Regression and uncertainty

Frank Edwards

Large sample (asymptotic) theorems, point estimates, and uncertainty

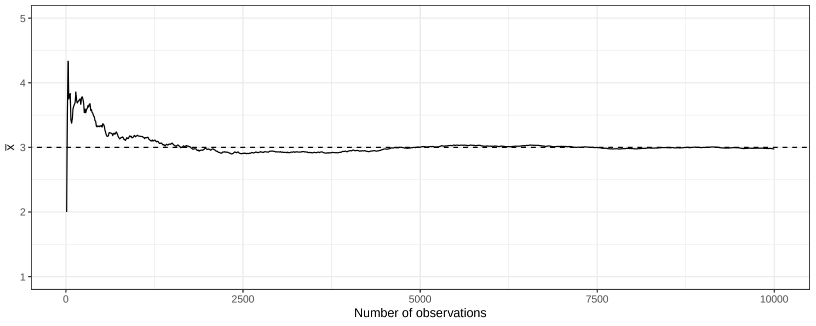
The law of large numbers

As a sample of draws from a random variable increases, the sample mean converges to the population mean $E(X)$

$$\bar{x}_n \to E(X)$$

The law of large numbers: point estimates converge to population parameters as n increases

A Monte Carlo simulation where we draw from Binomial(10, 0.3) 1 time up to 1000 times, then compute $\overline{x}$



The Central Limit Theorem

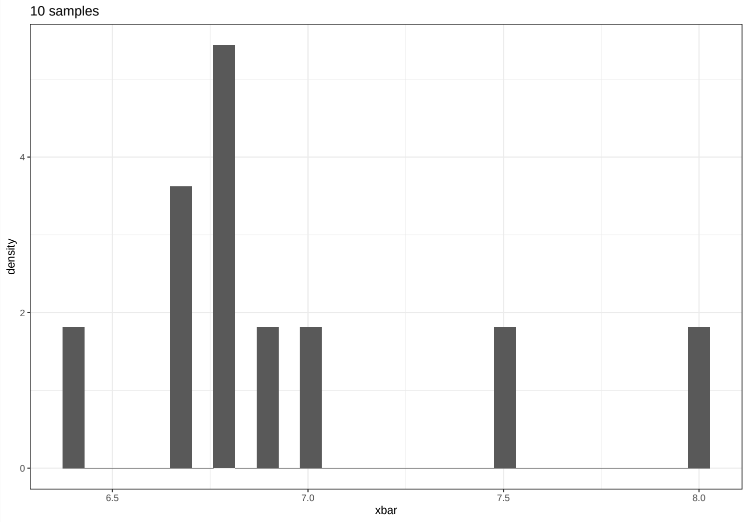
- If we draw independent random samples, as sample size n increases, the distribution of the sample mean $\bar{x}$ approaches a Normal distribution.

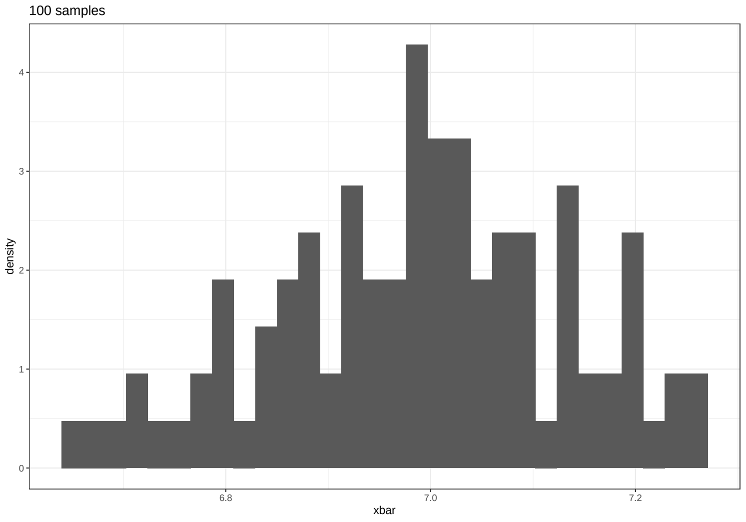
How many students eat pizza in a week?

