Homework 1

Group 5: Mikkel Corfitzen, Katja Johansen, Stefano Pellegrini, Rikke Stausholm

4/27/2020

Question 1

```
library(tidyverse)
library(babynames)
head(babynames)
```

```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

a) List the top 5 female baby names starting with P, regardless of year, as a table.

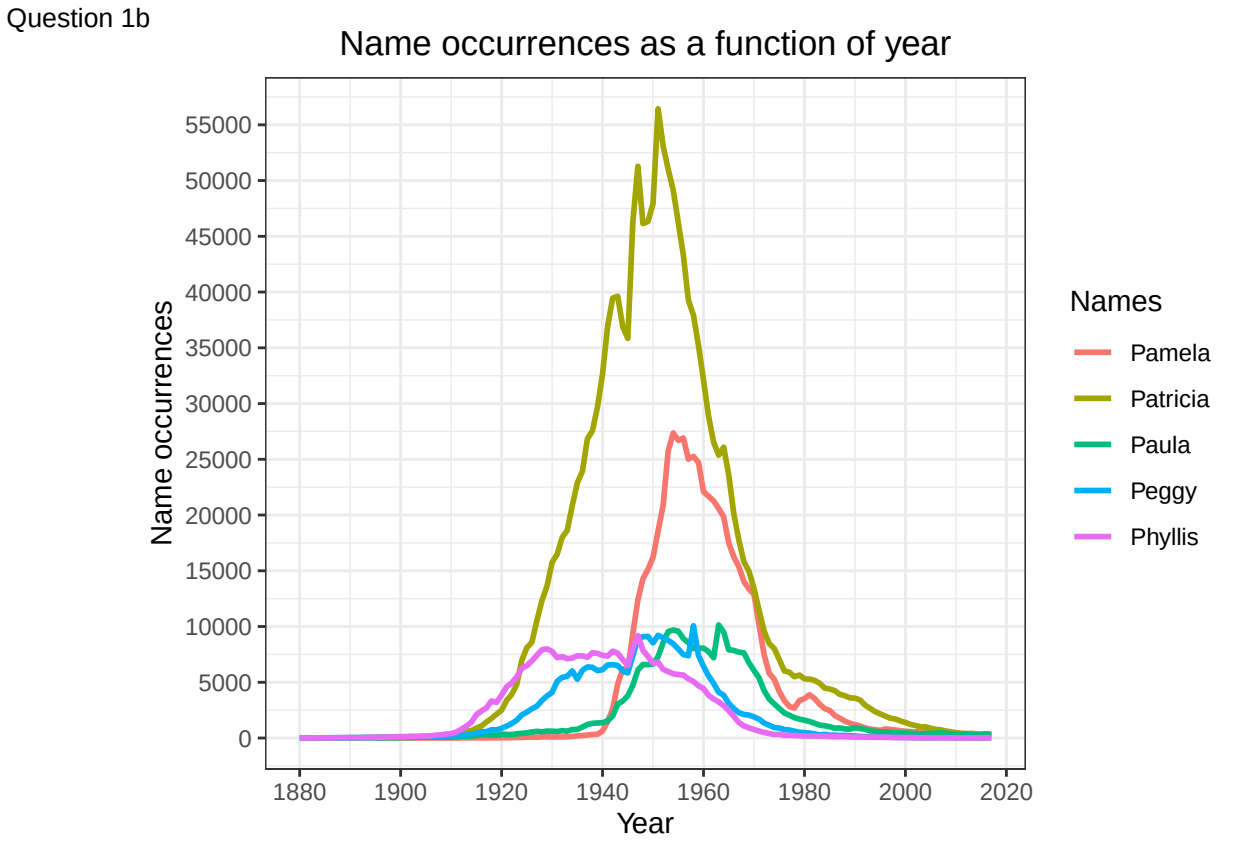
```
babynames %>%
  filter(sex == "F", str_detect(name, "^P")) %>% group_by(name) %>%
  summarise(total_count = sum(n)) %>% arrange(desc(total_count)) %>%
  head(5) -> top_P.female
top_P.female
```

```
## # A tibble: 5 x 2
##   name     total_count
##   <chr>          <int>
## 1 Patricia     1571692
## 2 Pamela        594174
## 3 Phyllis       322369
## 4 Peggy         292585
## 5 Paula         278003
```

b) Using the results from a, plot their occurrences as a function of year using a line plot. Comment on your results. If you get strange results, explain them and/or improve the plot.

