Sampling distributions, simulation, and the linear model

Frank Edwards

1/23/2024

School of Criminal Justice, Rutgers - Newark

Review HW1

- Challenges?

Sampling and sampling distributions

The sampling model and population inference

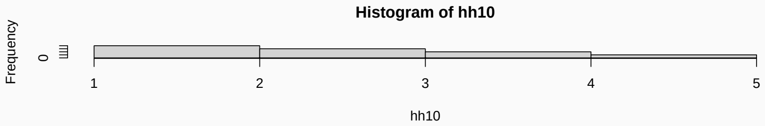
Under the **sampling model** we use a subset of the data to **infer** characteristics about the population.

I would like to know the average number of people living in a household in the United States.

Evaluating a sample

```r
draw_hh<-function(n){
  return(rpois(n, 1.53) + 1)
}

### sample 10 households
hh10<-draw_hh(10)
hist(hh10)
```



Infering population characteristics: mean

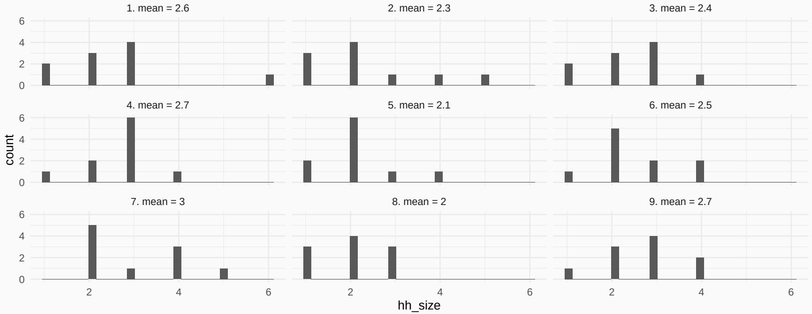
Let's assume this was a simple random sample (it was). We want to estimate μ, the population average household size. We've observed $\bar{hh}_{10}$, more commonly written as $\bar{x}$.

```
mean(hh10)
```

```
## [1] 2.8
```

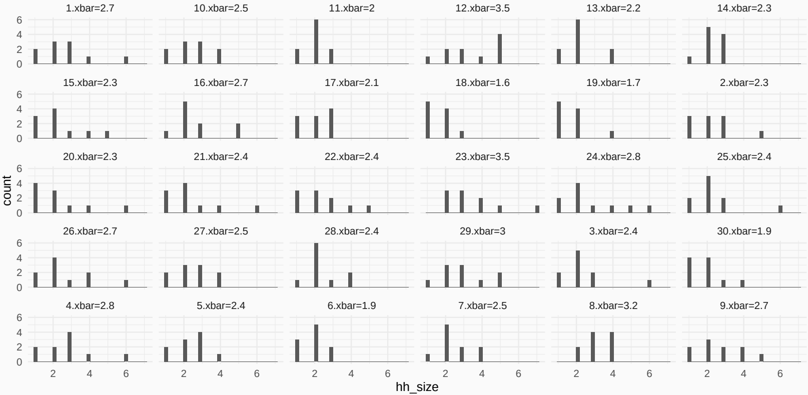
Describing uncertainty in our inference

We could have observed many possible samples of 10 households



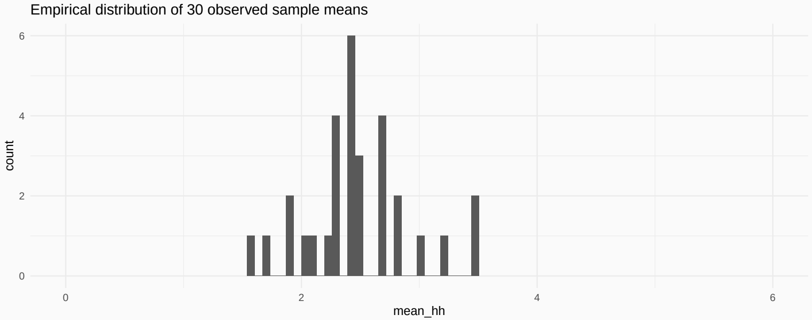
The approximate sampling distribution of hh_10

Each sample of 10 could draw any distribution of `hh_size`, here are 30 examples.



