Statistics for Data Science
Wintersemester 2024/25
Solutions to the 4^{th} exercise

1. (a) We obtain

$$\mathbb{E}[\overline{m}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} m = m.$$

(b) There holds

$$\begin{aligned}\operatorname{Var}(\overline{m}_n) &= \mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - m\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m)\right)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[(X_i - m)^2] = \frac{\sigma^2}{n}.\end{aligned}$$

(c) Since $X \sim \mathcal{N}(1,2)$, we deduce that

$$\operatorname{Var}(\overline{m}_n) \le 10^{-2} \quad \Leftrightarrow \quad \frac{2}{n} \le 10^{-2} \quad \Leftrightarrow \quad n \ge 200.$$

2. (a) Please see the script `w4e2.py` on the course webpage.

In the implementation, it may be useful to notice that if $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \operatorname{Ber}(p)$, then $X_1 + \cdots + X_n \sim \operatorname{Bin}(n,p)$.

(b) Since $\mathbb{E}[X] = p$ and $\operatorname{Var}(X) = p(1-p)$, the central limit theorem implies that

$$\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p\right) \xrightarrow{d} \mathcal{N}(0,1)$$

or, loosely speaking,

$$\frac{1}{n}\sum_{k=1}^{n} X_k \overset{d}{\approx} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \quad \text{for large } n.$$

In particular,

$$H = \sum_{k=1}^{n} X_k \overset{d}{\approx} \mathcal{N}\big(np, np(1-p)\big).$$

(c) Let $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. We wish to find $a > 0$ such that

$$\int_{\mu-a}^{\mu+a} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,\mathrm{d}x = 0.95.$$

Using the change of variables $y = \frac{x-\mu}{\sigma}$, where $\mathrm{d}x = \sigma \mathrm{d}y$, we obtain

$$\begin{aligned}
\int_{-a/\sigma}^{a/\sigma} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}y = 0.95 \quad &\Leftrightarrow \quad 2\int_0^{a/\sigma} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}y = 0.95 \\
&\Leftrightarrow \quad \int_0^{a/\sigma} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}y = \frac{0.95}{2} \\
&\Leftrightarrow \quad \Phi\left(\frac{a}{\sigma}\right) = \int_{-\infty}^{a/\sigma} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}y = \frac{1}{2} + \frac{0.95}{2} \\
&\Leftrightarrow \quad \frac{a}{\sigma} = \Phi^{-1}\left(\frac{1}{2} + \frac{0.95}{2}\right) \\
&\Leftrightarrow \quad a = \sigma\Phi^{-1}\left(\frac{1}{2} + \frac{0.95}{2}\right) \\
&\Leftrightarrow \quad a = \sqrt{np(1-p)}\Phi^{-1}\left(\frac{1}{2} + \frac{0.95}{2}\right),
\end{aligned}$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x$ is the CDF of $\mathcal{N}(0,1)$ and Φ^{-1} the corresponding quantile function. Plugging in the values $p = \frac{1}{3}$, $n = 10^3$, and $\Phi^{-1}\big(\frac{1}{2} + \frac{0.95}{2}\big) = 1.95996\ldots$ yields

$$a = 29.2174\ldots$$

and therefore

$$\mathbb{P}(304.116 \leq H \leq 362.551) \approx 0.95.$$

3. Please see the script `w4e3.py` on the course webpage.

By the central limit theorem,

$$\sqrt{\frac{n}{\sigma^2}}(Q_{d,n}(g) - I_d(g)) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\sigma^2 = \mathrm{Var}(g(X))$. Loosely speaking,

$$Q_{d,n}(g) - I_d(g) \overset{d}{\approx} \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \quad \text{for large } n.$$

Therefore one expects the Monte Carlo convergence rate

$$|Q_{d,n}(g) - I_d(g)| \approx \frac{\sigma}{\sqrt{n}}.$$

Especially, the expected convergence rate is $\mathcal{O}(n^{-1/2})$ regardless of $d \in \mathbb{N}$.

