Inference

Week 8

9 March, 2015

Prof. Andrew Eggers

Questions from last week

- "You showed a linear relationship between age and mortality but that's probably not right!" Yes. One option is to make it quadratic:

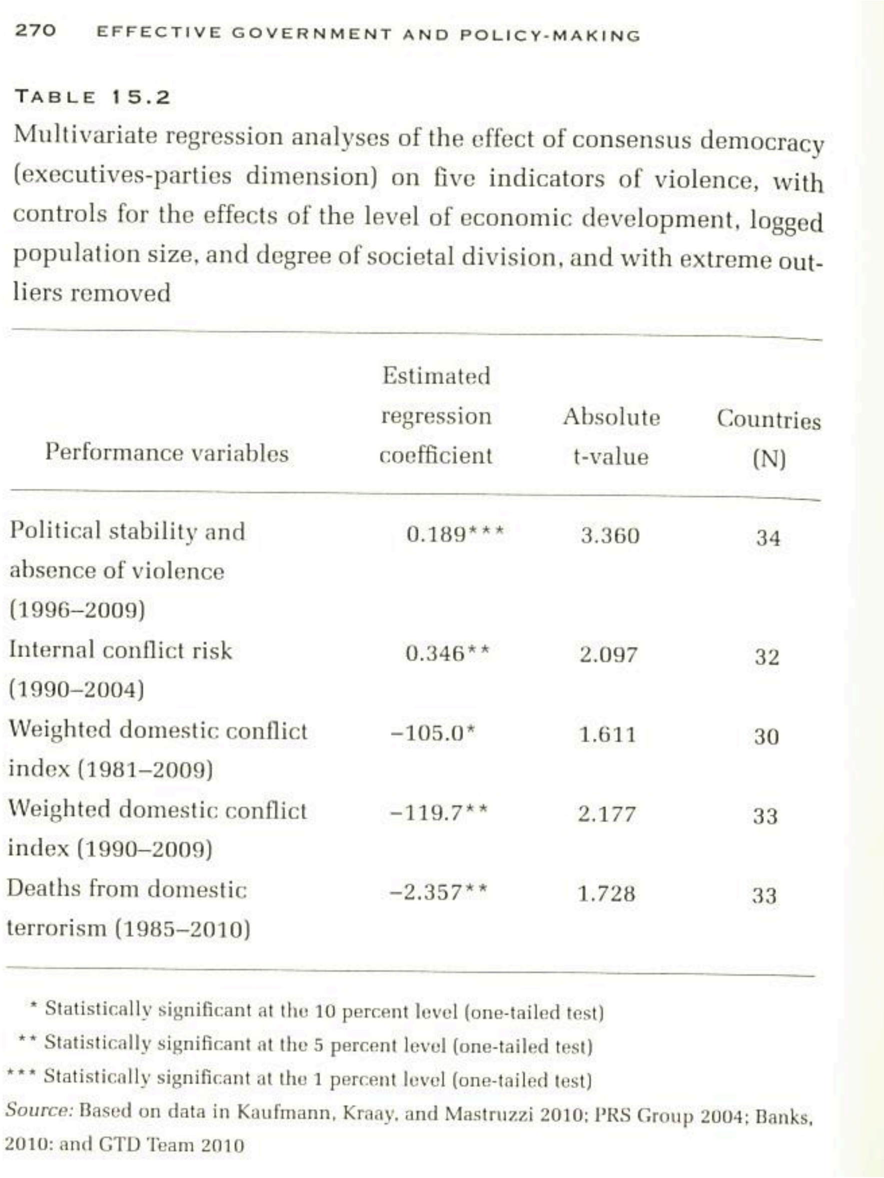
$$\mathrm{Mortality}_i = \beta_0 + \beta_1 \mathrm{PipeSmoker}_i + \beta_2 \mathrm{Age}_i$$

$$\mathrm{Mortality}_i = \beta_0 + \beta_1 \mathrm{PipeSmoker}_i + \beta_2 \mathrm{Age}_i + \beta_3 \mathrm{AgeSquared}_i$$

- "For my essay, can I add data from another source to evaluate one of Lijphart's claims?" Yes!

- "Why ordinary least **squares**? Why minimize sum of **squared** residuals rather than, say, sum of absolute residuals or cubed residuals?" Four answers:
  - Good point: sometimes you might want to minimize something else
  - The minimization works much better with squares
  - With categorical independent variables, OLS corresponds to the mean
  - OLS is best estimate (MLE) if the errors are normally distributed

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)

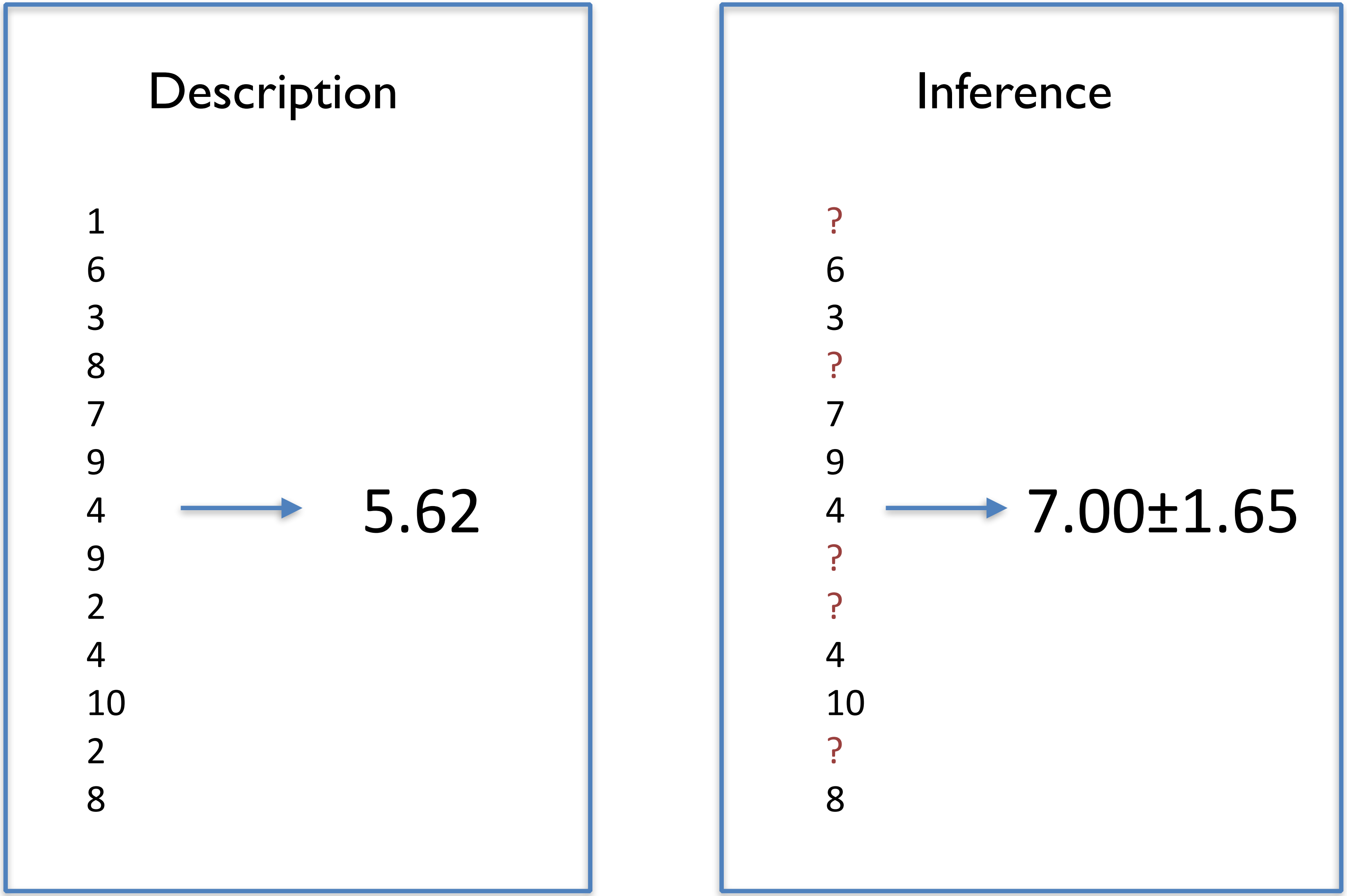
** Statistically significant at the 5 percent level (one-tailed test)

