# Probability and Statistics

## Exercise sheet 14

You do not have to hand in your solutions. There will be no corrections of Exercise sheet 14. The solution will be uploaded on the web page.

**Exercise 14.1**

Let $X_1 \dots X_n$ be i.i.d. with a continuous distribution $F$. The sign test is a test where the null hypothesis is that the median of $X$ is $m$, i.e.

$$F(m) = \frac{1}{2}.$$

Use the duality theorem (Theorem 8.3.1 in the lecture notes) to construct an approximate confidence interval for the median of $F$ at level 95%.

**Solution 14.1** We have to construct a test between the hypotheses:

$$H_0 : F^{\leftarrow}\left(\frac{1}{2}\right) = m,$$
$$H_1 : F^{\leftarrow}\left(\frac{1}{2}\right) \neq m.$$

We use the statistic $T_{n,m} = \sum_{i=1}^{n} I_{\{X_i \leq m\}}$ and the test is given by

$$\varphi(\mathbf{x}) = 1 \Leftrightarrow \left|T_{n,m} - \frac{n}{2}\right| > c(n, \alpha),$$

where $\mathbf{x} = (x_1 \dots x_n)$, $n$ is the size of the sample and $\alpha$ is the level of the test. We have that under $H_0$, $T_{n,m}$ follows a binomial law $Bin(n, \frac{1}{2})$. So if we define $k = \frac{n}{2} - c(n, \alpha)$, we should have, thanks to the symmetry of the binomial coefficients,

$$\mathbb{P}_{H_0}[T_{n,m} < k] = \sum_{j=0}^{k-1} \binom{n}{j} 0.5^n \leq \frac{\alpha}{2} < \sum_{j=0}^{k} \binom{n}{j} 0.5^n = \mathbb{P}_{H_0}[T_{n,m} \leq k]$$

and $\frac{n}{2} + c(n, \alpha) = n - k$. Take $A \subseteq \mathbb{R}^{n+1}$ to be

$$A := \left\{(\mathbf{x}, m) \in \mathbb{R}^n \times \mathbb{R} : k \leq \sum_{i=1}^{n} I_{\{x_i \geq m\}} \leq n - k\right\};$$

then we want $C(x_1, \dots, x_n) = \{m : (\mathbf{x}, m) \in A\}$ to produce a confidence interval of level $1 - \alpha$ for $F^{\leftarrow}(\frac{1}{2})$. Note that with the usual notation $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ for the ordered sample, we have

$$m \in [x_{(j)}, x_{(j+1)}) \Leftrightarrow \sum_{i=1}^{n} I_{\{x_i > m\}} = j,$$

so that $C(x_1, ..., x_n) = [x_{(k)}, x_{(n-k)}]$. Thanks to the central limit theorem,

$$\begin{aligned}\mathbb{P}_m\left[k \le \sum_{i=1}^n I_{\{X_i>m\}} \le n-k\right] &= \mathbb{P}_m\left[\frac{2}{\sqrt{n}}\left(k-\frac{n}{2}\right) \le \frac{2}{\sqrt{n}}\left(\sum_{i=1}^n I_{\{X_i>m\}}-\frac{n}{2}\right) \le \frac{2}{\sqrt{n}}\left(\frac{n}{2}-k\right)\right]\\ &\approx \Phi\left(\frac{2}{\sqrt{n}}\left(\frac{n}{2}-k\right)\right)-\Phi\left(\frac{2}{\sqrt{n}}\left(k-\frac{n}{2}\right)\right)\\ &= 2\Phi\left(\frac{2}{\sqrt{n}}\left(\frac{n}{2}-k\right)\right)-1.\end{aligned}$$

We want this to be at least $0.975 = \Phi\left(\Phi^{\leftarrow}(0.975)\right) = \Phi(1.96)$, so $k \approx \frac{n}{2} - \frac{1.96}{2}\sqrt{n} \approx \left\lfloor \frac{n}{2} - \sqrt{n}\right\rfloor$. Then

$$C((X_1, ..., X_n)) \approx \left[X_{\left(\lfloor \frac{n}{2}-\sqrt{n}\rfloor\right)}, X_{\left(\lfloor \frac{n}{2}+\sqrt{n}\rfloor\right)}\right],$$

is an approximate confidence interval with level 95%.

