Inference

Week 7

29 February, 2016

Prof. Andrew Eggers

What we’re trying to understand today

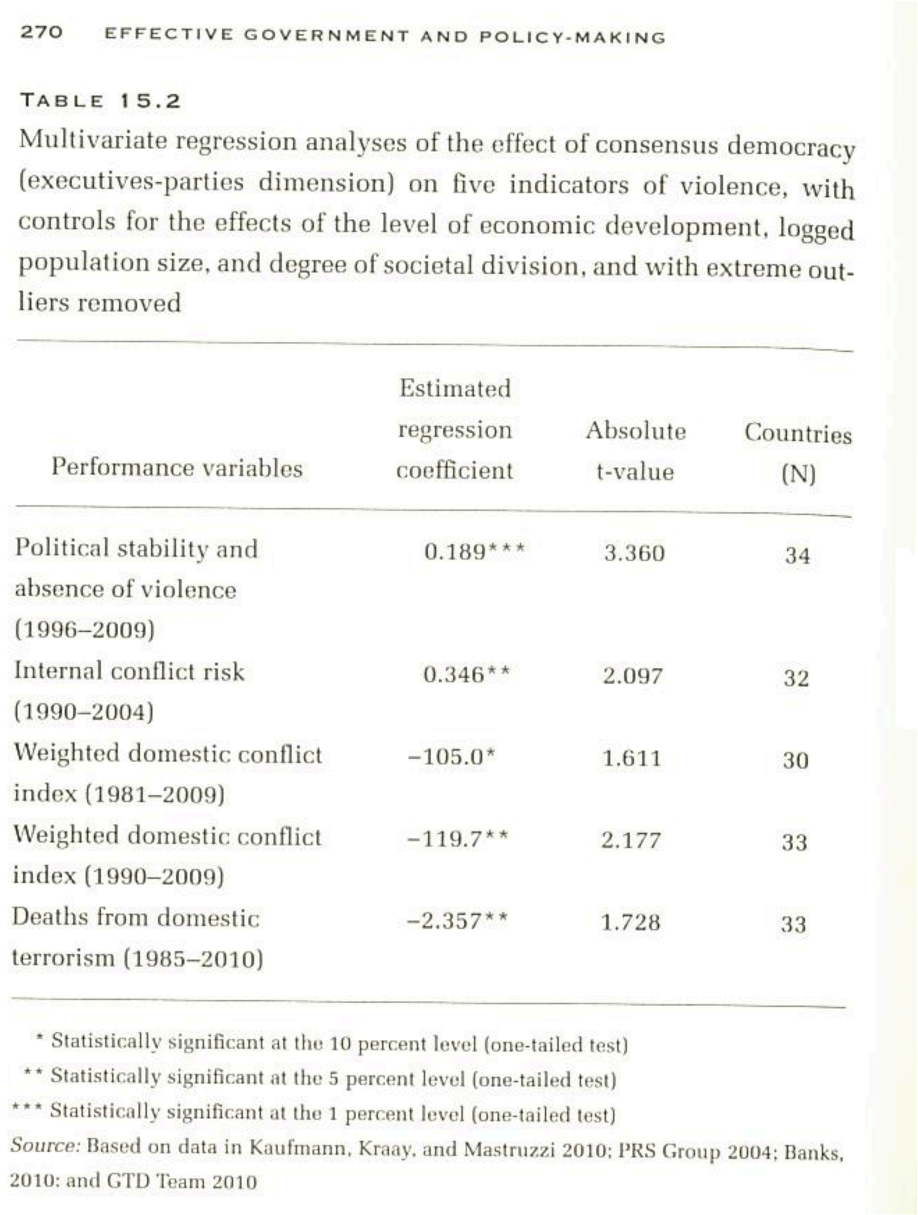
Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(1)	(2)	(3)
Intercept	-1.629* (0.509)	-3.166* (0.511)	-2.982* (0.527)
Chocolate consumption per capita (log scale)	2.092* (0.298)	1.026* (0.326)	0.709 (0.415)
GDP/capita (thousands of USD)		0.105* (0.024)	0.106* (0.024)
NW Europe			0.549 (0.452)
R^2	0.70	0.85	0.86
N	34	34	34

Standard errors in parentheses. * Indicates $p<0.05$

- What do the stars mean on regression tables?
- What is the “margin of error” of a poll?
- What statistical findings are reliable? Which might be just a fluke?

What we're trying to understand today



270 EFFECTIVE GOVERNMENT AND POLICY-MAKING

TABLE 15.2

Multivariate regression analyses of the effect of consensus democracy (executives-parties dimension) on five indicators of violence, with controls for the effects of the level of economic development, logged population size, and degree of societal division, and with extreme outliers removed

Performance variables	Estimated regression coefficient	Absolute t-value	Countries (N)
Political stability and absence of violence (1996–2009)	0.189***	3.360	34
Internal conflict risk (1990–2004)	0.346**	2.097	32
Weighted domestic conflict index (1981–2009)	−105.0*	1.611	30
Weighted domestic conflict index (1990–2009)	−119.7**	2.177	33
Deaths from domestic terrorism (1985–2010)	−2.357**	1.728	33

* Statistically significant at the 10 percent level (one-tailed test)

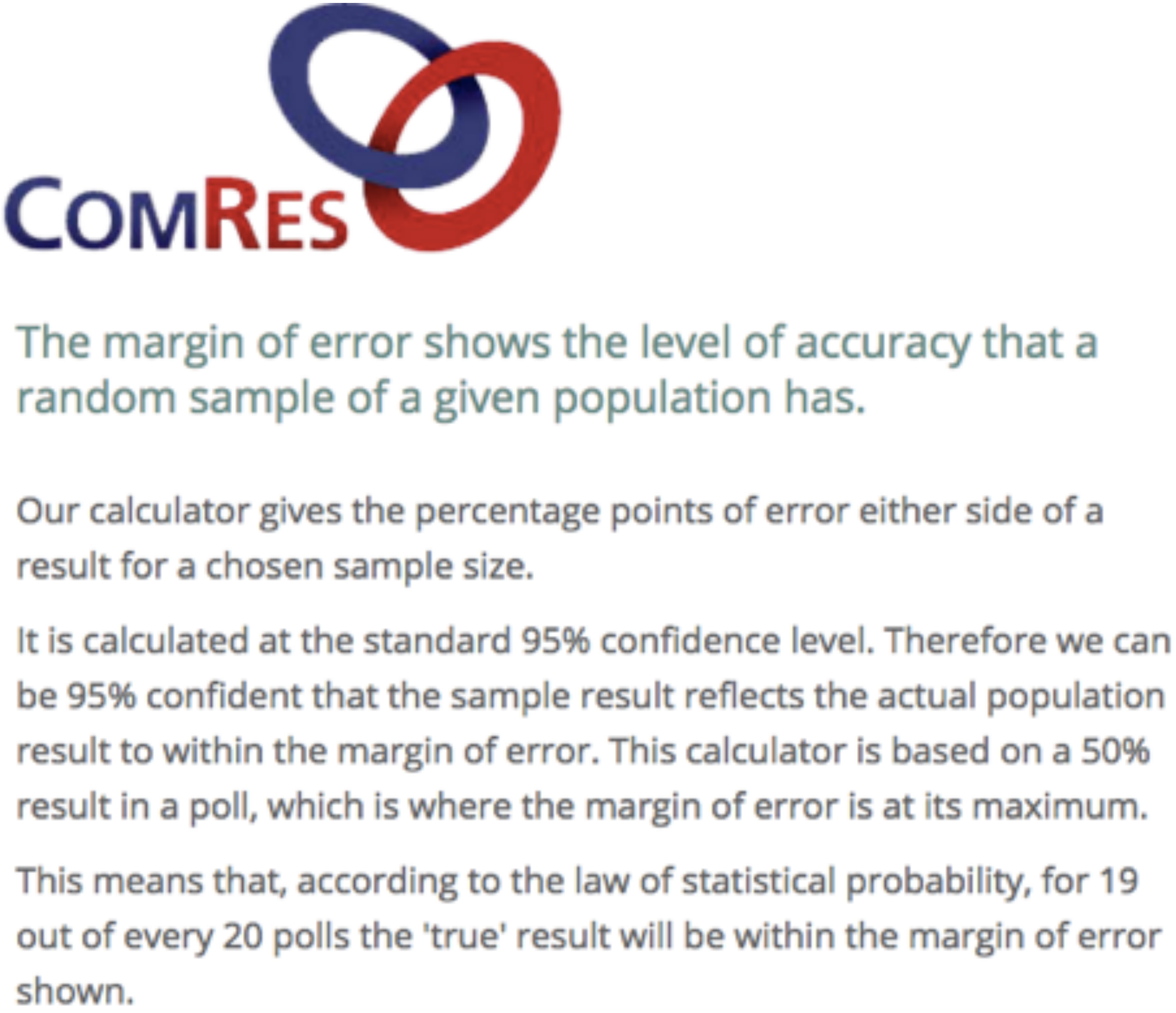
** Statistically significant at the 5 percent level (one-tailed test)

*** Statistically significant at the 1 percent level (one-tailed test)

Source: Based on data in Kaufmann, Kraay, and Mastruzzi 2010; PRS Group 2004; Banks, 2010; and GTD Team 2010

- What do the stars mean on regression tables?
- What is the "margin of error" of a poll?
- What statistical findings are reliable? Which might be just a fluke?

