Assignment 4

Mehnaz_meem

2025-10-05

##The file “gene_expression.tsv” contains RNA-seq count data for three samples of interest. ##Read in the file, making the gene identifiers the row names. Show a table of values for the first six genes.

```r
library("R.utils")
```

```
## Loading required package: R.oo

## Loading required package: R.methodsS3

## R.methodsS3 v1.8.2 (2022-06-13 22:00:14 UTC) successfully loaded. See ?R.methodsS3 for help.

## R.oo v1.27.1 (2025-05-02 21:00:05 UTC) successfully loaded. See ?R.oo for help.

##
## Attaching package: 'R.oo'

## The following object is masked from 'package:R.methodsS3':
##
##     throw

## The following objects are masked from 'package:methods':
##
##     getClasses, getMethods

## The following objects are masked from 'package:base':
##
##     attach, detach, load, save

## R.utils v2.13.0 (2025-02-24 21:20:02 UTC) successfully loaded. See ?R.utils for help.

##
## Attaching package: 'R.utils'

## The following object is masked from 'package:utils':
##
##     timestamp

## The following objects are masked from 'package:base':
##
##     cat, commandArgs, getOption, isOpen, nullfile, parse, warnings
```

```r
URL="https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/gene_expression.tsv"
download.file(URL,destfile="gene_expression.tsv")
```

```r
list.files()
```

```
##  [1] "ccoli_cds.fa"              "ccoli_cds.fa.gz"
##  [3] "ecoli_cds.fa"              "ecoli_cds.fa.gz"
##  [5] "gene_expression.tsv"       "growth_data.csv"
##  [7] "LICENSE"                   "MEHNAZ_MEEM_A4_part_1.pdf"
##  [9] "MEHNAZ_MEEM_A4_part_1.Rmd" "MEHNAZ_MEEM_A4_part_2.Rmd"
## [11] "MEHNAZ_RSTUDIO.Rproj"      "README.html"
## [13] "README.md"                 "week 8 test file.Rmd"
## [15] "week 8 test.Rmd"           "WEEK 8.Rmd"
## [17] "Week 9.Rmd"                "week-8-test-file.html"
## [19] "week-8-test.html"          "WEEK-8.html"
## [21] "Week-9.pdf"                "Week10.Rmd"
```

##The file “growth_data.csv” contains measurements for tree circumference growing at two sites

```
URL="https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/growth_data.csv"
download.file(URL,destfile="growth_data.csv")
```

```
list.files()
```

```
##  [1] "ccoli_cds.fa"              "ccoli_cds.fa.gz"
##  [3] "ecoli_cds.fa"              "ecoli_cds.fa.gz"
##  [5] "gene_expression.tsv"       "growth_data.csv"
##  [7] "LICENSE"                   "MEHNAZ_MEEM_A4_part_1.pdf"
##  [9] "MEHNAZ_MEEM_A4_part_1.Rmd" "MEHNAZ_MEEM_A4_part_2.Rmd"
## [11] "MEHNAZ_RSTUDIO.Rproj"      "README.html"
## [13] "README.md"                 "week 8 test file.Rmd"
## [15] "week 8 test.Rmd"           "WEEK 8.Rmd"
## [17] "Week 9.Rmd"                "week-8-test-file.html"
## [19] "week-8-test.html"          "WEEK-8.html"
## [21] "Week-9.pdf"                "Week10.Rmd"
```

##Question 1 ##Read the file data, first six genes

```
gene_data <- read.delim("gene_expression.tsv", row.names = 1, header = TRUE)
head(gene_data, 6)
```

```
##                                 GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000223972.5_DDX11L1                              0                        0
## ENSG00000227232.5_WASH7P                             187                      109
## ENSG00000278267.1_MIR6859-1                            0                        0
## ENSG00000243485.5_MIR1302-2HG                          1                        0
## ENSG00000237613.2_FAM138A                              0                        0
## ENSG00000268020.3_OR4G4P                               0                        1
##                                 GTEX.1117F.0526.SM.5EGHJ
## ENSG00000223972.5_DDX11L1                              0
## ENSG00000227232.5_WASH7P                             143
## ENSG00000278267.1_MIR6859-1                            1
## ENSG00000243485.5_MIR1302-2HG                          0
## ENSG00000237613.2_FAM138A                              0
## ENSG00000268020.3_OR4G4P                               0
```

##Question 2 ##New column with mean of other columns

```
gene_data$Mean <- rowMeans(gene_data)
```