```
babynames %>% filter(sex == "F", name %in% top_P.female$name) %>%
  ggplot(aes(x=year, y=n, col=name)) + geom_line(size=1) + theme_bw() +
  # Customize plot details
```

```
  ylab("Name occurrences") + xlab("Year") +
  labs(title="Name occurrences as a function of year",
       color="Names",
       tag = "Question 1b") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.tag = element_text(size = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10))
```



The most occurring female names from the 1880 to the 2017 are Pamela, Patricia, Paula, Peggy and Phyllis. It is possible to observe that Patricia was really popular from 1930 to 1970, while Pamela had a spike in popularity from 1950 to 1970. The remaining three names do not show any particular peak and there have been around 300.000 occurrences of each name from 1880 to 2017.

Question 2

In the same dataset, is the name Arwen significantly more (or less) common in 2004 vs 1990? Is the change significant? What is the likely cause? Do not use hard-coding.

```
# Filter, add a column for the total name count per year,
# add column for the total name per year with Arwen excluded
babynames %>%
  group_by(year) %>% mutate(Total_names = sum(n)) %>%
  filter(name == "Arwen", year %in% c("1990","2004")) %>%
  mutate(Total_no_Arwen = Total_names - n) -> arwen_2years
```

```
arwen_2years
```

```
## # A tibble: 2 x 7
## # Groups:   year [2]
##    year sex   name      n       prop Total_names Total_no_Arwen
##   <dbl> <chr> <chr> <int>      <dbl>       <int>          <int>
## 1  1990 F     Arwen    10 0.00000487     3950992        3950982
## 2  2004 F     Arwen   166 0.0000823      3818361        3818195
```

```
# Convert the tibble to a dataframe
m <- as.matrix(arwen_2years[c(4,7)])
rownames(m) <- arwen_2years$year
colnames(m)[1] <- "Arwen"
m
```

```
##      Arwen Total_no_Arwen
## 1990    10        3950982
## 2004   166        3818195
```

We choose to perform the Fisher exact test because it is the correct test to check if there is a difference in the odds ratio between two groups, in a 2 by 2 matrix. Here we are testing if the odds of the babys named Arwen between 1990 and 2004, divided by the odds of the total names (Arwen escluded) given in these two years, is significantly different than 1.

Null: the odds ratio between the two groups is 1. Meaning that the name Arwen is as common in 1990 as it is in 2004.

