Statistical Inference

Two simple examples

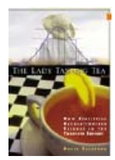
- Lady tasting tea
- Human energy fields

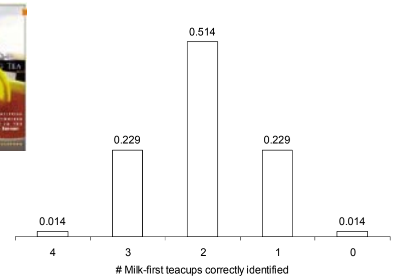
- These examples provide the intuition behind statistical inference

Fisher's exact test

- A simple approach to inference
- Only applicable when outcome probabilities known
- Lady tasting tea example
  - Claims she can tell whether the milk was poured first
  - In a test, 4/8 teacups had milk poured first
  - The lady correctly detects all four
  - Should we believe that she has milk-first detection ability?
- To answer this question, we ask, "What is the probability she did this by chance?"
  - If likely to happen by chance, then we shouldn't be convinced
  - If very unlikely, then we should maybe believe her
  - This is the basic question behind statistical inference
    - Null hypothesis
    - People seem poorly equipped to make these inferences, in part because they forget about failures, but notice success: e.g. Dog ESP, miracles
    - Other examples: fingerprints, DNA, HIV tests, regression coefficients, mean differences, etc.
  - Answer?
    - 70 ways of choosing four cups out of eight
    - How many ways can she do so correctly?

Lady tasting tea: Prob. of identifying by chance





By chance, she would only guess all four correctly with probability (1/70 =) 0.014. So, we can be quite confident in her milk-first detection ability.

Second simple example
Healing touch: human energy field detection

"**A Close Look at Therapeutic Touch**"

Linda Rosa; Emily Rosa; Larry Sarner; Stephen Barrett. 1998.

JAMA

(279: 1005 – 1010)

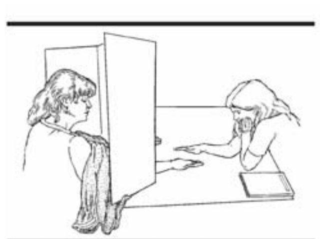


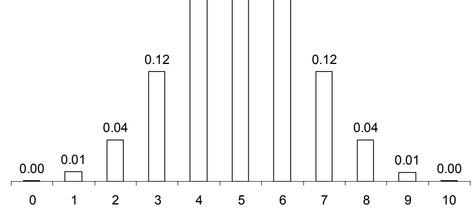
Figure 1.—Experimenter hovers hand over one of subject's hands. Draped towel prevents peeking. Drawing by Pat Linse, Skeptics Society

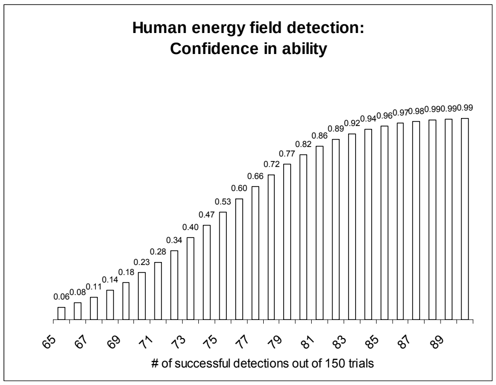
Human energy field: Prob. of success by chance

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

n (the number of trials) = 10
k (number of successes)
p (prob. of success for the null model) = .5

Binomial distribution





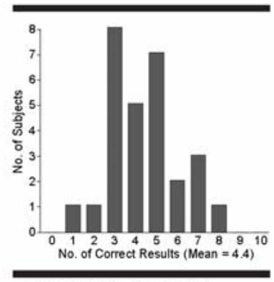


Figure 2.—Distribution of test results.