The sampling distribution of a parameter

Just as our sample has a theoretical sampling distribution, our estimate of the sample mean $\overline{x}$ has a sampling distribution.



Constructing a parameter estimate from a sampling distribution estimate

We can use the *central limit theorem* ($\bar{x} \sim N(\mu, \sigma)$ as $n \to \infty$) to estimate a sampling distribution for a parameter from our observed data.

We compute the sample mean ($\bar{x}$) and the *standard error* of the sample mean ($sd_x/\sqrt{n}$) to describe this distribution.

```
hh10 # the sample (x)
```

```
##  [1] 2 2 3 4 1 4 5 3 3 1
```

```
mean(hh10) # xbar
```

```
## [1] 2.8
```

```
sd(hh10) / sqrt(length(hh10)) # s_x
```

```
## [1] 0.4163332
```

Visualizing the sampling distribution of sample means

We can describe our uncertainty in the estimate of μ with the estimated sampling distribution for $\bar{x}$, or the possible values of the sample mean we *could have* observed based on these data.

Visualizing the sampling distribution of sample means

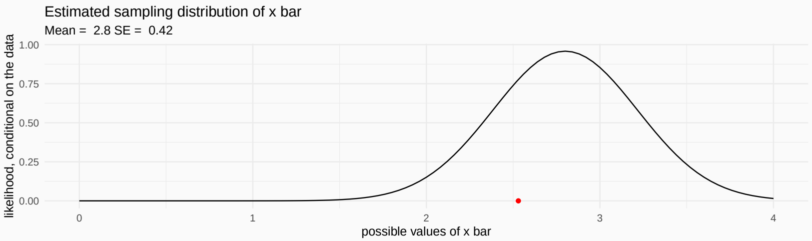
We can describe our uncertainty in the estimate of μ with the estimated sampling distribution for $\bar{x}$, or the possible values of the sample mean we *could have* observed based on these data.

$$\mu \sim Normal(\bar{x}, s_x)$$

Visualizing the sampling distribution of sample means

We can describe our uncertainty in the estimate of μ with the estimated sampling distribution for $\bar{x}$, or the possible values of the sample mean we *could have* observed based on these data.

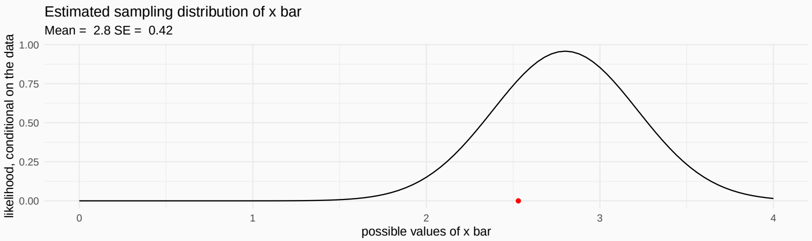
$$\mu \sim Normal(\bar{x}, s_x)$$



Visualizing the sampling distribution of sample means

We can describe our uncertainty in the estimate of μ with the estimated sampling distribution for $\bar{x}$, or the possible values of the sample mean we *could have* observed based on these data.

$$\mu \sim Normal(\bar{x}, s_x)$$

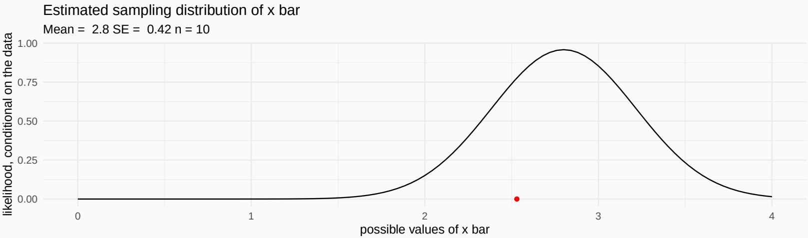


We use these estimates to describe our uncertainty in the value of the *population parameter* μ.

