BM5033 Report

Final exam: Statistical Inference Methods in Bioengineering

Akhil Gattu
Ameya Chatur
Sriram Gonella
Vibhuvan Reddy
Vidya Ajay

26 October 2022

Problem Statement:

A lab working on the methods of drug delivery have designed four drug delivery vehicles. These delivery methods are metallic (primarily silver), polymeric (synthetic), plant based ECM derived nanoparticles, or liposomes. In the in vitro test to check the efficacy of these methods the drug promoting wound healing was added to three cell cultures with scratch assay. The three cell lines used in the study were MDCK cells, human primary epithelial cells and fibroblast cells. After the addition of the drug with each delivery method the cultures were monitored to measure the time for the scratch to heal.

Further, in the animal study to look at the effectiveness of the delivery methods experiments were performed on mice. On each mouse one minor wound was made on each forelimbs. In one wound the drug was applied directly while in the other wound the drug was applied with the help of the delivery method. This experiment was done for eight mice each delivery method and time of the wounds healing was recorded. The data for this experiment is stored in the sheets having names same as the delivery method.

With this information,

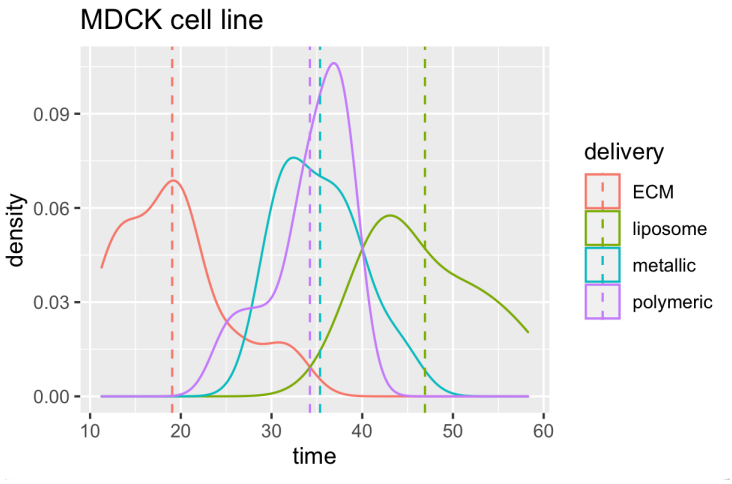
1. You are expected to explore the data and identify patterns, if any.
2. Perform appropriate statistical tests to infer the effectiveness of the delivery methods.
3. Is there any dependence between cell type and delivery method ?
4. Can you suggest any changes in the experimental methodology in this study.

1. Data analysis

Experiment 1: Effect of four delivery methods on three different cell lines – MDCK, HPEC and fibroblast cells.

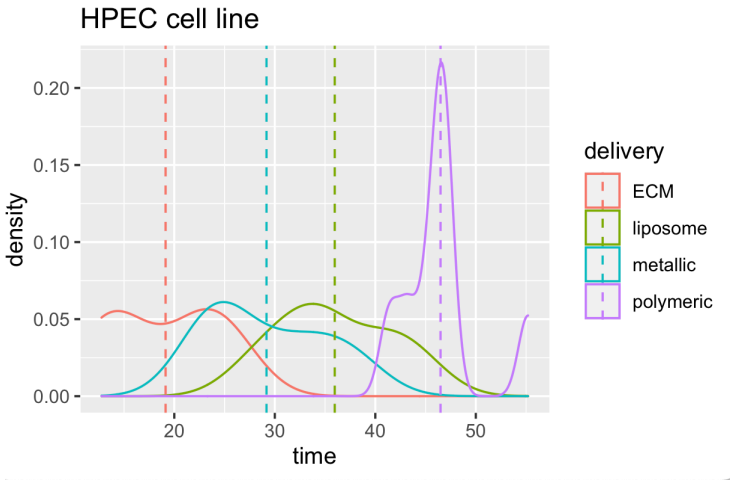
- The control parameters are nominal (cell type) and the output data is quantitative.
- In each cell type, four types of delivery methods were used (metallic, polymeric, ECM and liposomes); this implies the presence of four output groups.
- Since there are more than two output groups and only one output variable in each, further analysis on the data can be done using ANOVA.

Shown below are the density plots of the time taken for different delivery methods in each cell line. The dashed lines represent the mean values of each curve.



