Homework5

Isaiah Thompson Ocansey

2023-03-29

```
df<-read.csv(file="serialdat.csv",sep=",", header = TRUE)
dim(df);head(df)
```

```
## [1] 42  6
```

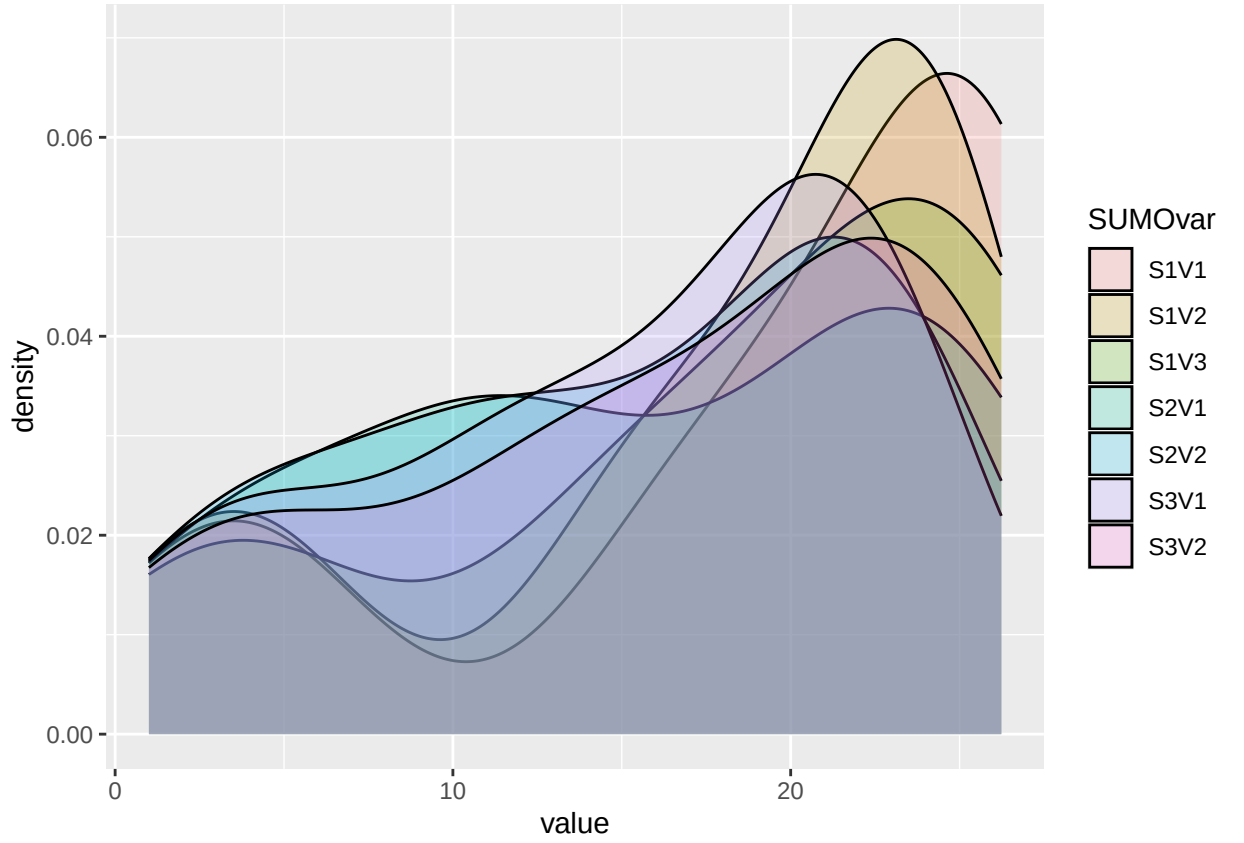
```
##   SUMOvar X10.x.copies Replicate.1 Replicate.2 Replicate.3 Average.Cq
## 1    S1V1            6    16.27132    16.19231    16.36603   16.27655
## 2    S1V1            5    20.14263    20.12184    20.05466   20.10638
## 3    S1V1            4    23.07819    23.10269    22.86079   23.01389
## 4    S1V1            3    25.53921    25.51511    25.41548   25.48993
## 5    S1V1            2    26.05758    25.99988    26.04024   26.03257
## 6    S1V1            1    26.23620    26.03428    26.19077   26.15375
```

The data contains information about gene variant transcriptions.There were three replications of the variant transcriptions and a final column where the three replications were averaged. The categorical variable included is SUMOver-this has seven classes of genes labelled in the format S1V1, S1V2, S1V3, S2V1, S2V2, S3V1, S3V2.In all, the data has 42 observations and 6 variables.

```
library(reshape)
library(ggplot2)
df1<-melt(df)
```