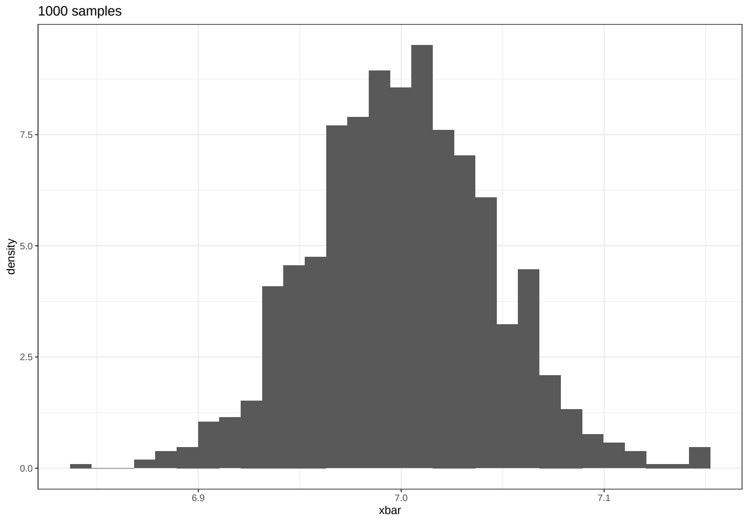
We want to estimate $\overline{pizza}$, the proportion of students who eat pizza per week on campus. Our approach: randomly select 10 classrooms, then randomly select 10 students from each class. Count the number who ate pizza within the prior week (0 = No pizza, 1 = pizza)

Monte Carlo simulations of a binomial variable p=0.7, n=10

1. Take 10 draws from $pizza \sim Binomial(10, 0.7)$
2. Compute $\overline{pizza}$
3. Repeat many times!


10 samples
density
xbar
0
2
4
6.5
7.0
7.5
8.0


100 samples
density
xbar
0
1
2
3
4
6.8
7.0
7.2


1000 samples
density
7.5
5.0
2.5
0.0
6.9
7.0
7.1
xbar

Of course, we generally don't replicate 1000 times.

Let's take 1 draw from the pizza study

```
n<-10 # 10 students in a class
p<-0.7 # 70% chance of a 1
pizza<-rbinom(n = n, p = p, size = 10)
```

What can we say about the proportion of students who eat pizza at Rutgers?

Our point estimate for the proportion is

```
mean(pizza)
```

```
## [1] 7.1
```

Our standard deviation for the study is

```
sd(pizza)
```

```
## [1] 1.449138
```

We can describe our estimate for the *sampling distribution* of $\overline{pizza}$ as a Normal distribution centered at 7.1 with a standard error of 0.46.

We can construct a 95% confidence interval to describe how uncertain we are about the location of $\overline{pizza}$. Here, that interval is:

```
## bounds = +/- 1.96 (Normal PDF for 95% mass)
se<-sd(pizza)/sqrt(10)
mean(pizza) + 1.96 * se
```

```
## [1] 7.998185
```

```
mean(pizza) - 1.96 * se
```

```
## [1] 6.201815
```

How should we interpret this interval?

Hypothesis testing

The problem

- I think that the true proportion of students who eat pizza is 0.5
- How can I use my data to evaluate this claim?

