Statistics for Data Science
Wintersemester 2024/25
Solutions to the 1$^{\text{st}}$ exercise

1. We have been given the following probabilities:

$$\begin{aligned} &\mathbb{P}(+|D) = 0.9, &&\mathbb{P}(-|D) = 0.1, \\ &\mathbb{P}(+|\neg D) = 0.1, &&\mathbb{P}(-|\neg D) = 0.9, \\ &\mathbb{P}(D) = 0.01, &&\mathbb{P}(\neg D) = 0.99. \end{aligned} \tag{1}$$

Here, $\mathbb{P}(+|\neg D)$ is called the *false positive rate* and $\mathbb{P}(-|D)$ is called the *false negative rate* of the test; $\mathbb{P}(+|D)$ is the *sensitivity* of the test and $\mathbb{P}(-|\neg D)$ is the *specificity*; $\mathbb{P}(D)$ is the *prevalence* of the disease in the population.

By Bayes' formula and the law of total probability:

$$\begin{aligned} \mathbb{P}(D|+) &= \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+)} \\ &= \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|\neg D)\mathbb{P}(\neg D)} \\ &= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.1 \cdot 0.99} = 0.08333\ldots \end{aligned}$$

The prevalence of the disease in the population is relatively low compared to the false positive rate of the test, so the test is not at all reliable in predicting whether the individual has the disease or not. If the false positive rate were, e.g., 0.01 instead, then the probability of being sick, given that the test is positive, would increase to 0.47619.

One can try to mitigate the situation by doing multiple tests. With the parameters (1), the probability of being sick, given that *two tests* come back positive, is

$$\frac{0.9^2 \cdot 0.01}{0.9^2 \cdot 0.01 + 0.1^2 \cdot 0.99} = 0.45.$$

While this is still not great, it is a significant improvement.

The quantity $\mathbb{P}(D|+)$ is the *positive predictive value* of a test. Another interesting quantity is the *negative predictive value* of a test $\mathbb{P}(\neg D|-)$, the probability that someone with a negative test does not have the disease. In the case of this example, we have

$$\begin{aligned} \mathbb{P}(\neg D|-) &= \frac{\mathbb{P}(-|\neg D)\mathbb{P}(\neg D)}{\mathbb{P}(-)} \\ &= \frac{\mathbb{P}(-|\neg D)\mathbb{P}(\neg D)}{\mathbb{P}(-|D)\mathbb{P}(D) + \mathbb{P}(-|\neg D)\mathbb{P}(\neg D)} \\ &= \frac{0.9 \cdot 0.99}{0.1 \cdot 0.01 + 0.9 \cdot 0.99} = 0.99887\ldots, \end{aligned}$$

so the negative predictive value of this test is excellent.

2. Please see the script `w1e2.py` on the course homepage.

3. (a) The CDF is the antiderivative of the PDF. For $x < 0$, there holds

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} 0 \, \mathrm{d}t = 0.$$

For $x \geq 0$, we obtain

$$\begin{aligned} F(x) = \mathbb{P}(X \leq x) &= \int_{-\infty}^{x} f(t) \, \mathrm{d}t = \int_{0}^{x} \underbrace{4t\mathrm{e}^{-t^2}}_{=-2\frac{\mathrm{d}}{\mathrm{d}t}(1+\mathrm{e}^{-t^2})} (1 + \mathrm{e}^{-t^2})^{-2} \, \mathrm{d}t \\ &= 2\Big[(1 + \mathrm{e}^{-t^2})^{-1}\Big]_{t=0}^{t=x} = \frac{2}{1 + \mathrm{e}^{-x^2}} - 1. \end{aligned}$$

Therefore

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{2}{1+\mathrm{e}^{-x^2}} - 1 & \text{if } x \geq 0. \end{cases}$$

(b) Let $q \in (0, 1)$. The mapping $F|_{(0,\infty)} \colon (0, \infty) \to (0, 1)$ is monotonically strictly increasing, thus invertible, so we obtain the expression for the inverse by solving x from

$$\begin{aligned} q = F(x) \quad &\Leftrightarrow \quad q = \frac{2}{1 + \mathrm{e}^{-x^2}} - 1 \\ &\Leftrightarrow \quad \frac{1-q}{q+1} = \mathrm{e}^{-x^2} \\ &\Leftrightarrow \quad \log \frac{1-q}{q+1} = -x^2 \\ &\Leftrightarrow \quad \log \frac{q+1}{1-q} = x^2 \\ &\Leftrightarrow \quad \sqrt{\log \frac{q+1}{1-q}} = x, \end{aligned}$$

from which we can read off the expression for $F^{-1}(q) = \sqrt{\log \frac{q+1}{1-q}}$, $q \in (0, 1)$.†

(c) To compute the probability $\mathbb{P}(0 < X < 1)$, we can use the CDF:

$$\mathbb{P}(0 < X < 1) = F(1) - F(0) = \frac{2}{1 + \mathrm{e}^{-1}} - 1 = \frac{\mathrm{e} - 1}{\mathrm{e} + 1} \approx 0.4621.$$

To obtain the value of $a \in \mathbb{R}$, note that

$$F(a) = \mathbb{P}(X \leq a) = 0.95 \quad \Leftrightarrow \quad a = F^{-1}(0.95) = \sqrt{\log \frac{1 + 0.95}{1 - 0.95}} = \sqrt{\log(39)} \approx 1.9140.$$

†If G is an inverse function of F, i.e., $F(G(q)) = q$ for all $q \in (0, 1)$, then the quantile function coincides with the function inverse since $F^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \geq q\} = G(q)$ for all $q \in (0, 1)$.

