Statistical Significance

Understanding Political Numbers

March 6, 2019

Review

Regression Review

1. **Assume: $\mathrm{E}[Y \mid X]$ is a line**

   Expected average y, conditional on its x value

2. **$\hat{y}_i$ is predicted y for observation i.**

   $\hat{y}_i = a + bx_i$

3. **y_i is the observed y (prediction + residual error)**

   $y_i = a + bx_i + e_i$

4. **Residual error: actual minus predicted**

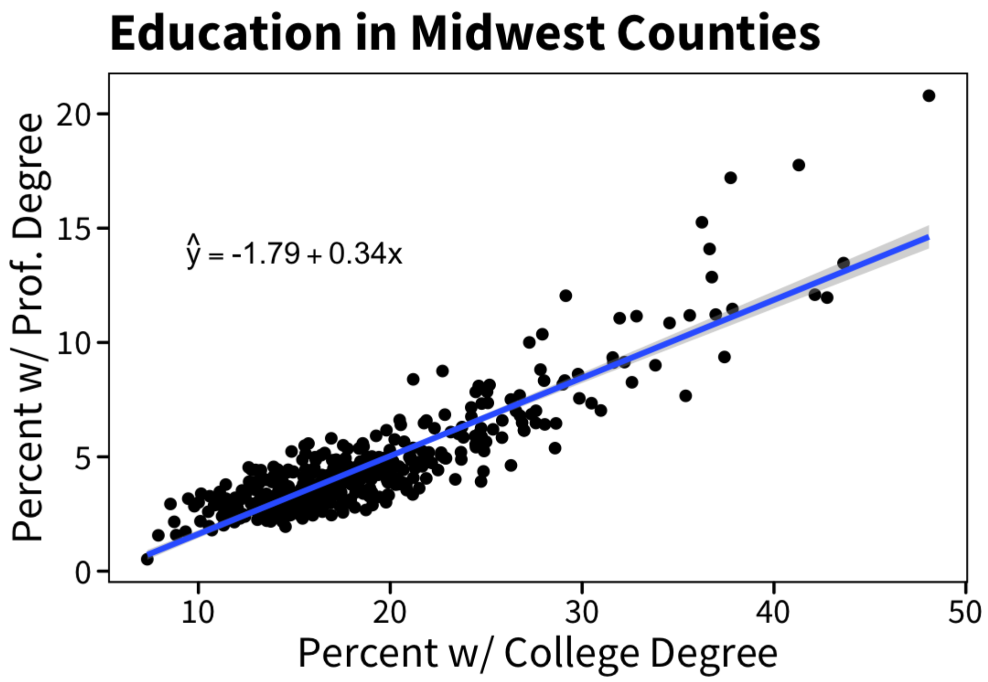
   $e_i = y_i - \hat{y}_i$

5. **"Ordinary least Squares" (OLS) estimation: pick a and b that minimize error**

   Technically, minimizing the "sum of squared error"

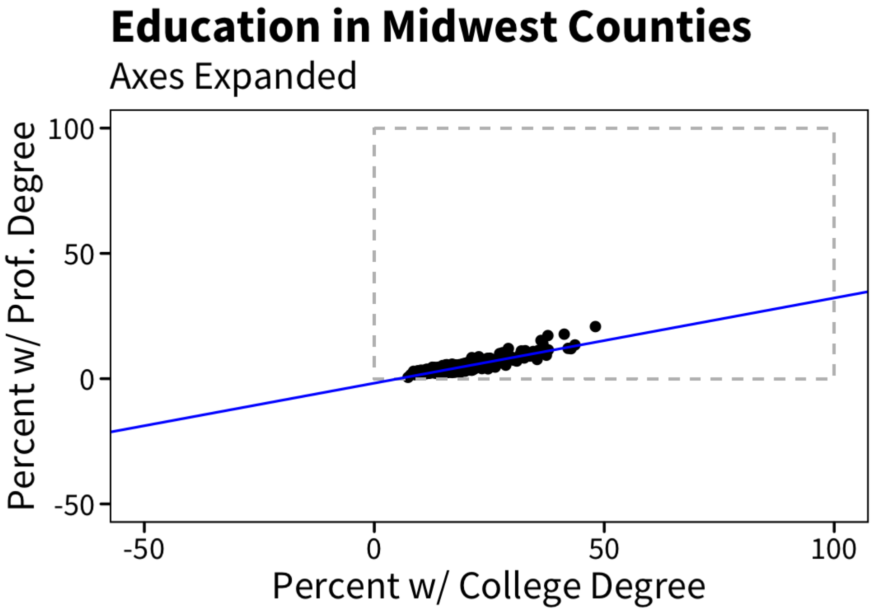
```r
library("tidyverse") # contains 'midwest' data

lm(percprof ~ percollege, data = midwest)
```

