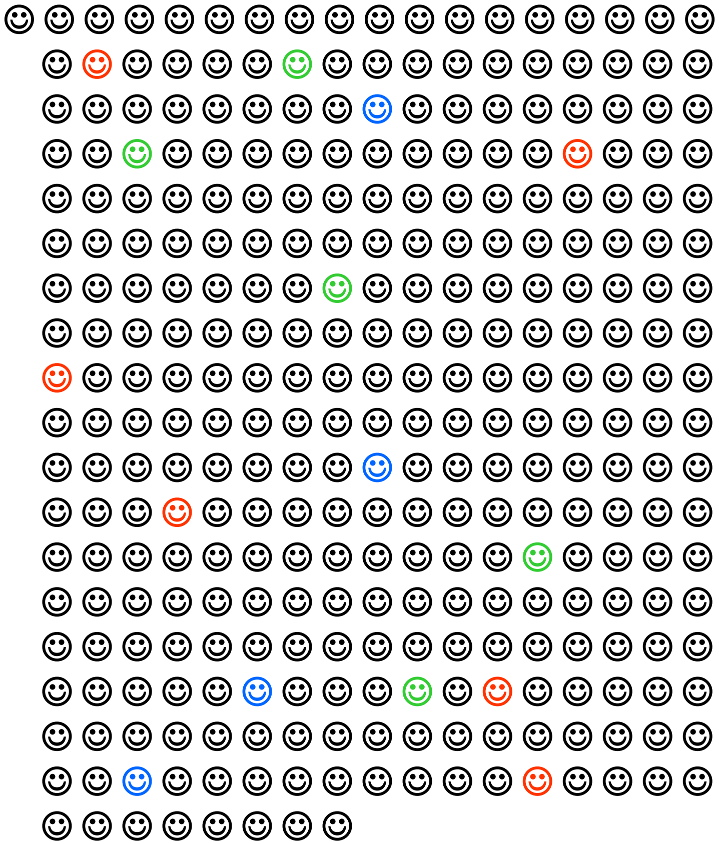
Foundations for Inferential Statistics

Portland State University
USP 634 Data Analysis I
Spring 2018

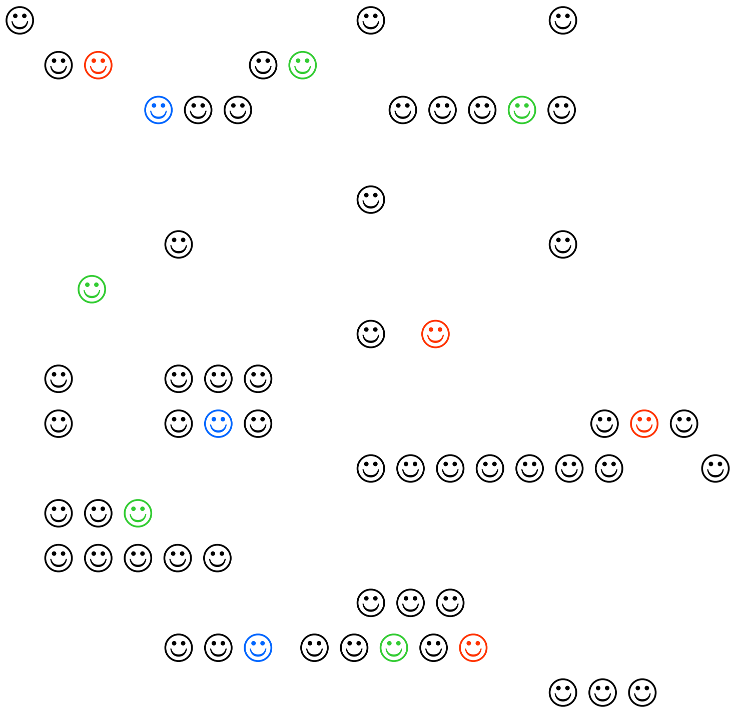
Inferential Statistics

The problem is…

- The *populations* we wish to study are almost always so large that we are unable to gather information from every case.



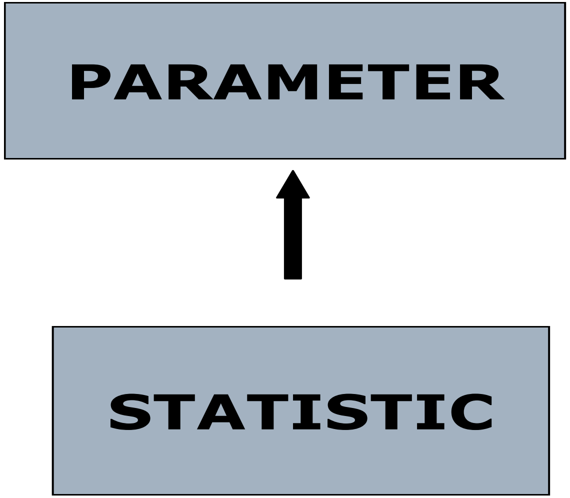
And the solution is…



- We choose a *sample* -- a carefully chosen subset of the population – and use information gathered from the cases in the sample to generalize to the population.

“Statistic” vs “parameter”

- *Statistics* are mathematical characteristics of samples.
- *Parameters* are mathematical characteristics of populations.
- Statistics are used to estimate parameters.



For Inference to Work

- Samples must be representative of the population.
- *Representative*: The sample has the same characteristics as the population.
- Samples drawn according to the rule of equal Probability of Selection Method

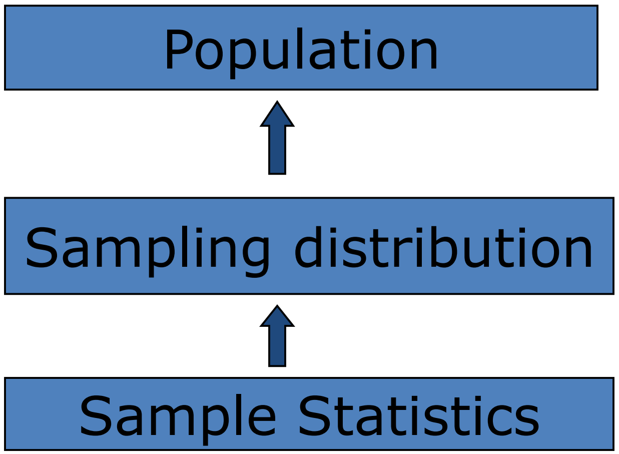
THE SAMPLING DISTRIBUTION

Sampling distribution

- The sampling distribution is a theoretical probability distribution of a sample statistic based on repeated samples of a given size.
- Allows us to
  - build confidence intervals (e.g., 95% confident that average auto use is between 3.5 and 4.1 trips per day)
  - test hypotheses (e.g., whether share of households in poverty is greater than 15 percent)