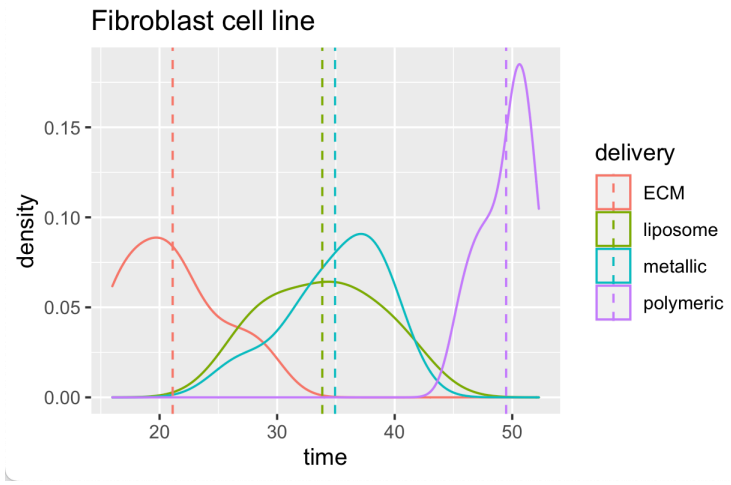
Delivery	*Mean Time*
ECM	19.05
liposome	46.91
metallic	35.35
polymeric	34.23

From this plot, it seems as though the ECM delivery method was the fastest in treating the wound in MDCK cells.



Delivery	*Mean Time*
ECM	19.13
liposome	35.96
metallic	29.17
polymeric	46.49

The polymer method shows a significant difference from the other methods by taking the most time by a huge margin (as seen in the spike).



Delivery	*Mean Time*
ECM	21.11
liposome	33.84
metallic	34.93
polymeric	49.48

Although most delivery methods performed relatively poorly with the fibroblast cells, the polymer method still takes much longer than the others.

Patterns observed in the data: It seems as though the ECM delivery method performs the best in all three cell lines observed here. It is difficult to make observations on the HPEC cell line at first sight because the readings seem to be spread out pretty evenly, except for the

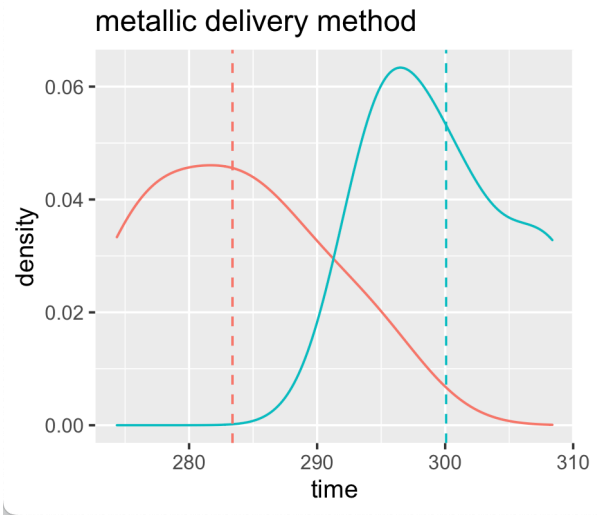
polymeric method. This method, on the other hand, seems to be the least effective in both HPEC and fibroblast cells (assuming that a shorter recovery time implies a more effective drug).

Another noticeable pattern is how almost all cells in a given cell line respond similarly to the polymeric method, be it a shorter or longer recovery time. This is evident in the large peak in all the curves of the polymer method.

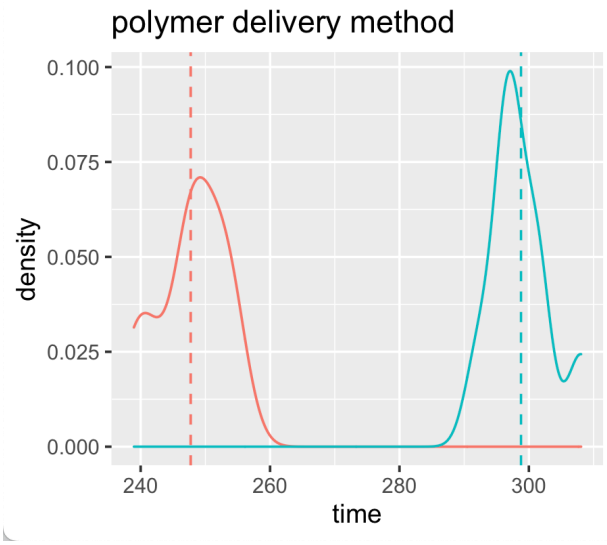
Experiment 2: Delivery method *vs.* No delivery method (mouse cells)

- This experiment measured the recovery times it took for mouse cells in (1) a control group that took the drug without the delivery method and (2) a group that took the drug *with* the delivery method for each of the four delivery methods (metallic, polymeric, ECM, liposome).
- The control parameters here are nominal and the output data is quantitative.
- Since there is only one output group and the sample sizes are the same, we can use paired tests (t-test or wilcoxon rank test, depending on the normality of the distribution) to analyze the data.

Shown below are the density plots of the time taken for both the control group and the delivery method group for each of the four delivery methods. The dashed lines represent the mean values of each curve.



