Statistics: Tutorial sheet 2

Mandatory exercises

Exercise 1. For each of the following distributions, indicate whether it is a member of the exponential family and, if yes, provide expressions for each component inside the definition of the exponential family.

a. Exponential(λ): $f(x|\lambda) = \lambda e^{-\lambda x}$; $0 \leq x < \infty$, $\lambda > 0$,

b. Gamma(α,β): $f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$; $0 \leq x < \infty$, $\alpha, \beta > 0$,

c. Uniform(0,θ): $f(x|\theta) = \frac{1}{\theta}$; $0 \leq x \leq \theta$, $\theta > 0$,

d. Poisson(λ): $f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$; $x = 0, 1, \ldots$, $\lambda > 0$.

Exercise 2. Let $g(x|\mu, \sigma)$ with cdf $G(x|\mu, \sigma)$ be a member of the location-scale family defined by $g(x)$ with cdf $G(x)$. Prove the following general results:

a. $G(x|\mu, \sigma) = G\left(\frac{x-\mu}{\sigma}\right)$,

b. $G^{-1}(x|\mu, \sigma) = \mu + \sigma G^{-1}(x)$.

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.

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.

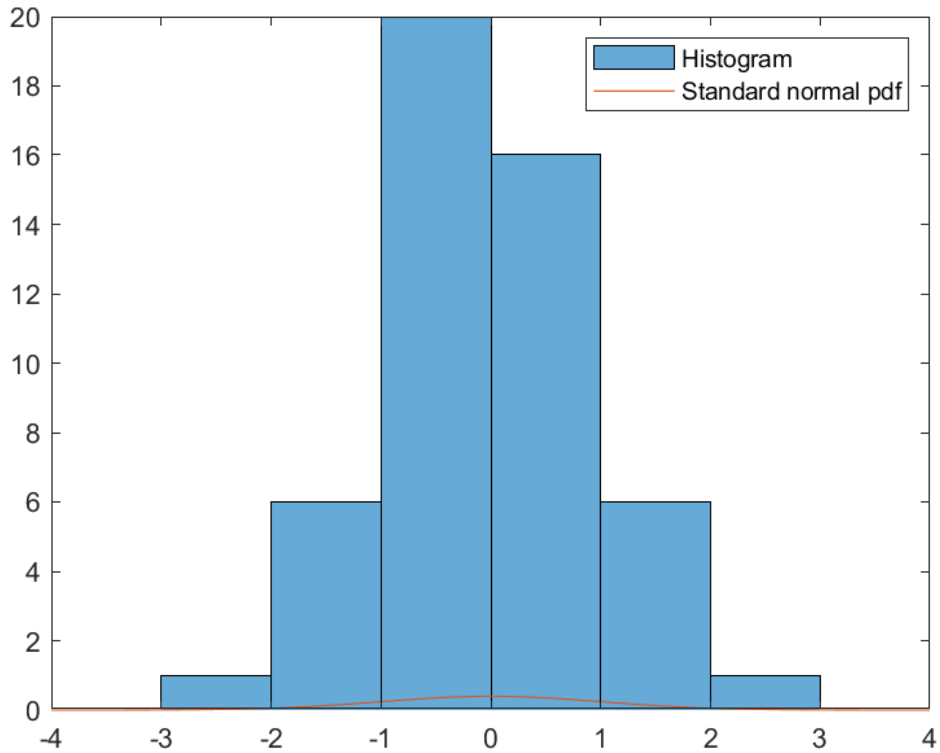

Figure 1: Estimated histogram

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

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

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

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 as in logistic regression[1] or feed forward neural networks.[2]

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

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

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

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?

b. Is the independence assumption reasonable here?