The basic approach

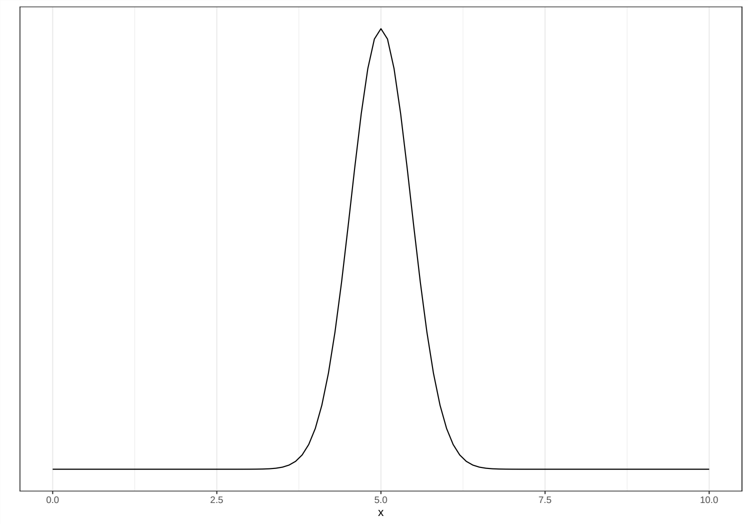
- Establish a hypothesis

1. $H_1 : E[pizza] = 5$

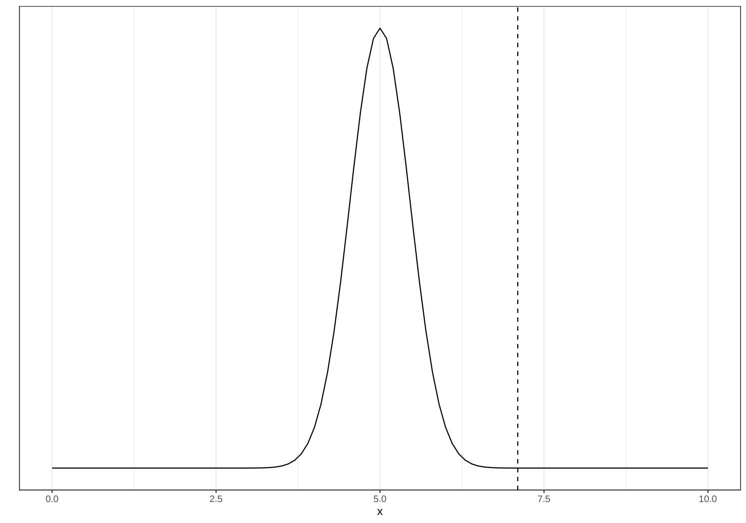
- Evaluate how likely our observations are under the hypothesis

1. We know via CLT that $\overline{pizza} \sim N(\mu, \sigma^2)$
2. How likely is 7.1 under a distribution with mean 5 and SE 0.46?

Our hypothesis for the sampling distribution of pizza habits



What we observed



How likely was our observation if H1 were true?

Use the Normal PDF to estimate

```
pnorm(q = mean(pizza), mean = 5, sd = sd(pizza)/sqrt(10))
```

```
## [1] 0.9999977
```

That's the proportion of observations that fall below our observation *if* H_1 is true. To convert this to how likely we are to observe our data *if* H_1 is true, subtract from 1

```
1 - pnorm(mean(pizza), 5, sd(pizza)/sqrt(10))
```

```
## [1] 2.296417e-06
```

What can we conclude?

Back to regression

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

ε: The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y.

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

ε: The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y.

The line $E(y_i) = \beta_0 + \beta_1 x_i$ provides an expected value for y_i based on the values of x_i.

The linear regression model and prediction

Remember, that we put a $h\hat{a}t$ on variables to indicate that they are estimated from the data, or predicted.

The linear regression model and prediction

Remember, that we put a *h$\hat{a}$t* on variables to indicate that they are estimated from the data, or predicted.

In other words, we try to learn about the 'true' *regression coefficients* β_1 and β_0 by estimating $\hat{\beta}_1$ and $\hat{\beta}_0$.

The linear regression model and prediction

Remember, that we put a *$\hat{\text{hat}}$* on variables to indicate that they are estimated from the data, or predicted.