Table 2.—Statistical Analysis

Statistical Function	Initial Test (n = 15)	Follow-up Test (n = 13)
Mean (95% confidence interval)	4.67 (3.67-5.67)	4.08 (3.17-4.99)
SD	1.74	1.44
α (1-tailed test)	.05	.05
t statistic	−0.7174	−2.222
Upper critical limit of Student *t* distribution	1.761	1.782
Alternative hypothesis, μ = 6.67	0.9559	0.9801
Alternative hypothesis, μ = 7.50	0.999644	0.999953

Linda Rosa; Emily Rosa; Larry Sarner; Stephen Barrett. 1998. "A Close Look at Therapeutic Touch" JAMA, 279: 1005 - 1010.

Null hypothesis

- In both cases, we calculated the probability of making the correct choice by chance and compared it to the observed results.
- Thus, our null hypothesis was that the lady and the therapists lacked any of their claimed ability.
- What's the null hypothesis that Stata uses by default for calculating p values?
- Always consider whether null hypotheses other than 0 might be more substantively meaningful.
  - E.g., testing whether the benefits from government programs outweigh the costs.

Assessing uncertainty

- With more complicated statistical processes, larger samples, continuous variables, Fisher's exact test becomes difficult or impossible
- Instead, we use other approaches, such as calculating standard errors and using them to calculate confidence intervals
- The intuition from these simple examples, however, extends to the more complicated one

Standard error: Baseball example

- In 2006, Manny Ramírez hit .321
- How certain are we that, in 2006, he was a .321 hitter? Confidence interval?
- To answer this question, we need to know how precisely we have estimated his batting average
- The standard error gives us this information, which in general is (where *s* is the sample standard deviation)
- Equation?

$$\text{std. err.} = \frac{s}{\sqrt{n}}$$

Baseball example

- The standard error (s.e.) for proportions (percentages/100) is?

$$\sqrt{\frac{p(1-p)}{n}}$$

- For n = 400, p = .321, s.e. = .023
- Which means, on average, the .321 estimate will be off by .023
  - His 95% confidence interval on his batting average ranges from
    - 298 to 344

Baseball example: postseason

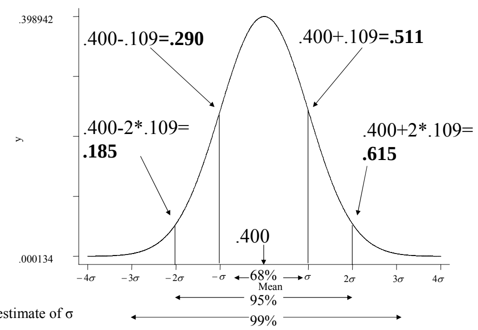
- 20 at-bats
  - N = 20, p = .400, s.e. = .109
  - Which means, on average, the .400 estimate will be off by .109
- 10 at-bats
  - N = 10, p = .400, s.e. = .159
  - Which means, on average, the .400 estimate will be off by .159

Using Standard Errors, we can construct "confidence intervals"

- **Confidence interval (ci)**: an interval between two numbers, where there is a certain specified level of confidence that a population parameter lies
- ci = sample parameter ± multiple * sample standard error

N = 20; avg. = .400; s = .489; s.e. =.109

Confidence interval



s.e. is estimate of σ

- Much of the time, we fail to appreciate the uncertainty in averages and other statistical estimates
  - Postseason statistics
  - Boardgames
  - Research
  - Life

Two types of inference

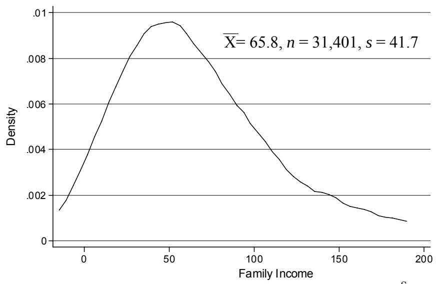
- Testing underlying traits
  - E.g., can lady detect milk-poured first?
  - E.g., does democracy improve human lives?
- Testing inferences about a population from a sample
  - What percentage of the population approves of President Bush?
  - What's average household income in the United States?