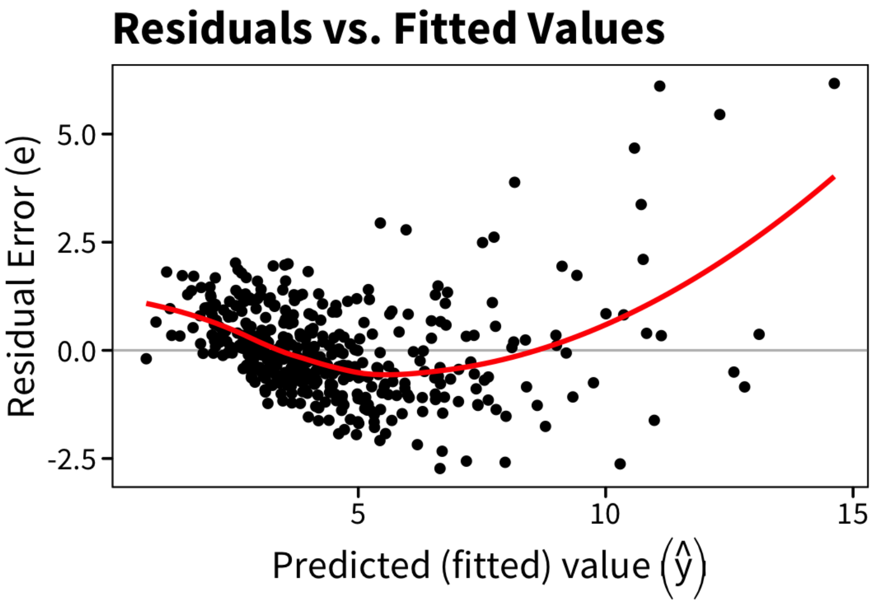
```
##
## Call:
## lm(formula = percprof ~ percollege, data = midwest)
##
## Coefficients:
## (Intercept)   percollege
##     -1.7899       0.3413
```



Warnings

1. Beware: Does a linear relationship make sense
2. Beware: extrapolation beyond data (top figure)
3. Beware: patterns in residuals (bottom figure)
4. Beware: influential outliers