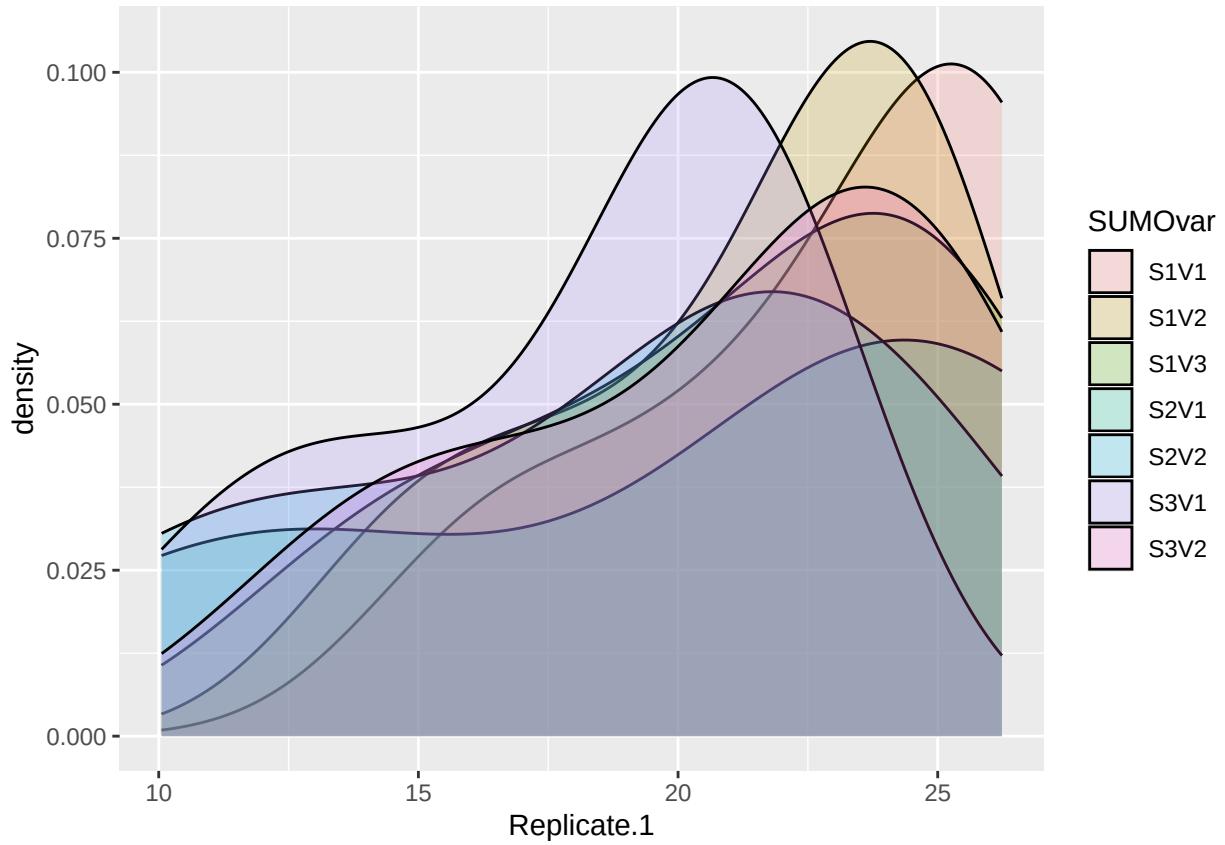
```
## Using SUMOvar as id variables
```

```
ggplot(df1, aes(x=value,fill = SUMOvar))+geom_density(alpha=0.2)
```

