Econometrics HW #1 Solutions

Ryan Safner

Due: Tuesday, September 25, 2018

Note: Answers may be longer than I would deem sufficient on an exam. Some might vary slightly based on points of interest, examples, or personal experience. These suggested answers are designed to give you both the answer and a short explanation of *why* it is the answer.

Theory & Concepts

For the following questions, please answer the questions completely but succinctly (2-3 sentences).

1. Explain the difference between exogeneity and endogeneity.

An **exogenous** model is one where the independent variable (X) is not associated with any other factors that affect the dependent variable (Y). If a model is truly exogenous, we can estimate the **causal effect** of X on Y.

An **endogenous** model is one where the independent variable (X) *is* associated with any other factors that affect the dependent variable (Y). If a model is endogenous, we have no accurately estimated the causal effect of X on Y, since other factors are getting entangled with X and Y.

---

2. Explain how conducting a randomized controlled experiment helps to identify the causal connection between two variables.

Randomized controlled experiments are where a pool of subjects representative of a population are randomly assigned into a treatment group (or into one of a number of groups given different levels of a treatment) or into a control group. The treatment group(s) is(are) given the treatment(s), the control group is given nothing (perhaps a placebo, though this is not always necessary), and then the average results of the two groups are compared to measure the true average effect of treatment.

The key is that the assignment must be random, which controls for all factors that potentially determine the outcome (e.g. when measuring individual outcomes, their height, family background, income, race, age, etc). If subjects are randomly assigned, then knowing anything about the individual (e.g. age, height, etc) tells us *nothing* about whether or not they got the treatment(s). The *only* thing that separates a member of the treatment group(s) from the control group is whether or not they were assigned to treatment. This ensures that the average person in the treatment group(s) looks like the average person in the control group, and that we are truly comparing apples to apples, rather than apples to oranges.

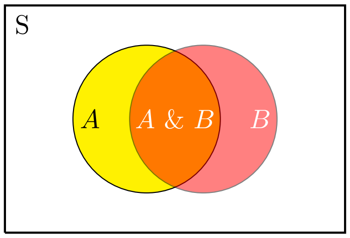
Theory Problems

For the following questions, please *show all work* and explain answers as necessary. You may lose points if you only write the correct answer. You may use R to verify your answers, but you are expected to reach the answers in this section "manually."

3. A college senior has applied for admission to two medical schools, *A* and *B*. She estimates the probability of acceptance at *A* at 0.7 and the probability of acceptance at *B* at 0.4 and the probability that she will be admitted to both at 0.2. What is the chance she will *not* be accepted at *either* school? (Consult the probability handout for basic probability rules.)

We are given $P(A) = 0.7, P(B) = 0.4$, and $P(A\&B) = 0.2$

Be careful, events A & B are *not* disjoint events! They can occur simultaneously, as you can get into multiple schools, so we must use the *general* addition rule of probability:



We are looking for the area in white, $1 - P(A \text{ or } B)$

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(\text{ and } B) \\ &= 0.7 + 0.4 - 0.2 \\ &= 0.9 \end{aligned}$$

This is the probability that she gets into both schools. The complement of this outcome, that she gets into neither school, is $P(\neg A \text{ and } \neg B) = 1 - P(A \text{ or } B)$

$$P(A \text{ nor } B) = 1 - 0.9 = 0.1$$

There is thus a 10% chance that she gets into neither school.

---

4. Suppose the probabilities of a visitor to Amazon's website and buying 0, 1, or 2 books are 0.2, 0.4, and 0.4 respectively. What are the *expected number* of books a visitor will purchase and the *standard deviation* of book purchases?