*** 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?
- Why is it difficult to evaluate the results of studies using "Big Data"?

Description versus inference



Description

1
6
3
8
7
9
4
9
2
4
10
2
8

→ 5.62

Inference

?
6
3
?
7
9
4
?
?
4
10
?
8

→ 7.00±1.65

Two kinds of statistical inference

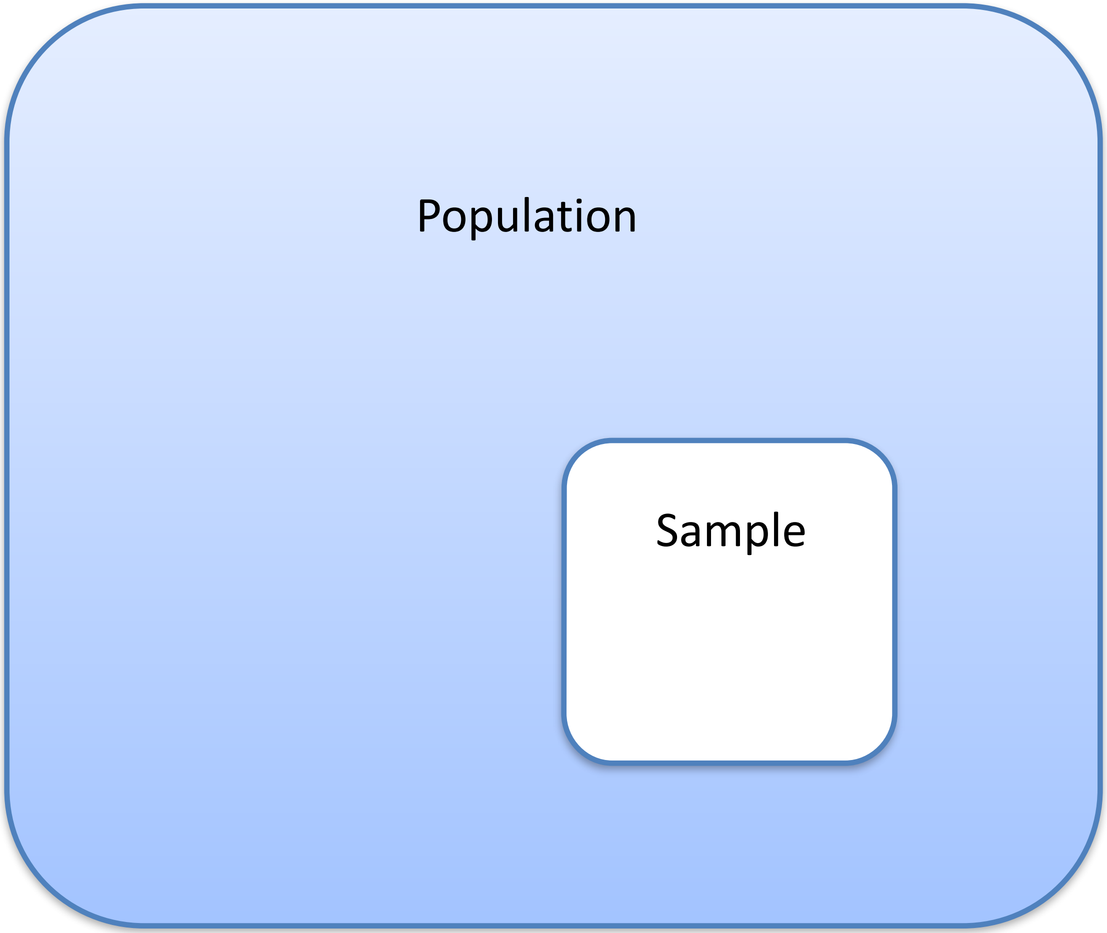
Population inference

Making statements about a **population** from a **sample**, i.e. describing the **whole** based on **part**.

Causal inference

Making statements about **something that didn't happen** based on **what did happen**.

Population inference



Random sampling is important and difficult

Ideally, the sample is a microcosm of the population => random sampling.

How would you obtain data from a random sample of

- PPE students?
- Oxford residents?
- UK voters?

How do you obtain inferences about a population if your sample is not random?

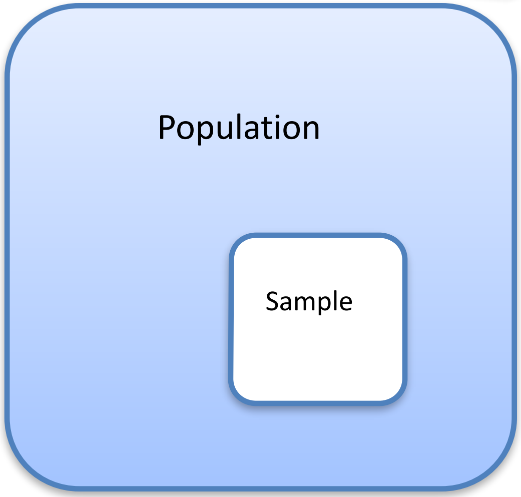
The logic of population inference with random sampling

You have the voting intentions of a **simple random sample** of 1000 UK voters: 333 say they will vote Conservative.

Simple random sample: all individuals have an equal and independent probability of being sampled.

What is your estimate of Conservative voting intentions in the **population**?

How precise is that estimate? How surprised should we be if the Conservatives only win 25%?