Example of second type of inference:

Testing inferences about a population from a sample

Family income in 2006

Certainty about mean of a population based on a sample: Family income in 2006



Source: 2006 CCES

Calculating the Standard Error on the mean family income of $65.8 thousand dollars

Equation?

$$\text{std. err.} = \frac{s}{\sqrt{n}}$$

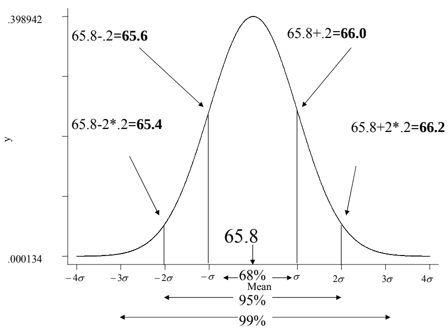
For the income example,

std. err. = 41.6/177.2 = $0.23 thousands of dollars

$\bar{X}$ = 65.8, n = 31401, s = 41.7

N = 31,401; avg. = 65.8; s = 41.6; s.e. = s/√n = .2

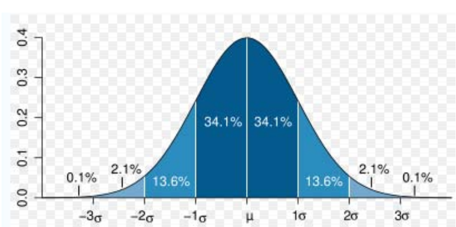
The Picture



Where does the bell-shaped curve come from?

That is, how do we know that two ± standard errors covers 95% of the distribution?

Could this possibly be right? Why?



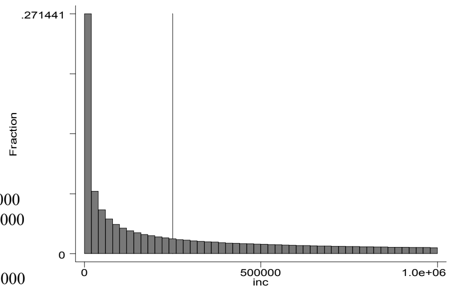
- **Central limit theorem**

Central Limit Theorem

As the sample size *n* increases, the distribution of the mean $\bar{X}$ of a random sample taken from **practically any population** approaches a *normal* distribution, with mean μ and standard deviation $\sigma/\sqrt{n}$

Illustration of Central Limit Theorem:
Exponential Distribution

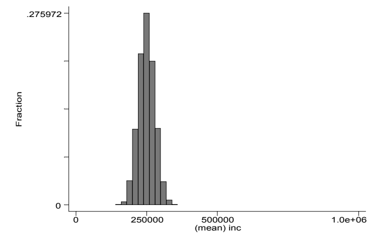
Mean = 250,000
Median=125,000
σ = 283,474
Min = 0
Max = 1,000,000



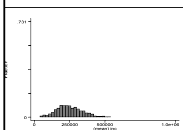
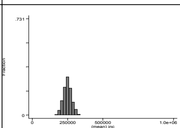
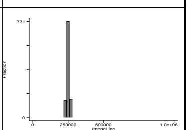
Consider 10,000 samples of n = 100

N = 10,000
Mean = 249,993
s = 28,559

What will the distribution of these means look like?



Consider 1,000 samples of various sizes

10	100	1000
		
Mean =250,105 s = 90,891	Mean = 250,498 s = 28,297	Mean = 249,938 s = 9,376

Convince yourself by playing with simulations

- http://onlinestatbook.com/stat_sim/sampling_dist/index.html
- http://www.kuleuven.ac.be/ucs/java/index.htm

Most important standard errors

Mean	$\frac{s}{\sqrt{n}}$
Proportion	$\sqrt{\frac{p(1-p)}{n}}$
Diff. of 2 means	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
Diff. of 2 proportions	$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$
Diff of 2 means (paired data)	$\frac{s_d}{\sqrt{n}}$
Regression (slope) coeff.	$\frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x}$

In small samples (n <30), these statistics are not normally distributed. Instead, they follow the t-distribution.