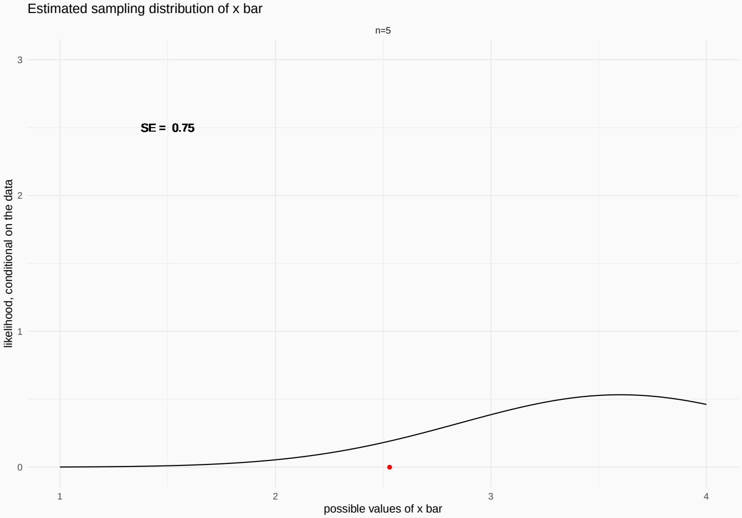
Question

Using this sampling distribution, compute a 95 percent confidence interval for μ.

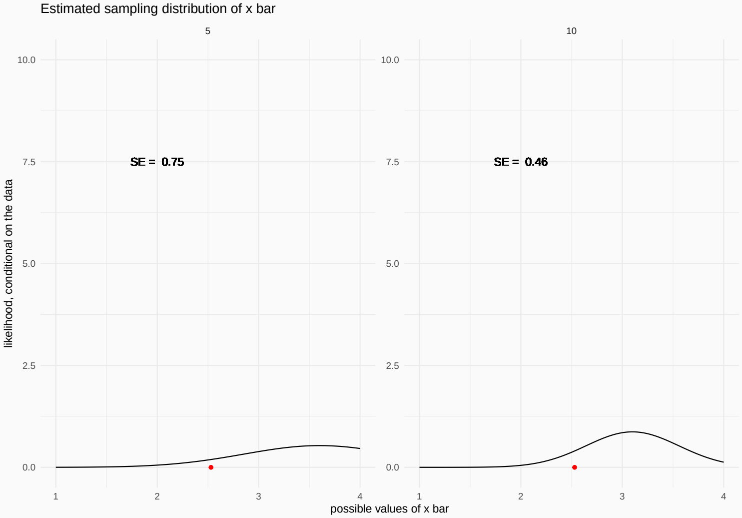
Hint: you can use `pnorm(0.025, 0, 1)` and `pnorm(0.975, 0, 1)` to obtain critical values for z.