Assessing accuracy with a thought experiment

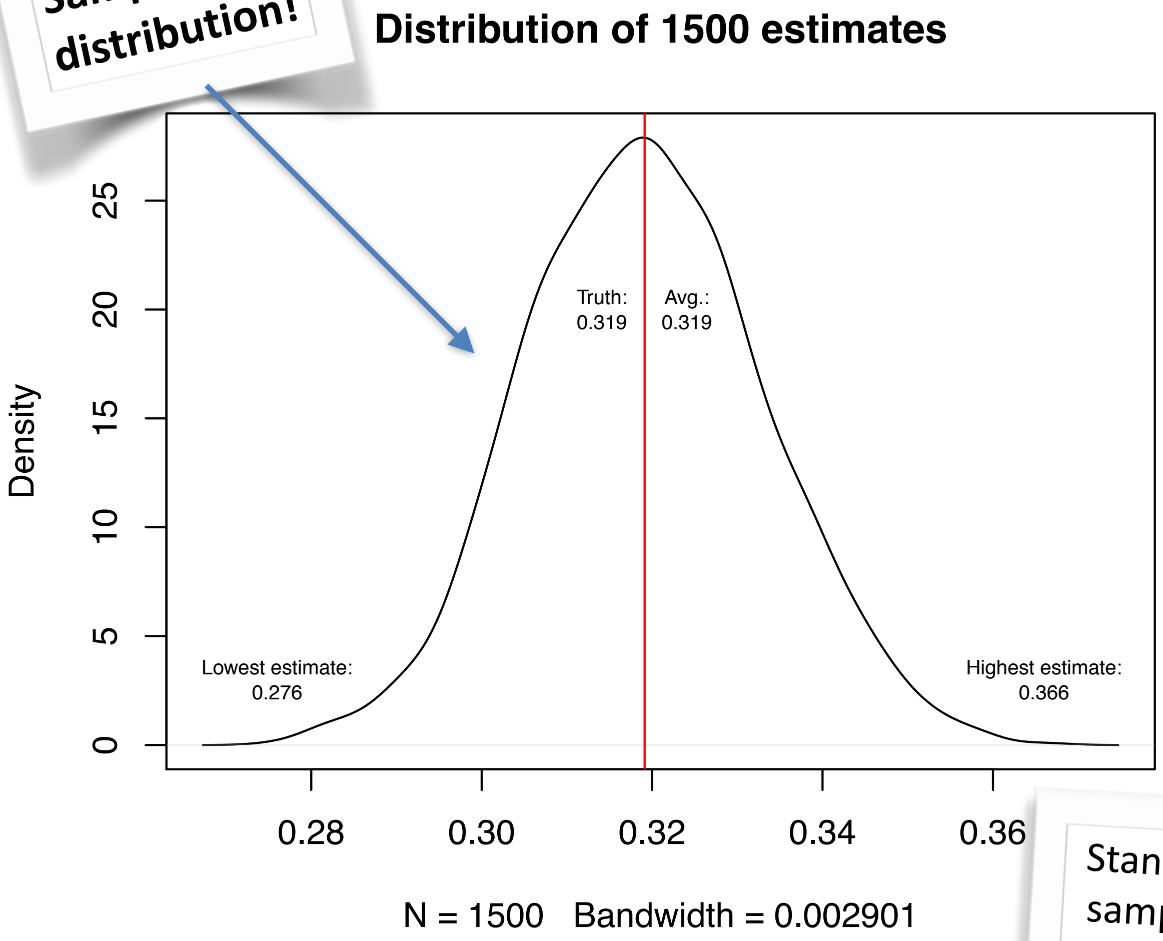
Suppose you had data from the whole population. If you sample just 1000 voters, how far off are you on average?

```
# First, create the “actual” data
# Suppose 9,573,631 Con voters out of
30 mill total

> con_voters = 9573631
> number_of_voters = 30000000
> intentions = c(rep(1, times =
con_voters), rep(0, times =
number_of_voters - con_voters))

# Make sure it makes sense

> mean(intentions)
[1] 0.319121
```

```
# Here’s how we take a single sample of
size n
> n = 1000
> a.sample = sample(x = intentions, size
= n)
> mean(a.sample)
[1] 0.286 # the answer this sample would
give us; varies from sample to sample

# Here’s how we do it 1500 times
> m = 1500
> estimates = c()  # for storage
> for(i in 1:m){
     estimates = c(estimates,
mean(sample(x = intentions, size = n)))
}
```

Assessing accuracy in the thought experiment

Sampling distribution!



How did we do?

Of 1500 samples (each with 1000 voters),

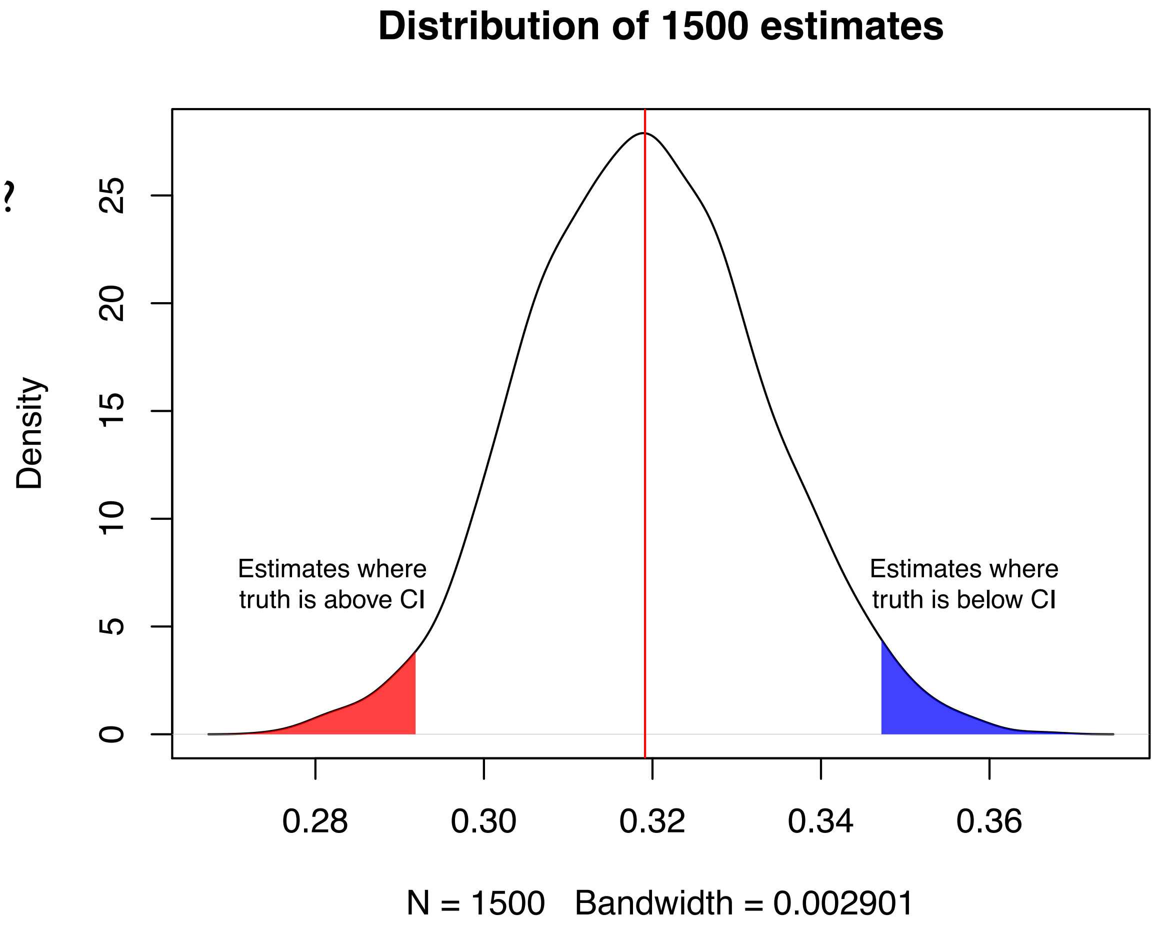
- 52% of samples gave an estimate within .01 of the truth
- 85% within .02
- 97% within .03
- 99.9% within .04