In other words, we try to learn about the 'true' *regression coefficients* β_1 and β_0 by estimating $\hat{\beta}_1$ and $\hat{\beta}_0$.

Standard errors of β

The standard error of β is the standard deviation of its sampling distribution.

Standard errors of β

The standard error of β is the standard deviation of its sampling distribution.

In other words: $\hat{\beta} \sim N(\beta, SE^2_\beta)$

Standard errors of β

The standard error of β is the standard deviation of its sampling distribution.

In other words: $\hat{\beta} \sim N(\beta, SE_{\beta}^2)$

The standard error of β is calculated as:

$$SE_{\beta} = \sqrt{\frac{\sum \varepsilon_i^2}{(n-2)\sum(x_i - \bar{x})^2}}$$

Note that the numerator captures variance in y and the denominator captures variance in x

Uncertainty and OLS

The Mark of a Criminal Record

```r
### read and format Pager data
cr<-read_csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/data/criminalrecord.csv")
cr<-cr %>%
  select(callback, crimrec)

head(cr)
```

```
## # A tibble: 6 x 2
##   callback crimrec
##      <dbl>   <dbl>
## 1        1       1
## 2        0       0
## 3        1       0
## 4        1       0
## 5        0       1
## 6        0       1
```

The Research question and the null hypothesis

- Does a criminal record make a callback less likely?

The Research question and the null hypothesis

- Does a criminal record make a callback less likely?
- Implied null hypothesis: No difference in callback rates

The Research question and the null hypothesis

- Does a criminal record make a callback less likely?
- Implied null hypothesis: No difference in callback rates

$$H_0 : E[Callback|Crimrec = T] - E[Callback|Crimrec = F] = 0$$

The Research question and the null hypothesis

- Does a criminal record make a callback less likely?
- Implied null hypothesis: No difference in callback rates

$$H_0 : E[Callback|Crimrec = T] - E[Callback|Crimrec = F] = 0$$

Written differently:

$$H_0 : E[Callback|Crimrec = T] = E[Callback|Crimrec = F]$$

Let’s estimate the model for the effect of crimrec on callback

```r
library(broom)
m0<-lm(callback ~ crimrec, data = cr)
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.226    0.0196     11.6  1.82e-28
## 2 crimrec       -0.125    0.0277     -4.53 6.87e- 6
```

Write this out as a regression equation.

- What does β_0 mean?
- What does β_1 mean?

Setting up our hypothesis test

H_0 : No effect of crimrec on callback.

What does this imply in terms of β?

Recall that our model says $E[callback] = \beta_0 + \beta_1 Crimrec$

Our null hypothesis for the central research question

$$H_0 : \beta_1 \sim N(0, SE^2_{\beta_1})$$

What do we observe?

```r
tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.226    0.0196     11.6  1.82e-28
## 2 crimrec       -0.125    0.0277     -4.53 6.87e- 6
```

Computing the null hypothesis test manually

How likely is -0.125 if H_0 is true?

Let's check our data against the Normal PDF for H_0

```
pnorm(-0.125, 0, 0.0277)
```

```
## [1] 3.201352e-06
```

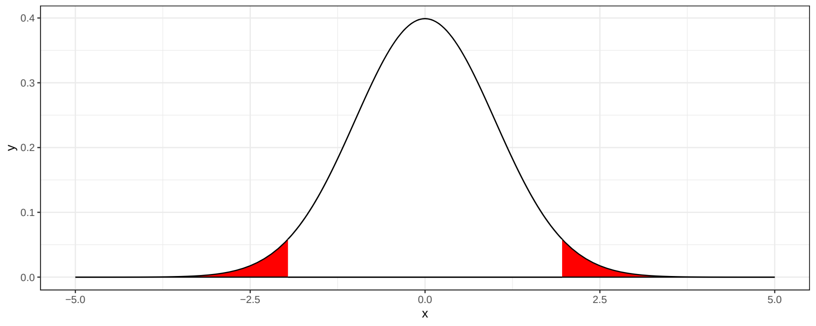
What do we think?