##Showing table of values for first six genes

```
head(gene_data, 6)
```

```
##                                 GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000223972.5_DDX11L1                              0                        0
## ENSG00000227232.5_WASH7P                             187                      109
## ENSG00000278267.1_MIR6859-1                            0                        0
## ENSG00000243485.5_MIR1302-2HG                        1                        0
## ENSG00000237613.2_FAM138A                            0                        0
## ENSG00000268020.3_OR4G4P                             0                        1
##                               GTEX.1117F.0526.SM.5EGHJ        Mean
## ENSG00000223972.5_DDX11L1                            0   0.0000000
## ENSG00000227232.5_WASH7P                           143 146.3333333
## ENSG00000278267.1_MIR6859-1                          1   0.3333333
## ENSG00000243485.5_MIR1302-2HG                        0   0.3333333
## ENSG00000237613.2_FAM138A                            0   0.0000000
## ENSG00000268020.3_OR4G4P                             0   0.3333333
```

##Question 3 ##List the 10 genes with the highest mean expression

```
top10_genes <- gene_data[order(gene_data$Mean, decreasing = TRUE), ]
head(top10_genes, 10)
```

```
##                            GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000198804.2_MT-CO1                     267250                  1101779
## ENSG00000198886.2_MT-ND4                     273188                   991891
## ENSG00000198938.2_MT-CO3                     250277                  1041376
## ENSG00000198888.2_MT-ND1                     243853                   772966
## ENSG00000198899.2_MT-ATP6                    141374                   696715
## ENSG00000198727.2_MT-CYB                     127194                   638209
## ENSG00000198763.3_MT-ND2                     159303                   543786
## ENSG00000211445.11_GPX3                      464959                    39396
## ENSG00000198712.1_MT-CO2                     128858                   545360
## ENSG00000156508.17_EEF1A1                    317642                    39573
##                            GTEX.1117F.0526.SM.5EGHJ     Mean
## ENSG00000198804.2_MT-CO1                     218923 529317.3
## ENSG00000198886.2_MT-ND4                     277628 514235.7
## ENSG00000198938.2_MT-CO3                     223178 504943.7
## ENSG00000198888.2_MT-ND1                     194032 403617.0
## ENSG00000198899.2_MT-ATP6                    151166 329751.7
## ENSG00000198727.2_MT-CYB                     141359 302254.0
## ENSG00000198763.3_MT-ND2                     149564 284217.7
## ENSG00000211445.11_GPX3                      306070 270141.7
## ENSG00000198712.1_MT-CO2                     122816 265678.0
## ENSG00000156508.17_EEF1A1                    339347 232187.3
```

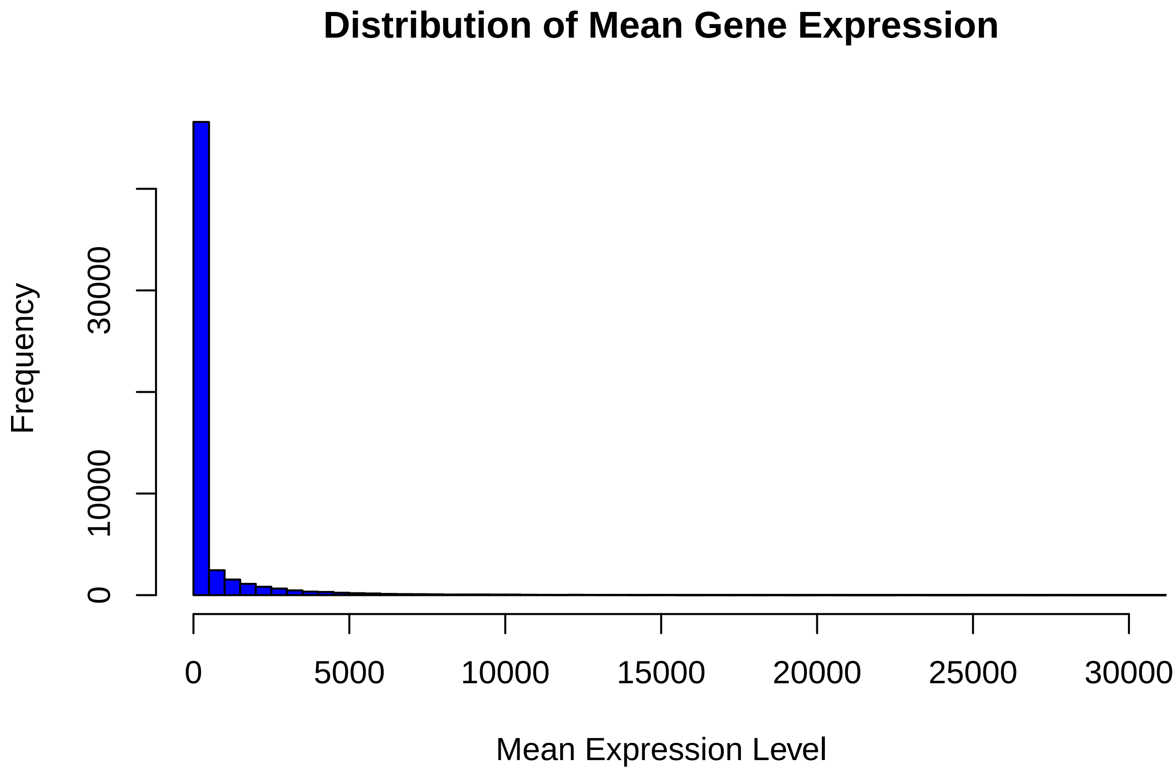
##Question 4 ##Determine the number of genes with a mean <10

```
num_genes_below_10 <- sum(gene_data$Mean < 10)
cat(num_genes_below_10)
```

```
## 35988
```

