Thinking Straight Friday , May 9

Morning Session

- Review of Assignment and Sampling
- Lecture/discussion on correlation and causation

Afternoon Session beginning at 1 pm

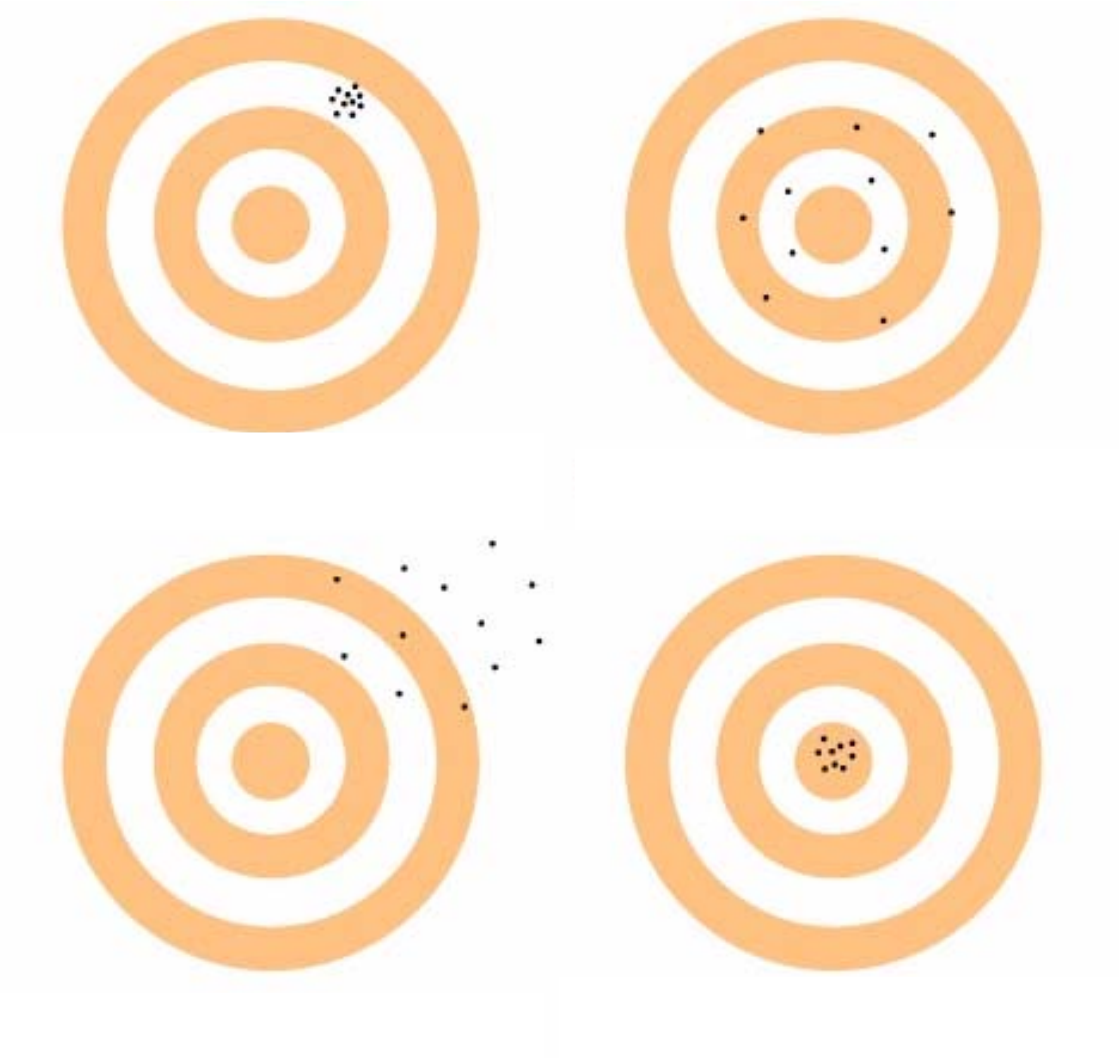
- **Continuation** of Lecture/discussion on correlation and causation
- Workshop on Theories

Be Sure to pick up handout on Virtue Ethics to read along with Rachels Ch. 12 for next Tuesday, May 13

Review Sampling Terminology

- Parameter
  - fixed, unknown number that describes the population
- Statistic
  - known value calculated from a sample
  - a statistic is used to estimate a parameter
- **Bias**
  - **in repeated samples, the sample statistic consistently misses the population parameter in the same direction**
- **Variability**
  - **different samples from the same population may yield different values of the sample statistic**

Large or small Bias? Large or small Variability?



Sampling Strategy

- To reduce bias, use random sampling
- To reduce variability, use larger samples
  - estimates from random samples will be closer to the true values in the population if the samples are larger
  - how close will they be?
    - margin of error

Margin of Error

- The amount by which the proportion obtained from the sample ($\hat{p}$) will differ from the true population proportion (p) rarely exceeds the *margin of error*.

- Typical margin of error: 1/sqrt(n)
  - In 95% of surveys, the sample proportion will not differ from the population proportion by any more than the margin of error. ("*95% confidence*")

Weekend Poll discussed Tuesday

Barack Obama	48%
Hillary Clinton	40%
Unsure	8%
Other	5%

Zogby Poll
Date: 5/3-4
North Carolina
Added: 5/5/08
Est. MoE = 3.9% [?]

Zogby comments

Barack Obama	51%
Hillary Clinton	37%
Unsure	12%
Other	

56%

43 %

with 12% allotted equally

55.99

41.74

Actual Vote

Zogby Poll
Date: 5/4-5
North Carolina
Added: 5/6/08
Est. MoE = 3.9% [?]
Sample 643 likely primary voters

Obama 50.1 - 57.9
Clinton 36.1 - 43.9

Last Poll before primary

PRESIDENTIAL PREFERENCE - DEM (Vote For 1) Vote Type Summary Contest Detail Map

100 of 100 Counties Reporting

	Percent	Votes
Hillary Clinton (DEM)	41.74%	652,824
Mike Gravel (DEM)	0.79%	12,409
Barack Obama (DEM)	55.99%	875,683
No Preference (DEM)	1.47%	23,042
		1,563,958

Survey Sample Size	Margin of Error Percent* % ±	Margin of Error Proportion* prop ±
100000	0.3	0.003
20000	1	0.007
10000	1	0.010
2000	2	0.022
1500	3	0.026
1000	3	0.032
900	3	0.033
800	3	0.035
700	4	0.038
600	4	0.041
500	4	0.045
400	5	0.050
300	6	0.058
200	7	0.071
100	10	0.100
50	14	0.141

*Assumes a 95% level of confidence

Size Matters in Sampling

Bigger Sample, less variability, narrower the margin of error

n=643

Sample Size	of Error Percent* % ±	of Error Percent** % ±	of Error Percent*** % ±
100000	0.3	0.4	0.3
20000	1	1	1
10000	1	1	1
2000	2	3	2
1500	3	3	2
1000	3	4	3
900	3	4	3
800	3	5	3
700	4	5	3
600	4	5	3
500	4	6	4
400	5	6	4
300	6	7	5
200	7	9	6
100	10	13	8
50	14	18	12
	*Assumes a 95% level of confidence	**Assumes a 99% level of confidence	***Assumes a 90% level of confidence

Holding size constant

The lower the confidence you can live with, the narrower the margin of error. For example, at 50% confidence MOE with sample size 600 is **± 1.4**

For those who have studied some statistics before

The margin of error reported with poll results is what is considered the 95% confidence level range. Meaning a pollster has a 95% confidence that the true measurement lies within the margin of error.

