Basic Statistical Inference for Survey Data

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

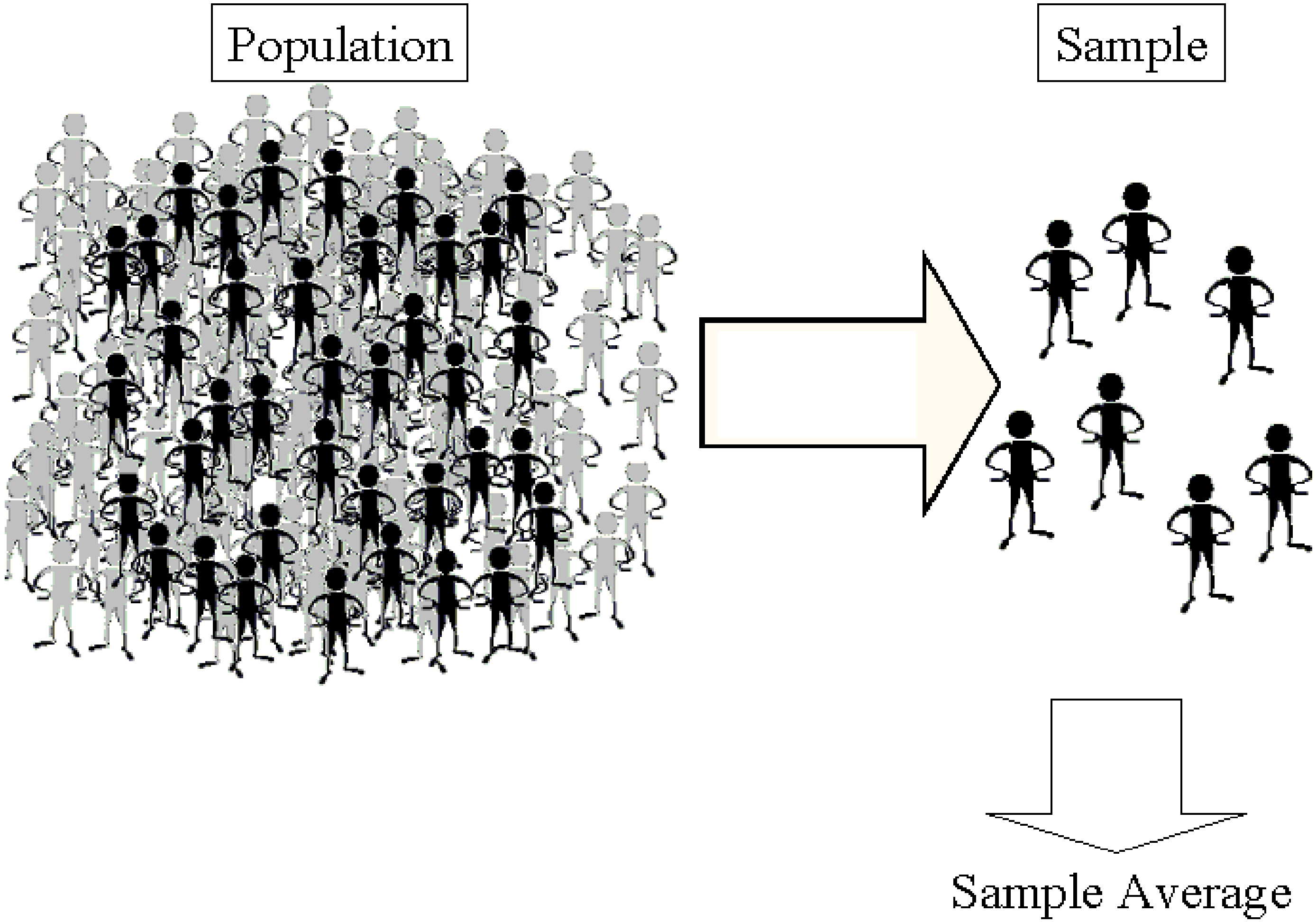
Goals for this Lecture

- Review of descriptive statistics
- Review of basic statistical inference
  - Point estimation
  - Sampling distributions and the standard error
  - Confidence intervals for the mean
  - Hypothesis tests for the mean
- Compare and contrast classical statistical assumptions to survey data requirements
- Discuss how to adapt methods to survey data with basic sample designs