The sampling distribution

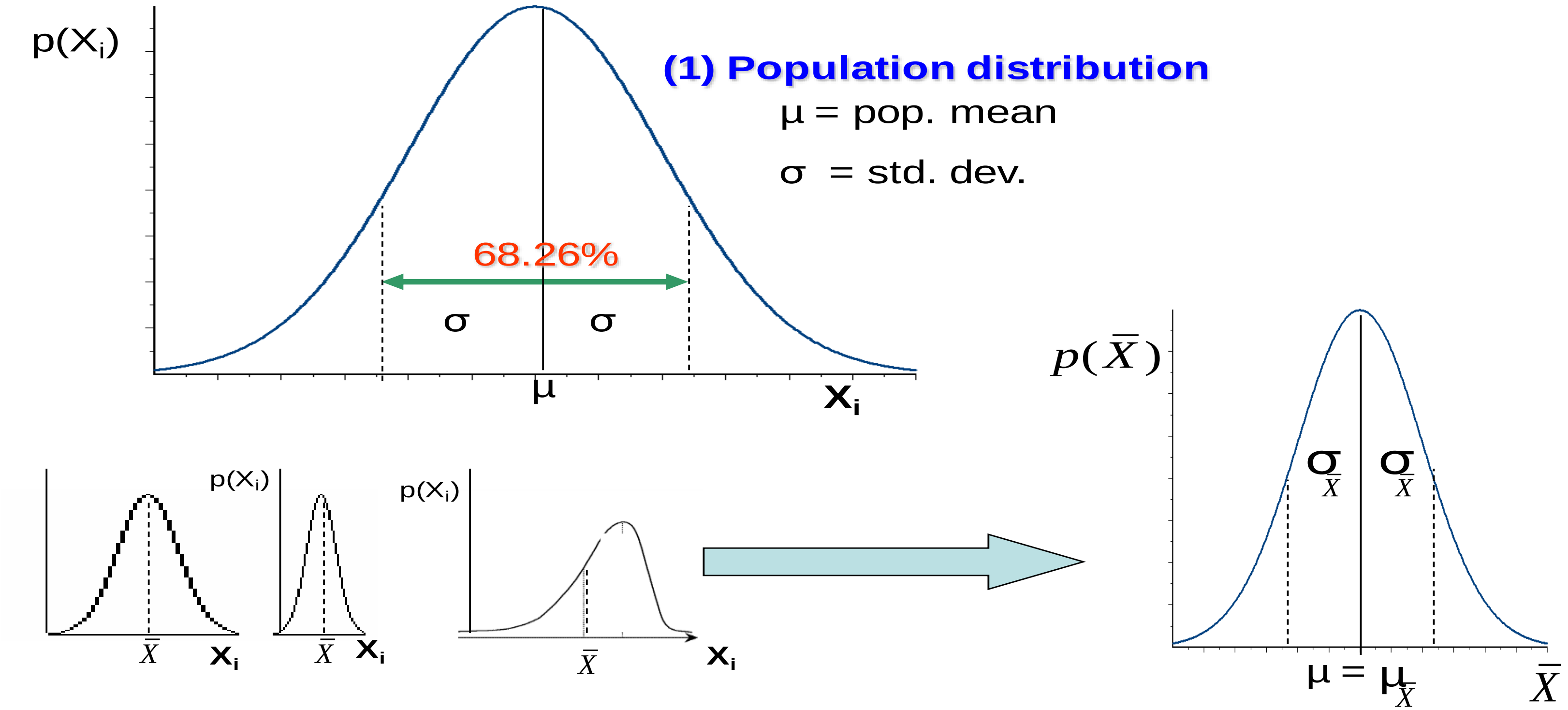
- Every application of inferential statistics involves 3 different distributions.
- Information from the sample is linked to the population via the sampling distribution.



Inferring population mean from sample mean

for normally distributed variables

- If repeated random samples of size *N* are drawn from a population with a normally distributed variable with mean μ and standard deviation σ, the distribution of sample means will be normal with
  - mean of μ
  - standard deviation of $\frac{\sigma}{\sqrt{N}}$



(1) Population distribution

μ = pop. mean

σ = std. dev.

(2) Several samples

$\bar{X}$ = sample mean

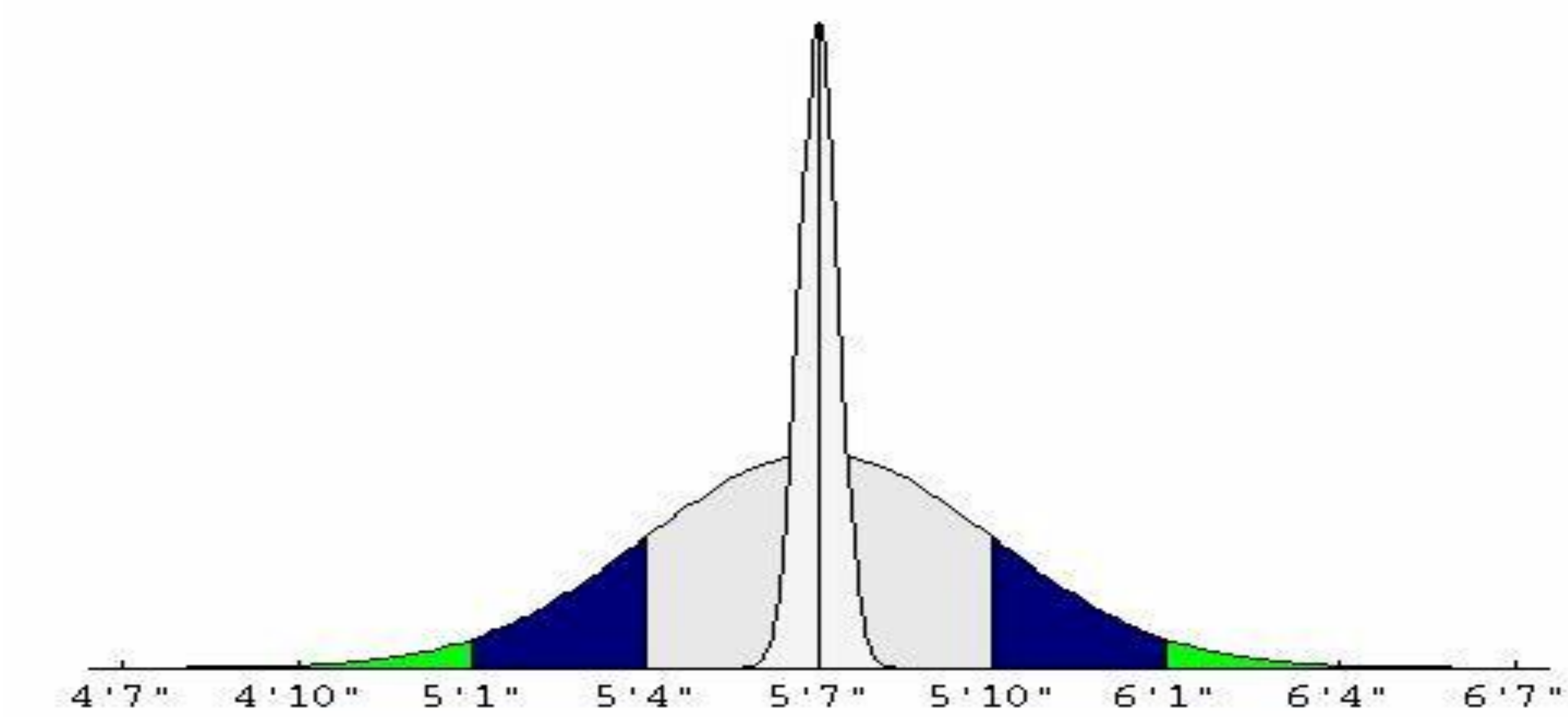
s = std. dev.