The standard error equation is shown below.

Standard error = $\sqrt{\frac{p*(1-p)}{n}}$

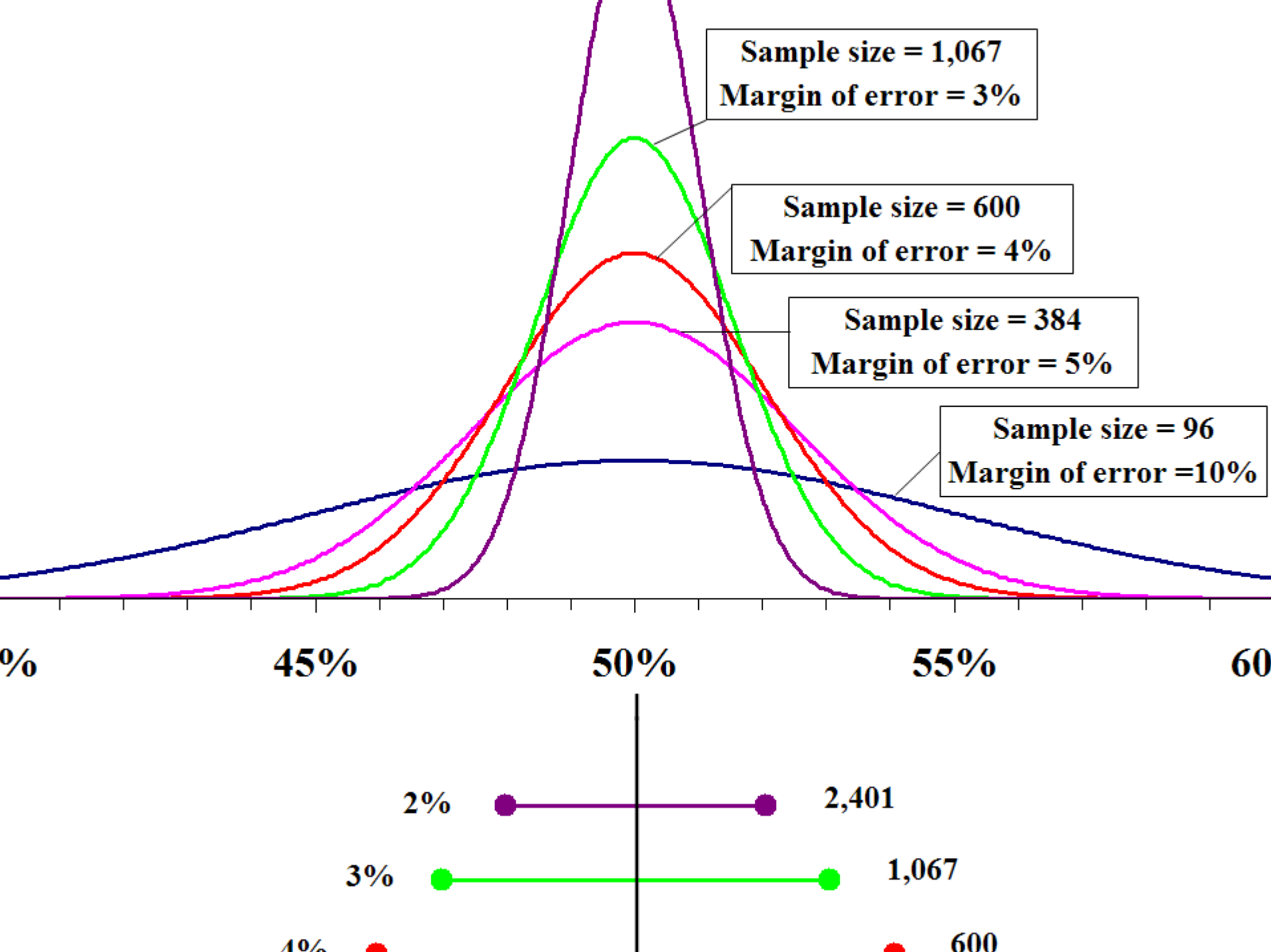
where p represent the support level of the poll and n is the number of voters polled.

And the 95% confidence interval is 1.96 * (standard error).

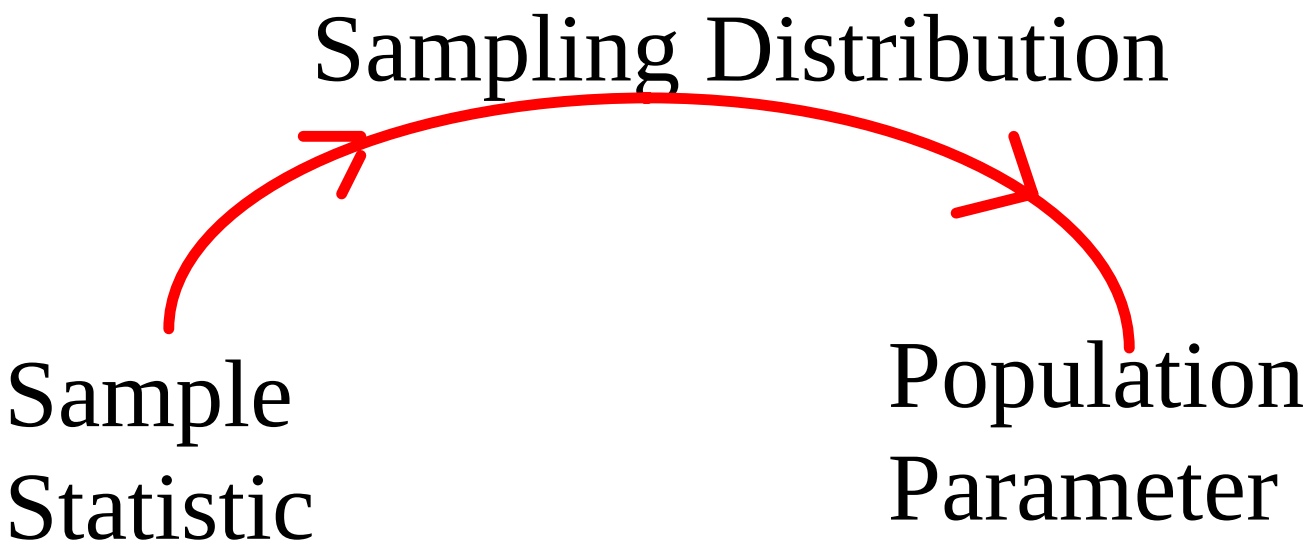
The maximum margin of error occurs when p = 50%.