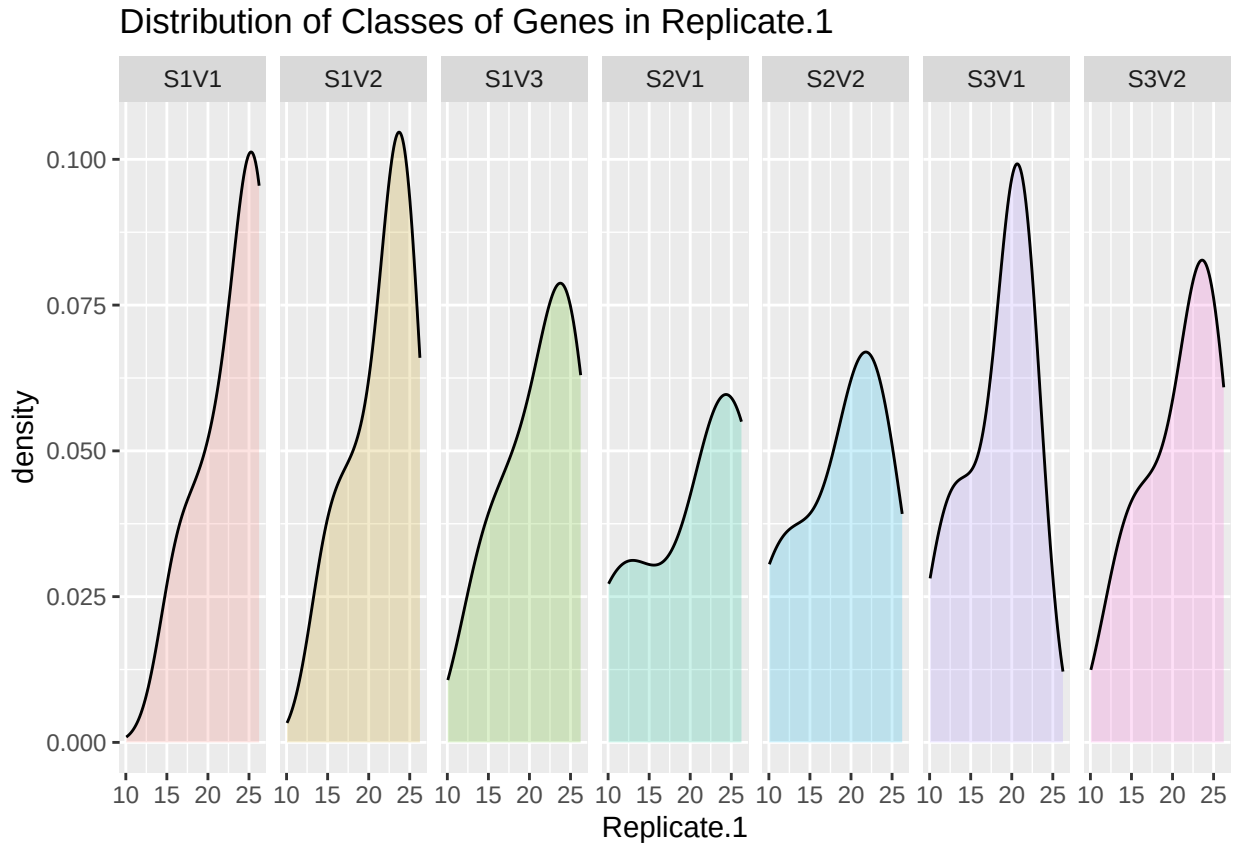


The above is the visualization of the distributions according to each class of gene. We can observe that the distribution of each class is negatively skewed which indicates deviation from normality.

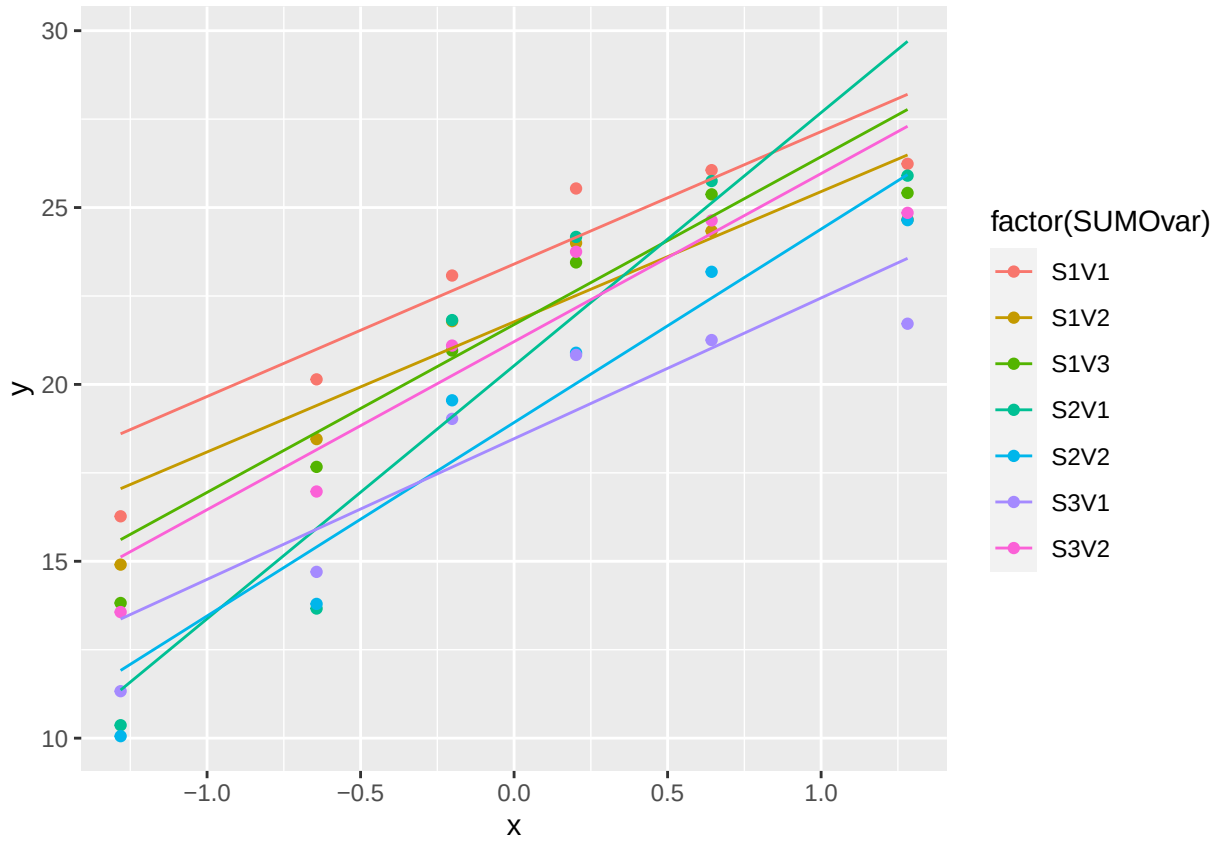
```
library(patchwork)
library(ggplot2)

par(mfrow=c(3,1))
ggplot(data=df, aes(x=Replicate.1,fill = SUMOvar)) + geom_density(alpha=0.2)
```




```
p1<-ggplot(data=df, aes(x=Replicate.1,fill = SUMOvar)) + geom_density(alpha=0.2)
p2<-ggplot(data=df, aes(x=Replicate.2,fill = SUMOvar)) + geom_density(alpha=0.2)
p3<-ggplot(data=df, aes(x=Replicate.3,fill = SUMOvar)) + geom_density(alpha=0.2)
p1 + facet_wrap( ~ SUMOvar , nrow = 1) + theme(legend.position = "none") +
  ggtitle("Distribution of Classes of Genes in Replicate.1 ")
```