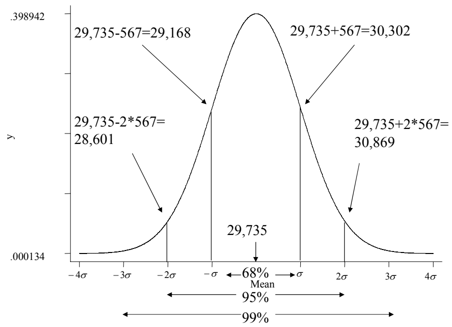
We'll discuss that complication next class.

Another example

- Let's say we draw a sample of tuitions from 15 private universities. Can we estimate what the average of all private university tuitions is?
- N = 15
- Average = $29,735
- s = 2,196
- s.e. = $\frac{s}{\sqrt{n}} = \frac{2{,}196}{\sqrt{15}} = 567$

N = 15; avg. = 29,735; s = 2,196; s.e. = s/√n = 567

The Picture



Confidence Intervals for Tuition Example

- 68% confidence interval
  - = \$29,735 ± 567
  - = [\$29,168 to \$30,302]
- 95% confidence interval
  - = \$29,735 ± 2*567
  - =[\$28,601 to \$30,869]
- 99% confidence interval
  - = \$29,735 ± 3*567
  - = [\$28,034 to \$31,436]

Using z-scores

The *z*-score or the "standardized score"

Equation? $$z = \frac{x - \bar{x}}{\sigma_x}$$

Using z-scores to assess how far values are from the mean

What if someone (ahead of time) had said, "I think the average tuition of private universities is \$25k"?

- Note that \$25,000 is well out of the 99% confidence interval, [28,034 to 31,436]
- Q: How far away is the \$25k estimate from the sample mean?
  - A: Do it in *z*-scores: (29,735-25,000)/567
    - = 8.35

More confidence interval calculations

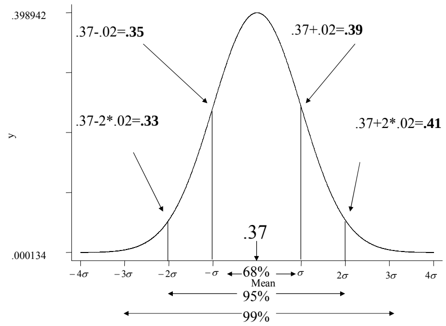
Proportions
Difference in means
Difference in proportions

Constructing confidence intervals of <u>proportions</u>

- Let us say we drew a sample of 1,000 adults and asked them if they approved of the way George Bush was handling his job as president. (March 13-16, 2006 Gallup Poll) Can we estimate the % of all American adults who approve?
- N = 1000
- p = .37
- s.e. = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.37(1-.37)}{1000}} = 0.02$

N = 1,000; p. = .37; s.e. = √p(1-p)/n = .02

The Picture



Confidence Intervals for Bush approval example

- 68% confidence interval = .37±.02 = [.35 to .39]
- 95% confidence interval = .37±2*.02 = [.33 to .41]
- 99% confidence interval = .37±3*.02 = [.31 to .43]

What if someone (ahead of time) had said, "I think Americans are equally divided in how they think about Bush."

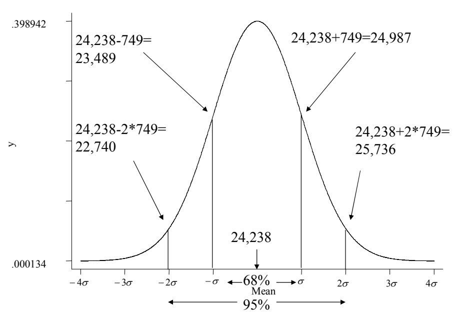
- Note that 50% is well out of the 99% confidence interval, [31% to 43%]
- Q: How far away is the 50% estimate from the sample proportion?
  - A: Do it in *z*-scores: (.37-.5)/.02 = -6.5