##Question 5 ##Make a histogram plot of the mean values

```
hist(gene_data$Mean,
breaks = 1000,
main = "Distribution of Mean Gene Expression",
xlab = "Mean Expression Level",
ylab = "Frequency",
xlim = c(0,30000),
col = "blue")
```



##Question 6 ##Import growth_data csv file into an R object

```
growth <- read.csv("growth_data.csv", header = TRUE)
```

##Column names of growth_data

```
colnames(growth)
```

```
## [1] "Site"             "TreeID"           "Circumf_2005_cm" "Circumf_2010_cm"
## [5] "Circumf_2015_cm" "Circumf_2020_cm"
```

##Question 7 ##Calculate the mean and standard deviation of tree circumference at the start and end of the study at both sites

```
# Mean and Standard Deviation (2005) & (2020)
growth_stats_raw <- aggregate(cbind(Circumf_2005_cm, Circumf_2020_cm)~ Site,
data = growth,
FUN = function(x) c(mean = mean(x, na.rm = TRUE),
sd = sd(x, na.rm = TRUE)))


# Split the matrix columns into separate columns
growth_stats_summary <- data.frame(
Site = growth_stats_raw$Site,
Mean_2005 = growth_stats_raw$Circumf_2005_cm[, "mean"],
SD_2005 = growth_stats_raw$Circumf_2005_cm[, "sd"],
Mean_2020 = growth_stats_raw$Circumf_2020_cm[, "mean"],
SD_2020 = growth_stats_raw$Circumf_2020_cm[, "sd"]
```

```r
)



# Display table results
print(growth_stats_summary)
```

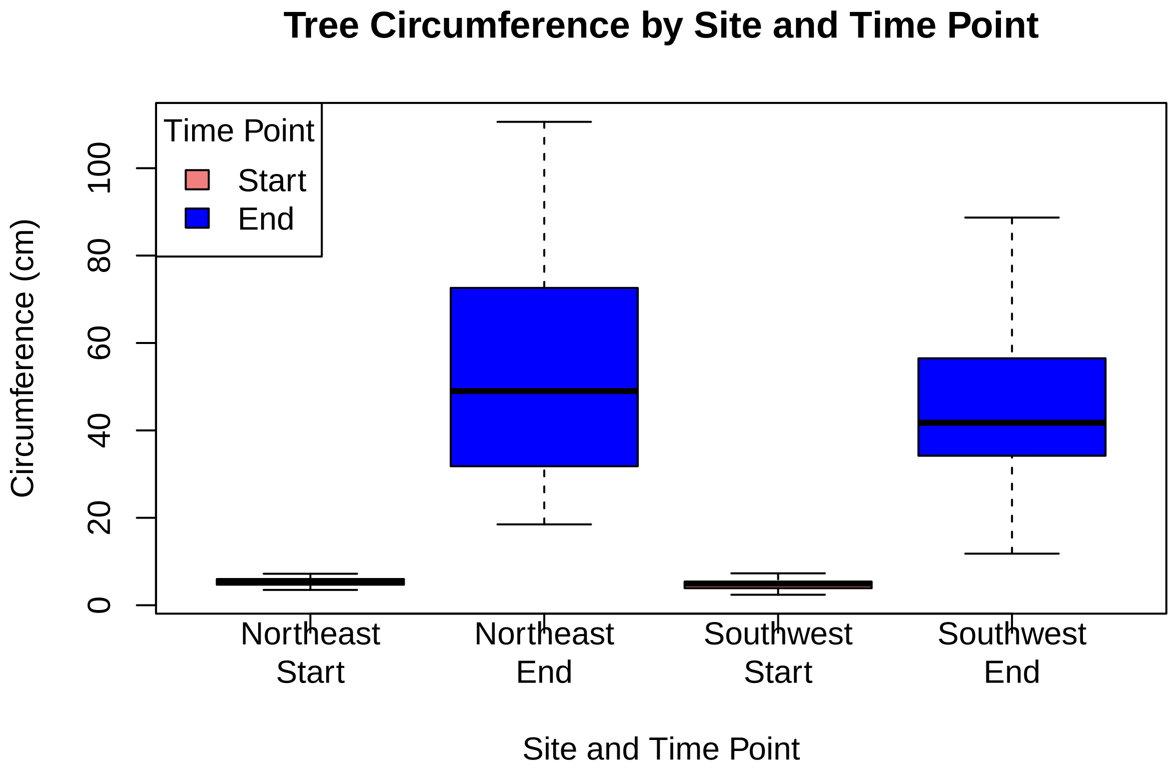
```
##        Site Mean_2005   SD_2005 Mean_2020  SD_2020
## 1 northeast     5.292 0.9140267    54.228 25.22795
## 2 southwest     4.862 1.1474710    45.596 17.87345
```