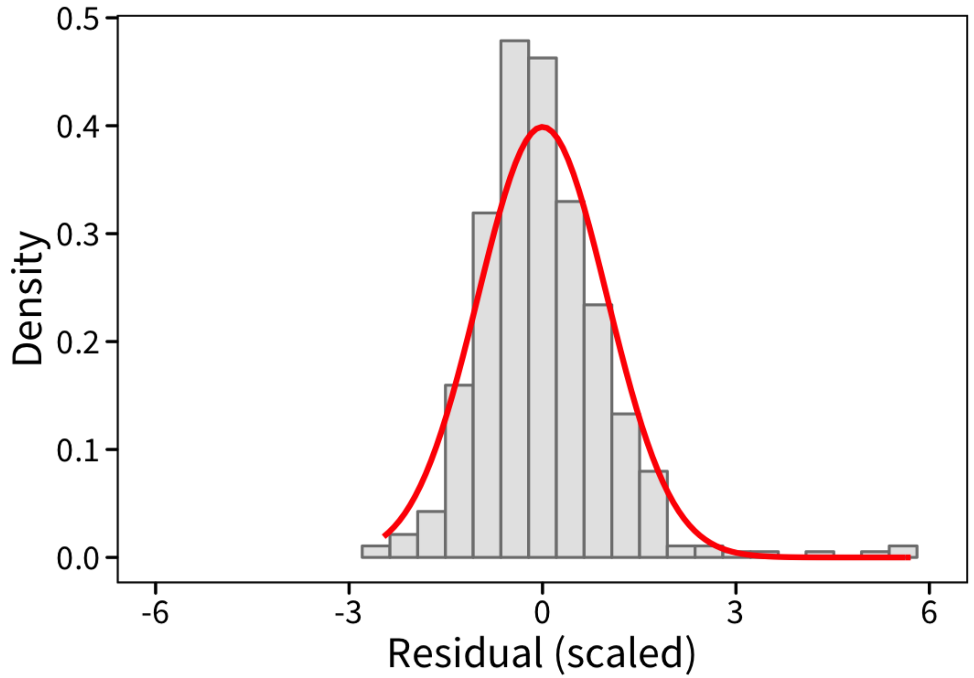




Assumptions about leftover error

We assume that error e_i is random noise

- ***After* accounting for x
- Only x affects y? No.
- e_i is the sum of "everything else"
- Accumulation of random noise → normal distribution
- Expected value of error is 0



Statistical Significance

A result is *statistically significant* if is was unlikely to have occurred by chance

The "True" Model

We estimate a and b, but estimates are noisy. What can we learn about the *true* equation?

The "True" Model

We estimate a and b, but estimates are noisy. What can we learn about the *true* equation?

The true equation

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Problem: We don't know α and β and never will

The "True" Model

We estimate a and b, but estimates are noisy. What can we learn about the *true* equation?

The true equation

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Problem: We don't know α and β and never will

The estimated equation

$$y_i = a + b x_i + e_i$$

a and b are imperfect estimates of α and β

The "True" Model

We estimate a and b, but estimates are noisy. What can we learn about the *true* equation?

The true equation

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Problem: We don't know α and β and never will

The estimated equation

$$y_i = a + b x_i + e_i$$

a and b are imperfect estimates of α and β

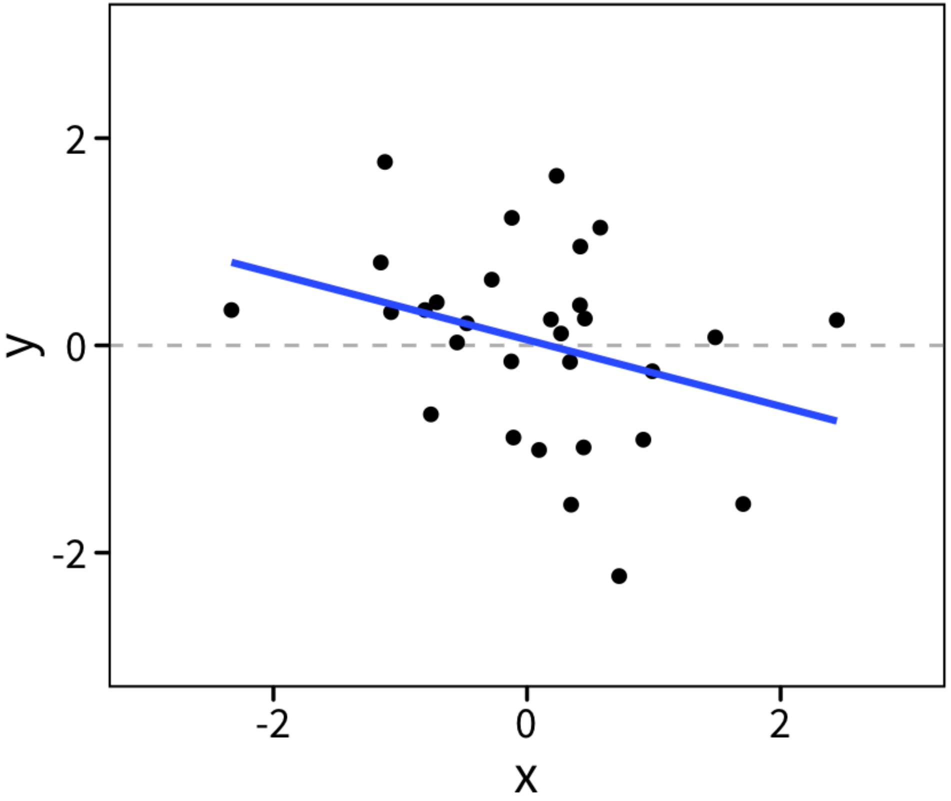
Statistical inference is "what conclusions can I draw about β even though I can't see it?"