First task: understanding margin of error



COMRES

The margin of error shows the level of accuracy that a random sample of a given population has.

Our calculator gives the percentage points of error either side of a result for a chosen sample size.

It is calculated at the standard 95% confidence level. Therefore we can be 95% confident that the sample result reflects the actual population result to within the margin of error. This calculator is based on a 50% result in a poll, which is where the margin of error is at its maximum.

This means that, according to the law of statistical probability, for 19 out of every 20 polls the 'true' result will be within the margin of error shown.

http://www.comres.co.uk/our-work/margin-of-error-calculator/

Understanding margin of error

Recall from the measurement lecture:

measured value = true value + bias + random error

To get rid of bias:

- in measuring concepts (week 2), we sought **valid** measures
- in selecting cases (week 4), we used **random sampling** or other approaches in which "criteria determining selection are not correlated with the outcome of interest"

"Margin of error" tries to summarize the **magnitude of random error due to sampling**.

Thought experiment

measured value = true value + bias + random error

Imagine you took a random sample of GB adults and asked whether they supported remaining in the EU.

Is the average support in your sample close to the true average support?

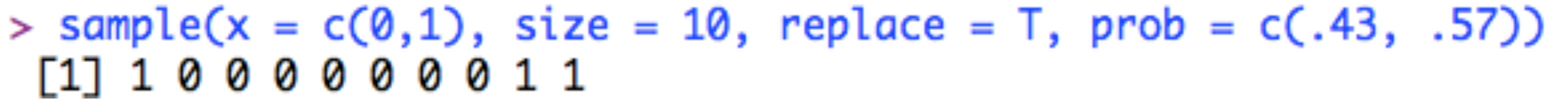
What would the magnitude of the **random error** depend on?

- size of sample (1,006 GB adults vs. 10,000,000)
- true level of support (what if 100% supported remaining in EU?)

Simulating the thought experiment in R

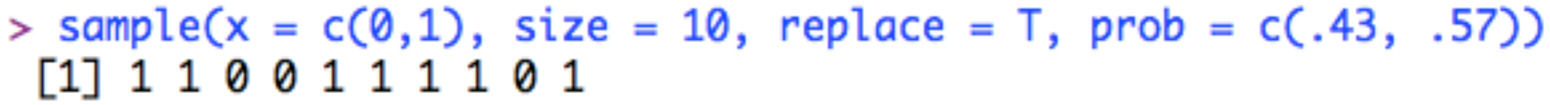
We **assume we know** that 57% of people support **remain**, and that we can randomly pick a sample of people to “survey”.

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.57:



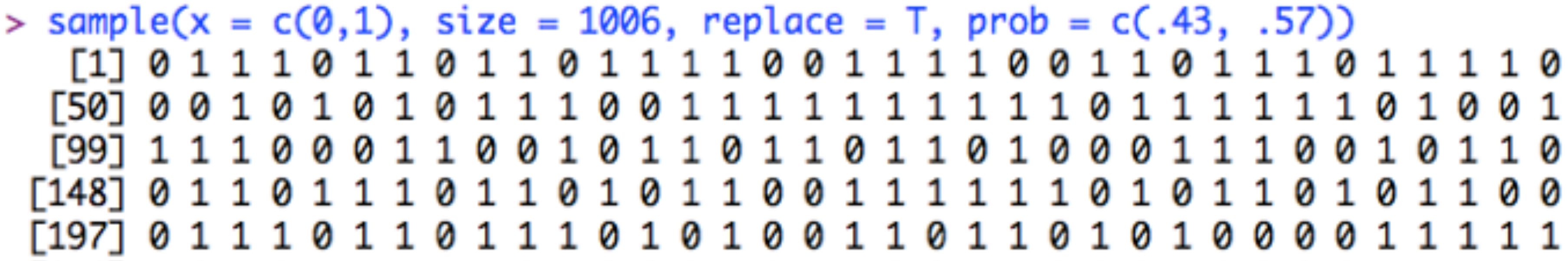
```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.43, .57))
 [1] 1 0 0 0 0 0 0 0 1 1
```

I can do it again:



```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.43, .57))
 [1] 1 1 0 0 1 1 1 1 0 1
```

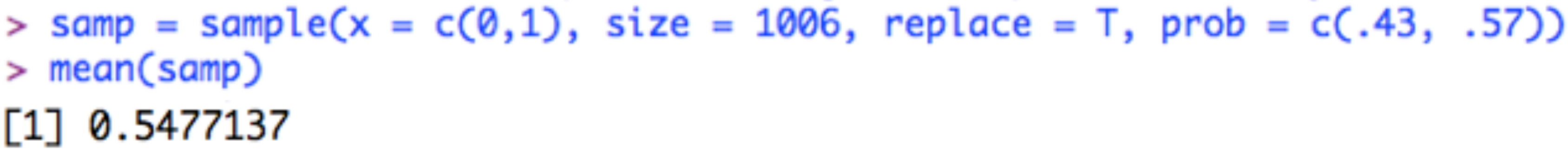
I can increase the number of “respondents” to 1,006:



```
> sample(x = c(0,1), size = 1006, replace = T, prob = c(.43, .57))
   [1] 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 0 1 1 1 0 1 1 1 1 0
  [50] 0 0 1 0 1 0 1 0 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0 0 1
  [99] 1 1 1 0 0 0 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 1 0 0 0 1 1 1 0 0 1 0 1 1 0
 [148] 0 1 1 0 1 1 1 0 1 1 0 1 0 1 1 0 0 1 1 1 1 1 1 0 1 0 1 1 0 1 0 1 1 0 0
 [197] 0 1 1 1 0 1 1 0 1 1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 0 0 0 0 1 1 1 1 1
```

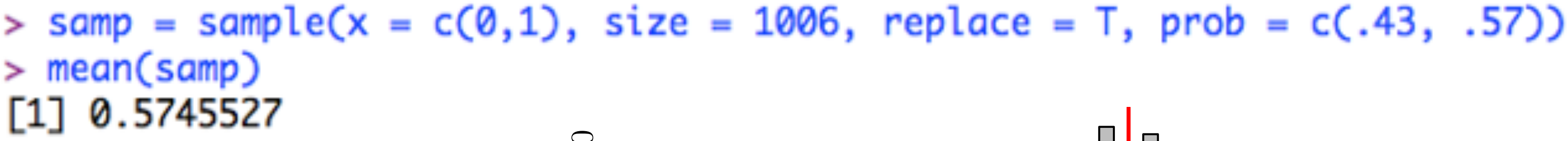
Simulating the thought experiment (2)

I can store the sample and take the mean:



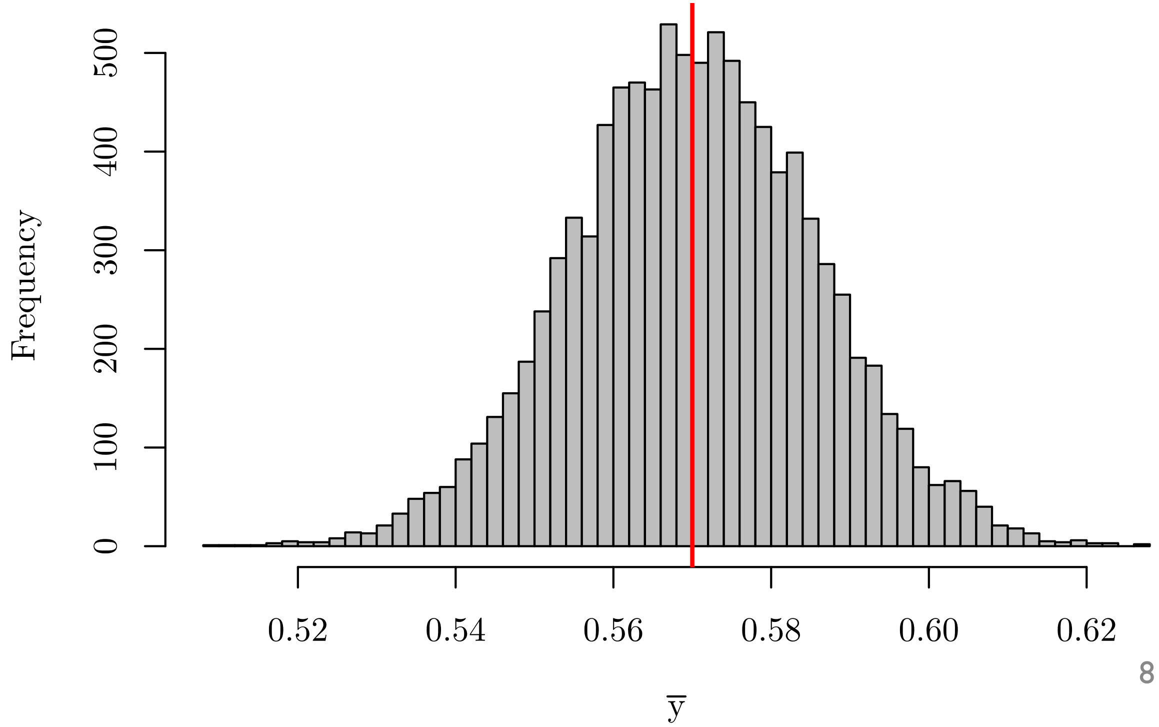
```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.43, .57))
> mean(samp)
[1] 0.5477137
```