4. (a) Let $k \in \mathbb{N}$. Clearly, $f_k(x) \geq 0$ for all $x \in \mathbb{R}$. On the other hand,

$$\int_{-\infty}^{\infty} f_k(x)\,\mathrm{d}x = \int_1^{\infty} \frac{k}{x^{k+1}}\,\mathrm{d}x = -\big[x^{-k}\big]_{x=1}^{x=\infty} = 1.$$

Therefore f_k is a probability density function for all $k \in \mathbb{N}$.

(b) Let $\ell \in \mathbb{N}$. Now

$$\mathbb{E}[X^\ell] = \int_1^\infty x^\ell \frac{k}{x^{k+1}}\,\mathrm{d}x = k\int_1^\infty x^{\ell-k-1}\,\mathrm{d}x.$$

If $0 \le \ell < k$, then

$$\mathbb{E}[X^\ell] = \frac{k}{\ell-k}\big[x^{\ell-k}\big]\Big|_{x=1}^{x=\infty} = \frac{k}{k-\ell} < \infty.$$

If $\ell \ge k$, then $x^{\ell-k-1} \ge x^{-1}$ for all $x \ge 1$ and thus

$$\mathbb{E}[X^\ell] \ge \int_1^\infty x^{-1}\,\mathrm{d}x = \infty.$$

(c) If $k = 1$, then X is not integrable and the sample average $\overline{X}_n$ of course diverges since $\mathbb{E}[X] = \infty$.

If $k = 2$, then X is integrable, but not square-integrable. The sample average $\overline{X}_n$ converges almost surely to $\mathbb{E}[X]$ by the (strong) law of large numbers. The central limit theorem does not hold.†

If $k \ge 3$, then X is square-integrable. The sample average $\overline{X}_n$ converges almost surely to $\mathbb{E}[X]$ by the (strong) law of large numbers. In fact,

$$\sqrt{\frac{n}{\sigma^2}}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\mu = \mathbb{E}[X] = \frac{k}{k-1}$ and $\sigma^2 = \mathrm{Var}(X) = \frac{k}{(k-2)(k-1)^2}$ for $k \ge 3$. The theoretically expected convergence rate is $|\overline{X}_n - \mathbb{E}[X]| = \mathcal{O}(n^{-1/2})$.

Appendix

The integral formula $I_d(g) = 2^d\cos(\frac{d}{2})(\sin\frac{1}{2})^d$ for $d \in \mathbb{N}$ in task 3 is a consequence of the following, more general result.

Lemma (Oscillatory integral family). Let $c_i \in \mathbb{R}\setminus\{0\}$ and $w_i \in \mathbb{R}$ for all $i \in \{1,\ldots,d\}$. Then

$$\int_0^1 \cdots \int_0^1 \cos\left(2\pi w_1 + \sum_{i=1}^d c_i x_i\right)\mathrm{d}x_d\cdots\mathrm{d}x_1 = 2^d\cos\left(2\pi w_1 + \frac{1}{2}\sum_{i=1}^d c_i\right)\prod_{i=1}^d \frac{\sin(\frac{c_i}{2})}{c_i}.$$

Proof. The claim can be proved *a fortiori* by first proving the more general integral formula

$$\int_0^1 \cdots \int_0^1 \cos\left(C + \sum_{i=1}^d c_i x_i\right)\mathrm{d}x_d\cdots\mathrm{d}x_1 = 2^d\cos\left(C + \frac{1}{2}\sum_{i=1}^d c_i\right)\prod_{i=1}^d \frac{\sin(\frac{c_i}{2})}{c_i}, \quad C \in \mathbb{R}, \tag{1}$$

†Numerically, the Gaussian approximation does seem to resemble $\overline{X}_n$ for large n to a certain degree. This is somewhat expected, since the density $f_{2+\varepsilon}(x) = \frac{2+\varepsilon}{x^{3+\varepsilon}}\mathbf{1}_{[1,\infty)}(x)$ actually *does* satisfy the conditions of the CLT for arbitrarily small $\varepsilon > 0$. However, the tails of the distribution corresponding to $f_{2+\varepsilon}$ are not perfectly captured when ε is small (in practice, this means that n may need to be an **extremely** large number for the CLT to kick into effect for $0 < \varepsilon \ll 1$).