The Normal PDF and hypothesis testing

The logic of a hypothesis test

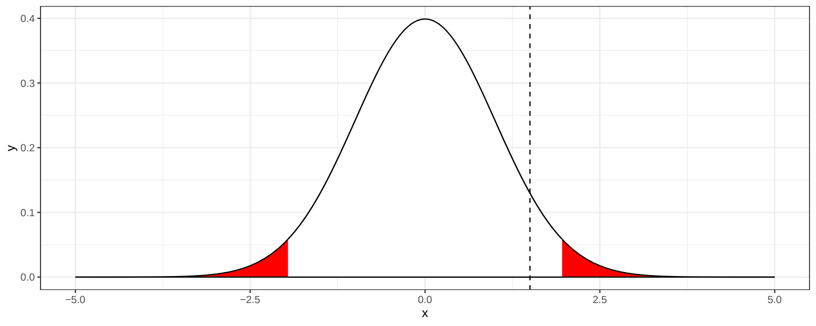
Assume $H_0 : Z(\beta) \sim N(0, 1)$ (standardized Beta follows a Z distribution)

We decide *a priori* that anything outside of the central 95% of the Normal PDF is inconsistent with H_0



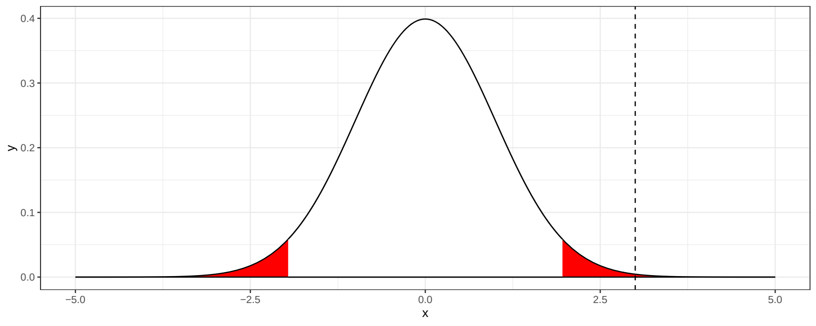
The logic of a hypothesis test

Now we observe our data and estimate our model. We find $\beta = 1.2$ and $SE = 0.8$. We convert that into a z-score $z = 1.2/0.8 = 1.5$ and check where it falls



The logic of a hypothesis test

Now we observe our data and estimate our model. We find $\beta = 1.2$ and $SE = 0.4$. We convert that into a z-score $z = 1.2/0.4 = 3$ and check where it falls



Null hypothesis testing: a recipe

1. Specify your null hypothesis (typically $Z(\beta) \sim N(0, 1)$)
2. Specify your critical value (sometimes called α), the threshold for finding 'statistical significance'
3. Estimate your model, compute a z statistic for β
4. Compute whether your $Z(\beta)$ falls in the critical region of the z distribution

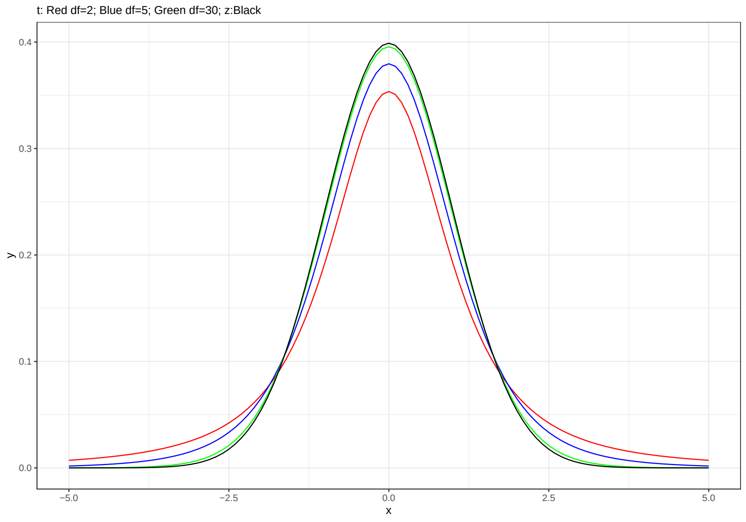
A note on t, z, and sample size

Technically, R will perform t tests, not z tests on our regression models.