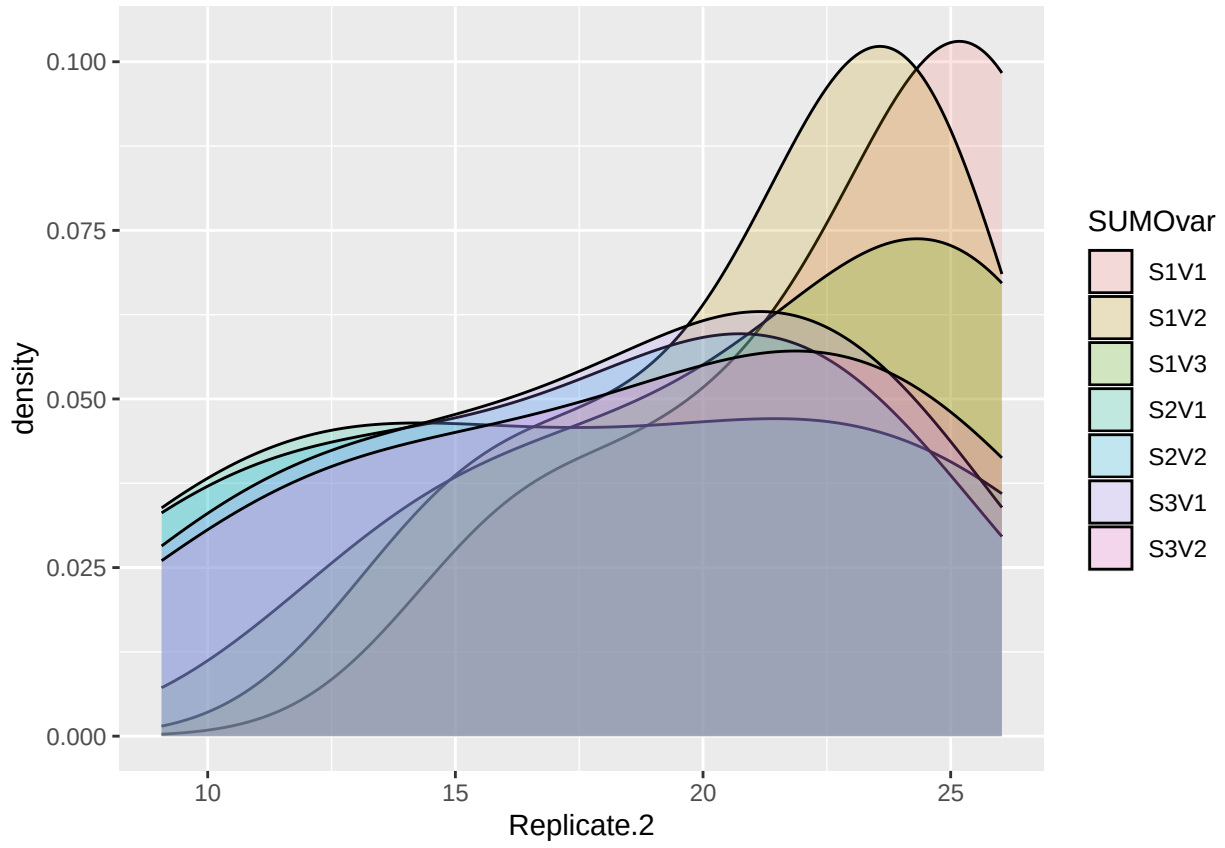



```
ggplot(df, aes(sample = Replicate.1, colour = factor(SUMOvar))) +
  stat_qq() +
  stat_qq_line()
```