4. (a) The different possibilities are

$X(\omega) = \omega_1 + \omega_2$	$\omega_1 = 0$	$\omega_1 = 1$
$\omega_2 = 0$	0	1
$\omega_2 = 1$	1	2

Therefore, the PMF takes the values

$$
\begin{aligned}
p(0) &= \mathbb{P}(X=0) = \mathbb{P}(X^{-1}(0)) = \frac{1}{4},\\
p(1) &= \mathbb{P}(X=1) = \mathbb{P}(X^{-1}(1)) = \frac{2}{4} = \frac{1}{2},\\
p(2) &= \mathbb{P}(X=2) = \mathbb{P}(X^{-1}(2)) = \frac{1}{4}.
\end{aligned}
$$

(b) If $x < 0$, then $F(x) = \sum_{a \le x} p(a) = 0$.
If $0 \le x < 1$, then $F(x) = \sum_{a \le x} p(a) = p(0) = \frac{1}{4}$.
If $1 \le x < 2$, then $F(x) = \sum_{a \le x} p(a) = p(0) + p(1) = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}$.
If $x \ge 2$, then $F(x) = \sum_{a \le x} p(a) = p(0) + p(1) + p(2) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1$.
Therefore

$$
F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{4} & \text{if } 0 \le x < 1 \\ \frac{3}{4} & \text{if } 1 \le x < 2 \\ 1 & \text{if } x \ge 2. \end{cases}
$$

(c) We now have

$$
\begin{aligned}
&F(x) = \frac{1}{4} \text{ for } 0 \le x < 1 &&\Rightarrow \quad \text{for all } q \in (0, \tfrac{1}{4}],\ F^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \ge q\} = 0,\\
&F(x) = \frac{3}{4} \text{ for } 1 \le x < 2 &&\Rightarrow \quad \text{for all } q \in (\tfrac{1}{4}, \tfrac{3}{4}],\ F^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \ge q\} = 1,\\
&F(x) = 1 \text{ for } x \ge 2 &&\Rightarrow \quad \text{for all } q \in (\tfrac{3}{4}, 1),\ F^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \ge q\} = 2.
\end{aligned}
$$

Therefore

$$
F^{-1}(q) = \begin{cases} 0 & \text{if } q \in (0, \frac{1}{4}] \\ 1 & \text{if } q \in (\frac{1}{4}, \frac{3}{4}] \\ 2 & \text{if } q \in (\frac{3}{4}, 1). \end{cases}
$$

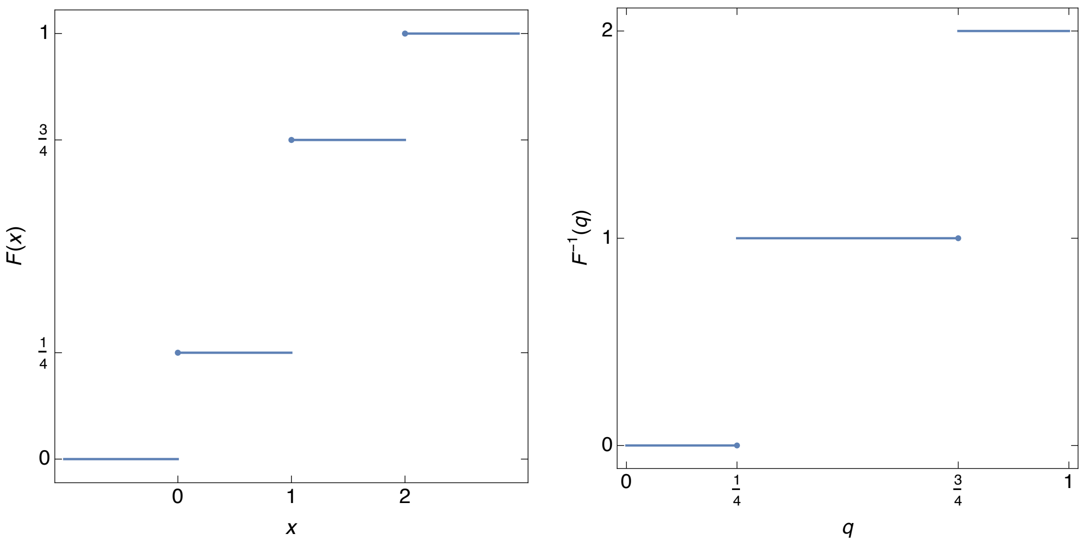

Figure 1: Plots of the CDF F from exercise 4b (left) and the quantile function F^{-1} from exercise 4c (right).