When our *degrees of freedom* are large, the t distribution converges to the z distribution (Normal(0,1)).

Degrees of freedom for regression are defined as $n - k$, where n is sample size, and k is the number of parameters we are estimating in our model.

The t and the z: convergence for large DF



Using the central limit theorem to calculate confidence intervals, compute p-values

If the sampling distribution for β is defined as:

$\hat{\beta} \sim N(\beta, SE_{\beta}^2)$

Using the central limit theorem to calculate confidence intervals, compute p-values

If the sampling distribution for β is defined as:

$\hat{\beta} \sim N(\beta, SE_{\beta}^{2})$

Then we can construct a 95 percent CI for β

$$\hat{\beta} \pm 1.96 \times SE_{\beta}$$

Using the central limit theorem to calculate confidence intervals, compute p-values

If the sampling distribution for β is defined as:

$\hat{\beta} \sim N(\beta, SE_{\beta}^2)$

Then we can construct a 95 percent CI for β

$$\hat{\beta} \pm 1.96 \times SE_{\beta}$$

And conduct a z test for $\hat{\beta}$ by evaluating how likely our estimated $\hat{\beta}$ is under the null hypothesis

$H_0 : \beta \sim N(0, SE_{\beta}^2)$

Using OLS to estimate the SATE

```r
cr_ols<-lm(callback ~
             crimrec,
           data = cr)

tidy(cr_ols)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.226    0.0196     11.6  1.82e-28
## 2 crimrec       -0.125    0.0277     -4.53 6.87e- 6
```

- What is the implied null hypothesis here?
- How do we compute ‘statistic’ (z-statistic)?
- How do we compute ‘p.value’?

Interpretation

```r
tidy(cr_ols)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.226    0.0196     11.6  1.82e-28
## 2 crimrec       -0.125    0.0277     -4.53 6.87e- 6
```

- What can we conclude about β_1?
- What does this tell us about our focal research question?

Statistical significance give us information about whether differences we observe for some outcome across levels of a predictor in our data are likely to occur if there were no underlying differences in the data-generating process.

Statistical significance: an interpretation guide

Statistical significance give us information about whether differences we observe for some outcome across levels of a predictor in our data are likely to occur if there were no underlying differences in the data-generating process.

Put simply: if there were no difference across levels, how likely would I be to observe what I did observe?

Statistical significance: an interpretation guide

Statistical significance give us information about whether differences we observe for some outcome across levels of a predictor in our data are likely to occur if there were no underlying differences in the data-generating process.

Put simply: if there were no difference across levels, how likely would I be to observe what I did observe?

What *t* tests and *z* tests do is provide a quick signal vs noise check for our data.

Statistical significance: CAUTIONS

Statistical significance testing is arbitrary. The null hypothesis is arbitrary, and 0.95 is arbitrary.

Statistical significance: CAUTIONS

Statistical significance testing is arbitrary. The null hypothesis is arbitrary, and 0.95 is arbitrary.

DO NOT USE SIGNIFICANCE TESTING ALONE TO INFORM YOUR SCIENTIFIC JUDGMENT

Statistical significance: CAUTIONS

Statistical significance testing is arbitrary. The null hypothesis is arbitrary, and 0.95 is arbitrary.

DO NOT USE SIGNIFICANCE TESTING ALONE TO INFORM YOUR SCIENTIFIC JUDGMENT

DO NOT USE SIGNIFICANCE TESTING ALONE TO DECIDE WHAT YOUR MODEL SHOULD BE