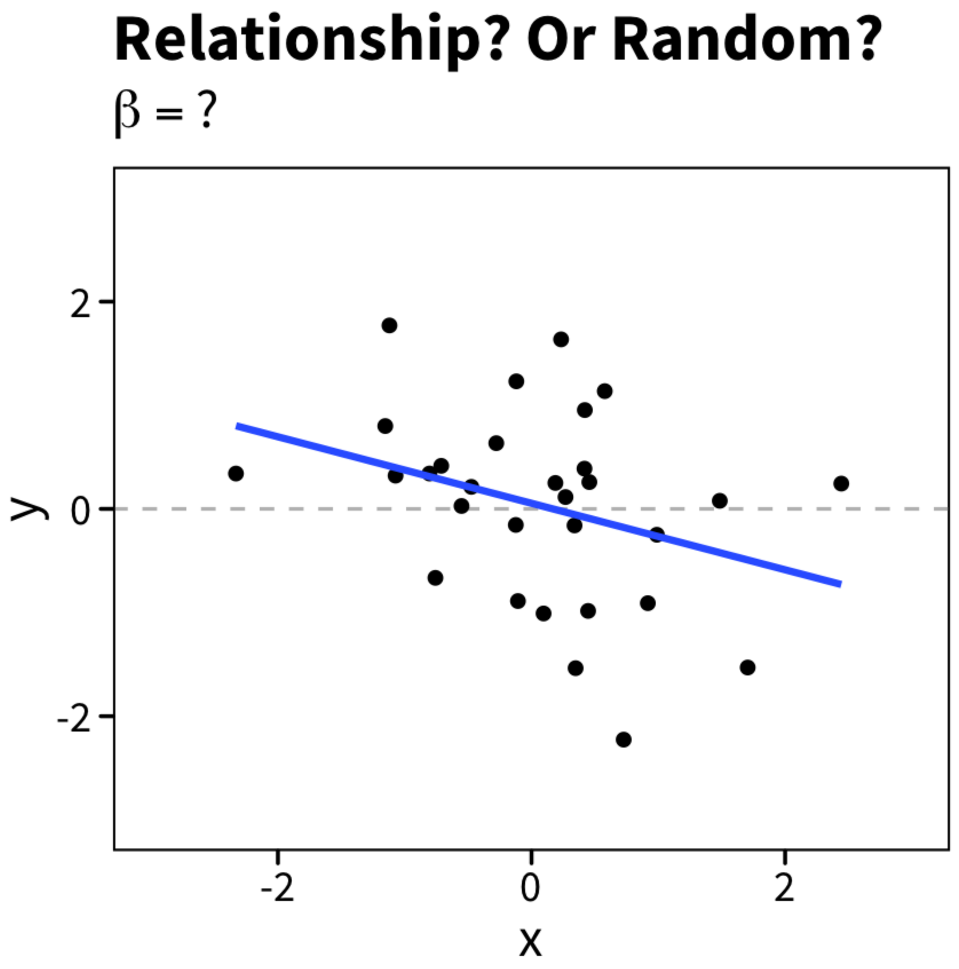
A result is *statistically significant* if is was unlikely to have occurred by chance

We want to make inferences about the "true" parameters, but we only observe a sample of data.