It's most helpful to make a table. The first column is the number of books bought (x_i), the second column is the associated probability of each amount of books being bought:

(1)	(2)	(3)	(4)	(5)
x_i	$P(x_i)$	$(x_i - \mu)$	$(x_i - \mu)^2$	$p_i(x_i - \mu)^2$
0	0.2	-1.2	1.44	0.228
1	0.4	-0.2	0.04	0.016
2	0.4	0.8	0.64	0.256
$\sum$				0.560

We first need to find the expected value:

$$\begin{aligned}\mu = E[X] &= \sum_{i=1}^{n} x_i p_i \\ E[X] &= 0(0.2) + 1(0.4) + 2(0.4) \\ E[X] &= (0) + (0.4) + (0.8) \\ E[X] &= 1.2\end{aligned}$$

Now that we have this, we can fill out the third column of the table, the deviation of each x_i value from the expected value of 1.2. Then, we square each deviation in the fourth column. The fifth column multiplies each squared deviation by its associated probability.

Finally, we add up all the probability-weighted squared deviations in column 5 to get the variance: 0.560.

Now we wanted the standard deviation of X, so we must square root the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.56} = 0.748$$

So our findings are that people will, on average, buy 1.2 books with a standard deviation of 0.75 books.

---

5. Scores on the SAT (out of 1600) are approximately normally distributed with a mean of 500 and standard deviation of 100.

a. What is the probability of getting a score between a 400 and a 600?

Let the random variable S be the SAT score. Then:

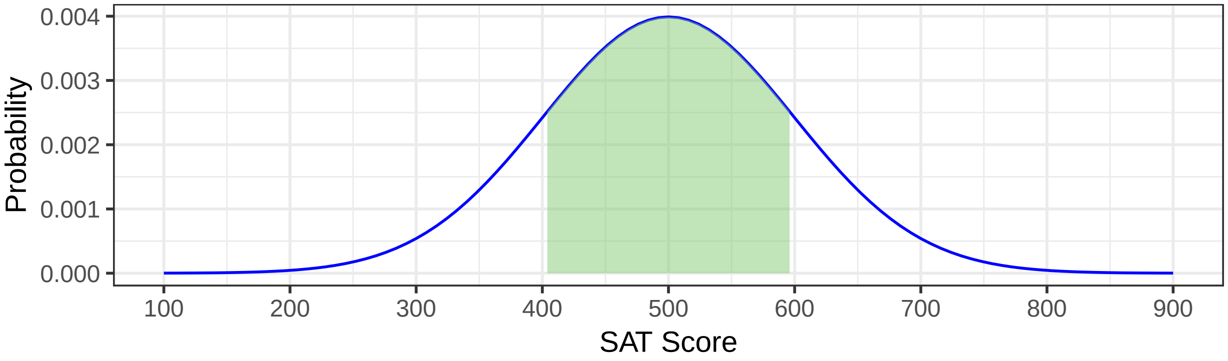
$$\begin{aligned}P(400 \leq S \leq 600) &= P\Big(\frac{400-500}{100} \leq \frac{S-500}{100} \leq \frac{600-500}{100}\Big) \\ &= P\big(-1 \leq Z \leq 1\big) \\ &\approx 0.68\end{aligned}$$

```
# you do not need to draw the pdf graph
# I draw it only so you can visualize what you are estimating

# It is easy to draw a graph in ggplot2
# It requires a bit more effort to shade in specific areas
# which is why the code looks more complicated than it need be
# create shaded area for normal function with mean 500, sd 100
# between 400 and 600

shade.a <- function(x) {
    y <- dnorm(x, mean = 500, sd = 100)
    y[x < 400 | x >600] <- NA
    return(y)
}

# create the plot
library("ggplot2") # we'll need ggplot2 for this
ggplot(data.frame(x=100:900),aes(x=x))+  # let the data be some variable x between 100-900
    # draw normal pdf with mean 500, sd 500
    stat_function(fun=dnorm, args=list(mean=500, sd=100), color="blue")+
    # add the shaded function defined above
    stat_function(fun=shade.a, geom="area", fill="#84CA72", alpha=0.5)+
    # manually define the ticks on x axis as sequence from 100 to 900 by 100
    scale_x_continuous(breaks=seq(100,900,100))+
    xlab("SAT Score")+ylab("Probability")+theme_bw() # fix labels on graph
```