(3) Sampling distribution

$\mu_{\bar{X}}$ = mean of sample means

$\sigma_{\bar{X}}$ = std. dev. of sample means (standard error)

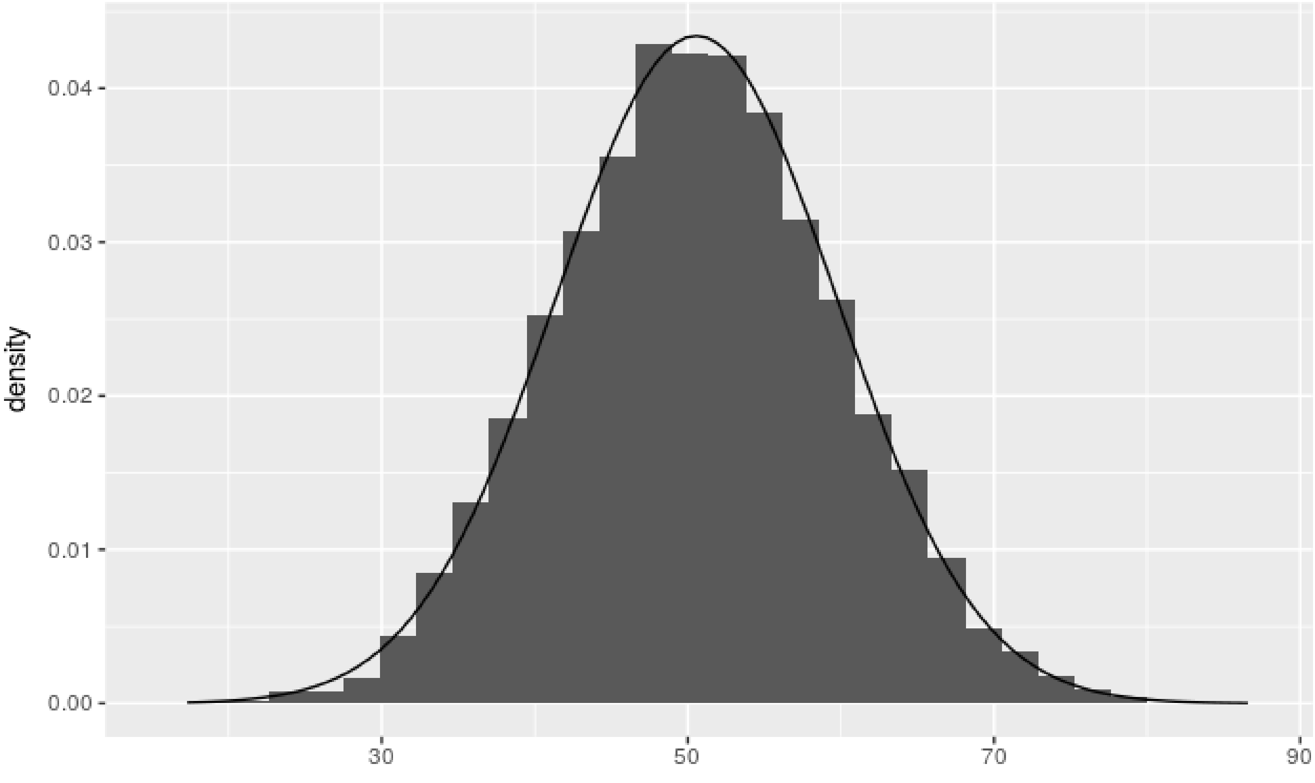
Population mean height 5’7”, sample size 80



Pop: 1-100

Sample: n=10, randon sample with replacement

Plot sample means of 10000 samples



Central Limit Theorem

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right),$$

where *SE* is represents standard error, which is defined as the standard deviation of the sampling distribution. If *σ* is unknown, use *s*.

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population mean.
- We won't go through a detailed proof of why $SE = \sigma / \sqrt{n}$, but note that as *n* increases *SE* decreases.
  - As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

CLT - conditions

Certain conditions must be met for the CLT to apply:

Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if

- random sampling / assignment is used, and
- if sampling without replacement, n < 10% of the population.

Sample size / skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.

- the more skewed the population distribution, the larger sample size we need for the CLT to apply
- for moderately skewed distributions **n > 30** is a widely used rule of thumb