Group	*Mean Time*
Drug	283.38
No drug	300.07



Group	*Mean Time*
Delivery	247.73
No delivery	298.78

Group	*Mean Time*
Delivery	280.53
No delivery	297.62

Group	*Mean Time*
Delivery	281.06
No delivery	280.52

Patterns observed in the data: In mouse cells, the delivery method that showed the most decrease in wound healing time is the polymeric method. Both metallic and ECM methods still show a significant difference, whereas negligible difference is reported in the liposome delivery.

It could be that the polymeric delivery method works specifically well in wound healing in mouse cells and not the three other cell lines (MDCK, HPEC and fibroblast), because in the previous experiment it was seen to have performed the worst.
In concordance with the first experiment, all the mouse cells tested also respond similarly to the polymeric delivery method (sharp characteristic peaks).

2. Statistical tests

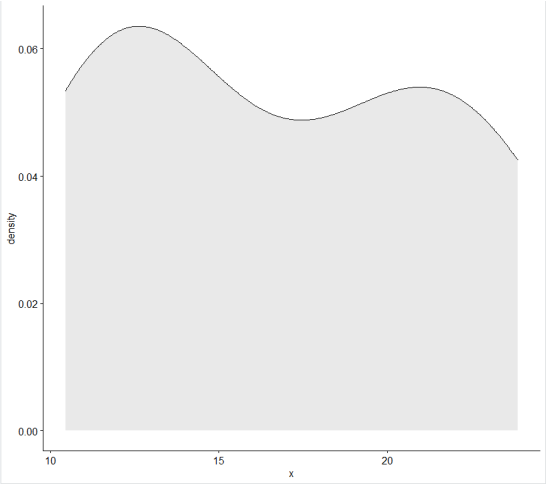
To determine the effectiveness of the delivery methods, we look at the data from the second experiment. Based on the normality of the difference of the pairs, we performed different statistical tests. The normality test we used was the Shapiro-Wilk test (null hypothesis: "sample distribution is normal")
All tests from here on will assume a significance level of 0.05.

Metallic delivery method:

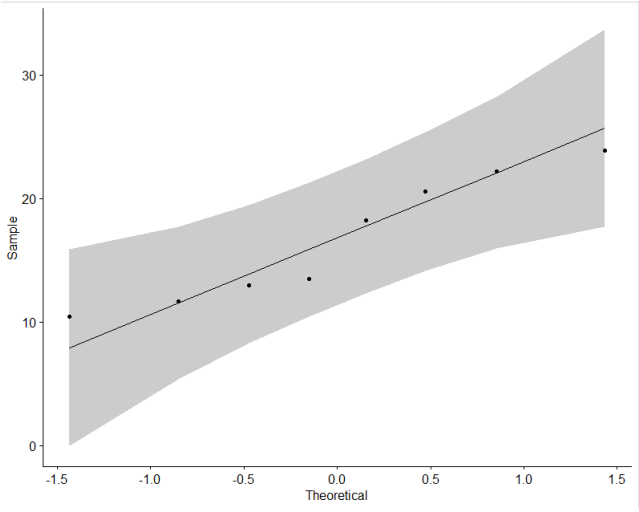
Shapiro-Wilk test:

```
> shapiro.test(metallic$Withoutdrug - metallic$WithDrug)

    Shapiro-Wilk normality test

data:  metallic$Withoutdrug - metallic$WithDrug
W = 0.9078, p-value = 0.3388
```

As p-value>0.05 and from the plots below, the data is not significantly different from the normal distribution.



Density plot of difference in values of response times with and without the "metallic" drug delivery method



QQ plot of difference in values of response times with and without the "metallic" drug delivery method

Performing one-tailed paired t-test on the data (μ_d is mean time with the delivery method μ_w is the mean time without the delivery method):

H_0: $\mu_d \geq \mu_w$

H_A: $\mu_d < \mu_w$

```
> t.test(with_method_metallic,without_method_metallic,alternative="less",paired=TRUE,var.equal=FALSE)

        Paired t-test

data:  with_method_metallic and without_method_metallic
t = -9.1117, df = 7, p-value = 1.968e-05
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
      -Inf -13.21737
sample estimates:
mean difference
      -16.68708
```

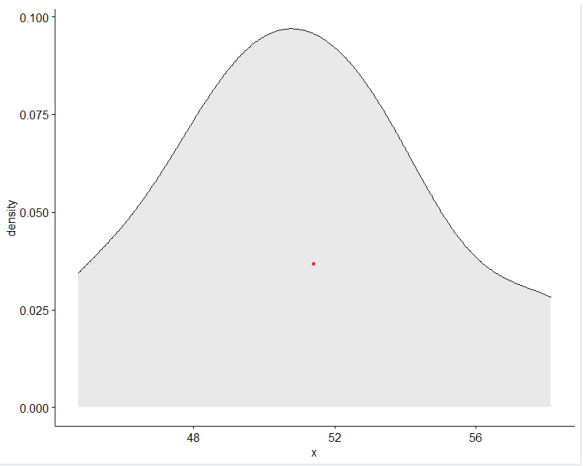
As p-value < 0.05, we reject the null hypothesis, we can say that the *response time decreases* through this delivery method ($\mu_d < \mu_w$), and it is

