Statistics: Tutorial sheet 1 - Solutions to Practice Exercises

Exercise 1. A student has gotten very excited from the histogram examples and has decided to test them out himself. He simulates a data set of fifty observations from the standard normal distribution. Then he creates a histogram and plots the standard normal pdf in one figure. The result is shown in Figure 1.

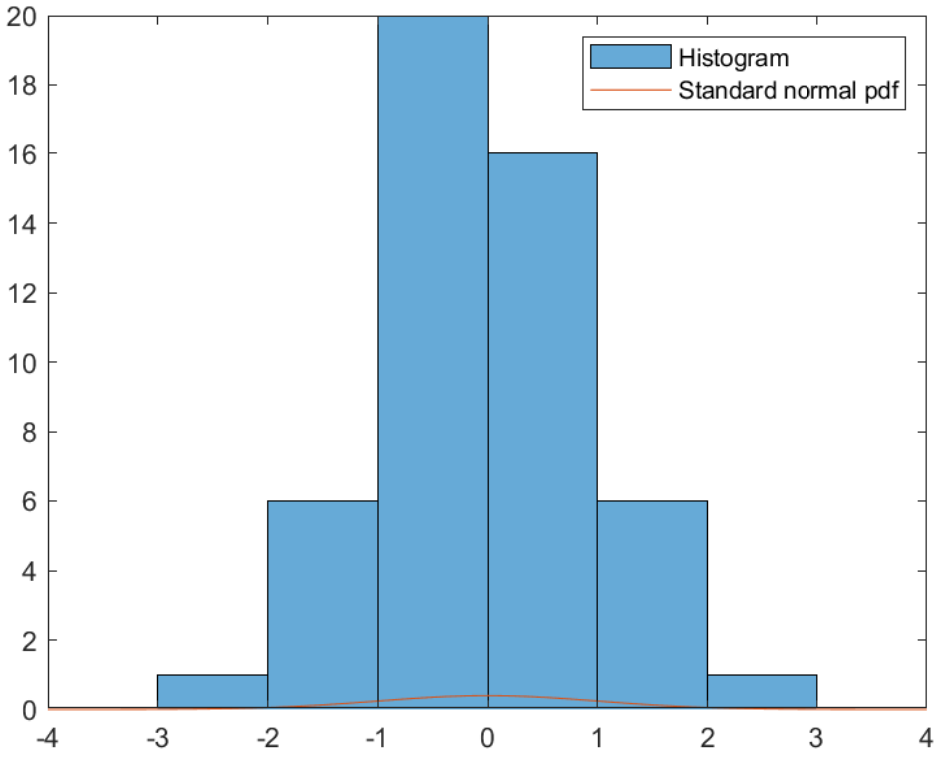


Figure 1: Estimated histogram

a. Recall that for $y \in (a_{j-1}, a_j]$ we defined the histogram function as

$$h_n(y) = \sum_{j=1}^{m} \mathbb{1}_{\{a_{j-1} < y \leq a_j\}} \left(\sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1} < x_i \leq a_j\}} \right).$$

Formally solve the integral

$$\int_{\mathbb{R}} h_n(y) dy.$$

b. Explain what mistake the student has made and provide a solution.

SOLUTION.

a. Note that $h_n(y)$ is zero below a_0 and above a_m. Therefore we have

$$\int_{\mathbb{R}} h_n(y)dy = \sum_{j=1}^{m} \int_{a_{j-1}}^{a_j} h_n(y)dy = \sum_{j=1}^{m} \int_{a_{j-1}}^{a_j} \left(\sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1}<x_i\leq a_j\}} \right) dy$$
$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1}<x_i\leq a_j\}} (a_j - a_{j-1}) = c \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}_{\{a_{j-1}<x_i\leq a_j\}}.$$

The next step is to swap the summation signs to obtain

$$c \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{\{a_{j-1}<x_i\leq a_j\}}.$$