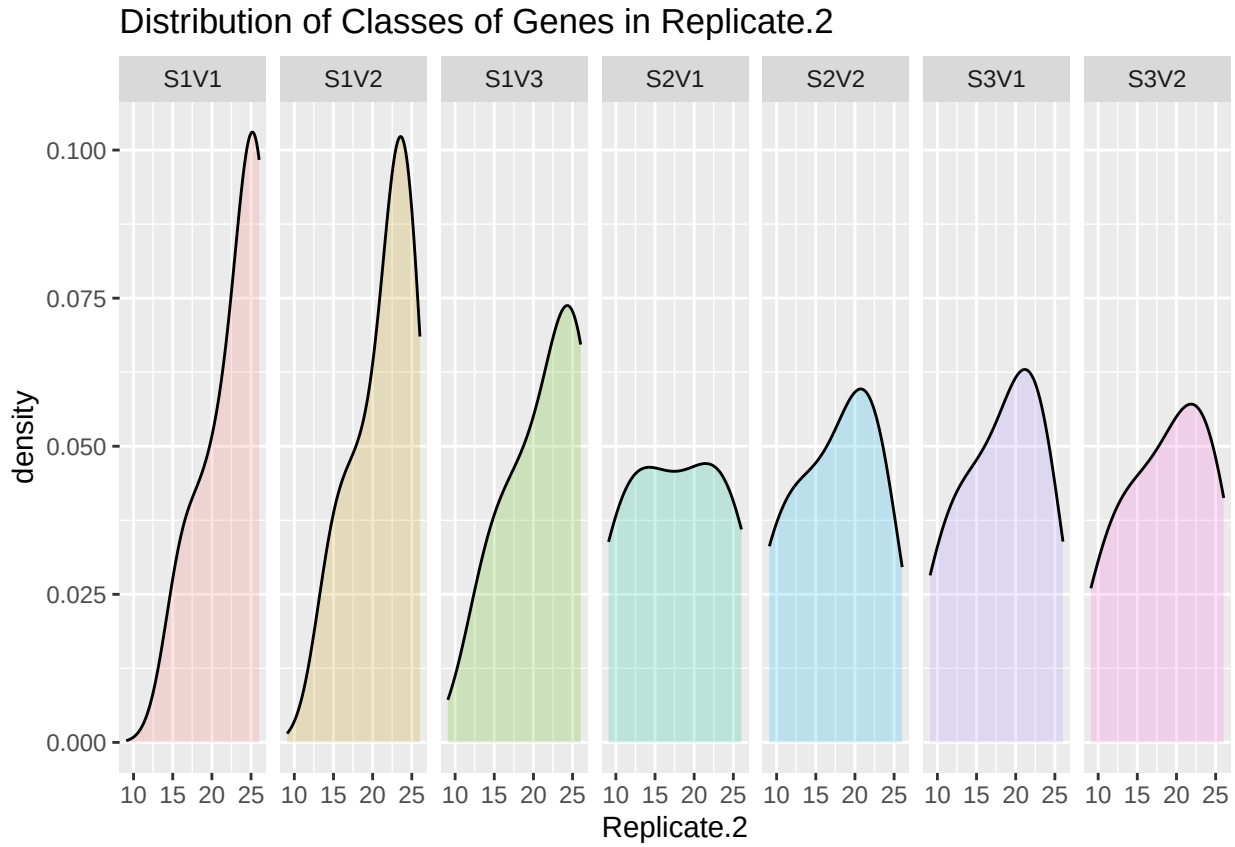


The above is the visualizations of each group across replication 1 in order to see if the transcription process introduced error.

