Home work 1

Homework 1

Names: Abdullah Faqih AlMubarok,August Nygaard Bodilsen, Natalie Christiansen, Maria Madrazo I Montoya

Group: 21

Question 1

Install the package babynames and look at the data babynames:

```
install.packages("babynames")
```

```
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.

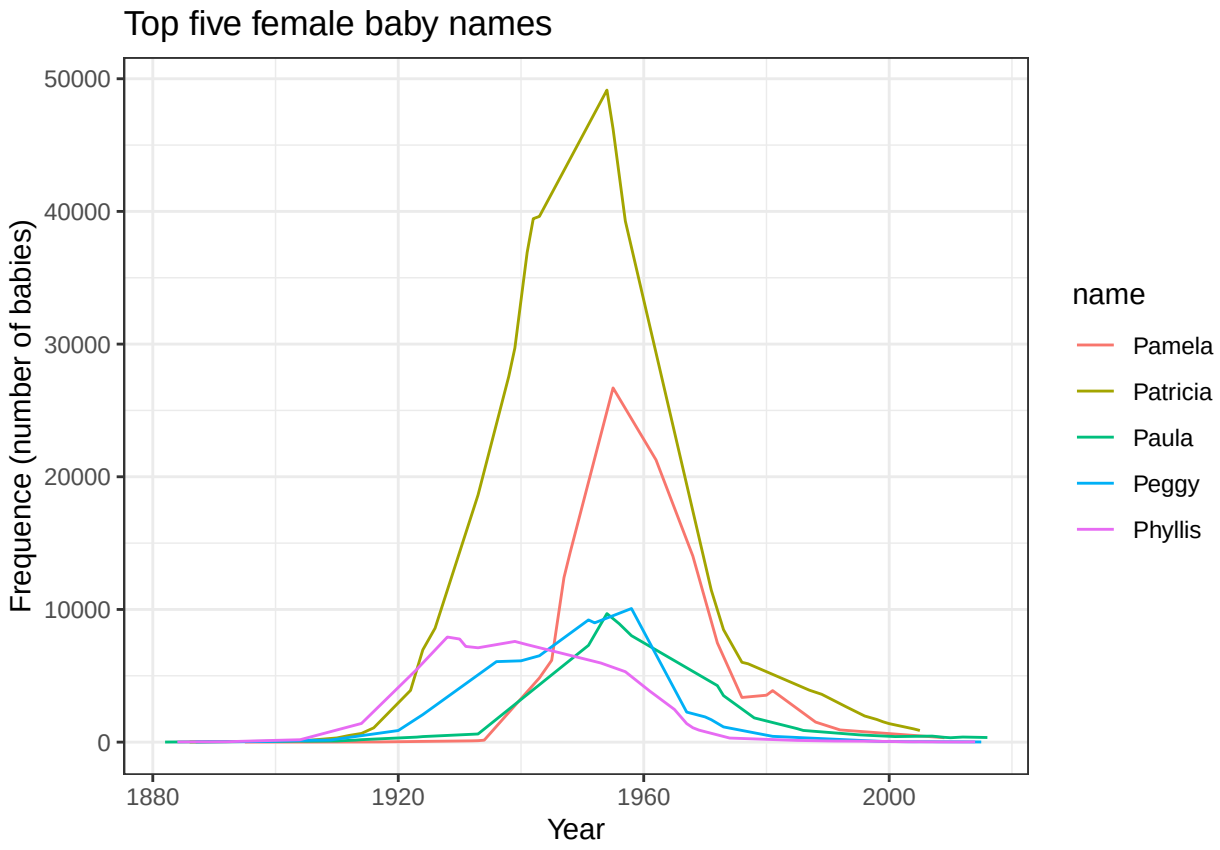
```
result1<-babynames %>%
  filter(sex=="F" & str_detect(name, "^P"))%>%
  group_by(name)%>%
  summarise(total = sum(n))%>%
  arrange(desc(total))%>%
  head(5)

result.table <- result1%>%
  as.matrix()%>%
  as.table()
row.names(result.table)<-c(1:5)
result.table
```

```
##   name     total
## 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.

