Biostatistics Project

Meetakshi Setiya, 2019253

```
cd ~/Documents/R/Project/
ls
```

```
## Code.Rmd
## Code.nb.html
## Code.pdf
## CodeToGenerateDatasets.R
## T0_data.csv
## TC1_data.csv
## TC2_data.csv
## TC3_data.csv
## TC4_data.csv
## TC5_data.csv
## untitled folder
```

```
library(hash)
```

```
t0_data <- read.csv("T0_data.csv")
tc1_data <- read.csv("TC1_data.csv")
tc2_data <- read.csv("TC2_data.csv")
tc3_data <- read.csv("TC3_data.csv")
tc4_data <- read.csv("TC4_data.csv")
tc5_data <- read.csv("TC5_data.csv")
colnames(t0_data) <- c('Genes','a','b','c','d')
colnames(tc1_data) <- c('Genes','a','b','c','d')
colnames(tc2_data) <- c('Genes','a','b','c','d')
colnames(tc3_data) <- c('Genes','a','b','c','d')
colnames(tc4_data) <- c('Genes','a','b','c','d')
colnames(tc5_data) <- c('Genes','a','b','c','d')
chemicals = list(t0_data, tc1_data, tc2_data, tc3_data, tc4_data, tc5_data)
```

I have assumed that the files are structured this way: 4 samples of the same gene were taken for each gene and they together made 4 groups a, b, c and d. This was done to ensure that the consistent results are obtained all across the 4 samples from the same genes.

All populations are assumed to be parametric since they are generated from rnorm()

Q1. Based on the transcriptomic data, which of these chemicals are structurally related to each other, and how did you come to this conclusion. Chemicals will be structurally related to each other if their biological activities across all four groups does not differ. This can be checked via a paired t-test.

```r
check_ttest <- function(dist1, dist2)
{
  p.a <- t.test(dist1$a,dist2$a,paired=TRUE)$p.value
  p.b <- t.test(dist1$b,dist2$b,paired=TRUE)$p.value
  p.c <- t.test(dist1$c,dist2$c,paired=TRUE)$p.value
  p.d <- t.test(dist1$d,dist2$d,paired=TRUE)$p.value

  count <- 0
  if(p.a>0.05)
    count <- count+1
  if(p.b>0.05)
    count <- count+1
  if(p.c>0.05)
    count <- count+1
  if(p.d>0.05)
    count <- count+1

  #check that the two chemicals show the same effect on at least 50% of the sample population
  return (list(count>=2, p.a, p.b, p.c, p.d))
}

sim <- hash()
sim[["TC1, TC2"]] <- check_ttest(tc1_data, tc2_data)[1]
sim[["TC1, TC3"]] <- check_ttest(tc1_data, tc3_data)[1]
sim[["TC1, TC4"]] <- check_ttest(tc1_data, tc4_data)[1]
sim[["TC1, TC5"]] <- check_ttest(tc1_data, tc5_data)[1]
sim[["TC2, TC3"]] <- check_ttest(tc2_data, tc3_data)[1]
sim[["TC2, TC4"]] <- check_ttest(tc2_data, tc4_data)[1]
sim[["TC2, TC5"]] <- check_ttest(tc2_data, tc5_data)[1]
sim[["TC3, TC4"]] <- check_ttest(tc3_data, tc4_data)[1]
sim[["TC3, TC5"]] <- check_ttest(tc3_data, tc5_data)[1]
sim[["TC4, TC5"]] <- check_ttest(tc4_data, tc5_data)[1]
```

```r
#check which chemicals are structurally similar
print_similar <- function(sim)
{
  print("Structurally Similar Chemicals are: ")
  for (k in keys(sim))
  {
    if(sim[[k]]==TRUE)
    {
      print(k)
    }
  }
}

print_similar(sim)
```

```
## [1] "Structurally Similar Chemicals are: "
## [1] "TC3, TC5"
```

Thus, TC3 and TC5 are structurally similar. How I reached this conclusion was through these steps:

- First, I performed paired t-test to check difference in means of two groups on each of the corresponding samples a, b, c and d for each chemical pair.
- Then, I found the p-values obtained for each chemical pair for samples a, b, c and d.
- Then, checked if the p-value was non-significant i.e. there is no difference in means for at least 50% of the corresponding sample distributions for those two chemicals i.e. any two of a, b, c and d.
- If it was, then those two chemicals showed similar biological activity and hence, have a similar chemical structure.

Q2. What genes are responsible for justifying structural similarity and dissimilarity? Provide the names of the top 10 genes in each condition. Structural similarity and dissimilarity would be defined by how consistent or inconsistent the activity of respective drugs is on the basal transcriptome respectively. What I have done here is calculated the effect size for all genes per drug with the basal transcriptome corresponding to the genes. This aggregates the effect of the drug across a,b,c and d.

Now, whether a gene has consistent effect sizes across all drugs can be found by calculating the variance of the observed drug effects. Finally, top 10 drugs with least effect variation across all drugs are the ones that can potentially lead to structural similarity. The top 10 drugs with the highest variation across all drugs lead to structural dissimilarity.