Relationship? Or Random?

$\beta = ?$





The "null hypothesis"

Assume that $\beta = 0$

Estimate the model on data

```
ex_reg <- lm(y ~ x, data = test_data) %>%
  print()
```

```
##
## Call:
## lm(formula = y ~ x, data = test_data)
##
## Coefficients:
## (Intercept)            x
##     0.05423     -0.32082
```

Assuming that $\beta = 0$, what's the probability (p) of observing a b this big *by random chance*?

A result is *statistically significant* if is was unlikely to have occurred by chance

We want to make inferences about the "true" parameters, but we only observe a sample of data.

What's the *probability* of observing our slope, *if the null were true* (p value)

Find the p-value

p value: The probability of observing a slope *at least this big* if the null hypothesis is true

Find the p-value

p **value:** The probability of observing a slope *at least this big* if the null hypothesis is true

Output from `tidy()` (from the `broom` package)

- a data frame!
- `estimate` : coefficients (a and b values)
- `std.error` : uncertainty of estimates
- `statistic` : standardized slope (estimate / std.err)
- `p-value` : self-explanatory

```r
# 'broom' pkg for model output
# install.packages("broom")

# load it
library("broom")

# info about model estimates
tidy(ex_reg)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   0.0542     0.165     0.330  0.744
## 2 x            -0.321      0.175    -1.83   0.0777
```

"Rejecting the null hypothesis"

Null hypothesis significance testing:

- "Assuming the null hypothesis is true, the probability of observing a slope at least this *extreme* is (p)"
- If p is really low, then it's unlikely that the data come from the null hypothesis
- "Statistical significance" means p is lower than some threshold
- Reject the null hypothesis at $(1 - p)$% confidence

"Rejecting the null hypothesis"

Null hypothesis significance testing:

- "Assuming the null hypothesis is true, the probability of observing a slope at least this *extreme* is (p)"
- If p is really low, then it's unlikely that the data come from the null hypothesis
- "Statistical significance" means p is lower than some threshold
- Reject the null hypothesis at $(1 - p)$% confidence

$p < 0.1$: significant at the 10% level (reject the null with 90% confidence)

$p < 0.05$: significant at the 5% level (reject the null with 95% confidence)

$p < 0.01$: significant at the 1% level (reject the null with 99% confidence)

Lower p values, stronger signal, more confident that $\beta \neq 0$

A result is *statistically significant* if is was unlikely to have occurred by chance

We want to make inferences about the "true" parameters, but we only observe a sample of data.

What's the *probability* of observing our slope, *if the null were true*

An estimate is *significant* if the probability of getting it, under the null, is "sufficiently low"

Where do p-values come from?