Constructing confidence intervals of <u>differences of means</u>

- Let's say we draw a sample of tuitions from 15 private and public universities. Can we estimate what the difference in average tuitions is between the two types of universities?
- N = 15 in both cases
- Average = 29,735 (private); 5,498 (public); diff = 24,238
- s = 2,196 (private); 1,894 (public)
- s.e. = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{4{,}822{,}416}{15} + \frac{3{,}587{,}236}{15}} = 749$

N = 15 twice; diff = 24,238; s.e. = 749

The Picture



Confidence Intervals for difference of tuition means example

- 68% confidence interval = 24,238±749 =

[23,489 to 24,987]

- 95% confidence interval = 24,238±2*749 = [22,740 to 25,736]
- 99% confidence interval =24,238±3*749 =
- [21,991 to 26,485]

What if someone (ahead of time) had said, “Private universities are no more expensive than public universities”

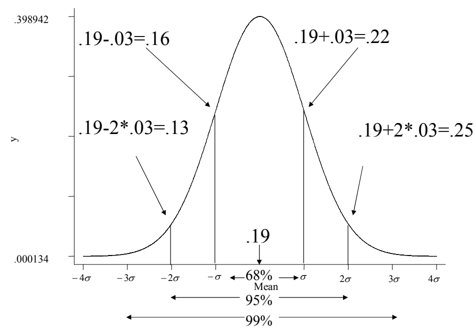
- Note that $0 is well out of the 99% confidence interval, [$21,991 to $26,485]
- Q: How far away is the $0 estimate from the sample proportion?
  - A: Do it in *z*-scores: (24,238-0)/749 = 32.4

Constructing confidence intervals of <u>difference of proportions</u>

- Let us say we drew a sample of 1,000 adults and asked them if they approved of the way George Bush was handling his job as president. (March 13-16, 2006 Gallup Poll). We focus on the 600 who are either independents or Democrats. Can we estimate whether independents and Democrats view Bush differently?
- N = 300 ind; 300 Dem.
- p = .29 (ind.); .10 (Dem.); diff = .19
- s.e. = $\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{.29(1-.29)}{300}+\frac{.10(1-.10)}{300}} = .03$

diff. p. = .19; s.e. = .03

The Picture



Confidence Intervals for Bush Ind/Dem approval example

- 68% confidence interval = .19±.03 =

[.16 to .22]

- 95% confidence interval = .19±2*.03 =

[.13 to .25]

- 99% confidence interval = .19±3*.03 =

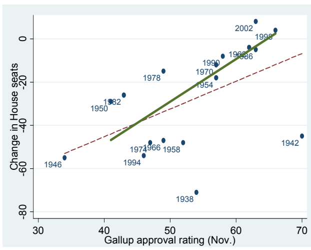
[.10 to .28]

What if someone (ahead of time) had said, “I think Democrats and Independents are equally unsupportive of Bush”?

- Note that 0% is well out of the 99% confidence interval, [10% to 28%]
- Q: How far away is the 0% estimate from the sample proportion?
  - A: Do it in *z*-scores: (.19-0)/.03 = 6.33

Constructing confidence intervals for regression coefficients

- Let's look at the relationship between the mid-term seat loss by the President's party at midterm and the President's Gallup poll rating



Slope = 1.97
N = 14
s.e.r. = 13.8
s_x = 8.14

$$\text{s.e.}_{\text{slope}} = \frac{s.e.r.}{\sqrt{n-1}} \times \frac{1}{s_x} = \frac{13.8}{\sqrt{13}} \times \frac{1}{8.14} = 0.47$$