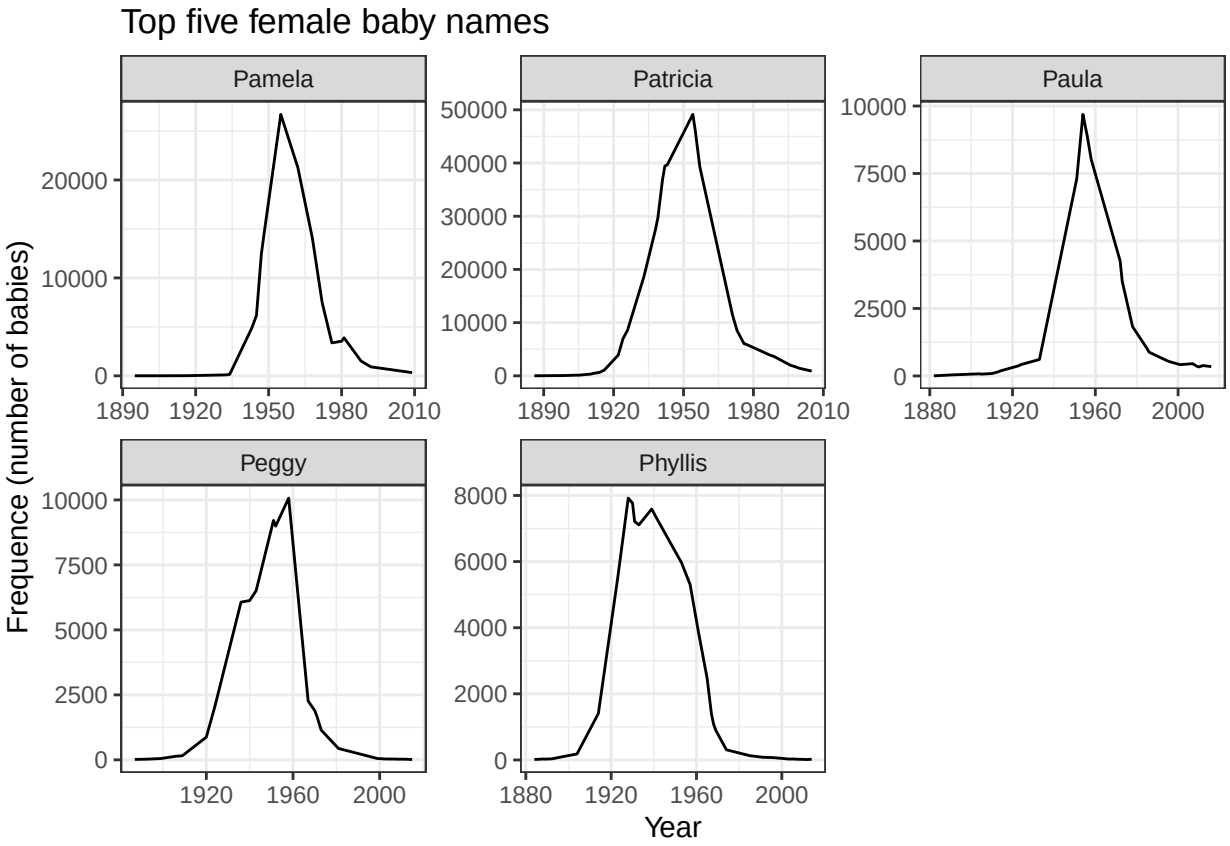
```
result2<- result1$name
top5 <- babynames %>%
         filter(name==result2 & sex=="F" )
ggplot(top5, aes(x = year, y = n)) +
  geom_line(aes(col = name)) +
  xlab("Year") +
  ylab("Frequence (number of babies)") +
  ggtitle("Top five female baby names") +
  theme_bw()
```



The plot shows that the most common female name starting with P was Patricia, between 1930 and 1950. In second place, it was Pamela around 1940 and 1970. The rest of the names had similar ocurrences. After 1950, the level of ocurrence for these names gradually dropped to very low levels. The absence of values before 1910 could be due to lack of registration or low popularity of the names.