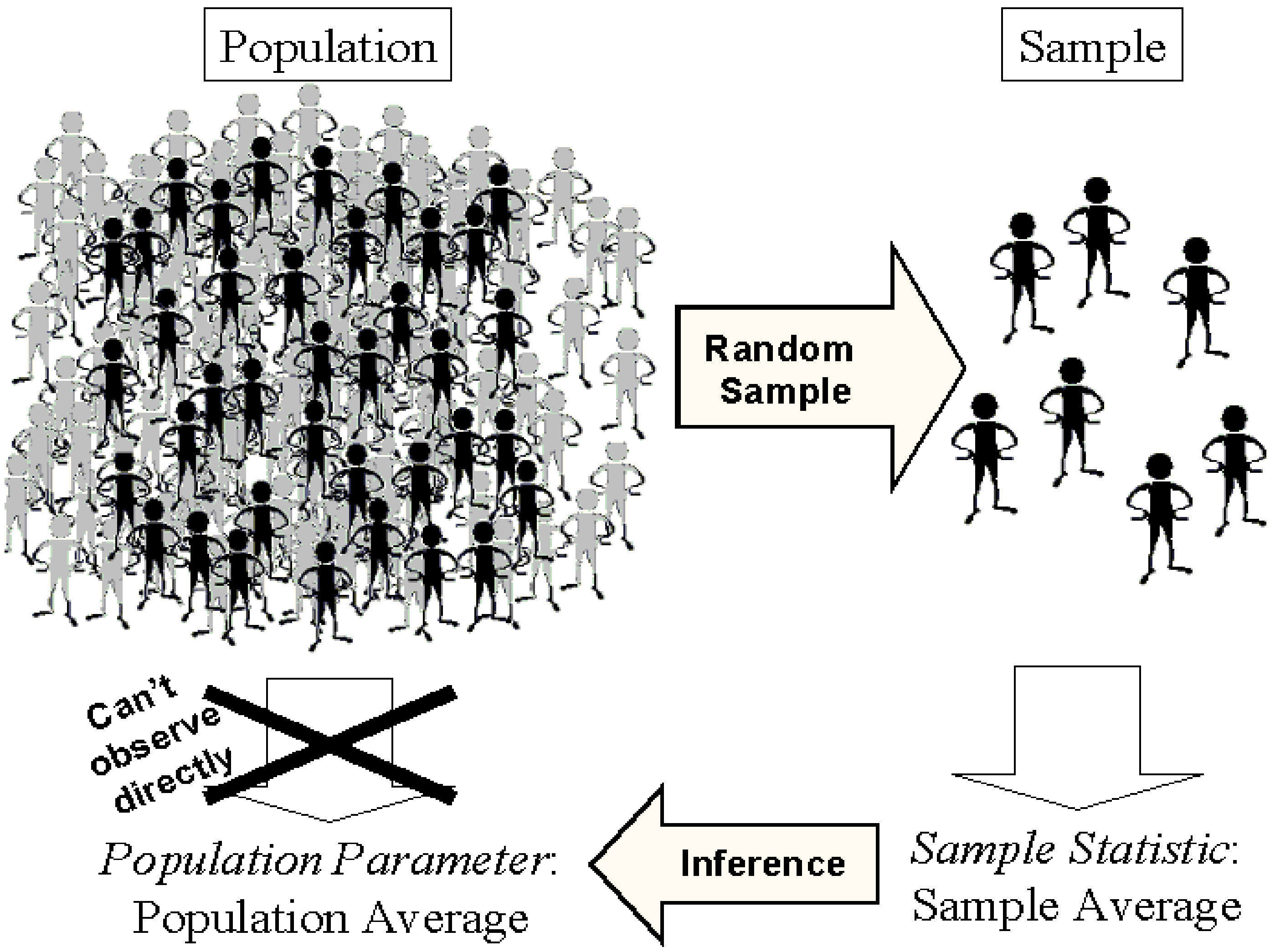
Two Roles of Statistics

- Descriptive: Describing a sample or population
  - Numerical: (mean, variance, mode)
  - Graphical: (histogram, boxplot)
- Inferential: Using a sample to *infer* facts about a population
  - Estimating (e.g., estimating the average starting salary of those with systems engineering Master's degrees)
  - Testing theories (e.g., evaluating whether a Master's degree increases income)
  - Building models (e.g., modeling the relationship of how an advanced degree increases income)

A Descriptive Statistics Question: *What was the average survey response to question 7?*

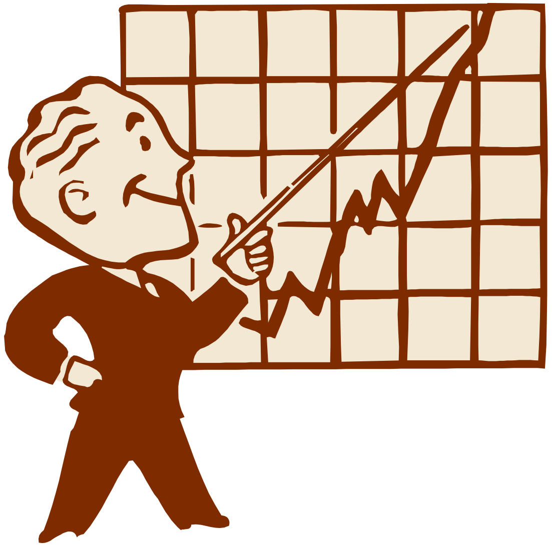


An Inferential Question: *Given the sample, what can we say about the average response to question 7 for the population?*



Lots of Descriptive Statistics

- Numerical:
  - Measures of location
    - Mean, median, trimmed mean, percentiles
  - Measures of variability
    - Variance, standard deviation, range, inter-quartile range
  - Measures for categorical data
    - Mode, proportions



- Graphical
  - Continuous: Histograms, boxplots, scatterplots
  - Categorical: Bar charts, pie charts

Continuous Data: Sample Mean, Variance, and Standard Deviation

- Sample average or sample mean is a measure of location or central tendency:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample variance is a measure of variability

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Standard deviation is the square root of the variance

$$s = \sqrt{s^2}$$

Statistical Inference

- Sample mean, $\overline{x}$, and sample variance, s^2, are statistics calculated from data
- Sample statistics used to estimate true value of population called estimators
- Point estimation: estimate a population statistic with a sample statistic
- Interval estimation: estimate a population statistic with an interval
  - Incorporates uncertainty in the sample statistic
- Hypothesis tests: test theories about the population based on evidence in the sample data