I can do it again:



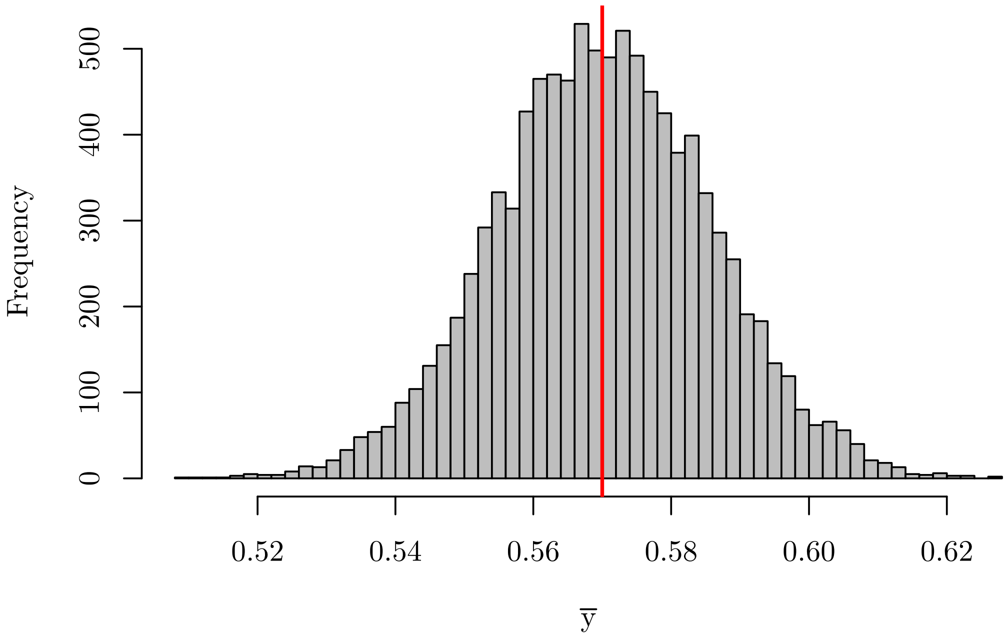
```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.43, .57))
> mean(samp)
[1] 0.5745527
```

I can do it 10,000 times and look at the histogram of support:



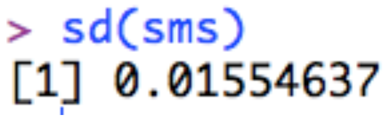
Simulating the thought experiment (3)

The results vary across our 10,000 “surveys” because of **sampling error.**



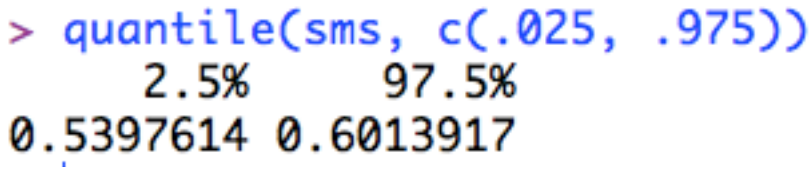
How much sampling error is there in our simulation?

The standard deviation:



```
> sd(sms)
[1] 0.01554637
```

95% of the samples had a mean between 0.54 and 0.60:

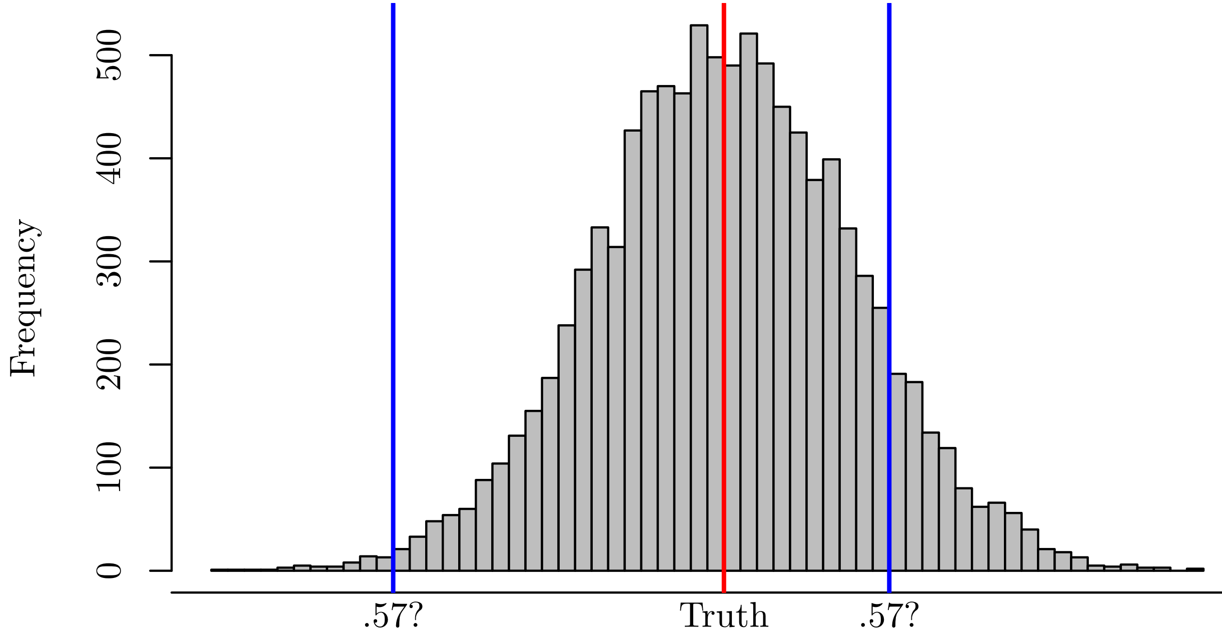


```
> quantile(sms, c(.025, .975))
     2.5%     97.5% 
0.5397614 0.6013917 
```

From thought experiment to margin of error

In a real survey, you get a **single number**.

The histogram from the thought experiment gives you a clue how close your number is to the “Truth”.



In our thought experiment (where we **know** the truth), 95% of the samples were within 0.031 of the truth.

In our actual survey (where **don’t know** the truth), we have 95% confidence that our estimate of 0.57 is within 0.031 of the truth.

Margin of error

Another way to get the margin of error

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

Our sample of 576 “Remains” and 430 “Leaves” and “Don’t knows” has a variance of .255.

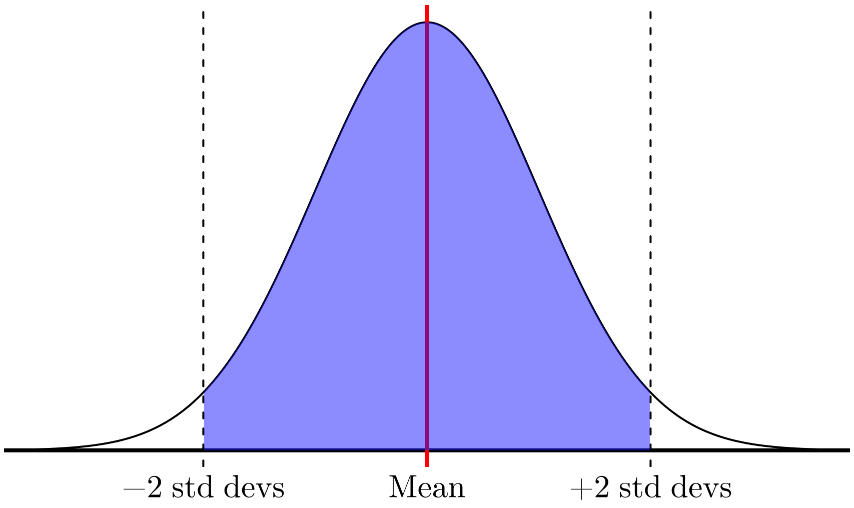
So the estimated standard deviation (**standard error**) of our estimate is:

$$\sqrt{\frac{.255}{1006}} = 0.0159$$

Compare: the standard deviation of our simulations was 0.0155

Another way to get the margin of error

In a Normal distribution, about 95% of the draws are within 2 standard deviations of the mean.



This indicates that in 95% of surveys we run, our answer should be within 2 standard deviations of the truth.

Given estimated standard deviation (standard error) of 0.016, we have a **margin of error** (2 times standard error) of .032.

Compare: our simulations implied a margin of error of 0.031.

Why were the polls wrong in the 2015 election?

measured value = true value + bias + random error

Random error?

No, because all of the polls were wrong in the same way.

It was **bias** (in the statistical sense):

- Conservative voters under-represented in surveys, Labour voters over-represented.
- Politically engaged over-represented.

Extremely difficult to get truly representative random sample.