```
# Perform Fisher exact test
fisher.test(m)
```

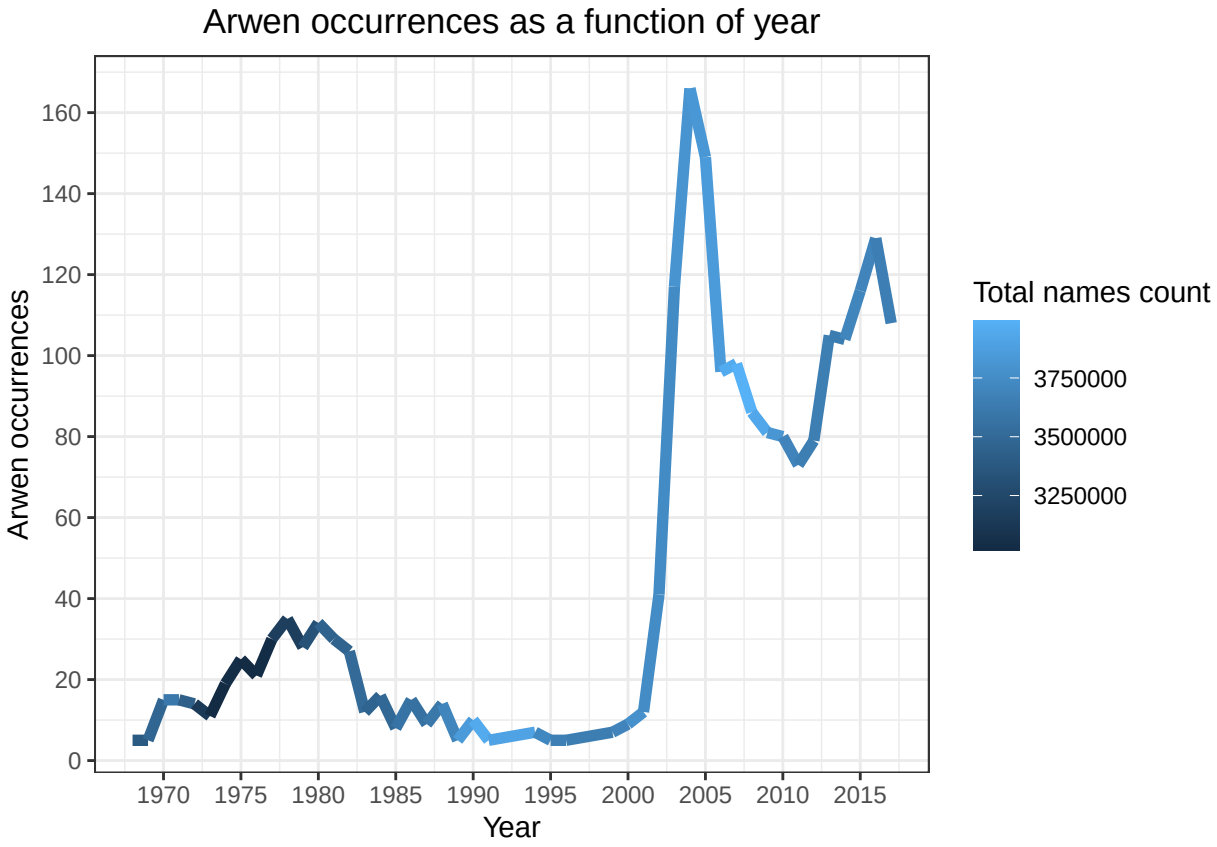
```
##
##  Fisher's Exact Test for Count Data
##
## data:  m
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02737909 0.10978171
## sample estimates:
## odds ratio
## 0.05820734
```

We can observe that the p-value is really low (2.2e-16) and that the odds ratio is 0.058. So we have enough proof to reject the null hypothesis, we can state that the odds ratio between the two groups is significantly different than 1 and that the name Arwen is significantly more common in 2004 than 1990.

2001, 2002, 2003 are the years of the release of The Lord of the Rings film series. It is likely that this is the cause of the result we obtained and, even if we can't prove it, the following plot give us strong belief that this is the case.

```
babynames %>%
  group_by(year) %>%
  mutate(Total_names = sum(n)) %>%
  filter(name == "Arwen") %>%
  mutate(sum(n)) %>%
  ggplot(aes(x=year, y=n, col=Total_names)) + geom_line(size=2) + theme_bw() +
  # Customize plot details
  ylab("Arwen occurrences") + xlab("Year") +
```

```
  labs(title="Arwen occurrences as a function of year", color="Total names count") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10))
```



Question 3

Produce the following plot starting from the flowers dataset. A potentially useful function that you may not have seen: bind_rows(), merges two tibbles by rows so that the joint tibble becomes longer, not wider.