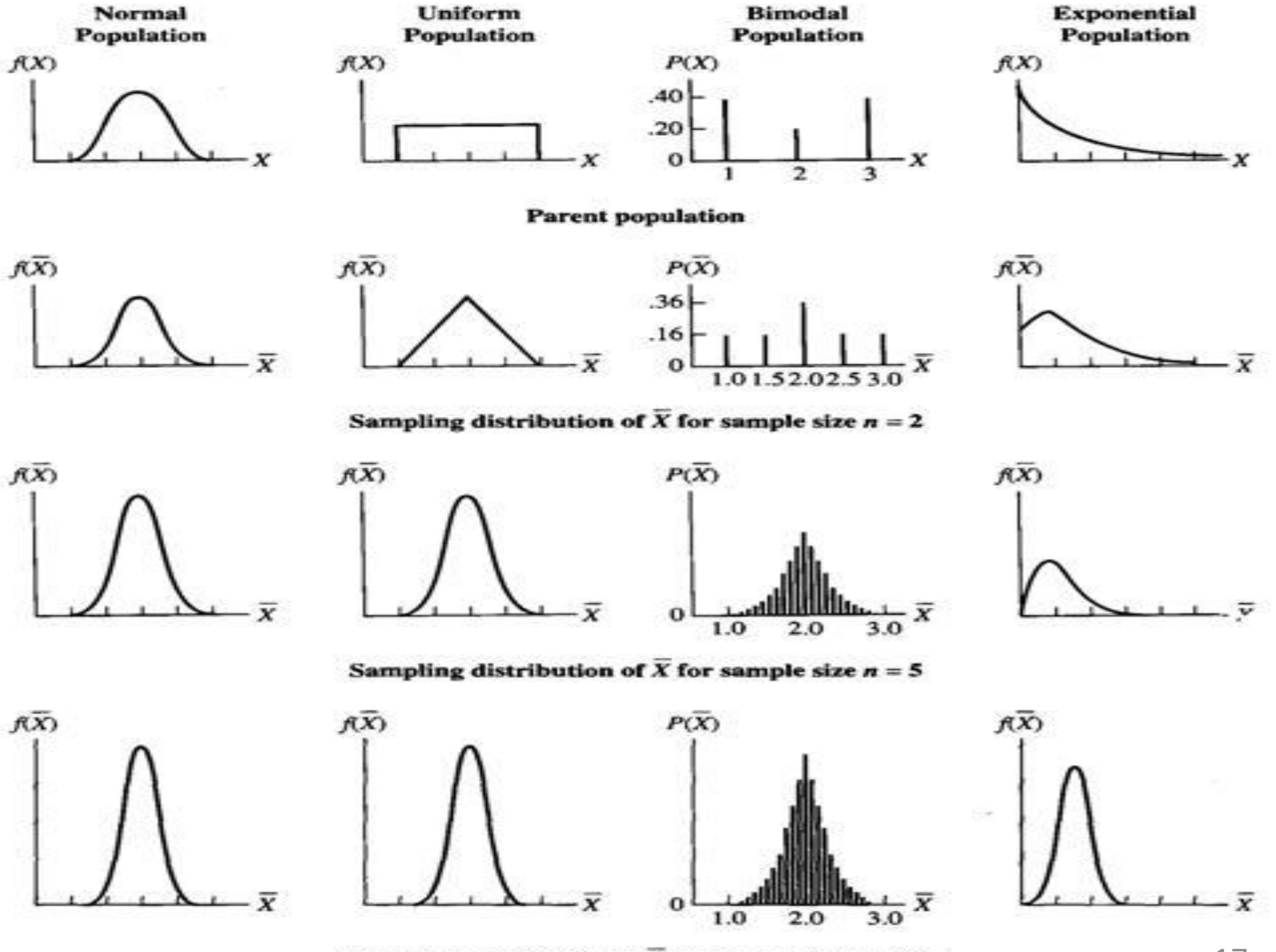
This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.

Population Distribution

n = 2

n = 5

n = 30



Parent population

Sampling distribution of $\bar{X}$ for sample size $n = 2$

Sampling distribution of $\bar{X}$ for sample size $n = 5$

Sampling distribution of $\bar{X}$ for sample size $n = 30$

Estimation Procedures

- **Estimates of Population Parameters:**

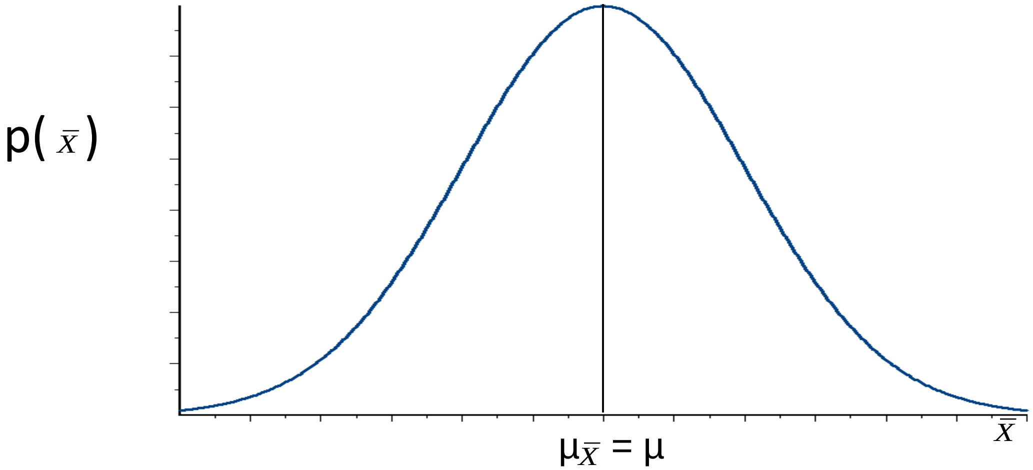
  - **POINT**: Best single value to estimate a parameter (e.g., $\bar{X}$ is the best single estimate of μ)

  - **INTERVAL**: Range of values we're confident at a certain probability (confidence) level that encompasses the true population parameter

2 Properties of an Estimator

- Use $\bar{X}$ to estimate μ. Why?

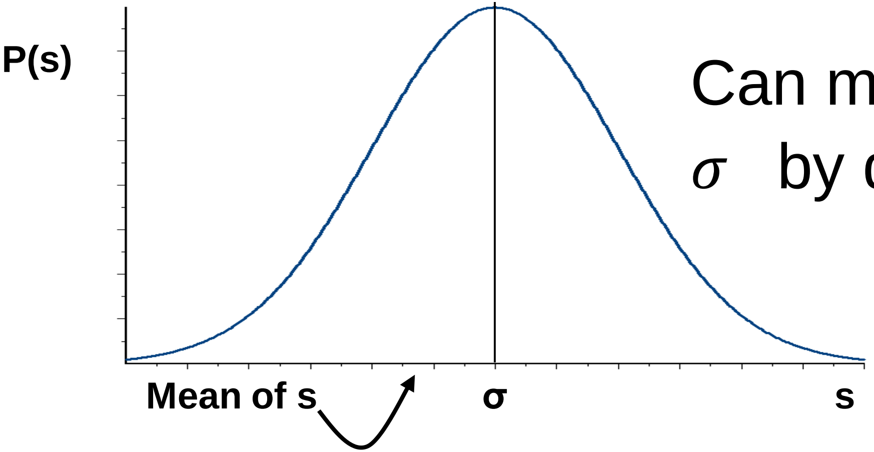
(1) It's an Unbiased estimator. CLT tells us…on average, the sample mean ($\bar{X}$) is on target.



