion of 50 percent

(a) Which potential issues would you expect to have to deal with when conducting such a survey?

- Making sure the samples are i.i.d. (e.g. non-independent answers when asking people from the same household or asking the same person twice).
- Collecting a sample that's representative of the entire population (Careful: The entire population is not everybody living in Germany, but everybody eligible to vote) Problems: Many people no longer have a regular telephone, some don't have access to the internet, some don't have either. The voting preferences in these groups might not be identical to the entire population → Results might have to be weighted by known frequency in actual population.
- Social desirability of answers (Voters might not acknowledge their real voting intentions, especially when talking to a real person on the phone → Automation through AI generated voice? Would this even help or create other problems?)

- Getting a sufficiently large sample size can be challenging. 1000 is usually enough though.
- Interest groups might try to influence the results (how easily this can be done depends on the way the survey is conducted)

(b) In 20 different and independent polls like the "Sonntagsfrage", one Confidence Interval (CI) each is determined for the proportion of SPD voters at level alpha = 0.05. What is the probability that 1) none of the CI's, 2) all of the CI's and 3) at least 3/4 of the CI's cover the true value?

While speaking strictly, we should not regard the confidence interval as a probabilistic statement about the parameter, we can still define a Bernoulli distributed random variable X: "The confidence interval of the i-th survey covers the true value" $\rightarrow \quad X_i \sim B(0.95)$
We have a sample of $X = \sum_{i=1}^{20} X_i$ random variables with $X \sim Bin(20, 0.95)$

We can thus easily calculate the probability, either using the density of the binomial distribution or using R as a calculator:

- None of the CIs covering the true value: $\mathbb{P}(X = 0) \approx 0$

```
> dbinom(0, 20, 0.95)
[1] 9.536743e-27
```

- All of the CIs covering the true value: $\mathbb{P}(X = 20) = 0.358$

```
> dbinom(20, 20, 0.95)
[1] 0.3584859
```

- 15 or more of the CIs covering the true value: $\mathbb{P}(X \geq 15) = 0.9997 \approx 1$

```
> pbinom(14, 20, 0.95, lower.tail = FALSE)
[1] 0.9996707
```

(c) Confirm the information on the margin of error given by the polling institute (i.e., Schwankungsbreite) for a proportion value of 10% and 50% at a confidence level of $1 - \alpha = 0.95$. State and explain your assumed distributional model and show your calculations.

Hint: The margin of error was derived (and rounded to the next integer) for the sample size of the "Sonntagsfrage". Look at the random variables X_i, $i = 1, ..., 1065$ as response of the i-th individual to the Sunday poll, where

$$X_i = \begin{cases} 1, & \text{Person votes for a party that has 10\% (or 50\%) of the votes,} \\ 0, & \text{Person votes for another party} \end{cases}$$

In essence, the exercise asks us to determine the width of the symmetrical $1 - \alpha = 0.95\%$ confidence interval around an estimate $\hat{\pi}$ for a binomially distributed variable $X \sim Bin(1000, \pi)$.

We know that the width of an (approximate) $(1 - \alpha)$ confidence interval for dichotomous variables is equal to the right bound minus the left bound:

$$b = \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} - \left(\hat{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right) = 2 \cdot z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

For fixed n and α, the width depends only on the product $\hat{\pi}(1-\hat{\pi})$ and thus on the estimate $\hat{\pi}$. For $\alpha = 0.05$, $\hat{\pi} = 0.1$ and $n = 1065$, we get:

$$b_{0.1} = 2 \cdot z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 2 \cdot 1.96 \cdot \sqrt{\frac{0.1 \cdot (1-0.1)}{1065}} = 0.037.$$