Classical Statistical Assumptions vs. Survey Practice / Requirements

- Basic statistical methods assume:
  - Population is of infinite size (or so large as to be essentially infinite)
  - Sample size is a small fraction of the population
  - Sample is drawn from the population via SRS
- In surveys:
  - Population always finite (though may be very large)
  - Sample could be sizeable fraction of the population
    - “Sizeable” is roughly > 5%
  - Sampling may be complex

Point Estimation (1)

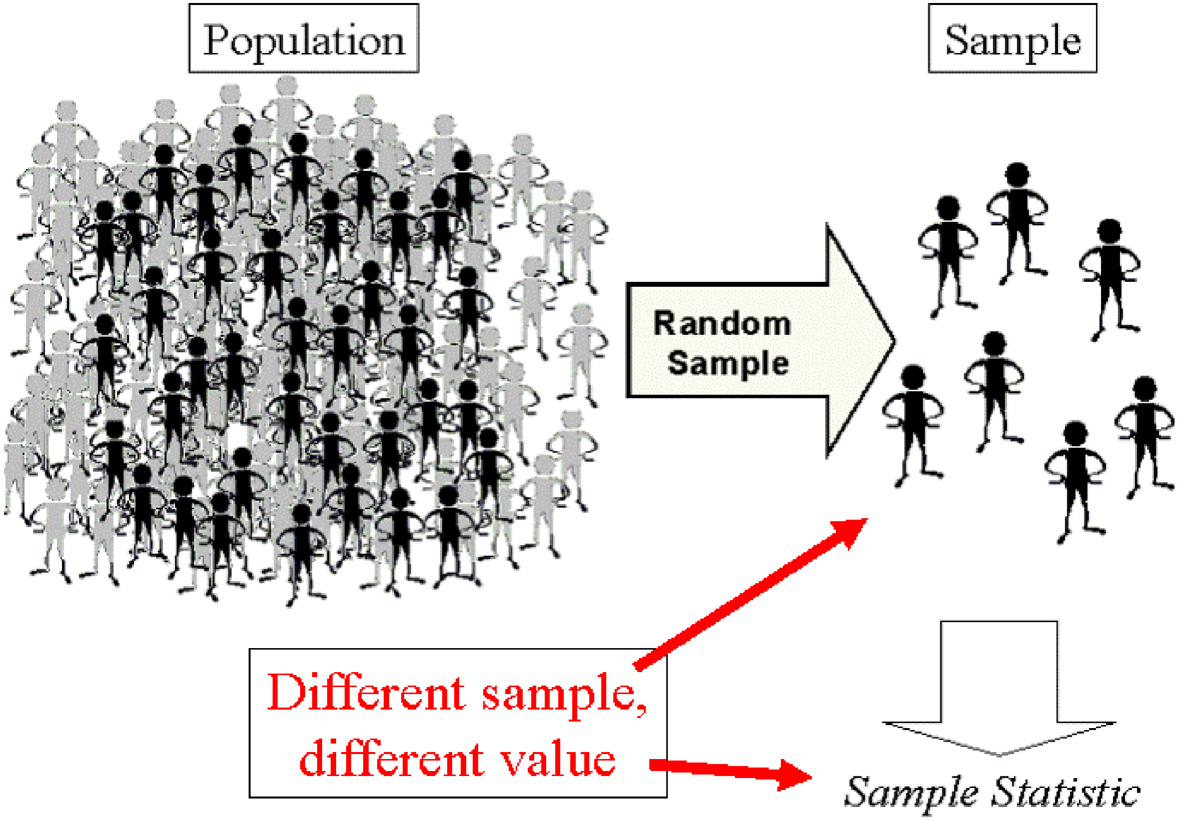
- Example: Use sample mean or proportion to estimate population mean or proportion
- Using SRS or a self-weighting sampling scheme, usual estimators <u>for the mean</u> calculated in all stat software packages are generally fine
  - Assuming no other adjustments are necessary
    - E.g., nonresponse, poststratification, etc
- Except under SRS, usual point estimates for standard deviation almost always wrong

Point Estimation (2)

- Naïve analyses just present sample statistics for the means and/or proportions
  - Perhaps some intuitive sense that the sample statistics are a measure of the population
  - But often don’t account for sample design
- However, when using point estimates, no information about sample uncertainty provided
  - If you did another survey, how much might its results differ from the current results?
- Also, even for mean, if sample design not self-weighting, need to adjust software estimators