Important: Margin of error captures random error (i.e. sampling error), not bias.

Quick recap: survey part

- **Margin of error**: an estimate of how much the estimate might vary due to random error (sampling error)
- In 95% of polls, the true value should be within margin of error (if no bias)
- Two ways we got the margin of error:
  - **Simulation** in R of 10,000 random samples of size 1,006 given a known level of support for "Leave"
  - **Central limit theorem:** approximation to a normal distribution
- Terminology:
  - **Sampling error**: variation in results from survey to survey due to variation in who gets randomly sampled
  - **Standard error**: our estimate of the standard deviation of the result across many surveys

Hypothesis testing

What does "**statistically significant**" mean? What is a "**p-value**"?

The logic of hypothesis testing is:

1. Calculate your statistic (e.g. support level, correlation, regression coefficient)

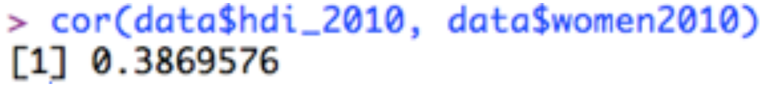
2. Define a "**null hypothesis**" (e.g. support is 50%, correlation is 0, regression coefficient is 0)

3. Calculate the **p-value**: probability of getting a statistic as large as yours if the null hypothesis were true (e.g. p=0.2, p=.002)

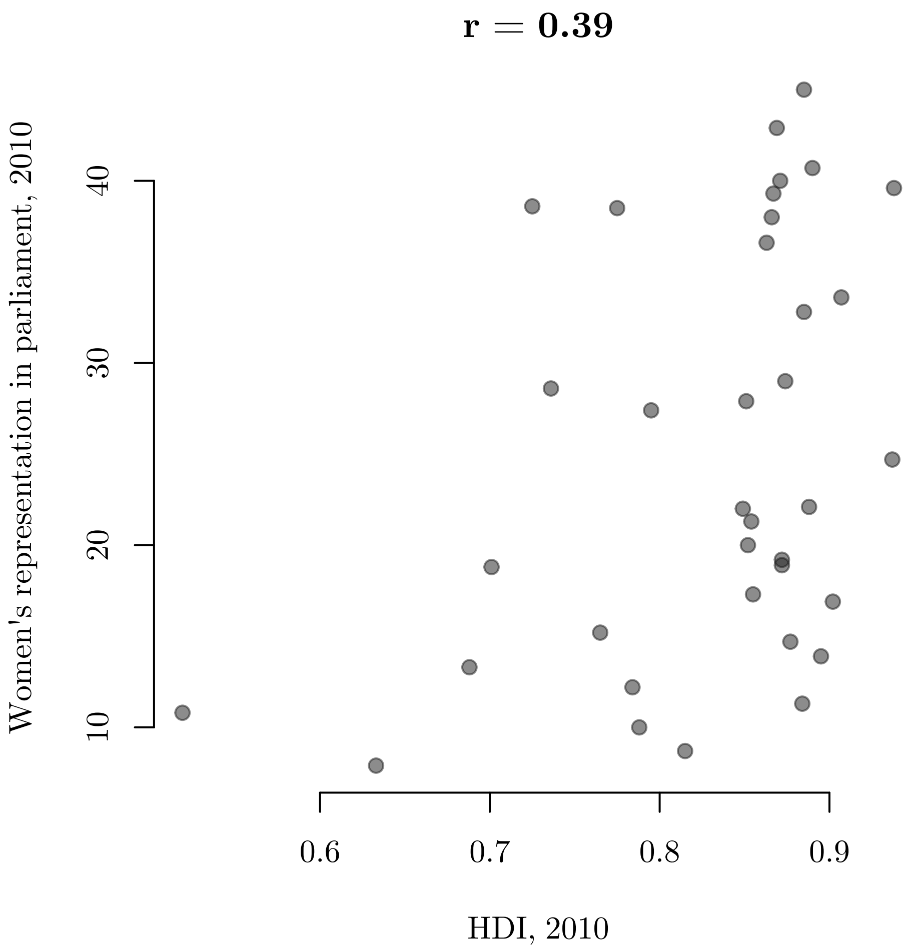
4. If p-value is low enough, reject null hypothesis, and say the correlation or regression coefficient is "statistically significant"

Example: correlation

Recall from Lab 2: in Lijphart's data there is a positive correlation across countries between the level of development and the proportion of women in parliament:



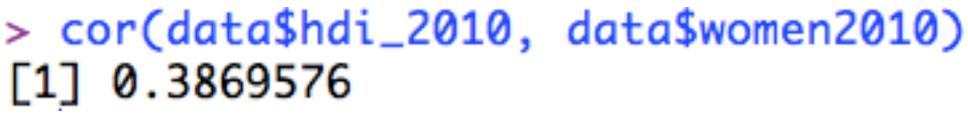
```
> cor(data$hdi_2010, data$women2010)
[1] 0.3869576
```



Example: correlation (2)

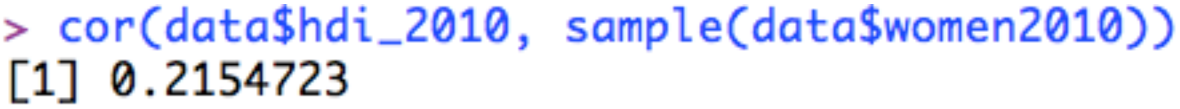
Our question: How likely are we to observe a correlation this large if there is actually no relationship?

Our approach: repeatedly reshuffle the data (so there actually *is* no relationship) and see how often we get a correlation as large as 0.39.



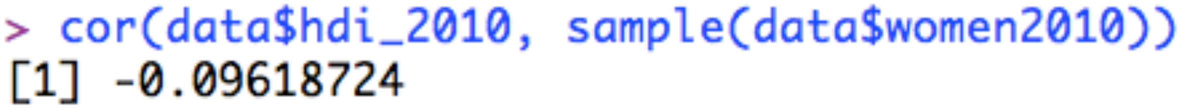
```
> cor(data$hdi_2010, data$women2010)
[1] 0.3869576
```

The actual data



```
> cor(data$hdi_2010, sample(data$women2010))
[1] 0.2154723
```

First reshuffle



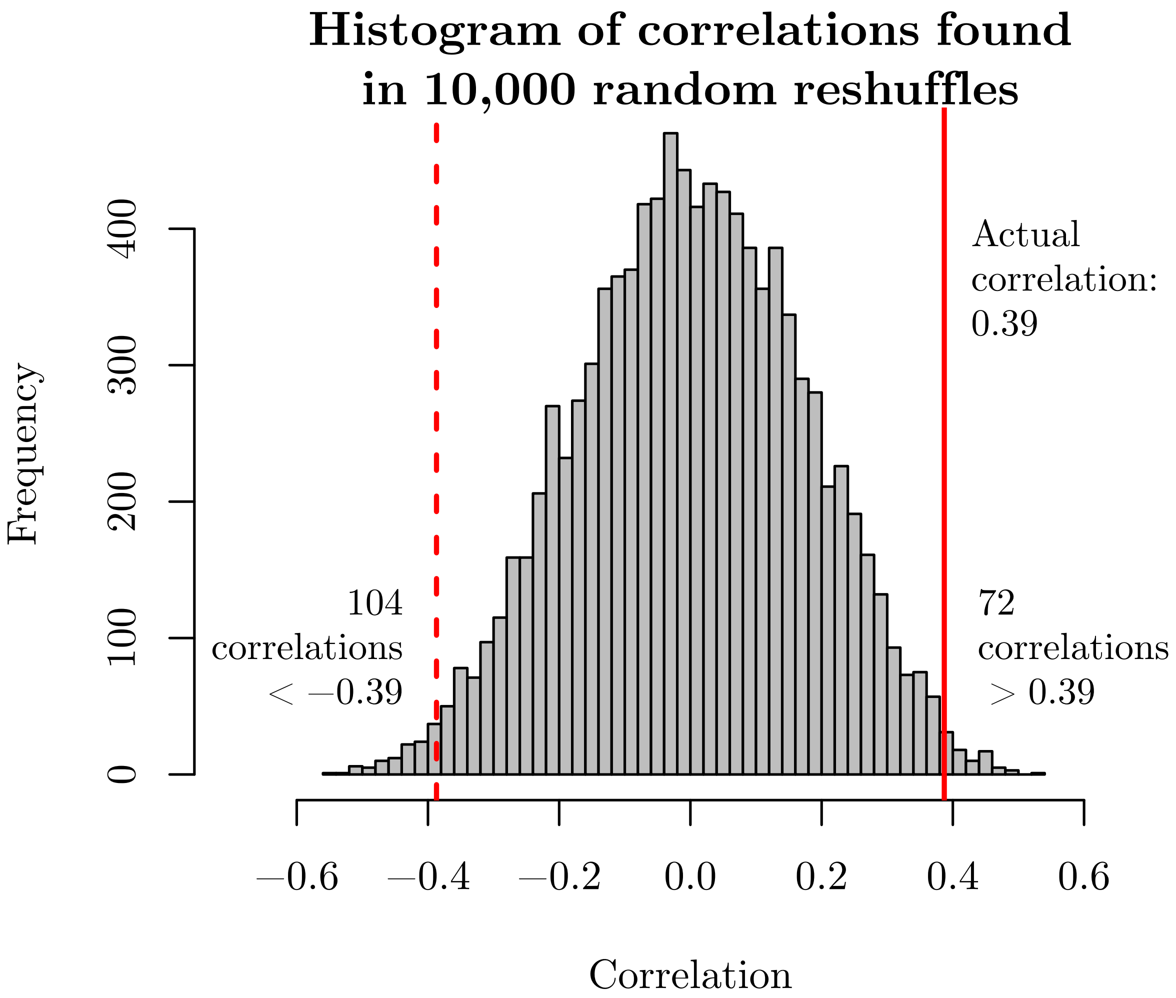
```
> cor(data$hdi_2010, sample(data$women2010))
[1] -0.09618724
```