(Maximum) margin of error (95%) = $1.96*\sqrt{\frac{.5*(1-.5)}{n}} = \frac{.98}{\sqrt{n}} \approx \frac{1}{\sqrt{n}}$

Margin of error at 99% confidence ≈ **1.29/√n**
Margin of error at 95% confidence ≈ **.98/√n**
Margin of error at 90% confidence ≈ **.82/√n**
Margin of error at 50% confidence ≈ **.35/√n**


Sample size = 1,067
Margin of error = 3%
Sample size = 600
Margin of error = 4%
Sample size = 384
Margin of error = 5%
Sample size = 96
Margin of error =10%
%
45%
50%
55%
60
2%
2,401
3%
1,067
600

Inferential Statistics uses the sampling distribution to bridge the gap between sample and population



End of Tuesday Lecture

Example

Satisfaction with Way Things Are Going in U.S.

MarketSearch SC Poll, September 1996

46% are satisfied

up 12 points from February and representing a record high

47% are dissatisfied

a record low

How the Poll was Conducted

The MarketSearch Poll of South Carolina is a semi-annual telephone survey of 800 consumers statewide. The Poll from which these findings are taken was conducted in late August and early September, 1996. The survey has a *random sampling error of ±3.5 percent*.

$$\text{random sampling error} = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{800}} = \frac{1}{28.3} = .035$$

Margin of error at 99% confidence ≈ **1.29/√n**
Margin of error at 95% confidence ≈ **.98/√n**
Margin of error at 90% confidence ≈ **.82/√n**

Case Study

Conclusion (Confidence statement)

For the proportion of the population who were satisfied, the sample proportion was $\hat{p} = .46$ (46%) and the margin of error was ±.035 (3.5%). We can then say that *"we are 95% confident that the proportion of the population who were satisfied was between .425 and .495 (42.5% and 49.5%)."*

Errors in Sample Surveys

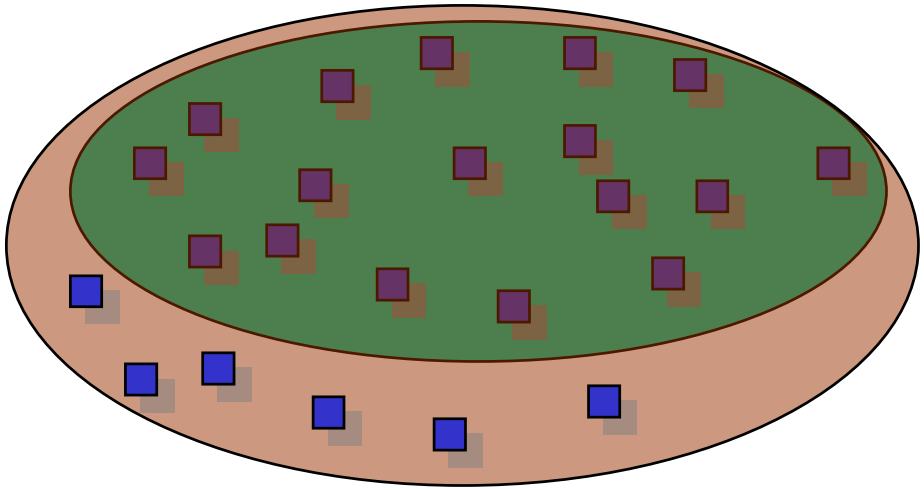
- Mistakes in Sampling
- Random sampling errors
  - measured by *margin of error*
- Nonsampling errors

Sampling Errors

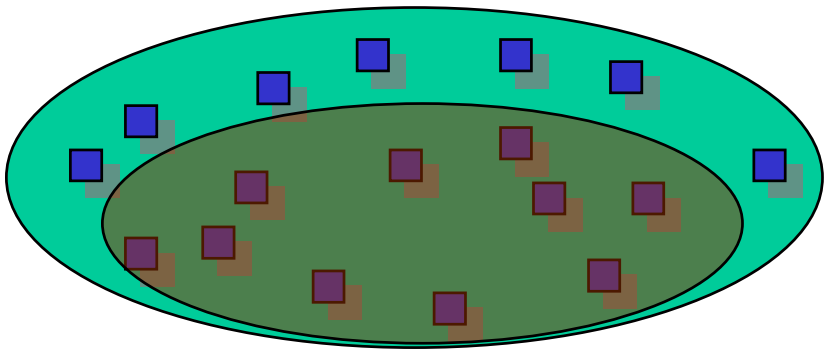
- Difficulties
  - Using the wrong sampling frame (next slide)

- Disasters
  - Using voluntary response (volunteer sample)
  - Using a convenience or haphazard sample
    - cannot extend results to the population of interest (need a broad cross-section of the population)

Using the Wrong Sampling Frame



Including some units not in the population.



Undercoverage: Excluding some units in the population.

Nonsampling Errors

- Difficulties
  - Processing errors (data entry, calculations)
  - Wording of questions / Response error
- Disasters
  - Nonresponse (cannot contact subjects or they do not respond)

An Common Source of Nonsampling error: The Pitfalls of Asking Survey Questions

- Deliberate bias
- Unintentional bias
- Desire to please
- Asking the uninformed
- Unnecessary complexity
- Ordering of questions
- Confidentiality and anonymity

Deliberate Bias

- “If you found a wallet with $20 in it, would you return the money?”
- “If you found a wallet with $20 in it, would you do the right thing and return the money?”

Unintentional Bias

- “I have taught several students over the past few years.”
  - How many students do you think I have taught?
  - How many years am I referring to?
- “Over the past few days, how many servings of fruit have you eaten?”
  - How many days are you considering?
  - What constitutes a serving?

Desire to Please

- “Is your instructor doing a good job presenting the course material in a clear and interesting way?”
    - Yes
    - No

Asking the Uninformed

A Case Study

Washington Post National Weekly Edition (April 10-16, 1995, p. 36)

- A 1978 poll done in Cincinnati asked people whether they "favored or opposed repealing the 1975 Public Affairs Act."
  - There was no such act!
  - About one third of those asked expressed an opinion about it.

Unnecessary Complexity

- “Do you sometimes find that you have arguments with your family members and co-workers?”
  - Arguments with family members
  - Arguments with co-workers

Ordering of Questions

- “How often do you normally go out on a date? about ___ times a month.”
- “How happy are you with life in general.”
  - Strong association between these questions.
  - If the ordering is reversed, then there would be no strong association between these questions

Confidentiality and Anonymity

- Confidential answer
  - respondent is known, but the information is a secret
- Anonymous answer
  - the respondent is not known, or cannot be linked to his/her response

Criticism of Sampling Arguments

√ 1. Attacking the evidence. Is the evidence cited in the premise true or can the data be disputed

√ 2. Questioning the representativeness of the sample.
   (a) Size of Sample
   (b) Sample Selection

3. Pointing to a shift in the unit of analysis

4. Challenging the truth of the conclusion.

Shifting the Unit of Analysis

Example 8.9:

(1) Most **courses** sampled at the university give exams

(likely) Most **teachers** in the university give exams

Example 8.10

(1) 20 percent of **schools** sampled across the United States fail to meet the Average Yearly requirement of the No Child Left Behind Act.