Sample Standard Deviation (*s*) is a Biased Estimate of Population Standard Deviation (σ)

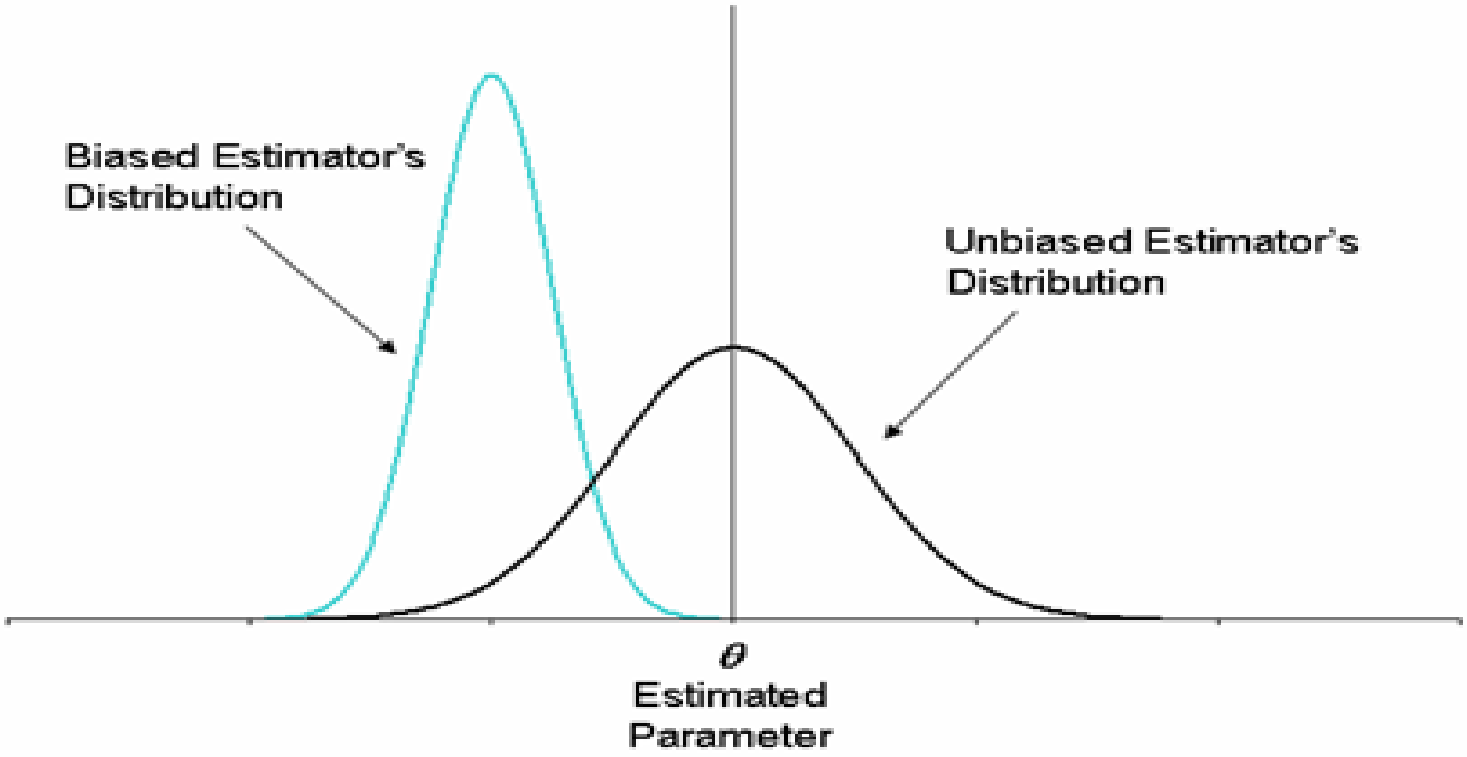
Standard Deviation

$$\boldsymbol{s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N}}} \quad \longrightarrow \quad \boldsymbol{\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}}$$



Can make s unbiased estimate of σ by dividing N-1:

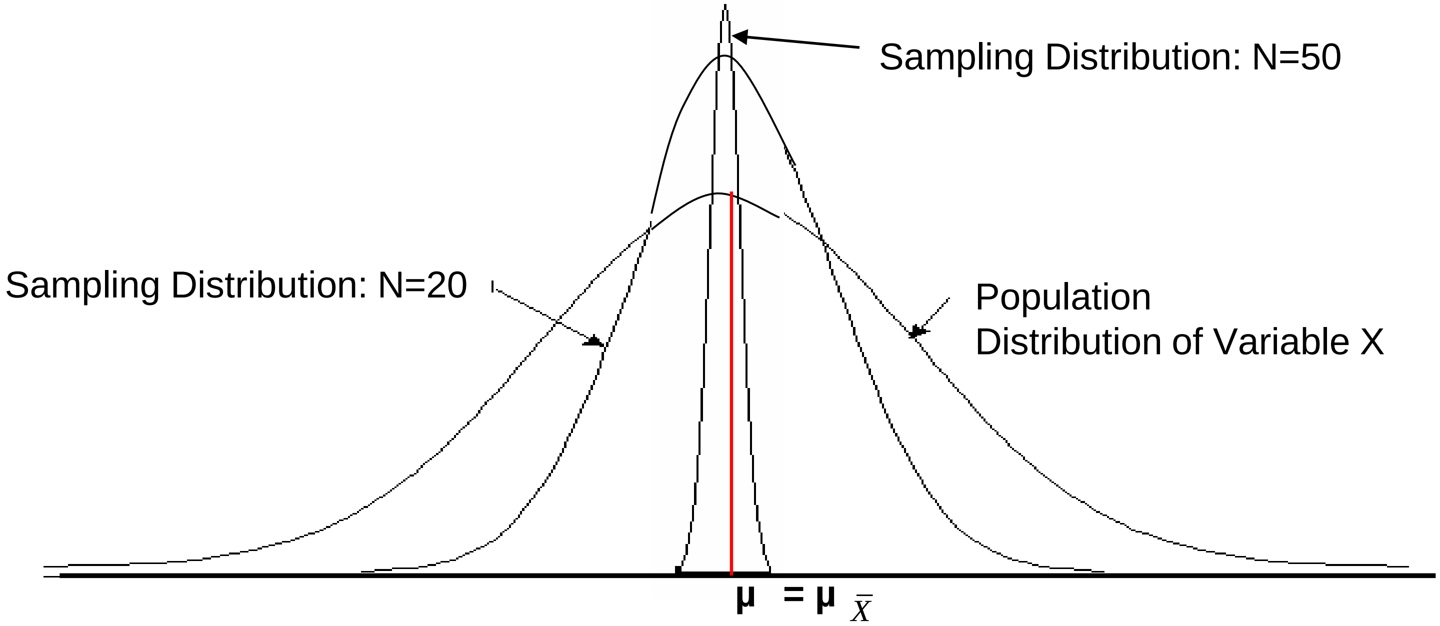
$$\boldsymbol{s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N - 1}}}$$


Biased Estimator's Distribution
Unbiased Estimator's Distribution
θ
Estimated Parameter

2 Properties of an Estimator

(2) **<u>Efficiency</u>**: The sampling distribution of the estimate (e.g., $\bar{X}$) is clustered around the true parameter (μ).