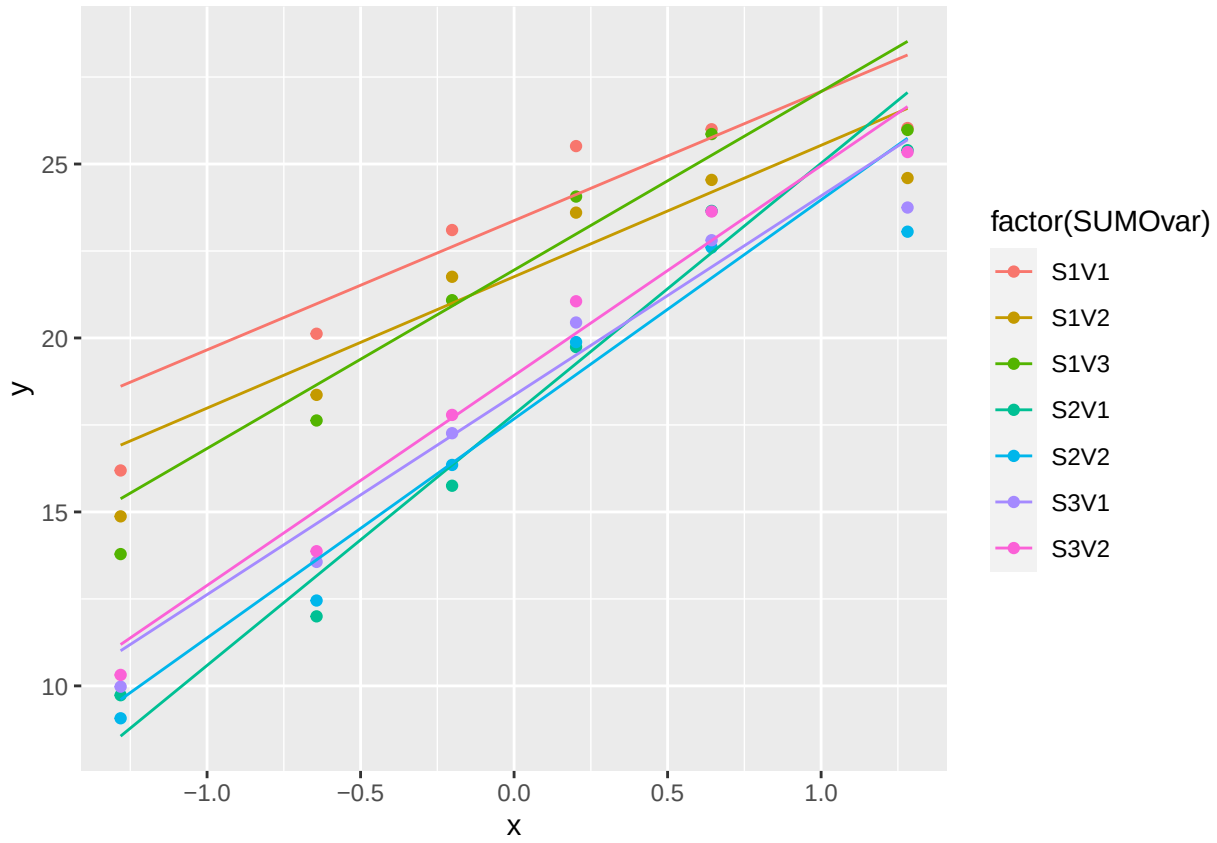
```
par(mfrow=c(3,1))
ggplot(data=df, aes(x=Replicate.2,fill = SUMOvar)) + geom_density(alpha=0.2)
```




```
p2 + facet_wrap( ~ SUMOvar , nrow = 1) + theme(legend.position = "none") +
  ggtitle("Distribution of Classes of Genes in Replicate.2 ")
```

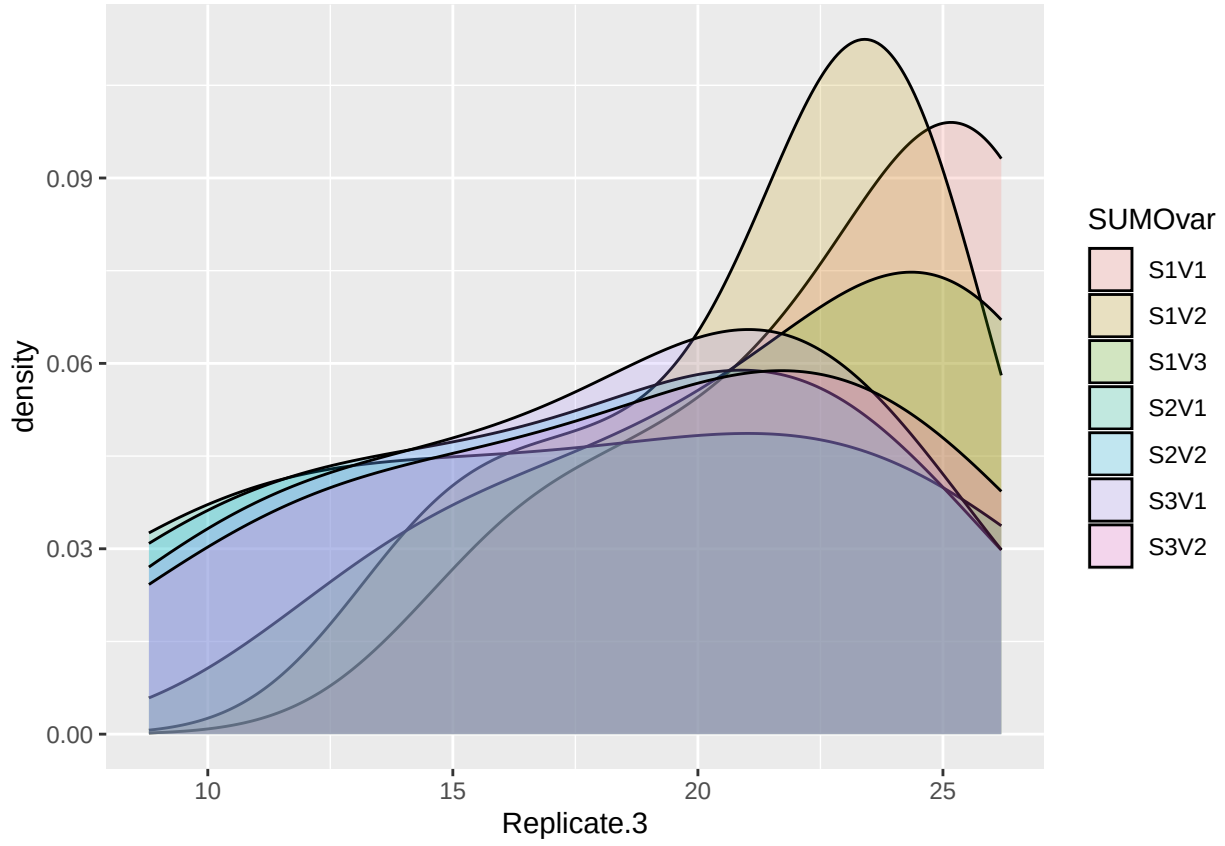



```
ggplot(df, aes(sample = Replicate.2, colour = factor(SUMOvar))) +
  stat_qq() +
  stat_qq_line()
```

