

# Assignment-4 part-1

Teja Sri

2025-09-19

```
# Gene expression file
url_gene <- "https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/gene_expression.tsv"
download.file(url_gene, destfile = "gene_expression.tsv", mode = "wb")
gene <- read.delim("gene_expression.tsv", row.names = 1, check.names = FALSE)
head(gene)
```

```
##                                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
```

```
# Growth data file
url_growth <- "https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/growth_data.csv"
download.file(url_growth, destfile = "growth_data.csv", mode = "wb")
growth <- read.csv("growth_data.csv", check.names = FALSE)
head(growth)
```

```
##      Site TreeID Circumf_2005_cm Circumf_2010_cm Circumf_2015_cm
## 1 northeast  A012             5.2             10.1             19.9
## 2 southwest  A039             4.9              9.6             18.9
## 3 southwest  A010             3.7              7.3             14.3
## 4 northeast  A087             3.8              6.5             10.9
## 5 southwest  A074             3.8              6.4             10.9
## 6 northeast  A008             5.9             10.0             16.8
##      Circumf_2020_cm
## 1                 38.9
## 2                 37.0
## 3                 28.1
## 4                 18.5
## 5                 18.4
## 6                 28.4
```

```
## 1. Show first six genes
# Display the first six rows of the dataset (gene expression counts).
```

```
head(gene)
```

```
##                                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
```

```
## 2. Add mean column
```

```
# Calculate the mean expression across the three samples for each gene.
```

```
# Store it in a new column called "mean_expr".
```

```
gene$mean_expr <- rowMeans(gene)
```

```
# Show the first six rows again, now including the new column.
```

```
head(gene)
```

```
##                                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_expr
## 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
```

```
## 3. Top 10 genes by mean expression
```

```
# Order the genes by descending mean expression and select the top 10.
```

```
top10 <- head(gene[order(-gene$mean_expr), ], 10)
```

```
top10
```

```
##                                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_expr
## 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
```

```
## 4. Number of genes with mean < 10
# Count how many genes have a mean expression value less than 10.
n_low <- sum(gene$mean_expr < 10)
n_low
```

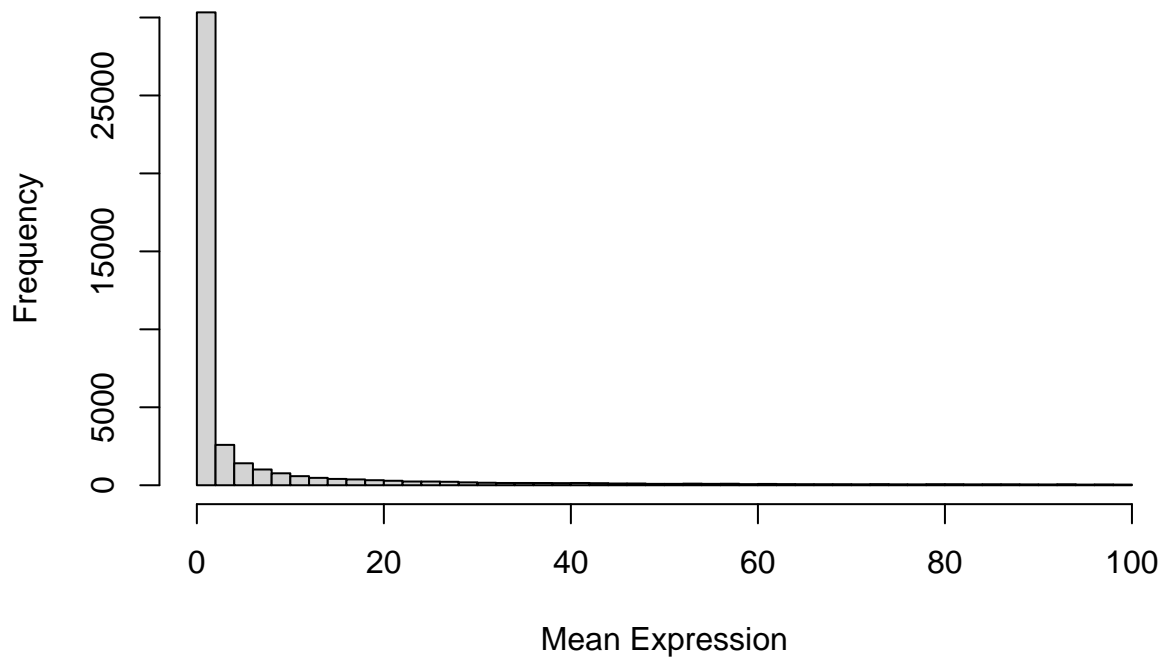
```
## [1] 35988
```

```
## 5. Histogram of mean values
# Plot a histogram to visualize the distribution of mean gene expression values.
# Compute mean expression per gene
mean_expression <- rowMeans(gene)
```

```
# Subset values below 100
mean_expression_sub <- mean_expression[mean_expression < 100]
```

```
# Plot histogram
hist(mean_expression_sub,
     breaks = 50,
     main = "Histogram of mean gene expression (<100)",
     xlab = "Mean Expression",
     col = "lightgray", border = "black")
```