Polymeric delivery method:

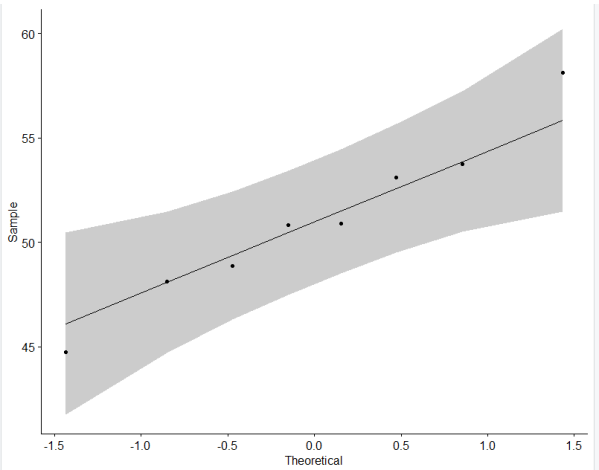
Shapiro-Wilk test:

```
> shapiro.test(polymer$Withoutdrug - polymer$WithDrug)

    Shapiro-Wilk normality test

data:  polymer$Withoutdrug - polymer$WithDrug
W = 0.98327, p-value = 0.9772
```

As p-value>0.05, the difference data is not significantly different from the normal distribution.



Density plot of difference in values of response times with and without the "polymeric" drug delivery method



QQ plot of difference in values of response times with and without the "polymeric" drug delivery method

Both the Shapiro-Wilk test and the QQ plot shows that the data is approximately normally distributed.

Performing one-tailed paired t-test on the data (μ_d is mean time with the delivery method μ_w is the mean time without the delivery method):

H_0: $\mu_d \geq \mu_w$

H_A: $\mu_d < \mu_w$

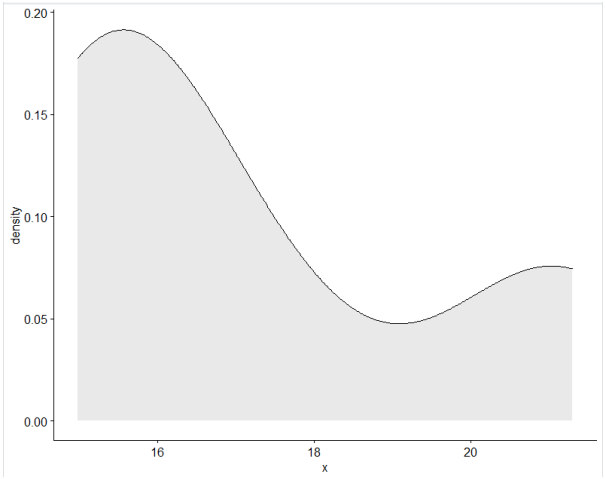
```
> t.test(with_method_polymeric,without_method_polymeric,alternative="less",paired=TRUE,var.equal=FALSE)

        Paired t-test

data:  with_method_polymeric and without_method_polymeric
t = -35.699, df = 7, p-value = 1.757e-09
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
     -Inf -48.3455
sample estimates:
mean difference
        -51.055
```

As p-value < 0.05, we reject the null hypothesis, hence, we can say that the *response time decreases* through the polymeric delivery method ($\mu_d < \mu_w$), and it is effective.

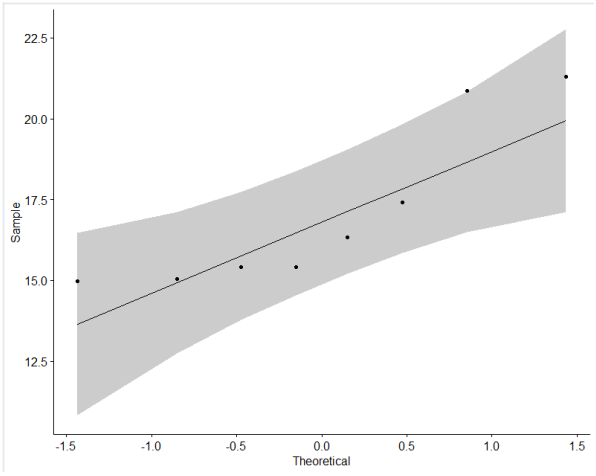
ECM delivery method:

Shapiro-Wilk test:

```
> shapiro.test(ecm$Withoutdrug - ecm$WithDrug)

    Shapiro-Wilk normality test

data:  ecm$Withoutdrug - ecm$WithDrug
W = 0.78394, p-value = 0.01922
```



Density plot of difference in values of response times with and without the "ECM" drug delivery method



QQ plot of difference in values of response times with and without the "ECM" drug delivery method

From the Shapiro-Wilk test (p-value<0.05) and both plots where the data deviate from the normal distribution (QQ plot is outside the bounds of the reference line), we conclude that the data is not normally distributed.