Sampling Distributions

- Abstract from people and surveys to random variables and their distributions



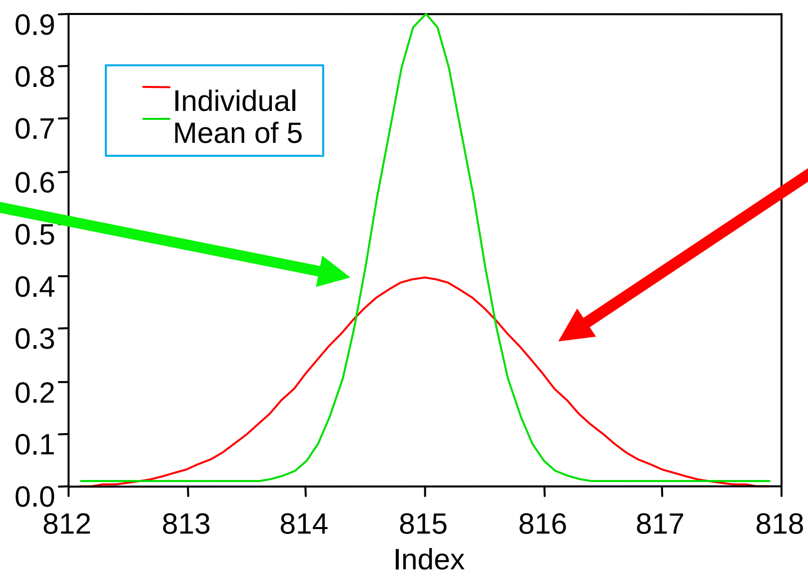
Sampling Distributions

- Sampling distribution is the probability distribution of a sample statistic

Sampling distribution of means of five obs

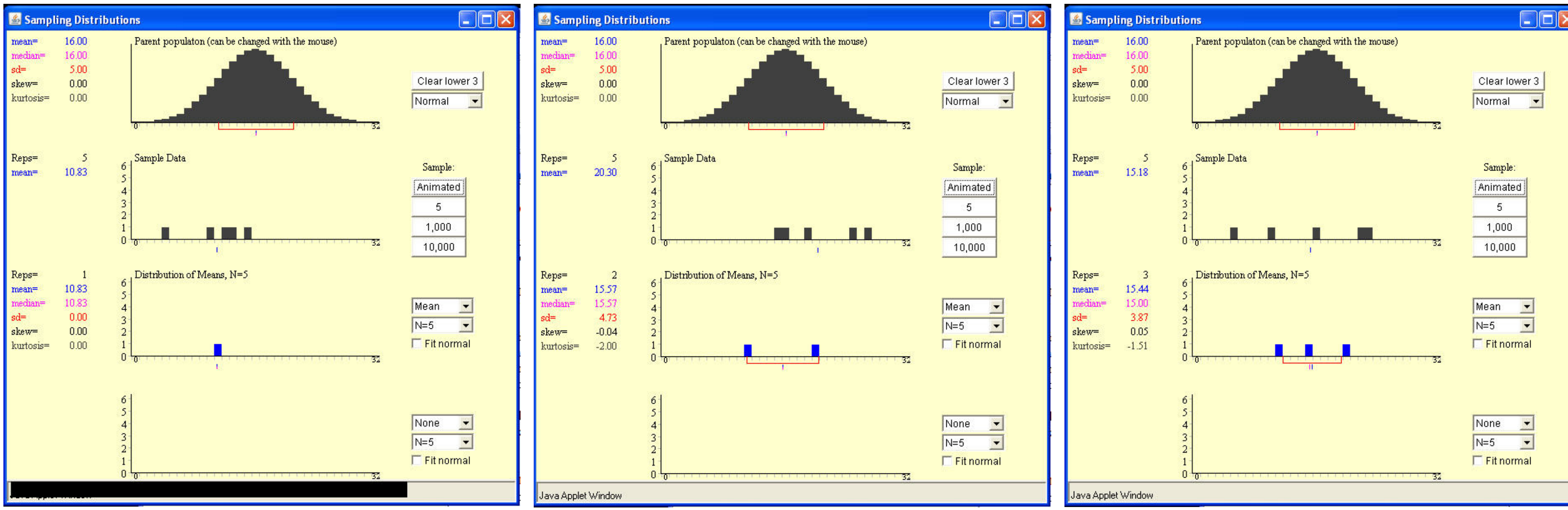
Standard error:

$$\sigma_{\bar{X}} = \sigma / \sqrt{5}$$



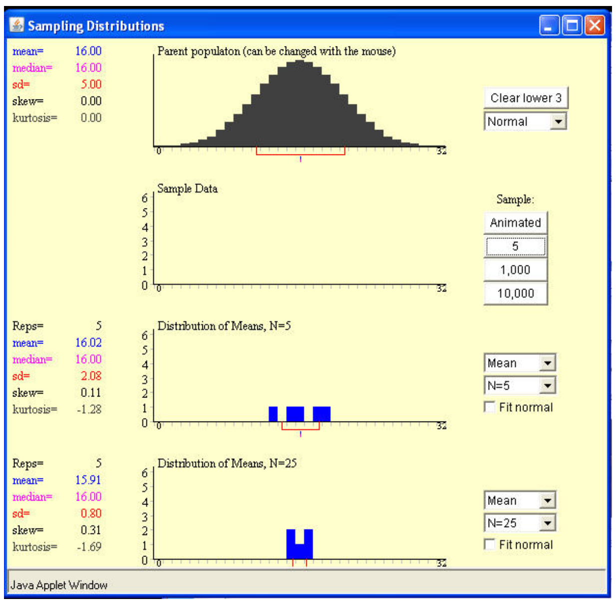
Distribution individual obs with standard deviation σ