Second reshuffle

Example: correlation (3)

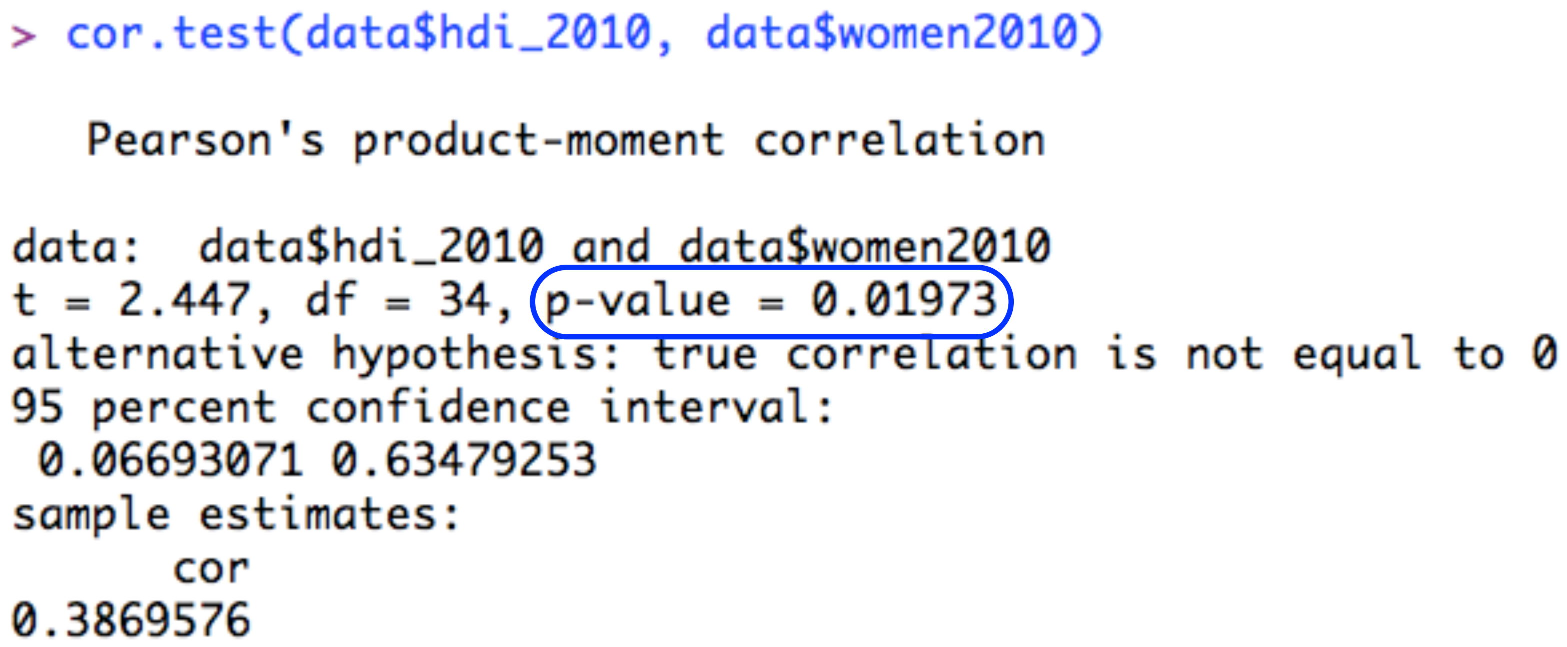
Our question: How likely are we to observe a correlation this large if there is actually no relationship?

Our answer: p = 0.0176. In 10,000 reshuffles, 176 had correlations larger than 0.39 or smaller than -0.39.



Example: correlation (4)

How you did this in Lab 2:



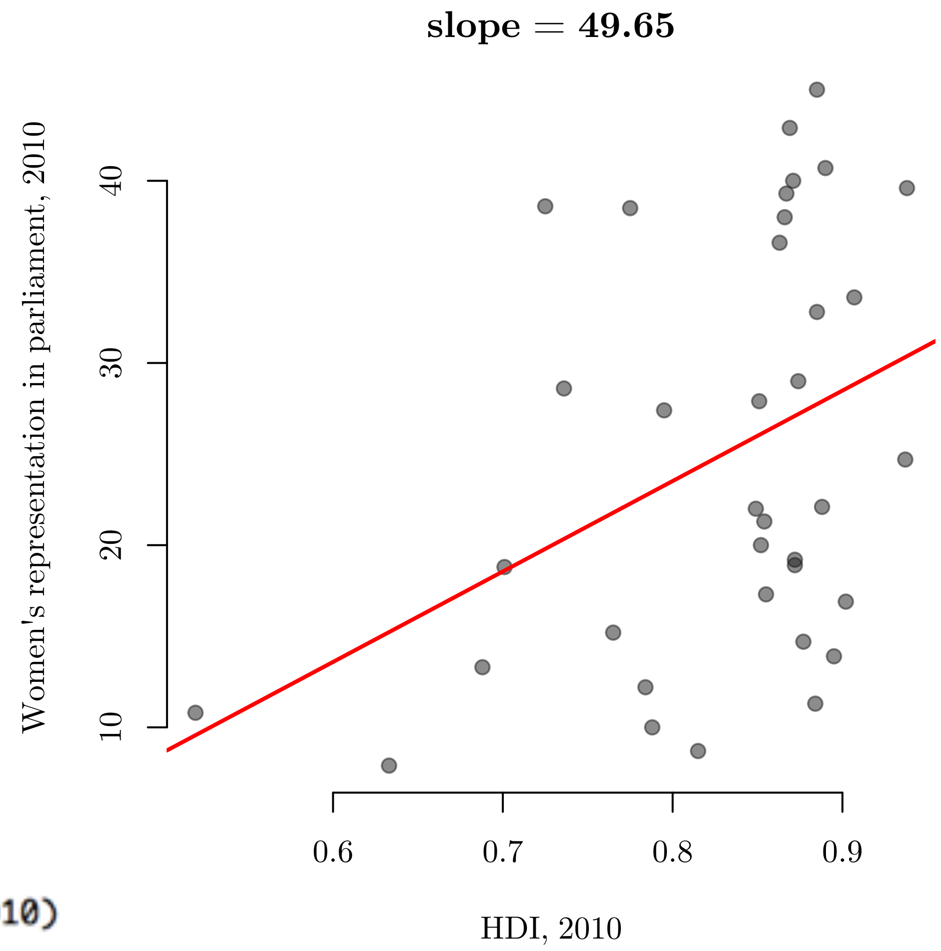
```
> cor.test(data$hdi_2010, data$women2010)

	Pearson's product-moment correlation

data:  data$hdi_2010 and data$women2010
t = 2.447, df = 34, p-value = 0.01973
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06693071 0.63479253
sample estimates:
      cor 
0.3869576 
```

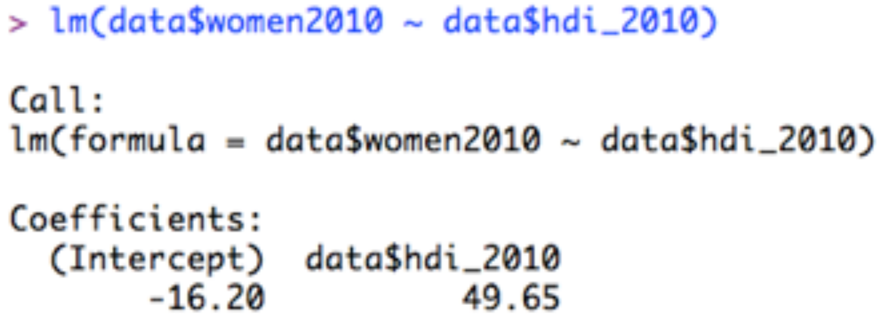
The hypothesis testing recipe applied

1. Calculate your statistic	Correlation is 0.39
2. Define a null hypothesis	No relationship
3. Calculate the **p-value**	0.0176
4. If p-value is low enough, reject null hypothesis	Null hypothesis rejected!