Perfroming non-parametric Wilcoxon Rank test on the data (μ_d is mean time with the delivery method μ_w is the mean time without the delivery method):

H_0: $\mu_d \geq \mu_w$

H_A: $\mu_d < \mu_w$

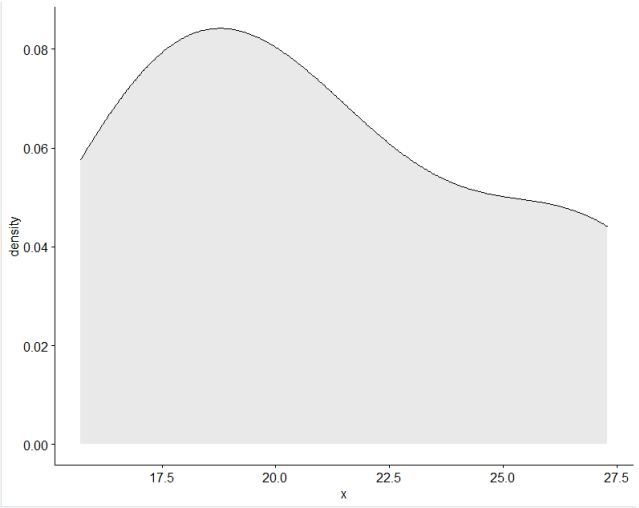
```
> wilcox.test(with_method_ecm,without_method_ecm,alternative="less",paired=TRUE,var.equal=FALSE)

        Wilcoxon signed rank exact test

data:  with_method_ecm and without_method_ecm
V = 0, p-value = 0.003906
alternative hypothesis: true location shift is less than 0
```

As p-value < 0.05, we reject the null hypothesis, hence, we can say that the *response time decreases* through the ECM delivery method ($\mu_d < \mu_w$), and it is effective.

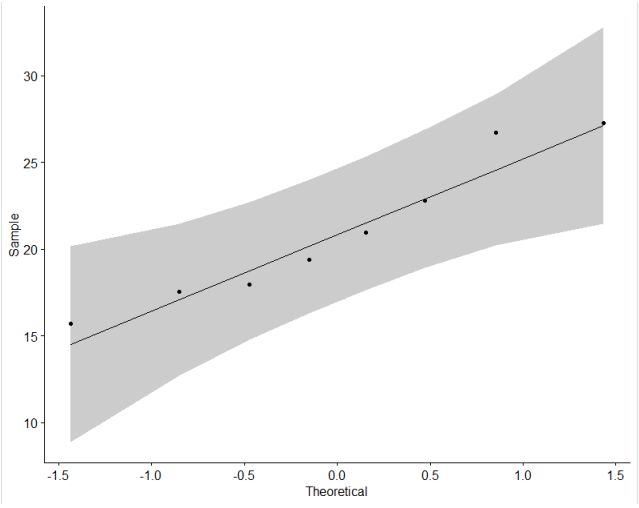
Liposome delivery method:

Shapiro-Wilk test:

```
> shapiro.test(lipo$Withoutdrug - lipo$WithDrug)

    Shapiro-Wilk normality test

data:  lipo$Withoutdrug - lipo$WithDrug
W = 0.92368, p-value = 0.4605
```



Density plot of difference in values of response times with and without the "liposome" drug delivery method



QQ plot of difference in values of response times with and without the "liposome" drug delivery method

As p-value>0.05, and the QQ plot is within the bounds of the reference line, the difference data is approximately normally distributed.

Performing one-tailed paired t-test on the data (μ_d is mean time with the delivery method μ_w is the mean time without the delivery method):

H_0: $\mu_d \geq \mu_w$

H_A: $\mu_d < \mu_w$

```
> t.test(with_method_liposome,without_method_liposome,alternative="less",paired=TRUE,var.equal=FALSE)

        Paired t-test

data:  with_method_liposome and without_method_liposome
t = -13.965, df = 7, p-value = 1.143e-06
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
      -Inf -18.19672
sample estimates:
mean difference
      -21.05297
```