```
my_flowers <- read_tsv("flowers.txt")
```

```
## Parsed with column specification:
## cols(
##   Sepal.Length = col_double(),
##   Sepal.Width = col_double(),
##   Petal.Length = col_double(),
##   Petal.Width = col_double(),
##   Species = col_character()
## )
```

```
# Generate a tibble for the sepal adding organ information
my_flowers[c(1,2,5)] %>% mutate(Organ = "Sepal") -> Sepal
colnames(Sepal)[1:2] <- c("length", "width")
```

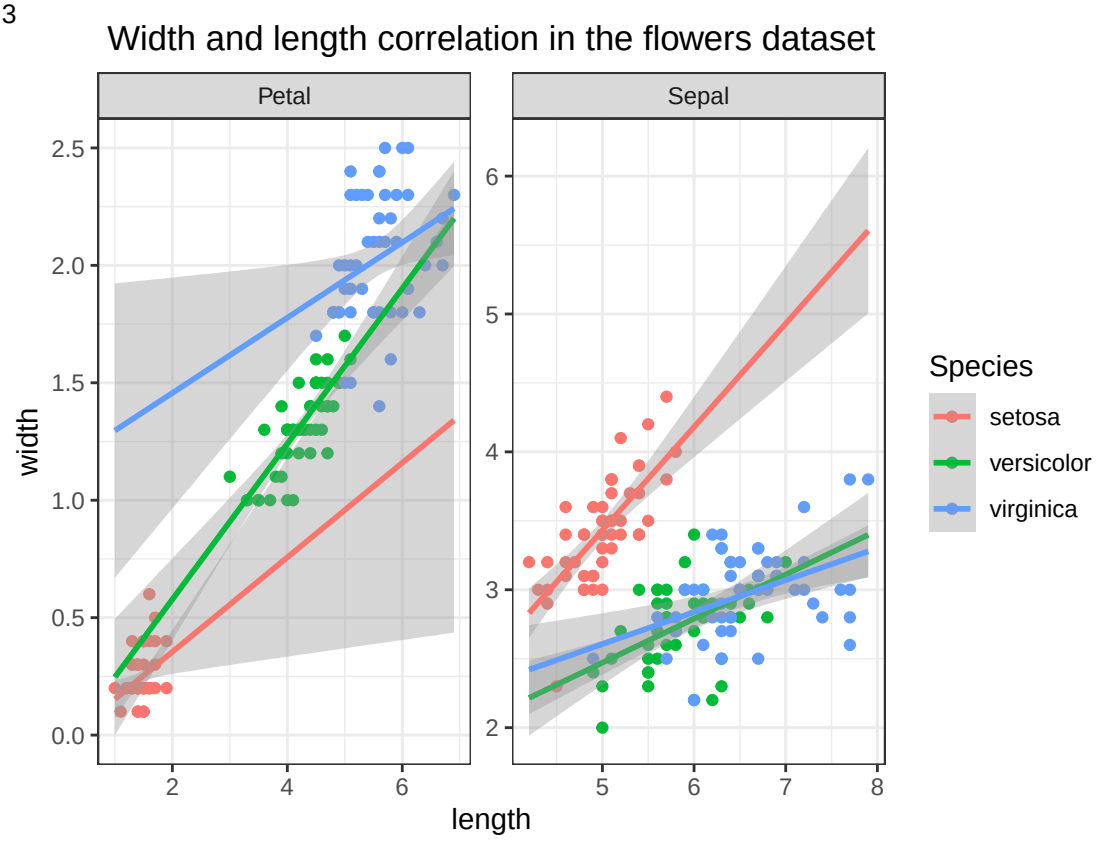
```
# Generate a tibble for the petal adding organ information
my_flowers[3:5] %>% mutate(Organ = "Petal") -> Petal
colnames(Petal)[1:2] <- c("length", "width")