---

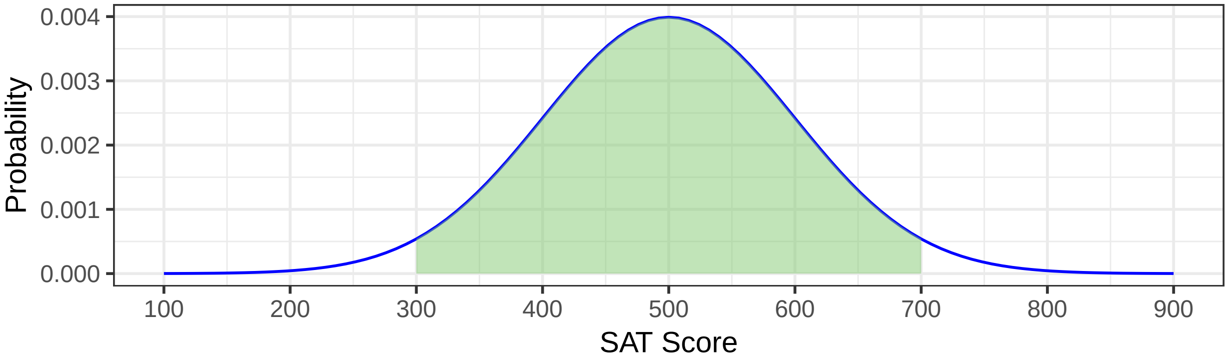
b. What is the probability of getting a score between a 300 and a 700?

$$\begin{aligned} P(300 \leq S \leq 700) &= P\Big(\frac{300-500}{100} \leq \frac{S-500}{100} \leq \frac{700-500}{100}\Big) \\ &= P\big(-2 \leq Z \leq 2\big) \\ &\approx 0.95 \end{aligned}$$

```
# create shaded area for normal function with mean 500, sd 100
# between 300 and 700

shade.b <- function(x) {
    y <- dnorm(x, mean = 500, sd = 100)
    y[x < 300 | x >700] <- NA
    return(y)
}

# create the plot
ggplot(data.frame(x=100:900),aes(x=x))+
    stat_function(fun=dnorm, args=list(mean=500, sd=100), color="blue")+
    stat_function(fun=shade.b, geom="area", fill="#84CA72", alpha=0.5)+
    scale_x_continuous(breaks=seq(100,900,100))+
    xlab("SAT Score")+ylab("Probability")+theme_bw()
```



---

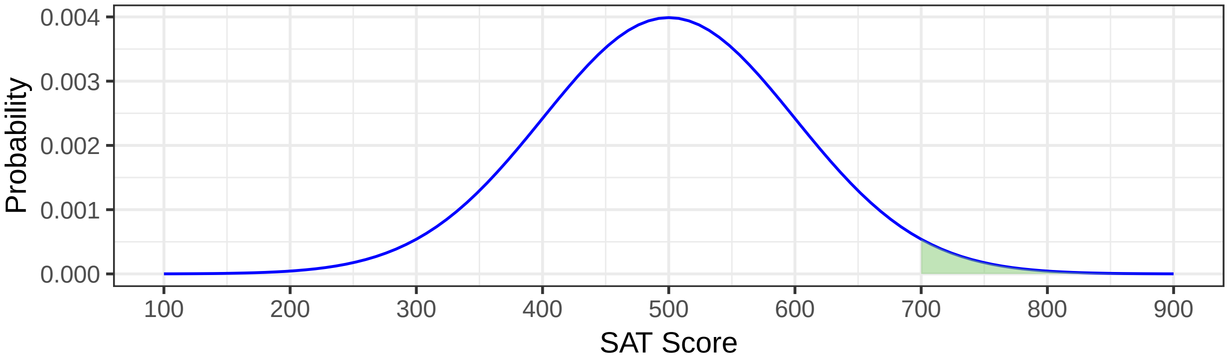
c. What is the probability of getting *at least* a 700?