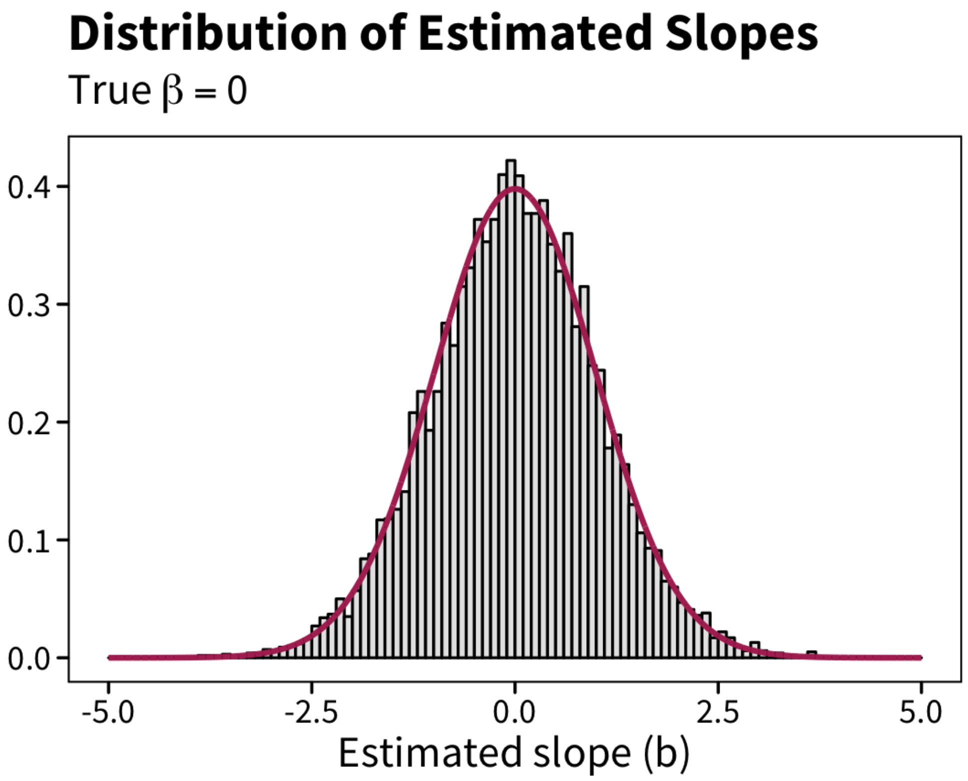
Let's do a `S I M U L A T I O N`

- Generate 10k datasets containing x and y
- In every dataset, the **true slope** is zero
- In every dataset, our **estimated slope** is not zero (thanks to random error e_i)

Where do p-values come from?

Let's do a `S I M U L A T I O N`

- Generate 10k datasets containing x and y
- In every dataset, the **true slope** is zero
- In every dataset, our **estimated slope** is not zero (thanks to random error e_i)



We know the theoretical distribution of "by-chance" slopes

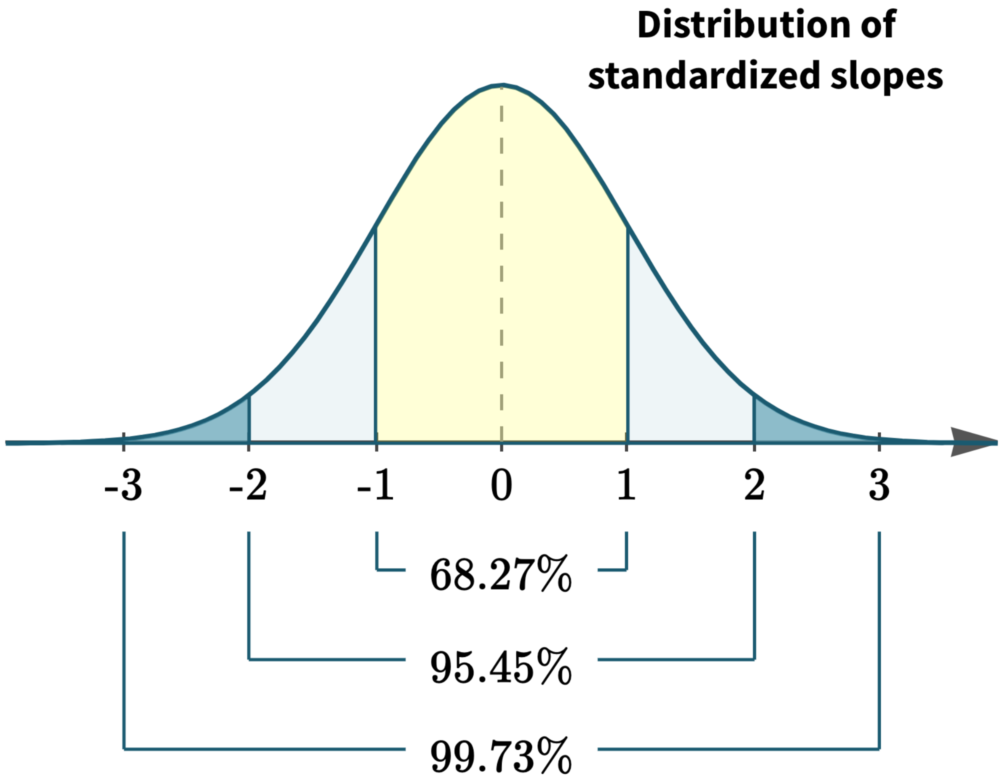
We know the distribution of "by-chance" slopes