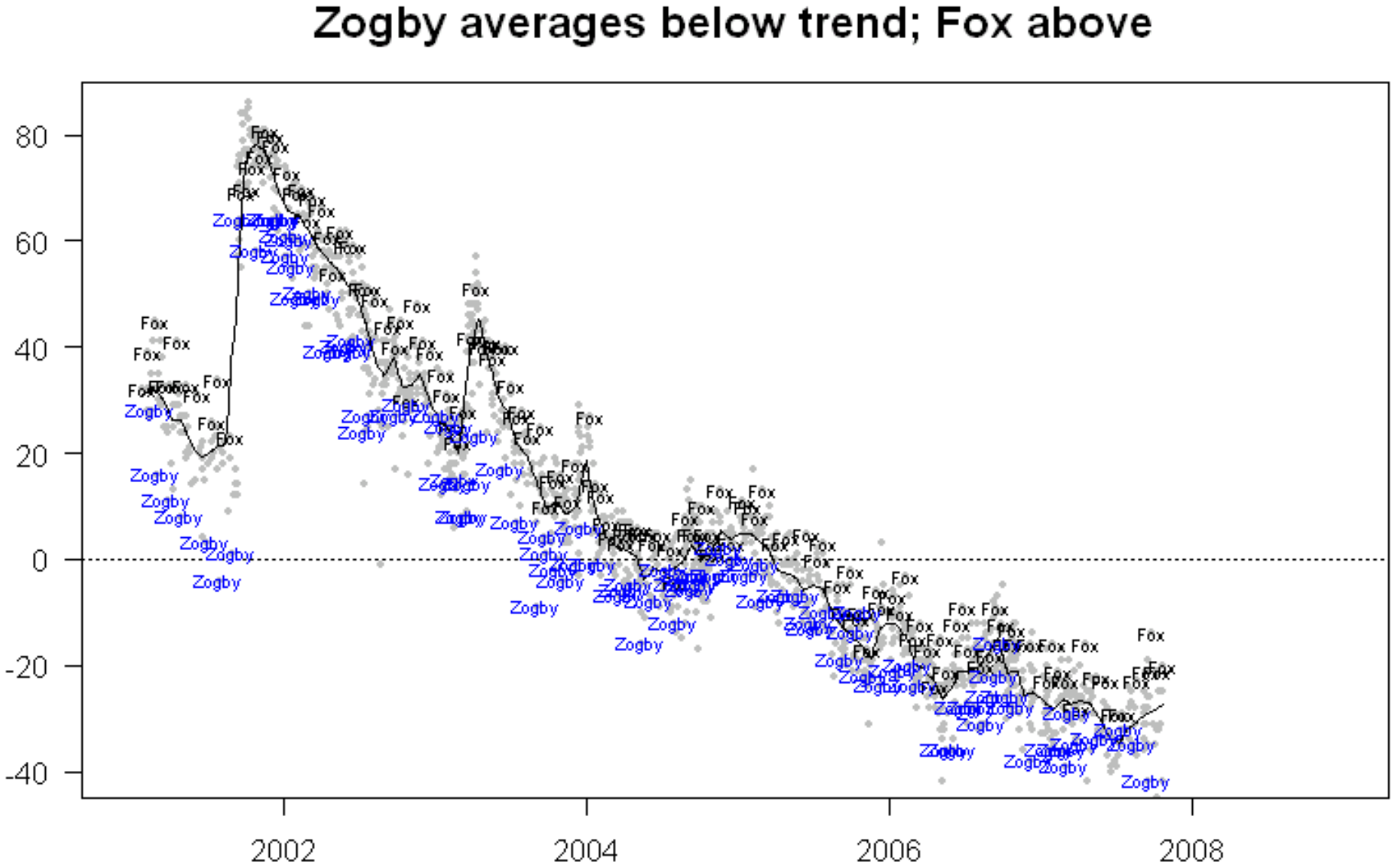
(likely) 20 percent of **schools districts** across the United States fail to meet the Average Yearly requirement of the No Child Left Behind Act.

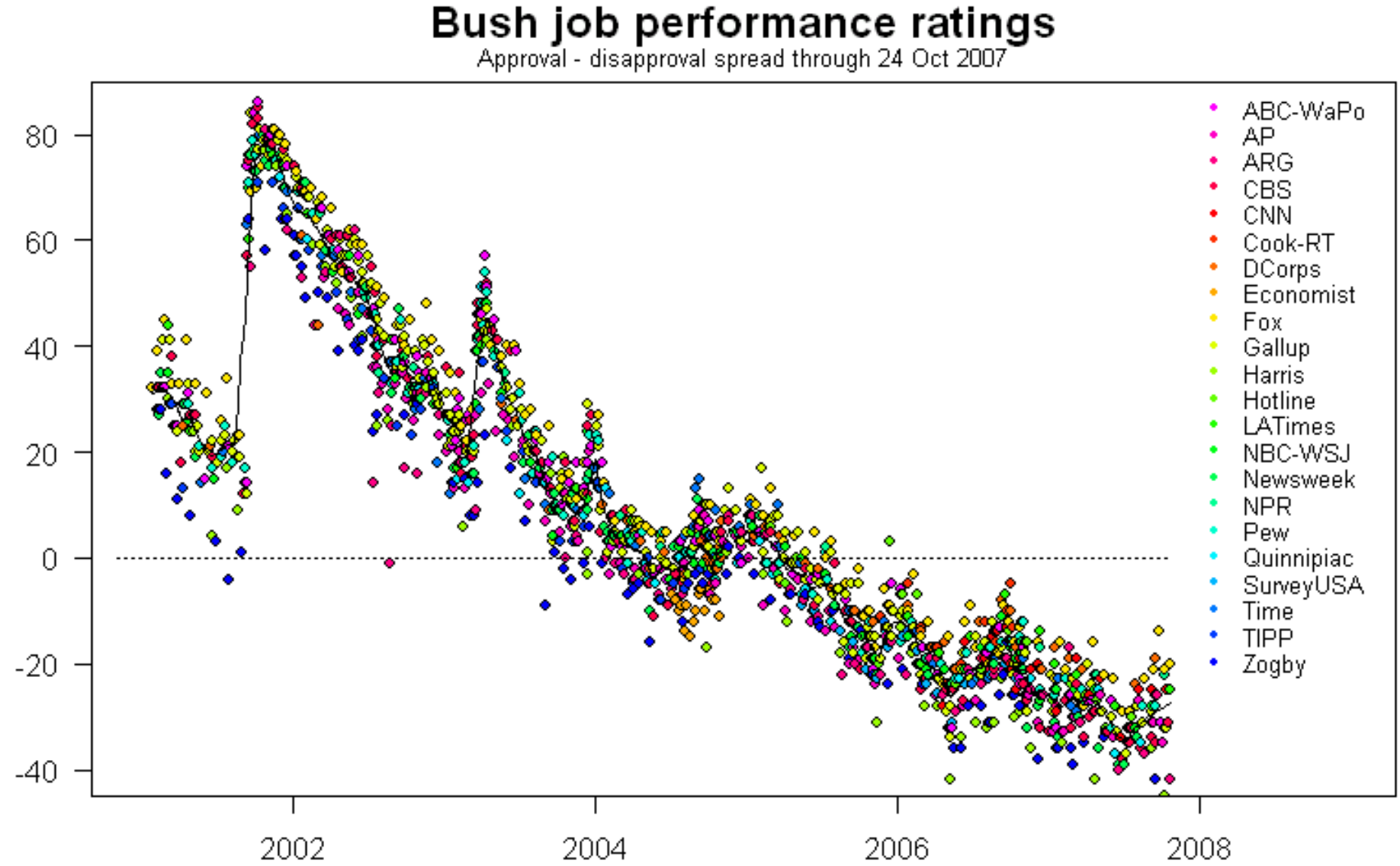
Directly Challenging the Truth of Conclusion