## Histogram of mean gene expression (<100)



```
# Inspect column names
```

```
colnames(growth)
```

```
## [1] "Site"          "TreeID"         "Circumf_2005_cm" "Circumf_2010_cm"
```

```
## [5] "Circumf_2015_cm" "Circumf_2020_cm"
```

```
# Define Start and End explicitly from the available columns
```

```
growth$Start <- growth$Circumf_2005_cm
```

```
growth$End   <- growth$Circumf_2020_cm
```

```
## 1) Column names (already printed by colnames)
```

```
# colnames(growth)
```

```
gene <- read.delim("gene_expression.tsv", row.names = 1, check.names = FALSE)
```

```
head(gene)
```

```
##                               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

## --- Prepare convenient Start/End columns (from the year columns) ---
growth$Start <- growth$Circumf_2005_cm
growth$End   <- growth$Circumf_2020_cm

## --- 2) Mean and SD at Start and End by Site (PRINT BOTH) ---
by_start <- aggregate(Start ~ Site, data = growth,
                      FUN = function(x) c(mean = mean(x, na.rm = TRUE),
                                           sd   = sd(x,   na.rm = TRUE)))
by_end   <- aggregate(End ~ Site, data = growth,
                      FUN = function(x) c(mean = mean(x, na.rm = TRUE),
                                           sd   = sd(x,   na.rm = TRUE)))

start_stats <- cbind(Site = by_start$Site, as.data.frame(by_start$Start))
end_stats   <- cbind(Site = by_end$Site,   as.data.frame(by_end$End))
colnames(start_stats)[2:3] <- c("mean_start", "sd_start")
colnames(end_stats)[2:3]   <- c("mean_end",   "sd_end")

start_stats  # <-- keep this printed

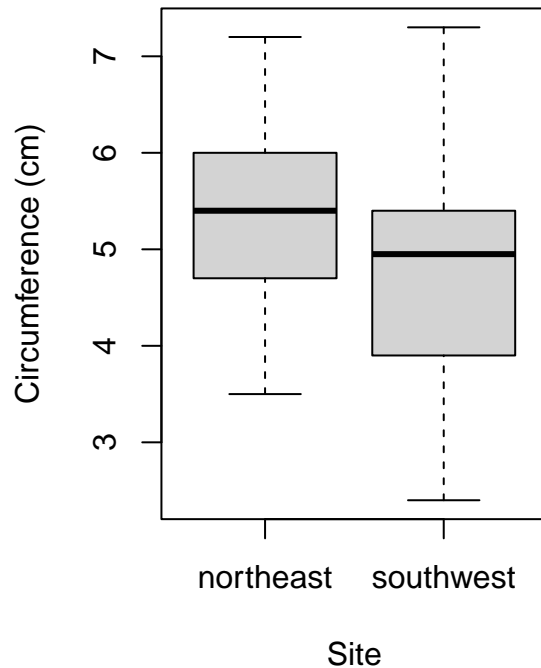
##           Site mean_start sd_start
## 1 northeast      5.292 0.9140267
## 2 southwest      4.862 1.1474710
end_stats    # <-- and this printed

##           Site mean_end sd_end
## 1 northeast    54.228 25.22795
## 2 southwest    45.596 17.87345

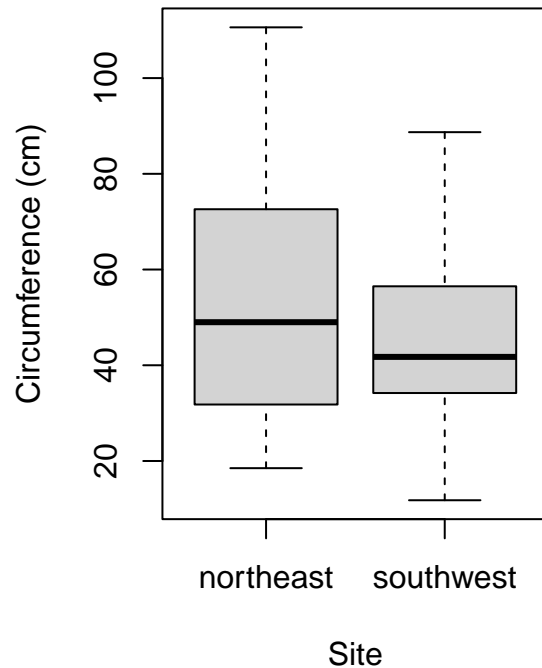
## --- 3) Boxplots (Start vs End by Site) ---
par(mfrow = c(1,2))
boxplot(Start ~ Site, data = growth,
        main = "Tree circumference at Start (2005)", ylab = "Circumference (cm)")
boxplot(End   ~ Site, data = growth,
        main = "Tree circumference at End (2020)",  ylab = "Circumference (cm)")

```

**Tree circumference at Start (200!**



**Tree circumference at End (2020**



```
par(mfrow = c(1,1))
```

```
## --- 4) Mean growth over the LAST 10 YEARS (2010 -> 2020) ---
growth$growth10 <- growth$Circumf_2020_cm - growth$Circumf_2010_cm
```

```
## site means (and SD if you want)
tapply(growth$growth10, growth$Site, mean, na.rm = TRUE)
```

```
## northeast southwest
##    42.94    35.49
```

```
# If you also want SD:
tapply(growth$growth10, growth$Site, sd, na.rm = TRUE)
```

```
## northeast southwest
##    22.81510    16.05704
```

```
## --- 5) t-test: is 10-year growth different between sites? ---
t.test(growth$growth10 ~ growth$Site)
```

```
##
## Welch Two Sample t-test
##
## data: growth$growth10 by growth$Site
## t = 1.8882, df = 87.978, p-value = 0.06229
## alternative hypothesis: true difference in means between group northeast and group southwest is not 0
## 95 percent confidence interval:
## -0.3909251 15.2909251
```

```
## sample estimates:
## mean in group northeast mean in group southwest
##           42.94           35.49
```