Standard deviation of sampling distribution = **standard error**

Building confidence intervals from the true sampling distribution

Suppose we build an interval [*Estimate - x, Estimate* + *x*] around every estimate. (x is **margin of error**.) For how many estimates does this interval include the truth?

x	Proportion of estimates for which interval includes truth
.01	*52%*
.02	*85%*
.0275	*95%*
.03	*97%*



Building confidence intervals from the **estimated** sampling distribution

What about when we don't know the truth?

1. We take a **random sample** from the population.
2. We make an **estimate** based on the sample.
3. *Using our sample*, we generate a **sampling distribution** (which shows how estimates vary from one sample to the next, and allows us to calculate a **confidence interval**).

Step 3 can be done via *simulation* as above (see next example), or with asymptotic approximations. (See readings.)

Key assumption: *shape* of the sampling distribution around the truth can be estimated using the sample.

Application to regression

This works for regression too!

e.g. in above example, could use OLS to estimate:

$$\mathrm{VoteCon_i} = \beta_0 + \beta_1 \mathrm{Educ_i} + u_i$$

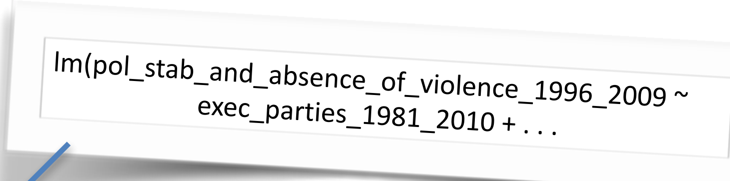
and generate a sampling distribution for $\hat{\beta}_1$ as we did above for the mean.

Can also do a **hypothesis test**: "if in fact there were no relationship in the population, how likely is it that we would draw a sample in which $\hat{\beta}_1$ is at least as large as the one we observed?"

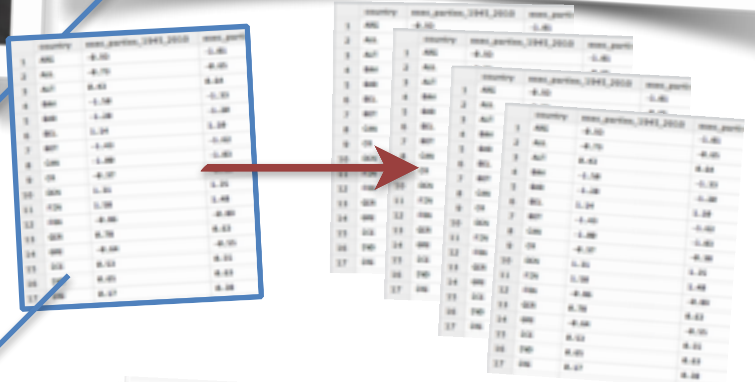
Application to Lijphart

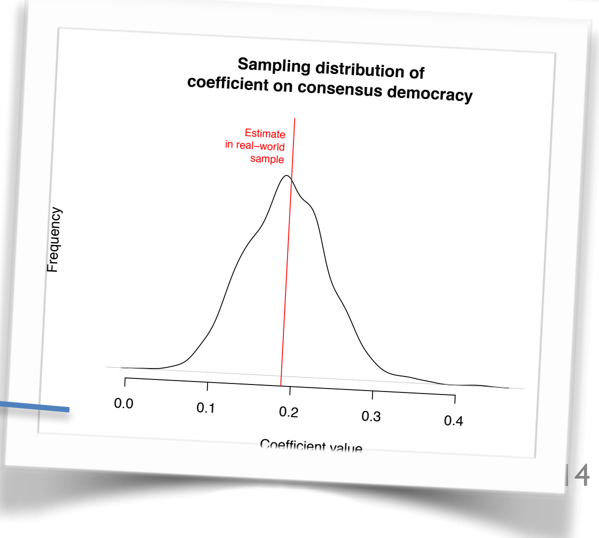
Procedure:

1. Observe measures of consensus democracy & violence from 34 democracies
2. Estimate the relationship between consensus democracy and absence of violence in the sample
3. Use the sample to generate a sampling distribution for your estimate
    1. Expand the sample into a "population" — many replicates of these 34 countries
    2. Sample 34 countries at random from this "population"; estimate the relationship; store and repeat

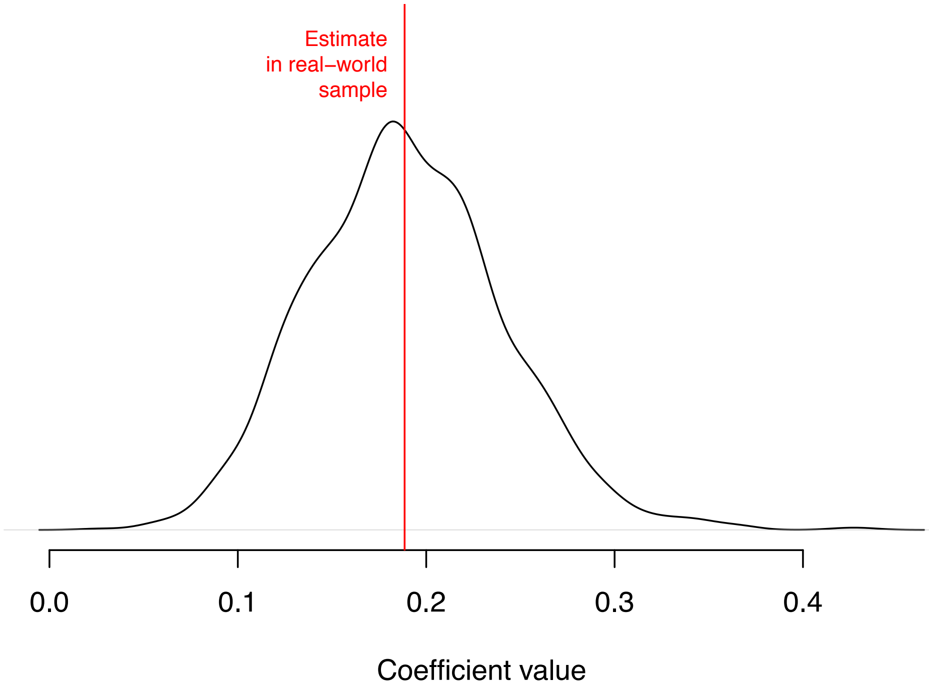


```
lm(pol_stab_and_absence_of_violence_1996_2009 ~
        exec_parties_1981_2010 + . . .
```

