HW5 : Distribution

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

March 28, 2023

Contents

---

1 Load data

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

2 Data

The new data (serialdat.csv), contains information about gene variant transcriptions. There were three replications of the variant transcriptions and a final column where the three replications

[*] enkum@miners.utep.edu

were averaged. The two categorial varialbes included are the SUMOver (which is classes of genes - S#V# format). Ignore the replication number that follows the S#V# groups (of which there are 6 groups). Try to visualize the distributions according to each group. You may also include visualizations across the replications to see if the transcription process introduced error overall.

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
## 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

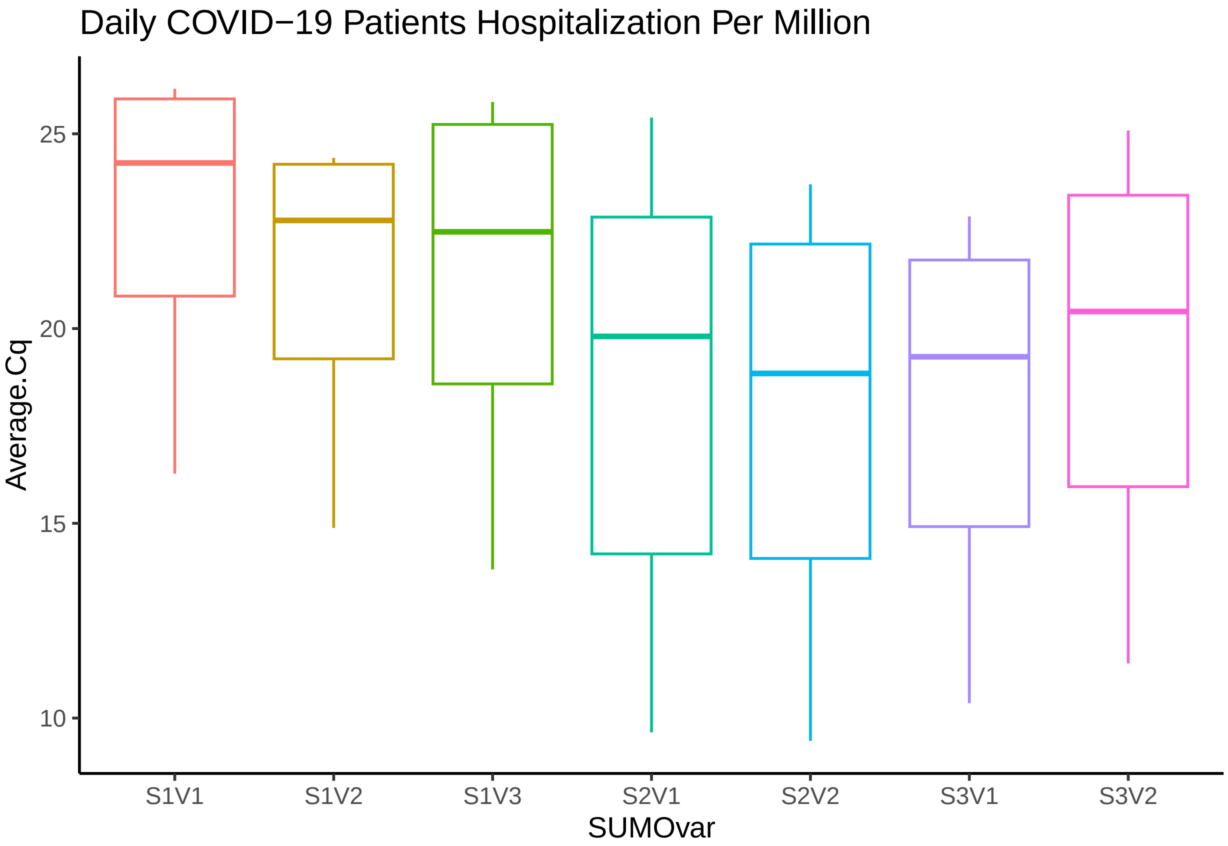
```
dat0 <- dat[-43,] %>% select(-X10.x.copies) %>%
                mutate(SUMOvar = sapply(str_split(SUMOvar,"-",),'[',1))

head(dat0)
```

```
##   SUMOvar Replicate.1 Replicate.2 Replicate.3 Average.Cq
## 1    S1V1    16.27132    16.19231    16.36603   16.27655
## 2    S1V1    20.14263    20.12184    20.05466   20.10638
## 3    S1V1    23.07819    23.10269    22.86079   23.01389
## 4    S1V1    25.53921    25.51511    25.41548   25.48993
## 5    S1V1    26.05758    25.99988    26.04024   26.03257
## 6    S1V1    26.23620    26.03428    26.19077   26.15375
```