**Exercise 14.2** We want to investigate the effect of an outlier on confidence intervals. Let $X_1, \dots, X_n$ be i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ with unknown $\sigma$.

(a) Give the two-sided confidence interval for the unknown parameter $\mu$ with level $\alpha$.

(b) How does the realized confidence interval behave for $x_1 \to \infty$ and fixed $x_2, \dots, x_n$ ?
Hint: Show first for every $c \in \mathbb{R}$ that $\sum_{i=1}^n (x_i - c)^2 = \sum_{i=1}^n (x_i - \overline{x}_n)^2 + n(c - \overline{x}_n)^2$.

**Solution 14.2**

(a) $\left[\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}},\ \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right]$ from the lecture notes.

(b) The right endpoint clearly diverges to $\infty$, because it is at least $\overline{x}_n$.
For the left endpoint, we set $a := \frac{1}{n}\sum_{i=2}^n x_i$, and with the hint, we obtain

$$\begin{aligned} s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2 &= \frac{1}{n-1}\sum_{i=1}^n (x_i - a)^2 - \underbrace{\frac{n}{n-1}(a-\overline{x}_n)^2}_{=\frac{1}{n(n-1)}x_1^2}\\ &= \underbrace{\frac{1}{n-1}\sum_{i=2}^n (x_i-a)^2}_{=:b} + \frac{1}{n-1}(x_1-a)^2 - \frac{1}{n(n-1)}x_1^2\\ &= b + \frac{1}{n-1}(x_1^2 - 2x_1 a + a^2) - \frac{1}{n(n-1)}x_1^2\\ &= x_1^2\left(\frac{1}{n-1} - \frac{1}{n(n-1)} - \frac{2a}{x_1}\frac{1}{n-1} + \frac{1}{n-1}\frac{a^2}{x_1^2} + \frac{b}{x_1^2}\right) =: x_1^2 f(x_1).\end{aligned}$$

and $f(x_1) \to \frac{1}{n-1} - \frac{1}{n(n-1)} = \frac{1}{n}$ as $x_1 \to \infty$.

Moreover, we have $\overline{x}_n = \frac{1}{n}x_1 + a = x_1\left(\frac{1}{n} + \frac{a}{x_1}\right) =: x_1 g(x_1)$ with $g(x_1) \to \frac{1}{n}$ as $x_1 \to \infty$. This gives

$$\begin{aligned}\overline{x}_n - \frac{s_n}{\sqrt{n}}t_{n-1,1-\frac{\alpha}{2}} = \overline{x}_n\left(1 - \frac{s_n}{\overline{x}_n\sqrt{n}}t_{n-1,1-\frac{\alpha}{2}}\right) &= \overline{x}_n\left(1 - \frac{x_1\sqrt{f(x_1)}}{x_1 g(x_1)\sqrt{n}}t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \overline{x}_n\left(1 - \frac{\sqrt{f(x_1)}}{g(x_1)\sqrt{n}}t_{n-1,1-\frac{\alpha}{2}}\right) =: \overline{x}_n h(x_1)\end{aligned}$$

and $h(x_1) \to 1 - t_{n-1,1-\frac{\alpha}{2}}$ as $x_1 \to \infty$. For all levels used in practice, the $t$-quantile is strictly larger than 1, so that the left endpoint converges to $-\infty$. This means that for extreme values of one single data point, the confidence interval does not give any information any more—every value for $\mu = \mathbb{E}[X]$ is plausible.