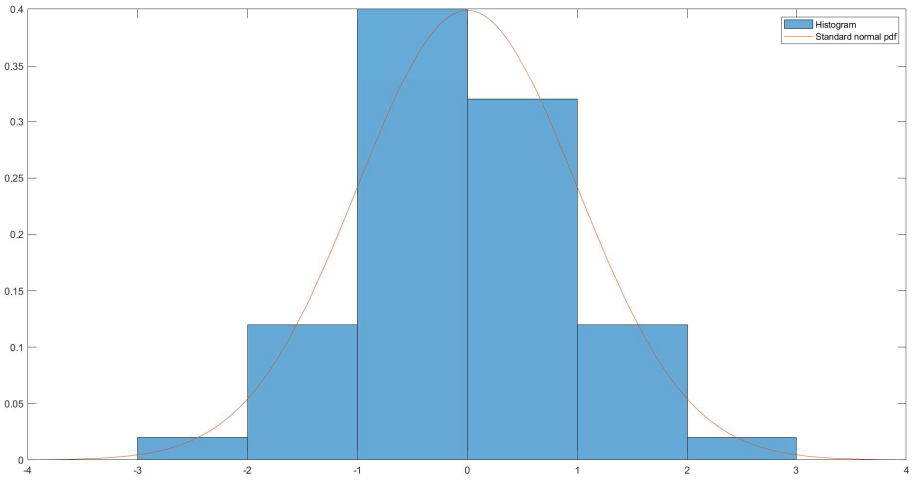
Note that x_1 is in an interval (a_{j-1}, a_j) for exactly one unique j, for instance it could be in $(a_3, a_4]$. The same is true for each x_i and so the inner summation

$$\sum_{j=1}^{m} \mathbb{1}_{\{a_{j-1}<x_i\leq a_j\}} = 1.$$

We conclude that

$$\int_{\mathbb{R}} h_n(y)dy = c \sum_{i=1}^{n} 1 = cn.$$

b. We have shown that $\int_{\mathbb{R}} h_n(y)dy = cn = 50$ in our example, while for any pdf $g(x)$ we have $\int_{\mathbb{R}} g(y)dy = 1$. The student has not rescaled his histogram, which makes the density appear completely at the bottom. Instead he should have plotted the histogram $\tilde{h}_n(y) = \frac{h_n(y)}{cn}$ which would have resulted in the following figure.



Exercise 2. Let $X_1, \ldots, X_n$ be a random sample with population $g(x)$ and let $Y \sim g(x)$ be independent of each X_i.

a. Prove that $P(Y \leq X_{(1)}) = \frac{1}{n+1}$.

b. Prove for $k \in \{2, 3, \ldots, n\}$ that we have $P(X_{(k-1)} < Y \leq X_{(k)}) = \frac{1}{n+1}$.

c. Use this to show that $P(Y \leq X_{(k)}) = \frac{k}{n+1}$.

SOLUTION.

a. The statement $Y \leq X_{(1)}$ is equal to saying that Y is the smallest random variable out of the sequence $Y, X_1, X_2, \ldots, X_n$. Since each element in this sequence is independent with the same distribution we must have

$$P(Y \text{ is the smallest}) = P(X_1 \text{ is the smallest}) = \cdots = P(X_n \text{ is the smallest}).$$

Next, note that one of these random variables has to be the smallest one, so that

$$P(Y \text{ is the smallest}) + \sum_{i=1}^{n} P(X_i \text{ is the smallest}) = 1.$$

These two equations solve for $P(Y \text{ is the smallest}) = P(Y \leq X_{(1)}) = \frac{1}{n+1}$.

b. Repeat the argument above by noting that the statement $X_{(k-1)} < Y \leq X_{(k)}$ is equal to saying that Y is the k'th random variable out of the sequence $Y, X_1, X_2, \ldots, X_n$.

c. This immediately follows from

$$\begin{aligned} P(Y \leq X_{(k)}) &= P(Y \leq X_{(1)}) + \sum_{i=2}^{k} P(X_{(i-1)} < Y \leq X_{(i)}) \\ &= \frac{1}{n+1} + \sum_{i=2}^{k} \frac{1}{n+1} = \frac{1 + (k-1)}{n+1} = \frac{k}{n+1}. \end{aligned}$$