The Stata output

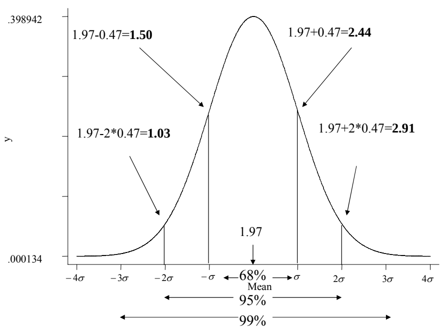
```
. reg loss gallup if year>1948

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  1,    12) =   17.53
       Model |  3332.58872     1  3332.58872           Prob > F      =  0.0013
    Residual |  2280.83985    12  190.069988           R-squared     =  0.5937
-------------+------------------------------           Adj R-squared =  0.5598
       Total |  5613.42857    13  431.802198           Root MSE      =  13.787

------------------------------------------------------------------------------
        loss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gallup |    1.96812   .4700211     4.19   0.001     .9440315    2.992208
       _cons |  -127.4281   25.54753    -4.99   0.000    -183.0914   -71.76486
------------------------------------------------------------------------------
```


The Picture

N = 14; slope=1.97; s.e. = 0.47



Confidence Intervals for regression example

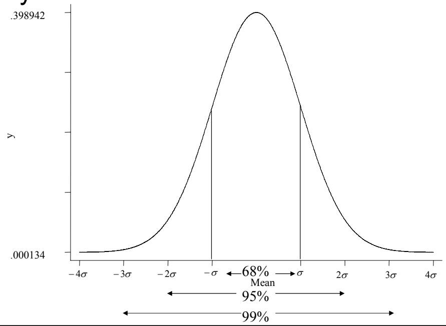
- 68% confidence interval = 1.97± 0.47= [1.50 to 2.44]
- 95% confidence interval = 1.97± 2*0.47 = [1.03 to 2.91]
- 99% confidence interval = 1.97±3*0.47 = [0.62 to 3.32]

What if someone (ahead of time) had said, "There is no relationship between the president's popularity and how his party's House members do at midterm"?

- Note that 0 is well out of the 99% confidence interval, [0.62 to 3.32]
- Q: How far away is the 0 estimate from the sample proportion?
  - A: Do it in *z*-scores: (1.97-0)/0.47 = 4.19

z vs. *t*

If *n* is sufficiently large, we know the distribution of sample means/coeffs. will obey the normal curve



- When the sample size is large (i.e., > 150), convert the difference into *z* units and consult a *z* table

$Z = (H_1 - H_0) / \text{s.e.}$

Reading a *z* table

Regression example

$Z = (H_1 - H_{null}) / \text{s.e.}$

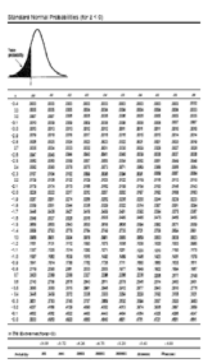
Large sample (n = 1000)
Slope (b) = 2.1
s.e. = 0.9

Calculate p-value for one-tailed test $H_{null} = 0$

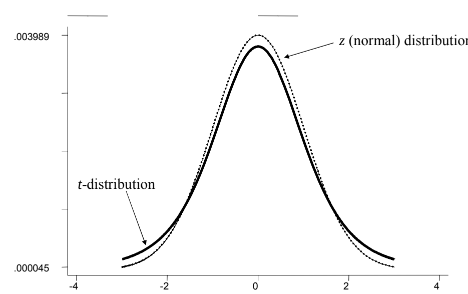
Z = (2.1 – 0)/0.9
Z =2.3
p-value (using handout)
Pr(Z >2.3) < 0.5 -.4893
Pr(Z >2.3) < 0.011

Interpretation: probability that we would observe a coefficient of 2.1 by chance is less than 0.011.

For two-tailed test: Pr(|Z| >2.3) <1 – 2*.4893
(calculations differ by table)



t
(when the sample is small)



- When the sample size is small (i.e., <150), convert the difference into *t* units and consult a *t* table

$t = (H_1 - H_{null}) / \text{s.e.}$

Mid-term seat loss example

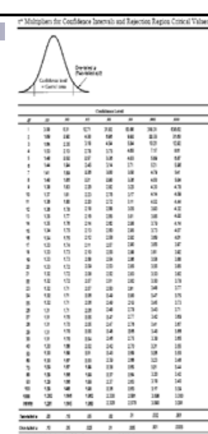
Slope = 1.97
$\text{s.e.}_{slope} = 0.47$

What's H_1?