**Proof of the hint:** Squaring out and using $n\overline{x}_n = \sum_{i=1}^n x_i$ so that $\sum_{i=1}^n (x_i - \overline{x}_n) = 0$ gives

$$\begin{aligned}\sum_{i=1}^n (x_i - c)^2 &= \sum_{1=1}^n (x_i - \overline{x}_n + \overline{x}_n - c)^2 \\ &= \sum_{1=1}^n (x_i - \overline{x}_n)^2 + 2(\overline{x}_n - c) \sum_{1=1}^n (x_i - \overline{x}_n) + n(\overline{x}_n - c)^2 \\ &= \sum_{1=1}^n (x_i - \overline{x}_n)^2 + n(\overline{x}_n - c)^2,\end{aligned}$$

as claimed.

**Exercise 14.3** In einer Nagelfabrik will man die Länge der produzierten Nägel möglichst genau schätzen. Man nimmt an, dass die Längen der Nägel $X_i, i = 1, ..., n$, unabhängig und approximativ normalverteilt sind mit Varianz 1 mm$^2$ und unbekanntem Erwartungswert $\mu$ mm. Wieviele Nägel muss man mindestens messen, damit das zweiseitige 95%-Konfidenzintervall für $\mu$ höchstens Länge 0.5 mm hat?

**Solution 14.3** Mit den gemessenen Nägellängen $X_i$ haben wir für das Konfidenzintervall die Grenzen

$$\frac{1}{n}\sum_{i=1}^n X_i \pm \sqrt{\frac{\sigma^2}{n}}\Phi^{\leftarrow}\Big(1 - \frac{\alpha}{2}\Big).$$

Somit erhält man für die Länge des Vertrauensintervalls $C(\mathbf{X})$

$$|C(\mathbf{X})| = \frac{2\Phi^{\leftarrow}\Big(1 - \frac{\alpha}{2}\Big)}{\sqrt{n}}$$

und möchte $|C(\mathbf{X}| \leq 0.5$ haben, also $\Phi^{\leftarrow}\left(1 - \frac{\alpha}{2}\right)\sigma \leq 0.5\sqrt{n}$. Nach n aufgelöst und für $\alpha = 0.05$, $\sigma^2 = 1$ eingesetzt gibt das $n \geq \left(4\Phi^{\leftarrow}\left(1 - \frac{\alpha}{2}\right)\sigma\right)^2 = 16 \times (1.96)^2 = 61.47$. Also muss man mindestens 62 Nägel messen.

**Exercise 14.4** Wie oft muss man eine Münze werfen, damit das 99%-Konfidenzintervall für die Erfolgswahrscheinlichkeit $p$ (Kopf wird als Erfolg interpretiert) höchstens Länge 0.01 hat?
**Hinweis:** Benutzen Sie die Normalapproximation. Die Intervallgrenzen hängen noch von $p$ ab. Maximieren Sie nun über $p$.

**Solution 14.4** Wegen dem zentralen Grenzwertsatz ist

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - p}{\sqrt{\frac{1}{n}p(1-p)}}$$

approximativ standardnormalverteilt. Mit der Normalapproximation erhalten wir also für das approximative Konfidenzintervall zum Niveau $\alpha$ die Grenzen

$$\frac{1}{n}\sum_{i=1}^n X_i \pm \Phi^{\leftarrow}\Big(1 - \frac{\alpha}{2}\Big)\sqrt{\frac{1}{n}p(1-p)}.$$

Die Intervallgrenzen müssen von $p$ unabhängig sein, deshalb maximieren wir $p(1-p)$ über $p$. Das Maximum wird für $p = 0.5$ angenommen. Damit erhalten wir:

$$|C(\mathbf{X})| \leq \frac{2\Phi^{\leftarrow}\left(1 - \frac{\alpha}{2}\right)}{2\sqrt{n}}$$

und möchten $|C(\mathbf{X})| \leq 0.01$ haben. Das haben wir approximativ für $0.01\sqrt{n} \geq \Phi^{\leftarrow}\left(1 - \frac{\alpha}{2}\right)$, also $n \geq \left(100\Phi^{\leftarrow}\left(1 - \frac{\alpha}{2}\right)\right)^2 = (100 \times 2.58)^2 = 66\,564$.

If you have feedback regarding the exercise sheets, please send a mail to Jakob Heiss.