The graphs in the plot overlap, this could be solved by separating the graphs based on name.

```
ggplot(top5, aes(x = year, y = n)) +
  geom_line() +
  xlab("Year") +
  ylab("Frequence (number of babies)") +
  ggtitle("Top five female baby names") +
  facet_wrap(~name, scales="free")+
  theme_bw()
```



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.

We use fischers exact test to compare if the count of girls named Arwen in 2004 significantly differs from the count in 1990. In other words, we will test if the relative proportion of babies named Arwen in 2004 is different than in 1990. We use fishers exact test instead of the chi-square test, due to the count of people named Arwen being very low.'

H0: The odds ratio is 1. There is no relationship between rows and columns in our test matrix.
Halt: The odds ratio is not 1.
alpha= 0.05

```
babynames %>%
  group_by(year) %>%
  mutate(other_name = sum(n)-n) %>%
  filter(name == "Arwen", year == 1990|year == 2004) %>%
  arrange(desc(year)) %>%
  select(year, n, other_name) -> edited_babynames

fisher_test_matrix = as.matrix(column_to_rownames(edited_babynames, "year"))
fisher_test_matrix
```

```
##        n other_name
## 2004 166    3818195
## 1990  10    3950982
```

```
fisher.test(fisher_test_matrix)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  fisher_test_matrix
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   9.108985 36.524229
## sample estimates:
## odds ratio
##   17.17996
```

Since we obtain a p-value $<$ 2.2e-16, which is $<$ alpha, we can reject the null hypothesis. So, there is a significant difference (an increase) in the number of babies named Arwen in 2004 compared to 1990. This could be due to the release of the movies "Lord of The Rings", since one of the characters had this name.

Conclusion: The P-value is $<$ 2.2e-16, so we can be very confident that the odds ratio is not 1. Our estimated odds ratio is 17.17996, indicating that there is a much larger chance of being named Arwen when born in 1990 compared to being born in 2004. This could be due to the release of the movies "Lord of The Rings", since one of the characters had this name.

