Question Viewer

Correct answers are shown in the answer boxes and for the list options.

Area: B6 : Hypothesis test for the difference between two means
Application: Iris Petal Lengths

All versions

Question Background

This exercise is designed to be used with the statistical package R within RStudio. RStudio is on the lab computers (MA124) or you can download it to your own computer.

The data file for this question is a sample from a built-in R data set, **iris**. The data set contains the measurements, in centimeters, of the variables **Sepal_length**, **Sepal_width**, **Petal_length** and **Petal_width**, respectively, for flowers from each of 3 species of iris. The **Species** variable has the factor levels; Iris setosa (**1**), versicolor (**2**), and virginica (**3**).

Using the link below, save a text version (file extension .txt) of the datafile to a folder on your computer. Then open RStudio, set your working directory to the folder where you saved the data, and load the data using the **read.table()** function. Remember to assign the data to an R object (i.e. give your dataset a name e.g. **mydata**). Note: this datafile has a header in it, so you should set **header=TRUE**. Ensure you set the working directory to where you saved the file in order to read it in, using double "\\" and not single "\" in the filepath name. Further instructions on using R and loading in data can be found in the Introduction to R document on the resources page and in the course slides.

Download this data file: » text file

Tutors: Students should have the Intro2R.pdf document and their notes open in another browser window or tab before any further assistance is given.

Once you have loaded in the data, use the command **head(mydata)**. This allows you to see the variable names and the format of the data.

Editor: 1/21 AL to be checked.

Version 1: Assignment

Tutors: Randomised data sets

Editor: MD 1/21 solutions formatted and checked

Use the commands **virginica<-mydata$Sepal_Length[mydata$Species==3]** and **setosa<-mydata$Sepal_Length[mydata$Species==1]** to store the sepal lengths of the Virginica and Setosa species. In this question you will compare the sample means to estimate the true difference in sepal length means of these species.

a) If V represents "Virginica" and S represents "Setosa", which one of the following is the appropriate pair of hypotheses? 1

- $H_0 : \mu_V - \mu_S \neq 0; \quad H_A : \mu_V - \mu_S = 0$
- $H_0 : \bar{x}_V - \bar{x}_S = 0; \quad H_A : \bar{x}_V - \bar{x}_S \neq 0$
- $H_0 : \bar{x}_V = \bar{x}_S; \quad H_A : \bar{x}_V \neq \bar{x}_S$
- $H_0 : \mu_V \neq \mu_S; \quad H_A : \mu_V = \mu_S$
- (●) $H_0 : \mu_V = \mu_S; \quad H_A : \mu_V \neq \mu_S$

b) What is the standard error of the estimated difference in means? Hint: To calculate this, it might be helpful to use the commands length(virginica) and length(setosa) to know how many virginica and setosa observations there are in this sample. 0.1499 ≥ 4DP 1

Solution: The standard error is $\sqrt{\frac{s_S^2}{n_S} + \frac{s_V^2}{n_V}} = \sqrt{\frac{0.1327}{18} + \frac{0.3169}{21}}$

where s_S and s_V are the sample standard deviations for setosa and virginica species, and n_S and n_V are the number of setosa and virginica observations in the sample.

c) Using the order of differencing $(\bar{x}_V - \bar{x}_S)$, calculate the test statistic for this test. 11.019 ≥ 3DP 1

Solution: $\frac{(\bar{x}_V - \bar{x}_S) - 0}{SE} = \frac{6.6238 - 4.9722}{0.1499} = 11.019$

d) Using the estimated degrees of freedom, $\nu = 34.60$, the p-value is: 0.0000 ≥ 4DP 1

Solution: **2*(1-pt(11.019,34.60))** OR **2*pt(11.019,34.60,lower.tail=FALSE)**

General method for finding p-values using a t-distribution: Sketch the t-distribution and mark the mean (the mean of the t-distribution is always zero) add your test statistic. The p-value is the probability under the t-distribution of your test statistic *or a value more extreme*, ie away from the mean. You then need to double this probability to include the other tail (this is because prior to collecting your data it is possible that deviations from the mean could occur in either direction from the mean).

If your test statistic (TS) is *negative* use:
2*pt(TS, ν, lower.tail=TRUE)
equivalently
2*pt(TS, ν)
because lower.tail=TRUE is default and so does not need to be provided
If your test statistic is *positive* use:
2*pt(TS, ν, lower.tail=FALSE) or **2*(1-pt(TS, ν)**

Editor: With present data p-value will be 0 (to 4dp) for all datasets. In future versions want to perhaps tweak datasets so this does not occur - comparing Setosa and Versicolor might be better.

e) Assuming the necessary assumptions are satisfied, select the appropriate conclusion given the p-value you have calculated. 1

- The p-value is greater than or equal to 0.05. We therefore do not have sufficient evidence to reject H_0. The difference between the sample means is consistent with what we might expect to see from random variation under the null hypothesis.
- The p-value is greater than or equal to 0.05. We therefore have evidence to reject H_0 and can say that there is evidence of a difference in sepal length between the Virginica and Setosa species at the 5% level.

- (●) The *p*-value is less than 0.05. We therefore have evidence to reject H_0 and can say that there is evidence of a difference in sepal length between the Virginica and Setosa species at the 5% level.
- ( ) Because the *p*-value is smaller than 0.95 there will not be any significant result from this analysis.
- ( ) The *p*-value is less than 0.05. We therefore do not have sufficient evidence to reject H_0. The difference between the sample means is consistent with what we might expect to see from random variation under the null hypothesis.