using induction with respect to the dimension $d \in \mathbb{N}$. The basis $d = 1$ of the induction argument follows by first computing

$$\begin{aligned}
\int_0^1 \cos(C + c_1 x_1)\,\mathrm{d}x_1 &= \frac{1}{c_1}\big[\sin(C + c_1) - \sin(C)\big] \\
&= \frac{1}{c_1}\big[\cos(C + c_1 - \tfrac{\pi}{2}) - \cos(C - \tfrac{\pi}{2})\big] \\
&= \frac{1}{c_1}\big[\cos(C + \tfrac{1}{2}c_1 + \tfrac{1}{2}c_1 - \tfrac{\pi}{2}) - \cos(C + \tfrac{1}{2}c_1 - \tfrac{1}{2}c_1 - \tfrac{\pi}{2})\big] \\
&= 2\cos(C + \tfrac{1}{2}c_1)\frac{\sin(\frac{c_1}{2})}{c_1},
\end{aligned}$$

where the final step follows from the trigonometric identity $\cos(x+y) - \cos(x-y) = 2\cos(x + \frac{\pi}{2})\sin(y)$ with $x = C + \frac{1}{2}c_1 - \frac{\pi}{2}$ and $y = \frac{1}{2}c_1$.

Suppose that (1) holds for some $d \in \mathbb{N}$. Then it follows that

$$\begin{aligned}
&\int_0^1 \cdots \int_0^1 \cos\left(C + \sum_{i=1}^{d+1} c_i x_i\right) \mathrm{d}x_{d+1} \cdots \mathrm{d}x_1 \\
&= \int_0^1 \cdots \int_0^1 \left(\int_0^1 \cos\left(C + \sum_{i=1}^{d} c_i x_i + c_{d+1}x_{d+1}\right) \mathrm{d}x_{d+1}\right) \mathrm{d}x_d \cdots \mathrm{d}x_1 \\
&= \frac{1}{c_{d+1}} \int_0^1 \cdots \int_0^1 \left[\sin\left(C + \sum_{i=1}^{d} c_i x_i + c_{d+1}\right) - \sin\left(C + \sum_{i=1}^{d} c_i x_i\right)\right] \mathrm{d}x_d \cdots \mathrm{d}x_1 \\
&= \frac{1}{c_{d+1}} \int_0^1 \cdots \int_0^1 \left[\cos\left(C + \sum_{i=1}^{d} c_i x_i + c_{d+1} - \frac{\pi}{2}\right) - \cos\left(C + \sum_{i=1}^{d} c_i x_i - \frac{\pi}{2}\right)\right] \mathrm{d}x_d \cdots \mathrm{d}x_1 \\
&= \frac{1}{c_{d+1}} \left[2^d \cos\left(C + \frac{1}{2}\sum_{i=1}^{d} c_i + c_{d+1} - \frac{\pi}{2}\right) \prod_{i=1}^{d} \frac{\sin(\frac{c_i}{2})}{c_i}\right. \\
&\qquad\left. - 2^d \cos\left(C + \frac{1}{2}\sum_{i=1}^{d} c_i - \frac{\pi}{2}\right) \prod_{i=1}^{d} \frac{\sin(\frac{c_i}{2})}{c_i}\right] \\
&= \frac{2^d}{c_{d+1}} \left[\cos\left(C + \frac{1}{2}\sum_{i=1}^{d} c_i + c_{d+1} - \frac{\pi}{2}\right) - \cos\left(C + \frac{1}{2}\sum_{i=1}^{d} c_i - \frac{\pi}{2}\right)\right] \prod_{i=1}^{d} \frac{\sin(\frac{c_i}{2})}{c_i} \\
&= 2^{d+1} \cos\left(C + \frac{1}{2}\sum_{i=1}^{d+1} c_i\right) \prod_{i=1}^{d+1} \frac{\sin(\frac{c_i}{2})}{c_i},
\end{aligned}$$

where the final step follows once again from the trigonometric identity $\cos(x+y) - \cos(x-y) = 2\cos(x + \frac{\pi}{2})\sin(y)$, this time by choosing $x = C + \frac{1}{2}\sum_{i=1}^{d+1} c_i - \frac{\pi}{2}$ and $y = \frac{1}{2}c_{d+1}$. This proves the assertion. □