Example: bivariate regression

Recall from Lab 2: in Lijphart's data, positive relationship between development and women's representation in parliament:



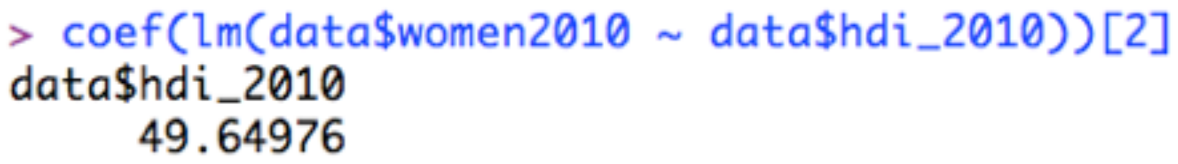


```
> lm(data$women2010 ~ data$hdi_2010)

Call:
lm(formula = data$women2010 ~ data$hdi_2010)

Coefficients:
  (Intercept)  data$hdi_2010
       -16.20          49.65
```

Example: bivariate regression (2)

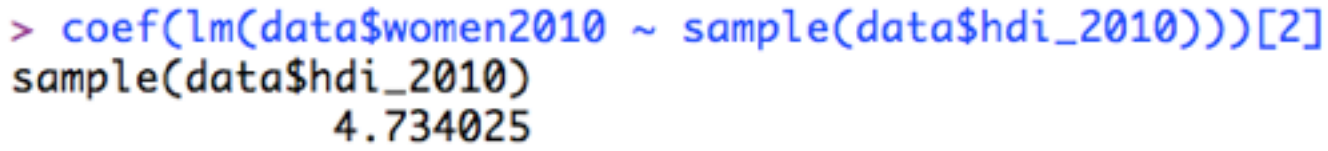
Our question: How likely are we to observe a slope this large if there is actually no relationship?

Our approach: repeatedly reshuffle one of the variables (so there actually *is* no relationship) and see how often we get a slope as large as 49.65.



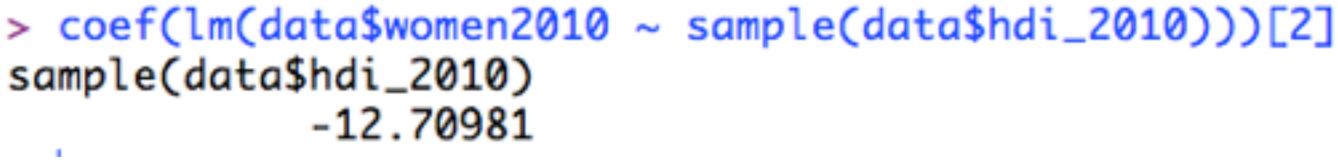
```
> coef(lm(data$women2010 ~ data$hdi_2010))[2]
data$hdi_2010
     49.64976
```

The actual data



```
> coef(lm(data$women2010 ~ sample(data$hdi_2010)))[2]
sample(data$hdi_2010)
             4.734025
```

First reshuffle



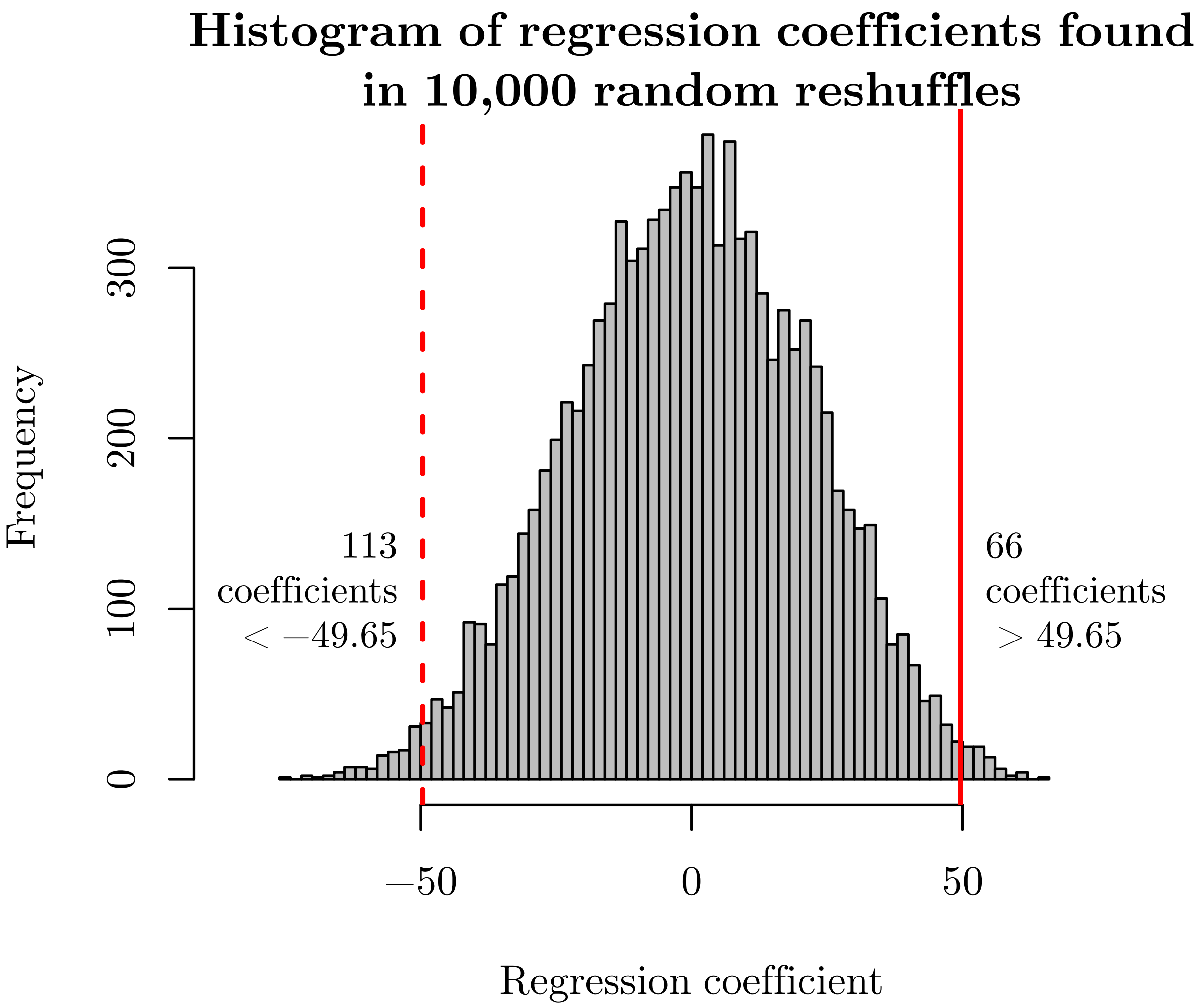
```
> coef(lm(data$women2010 ~ sample(data$hdi_2010)))[2]
sample(data$hdi_2010)
            -12.70981
```

Second reshuffle

Example: bivariate regression (3)

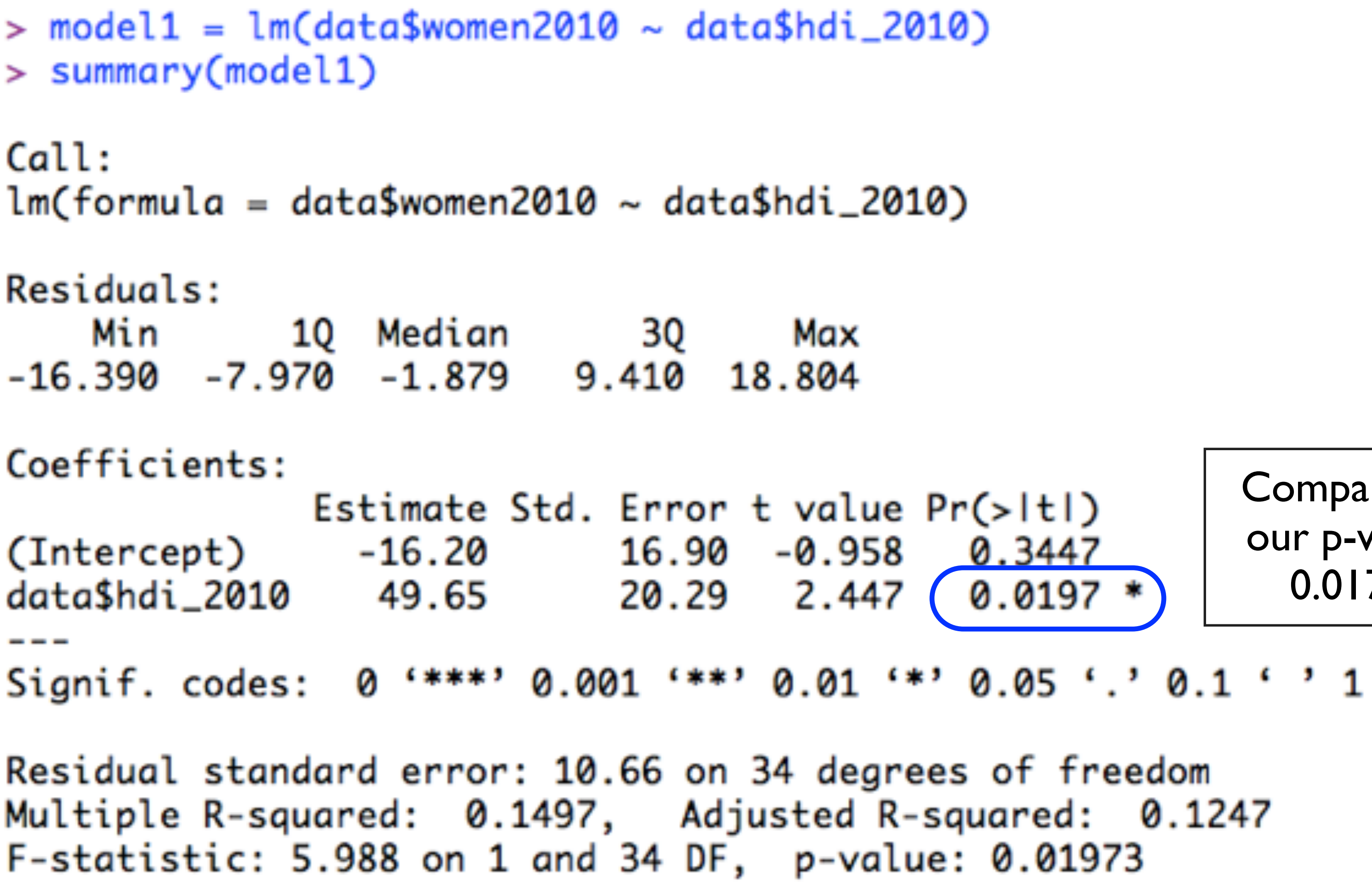
Our question: How likely are we to observe a slope this large if there is actually no relationship?