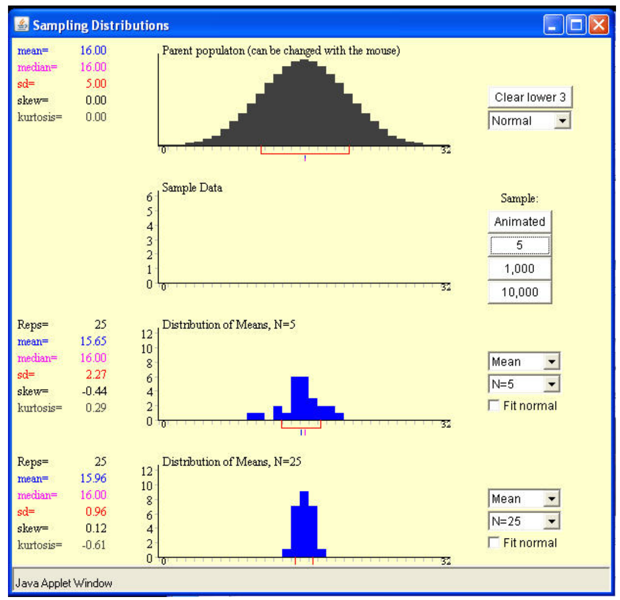
Demonstrating Randomness

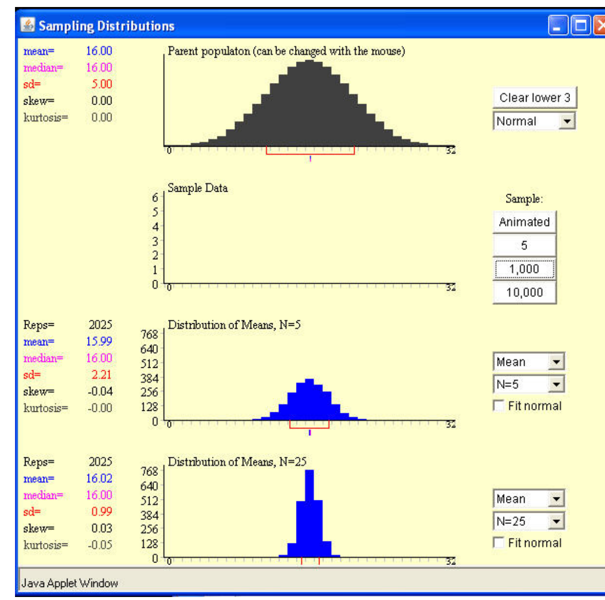


http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

Simulating Sampling Distributions





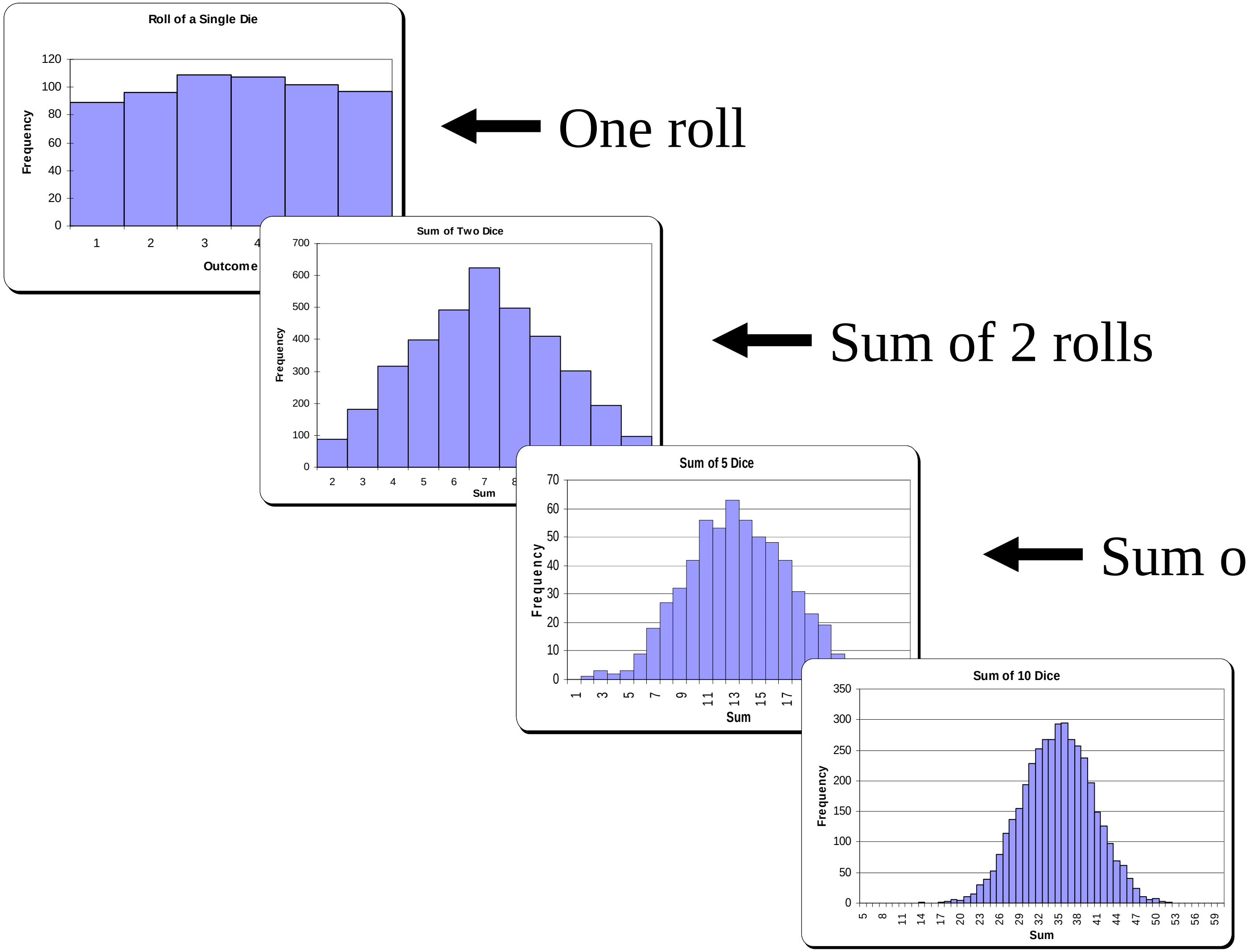


http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

Central Limit Theorem (CLT) for the Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and standard deviation σ
- For large sample size n, the distribution of the sample mean has approximately a normal distribution
  - with mean μ, and
  - standard deviation $\sigma/\sqrt{n}$
- The larger the value of n, the better the approximation