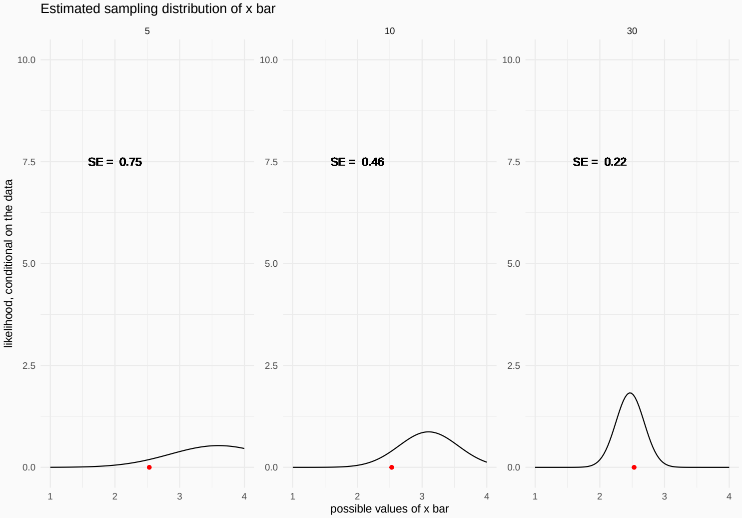
The sampling distribution of the mean



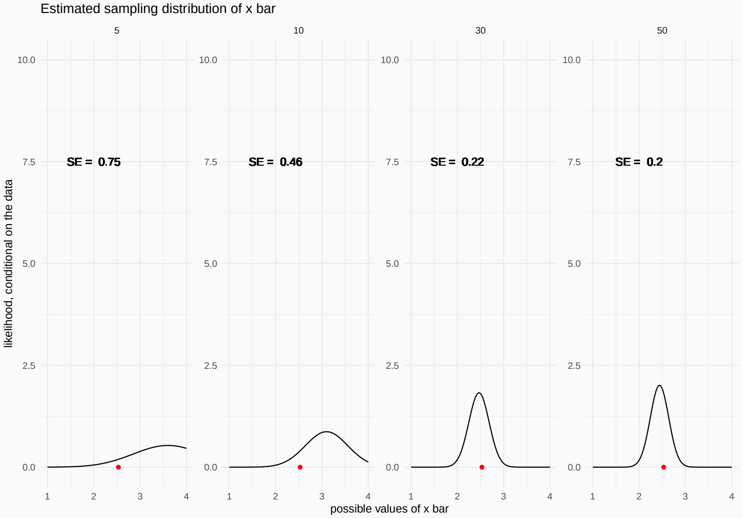
The sampling distribution of the mean



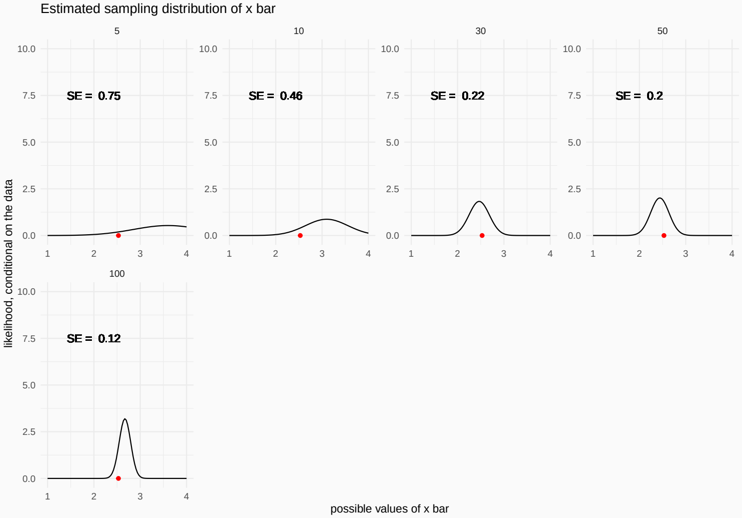
The sampling distribution of the mean



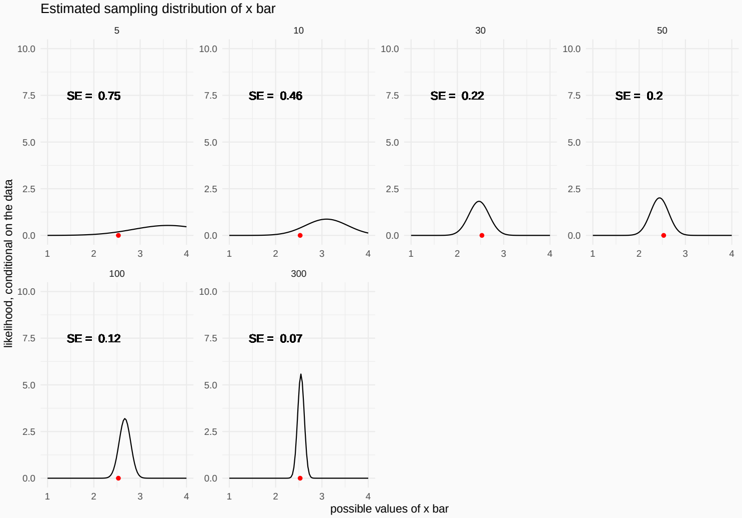
The sampling distribution of the mean



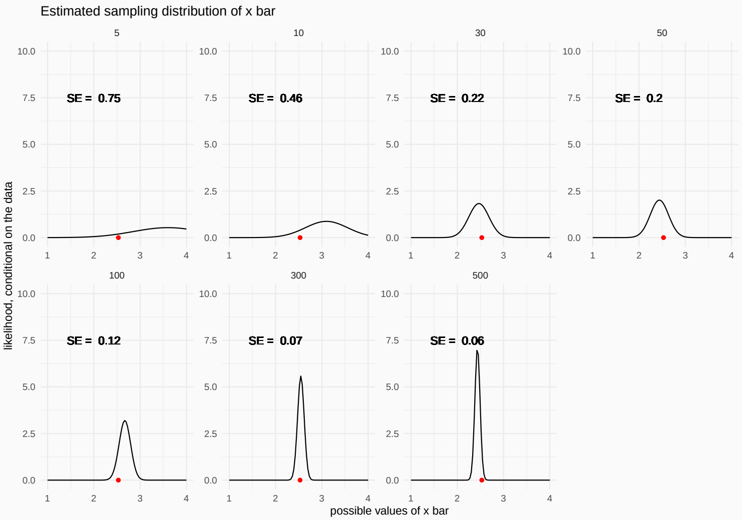
The sampling distribution of the mean



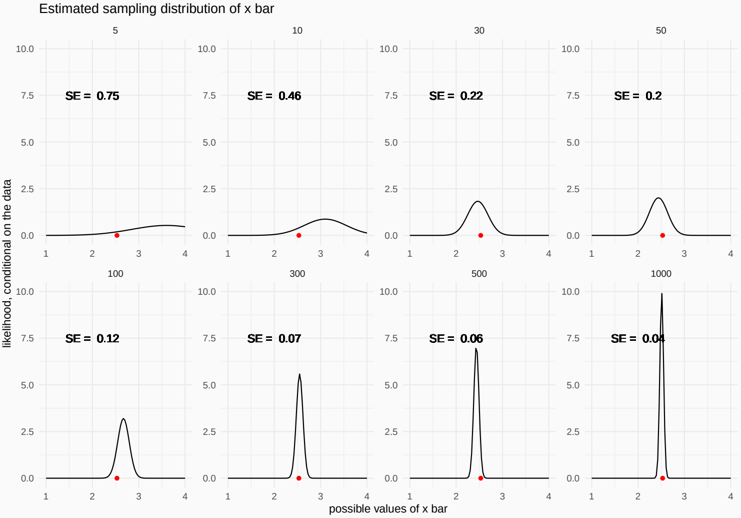
The sampling distribution of the mean



The sampling distribution of the mean



The sampling distribution of the mean



1. What is a parameter?

1. What is a parameter?
2. What is the difference between $\bar{x}$ and μ?

Review

1. What is a parameter?
2. What is the difference between $\bar{x}$ and μ?
3. What is the difference between a sample and a sampling distribution?

Review

1. What is a parameter?
2. What is the difference between $\bar{x}$ and μ?
3. What is the difference between a sample and a sampling distribution?
4. Briefly explain the logic of a confidence interval through the logic of a sampling distribution

1. Let's draw 50 samples with 100 households sampled

```
samp_hh<-data.frame(sample_n = rep(1:50, each = 100))
temp<-draw_hh(100)
for(i in 2:50){
  temp<-c(temp,
          draw_hh(100))
}

samp_hh <- samp_hh %>%
  mutate(hh_size = temp)
```

Confidence intervals and sampling distributions

1. Let's draw 50 samples with 100 households sampled
2. Let's compute 95 percent confidence intervals for $\bar{x}$ for each sample

```
samp_ci<-samp_hh %>%
  group_by(sample_n) %>%
  summarise(xbarhat = mean(hh_size),
            se = sd(hh_size)/sqrt(100)) %>%
  mutate(ci_lwr = xbarhat - 1.96 * se,
         ci_upr = xbarhat + 1.96 * se)
```

