Assignment 6 for Statistical Computing and Empirical Mathods

Dr. Henry WJ Reeve

Teaching block 1 2021

Introduction

This document describes your sixth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommend that you first watch video lectures 13 and 14.

Begin by creating an Rmarkdown document with html output. You are not expected to hand in this piece of work, but it is a good idea to get used to using Rmarkdown.

1 A Gaussian model for Red tailed hawks

In this question we will fit a Gaussian model to a Red-Tailed hawk data set.

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.

Display the first five rows of the "RedTailedDf". The resulting subset of the data frame should look as follows:

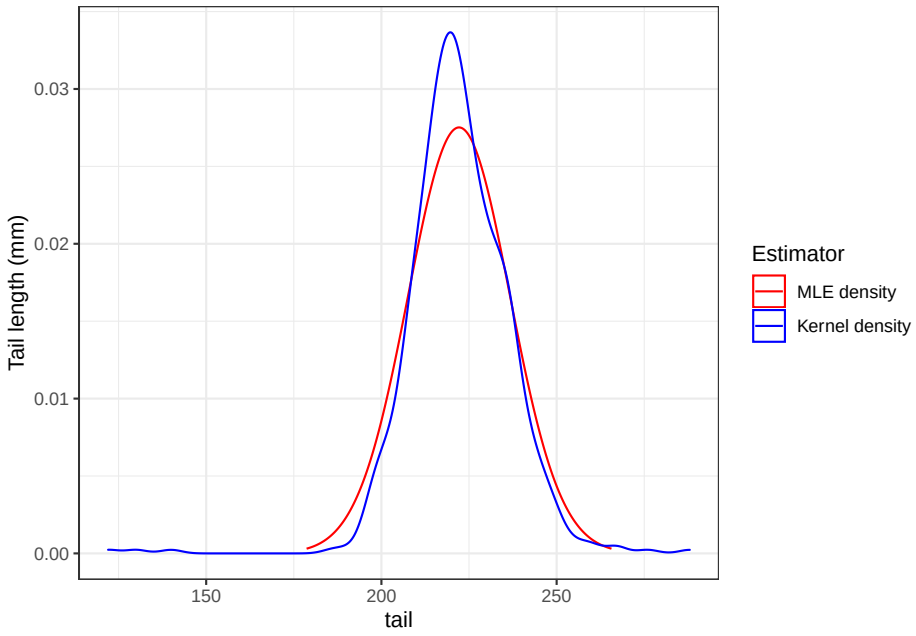
```
##   Weight Tail Wing
## 1    920  219  385
## 2    930  221  376
## 3    990  235  381
## 4   1090  230  412
## 5    960  212  370
```

We now model the vector of tail lengths from "RedTailedDf" as a sequence $X_1, \cdots, X_n \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ consisting of independent and identically distributed with unknown population mean μ_0 and population variance σ_0^2.

The maximum likelihood estimates for μ_0 is given by $\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the maximum likelihood estimate for σ_0^2 is given by $\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}_{\mathrm{MLE}}\right)^2$.

Apply the maximum likelihood method to compute estimates $\hat{miu}_{\mathrm{MLE}}$ for μ_0 and $\hat{\sigma}^2_{\mathrm{MLE}}$ for σ^2_0.

Next generate a plot which compares the probability density function for your fitted Gaussian model for the tail length of the Red-Tailed hawks with a kernel density plot. Your plot should look as follows:



2 Location estimators with Gaussian data

In this question we compare two estimators for the population mean μ_0 in a Gaussian setting in which we have independent and identically distributed data $X_1, \cdots, X_n \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$.

What is the population median of a Gaussian random variable $X_i \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$?