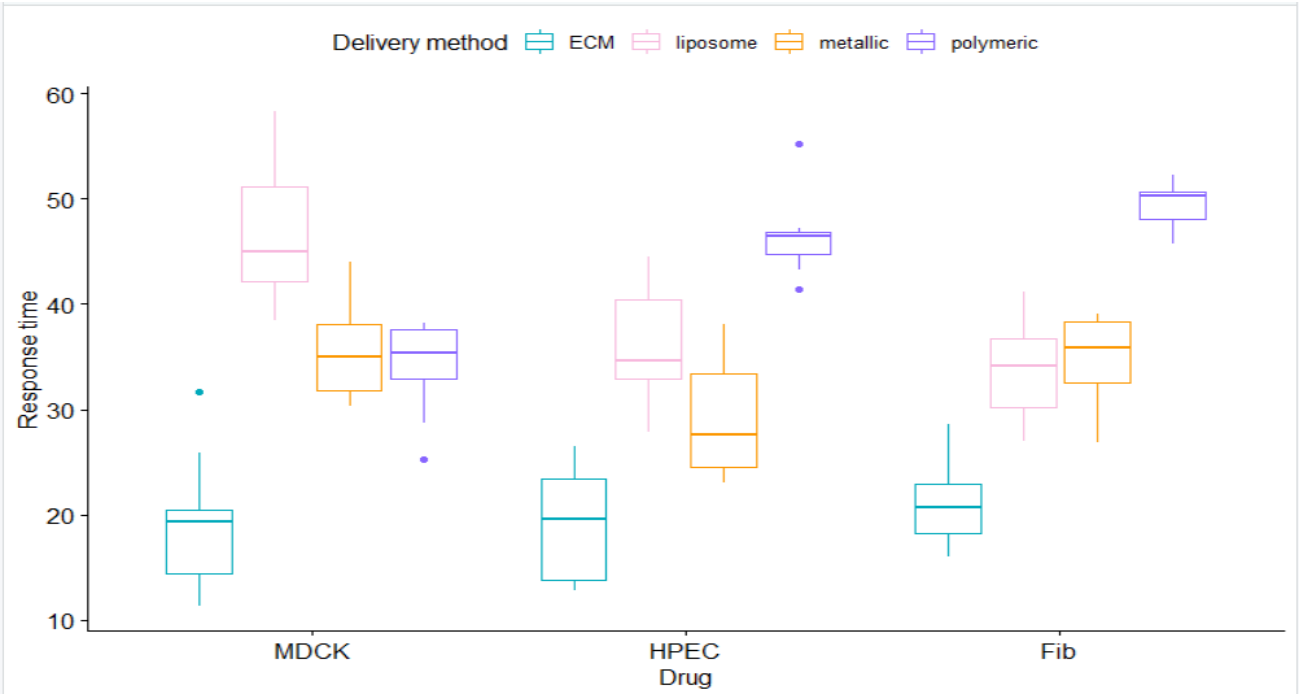
As p-value < 0.05, we reject the null hypothesis, hence, we can say that the response time decreases through the liposome delivery method ($\mu_d < \mu_w$), and it is effective.

3. Cell type - Delivery Method Dependence

For this section, we have combined data from all the sheets in the given excel file and combined them into a single file for ease of performing analysis of variance.

We will now look at the boxplot of the data with response time being on y-axis and the drug type being on x-axis under 4 different delivery methods to understand the minimum, maximum, median, 25 percentile and 75 percentile data:

All tests from here on will assume a significance level of 0.05.



Normality and homogeneity of variance testing:

Bartlett's test: One of the important assumptions of two way ANOVA is the assumption of homoscedasticity, i.e the variance for each set of data should not vary significantly for all the groups. So, to test homoscedasticity, we use Bartlett's test (null hypothesis: "The variance is homogeneous in all the groups"), the result for which is as follows:

```
> bartlett.test(data$`Response time` ~ data$`Cell type`)

        Bartlett test of homogeneity of variances

data:  data$`Response time` by data$`Cell type`
Bartlett's K-squared = 0.055812, df = 2, p-value = 0.9725
```

Since the p-value 0.97 is greater than 0.05, the null hypothesis cannot be rejected. Clearly, there are two factors — cell type and delivery method — for which we have to check for dependence. This calls for the two-factor ANOVA test, where we will be testing three null hypotheses simultaneously:

1. H_0: There is no difference between response time due to cell types.
   H_A: There is a difference between response time due to cell types.
2. H_0: There is no difference between response time due to delivery methods.
   H_A: There is a difference between response time due to delivery methods.

3. H_0: There is no interaction between cell type and delivery method.

   H_A: There is interaction between cell type and delivery method.