Confidence intervals and sampling distributions

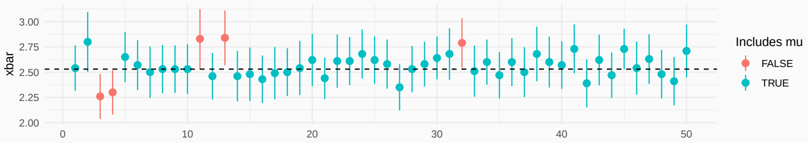
1. Let's draw 50 samples with 100 households sampled
2. Let's compute 95 percent confidence intervals for $\bar{x}$ for each sample
3. Let's add a binary variable indicating whether the interval includes μ (2.53)

```
samp_ci<- samp_ci %>%
  mutate(sig_test.95 = ci_lwr<2.53 & ci_upr>2.53)
```

Confidence intervals and sampling distributions

1. Let's draw 50 samples with 100 households sampled
2. Let's compute 95 percent confidence intervals for $\overline{x}$ for each sample
3. Let's add a binary variable indicating whether the interval includes μ (2.53)
4. Plot it!

```
ggplot(samp_ci,
       aes(ymin = ci_lwr, ymax = ci_upr,
           y = xbarhat, x = sample_n,
           color = sig_test.95)) +
  geom_pointrange() +
  geom_hline(yintercept = 2.53, lty = 2) +
  labs(x = "", y = "xbar", color = "Includes mu")
```



Break

Sampling distributions and regression parameters

We can apply the exact same logic to regression parameters. Let's use the `mpg` data to estimate the relationship between engine size (`displ`) and fuel efficiency (`hwy`).

```
glimpse(mpg)
```

```
## Rows: 234
## Columns: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "audi", "~
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 quattro", "~
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0, 2.0, 2.~
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1999, 200~
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 8, 8, ~
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)", "auto~
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4", "4", "4~
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 17, 17, 1~
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 25, 25, 2~
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p~
## $ class        <chr> "compact", "compact", "compact", "compact", "compact", "c~
```

Estimate the model

We model fuel efficiency as a linear function of engine size with the model

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x$$

```r
m0<-lm(hwy ~ displ, data = mpg)
```

What have we estimated?

```r
library(broom)
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. How does the `estimate` relate to the population mean?

What have we estimated?

```r
library(broom)
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. How does the `estimate` relate to the population mean?
2. What does the standard error tell us?

What have we estimated?

```r
library(broom)
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. How does the `estimate` relate to the population mean?
2. What does the standard error tell us?
3. What is `statistic`?

What have we estimated?

```
library(broom)
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. How does the `estimate` relate to the population mean?
2. What does the standard error tell us?
3. What is `statistic`?
4. What about that p value?

Let’s interpret the model

```r
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. What is the difference between $\hat{\beta}$ and β?

Let's interpret the model

```r
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displl          -3.53     0.195     -18.2 2.04e- 46
```

1. What is the difference between $\hat{\beta}$ and β?
2. What is β_0?

Let’s interpret the model

```r
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. What is the difference between $\hat{\beta}$ and β?
2. What is β_0?
3. What is β_1?

Let's interpret the model

```r
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. What is the difference between $\hat{\beta}$ and β?
2. What is β_0?
3. What is β_1?
4. Describe the relationship between engine size and fuel efficiency in terms of magnitude (M) and sign (S).

Let's interpret the model

```r
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. What is the difference between $\hat{\beta}$ and β?
2. What is β_0?
3. What is β_1?
4. Describe the relationship between engine size and fuel efficiency in terms of magnitude (M) and sign (S).
5. How certain are we in these findings? How precise are you willing to be?

Let's interpret the model

```r
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    35.7      0.720      49.6 2.12e-125
## 2 displ          -3.53     0.195     -18.2 2.04e- 46
```

1. What is the difference between $\hat{\beta}$ and β?
2. What is β_0?
3. What is β_1?
4. Describe the relationship between engine size and fuel efficiency in terms of magnitude (M) and sign (S).
5. How certain are we in these findings? How precise are you willing to be?
6. What assumptions have we made?

Homework

1. complete Chapters 2, 5, 6, and 7 from STAT 545 (`https://stat545.com/r-basics.html`)
2. complete Introduction to R Markdown (`https://rmarkdown.rstudio.com/articles_intro.html`)
3. Write a brief RMarkdown report explaining how you are feeling about writing R and markdown code, and explaining areas where you feel you need support.