Question 3

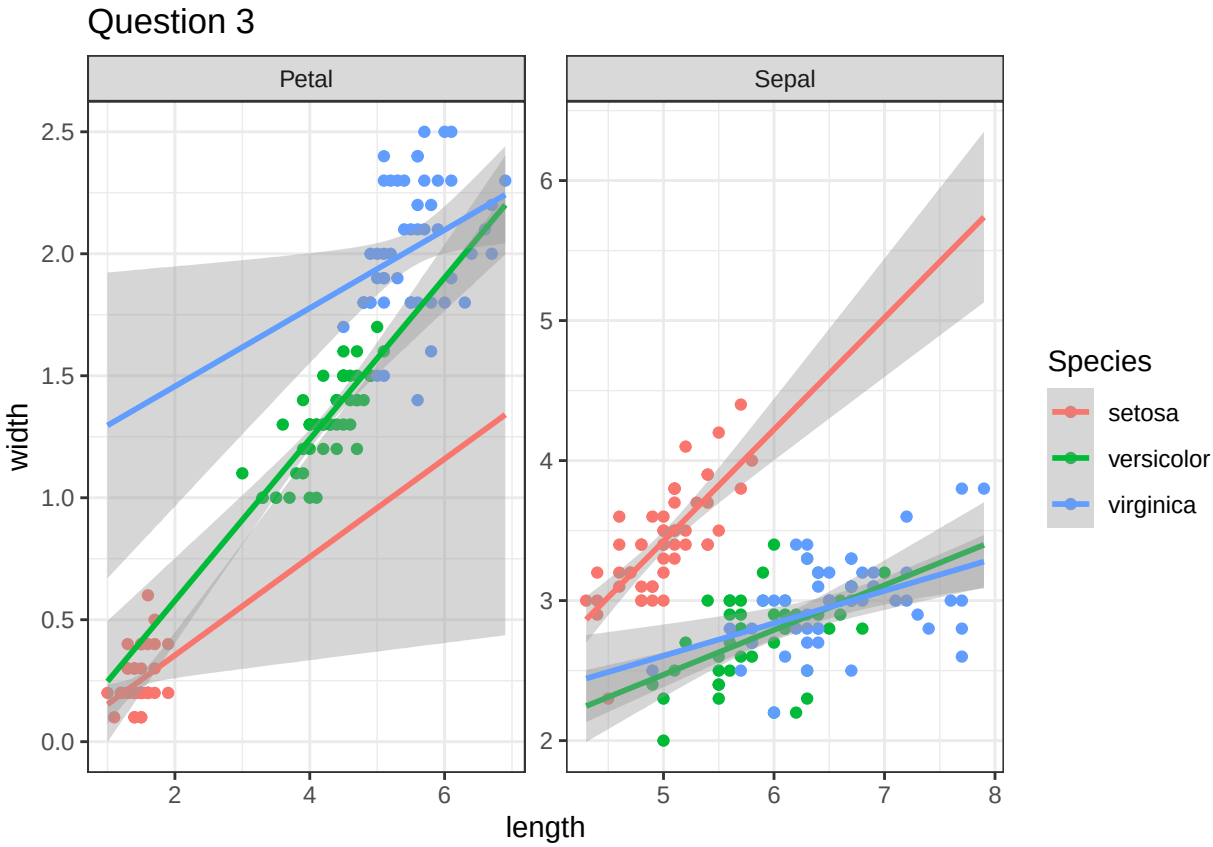
Produce the following plot starting from the flowers dataset. A potentially useful function that you may not have seen: bind_rows(): merges two tibbles by rows so that the joint tibble becomes longer, not wider

```
iris_petal <- iris %>% select(starts_with("Petal"), Species) %>%
  mutate(Dimension = "Petal", width = Petal.Width, length = Petal.Length) %>%
  select(Species, length, width, Dimension)

iris_sepal <- iris %>% select(starts_with("Sepal"), Species) %>%
  mutate(Dimension = "Sepal", width = Sepal.Width, length = Sepal.Length) %>%
  select(Species, length, width, Dimension)

bind_rows(iris_petal, iris_sepal) %>%
  ggplot(aes(y=width, x=length, col=Species)) + geom_point() +
  geom_smooth(method="lm", fullrange=TRUE) + facet_wrap(~Dimension, scales="free") +
  theme_bw()+ggtitle("Question 3")
```

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



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 hypotheses:

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

Can you prove/disprove these two hypotheses statistically?

```
chip<- read_tsv("chip_mm5.txt")
```

```
## Rows: 5415 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): chr
## dbl (3): start, end, score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Hypothesis 1: Binding scores are dependent on chromosome

To answer the first hypothesis, we need to compare if any of the chromosomes are different in means for binding scores.

H0= There are no differences in mean binding score between chromosomes

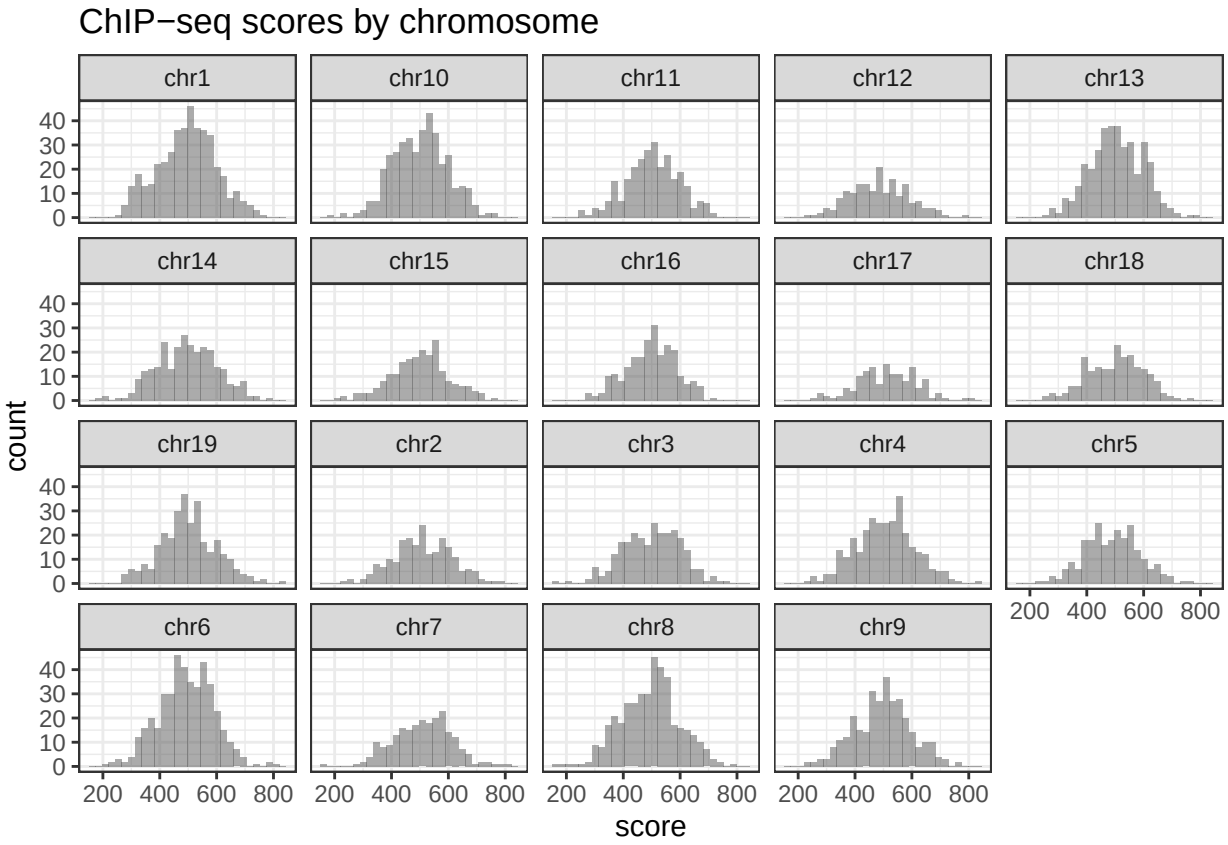
Halt= The mean binding scores differ between the chromosomes,
alpha= 0.05

First, we inspect the distribution of scores for the individual chromosomes:

```
chip %>% ggplot(aes(x=score)) + geom_histogram(alpha=0.5) + facet_wrap(~chr) +
theme_bw()+
ggtitle("ChIP-seq scores by chromosome")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing non-finite values (`stat_bin()`).
```



The test of choice will be one-way ANOVA test because we are working with a continuous dependent variable (score) and one categorical predictor (chromosome), as well as the assumption of normally distributed scores.

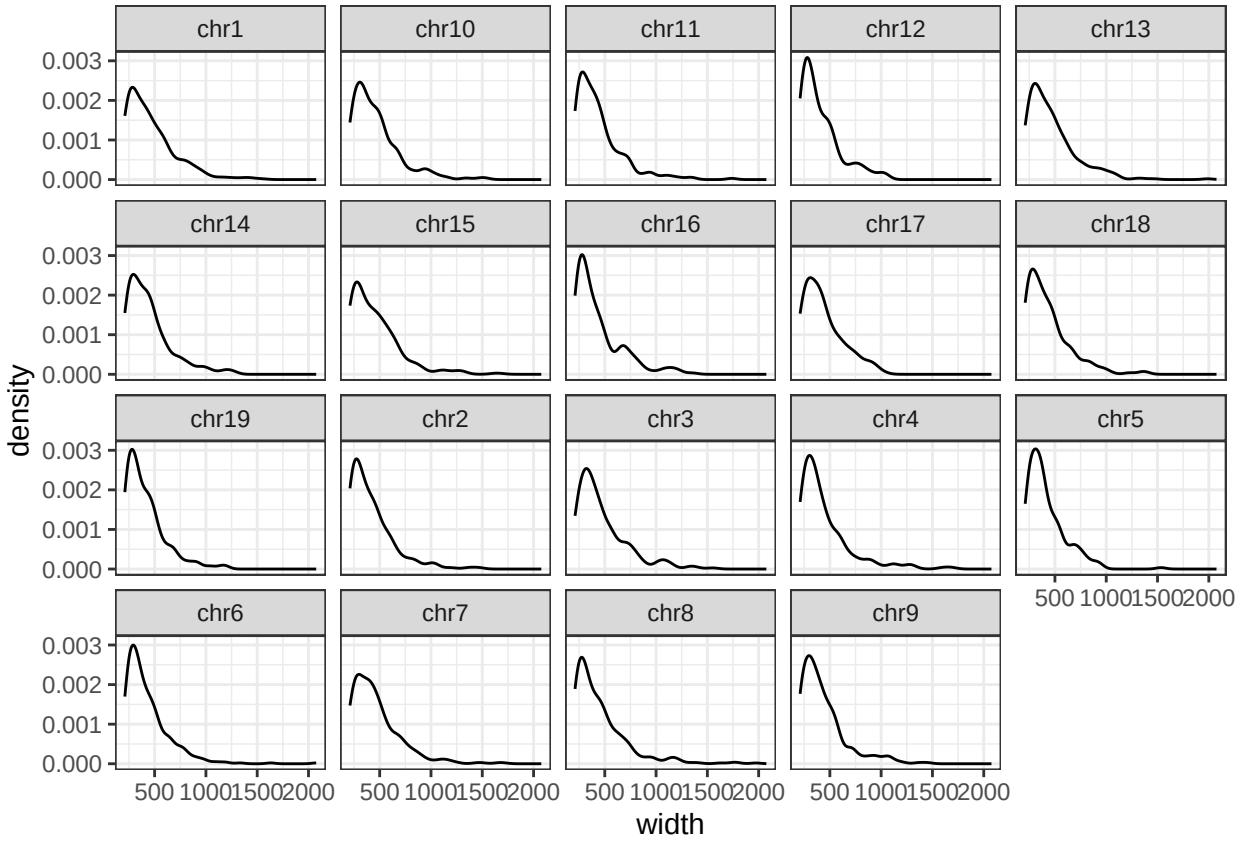
```
oneway.test(score ~ as.factor(chr), data=chip)
```

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

Binding score does not seem to be dependent on chromosome. With a p-value of 0.4298 (above our threshold), we cannot reject the null hypothesis.

Hypothesis 2: Binding site widths (end-start) are dependent on chromosome
To answer the second hypothesis, we will create a new column called "width", which takes into account the start and end of the binding. Once again, we will test if width follows a normal-like distribution on the chromosomes.

```
chip=mutate(chip, width=end-start)
ggplot(chip,aes(x=width))+
  geom_density()+theme_bw()+facet_wrap(~chr)
```



In this case, the binding site width does not follow a normal distribution. Because we have more than three chromosome groups to analyze, we will be using the non-parametric version: Kruskal Wallis.

H0= The mean ranks of binding site widths are equal between chromosomes
Halt= The mean ranks of binding site widths are not equal between chromosomes, alpha = 0.05

```
kruskal.test(width ~ as.factor(chr), data=chip)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  width by as.factor(chr)
## Kruskal-Wallis chi-squared = 38.536, df = 18, p-value = 0.003288
```

We get a p-value of 0.003288 (below our threshold), so we reject the null hypothesis. Therefore, binding site width does seem to be dependent on chromosome.