The two-factor ANOVA test gave the following result in R:

```
> aovOutput = aov(data$`Response time` ~ data$`Cell type`*data$`Delivery method`)
> summary(aovOutput)
                                    Df Sum Sq Mean Sq F value   Pr(>F)
data$`Cell type`                     2     74    37.2   1.482    0.233
data$`Delivery method`               3   8076  2692.0 107.344  < 2e-16 ***
data$`Cell type`:data$`Delivery method`  6   2253   375.6  14.975 7.02e-12 ***
Residuals                           92   2307    25.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA gave a p-value much less than 0.05 (significance level) for both the second and the third hypotheses which indicates that (1) there is a difference in the response time due to the delivery methods, and (2) there is an interaction between cell type and delivery method.

Conclusion: There is a a dependence between cell type and delivery method. Further analysis can be seen in the output of the Tukey HSD code appended at the end of the report.

4. Conclusion and Possible Changes in the Experimental Methodology

From the above analysis, we can conclude that every delivery method is effective since drug delivery through each method is able to lower the response time of the cell for healing. Even though the liposome delivery method at first look seemed to not be effective, statistical analysis proved that it does indeed reduce the healing time. Also, we have observed that there is a dependence between delivery method and cell type. We have also suggested a few changes below which we can make to this experimental setup in order to better analyze the effects of the different drug delivery methods to specific cell lines.

- For both experiments in this study, the sample size was 8, which is quite small to accurately reflect the population statistics. As is the case with every experiment, increasing the sample size may help infer outcomes of a population better.
- The type of mice used for the second experiment was not mentioned, so it is safe to assume that the mice were randomly chosen. With such a small sample size it is difficult to verify that the sample was diverse/ random in terms of their genetic makeup. Mice from a certain parentage may posses genes that make them sensitive to certain delivery

methods, but this might not be a statistically significant representative of population summary.

- The experiment done on the mice (with the control groups) could also have been performed on the three cell lines (MDCK, HPEC and fibroblast) to get a comprehensive report on whether the delivery method would actually have a difference in each cell line. With the information available to us now, it is difficult to conclude whether the delivery methods were “of any use”, seeing as to how it is not compared to the no-delivery-method group. A control group is always beneficial in these studies.

Appendix

All R codes (with the output) and edited excel files can be found in the drive link here. The scripts are also appended below.

Resources used:

- http://www.sthda.com/english/wiki/r-basic-statistics
- http://www.sthda.com/english/wiki/ggplot2-essentials

DENSITY PLOTS:

```
library(readxl)
library(plyr)
library(ggplot2)
mdck = read_excel("q3d1.xlsx", sheet = 'newmdck')
mu = ddply(mdck, 'delivery', summarise, grp.mean = mean(time))
head(mu)
p1 = ggplot(mdck, aes(x=time, color = delivery)) + geom_density() + geom_vline(data = mu, aes(xintercept =
grp.mean, color = delivery), linetype = 'dashed') + ggtitle('MDCK cell line')
p1
hpec = read_excel("q3d1.xlsx", sheet = 'newhpec')
mu2 = ddply(hpec, 'delivery', summarise, grp.mean = mean(time))
head(mu2)
p2 = ggplot(hpec, aes(x=time, color = delivery)) + geom_density() + geom_vline(data = mu2, aes(xintercept =
grp.mean, color = delivery), linetype = 'dashed') + ggtitle('HPEC cell line')
p2
fib = read_excel('q3d1.xlsx', sheet = 'newfib')
mu = ddply(fib, 'delivery', summarise, grp.mean = mean(time))
head(mu)
p3 = ggplot(fib, aes(x=time, color = delivery)) + geom_density() + geom_vline(data = mu, aes(xintercept = grp.mean,
color = delivery), linetype = 'dashed') + ggtitle('Fibroblast cell line')
p3

####################

metallic = read_excel('q3d1.xlsx', sheet = 'newmet')
mu = ddply(metallic, 'group', summarise, grp.mean = mean(time))
head(mu)
p4 = ggplot(metallic, aes(x=time, color = group)) + geom_density() + geom_vline(data = mu, aes(xintercept =
grp.mean, color = group), linetype = 'dashed') + ggtitle('metallic delivery method')
p4

polymer = read_excel('q3d1.xlsx', sheet = 'newpoly')
mu = ddply(polymer, 'group', summarise, grp.mean = mean(time))
head(mu)
p5 = ggplot(metallic, aes(x=time, color = group)) + geom_density() + geom_vline(data = mu, aes(xintercept =
grp.mean, color = group), linetype = 'dashed') + ggtitle('polymer delivery method')
```

```
p5

ecm = read_excel('q3d1.xlsx', sheet = 'newecm')
mu = ddply(ecm, 'group', summarise, grp.mean = mean(time))
head(mu)
p6 = ggplot(ecm, aes(x=time, color = group)) + geom_density() + geom_vline(data = mu, aes(xintercept = grp.mean,
color = group), linetype = 'dashed') + ggtitle('ECM delivery method')
p6

liposome = read_excel('q3d1.xlsx', sheet = 'newlipo')
mu = ddply(liposome, 'group', summarise, grp.mean = mean(time))
head(mu)
p7 = ggplot(liposome, aes(x=time, color = group)) + geom_density() + geom_vline(data = mu, aes(xintercept =
grp.mean, color = group), linetype = 'dashed') + ggtitle('Liposome delivery method')
p7
```

STATISTICAL TESTS:

```
library(readxl)
library(moments)
library("ggpubr")
#For metallic
Metallic=read_excel("q3d1.xlsx",sheet="Metallic",col_names = TRUE)
without_method_metallic=Metallic$"Without drug"
with_method_metallic=Metallic$"With Drug"

shapiro.test(without_method_metallic)
shapiro.test(with_method_metallic)

skewness(without_method_metallic-with_method_metallic)
kurtosis(without_method_metallic-with_method_metallic)
shapiro.test(without_method_metallic-with_method_metallic)
ggdensity(without_method_metallic-with_method_metallic, fill = "lightgray")
# QQ plot
ggqqplot(without_method_metallic-with_method_metallic)