$$
\begin{aligned}
P(S \geq 700) &= P\Big(\frac{S-500}{100} \geq \frac{700-500}{100}\Big) \\
&= P(Z \geq 2) \\
&\approx 0.025
\end{aligned}
$$

```
# create shaded area for normal function with mean 500, sd 100
# between 700 and 900

shade.c <- function(x) {
    y <- dnorm(x, mean = 500, sd = 100)
    y[x < 700 | x >900] <- NA
    return(y)
}

# create the plot
ggplot(data.frame(x=100:900),aes(x=x))+
    stat_function(fun=dnorm, args=list(mean=500, sd=100), color="blue")+
    stat_function(fun=shade.c, geom="area", fill="#84CA72", alpha=0.5)+
    scale_x_continuous(breaks=seq(100,900,100))+
    xlab("SAT Score")+ylab("Probability")+theme_bw()
```



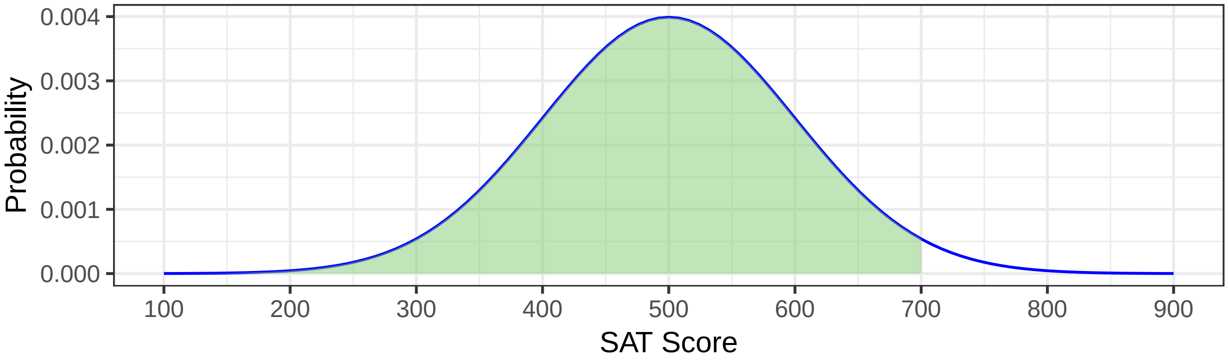
d. What is the probability of getting *at most* a 700?

$$
\begin{aligned}
P(S \leq 700) &= P\big(\frac{S-500}{100} \leq \frac{700-500}{100}\big) \\
&= P(Z \leq 2) \\
&\approx 0.975
\end{aligned}
$$

```
# create shaded area for normal function with mean 500, sd 100
# between 100 and 700

shade.d <- function(x) {
    y <- dnorm(x, mean = 500, sd = 100)
    y[x < 100 | x >700] <- NA
    return(y)
}

# create the plot
ggplot(data.frame(x=100:900),aes(x=x))+
    stat_function(fun=dnorm, args=list(mean=500, sd=100), color="blue")+
    stat_function(fun=shade.d, geom="area", fill="#84CA72", alpha=0.5)+
    scale_x_continuous(breaks=seq(100,900,100))+
    xlab("SAT Score")+ylab("Probability")+theme_bw()
```



e. What is the probability of getting exactly a 500?

This is a trick question to make sure you were paying attention about the pdf of a continuous variable. The probability of any particular value is so small, we define it as 0. We only can calculate the probability of a *range* of values. So

$$P(S = 500) = 0$$

R Problems

For the following problems, please attach/write the answers to each question on the same document as the previous problems, but also include a printed/attached (and commented!) *.R* script file of your commands to answer the questions.