Example: Sums of Dice Rolls



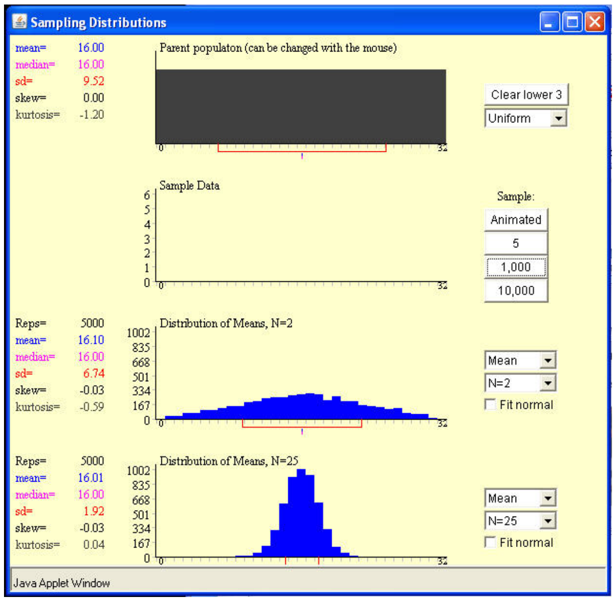
One roll

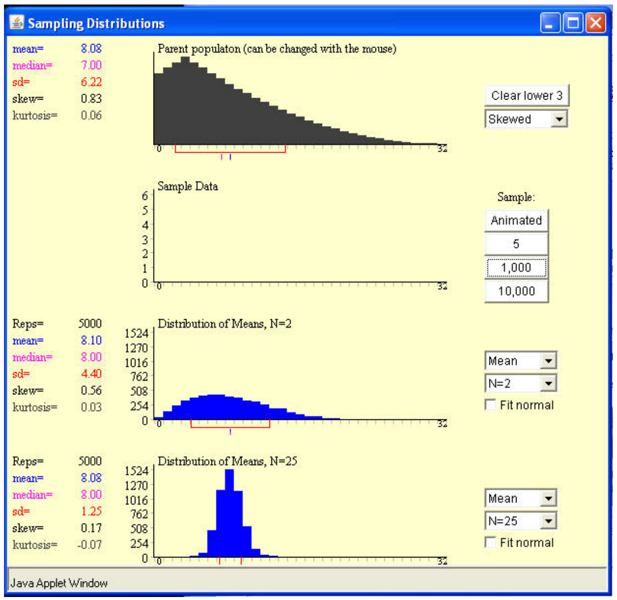
Sum of 2 rolls

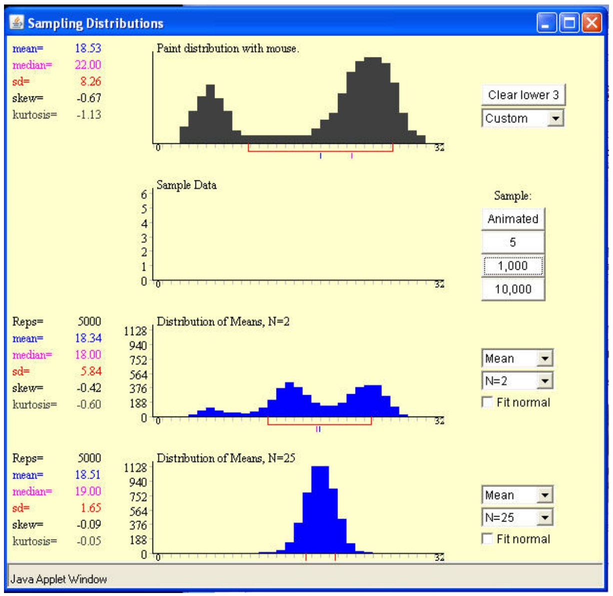
Sum of 5 rolls

Sum of 10 rolls

Demonstrating Sampling Distributions and the CLT







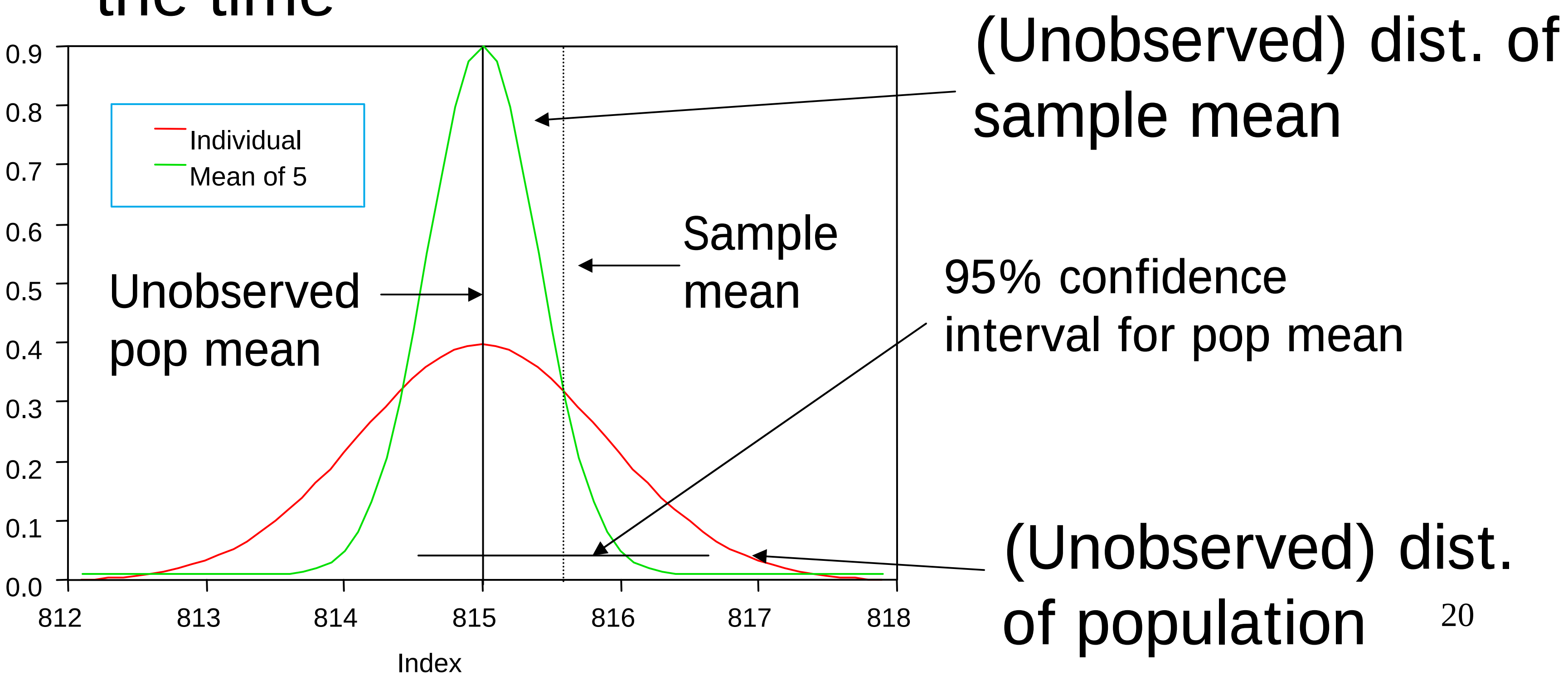
http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

Interval Estimation for μ

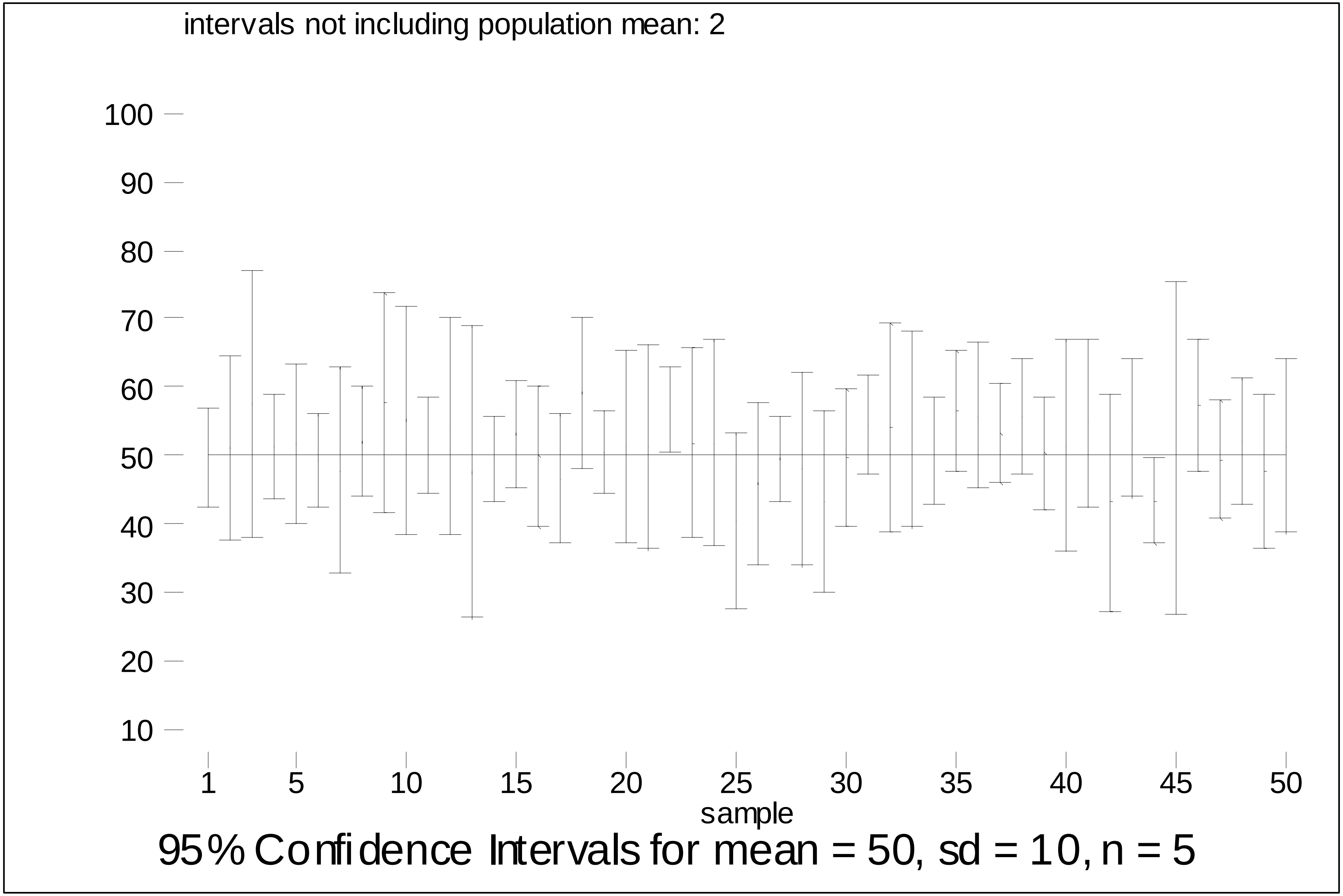
- Best estimate for μ is $\overline{X}$
- But $\overline{X}$ will never be *exactly* μ
  - Further, there is no way to tell how far off
- BUT can estimate μ's location with an interval and be right some of the time
  - Narrow intervals: higher chance of being wrong
  - Wide intervals: less chance of being wrong, but also less useful
- AND with confidence intervals (CIs) can define the probability the interval "covers" μ!