Our answer: p = 0.0179. In 10,000 reshuffles, 179 had slopes larger than 49.65 or smaller than -49.65.



Example: bivariate regression (4)

How this looked in Lab 3:



```
> model1 = lm(data$women2010 ~ data$hdi_2010)
> summary(model1)

Call:
lm(formula = data$women2010 ~ data$hdi_2010)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.390  -7.970  -1.879   9.410  18.804 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)     -16.20      16.90  -0.958   0.3447  
data$hdi_2010    49.65      20.29   2.447   0.0197 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 10.66 on 34 degrees of freedom
Multiple R-squared:  0.1497,	Adjusted R-squared:  0.1247 
F-statistic: 5.988 on 1 and 34 DF,  p-value: 0.01973
```

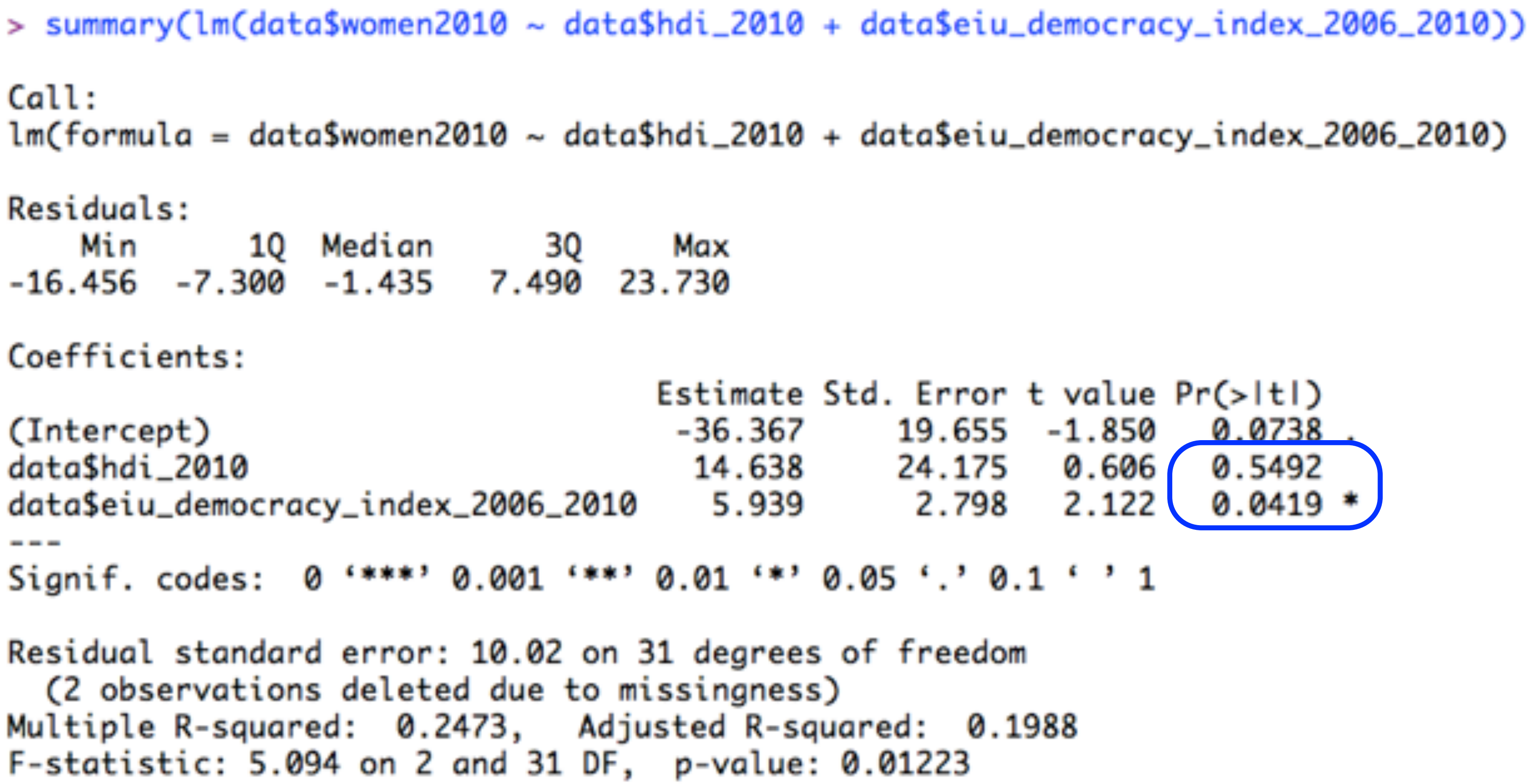
Compare to our p-value: 0.0179

The hypothesis testing recipe applied

1. Calculate your statistic	Slope is 49.65
2. Define a null hypothesis	Slope = 0 (no relationship)
3. Calculate the **p-value**	0.0179
4. If p-value is low enough, reject null hypothesis	Null hypothesis rejected!

Extending to multivariate regression

Multivariate regression is more complicated, but interpretation of the p-values is same: “If this variable were *really not related* to the outcome, how unusual would it be to see a slope this big?”



```
> summary(lm(data$women2010 ~ data$hdi_2010 + data$eiu_democracy_index_2006_2010))

Call:
lm(formula = data$women2010 ~ data$hdi_2010 + data$eiu_democracy_index_2006_2010)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.456  -7.300  -1.435   7.490  23.730 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)  
(Intercept)                         -36.367     19.655  -1.850   0.0738 .
data$hdi_2010                        14.638     24.175   0.606   0.5492  
data$eiu_democracy_index_2006_2010    5.939      2.798   2.122   0.0419 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 10.02 on 31 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.2473,	Adjusted R-squared:  0.1988 
F-statistic: 5.094 on 2 and 31 DF,  p-value: 0.01223
```

"If this variable were *really not related* to the outcome, how unusual would it be to see a slope this big?"

What do we mean by **really not related**?

The null hypothesis is a claim about the **truth** that we test with our **sample**.

This makes sense in a survey:

- Given a random **sample** of n=1,006 in which 57% said "Remain",
- can we reject the null hypothesis that actually only 50% of **all GB adults** support "Remain"?

The **truth** is the number in the population.

What about when we're talking about 36 democracies in Lijphart's data? This isn't a sample! What is the **truth** there? What do the p-values, standard errors mean?