6. Using the "table" method of finding standard deviation for a random variable discussed in class, use *R* to find the standard deviation of the following discrete random variable, X, that has the following pdf:

x_i	p_i
0	0.30
5	0.50
10	0.20

To jog your memory, standard deviation is the square root of the sum of the squared deviations from the mean weighted by the probability of the associated value of X.

```r
x.i<-c(0,5,10) # make vector of values of X
p.i<-c(0.3,0.5,0.2) # make vector of probabilities of each X values
rv<-data.frame(x.i,p.i) # make data frame of both

# find expected value: sum of values weighted by probability
sum(rv$x.i*rv$p.i) #it's 4.5
```

```
## [1] 4.5
```

```r
# find variance

# create a column of deviations from mean
rv$devs<-rv$x.i-4.5

# create a column of squared deviations
rv$devsq<-rv$devs^2

# create a column of the squared deviations weighted by probability
rv$w.devsq<-(rv$devsq*p.i)

# let's just check any make sure the table looks like it should
rv
```

```
##   x.i p.i devs devsq w.devsq
## 1   0 0.3 -4.5 20.25   6.075
## 2   5 0.5  0.5  0.25   0.125
## 3  10 0.2  5.5 30.25   6.050
```

```r
# take sum of probability-weighted squared deviations to find variance
variance<-sum(rv$w.devsq)

variance
```

```
## [1] 12.25
```

```r
# take square root of variance for standard deviation
sqrt(variance)
```

```
## [1] 3.5
```

---

7. Redo question 5 parts a-d using the `pnorm()` command in `R`.

```r
# See question 5 answers for the graphs again for reference

# part a

# note we are calculating the area between two values
# pnorm() actually takes the cdf of a value, i.e.
# the area of everything to the LEFT of a value "x", P(Z<x)
# so to get the probability between two values, take the
# cdf of the right value minues the cdf of the left value:
pnorm(600,mean=500,sd=100)-pnorm(400,mean=500,sd=100)
```

```
## [1] 0.6826895
```

```r
# part b
pnorm(700,mean=500,sd=100)-pnorm(300,mean=500,sd=100)
```

```
## [1] 0.9544997
```

```r
# part c
pnorm(700,mean=500,sd=100, lower.tail=TRUE)
```

```
## [1] 0.9772499
```

```r
# part d
pnorm(700, mean=500,sd=100, lower.tail=FALSE)
```

```
## [1] 0.02275013
```

```r
# or alternatively, since lower tail by default is set to TRUE
1-pnorm(700,mean=500,sd=100)
```

```
## [1] 0.02275013
```

---

8. We will use the dataset `mpg`, which is a part of the `ggplot2` package, and describes fuel economy data from the EPA on models of cars released between 1999-2008. Load (or install, if you don't have) the `ggplot2` package in order to use `mpg`.

a. What variables are included in the `mpg` data? (You don't need to explain them, they aren't well-documented, only write down they are).

```
#install.packages("ggplot2") # uncomment this (delete the first # in line)
# if you don't have ggplot2 installed

# you must install ggplot2 only once on a computer
# each time you want to use a package in a session, you
# must load it with library() once
library("ggplot2")
str(mpg) # get structure of mpg dataset
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

---

b. Find summary statistics for `hwy` and `cty` (miles per gallon on the highway and in the city, respectively).

```
# get summary statistics of 'hwy' variable in mpg dataset
summary(mpg$hwy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   18.00   24.00   23.44   27.00   44.00
```

```
# get summary statistics of 'cty' variable in mpg dataset
summary(mpg$cty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   14.00   17.00   16.86   19.00   35.00
```

---

c. How many different manufacturers are in the data, and how many cars from each manufacturer?