# Merge the two tibbles
new_tibble <- bind_rows(Sepal, Petal)

# Plot
new_tibble %>%
  ggplot(aes(x=length, y=width, col=Species)) + geom_point(size=1.5) +
  facet_wrap(~ Organ, scale ="free") +
  geom_smooth(method="lm", size=1, fullrange = TRUE) + theme_bw() +
  labs(title="Width and length correlation in the flowers dataset",
       tag = "Question 3") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.tag = element_text(size = 10))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Question 3




```
# Alternative way in just one line
bind_rows(iris %>%
  select(Species , Petal.Width , Petal.Length) %>%
  add_column(flower_part = "Petal") %>%
  rename(Length = Petal.Length) %>%
  rename(Width = Petal.Width),
  iris %>%
```

```
  select(Species , Sepal.Width , Sepal.Length) %>%
  add_column(flower_part = "Sepal") %>%
  rename(Length = Sepal.Length) %>%
  rename(Width = Sepal.Width)) %>%
ggplot(aes (x=Length, y=Width , col=Species)) +
  geom_point() +
  geom_smooth(method="lm", fullrange = TRUE) +
  theme_bw() +
  facet_wrap(~flower_part, scales = "free")
```

Question 4

We are given a file with binding sites of a certain transcription factor, made with the ChIP-seq technique (you will hear a lot more about the technique later in the course) by a collaborator. In the homework directory, there is a data file 'chip_mm5.txt' from the collaborator, representing binding sites from a Chip-chip experiment, with a column for chromosome, start, end, and score, where score is how 'good' the binding is. Our collaborator has two hypothesis:

1: Binding scores are dependent on chromosome

2: Binding site widths (end-start) are dependent on chromosome

a) Hypothesis 1

```
# Load data and drop rows containing NA values
my_chip <- read_tsv("chip_mm5.txt")
```

```
## Parsed with column specification:
## cols(
##   chr = col_character(),
##   start = col_double(),
##   end = col_double(),
##   score = col_double()
## )
```

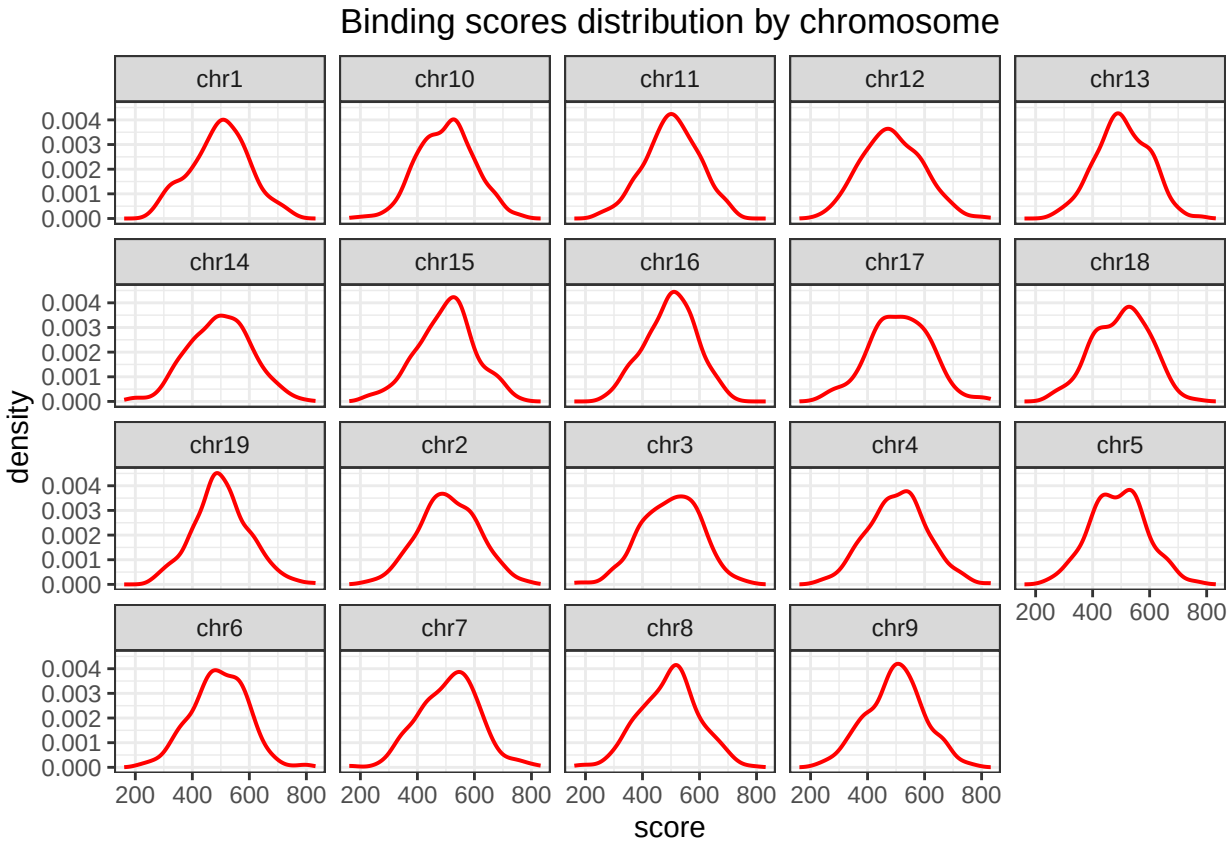
```
dim(my_chip)
```

```
## [1] 5415    4
```

```
my_chip <- drop_na(my_chip)
dim(my_chip)
```

```
## [1] 5410    4
```

```
# Plot the binding score ditributions by chromosome
my_chip %>% ggplot(aes(x=score)) + geom_line(col="red", size=0.7, stat="density") +
    facet_wrap(~chr) + theme_bw() +
    labs(title="Binding scores distribution by chromosome") +
    theme(plot.title = element_text(hjust = 0.5))
```