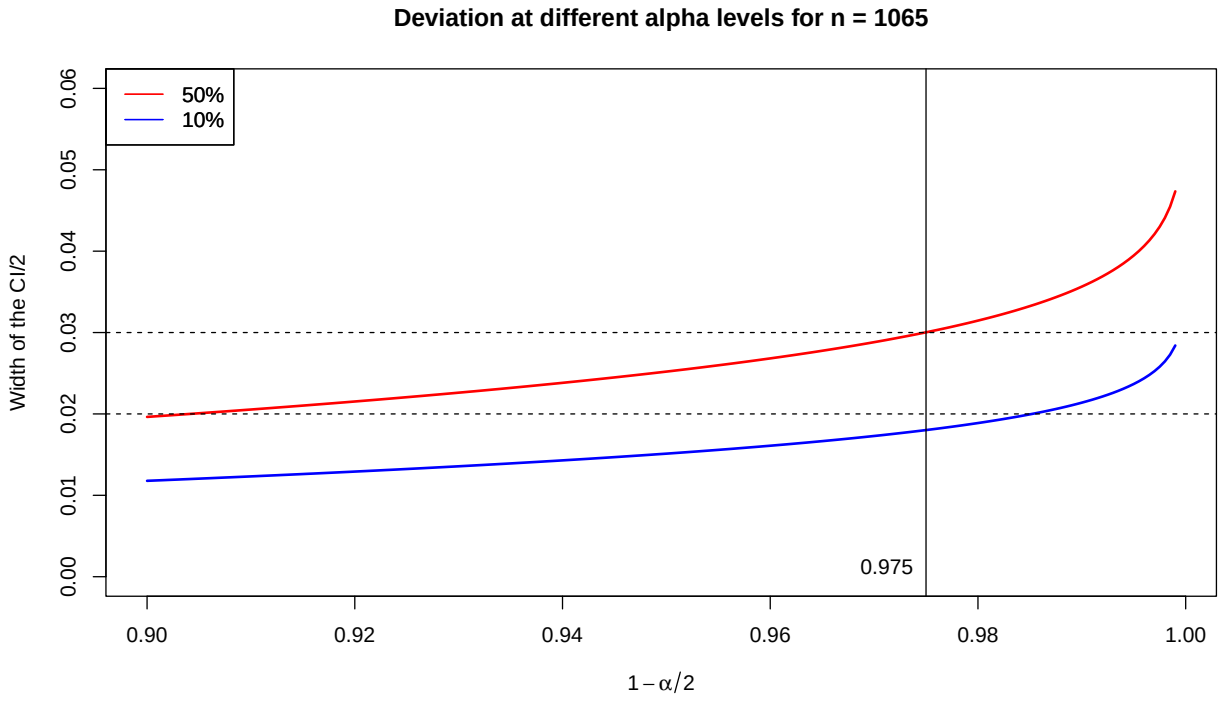
This is not the value we expected. The values given by infratest dimap are meant to be interpreted as deviation in only one direction $(\pi \pm x)$, such that the factor 2 in the formula has to be removed. We're only interested in the width of the one sided interval $[\hat{\pi}, \hat{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}]$ (or in the other direction). Thus:

$$b_{0.1} = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 1.96 * \sqrt{\frac{0.1 * (1-0.1)}{1065}} = 0.018 \approx 0.02.$$

and for $\pi = 0.5$:

$$b_{0.5} = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 1.96 * \sqrt{\frac{0.5 * (1-0.5)}{1065}} = 0.030.$$

Which matches our expectation. As a function of alpha we can see:



For a level of $\alpha = 0.05$, the specified deviance is approximately correct. For a party with 50% of the votes it is exactly 3 percentage points and for a party with 10% it is a bit lower than 2 percentage points, but since the margins of errors were probably rounded, this confirms the interval widths stated by infratest dimap.

Exercise 6.3 - Confidence Intervals: R Exercise

The local administration of the town of Gütersloh is interested in estimating the proportion of car owners among the citizens of the city to help with road planning. For cost reasons, $n = 50$ inhabitants of the city are first asked whether they own a car, of which 40 say they do own a car.

(a) First, state your assumed model and estimate the proportion of car owners in Gütersloh. Calculate an **approximate** 95 % confidence interval for the proportion of car owners.