##Question 8 ##Make a box plot of tree circumference at the start and end of the study at both sites.

```r
# Prepare data for boxplot
northeast_start <- growth$Circumf_2005_cm[growth$Site== "northeast"]
northeast_end <- growth$Circumf_2020_cm[growth$Site== "northeast"]
southwest_start <- growth$Circumf_2005_cm[growth$Site== "southwest"]
southwest_end <- growth$Circumf_2020_cm[growth$Site== "southwest"]


# Create boxplot
boxplot(northeast_start, northeast_end, southwest_start, southwest_end,
names = c("Northeast\nStart", "Northeast\nEnd", "Southwest\nStart", "Southwest\nEnd"),
main = "Tree Circumference by Site and Time Point",
xlab = "Site and Time Point",
ylab = "Circumference (cm)",
col = c("lightcoral", "blue", "lightcoral", "blue"),
border = "black")


# Add a legend
legend("topleft",
legend = c("Start", "End"),
fill = c("lightcoral", "blue"),
title = "Time Point")
```



##Question 9 ##Calculate the mean growth over the last 10 years at each site.

```
# Growth from 2010 to 2020
growth$Growth_10yr <- growth$Circumf_2020_cm- growth$Circumf_2010_cm


# Mean 10 year growth for each site
mean_growth_by_site <- aggregate(Growth_10yr~ Site, data = growth, FUN = mean)


# Show results
print(mean_growth_by_site)
```

```
##        Site Growth_10yr
## 1 northeast       42.94
## 2 southwest       35.49
```

##Question 10 ##Using T-test to estimate the growth beteen two sites

```
t_test_result <- t.test(Growth_10yr~ Site, data = growth)

## Print t-test results with interpretation
cat("\n T-Test: Comparing 10-Year Growth Between Sites\n")
```

```
##
##  T-Test: Comparing 10-Year Growth Between Sites
```

```r
cat("--------------------------------------------------\n")
```

```
## --------------------------------------------------
```

```r
cat("P-value:", round(t_test_result$p.value, 4), "\n")
```

```
## P-value: 0.0623
```

```r
cat("95% Confidence Interval:",
round(t_test_result$conf.int[1], 2), "to",
round(t_test_result$conf.int[2], 2), "\n")
```

```
## 95% Confidence Interval: -0.39 to 15.29
```

```r
cat("Mean Growth - Northeast:", round(t_test_result$estimate["mean in group Northeast"], 2), "cm\n")
```

```
## Mean Growth - Northeast: NA cm
```

```r
cat("Mean Growth - Southwest:", round(t_test_result$estimate["mean in group Southwest"], 2), "cm\n")
```

```
## Mean Growth - Southwest: NA cm
```

```r
# Corrected R Code for Interpretation Block
if (t_test_result$p.value < 0.05) {
  cat("Conclusion: There is a statistically significant difference in 10-year growth between sites (p <
} else {
  # Replacing the Unicode character (\u2265) with the ASCII text "> ="
  cat("Conclusion: No statistically significant difference in 10-year growth between sites (p >= 0.05).\
}
```

```
## Conclusion: No statistically significant difference in 10-year growth between sites (p >= 0.05).
```