From CLT, we know $\bar{X}$ is an efficient estimate of μ, because as N gets large, $\sigma_{\bar{X}}$ gets small



Sampling Distribution: N=50

Sampling Distribution: N=20

Population
Distribution of Variable X

$\mu = \mu_{\bar{X}}$

Point Estimates from Sample

- The sample mean ($\bar{X}$) is an unbiased estimator of the population mean
  - Its average is the population mean
  - Its standard error can be small with large samples
- Sample standard deviation *s* is a biased estimator of the population standard deviation

$$\sigma \sim s = \sqrt{\frac{\sum_n (X - \bar{X})^2}{N}}$$

s = $\sqrt{\frac{\sum_n (X-\bar{X})^2}{N-1}}$ (unbiased)

Point Estimates for Proportions

- The sample proportion (P_s) is an unbiased and efficient estimator of the population proportion (P_u):

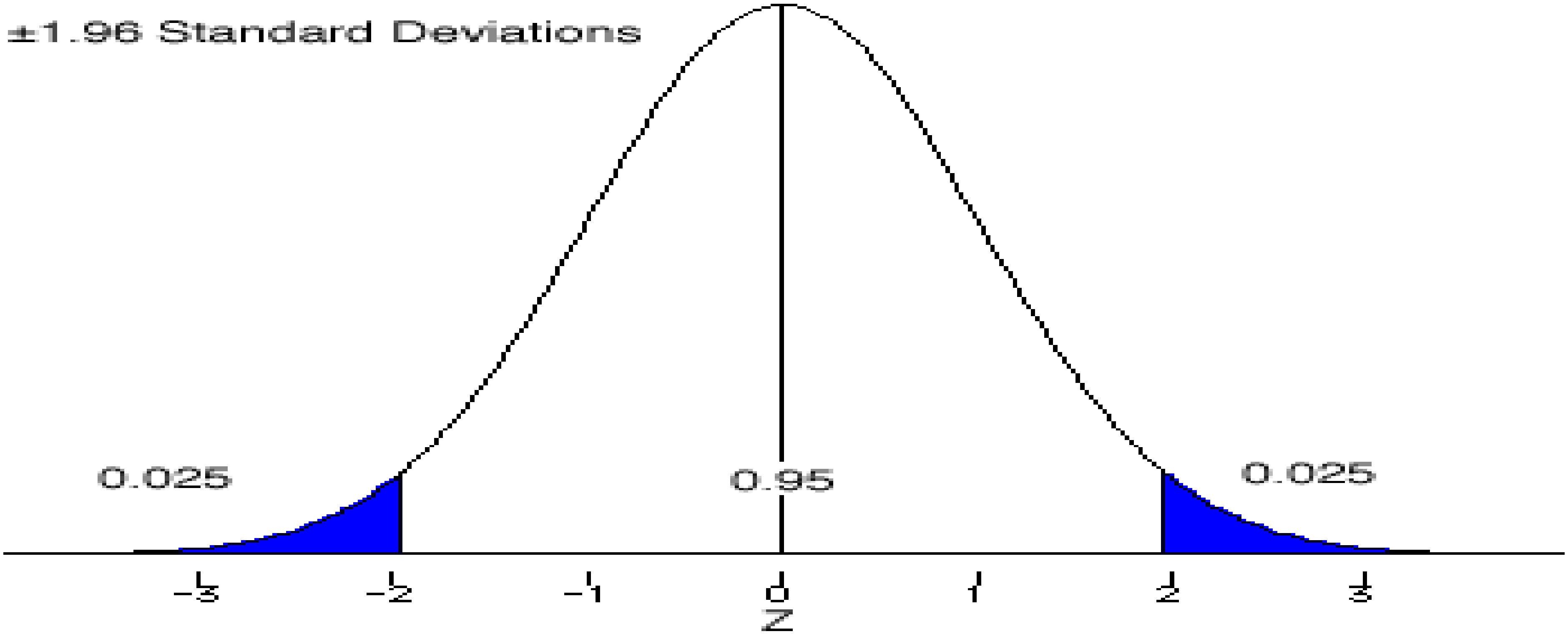
$$\mu_p = P_s$$

$$\sigma_p = \sqrt{P_s(1 - P_s)}$$

CONFIDENCE INTERVAL

Confidence intervals

- Rather than a single point estimate, find a range of values likely to include the true mean or proportion – a “confidence interval”
- Common to choose a 95% confidence interval, which means 5% of the time we’ll be in error … “critical value” $\alpha = .05$
- Taking 95% of the area around a value leaves 2.5% in each tail. The Z value for 95% area is 1.96. $Z_{\alpha/2}$=.025 = 1.96, or approximately 2.


±1.96 Standard Deviations
0.025
0.95
0.025
−3
−2
−1
0
1
2
3
Z

Deriving the confidence interval formula

$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} \approx \frac{s}{\sqrt{N}}$

$\bar{X} \sim N(\mu, \sigma_{\bar{X}})$

$P($-1.96$\leq Z \leq$ 1.96)=.95, and $Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$

$P(\bar{X} - 1.96\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{X}})$=.95

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$\bar{x} = 3.2$ $\qquad\qquad$ $s = 1.74$

The approximate 95% confidence interval is defined as

point estimate ± 2 x SE

$SE = s / \sqrt{n} = 1.74 / \sqrt{50} \approx 0.25$

$\bar{x} \pm 2 \times SE \quad \rightarrow \quad 3.2 \pm 2 \times 0.25$

$\rightarrow \quad (3.2 - 0.5, 3.2 + 0.5)$

$\rightarrow \quad (2.7, 3.7)$

Practice

Which of the following is the correct interpretation of this confidence interval?
We are 95% confident that
(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.
(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.
(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.
(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