$X \sim Bin(50, \pi), \quad \widehat{\pi} = \frac{40}{50} = 0.8$
We will use the approximation to the normal, even though $n\widehat{\pi}(1-\widehat{\pi}) = 8 < 9$:

$$\begin{aligned}
[G_u, G_o] &= \left[\hat{\pi} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\,,\ \hat{\pi} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right] \\
&= \left[0.8 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.8(1-0.8)}{50}}\,,\ 0.8 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.8(1-0.8)}{50}}\right] \\
&= [0.689, 0.911]
\end{aligned}$$

(b) Another random sample is collected, in which $n = 250$ people are now interviewed. This resulted in a proportion of 80% of car owners. Again, give a 95 % confidence interval and compare it to the one calculated in (a). Why do they differ?

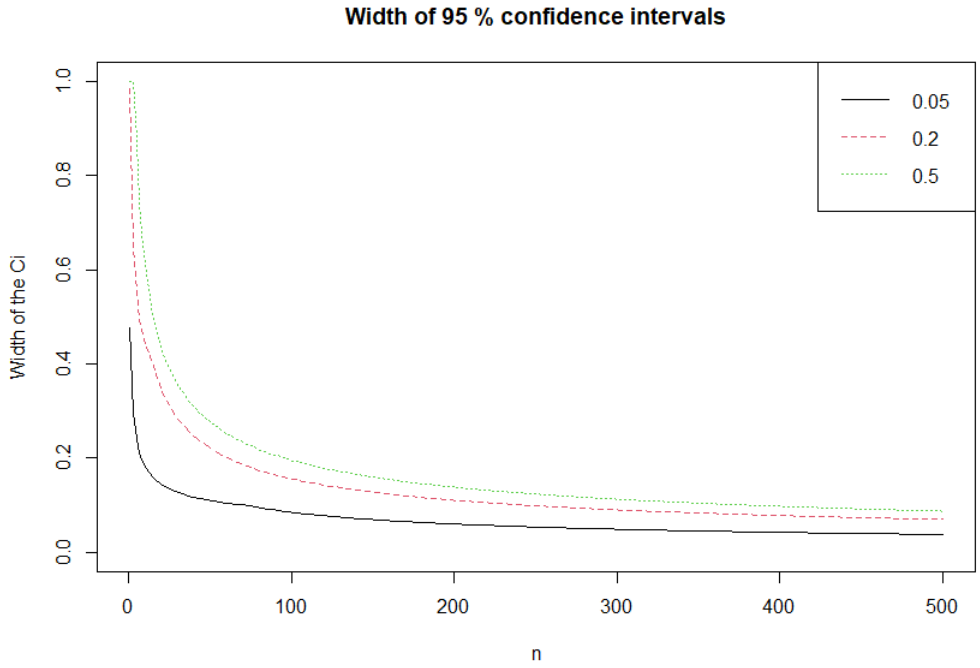
$$\begin{aligned}
[G_u, G_o] &= \left[0.8 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.8(1-0.8)}{250}}\,,\ 0.8 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.8(1-0.8)}{250}}\right] \\
&= [0.750, 0.850]
\end{aligned}$$

Since we have a larger sample, there is more evidence to support the assumed value of π and we are more confident in the result, thus the CI gets smaller.

(c) Implement a function `binomial_ci(n, pi)` in R that computes the **approximate** 95% confidence interval for the estimate *pi* in a a binomial distribution given the sample size n, as you also did in (a) and (b). The function should be able to handle a vectorised input of n (multiple values for n at once) and return a list with two elements: `[lower bound(s), upper bound(s)]`. *Note:* Keep in mind that the true value of *pi* can only be in the range $[0, 1]$.

See `R` file.

(d) Construct a plot that visualises the relationship between the width of the CI (make use of your function from (c)) and the sample size n. Plot the curve for different values of π, e.g. for $\pi = 0.05$, $\pi = 0.2$ and $\pi = 0.5$. You might want to make use of the function `manipulate` for interactively changing n and/or π. What can you tell about the width of the CI for different values of π?



We observe that the closer the value of π is to 0.5, the larger the width of the confidence interval is. The curves of the widths of the confidence intervals never intersect for different values of π. With increasing n, the width of all confidence intervals decreases. The larger n gets, the less an additional increase in n further decreases the uncertainty.