Exercise 3. In this exercise we study a new distribution called the *Logistic distribution* which for $\theta = (\mu, \sigma)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$ has cdf

$$G(x \mid \theta) = \frac{1}{1 + e^{-(x-\mu)/\sigma}} \qquad \text{if } x \in \mathbb{R}.$$

The logistic distribution has become very popular in machine learning due to the sharp S shape and easy differentiability of its cdf. This makes it one of the most common distributions to model probabilities of outcomes for categorical variables as in logistic regression[1] or feed forward neural networks.[2]

a. Derive the location-scale family of the Logistic$(0, 1)$ distribution.

b. Verify that $G(x \mid \mu, \sigma)^{-1} = \mu + \sigma G(x \mid 0, 1)^{-1}$.

SOLUTION.

a. We have $G(x \mid 0, 1) = \frac{1}{1+e^{-x}}$, so for $\mu \in \mathbb{R}$ and $\sigma > 0$ it follows that

$$G\left(\frac{x-\mu}{\sigma} \mid 0, 1\right) = \frac{1}{1 + e^{-(x-\mu)/\sigma}} = G(x \mid \mu, \sigma).$$

Therefore the set of Logistic distributions is the location-scale family of $G(x \mid 0, 1)$.

b. We find the inverse of $G(x \mid \mu, \sigma)$ as follows

$$\begin{aligned} y = G(x \mid \mu, \sigma) = \frac{1}{1 + e^{-(x-\mu)/\sigma}} \qquad &\Leftrightarrow \qquad \frac{1}{y} = 1 + e^{-(x-\mu)/\sigma} \\ &\Leftrightarrow \qquad -\log\left(\frac{1}{y} - 1\right) = \frac{x-\mu}{\sigma} \\ &\Leftrightarrow \qquad \mu - \sigma \log\left(\frac{1}{y} - 1\right) = x \end{aligned}$$

It follows that

$$G(y \mid \mu, \sigma)^{-1} = \mu - \sigma \log\left(\frac{1}{y} - 1\right) = \mu + \sigma\left[-\log\left(\frac{1}{y} - 1\right)\right] = \mu + \sigma G(y \mid 0, 1)^{-1}.$$

Exercise 4. Suppose that a marine biologist wants to know the number of fish N_0 that live in a lake. The water is not very clear, so she cannot count them from a helicopter or use other such practical methods. Instead she comes up with a different approach. She starts fishing in the lake, catches r different fish, puts a mark on each of them and then throws them back into the water. One week later she comes back and performs an experiment by catching a fish, writing down if its marked or not and throwing it back. After repeating this for a total of n times she has obtained a dataset $\boldsymbol{x} = (x_1, \ldots, x_n)$.

a. Assume that the random variables $X_1, \ldots, X_n$ are independent. Formulate a statistical model for $\boldsymbol{X} = (X_1, \ldots, X_n)$. What would be an intuitive estimator for N_0?

[1] https://en.wikipedia.org/wiki/Logistic_regression

[2] https://en.wikipedia.org/wiki/Feedforward_neural_network

b. Is the independence assumption reasonable here?

SOLUTION.

a. Every time we grab a fish there are r marked fishes out of N_0 total ones, and each observation is independent. Therefore we have a random sample from a population in $\{\text{Bernoulli}(p) \mid p \in [0,1]\}$. An intuitive estimator for p_0 is given by $\hat{p} = \overline{X}$. Therefore, since $p_0 = r/N_0$ we choose $\hat{N} = r/\hat{p} = r/\overline{X}$.

b. The independence assumption is probably not reasonable here. Once the biologist catches a fish and throws it back, she is much more likely to catch that same fish again compared to other ones in the lake. Therefore $P(X_2 = X_1) > 0$ and thus the observations are not independent.