What's H_{null}?

$t = (H_1 - H_{null}) / \text{s.e.}$
$t = (1.97 - 0) / 0.47$
$t = 4.19$

Reading a *t* table



Testing hypotheses in Stata with `ttest`

What if someone (ahead of time) said, "Private university tuitions did not grow from 2003 to 2004"

- Mean growth = $1,632
- Standard deviation on growth = 229
- Note that $0 is well out of the 95% confidence interval, [$1,141 to $2,122]
- Q: How far away is the $0 estimate from the sample proportion?
  - A: Do it in *z*-scores: (1,632-0)/229 = 7.13

The Stata output

```
. gen difftuition=tuition2004-tuition2003
. ttest diff = 0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
difftu~n |      15      1631.6    228.6886     885.707    1141.112    2122.088
------------------------------------------------------------------------------
    mean = mean(difftuition)                                      t =   7.1346
Ho: mean = 0                                     degrees of freedom =       14

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

You could test difference in means with

```
ttest tuition2004 = tuition2003
```


A word about standard errors and collinearity

- The problem: if X_1 and X_2 are highly correlated, then it will be difficult to precisely estimate the effect of either one of these variables on *Y*

How does having another *collinear* independent variable affect standard errors?

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{1}{N-n-1}\frac{S_Y^2}{S_{X_1}^2}\frac{1-R_Y^2}{1-R_{X_1}^2}}$$

R^2 of the "auxiliary regression" of X_1 on all the other independent variables

Example: Effect of party, ideology, and religiosity on feelings toward Bush

	Bush Feelings	Conserv.	Repub.	Religious
Bush Feelings	1.0	.39	.57	.16
Conserv.		1.0	.46	.18
Repub.			1.0	.06
Relig.				1.0

Regression table

	(1)	(2)	(3)	(4)
Intercept	32.7 (0.85)	32.9 (1.08)	32.6 (1.20)	29.3 (1.31)
Repub.	6.73 (0.244)	5.86 (0.27)	6.64 (0.241)	5.88 (0.27)
Conserv.	---	2.11 (0.30)	---	1.87 (0.30)
Relig.	---	---	7.92 (1.18)	5.78 (1.19)
N	1575	1575	1575	1575
R^2	.32	.35	.35	.36

Pathologies of statistical significance

Understanding and using "significance"

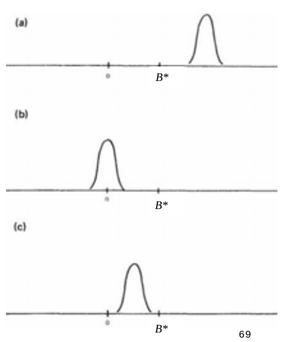
Substantive versus statistical significance

- Which variable is more statistically significant?
- X_1
- Which variable is more important?
- X_2
- Importance (size) is often more relevant

	(1)	(2)
Intercept	0.002 (0.005)	0.003 (0.008)
X_1	0.500* (0.244)	0.055** (0.001)
X_2	0.600 (0.305)	0.600 (0.305)
N	1000	1000
R^2	.32	.20
*p<.05, **p <.01		

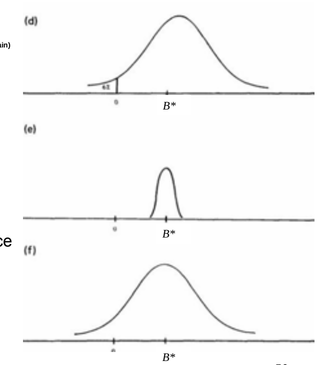
Substantive versus statistical significance (again)

- Think about point estimates, such as means or regression coefficients, as the center of distributions
- Let B^* be of value of a regression coefficient that is large enough for substantive significance
- Which is substantively significant?
- (a)



Substantive versus statistical significance (again)

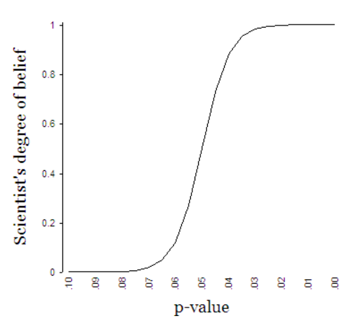
- Which is more substantively significant? That is, which is larger?
  - Depends, but probably (d)
- Don't confuse lack of statistical significance with no effect
  - Lack of statistical significance usually implies uncertainty, not no effect



Degree of significance

- We often use 95% confidence intervals, which correspond with p<.05