```
# create a frequency table of 'manufacturer' categorical variable in mpg dataset
table(mpg$manufacturer)
```

```
##
##       audi  chevrolet      dodge       ford      honda    hyundai
##         18         19         37         25          9         14
##       jeep land rover    lincoln    mercury     nissan    pontiac
##          8          4          3          4         13          5
##     subaru     toyota volkswagen
##         14         34         27
```

---

d. How many different classes are in the data, and how many cars from each class?

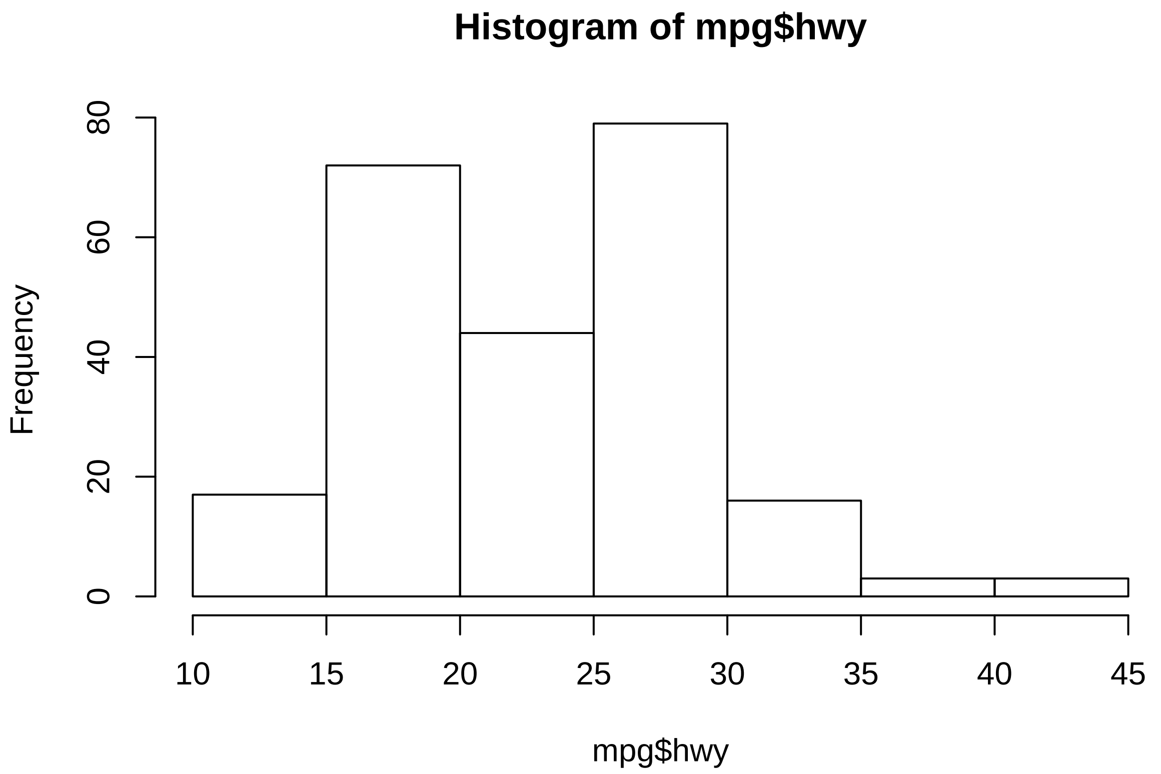
```
# create a frequency table of 'class' categorical variable in mpg dataset
table(mpg$class)
```