4 Visualizing distribution

5 Overall distribution among each group using the average replications

```
ggplot(dat0, aes(x= SUMOvar, y=Average.Cq, color = SUMOvar)) +
      geom_boxplot()+
     theme_classic() +
    ggtitle("Daily COVID-19 Patients Hospitalization Per Million")+
     theme(legend.position = "none")
```

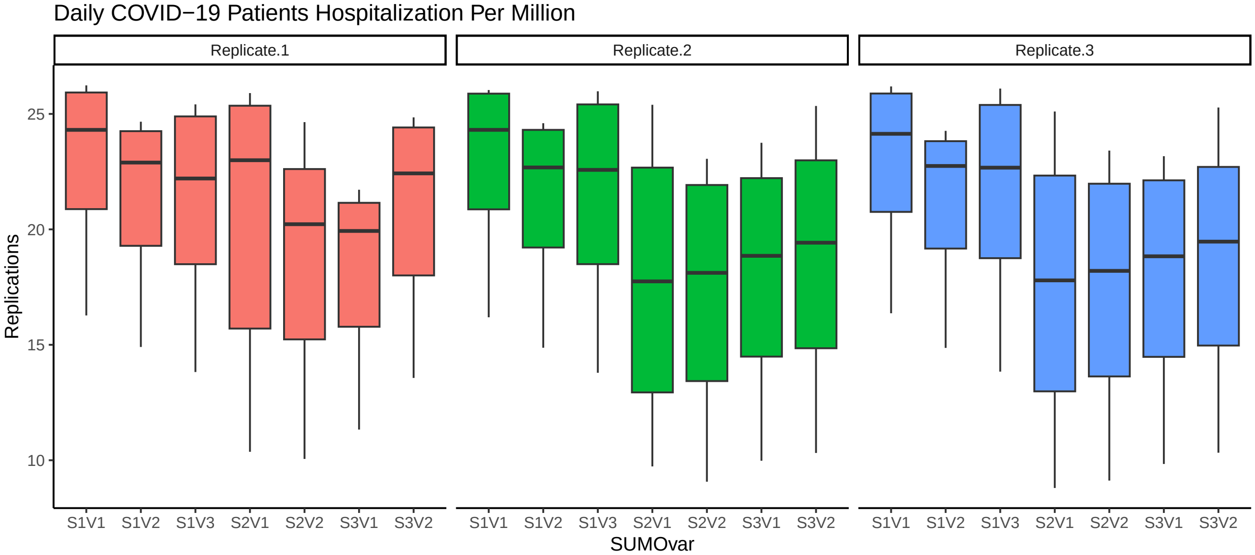


Here see that, the gene class S2V1 has more variability than all of the other classes. And more than half of its average replications lies below 19. The highest median average replication appears under the gene class S1V1. It has a median average replication of 24.

```
dat_reshape <- melt(dat0[,-5], id = "SUMOvar")

colnames(dat_reshape) <- c("SUMOvar","variable", "Replications")
```

6 Distributions by each group

```
ggplot(dat_reshape, aes(x=SUMOvar, y= Replications, fill = variable )) +
            geom_boxplot() +
            theme_classic() +
            ggtitle("Daily COVID-19 Patients Hospitalization Per Million")+
            theme(legend.position = "none")+
            facet_wrap(~variable)
```



We observe that under each replication, S2V1 has more variability than rest of the gene classes. And also, S1V1 has the highest median average replications in all the three replications.