This is inappropriate for deductive arguments

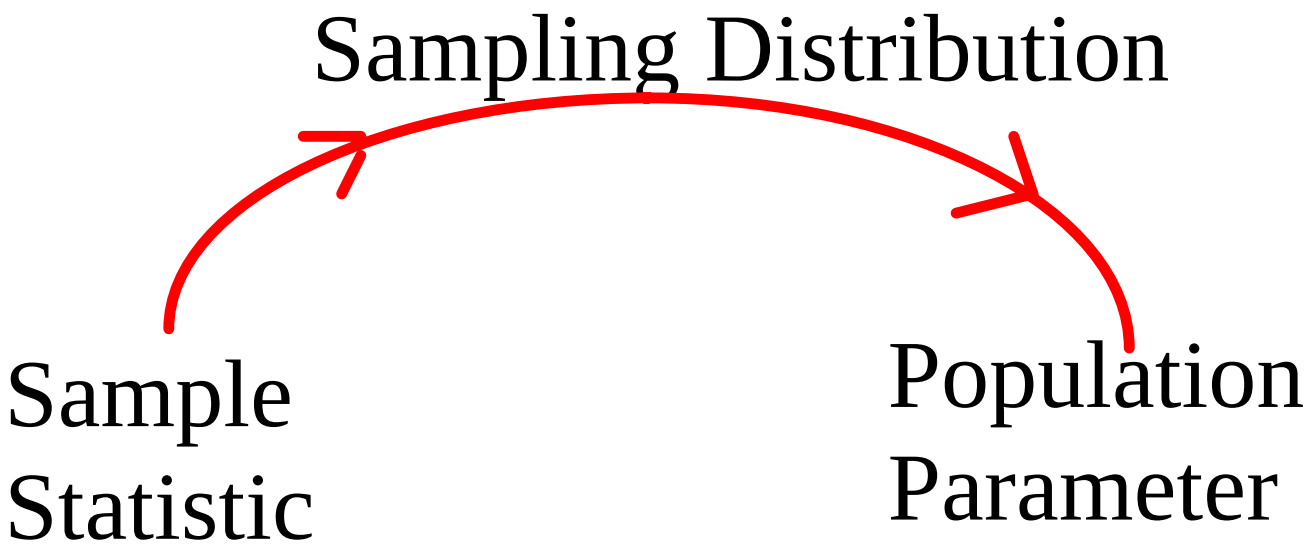
It might include counter evidence from other samples or other studies

Tracking polls on Bush's job approval rating




Bush job performance ratings
Approval - disapproval spread through 24 Oct 2007
80
60
40
20
0
-20
-40
2002
2004
2006
2008
ABC-WaPo
AP
ARG
CBS
CNN
Cook-RT
DCorps
Economist
Fox
Gallup
Harris
Hotline
LATimes
NBC-WSJ
Newsweek
NPR
Pew
Quinnipiac
SurveyUSA
Time
TIPP
Zogby

Inferential Statistics uses the sampling distribution to bridge the gap between sample and population



Recall

- **Parameter**
  - fixed, unknown number that describes the population
- **Statistic**
  - known value calculated from a sample
  - a statistic is used to estimate a parameter
    - **Sampling Variability**
      - » different samples from the same population may yield different values of the sample statistic
      - » estimates from samples will be closer to the true values in the population if the samples are larger

- Example:
  - The amount by which the proportion obtained from the sample ($\hat{p}$) will differ from the true population proportion (p) rarely exceeds the *margin of error*.

- **Sampling Distribution**
  - tells what values a statistic takes and how often it takes those values in repeated sampling.

- Example:
  - sample proportions ($\hat{p}$'s) from repeated sampling would have a normal distribution with a certain mean and standard deviation.

Example Fingerprints

- Fingerprints are a “sexually dimorphic trait…which means they are among traits that may be influenced by prenatal hormones.”
- It is known…
  - Most people have more ridges in the fingerprints of the right hand. (People with more ridges in the left hand have “leftward asymmetry.”)
  - Women are more likely than men to have leftward asymmetry.
- Compare fingerprint patterns of straight and gay men.

Fingerprint Study Results

- 66 gay men were studied.
  - 20 (30%) of the gay men showed left asymmetry.
- 186 straight men were also studied
  - 26 (14%) of the straight men showed left asymmetry.

Fingerprint Study Question

Assume that the proportion of all men who have the asymmetry is 15%.

Is it unusual to observe a sample of 66 men with a sample proportion ($\hat{p}$) of 30% if the true population proportion (p) is 15%?

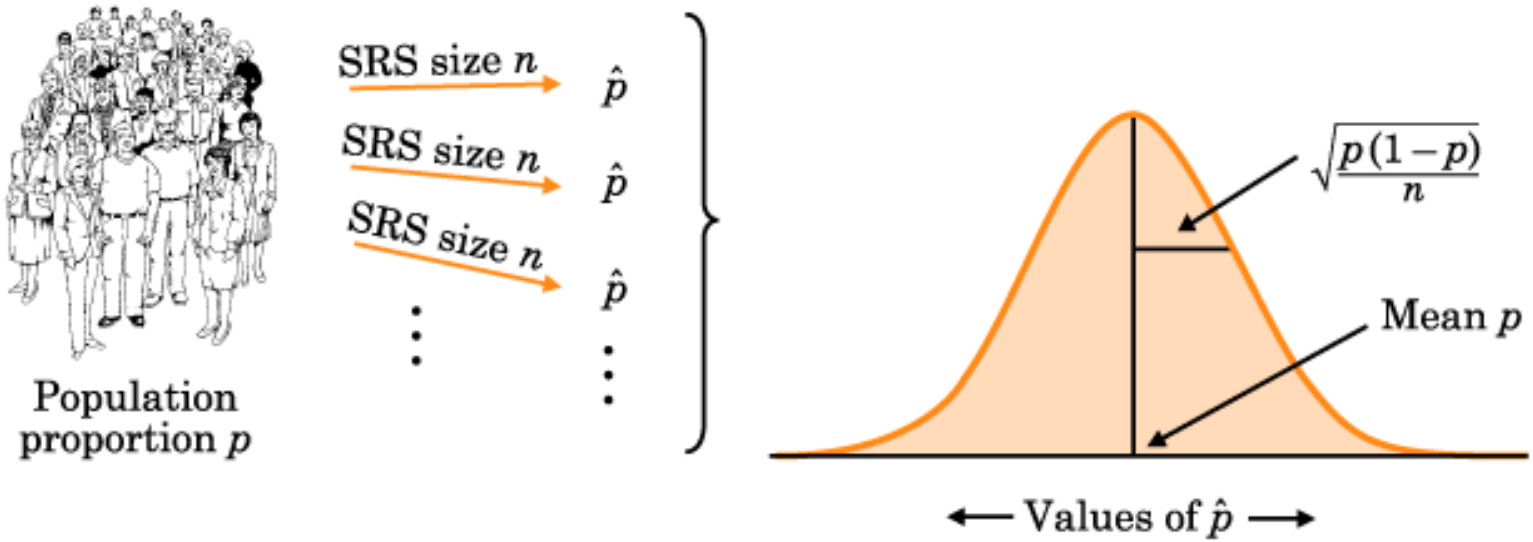
The Rule for Sample Proportions

If numerous samples or repetitions of size *n* are taken, the frequency curve of the sample proportions $\hat{p}$ from various samples will be *approximately bell-shaped*. The *mean* of those sample proportions will be ***p*** (the population proportion). The *standard deviation* will be:

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

Rule Conditions and Illustration

- For rule to be valid, must have
  - Random sample
  - 'Large' sample size



The Rule for Sample Proportions Applied to the Case Study

$$p = 0.15 \ (= \text{mean}); \quad n = 66$$

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.15(1-0.15)}{66}}$$

$$= 0.044 \ (= \text{s.d.})$$

Where should 95% of the sample proportions lie?

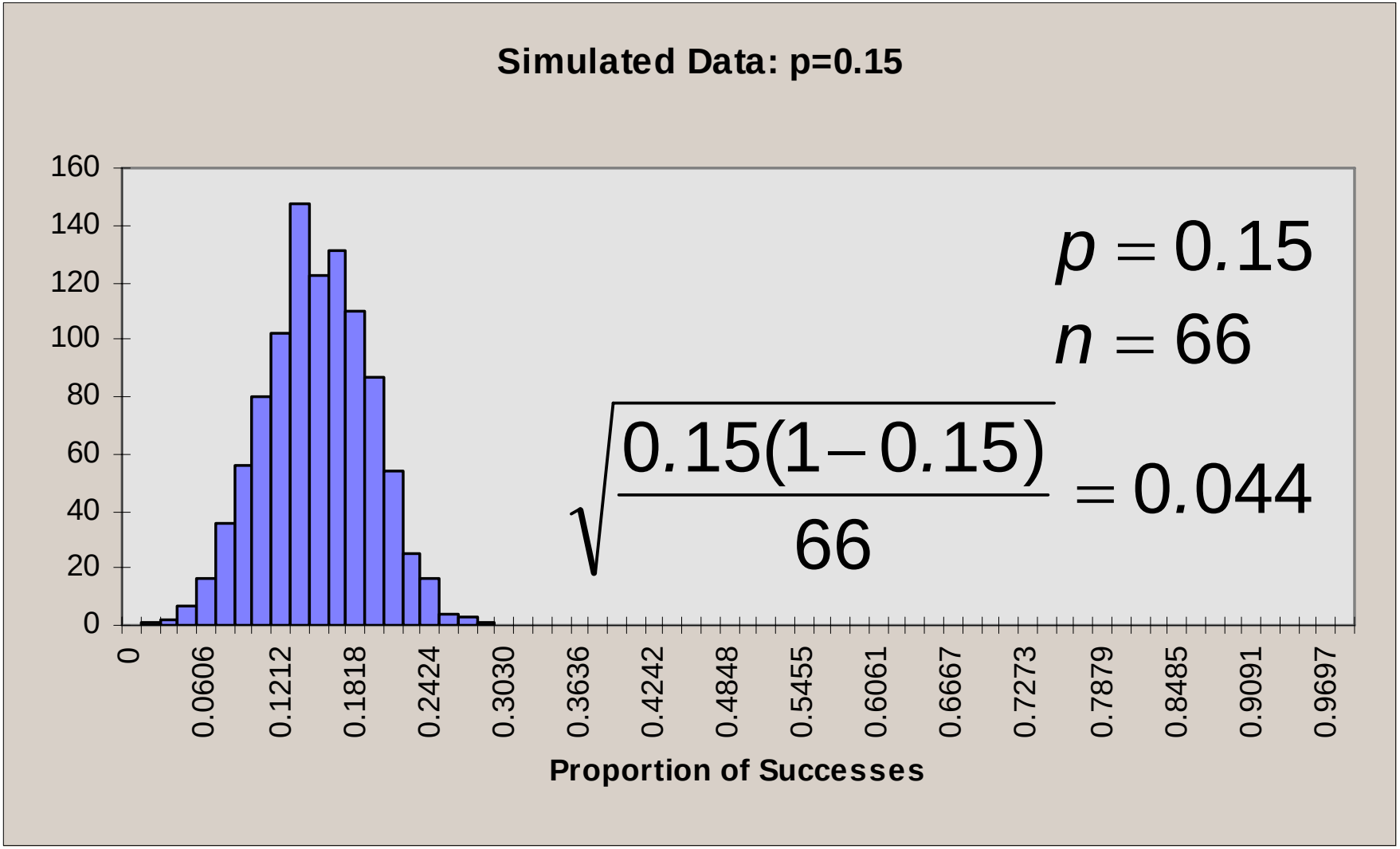
- mean plus or minus about two (1.96) standard deviations