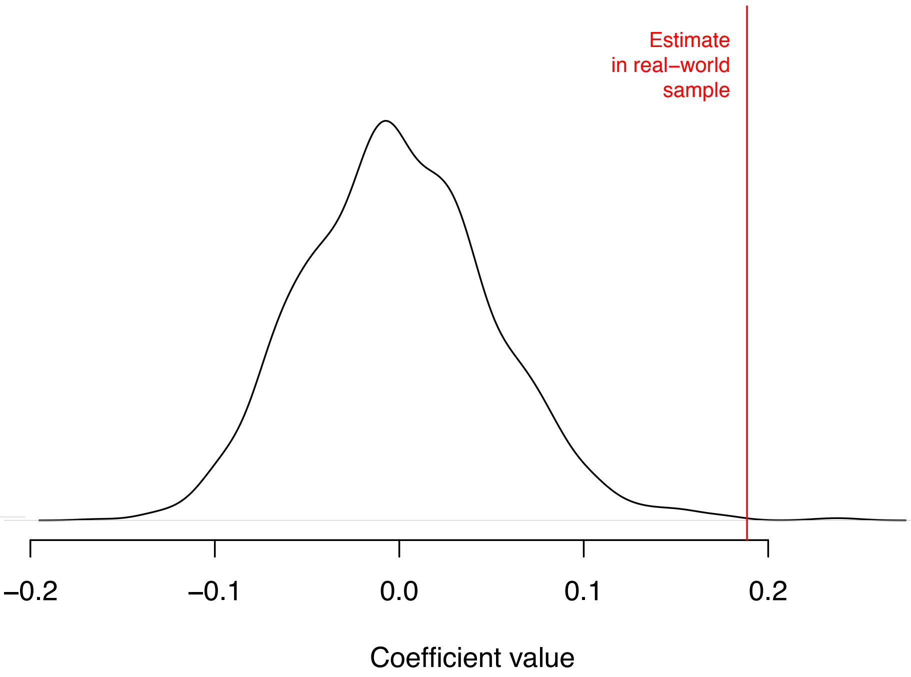




Hypothesis testing using the sampling distribution



Sampling distribution of coefficient on consensus democracy



Assumed sampling distribution of coefficient on consensus democracy under null hypothesis

Is it likely that we would get an estimate as large as .189 if true coefficient is 0?

If the shape of the sampling distribution does not depend on the true value of the coefficient, we can recenter the sampling distribution at zero (as above) and compute the proportion of samples producing an estimate larger than .189.

This proportion (here, 2/1500 = .0013) is the (one-tailed) **p-value**.

Bringing it all together: the logic

We have a **question** about a population (week 1); we work out issues of design & measurement (weeks 2 & 4)

$\beta = ?$

We have a **sample** from the **population** (week 3).

We use the sample to produce an **estimate** (weeks 5-7).

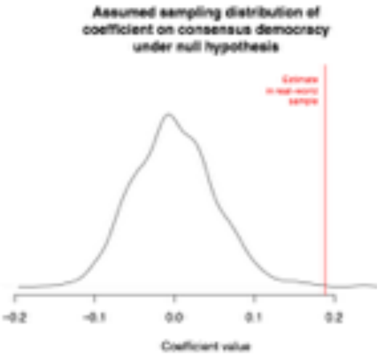
$\hat{\beta}$

We want to know how confident to be about this estimate (week 8).

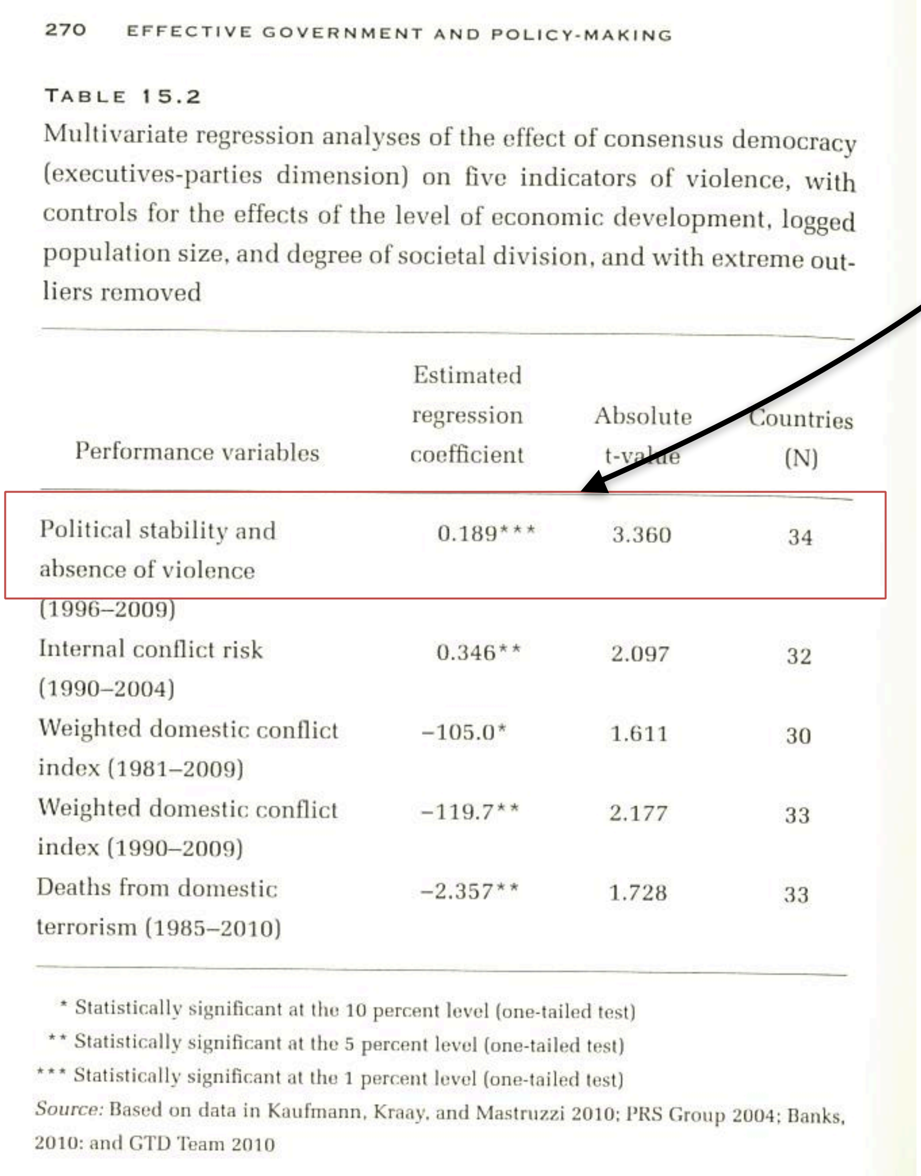
Confidence interval: interval such that 95% of similar samples would include the true population parameter.

$\hat{\beta} \pm x$

p-value: probability of getting an estimate at least this far from zero if the true parameter were in fact zero.