```
##
##    2seater    compact    midsize    minivan     pickup subcompact
##          5         47         41         11         33         35
##        suv
##         62
```

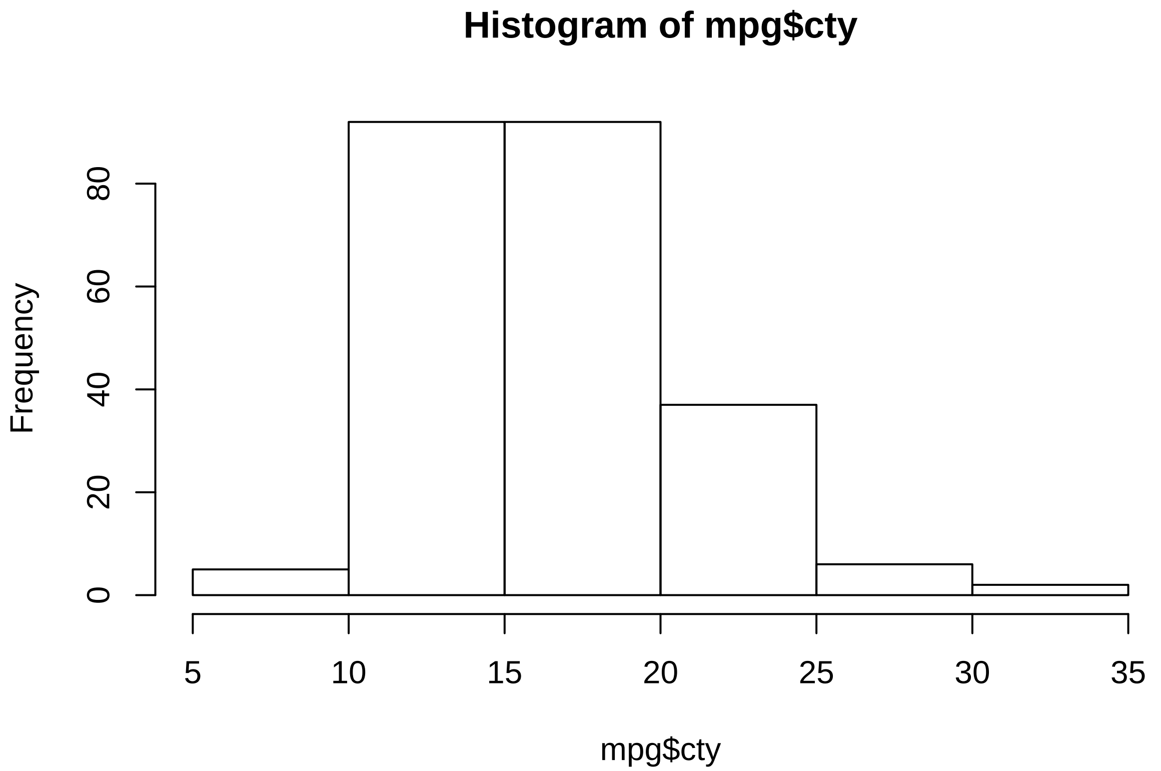
---

e. Plot two histograms, one of `hwy` and one of `cty`

```
hist(mpg$hwy) # create histogram of 'hwy' variable from mpg dataset
```

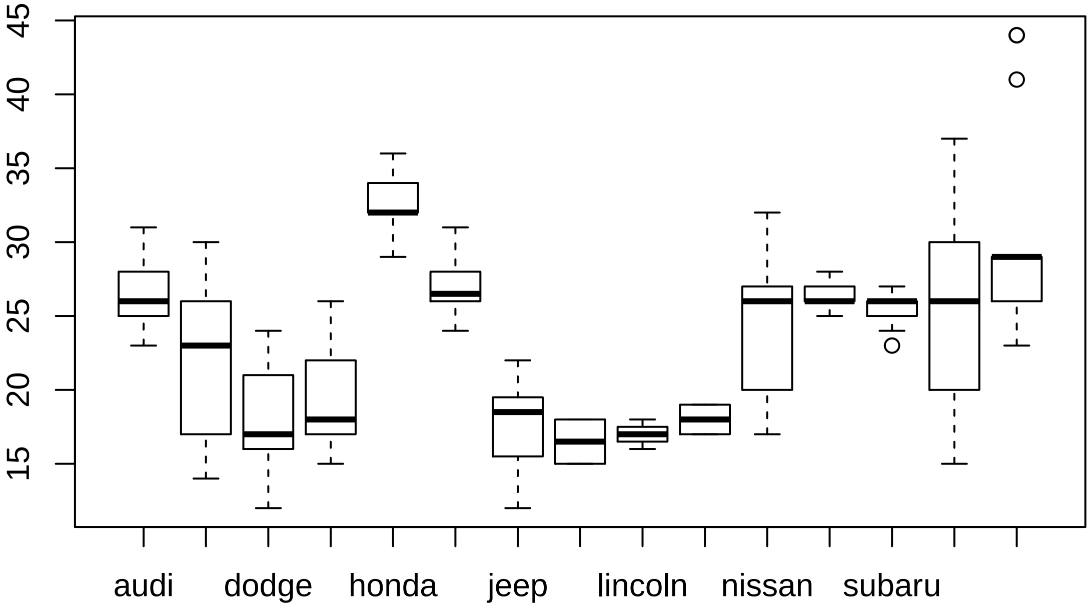


