```
---
title: "Lab Chapter 9"
author: "Your Name Here"
date: ""
output: html_document
---

##### Remember to change the `author: ` field on this Rmd file to your own name.

We'll begin by loading all the packages we might need.
```{r}
library(MASS)
library(plyr)
library(reshape)  # You may need to install this one first!
library(ggplot2)
```

Testing means between two groups

(a) Using the Cars93 data and the `t.test()` function, run a t-test to see if average `MPG.highway` is different between US and non-US vehicles.

Try doing this both using the formula style input and the `x`, `y` style input. `MPG.highway ~ Origin`

```{r}
# # Formula version
mpg.t.test <- t.test(    )
mpg.t.test

# x, y version # Complete the y
with(Cars93, t.test(x = MPG.highway[Origin == "USA"], y =  ))
```

(b) What is the confidence interval for the difference?

```{r}
# Edit me: Use mpg.t.test$conf.int

```

(c) Repeat part (a) using the `wilcox.test()` function.

```{r}
# Edit me
mpg.wilcox.test <- wilcox.test(  , data = Cars93)
mpg.wilcox.test
```

(d) Are your results for (a) and (c) very different?

Is the data normal?

(a) Construct histograms of `MPG.highway`, one plot for each `Origin`
```

category.

```{r}
# Edit me
qplot(x = , data = Cars93, facets = ~Origin, geom = "histogram", fill =
Origin, binwidth = 2)
```

(b) Does the data look to be normally distributed?

(c) Construct qqplots of `MPG.highway`, one plot for each `Origin` category. Overlay a line on each plot using with `qqline()` function.

```{r}
# Edit me
par(mfrow = c(1,2))
# USA cars
with(Cars93, qqnorm(MPG.highway[Origin == "USA"]))
with(Cars93, qqline(MPG.highway, col = "blue"))
# Foreign cars
with(Cars93, qqnorm(MPG.highway[Origin == "non-USA"]))
with(Cars93, qqline(MPG.highway, col = "blue"))

```

(d) Does the data look to be normally distributed?

Testing 2 x 2 tables

Doll and Hill's 1950 article studying the association between smoking and lung cancer contains one of the most important 2 x 2 tables in history.

Here's their data:

```{r}
smoking <- as.table(rbind(c(688, 650), c(21, 59)))
dimnames(smoking) <- list(has.smoked = c("yes", "no"),
                    lung.cancer = c("yes","no"))
smoking
```

(a) Use `fisher.test()` to test if there's an association between smoking and lung cancer.

```{r}
# Edit me
smoking.fisher.test <- fisher.test(smoking)
smoking.fisher.test
```

(b) What is the odds ratio?

```{r}
# Edit me use smoking.fisher.test$estimate
```

(c) Are your findings significant?

```{r}
# Edit me use smoking.fisher.test$p.value
```

Plotting error bars

This exercise will take a long time if you choose to pursue it.

Using Doll and Hill's smoking data and, construct a bar graph with accompanying error bars showing the proportion of study participants with lung cancer.

To succeed in this exercise, you'll have to follow along careful with the lecture notes. Please read the section titled "Plotting the table values with confidence".

```{r}
# 'melt' table into a data frame
smoking.melt <- melt(smoking, id=c("lung.cancer", "has.smoked"))

# Add a row sums column giving denominators (# smoked, # not smoked) for proportions
smoking.melt <- transform(smoking.melt, totals = rowSums(smoking)[rev(has.smoked)])

# define list of prop, and lower and upper endpoints
conf.ints <- list(prop = NULL, lower = NULL, upper = NULL)
for(i in 1:nrow(smoking.melt)) {
  numerator <- smoking.melt$value[i]
  denominator <- smoking.melt$totals[i]
  prop.test.out <- prop.test(numerator, denominator)

  # Add estimate of proportion to list
  conf.ints[["prop"]][i] <- prop.test.out$estimate
  # Grab confidence interval
  interval <- prop.test.out$conf.int
  # Add estimate and endpoints to conf.ints list
  conf.ints[["lower"]][i] <- interval[1]
  conf.ints[["upper"]][i] <- interval[2]
}

# Append confidence interval information to melted data frame
smoking.toplot <- cbind(smoking.melt, conf.ints)

# Use ggplot to construct plot
ggplot(smoking.toplot[smoking.toplot$lung.cancer == "yes",],
aes(x=has.smoked, y=prop, fill=has.smoked)) +
  geom_bar(position="dodge", stat="identity") +
  geom_errorbar(aes(ymin=lower, ymax=upper),
                width=.2,                    # Width of the error bars
                position=position_dodge(0.9)) +
  ylab("proportion of participants with lung cancer")
```

ANOVA with birthwt data

Let's form our favourite birthwt data set.

```{r}
# Rename the columns to have more descriptive names
colnames(birthwt) <- c("birthwt.below.2500", "mother.age",
"mother.weight",
    "race", "mother.smokes", "previous.prem.labor", "hypertension",
"uterine.irr",
    "physician.visits", "birthwt.grams")

# Transform variables to factors with descriptive levels
birthwt <- transform(birthwt,
            race = as.factor(mapvalues(race, c(1, 2, 3),
                              c("white","black", "other"))),
            mother.smokes = as.factor(mapvalues(mother.smokes,
                              c(0,1), c("no", "yes"))),
            hypertension = as.factor(mapvalues(hypertension,
                              c(0,1), c("no", "yes"))),
            uterine.irr = as.factor(mapvalues(uterine.irr,
                              c(0,1), c("no", "yes")))
            )
```

(a) Create a new factor that categorizes the number of physician visits into three levels: 0, 1, 2, 3 or more.

```{r}
# Edit me
```

```{r}
phys.visit.binned <- birthwt$physician.visits
phys.visit.binned[phys.visit.binned >= 3] <- "3.or.more"
birthwt <- transform(birthwt, phys.visit.binned =
as.factor(phys.visit.binned))

birthwt$phys.visit.binned
```

```

**Hint**: One way of doing this is with mapvalues, by mapping all instances of 3, 4,... etc, to "3 or more".

**(b)** Run an ANOVA to determine whether the average birth weight varies across number of physician visits.

```{r}
# Edit me

summary(aov(birthwt.grams ~ phys.visit.binned, data = birthwt))
```

The p-value is greater than 0.05, so the variation in birthweight across
```

number of physician visits is not statistically significant.