HW5 : Distribution

Ebenezer Nkum* University of Texas at El Paso (UTEP)

April 12, 2023

Contents

---

1 Load data

```
# Load data
dat <- read.csv("serialdat.csv", header = T)
```

2 Data

The data contains information about gene variant transcriptions. There are three replications of the variant transcriptions and a final column where the three replications were averaged. The categorical variables included are the SUMOver (which is classes of genes - S#V# format).

3 Step 1: Inspect data, assess it for completeness, good formatting, and any errors

3.1 Inspect data

```
##     SUMOvar X10.x.copies Replicate.1 Replicate.2 Replicate.3 Average.Cq
## 1 S1V1-10^6            6    16.27132    16.19231    16.36603   16.27655
```

*enkum@miners.utep.edu

```
## 2 S1V1-10^5            5    20.14263    20.12184    20.05466    20.10638
## 3 S1V1-10^4            4    23.07819    23.10269    22.86079    23.01389
## 4 S1V1-10^3            3    25.53921    25.51511    25.41548    25.48993
## 5 S1V1-10^2            2    26.05758    25.99988    26.04024    26.03257
## 6 S1V1-10^1            1    26.23620    26.03428    26.19077    26.15375
```

```
## [1] 43  6
```

3.2 Remove the replication number that follows the S#V# groups

```
dat0 <- dat[-43,] %>% select(-X10.x.copies) %>%
                mutate(SUMOvar = sapply(str_split(SUMOvar,"-",),'[',1))

dat0 <- dat0 %>% mutate(SUMOvar = substr(SUMOvar,1,2))


head(dat0)
```

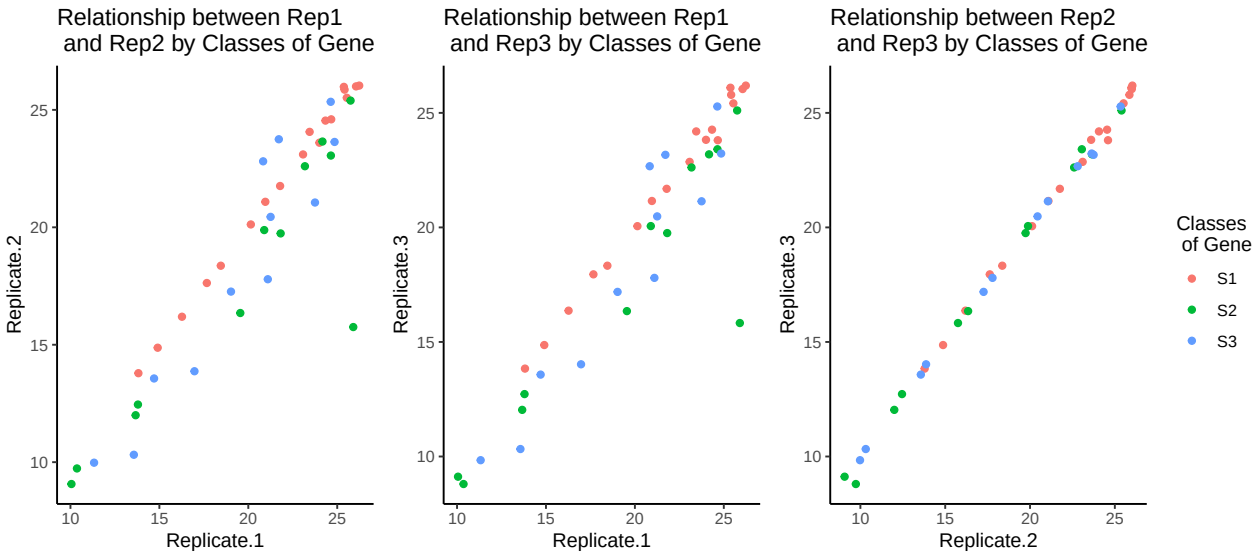
```
##   SUMOvar Replicate.1 Replicate.2 Replicate.3 Average.Cq
## 1      S1    16.27132    16.19231    16.36603   16.27655
## 2      S1    20.14263    20.12184    20.05466   20.10638
## 3      S1    23.07819    23.10269    22.86079   23.01389
## 4      S1    25.53921    25.51511    25.41548   25.48993
## 5      S1    26.05758    25.99988    26.04024   26.03257
## 6      S1    26.23620    26.03428    26.19077   26.15375
```

4 Visualizing distribution

5 Overall distribution among each group using the average replications

```
par(mfrow=c(2,2))
p <- ggplot(dat0, aes(x= Replicate.1, y=Replicate.2, color = SUMOvar)) +
      geom_point()+
      theme_classic() +
      ggtitle("Relationship between Rep1\n and Rep2 by Classes of Gene")+
      theme(legend.position = "none")



p1 <- ggplot(dat0, aes(x= Replicate.1, y=Replicate.3, color = SUMOvar)) +
      geom_point()+
      theme_classic() +
      ggtitle("Relationship between Rep1\n and Rep3 by Classes of Gene")+
      theme(legend.position = "none")
```

```
p2 <- ggplot(dat0, aes(x= Replicate.2, y=Replicate.3, color = SUMOvar)) +
      geom_point()+
      theme_classic() +
      ggtitle("Relationship between Rep2\n and Rep3 by Classes of Gene")

p2$labels$colour<-"Classes \n of Gene"

p+p1+p2
```



Upon examining the associations between the three transcription entities, it is apparent that there exists a significant degree of variability amongst the s2 and s3 gene classes in relation to the association between replication 1 and 2. A similar observation can be made regarding the correlation between replication 1 and replication 3. However, it is worth noting that there is very minimal variability or nearly perfect association with regards to the association between replication 2 and replication 3.