Compare slopes by *standardizing* them:

$$t = \frac{b}{std.err(b)}.$$

"Big" t values are unlikely

p value is the probability of getting an even "bigger" t value



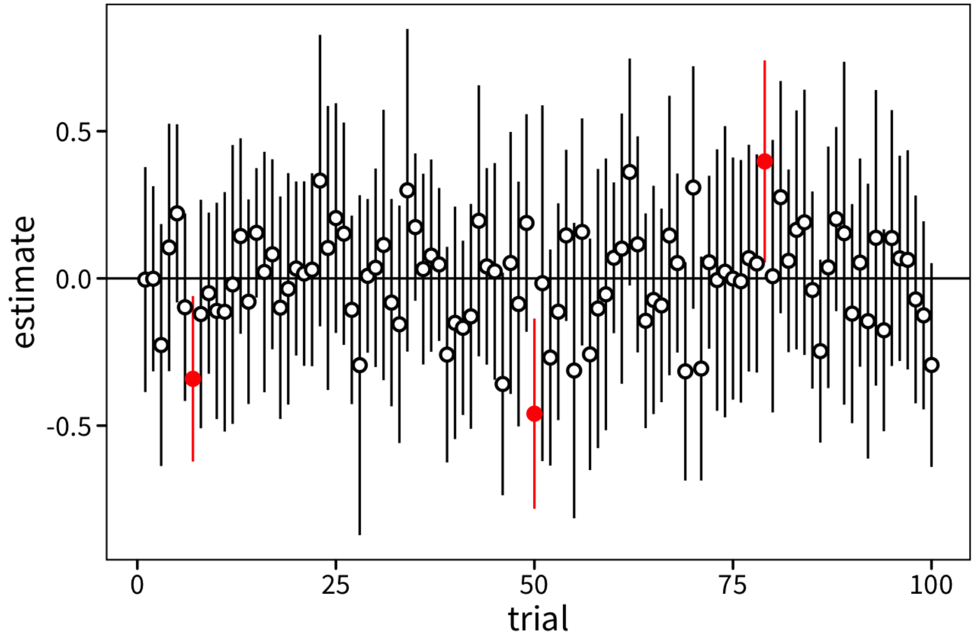
Confidence levels and p-values

95% Interval = $b \pm 1.96(se(b))$

Naive interpretation: 95% chance that the true value is within the interval

Better interpretation: The parameter is in the interval or it's not. The interval contains the true value in 95% of samples (if you could take an infinite number of samples, which, you can't)

Practical interpretation: Interval contains all the values I can't reject. if it doesn't contain zero, you can reject zero



Inference issues with *p* values

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

It takes *lots* of data to estimate small effects w/ statistical significance

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

It takes *lots* of data to estimate small effects w/ statistical significance

Insignificance does *not* mean "no relationship," only that there wasn't enough data to reject the null hypothesis

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

It takes *lots* of data to estimate small effects w/ statistical significance

Insignificance does *not* mean "no relationship," only that there wasn't enough data to reject the null hypothesis

Relationships are everywhere, we just need enough data to make confident inferences about what they are

A result is *statistically significant* if is was unlikely to have occurred by chance

We want to make inferences about the "true" parameters, but we only observe a sample of data.

What's the *probability* of observing our slope, *if the null were true*

An estimate is *significant* if the probability of getting it, under the null, is "sufficiently low"

Null relationships can still "pop" as significant, and "non-null" relationships may fail to show insignificance