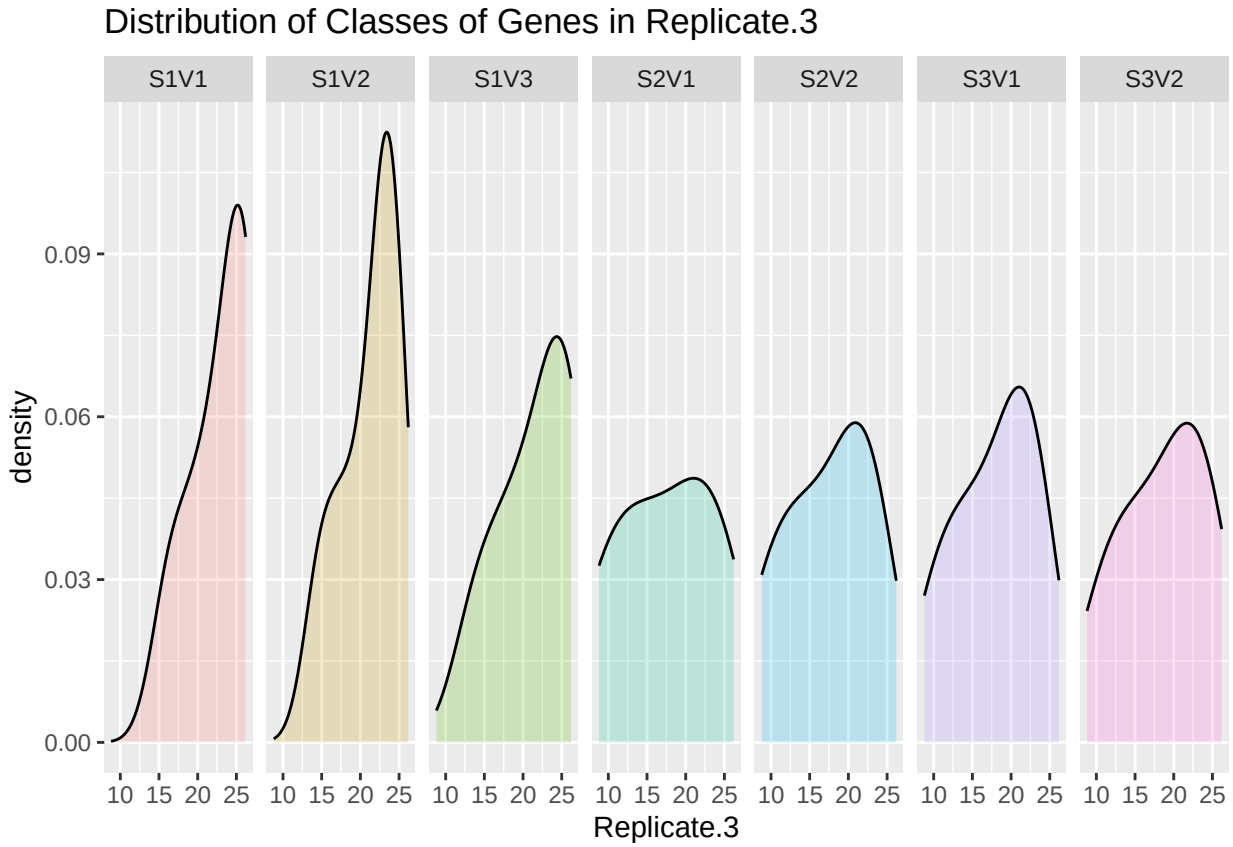


The above is the visualizations of each group across replication 2.