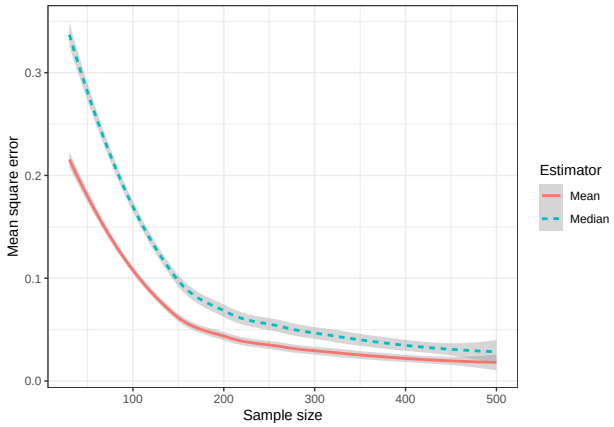
The following code generates a data frame consisting which estimates the mean squared error of the sample median as an estimator of μ_0.

```
set.seed(0)
num_trials_per_sample_size<-100
min_sample_size<-5
max_sample_size<-1000
sample_size_inc<-5
mu_0<-1
sigma_0<-3

simulation_df<-crossing(trial=seq(num_trials_per_sample_size),
                        sample_size=seq(min_sample_size,
                                        max_sample_size,sample_size_inc))%>%
  # create data frame of all pairs of sample_size and trial
  mutate(simulation=pmap(.l=list(trial,sample_size),
                         .f=~rnorm(.y,mean=mu_0,sd=sigma_0)))%>%
  # simulate sequences of Gaussian random variables
  mutate(sample_md=map_dbl(.x=simulation,.f=median))%>%
  # compute the sample medians
  group_by(sample_size)%>%
    summarise(msq_error_md=mean((sample_md-mu_0)^2))
```

Modify the above code to include estimates of the mean square error of the sample mean.

Generate a plot which includes both the mean square error of the sample mean and the sample median as a function of the sample size. Your plot might look something like the following:



3 Unbiased estimation of the population variance

In this question we consider samples $X_1, \cdots, X_n \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ consisting of independent and identically distributed with unknown population mean μ_0 and unknown population variance σ_0^2.

Let $\overline{X}$ be the sample mean, let $\hat{V}_{\mathrm{MLE}} := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ and let $\hat{V}_{\mathrm{U}} := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

Conduct a simulation study which compares the bias of $\hat{V}_{\mathrm{MLE}}$ as an estimator of the population variance σ_0^2 with the bias of $\hat{V}_{\mathrm{U}}$ as an estimator for the population variance σ_0^2.

Is $\sqrt{\hat{V}_{\mathrm{U}}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$ an unbiased estimator for σ_0?

As an optional extra, give an analytic formula for the bias of $\hat{V}_{\mathrm{MLE}}$ and $\hat{V}_{\mathrm{U}}$ as estimators of σ_0^2.

4 Maximum likelihood estimators 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 population mean μ_0 and unknown population standard deviation σ_0.

Given an expression for the likelihood function $\ell(\mu, \sigma^2)$ based upon a sample $X_1, \ldots, X_n$.

Derive a formula for the derivative of the log-likelihood $\frac{\partial}{\partial\lambda}\log\ell(\lambda)$.

Show that $\overline{X} := \frac{1}{n}\sum_{i=1} X_i$ is the maximum likelihood estimator for μ_0 and $S^2 := \frac{1}{n}\sum_{i=1}(X_i - \overline{X})^2$ is the maximum likelihood estimator for σ_0^2.

What is the maximum likelihood estimator for σ_0?

5 Maximum likelihood estimation with the Poisson distribution

In this question we shall consider the topic of maximum likelihood estimation for a an independent and identically distributed sample from a Poisson random variable. Recall that Poisson random variables are a family of discrete random variables with distributions supported on $\mathbb{N}_0 := \{0, 1, 2, 3, \cdots\}$. Poisson random variables are frequently used to model the number of events which occur at a constant rate in situations where the occurrence of individual events are independent. For example, we might use the Poisson distribution to model the number of mutations of a given strand of DNA per time unit, or the number of customers who arrive at store over the course of a day. A classic example of statistical modelling based on a Poisson distribution is due to the statistician Ladislaus Josephovich Bortkiewicz. Bortkiewicz used the the Poisson distribution to model the number of fatalities due to horse-kick per year for each group of cavalary. We shall apply maximum likelihood estimation to Bortkiewicz's data. First let's explore maximum likelihood estimation for Poisson random variables.

A Poisson random variable has a probability mass function $p_\lambda : \mathbb{R} \to (0, \infty)$ with a single parameter $\lambda > 0$. The probability mass function $p_\lambda : \mathbb{R} \to (0, \infty)$ is defined for $x \in \mathbb{R}$ by

$$p_\lambda(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x \in \mathbb{N}_0 \\ 0 & \text{for } x \notin \mathbb{N}_0. \end{cases}$$

Suppose that you have a sample of independent and identically distributed random variables $X_1, \ldots, X_n \sim p_{\lambda_0}$ i.e. $X_1, \ldots, X_n$ are independent and each has probability mass function p_{λ_0}.

Show that for a sample $X_1, \ldots, X_n$, the likelihood function $\ell : (0, \infty) \to (0, \infty)$ is given by

$$\ell(\lambda) := e^{-n \cdot \lambda} \cdot \lambda^{n \cdot \overline{X}} \cdot \left(\prod_{i=1}^{n} \frac{1}{X_i!} \right),$$

where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the sample mean.