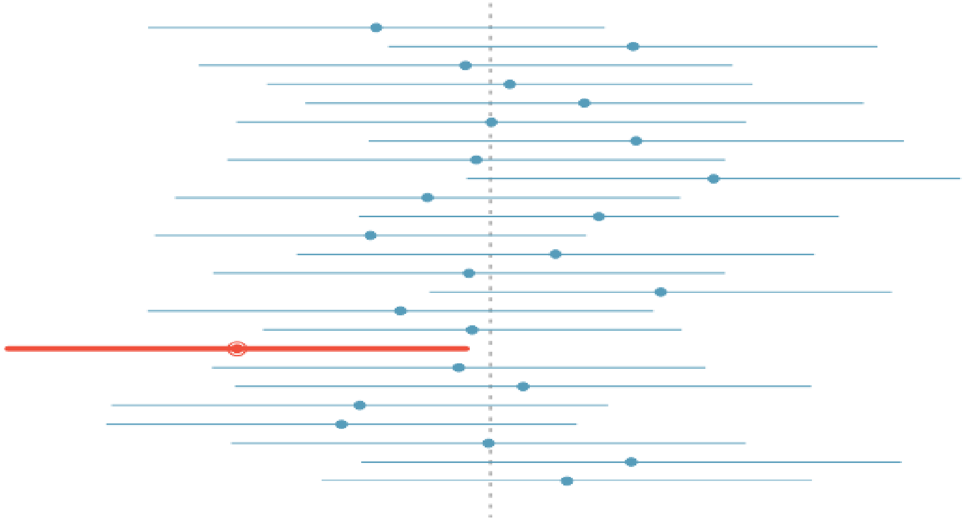
(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate ± 1.96 x SE*.

Then about 95% of those intervals would contain the true population mean (μ).

The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



An interactive demonstration of CI:

http://rpsychologist.com/d3/CI/

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?

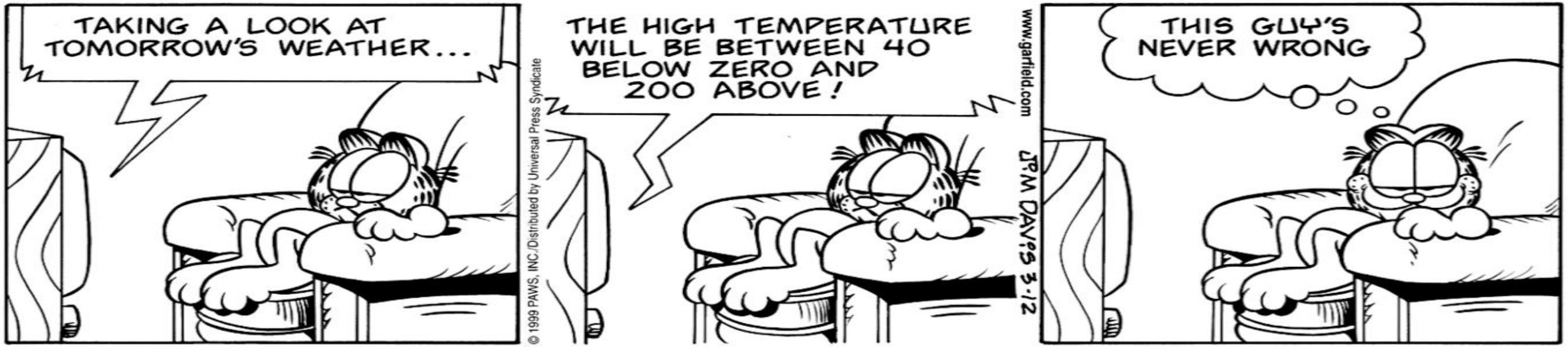

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVIS 3-12

Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Changing the confidence level

point estimate ± z x SE*

- In a confidence interval, *z* x SE* is called the margin of error, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust z* in the above formula.
- For a 95% confidence interval, z* = 1.96.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- However, using the standard normal (z) distribution, it is possible to find the appropriate z* for any confidence level.

Practice

Which of the below Z scores is the appropriate z* when calculating a 98% confidence interval?

(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = -2.33

(e) Z = -1.65

Practice

Which of the below Z scores is the appropriate z* when calculating a 98% confidence interval?

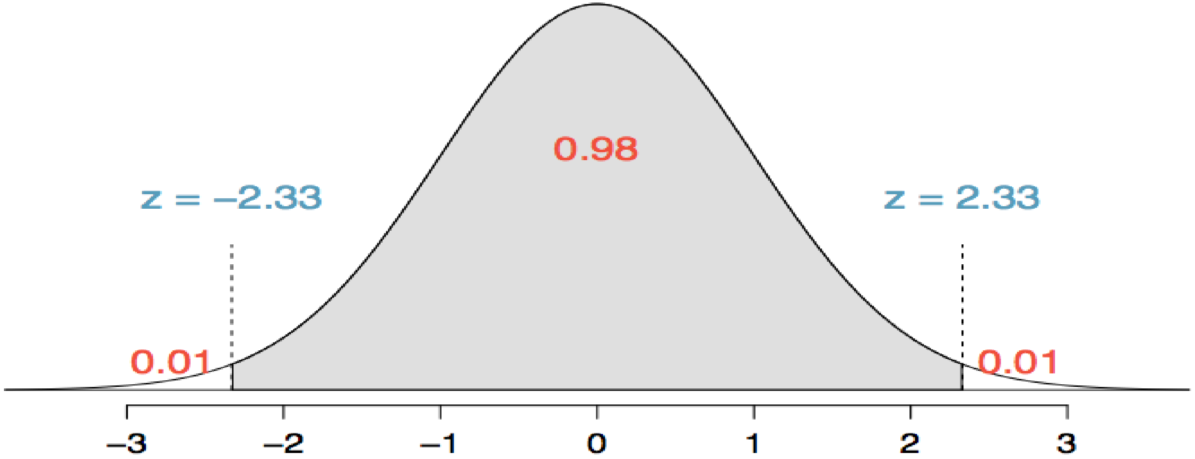
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = -2.33

(e) Z = -1.65



You can find the value with R: qnorm(0.01); using the Z-table (e.g.: http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf) or

the applet

Confidence interval for proportions (large samples)

- In random sample of 1,000 residents, 55% support regional veto power over local land-use decisions and 45% oppose it. Would a regional land-use referendum likely to pass?
- A different type of problem: we need to build a confidence interval around a bi-nominal variable's proportion, not an interval variable's mean. $P_s = .58$; N=1,000

Confidence interval for proportions (large samples)

- The Central Limit theorem still applies: The sampling distribution of sample proportions is a normal distribution with

$$\mu_P = P_u \text{ and } \sigma_P = \sqrt{\frac{P_u(1-P_u)}{N}}$$

$$\text{Confidence Intervals: } P_s \pm Z_{\alpha/2}\sqrt{\frac{P_u(1-P_u)}{N}}$$

- *What is P_u?*
  1. Not knowing $P_{\mathbf{u}}$ (but knowing $P_{\mathbf{s}}$), what is our "best guest" of $P_{\mathbf{u}}$. Could use $P_{\mathbf{s}}$.
  2. Assume we know nothing about the issue – i.e. a 50-50 odds, or $P_{\mathbf{u}}$ = .5. This "no knowledge" estimate also produces the biggest c.i. (A conservative estimate that produces the largest Standard Error.)

