Assignment 8

EMATM0061: Statistical Computing and Empirical Methods, TB1, 2024

Dr. Rihuan Ke

Introduction

This is the 8th assignment for Statistical Computing and Empirical Methods. This assignment is mainly based on Lectures 21, 22, 23 and 24 (see the Blackboards). Please note that you don't need to submit this assignment.

Load the tidyverse package:

```
library(tidyverse)
```

1. A chi-squared test of population variance

Suppose we have an i.i.d. sample $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and a conjectured value for the population variance σ_0^2. We wish to test the null hypothesis that σ_0^2.

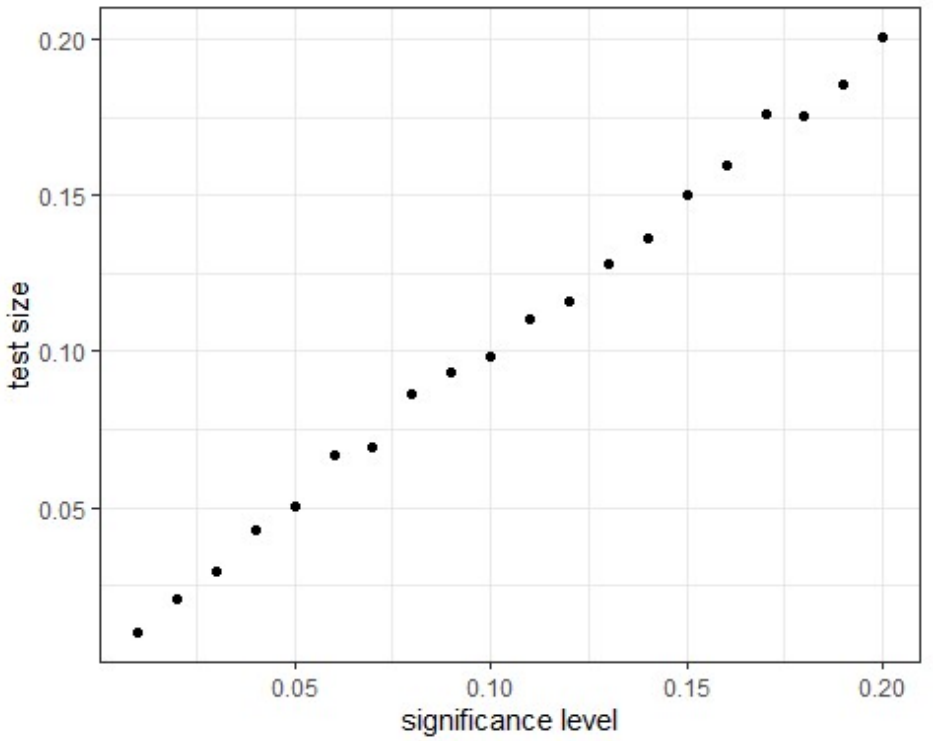
(Q1)

Implement a function called "chi_square_test_one_sample_var" which takes as input a sample called ""sample"" and a null value for the variance called "sigma_square_null".

(Q2)

Conduct a simulation study to see how the size of the test varies as a function of the significance level. You can consider a sample size of 100, $\mu = 1$, $\sigma^2 = 4$.

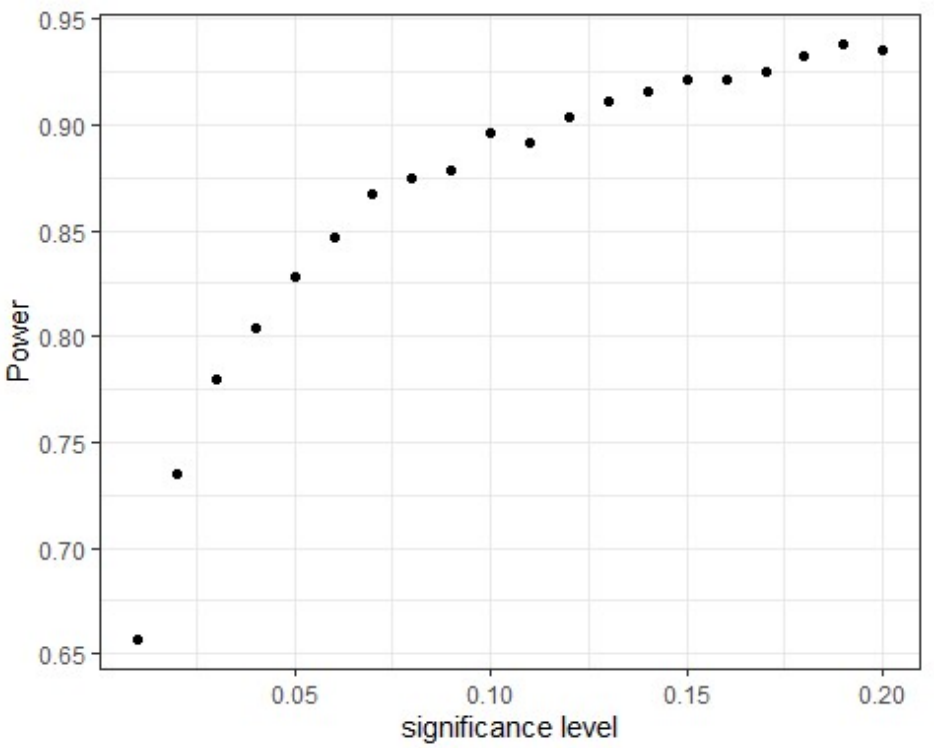
For example, you can create a plot similar to the one below.