Replicating Lijphart’s analysis & interpreting the results



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)

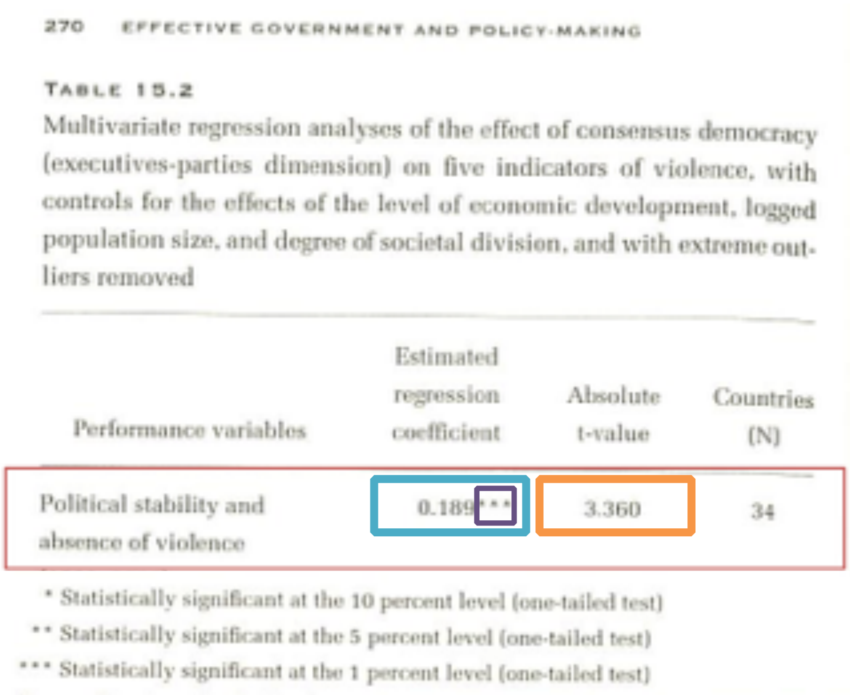
Source: Based on data in Kaufmann, Kraay, and Mastruzzi 2010; PRS Group 2004; Banks, 2010; and GTD Team 2010

We’re going to reproduce this result (almost)!

Checklist:

- download data
- make “logged population variable
- identify outliers from text and in data
- run regression controlling for specified variables
- look at `summary()`

Replicating Lijphart's analysis & interpreting the results

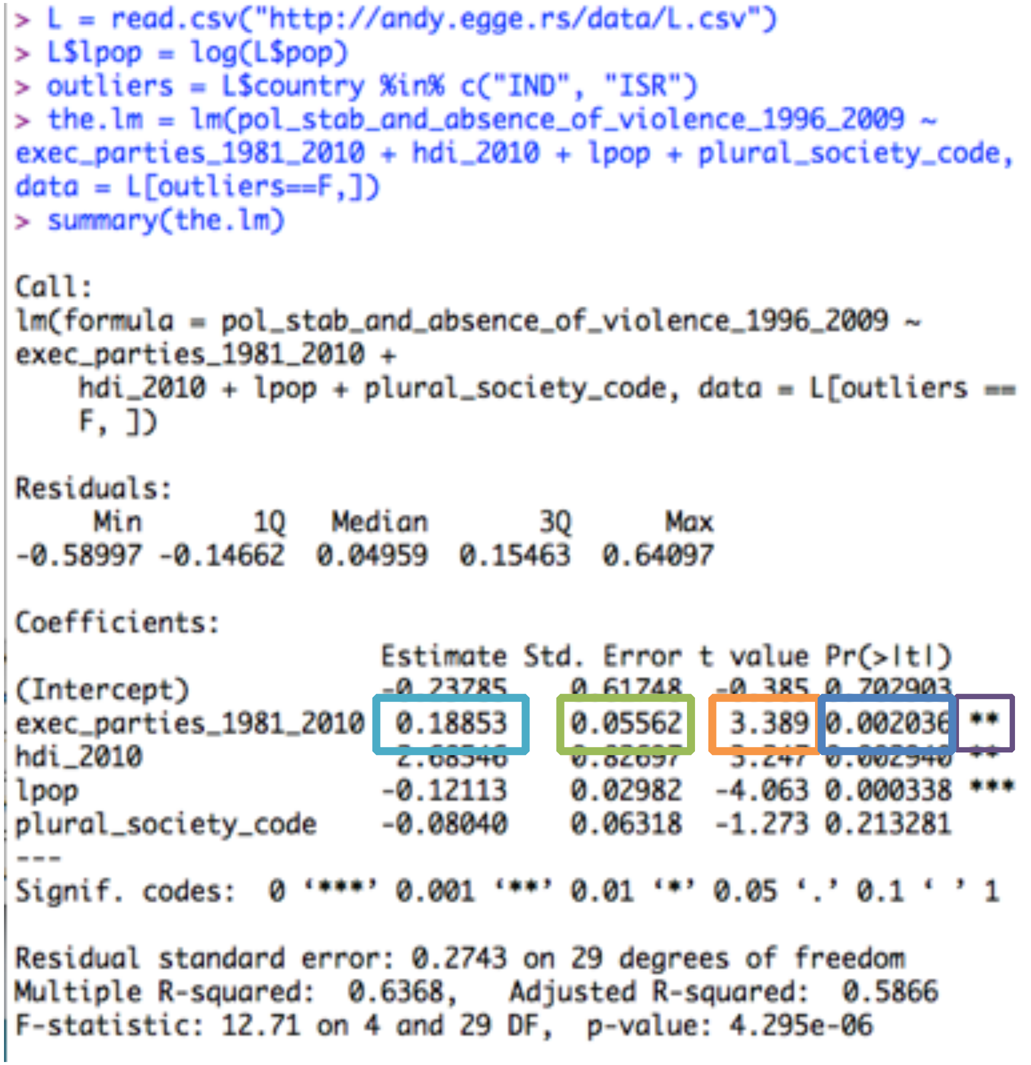


270 EFFECTIVE GOVERNMENT AND POLICY-MAKING

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	0.189***	3.360	34

* 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)



```
> L = read.csv("http://andy.egge.rs/data/L.csv")
> L$lpop = log(L$pop)
> outliers = L$country %in% c("IND", "ISR")
> the.lm = lm(pol_stab_and_absence_of_violence_1996_2009 ~
exec_parties_1981_2010 + hdi_2010 + lpop + plural_society_code,
data = L[outliers==F,])
> summary(the.lm)

Call:
lm(formula = pol_stab_and_absence_of_violence_1996_2009 ~
exec_parties_1981_2010 +
    hdi_2010 + lpop + plural_society_code, data = L[outliers ==
    F, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.58997 -0.14662  0.04959  0.15463  0.64097

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            -0.23785    0.61748  -0.385 0.702903
exec_parties_1981_2010  0.18853    0.05562   3.389 0.002036 **
hdi_2010                2.68546    0.82697   3.247 0.002940 **
lpop                   -0.12113    0.02982  -4.063 0.000338 ***
plural_society_code    -0.08040    0.06318  -1.273 0.213281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2743 on 29 degrees of freedom
Multiple R-squared:  0.6368,	Adjusted R-squared:  0.5866
F-statistic: 12.71 on 4 and 29 DF,  p-value: 4.295e-06
```

Notes:

- t-value is the coefficient divided by its standard error
- asterisks refer to p-value being below certain cutoffs

What makes a coefficient estimate more precise?