After preprocessing the data (removing rows containing at least an NA values), we plotted the binding scores distribution of each chromosome, and they seem to be all normally distributed.

Next, since the binding scores are normally distributed, we decided to perform an ANOVA test to check if there is a difference between the means of the scores distributions between chromosomes.

Null: there is no differene in the means of the binding score distributions between the chromosome.

```
# Multisample ANOVA
oneway.test(score ~ chr, data=my_chip)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  score and chr
## F = 1.0228, num df = 18.0, denom df = 1797.5, p-value = 0.4298
```

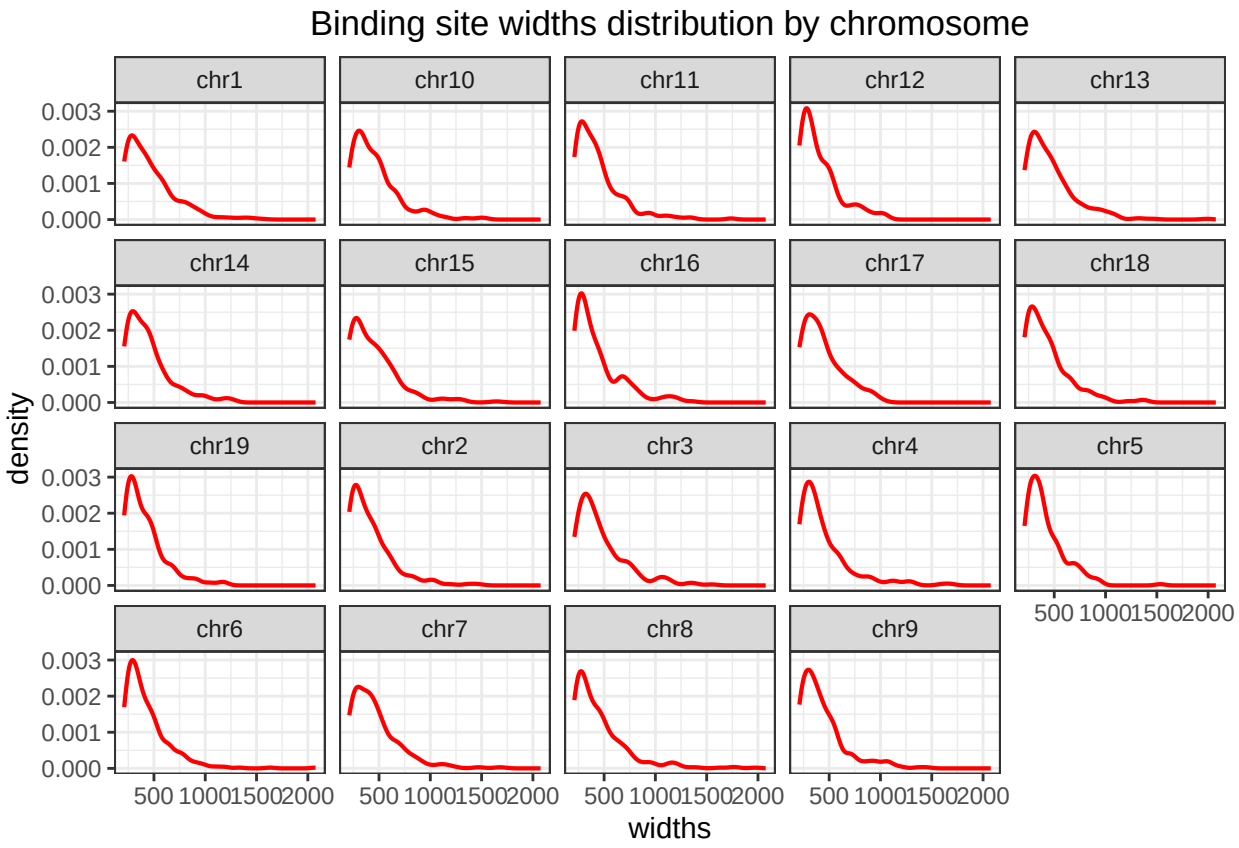
The null hypothesis is that there is no difference in the means of the distributions between the groups. The p-value is large (0.4298), we don't have enough evidence to reject the null hypothesis, so we can't state that the means of the binding scores distribution between chromosome are different. We don't have enough proof to state that the binding scores are dependent on the chromosomes.

b) Hypothesis 2

```
# Add binding site widths column to the tibble
my_chip %>% mutate(widths = end - start) -> my_chip

# Plot the binding widths ditributions by chromosome
```

```r
my_chip %>% ggplot(aes(x=widths)) +
  geom_line(col="red", size=0.75, stat="density") + theme_bw() +
    facet_wrap(~chr) +
    labs(title="Binding site widths distribution by chromosome") +
    theme(plot.title = element_text(hjust = 0.5))
```



To test the second hypothesis, as a preprocessing step, we added the binding site widths column to the ChIP-seq data. Then, we the plotted the widths binding site distribution of each chromosome, and they seem to be truncated normally distributed.

Since we are not sure if we can use the ANOVA test with truncated normally distributed data, we will use the Kruskal-Wallis test, which is similar to oneway ANOVA but it doesn't require normally distributed data.

Null: Binding site widths between chromosomes have a location shift of 0.

```r
# Kruskal-Wallis test
kruskal.test(widths ~ chr, data=my_chip)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  widths by chr
## Kruskal-Wallis chi-squared = 38.411, df = 18, p-value = 0.003416
```

Kruskal-Wallis test is the analogous of the wilcoxon test, the null hypothesis is that the distributions have U difference, or location shift, of 0. The obtained p-value is smaller than 0.05 (0.003416). We can reject the null hypothesis and we can state that the groups don't follow the same distributions, at least one pair in the group should have a different location shift, meaning that the gene width is dependent on chromosome.

To determine which chromosomes are the most different we will perform a pairwise wilkox test.