(Q4)

Conduct a simulation study to see how the statistical power of the test varies as a function of the significance level. You can consider a sample size of 100, $\mu = 1$, $\sigma^2 = 6$ and $\sigma_0^2 = 4$.

For example, you can create a plot similar to the one below.



(Q5)

Load the "Palmer penguins" library and extract a vector called ""bill_adelie"" consisting of the bill lengths of the Adelie penguins belonging to the Adelie species.

Suppose we model the sequence of bill lengths as a sample of independent and identically distributed Gaussian random variables $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with a population mean μ and population standard deviation σ.

Now apply your function ""chi_square_test_one_sample_var"" to test the null hypothesis that the population standard deviation is 3 mm at a significance level of $\alpha = 0.1$.

2. Obstacles to valid scientific inference

(Q1)

The following concepts have been introduced in Lecture 21. For each of the following concepts, give

(A) an explanation of the concept and
(B) an example of a situation (real or hypothetical) where they create a barrier to drawing scientific conclusions based on data.

You are encouraged to discuss these concepts with your colleagues.

1. Measurement distortions
2. Selection bias
3. Confounding variables

3. Multivariate distributions and parameter estimation

Suppose that we have a sample of red-tailed hawks and we want to investigate the distribution of several features (Wing, Weight and Tail) of red-tailed hawks. We model the Wing, Weight and Tail with a multivariate Gaussian distribution. First, load the "Hawks" data frame from the "Stat2Data" library.

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

(Q1)

Now extract a subset of the data frame called ""hawks_rt"" that contains only the rows corresponding to hawks from the "Red-tailed" (RT) species and three columns - ""Wing"", ""Weight"", ""Tail"". Remove any rows of ""hawks_rt"" with missing values from one of the relevant columns.

(Q2)

Now, lets model the three features ""Wing"", ""Weight"" and ""Tail"" with a multivariate Gaussian distribution. Suppose that your data frame "hawks_rt" consists of a i.i.d. sample $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \Sigma)$. Here we model the three features ""Wing"", ""Weight"" and ""Tail"" with a multivariate Gaussian distribution with population mean μ and population covariance matrix Σ. Compute the minimum variance unbiased estimates (MVUE) of the μ and Σ.

4. Basic concepts in classification

In lecture 24, we introduced some concepts in classification. Try to refresh your memory of these concepts by explaining them.

(Q1) Write down your explanation of each of the following concepts. Give an example where appropriate.

1. A classification rule
2. A learning algorithm
3. Training data
4. Feature vector
5. Label

6. Expected test error
7. Train error
8. The train test split

5. The train test split

Suppose you want to build a classifier to predict whether a hawk belongs to either the “Sharp-shinned” or the “Cooper’s” species of hawks. The feature vector will be a four-dimensional row vector containing the weight, and the lengths of the wing, the tail and the hallux. The labels will be binary: 1 if the hawk is “Sharp-shinned” and 0 if the hawk belongs to “Cooper’s” species.

(Q1)

Begin by loading the “Hawks” data frame from the “Stat2Data” library. Now extract a subset of the data frame called ““hawks_total”” with five columns - ““Weight””, ““Wing””, ““Hallux””, ““Tail”” and ““Species””. The data frame should only include rows corresponding to hawks from either the “Sharp-shinned” (SS) or the “Cooper’s” (CH) species, and not the “Red-tailed” (RT) species. Convert the Species column to a binary variable with a 1 if the hawk belongs to the sharp-shinned species and 0 if the hawk belongs to Cooper’s species. Finally, remove any rows with missing values from one of the relevant columns.

(Q2)

Now implement a train test split for your ““hawks_total”” data frame. You should use 60% of your data within your training data and 40% in your test data. You should create a data frame consisting of training data called ““hawks_train”” and a data frame consisting of test data called ““hawks_test””. Display the number of rows in each data frame.

(Q3)

Next extract a data frame called ““hawks_train_x”” from your training data (from ““hawks_train””) containing the feature vectors and no labels. In addition, extract a vector called ““hawks_train_y”” consisting of labels from your training data. Similarly, create data frames called ““hawks_test_x”” and ““hawks_test_y”” corresponding to the feature vectors and labels within the test set, respectively.

(Q4)

Now let’s consider a very simple (and not very effective) classifier which entirely ignores the feature vectors. The classifier is defined as follows.

Let $\hat{y} \in \{0,1\}$ be a fixed value. For any input x, the output of classifier is always the fixed value $\hat{y}$. In other words, your classifier is of the form $\phi_{\hat{y}}(x) \equiv \hat{y}$ for all $x \in \mathbb{R}^4$ (note that $\hat{y}$ is fixed).

Based on the training data from the previous questions, your task is to choose a value $\hat{y}$ from $\{0,1\}$ such that the classifier has a smaller training error.

(Q5)

Next compute the train and test error of $\phi_{\hat{y}}$.

In general, $\phi_{\hat{y}}$ performs poorly, as it does not use any information of the feature vector. However, in the example, the train error and test error seems relatively low (much less than 50%). Try to explain why the errors are relatively low. In which cases we might have a error of $\phi_{\hat{y}}$ close to 50%?