$$0.15 - 2(0.044) = 0.062$$

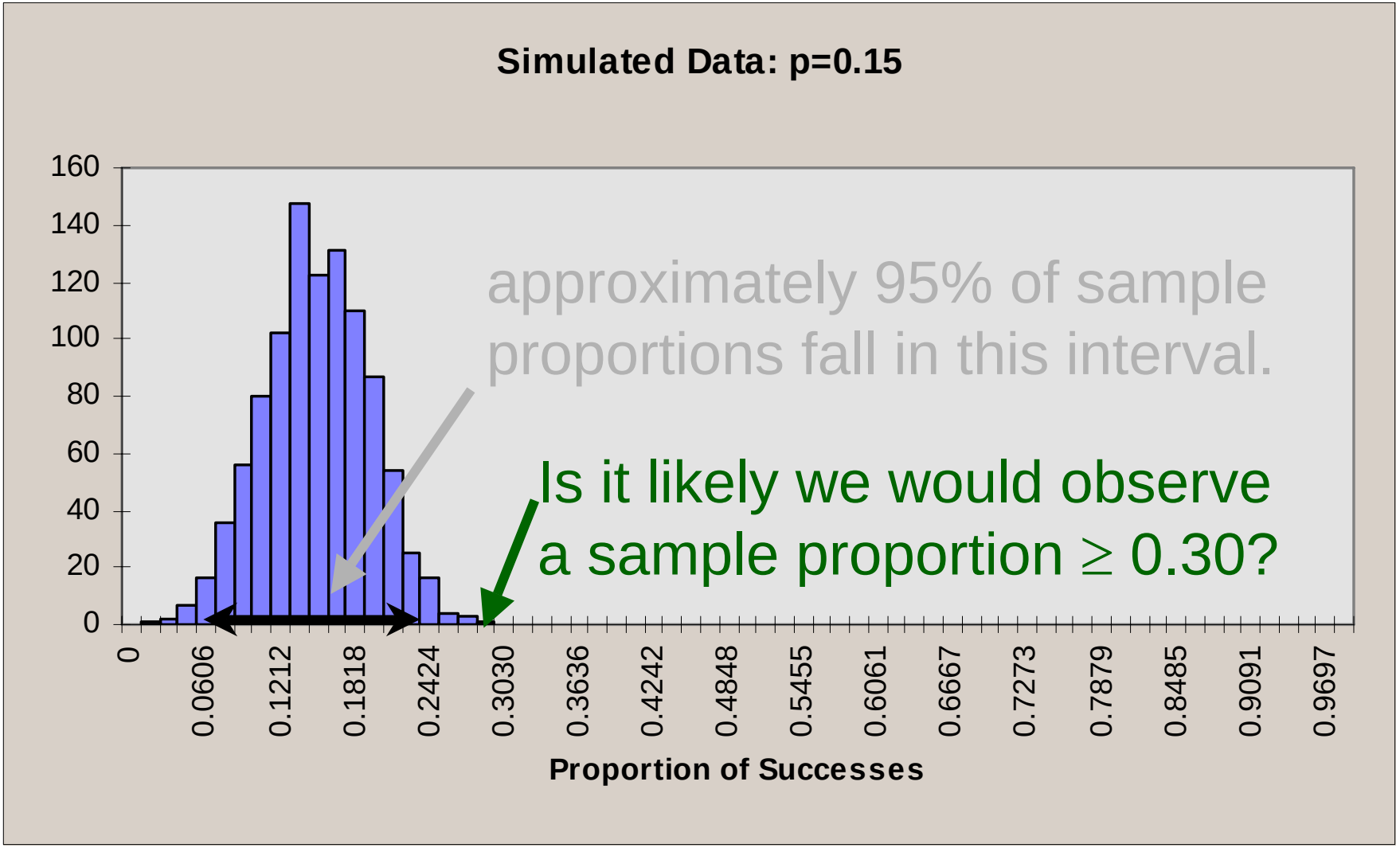
$$0.15 + 2(0.044) = 0.238$$

- 95% should fall between 0.062 & 0.238

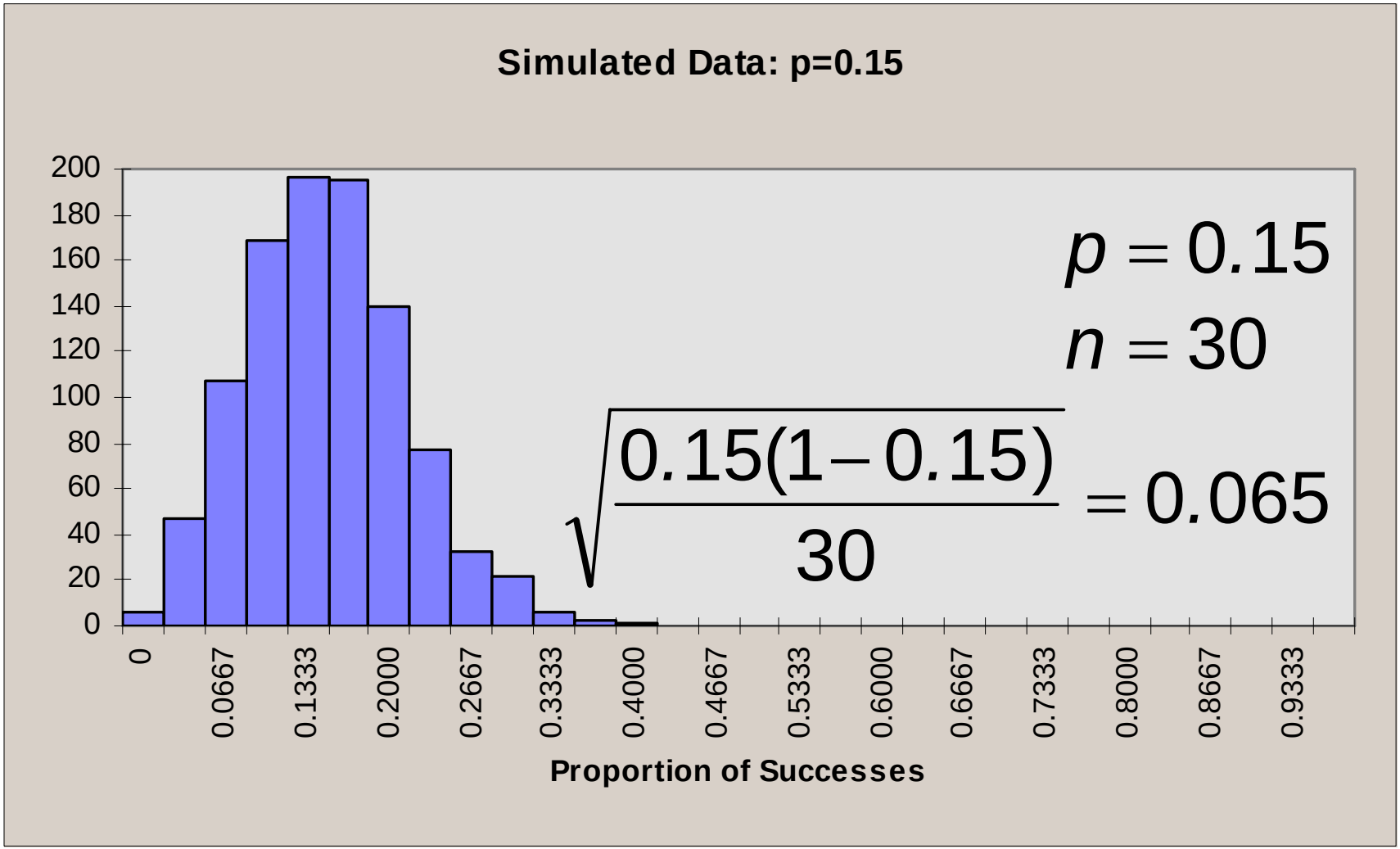
1000 Simulated Samples (n=66)



