Assignment 9 for Statistical Computing and Empirical Methods: Statistical estimation and parametric modelling

Henry Reeve

Introduction

This document describes your eighth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommended that you first watch video lecture 9 entitled “Estimation and parametric modelling”.

1 Bias and variance

In this question we will carry out a simulation study to investigate the bias and variance of an estimate $\hat{q}_n$ for a population parameter q. Let’s consider a Bernoulli random variable $\mathcal{B}(q)$ with $q \in (0,1)$. Given a sample $X_1, \cdots, X_n \sim \mathcal{B}(q)$ we estimate q with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Firstly, when carrying out a simulation study involving randomness we should start by setting a random seed. This ensures that our code is reproducible and re-running the code leads to the same results.

We can create a randomised function called **sample_mean_Bernoulli()** which takes an input two parameters - a parameter $q \in (0,1)$ and a natural number $n \in \mathbb{N}$, generates a Bernoulli sample $X_1, \cdots, X_n \sim \mathcal{B}(q)$, then computes the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and returns $\bar{X}$.

```
sample_mean_Bernoulli<-function(q,n){

  bernoulli_sample<-rbinom(n,1,q) # generate Bernoulli sample

  return(mean(bernoulli_sample)) # return the mean

  }
```

Use your understanding of the Bernoulli distribution, the Binomial distribution and the function **rbinom()** to understand why this works.

Next set $q = 0.7$, $n = 25$ and $\text{num\_trials} = 1000$.

```
q<-0.7
n<-10
num_trials<-1000
```

We can use the **map_dbl()** function combined with the **sample_mean_Bernoulli()** function to generate a vector of size $\text{num\_trials} = 1000$ such that every element of the vector is an output of the function **sample_mean_Bernoulli()**.

```
sample_mean_Bernoulli_vect<-map_dbl(seq(num_trials),~sample_mean_Bernoulli(n=n,q=q))
```

We can estimate the bias, the variance and the mean square error of the sample mean as follows:

```
sample_mean_bias <- mean(sample_mean_Bernoulli_vect)-q
sample_mean_variance <-var(sample_mean_Bernoulli_vect)
sample_mean_mse <- mean((sample_mean_Bernoulli_vect-q)**2)
```

Print out the mean squared error and the sum of the variance and the square bias.

```
sample_mean_mse # mean squared error of our estimate
```

```
## [1] 0.01892
```

```
sample_mean_variance+sample_mean_bias**2 # variance plus squared bias
```

```
## [1] 0.01893892
```

Explain the relationship between the above two numbers via the bias-variance decomposition.

Now use the above method to investigate the sample standard deviation $S := \sqrt{\frac{1}{n-1}\left(X_i - \overline{X}\right)^2}$ used as an estimate of the population standard deviation $\sqrt{q(1-q)}$.

What is the bias of this estimate? What is the variance of this estimate? What is the mean square error? How are they related?

2 Maximum likelihood estimation for the Gaussian distribution

Suppose that $X_1, \cdots, X_n \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ are independent and identically distributed with unknown μ_0 and σ_0. Show that $\overline{X} := \frac{1}{n}\sum_{i=1} X_i$ is the maximum likelihood estimator for μ_0 and $S := \frac{1}{n}\sum_{i=1}(X_i - \overline{X})^2$ is the maximum likelihood estimator for σ_0.

3 Maximum likelihood estimation and the exponential distributin

Recall from Assignment 8 that given a positive real number $\lambda > 0$, an exponential random variable X with parameter λ is a continuous random variable with density $p_\lambda : \mathbb{R} \to (0, \infty)$ defined by

$$p_\lambda(x) := \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

Suppose that $X_1, \cdots, X_n$ is an i.i.d sample from the exponential distribution with an unknown parameter $\lambda > 0$. What is the maximum likelihood estimate for λ?

4 Multivariate Gaussian for some parameters of the Hawks distribution

In this exercise we shall model data from the Hawks data set as a multivariate Gaussian distribution.

First load the Hawks data set as follows.

```
library(Stat2Data)
data("Hawks")
```

Now use your data wrangling skills to filter extract a subset of the Hawks data set so that every Hawk belongs to the “Red-Tailed” species, and extract the “Weight”, “Tail” and “Wing” columns. The returned output should be a data frame called “RedTailedDf” with three numerical columns and 577 examples.

Next model the rows of the data frame “RedTailedDf” as random i.i.d. vectors $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^3$ is the population mean vector and $\Sigma \in \mathbb{R}^{3\times 3}$ is the population covariance matrix.

Compute the maximum likelihood estimates for μ and Σ. Are these unbiased?

5 Population median

Suppose that a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is drawn from a univarate (one-dimensional) Gaussian with mean μ and variance σ^2. What is its population median?

Give an example of a random variable with more than one median.