#Polymeric
Polymeric=read_excel("q3d1.xlsx",sheet="Polymeric",col_names = TRUE)
without_method_polymeric=Polymeric$"Without drug"
with_method_polymeric=Polymeric$"With Drug"

#plot(without_method_polymeric-with_method_polymeric, type = "line")
ggdensity(without_method_polymeric-with_method_polymeric, fill = "lightgray")
# QQ plot
ggqqplot(without_method_polymeric-with_method_polymeric)
skewness(without_method_polymeric-with_method_polymeric)
kurtosis(without_method_polymeric-with_method_polymeric)
wilcox.test(without_method_polymeric, with_method_polymeric, paired = TRUE, alternative = "greater")

shapiro.test(without_method_polymeric-with_method_polymeric)
shapiro.test(without_method_polymeric)
shapiro.test(with_method_polymeric)

#ECM
ECM=read_excel("q3d1.xlsx",sheet="ECM",col_names = TRUE)
without_method_ecm=ECM$"Without drug"
with_method_ecm=ECM$"With Drug"

plot(without_method_ecm - with_method_ecm, type = "line")
wilcox.test(without_method_ecm, with_method_ecm, paired = TRUE, alternative = "less")
ggdensity(without_method_ecm-with_method_ecm, fill = "lightgray")
# QQ plot
ggqqplot(without_method_ecm-with_method_ecm)
shapiro.test(without_method_ecm - with_method_ecm)
shapiro.test(without_method_ecm)
shapiro.test(with_method_ecm)
#Liposome
Liposome=read_excel("q3d1.xlsx",sheet="Liposome",col_names = TRUE)
without_method_liposome=Liposome$"Without drug"
with_method_liposome=Liposome$"With Drug"

shapiro.test(without_method_liposome)
shapiro.test(with_method_liposome)
ggdensity(without_method_liposome-with_method_liposome, fill = “lightgray")
# QQ plot
ggqqplot(without_method_liposome-with_method_liposome)
shapiro.test(without_method_liposome - with_method_liposome)
wilcox.test(without_method_liposome, with_method_liposome, paired = TRUE, var.equal = TRUE, alternative =
"greater")
skewness(without_method_liposome - with_method_liposome)
kurtosis(without_method_liposome - with_method_liposome)

#Hypothesis Testing - Parametric since data is noramlly distributed

#Metallic
t.test(with_method_metallic,without_method_metallic,alternative="less",paired=TRUE,var.equal=FALSE)
```

```
#Polymeric
t.test(with_method_polymeric,without_method_polymeric,alternative="less",paired=TRUE,var.equal=FALSE)

#ECM
wilcox.test(with_method_ecm,without_method_ecm,alternative="less",paired=TRUE,var.equal=FALSE)
#Liposome
t.test(with_method_liposome,without_method_liposome,alternative="less",paired=TRUE,var.equal=FALSE)

#ANOVA - We only have one factor

#Pairwise Comparison of Delivery Methods

#Metallic vs Polymeric
t.test(with_method_metallic,with_method_polymeric,alternative="less")
#  283.3879  247.7323
#Metallic vs ECM
t.test(with_method_metallic,with_method_ecm,alternative="less")
#Metallics vs Liposome
t.test(with_method_metallic,with_method_liposome,alternative="less")
#Polymeric vs ECM
t.test(with_method_polymeric,with_method_ecm,alternative="less")
#Polymeric vs Liposome
t.test(with_method_polymeric,with_method_liposome,alternative="less")
#ECM vs Liposome
t.test(with_method_ecm,with_method_liposome,alternative=“less”)
```

TUKEY HSD

```
library(readxl)
library(ggpubr)
#data_MDCK = read_excel("q3d1.xlsx", sheet = "MDCK")
#data_HPEC = read_excel("q3d1.xlsx", sheet = "HPEC")
#data_fib = read_excel("q3d1.xlsx", sheet = "Fib")
library(ggpubr)
data = read_excel("combined.xlsx")
ggboxplot(data, x = "Cell type", y = "Response time",
          color = "Delivery method",
          order = c("MDCK","HPEC","Fib"),
          ylab = "Response time", xlab = "Drug")



# homogeneity of variance

bartlett.test(data$`Response time` ~ data$`Cell type`)

aovOutput = aov(data$`Response time` ~ data$`Cell type`*data$`Delivery method`)
#summary(aovOutput)
tukOut = TukeyHSD(aovOutput)
print(tukOut)
```

Group members’ contribution score:

Akhil Gattu: 10
Ameya Chatur: 10
Sriram Gonella: 10
Vibhuvan Reddy: 10
Vidya Ajay: 10