```r
hist(mpg$cty) # create histogram of 'cty' variable from mpg dataset
```



f. Plot a boxplot of `hwy` mpg by `manufacturer`. Which manufacturer appears to have the highest average highway mpg? Which appears to have the largest variance? Which appears to have outliers? (There are too many manufacturers for `R` to print them on the axis by default, but they are listed in alphabetical order, check using your answers to part c.)

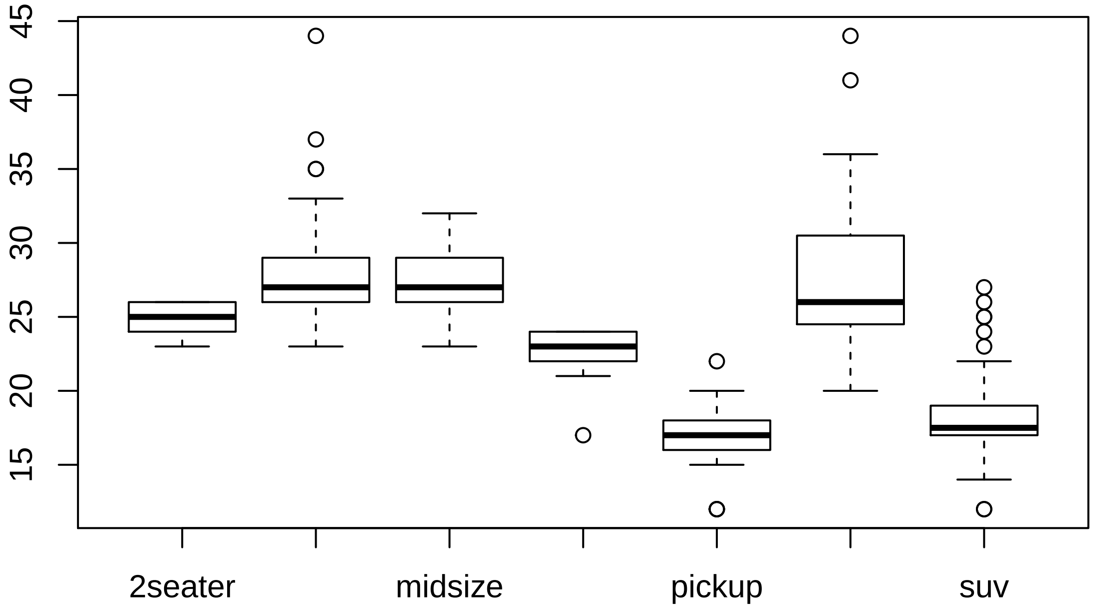
```r
# create a separate boxplot of 'hwy' over every category of 'manufacturer' variable, both from mpg data
boxplot(hwy~manufacturer,data=mpg)
```



---

g. Plot a boxplot of `hwy` by `class`. Which car classes appear to have the highest average highway mpg? Highest variance?

```r
# create a separate boxplot of 'hwy' over every category of 'class' variable, both from mpg dataset
boxplot(hwy~class,data=mpg)
```


45
40
35
30
25
20
15
2seater
midsize
pickup
suv

***