Derive a formula for the derivative of the log-likelihood $\frac{\partial}{\partial \lambda} \log \ell(\lambda)$.

Show that $\lambda \mapsto \log \ell(\lambda)$ reaches its maximum at the single point at which $\lambda = \overline{X}$. Hence, the maximum likelihood estimate for the true parameter λ_0 is $\hat{\lambda}_{\mathrm{MLE}}$.

Now conduct a simulation experiment which explores the behavior of $\hat{\lambda}_{\mathrm{MLE}}$ on simulated data. You may wish to consider a setting in which $\lambda_0 = 0.5$ and generate a plot of the mean squared error as a function of the sample size.

Now that we have explored maximum likelihood estimation with a Poisson distribution for simulated data we shall return to Poission modelling with real data. Let's take a look at the famous horse-kick fatality data set explored by Ladislaus Josephovich Bortkiewicz. A csv file containing this data is available within Blackboard.

Download the csv file and load the file into an R data frame. You may wish to use the `read.csv()` function.

The count data for horse fatalities per year, per cavalary corps are given in the "fatalities" column. Model the values in this column as independent random variables $X_1, \ldots, X_n$ from a Poisson distribution with parameter λ_0 and compute the maximum likelihood estimate $\hat{\lambda}_{\mathrm{MLE}}$ for λ_0.

Use your fitted Poisson model to give an estimate for the probability that a single cavalry corps has no fatalities due to horse kicks in a single year. You may want to use the `dpois` function.

As an optional extra give a formula for $\mathcal{I}(\lambda) := -\mathbb{E}\left(\frac{\partial^2}{\partial \lambda^2} \log \ell(\lambda) \right)$. Next generate a simulation involving random samples of size 1000 from a Poisson random variable with parameter $\lambda_0 = 0.5$. Give a kernel density plot of $\sqrt{n \cdot \mathcal{I}(\lambda_0)} \left(\hat{\lambda}_{\mathrm{MLE}} - \lambda_0 \right)$.

6 Maximum likelihood estimation for the exponential distribution

Recall from Assignment 5 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 $\lambda_0 > 0$. What is the maximum likelihood estimate for λ_0?

We shall now use the exponential distribution to model the differences in purchase times between customers at a large supermarket. In Blackboard you will find the "CustomerPurchase" csv file.

Download the "CustomerPurchase" csv file and load the file into an R data frame. You may wish to use the `read.csv()` function. The first column is the purchase time given in seconds since the store opens.

Add a new column in your data frame called "time_diffs" which gives the time in seconds until the next customer's purchase. That is, letting $Y_1, Y_2, \ldots, Y_{n+1}$ denote the sequence of arrival times in seconds, the `time_diffs`

column contiains $X_1, \ldots, X_n$ where $X_i = Y_{i+1} - Y_i$ for each $i = 1, \ldots, n$. You may want to use the `lead()` function.

Model the sequence of differences in purchase times $X_1, \ldots, X_n$ as independent and identically distributed exponential random variables. Compute the maximum likelihood estimate of the rate parameter $\hat{\lambda}_{\mathrm{MLE}}$.

Use your fitted exponential model to give an estimate of the probability of an arrival time in excess of one minute. You may wish to make use of the `pexp()` function.

7 (**) MLE for the capture and recapture model

As an optional extra we return to the capture and recapture model discussed in lecture 9.

Recall that in this example there are n_0 squirrels living on an island. A conservationist captures t squirrels at random before tagging and releasing them. A week later, the conservationist captures k squirrels at random again, and counts how many have already been tagged. For simplicity we assume that the population of n squirrels is constant over the time period, and on both occasions the squirrels are selected purely at random.

We let Z be the random variable corresponding to the number of recaptured squirrels.

In lecture 9 we showed that for $q \leq \min\{t, k\}$ we have $\mathbb{P}(Z = q) = \frac{\binom{t}{q} \cdot \binom{n_0 - t}{k-q}}{\binom{n_0}{k}}$ and $\mathbb{P}(Z = q) = 0$ for $q > \min\{t, k\}$. Hence, the likelihood function $\ell : \mathbb{N} \to [0, 1]$ is given by

$$\ell(n) = \frac{\binom{t}{Z} \cdot \binom{n-t}{k-Z}}{\binom{n}{k}},$$

for all $n \in \mathbb{N}$. Give a formula for the maximum likelihood estimate $\hat{n}_{\mathrm{MLE}}$ of n_0.