- Is an effect statistically significant if it is p<.06? (that is, 95% CI encompasses zero)
  - Yes!
  - For many data sets, anything less than p<.20 is informative
  - Treat significance as a continuous variable
    - E.g., if p<.20, we should be roughly 80% sure that the coefficient is different from zero. If p<.10, we should be roughly 90% sure that the coefficient is different from zero. Etc.

Don't make this mistake



Understanding and using "significance"

Summary

- Focus on substantive significance (effect size), not statistical significance
- Focus on degree of uncertainty, not on the arbitrary cutoff of p =.05
  - Confidence intervals are preferable to p-values
  - Treat p-values as a continuous variables
- Don't confuse lack of statistical significance with no effect (that is, p >.05 does not mean b = 0)
  - Lack of statistical significance usually implies uncertainty, <u>not no effect!</u>

What to present

- Standard error
- CI
- t-value
- p-value
- Stars
- Combinations?
- Different disciplines have different norms, I prefer
  - Graphically presenting CIs
  - Coefficients with standard errors in parentheses
  - No stars
  - (Showing data through scatter plots more important)

TABLE 1. Explaining Democratic Lower House Seat Shares, 1946–90: Is the Nation Homogeneous?

	Main Effects		Southern Interactions		Border Interactions	
Variable	Coefficient	*t*-value	Coefficient	*t*-value	Coefficient	*t*-value
South	242.120	2.14	—		—	
Border	262.120	1.32	—		—	
Year	.032	.91	−.151	−2.28*	−.142	−1.31
Democrats (t − 1)	.508	20.10**	.435	7.70**	−.067	−.52
Compensation	.007	2.08*	−.005	−1.03	−.007	−.59
Presidential year	−27.079	−11.10**	22.778	6.19**	11.520	1.35
Presidential vote	.498	9.40**	−.452	5.85**	−.234	−1.39
Gubernatorial year	−.18.540	−7.86**	15.111	4.08**	−.440	−.04
Gubernatorial vote	.364	8.10**	−.328	−5.52**	−.044	−.20
Off year	−10.042	−10.62**	6.800	3.80**	2.818	.82
GNP growth	.149	2.74**	−.157	−1.49	−.148	−.79

Note: Thirty-one state intercepts not shown. Coefficients are unstandardized.
n = 1,035, adjusted R^2 = .89, *SEE* = 7.94
*p < .05; **p < .01

Statistical monkey business

(tricks to get p <.05)

- Bonferroni problem: using p <.05, one will get significant results about 5% (1/20) of the time by chance alone
- Reporting one of many dependent variables or dependent variable scales
  - Football mascots examples
  - Healing-with-prayer studies
  - Psychology lab studies
- Repeating an experiment until, by chance, the result is significant
  - Drug trials
  - Called file-drawer problem

Statistical monkey business

(tricks to get p <.05)

- Specification searches
  - Adding and removing control variables until, by chance, the result is significant
  - Exceedingly common

Statistical monkey business

Solutions

- With many dependent variables, test hypotheses on a simple unweighted average
- Bonferroni correction
  - If testing *n* independent hypotheses, adjusts the significance level by 1/*n* times what it would be if only one hypothesis were tested
  - E.g., testing 5 hypotheses at p < .05 level, adjust significance level to p/5 < .05/5 <.01
- Show bivariate results
- Show many specifications
- Model averaging
- Always be suspicious of statistical monkey business!