Null: Binding site widths between pairs of chromosomes have a location shift of 0.

```
# Pairwise wilkox test
pairwise.wilcox.test(my_chip$widths, my_chip$chr, p.adjust.method ="fdr")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  my_chip$widths and my_chip$chr
##
##       chr1  chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr2
## chr10 0.937 -     -     -     -     -     -     -     -     -     -     -
## chr11 0.410 0.370 -     -     -     -     -     -     -     -     -     -
## chr12 0.129 0.092 0.405 -     -     -     -     -     -     -     -     -
## chr13 0.797 0.901 0.251 0.092 -     -     -     -     -     -     -     -
## chr14 0.797 0.719 0.671 0.207 0.591 -     -     -     -     -     -     -
## chr15 0.717 0.655 0.839 0.362 0.521 0.901 -     -     -     -     -     -
## chr16 0.237 0.187 0.716 0.771 0.137 0.370 0.619 -     -     -     -     -
## chr17 0.797 0.719 0.797 0.302 0.608 0.937 0.943 0.548 -     -     -     -
## chr18 0.393 0.302 0.914 0.548 0.237 0.614 0.771 0.817 0.719 -     -     -
## chr19 0.137 0.092 0.548 0.816 0.092 0.237 0.405 0.901 0.405 0.702 -     -
## chr2  0.144 0.123 0.521 0.937 0.092 0.251 0.395 0.806 0.381 0.662 0.910 -
## chr3  0.816 0.875 0.302 0.092 0.998 0.655 0.546 0.154 0.667 0.272 0.092 0.117
## chr4  0.381 0.272 0.940 0.405 0.222 0.601 0.806 0.719 0.739 0.946 0.593 0.556
## chr5  0.237 0.207 0.806 0.591 0.144 0.405 0.680 0.864 0.566 0.914 0.771 0.719
## chr6  0.256 0.207 0.901 0.411 0.144 0.521 0.741 0.741 0.683 0.993 0.594 0.546
## chr7  0.943 0.993 0.405 0.137 0.901 0.774 0.683 0.237 0.771 0.381 0.144 0.157
## chr8  0.405 0.302 0.993 0.405 0.237 0.667 0.843 0.719 0.801 0.914 0.528 0.518
## chr9  0.284 0.237 0.901 0.490 0.187 0.528 0.717 0.801 0.683 0.996 0.671 0.657
##       chr3  chr4  chr5  chr6  chr7  chr8
## chr10 -     -     -     -     -     -
## chr11 -     -     -     -     -     -
## chr12 -     -     -     -     -     -
## chr13 -     -     -     -     -     -
## chr14 -     -     -     -     -     -
## chr15 -     -     -     -     -     -
## chr16 -     -     -     -     -     -
## chr17 -     -     -     -     -     -
## chr18 -     -     -     -     -     -
## chr19 -     -     -     -     -     -
## chr2  -     -     -     -     -     -
## chr3  -     -     -     -     -     -
## chr4  0.256 -     -     -     -     -
## chr5  0.187 0.876 -     -     -     -
## chr6  0.190 0.950 0.884 -     -     -
## chr7  0.901 0.370 0.239 0.272 -     -
## chr8  0.272 0.963 0.806 0.929 0.405 -
## chr9  0.237 0.939 0.901 0.978 0.302 0.876
##
## P value adjustment method: fdr
```

The p-value from the Wilcoxon-test are all above 0.05, so we cannot reject the null hypothesis, that there is no difference in location shift in the distribution between any pair of chromosome. The p-values is larger with the Wilcoxon-test compared to the Kruskal-Wallis due to the p-value adjustment, derived from the number

of paired test conducted. Based on the number of tests performed, we choose to used the false discovery rate p-value adjustment, which is less conservative than the others correction methods.

With the conflicting result from the Kruskal-Wallis test and the Wilcoxon-test we can conclude: The width of the gene is dependent of the chromosome as concluded from the Kruskal-Wallis test. But with the Wilcoxon-test we cannot conclude which of the chromosome pairs shows significantly different and we would need more data to prove which chromosome has significantly different genes width.