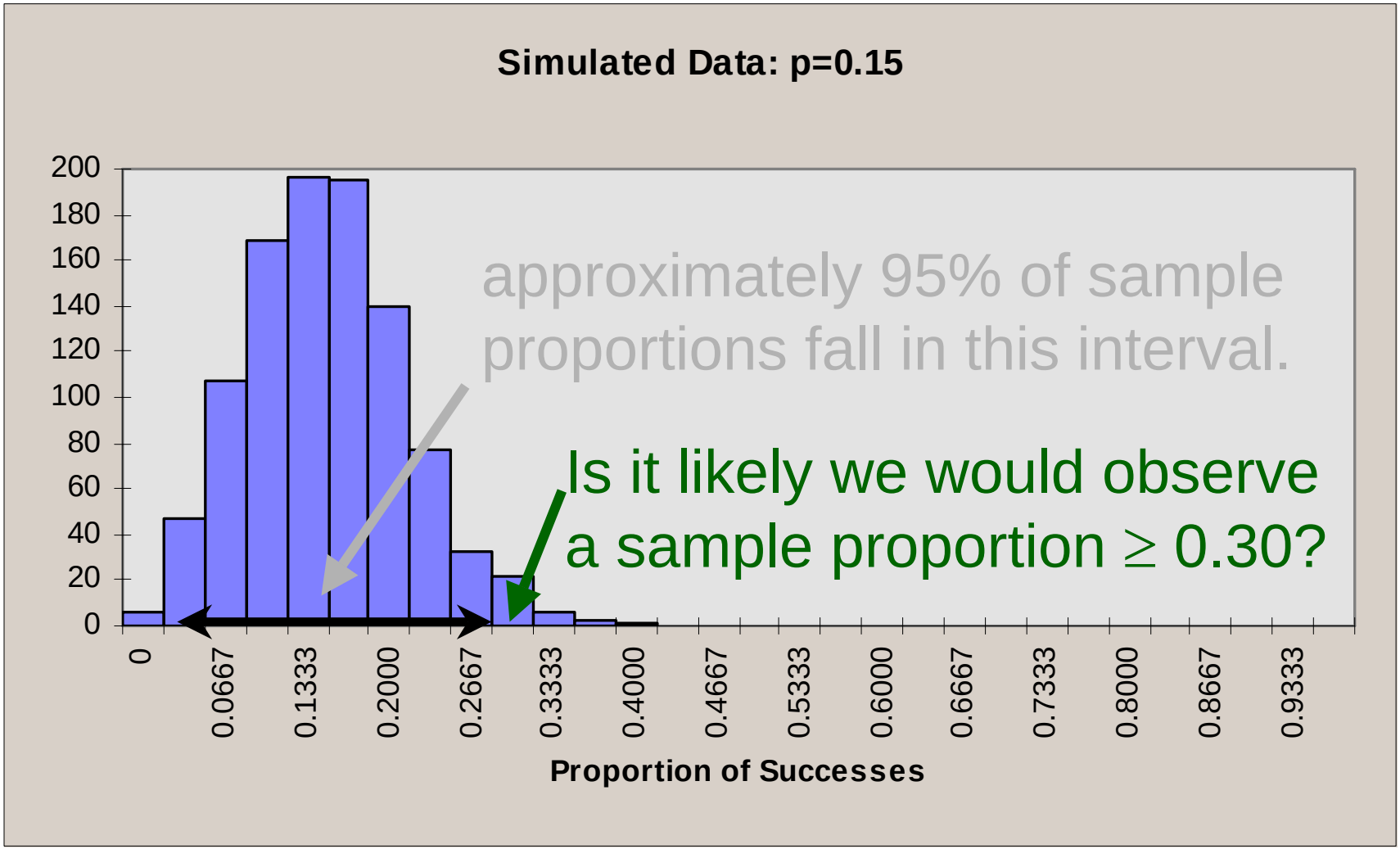
1000 Simulated Samples (n=66)



1000 Simulated Samples (n=**30**)



1000 Simulated Samples (n=**30**)



Confidence Interval for a Population Proportion

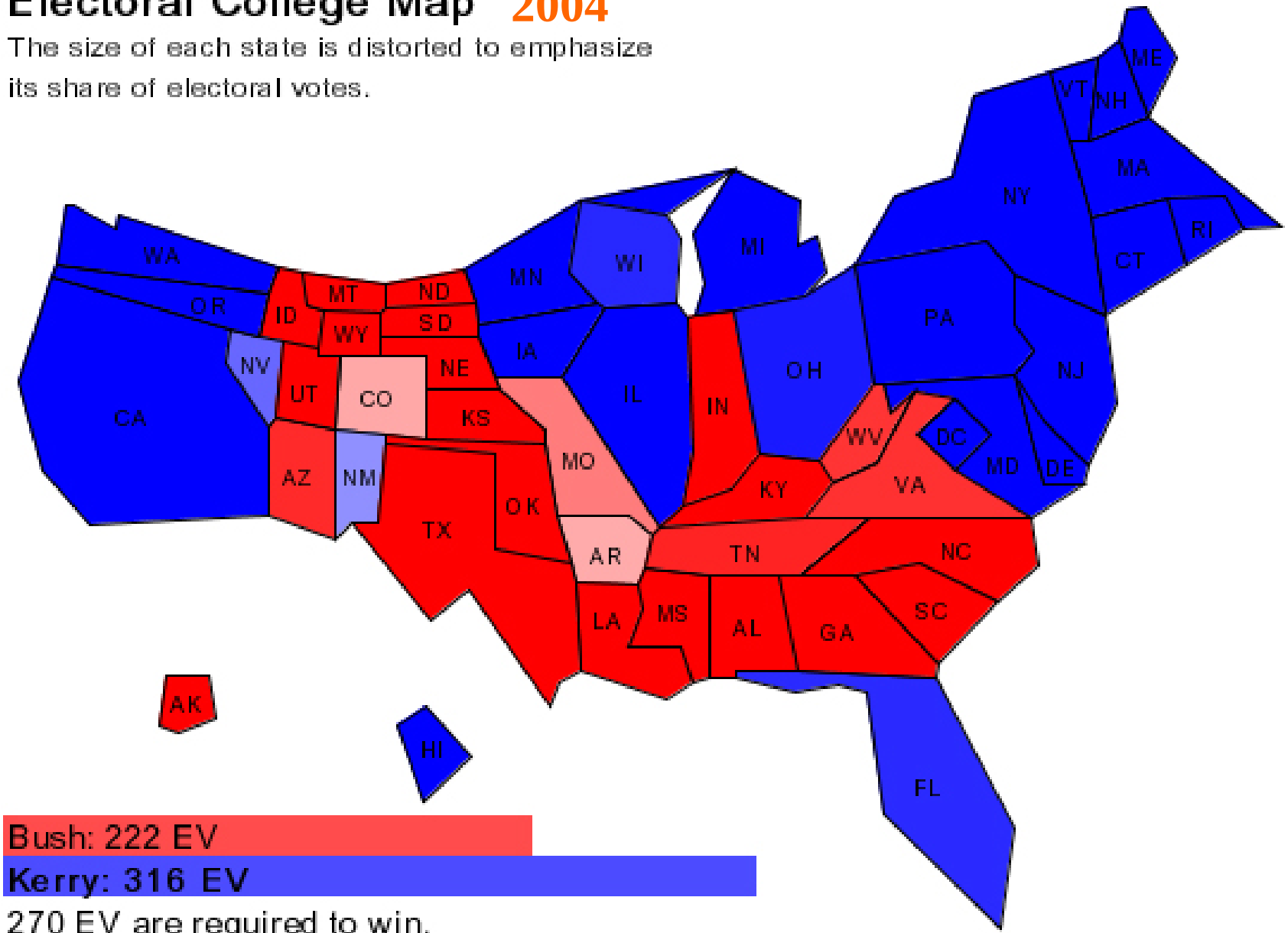
- An interval of values, computed from sample data, that is almost sure to cover the true population proportion.
- *"We are 'highly confident' that the true population proportion is contained in the calculated interval."*

Formula for a 95% Confidence Interval for the Population Proportion (Empirical Rule)

- sample proportion plus or minus two standard deviations of the sample proportion: $\hat{p} \pm 2\sqrt{\frac{p(1-p)}{n}}$
- since we don't know the population proportion p (needed to calculate the standard deviation) we will use the sample proportion $\hat{p}$ in its place.

Electoral College Map 2004

The size of each state is distorted to emphasize its share of electoral votes.



Bush: 222 EV

Kerry: 316 EV

270 EV are required to win.


That's All Folks