AT-3

Koushik

2022-05-25

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

URL= “<https://raw.githubusercontent.com/markziemann/SLE712_files/master/assessment_task3/bioinfo_asst3_part1_files/gene_expression.tsv>” NAME= “gene\_expression.tsv” download.file(URL,destfile=NAME)

URL= “<https://raw.githubusercontent.com/markziemann/SLE712_files/master/assessment_task3/bioinfo_asst3_part1_files/growth_data.csv>” NAME= “growth\_data.csv” download.file(URL,destfile=NAME) ### HEAD

read.delim(“gene\_expression.tsv”) head(read.delim(“gene\_expression.tsv”)) str(read.delim(“gene\_expression.tsv”))

cds = read.delim(“gene\_expression.tsv”)

head(cds) row.names(cds) row.names(cds) = cds$Name\_Description row.names(cds) ##Change the name of the row from a number to the name of the gene

head(cds,6) ##Display the values of the first six data

head(rowMeans(cds[2:4])) cds$mean = rowMeans(cds[2:4]) head(cds,6) ##Add a column to show the average of the gene values and display the first six sets of data

order(-cdsmean),] sorted[,c(4,ncol(sorted))] head(sorted[,c(4,ncol(sorted))],10) ##List the 10 genes with the highest mean expression

subset(cds,mean < 10) nrow(subset(cds,mean < 10)) nrow(cds)

##List the number of rows with gene means less than 10 and calculate how many rows there are and what the total number of rows is

hist(cds$mean,xlab=“mean”, main=“Histogram of means”)

###The following is an analysis of the data ‘growth\_data.csv’

read.csv(“growth\_data.csv”)

cdss = read.csv(“growth\_data.csv”) head(cdss) colnames(cdss) ##Read the document,and displays the column name

cdss[cdss$Site=="northeast",] site1 = cdss[cdss$Site==“northeast”,] cdss[cdss$Site=="southwest",] site2 = cdss[cdss$Site==“southwest”,] mean(site1[,3]) sd(site1[,3]) mean(site1[,6]) sd(site1[,6]) mean(site2[,3]) sd(site2[,3]) mean(site2[,6]) sd(site2[,6]) ##site1 is all the data in the northeast location tree, site2 is the data in the southwest location tree, determine the data and calculate the standard deviation and mean respectively

boxplot(site1Circumf\_2020\_cm, ylab=“circumference”, main=“The northeast position starts and ends with the data”, names = c(“start”,“end”))

boxplot(site2Circumf\_2020\_cm, ylab=“circumference”, main=“The southwest position starts and ends with the data”, names = c(“start”,“end”)) ##The data has already been sorted in the previous step, so you just need to do the box plot in this step.

site1gr = mean(site1[,6])-mean(site1[,4]) site1gr/mean(site1[,4])

site2gr = mean(site2[,6])-mean(site2[,4]) site2gr/mean(site2[,4])

##Calculation of the growth rate ##step1: The present value is subtracted from the past value ##step2: the result is divided by the past value

myfunc1 = function(x,y){ t.test(x,y)

rest = t.test(x,y) PVAL = rest$p.value boxplot(x,y) HEADER=paste(“P:”,PVAL) mtext(HEADER) }

t.test(cdssCircumf\_2010\_cm~cdssCircumf\_2020\_cm - cdssSite,xlab = “location”,ylab = “growth”) PVAL = t.test(cdssCircumf\_2010\_cm~cdssp.value HEADER=paste(“P:”,PVAL) mtext(HEADER)

wilcox.test(cdssCircumf\_2010\_cm~cdssCircumf\_2020\_cm - cdssSite,xlab = “location”,ylab = “growth”) PVAL = wilcox.test(cdssCircumf\_2010\_cm~cdssp.value HEADER=paste(“P:”,PVAL) mtext(HEADER)

##The p-values are calculated using ttest and wilcox.test respectively, and it can be seen that the p-values calculated using ttest are smaller, and the differences in the data over 10 years are directly applied in the calculation, and no mean values are involved