Top 10 genes responsible for justifying structural similarity: These will be genes whose mean effect remains consistent over the chemicals.

```
variation_in_effect <- function(includet4)
{
  eff_var <- list(length(1000))
  for (i in 1:nrow(tc3_data))
  {
    pop0 <- as.numeric(unlist(t0_data[i,][-1]))
    pop1 <- as.numeric(unlist(tc1_data[i,][-1]))
    pop2 <- as.numeric(unlist(tc2_data[i,][-1]))
    pop3 <- as.numeric(unlist(tc3_data[i,][-1]))
    pop4 <- as.numeric(unlist(tc4_data[i,][-1]))
    pop5 <- as.numeric(unlist(tc5_data[i,][-1]))
    eff1 <- cohen.d(pop0, pop1)$estimate
    eff2 <- cohen.d(pop0, pop2)$estimate
    eff3 <- cohen.d(pop0, pop3)$estimate
    eff4 <- cohen.d(pop0, pop4)$estimate
    eff5 <- cohen.d(pop0, pop5)$estimate
    if(includet4)
      eff_var[i] <- var(c(eff1, eff2, eff3, eff4, eff5))
    else
      eff_var[i] <- var(c(eff1, eff2, eff3, eff5))
  }
  df <- data.frame(Genes = t0_data$Genes)
  df$EffectVariation <- eff_var
  df <- as.data.frame(lapply(df, unlist))
  return (df)
}
```

Top 10 genes responsible for justifying structural similarity: These will be genes whose mean effect is the most consistent across different chemicals.

```
df <- variation_in_effect(TRUE)
df <- df[order(df$EffectVariation, decreasing = FALSE),]
print(df[1:10,])
```

```
##       Genes EffectVariation
## 116 GENE116       0.1365437
## 192 GENE192       0.1784257
## 30   GENE30       0.1915651
## 388 GENE388       0.2448784
## 584 GENE584       0.2543716
## 746 GENE746       0.2731821
## 321 GENE321       0.2756027
## 71   GENE71       0.2793289
## 492 GENE492       0.2903409
## 441 GENE441       0.3161354
```

Top 10 genes responsible for justifying structural dissimilarity: These will be drugs whose mean effect size is vastly inconsistent across different chemicals.

```
print(df[1000:991,])
```

```
##       Genes EffectVariation
## 352 GENE352        29.33699
## 206 GENE206        26.96362
## 501 GENE501        26.31237
## 704 GENE704        21.49084
## 676 GENE676        19.70704
## 121 GENE121        19.25909
## 404 GENE404        18.85438
## 867 GENE867        18.65000
## 142 GENE142        18.27376
## 103 GENE103        18.02477
```

Q3. Assuming that the experimentalist has done some mistake by forgetting to add compound C4 on the cells, how will the results for questions 1 and 2 change? Let us check if there is a statistical difference in the basal transcriptome data and data after adding C4

```
sim.t0_tc4 <- check_ttest(t0_data, tc4_data)
print(sprintf("p-value after t.test on population a: %f", sim.t0_tc4[2]))
```

```
## [1] "p-value after t.test on population a: 0.219942"
```

```
print(sprintf("p-value after t.test on population b: %f", sim.t0_tc4[3]))
```

```
## [1] "p-value after t.test on population b: 0.674488"
```

```
print(sprintf("p-value after t.test on population c: %f", sim.t0_tc4[4]))
```

```
## [1] "p-value after t.test on population c: 0.595211"
```

```
print(sprintf("p-value after t.test on population d: %f", sim.t0_tc4[5]))
```

```
## [1] "p-value after t.test on population d: 0.517057"
```

All of these are greater than 0.05 which means there is no statistical difference between the current basal transcriptome data and that in the file TC4_data. Since the experimentalist has forgotten to add compound C4, we can remove the data tc4 from consideration.

The answer to Q1 then becomes:

```
sim.pt3 <- hash()
sim.pt3[["TC1, TC2"]] <- check_ttest(tc1_data, tc2_data)[1]
sim.pt3[["TC1, TC3"]] <- check_ttest(tc1_data, tc3_data)[1]
sim.pt3[["TC1, TC5"]] <- check_ttest(tc1_data, tc5_data)[1]
sim.pt3[["TC2, TC3"]] <- check_ttest(tc2_data, tc3_data)[1]
sim.pt3[["TC2, TC5"]] <- check_ttest(tc2_data, tc5_data)[1]
sim.pt3[["TC3, TC5"]] <- check_ttest(tc3_data, tc5_data)[1]

print_similar(sim.pt3)
```

```
## [1] "Structurally Similar Chemicals are: "
## [1] "TC3, TC5"
```

(no change, because similar chemicals were found to be TC3 and TC5, no role played by TC4 here).

The answer to Q2 becomes:

```
df.pt3 <- variation_in_effect(FALSE)
```

Top 10 genes responsible for justifying structural similarity:

```
df.pt3 <- df.pt3[order(df.pt3$EffectVariation, decreasing = FALSE),]
print(df.pt3[1:10,])
```

```
##       Genes EffectVariation
## 175 GENE175      0.03476980
## 345 GENE345      0.06136282
## 601 GENE601      0.07882452
## 237 GENE237      0.08223950
## 597 GENE597      0.10950418
## 452 GENE452      0.11803252
## 891 GENE891      0.12322233
## 215 GENE215      0.13326048
## 30   GENE30      0.13402150
## 116 GENE116      0.14331060
```

Top 10 genes responsible for justifying structural dissimilarity:

```
print(df.pt3[1000:991,])
```

```
##       Genes EffectVariation
## 352 GENE352        34.88127
## 501 GENE501        24.82209
## 704 GENE704        22.73867
## 206 GENE206        21.17417
## 142 GENE142        19.10302
## 103 GENE103        17.44325
## 404 GENE404        17.16766
## 870 GENE870        16.30134
## 231 GENE231        16.02936
## 448 GENE448        15.87598
```