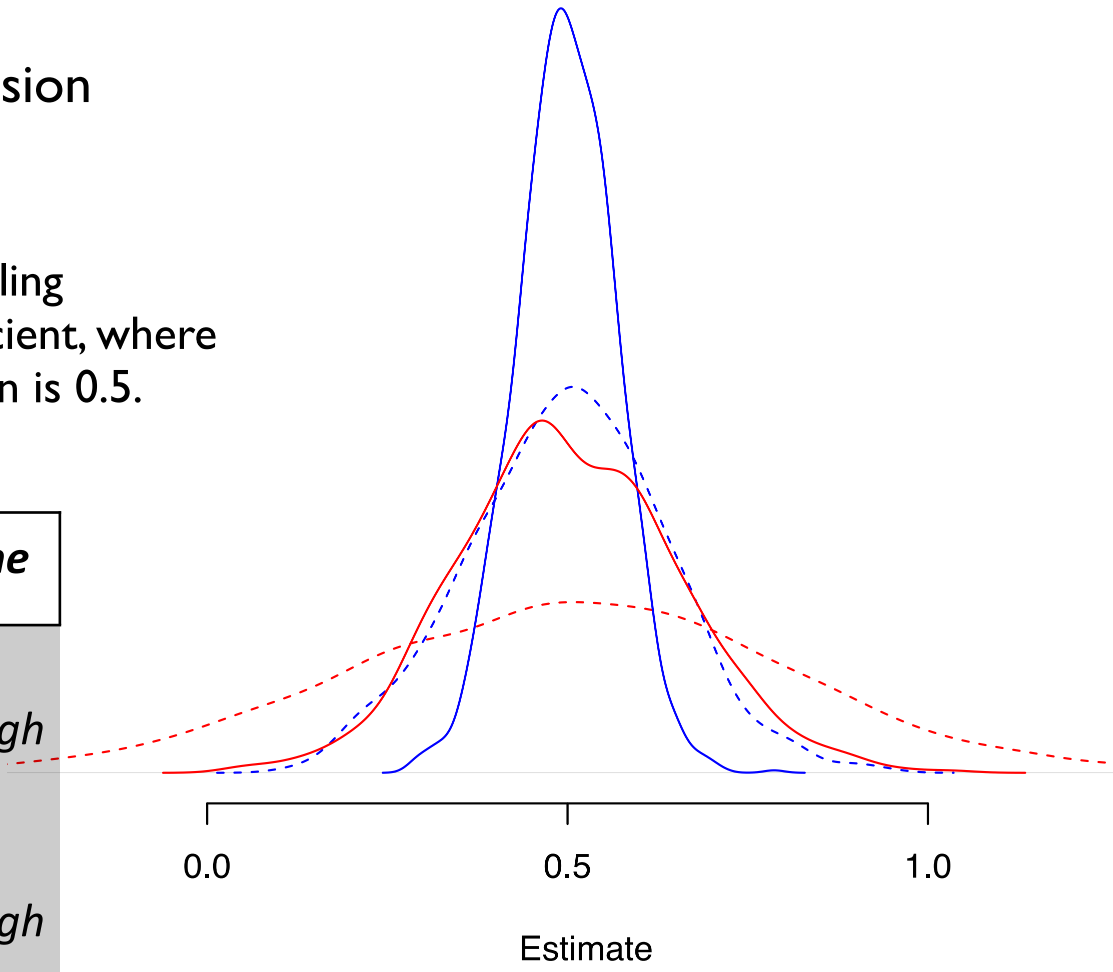
Two key factors:

- larger sample
- less noise/error in the regression

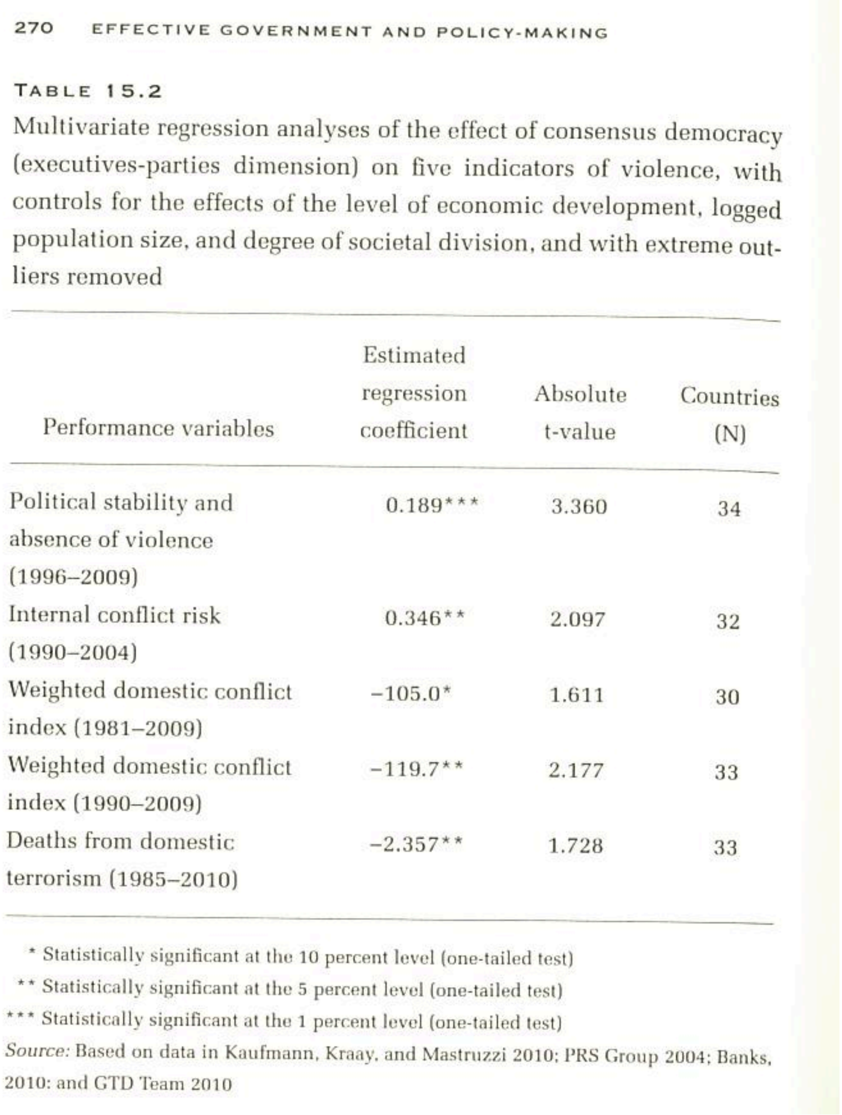
Illustration:

The four densities at right are sampling distributions for a regression coefficient, where the true coefficient in the population is 0.5.

	Solid line	***Dashed line***
Blue	*Large sample, low noise*	*Large sample, high noise*
Red	*Small sample, low noise*	*Small sample, high noise*



Wait. What is missing?