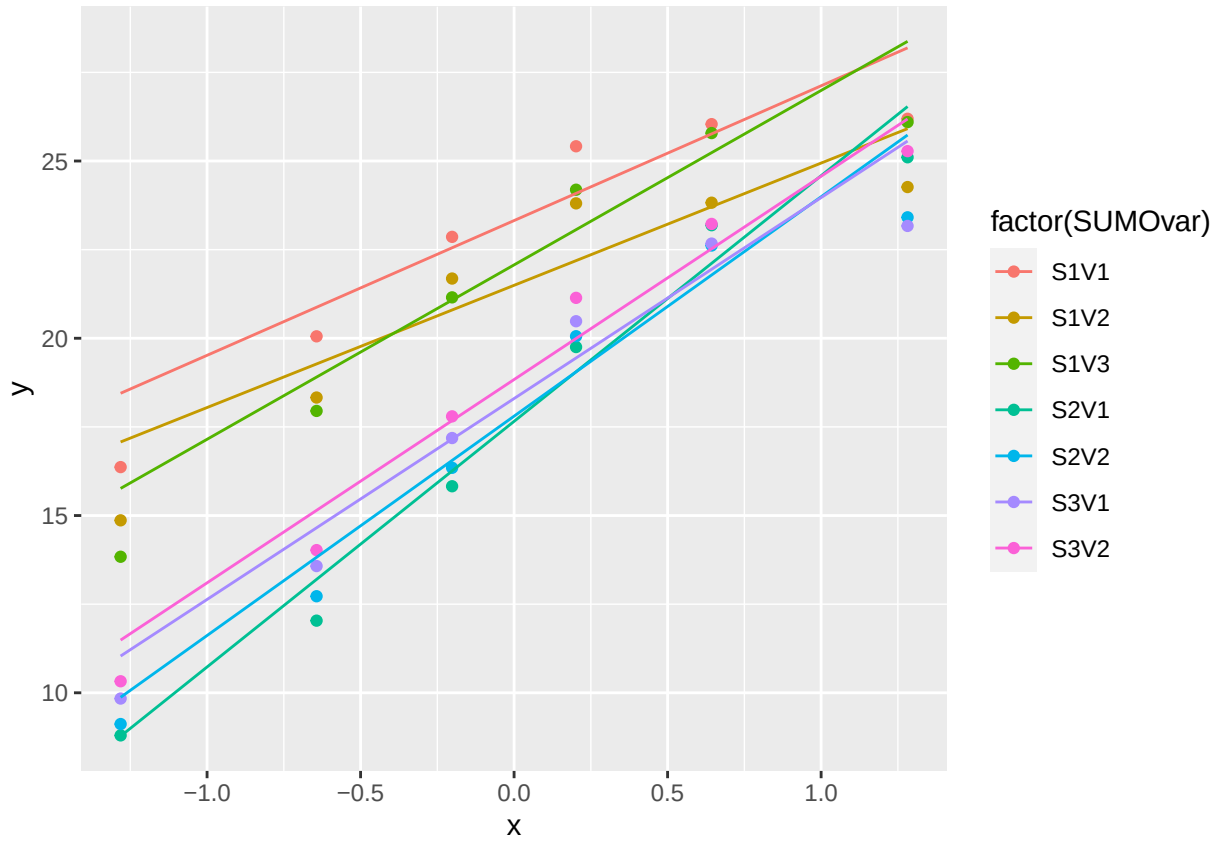
```
par(mfrow=c(3,1))
ggplot(data=df, aes(x=Replicate.3,fill = SUMOvar)) + geom_density(alpha=0.2)
```




```
p3 + facet_wrap( ~ SUMOvar , nrow = 1) + theme(legend.position = "none") +
  ggtitle("Distribution of Classes of Genes in Replicate.3 ")
```




```
ggplot(df, aes(sample = Replicate.3, colour = factor(SUMOvar))) +
  stat_qq() +
  stat_qq_line()
```



The above is the visualizations of each group across replication 3

CONCLUSION *We can observed that the distributions of the classes of gene across the three replicates (i.e Replicate.1, Replicate.2, Replicate.3) are negatively skewed which shows deviation from normality and this is confirmed using the qqplot in each replicate to test normality of the distributions. From the above visualizations of the distributions, we can say that the transcription process introduced error overall*