Three ways you can view standard errors, p-values in Lijphart (and other research not based on analysis of random samples)

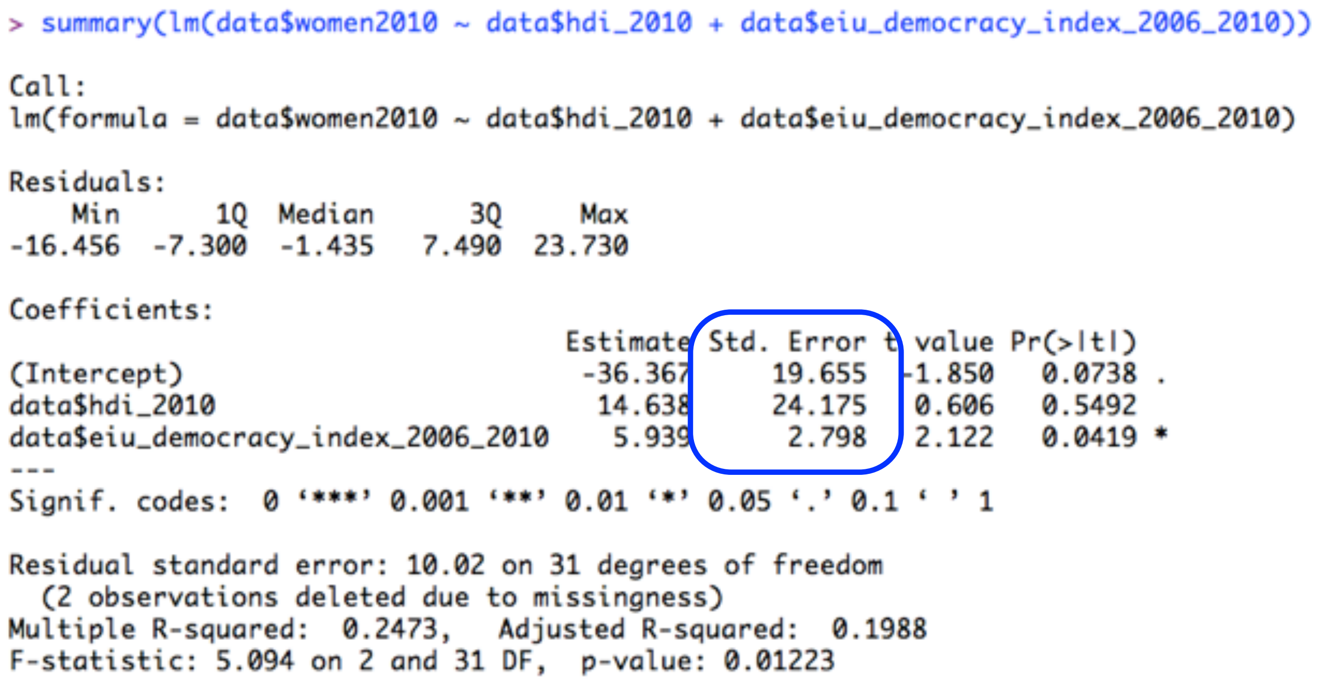
1. This **is** like a sample: The 36 countries we observe can be viewed as a sample from a population of **hypothetical countries**.
2. This **is** like a sample: The **countries** aren't sampled, but the **residuals** can be thought of as having been sampled.
3. This is **not** like a sample, and it's philosophically inappropriate to apply a framework for statistical inference developed for surveys to situations like this…
    A. …so we should use Bayesian statistics.
    B. …but we use the conventional (frequentist) approach anyway.

Kellstedt and Whitten: "no clear scientific consensus" (141)

Finally, back to margin of error

Recall the margin of error (= 2 times standard error) gave us a sense of how much the estimate would vary across many surveys.

The standard error in regression output plays the same role: in 95% of surveys/repeated samples, the difference between our estimate and the true value is less than 2 times the standard error.



```
> summary(lm(data$women2010 ~ data$hdi_2010 + data$eiu_democracy_index_2006_2010))

Call:
lm(formula = data$women2010 ~ data$hdi_2010 + data$eiu_democracy_index_2006_2010)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.456  -7.300  -1.435   7.490  23.730 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)  
(Intercept)                         -36.367     19.655  -1.850   0.0738 .
data$hdi_2010                        14.638     24.175   0.606   0.5492  
data$eiu_democracy_index_2006_2010    5.939      2.798   2.122   0.0419 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 10.02 on 31 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.2473,	Adjusted R-squared:  0.1988 
F-statistic: 5.094 on 2 and 31 DF,  p-value: 0.01223
```

Now you should understand:

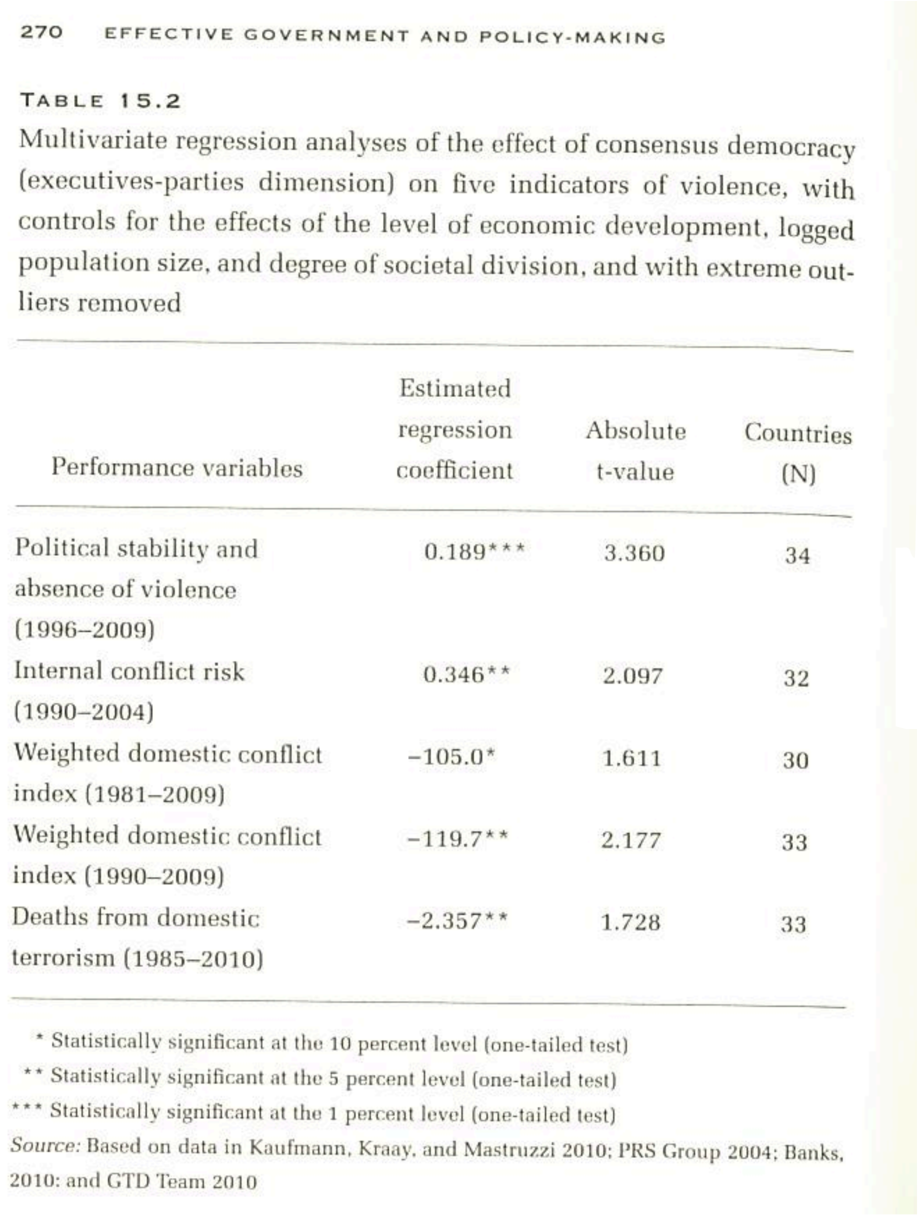
Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(1)	(2)	(3)
Intercept	-1.629* (0.509)	-3.166* (0.511)	-2.982* (0.527)
Chocolate consumption per capita (log scale)	2.092* (0.298)	1.026* (0.326)	0.709 (0.415)
GDP/capita (thousands of USD)		0.105* (0.024)	0.106* (0.024)
NW Europe			0.549 (0.452)
R^2	0.70	0.85	0.86
N	34	34	34

Standard errors in parentheses. * Indicates $p<0.05$

- what a dependent variable is
- what an independent variable is
- what the coefficients mean (intercept, slopes)
- what the stars mean (i.e. what $p<0.05$ means)
- what the standard errors mean

And this too!



270 EFFECTIVE GOVERNMENT AND POLICY-MAKING

TABLE 15.2

Multivariate regression analyses of the effect of consensus democracy (executives-parties dimension) on five indicators of violence, with controls for the effects of the level of economic development, logged population size, and degree of societal division, and with extreme outliers removed

Performance variables	Estimated regression coefficient	Absolute t-value	Countries (N)
Political stability and absence of violence (1996–2009)	0.189***	3.360	34
Internal conflict risk (1990–2004)	0.346**	2.097	32
Weighted domestic conflict index (1981–2009)	−105.0*	1.611	30
Weighted domestic conflict index (1990–2009)	−119.7**	2.177	33
Deaths from domestic terrorism (1985–2010)	−2.357**	1.728	33

* Statistically significant at the 10 percent level (one-tailed test)

** Statistically significant at the 5 percent level (one-tailed test)

*** Statistically significant at the 1 percent level (one-tailed test)

Source: Based on data in Kaufmann, Kraay, and Mastruzzi 2010; PRS Group 2004; Banks, 2010; and GTD Team 2010

- what the dependent and independent variables are
- what Lijphart means by “controlling for” three other variables
- what the stars mean

Looking ahead

- Lecture next week: applying what you've learned to readings, tutorial essays, exams
- Labs next week: multivariate regressions useful for essays
- Essays due week 2 of TT
  - Look for detailed guidelines on WebLearn
  - Drop-in sessions first week of TT (look for emails)

- **Speaker series:** 4pm Wed, March 9, Simon Jackman (Stanford University) on how social science methods are used outside of academia (MRB Lecture Theatre)