Confidence interval for proportions (large samples)

- In random sample of 1,000 residents, 54% support regional veto power over local land-use decisions and 46% oppose it. Would a regional land-use referendum likely to pass?

Point estimate ± 1.96 x SE

Ps = .55; N = 1000

SE = sqrt(.55 * (1-.55)/1000) = .0157

$P_s \pm Z_{\alpha/2}$ → .55 $\pm$ 1.96 * .0157

→ .55 $\pm$ .031

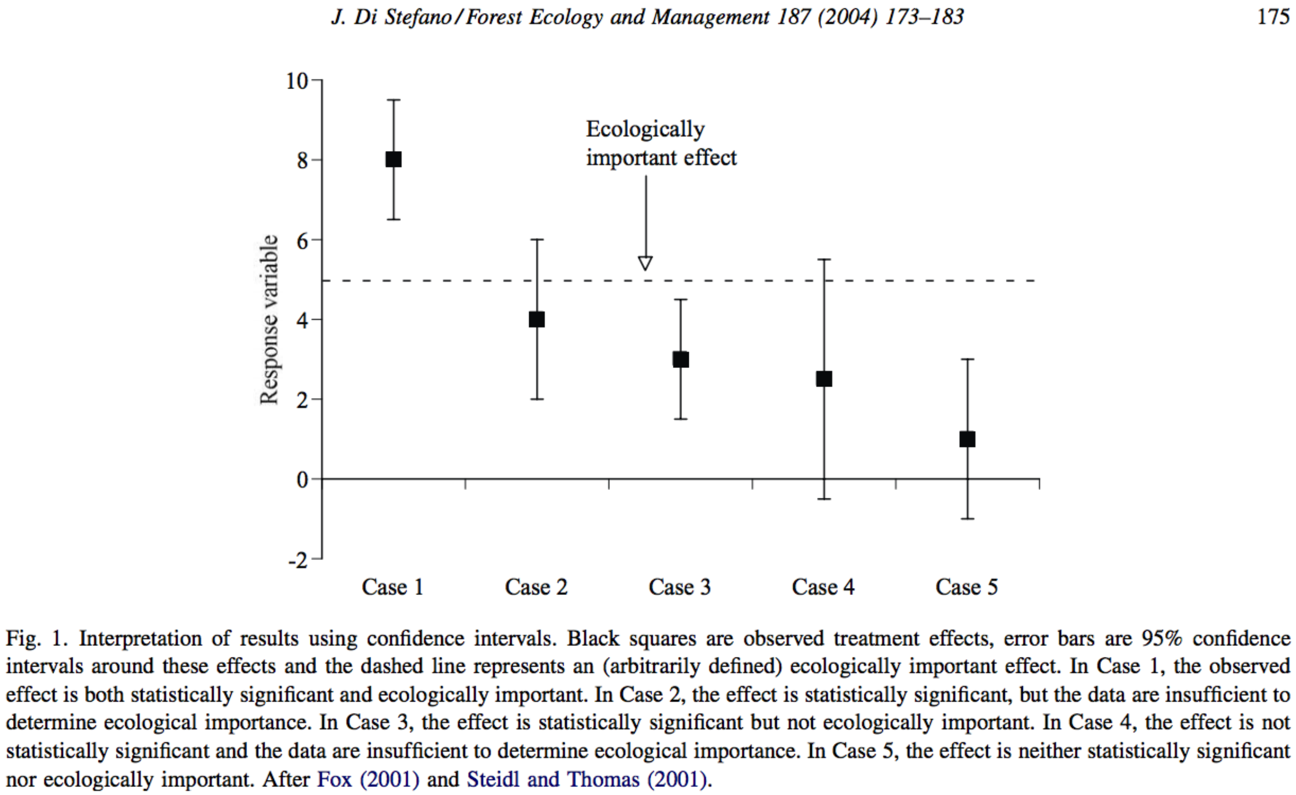
→ (.55 - .031, .55+.031)

→ (.519, .581)

If we want to be conservative

SE=sqrt(.50 * (1 - .50)/1000)=.0158

A Confidence Interval Approach to Data Analysis



J. Di Stefano / Forest Ecology and Management 187 (2004) 173–183 175

Fig. 1. Interpretation of results using confidence intervals. Black squares are observed treatment effects, error bars are 95% confidence intervals around these effects and the dashed line represents an (arbitrarily defined) ecologically important effect. In Case 1, the observed effect is both statistically significant and ecologically important. In Case 2, the effect is statistically significant, but the data are insufficient to determine ecological importance. In Case 3, the effect is statistically significant but not ecologically important. In Case 4, the effect is not statistically significant and the data are insufficient to determine ecological importance. In Case 5, the effect is neither statistically significant nor ecologically important. After Fox (2001) and Steidl and Thomas (2001).

Distributions and Key Statistics

	Population (Mean)	Population (Proportion)	Sample (Mean)	Sample (Proportion)
Observations			N	N
Mean	μ	P_μ	$\bar{X}$	P_s
Standard Deviation	σ	σ	s	$\sqrt{P_s\,(1-P_s\,)}$
Sampling Distribution			$N(\bar{X},\frac{s}{\sqrt{N}})$	$N\left(P_s,\frac{\sqrt{P_s\,(1-P_s\,)}}{\sqrt{N}}\right)$

SAMPLE SIZE

Choosing sample sizes to obtain desired accuracy for proportions

- Confidence interval: $P_s \pm Z_{\alpha/2}\sqrt{\frac{P_u(1-P_u)}{N}}$
- Margin of error: $E = Z_{\alpha/2}\sqrt{\frac{P_u(1-P_u)}{N}}$
- At 95% confidence, assuming a maximum standard error (P_u=0.5):

$E = 1.96\sqrt{(.5)(.5)/N}$ ⇨ $\sqrt{N} = 1.96 * .5/E$

⇨ $N = \left[\frac{1.96 * 0.5}{E}\right]^2 \approx \frac{1}{E^2}$

Sample size rule of thumb for proportions

$$N = \left[\frac{1.96 * 0.5}{E}\right]^2 \approx \frac{1}{E^2}$$

- With public opinion surveys, one wants to be precise (at least + .03)
  - “A referendum will gain 53% of votes +/- 5%” - ??

E **(Margin of Error)**	**# Cases**
± .10	100
± .05	400
± .03	1000
± .01	10,000