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)

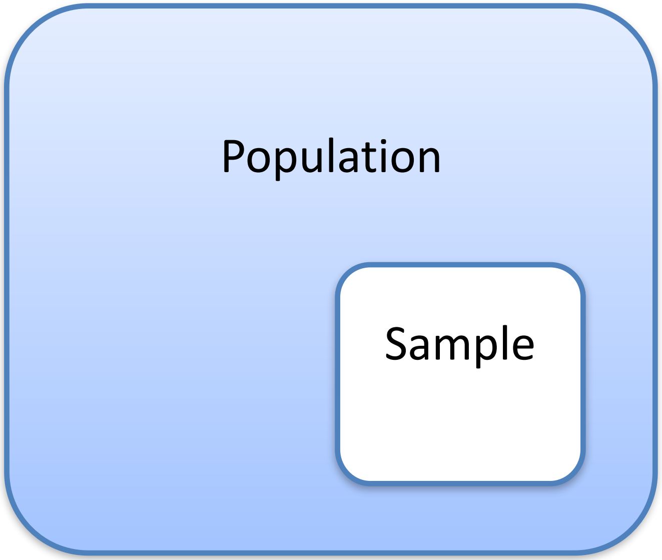
** Statistically significant at the 5 percent level (one-tailed test)

*** 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

Lijphart studies "the 36 countries (with pop. > 250k) that were democratic in the middle of 2010 and that had been continuously democratic since 1989 or earlier."

Why are there stars, t-values, standard errors, hypothesis tests etc.?



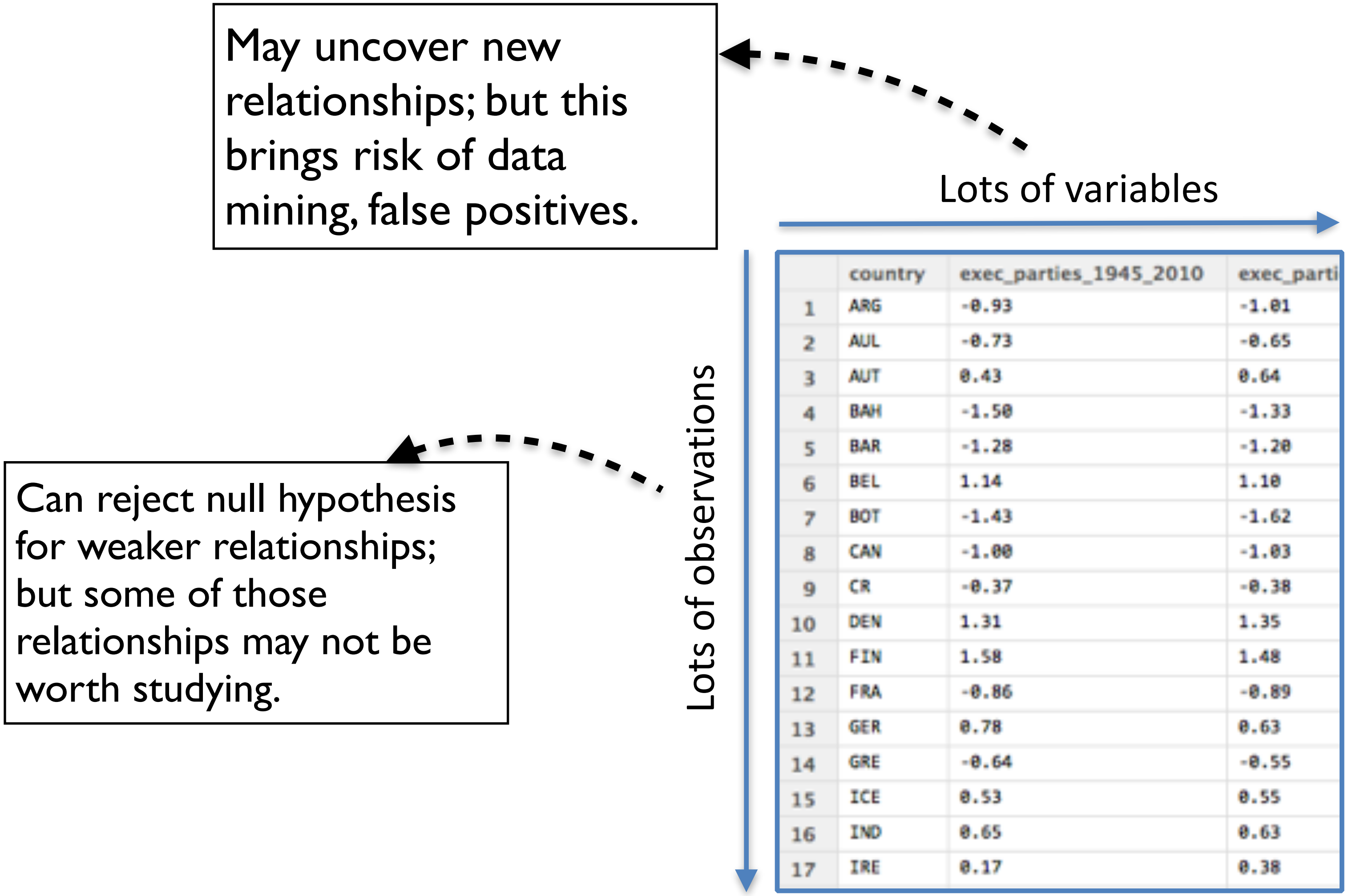
Three views on whether there should be stars in Lijphart (and most research in comparative politics and international relations)

1. (Rare) "No. You have the population. No need for statistical inference."
2. (Common) "Yes, because this **is** population inference."
    1. Technical version: "We are making inferences about a **super-population** (or **data generating process**) from which these 36 countries were sampled."
    2. Non-technical version: "The standard errors and hypothesis tests tell us whether the results we see might be merely a fluke."
3. (Rare, new) "Yes, because you are making causal claims. The **potential outcomes** are missing."

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

Is Big Data good for social science?

The risks depend on which part of the dataset gets “bigger”:



May uncover new relationships; but this brings risk of data mining, false positives.

Lots of variables

Can reject null hypothesis for weaker relationships; but some of those relationships may not be worth studying.

Lots of observations

	country	exec_parties_1945_2010	exec_parti
1	ARG	-0.93	-1.01
2	AUL	-0.73	-0.65
3	AUT	0.43	0.64
4	BAH	-1.50	-1.33
5	BAR	-1.28	-1.20
6	BEL	1.14	1.10
7	BOT	-1.43	-1.62
8	CAN	-1.00	-1.03
9	CR	-0.37	-0.38
10	DEN	1.31	1.35
11	FIN	1.58	1.48
12	FRA	-0.86	-0.89
13	GER	0.78	0.63
14	GRE	-0.64	-0.55
15	ICE	0.53	0.55
16	IND	0.65	0.63
17	IRE	0.17	0.38

Two announcements:

- Labs this week: regression commands useful for essay
- First week of TT: afternoon drop-in session in OQC lab (email to be sent out with details)

Thanks for your attention
and efforts this term!