Confidence Intervals: Main Idea

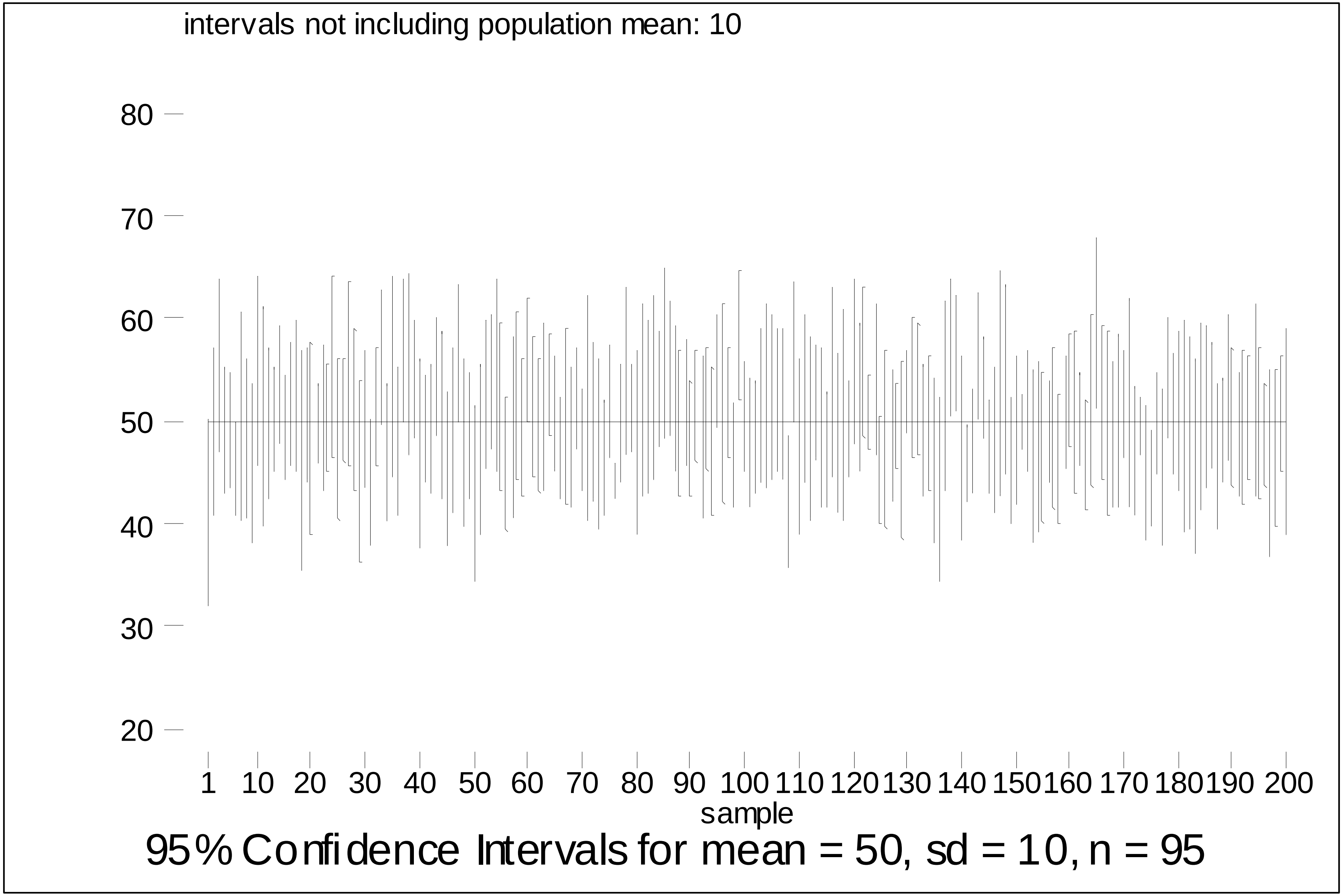
- Based on the normal distribution, we know $\bar{X}$ is within 2 s.e.s of μ 95% of the time
- Alternatively, μ is within 2 s.e.s of $\bar{X}$ 95% of the time



A Simulation



Another Simulation



Confidence Interval for μ (1)

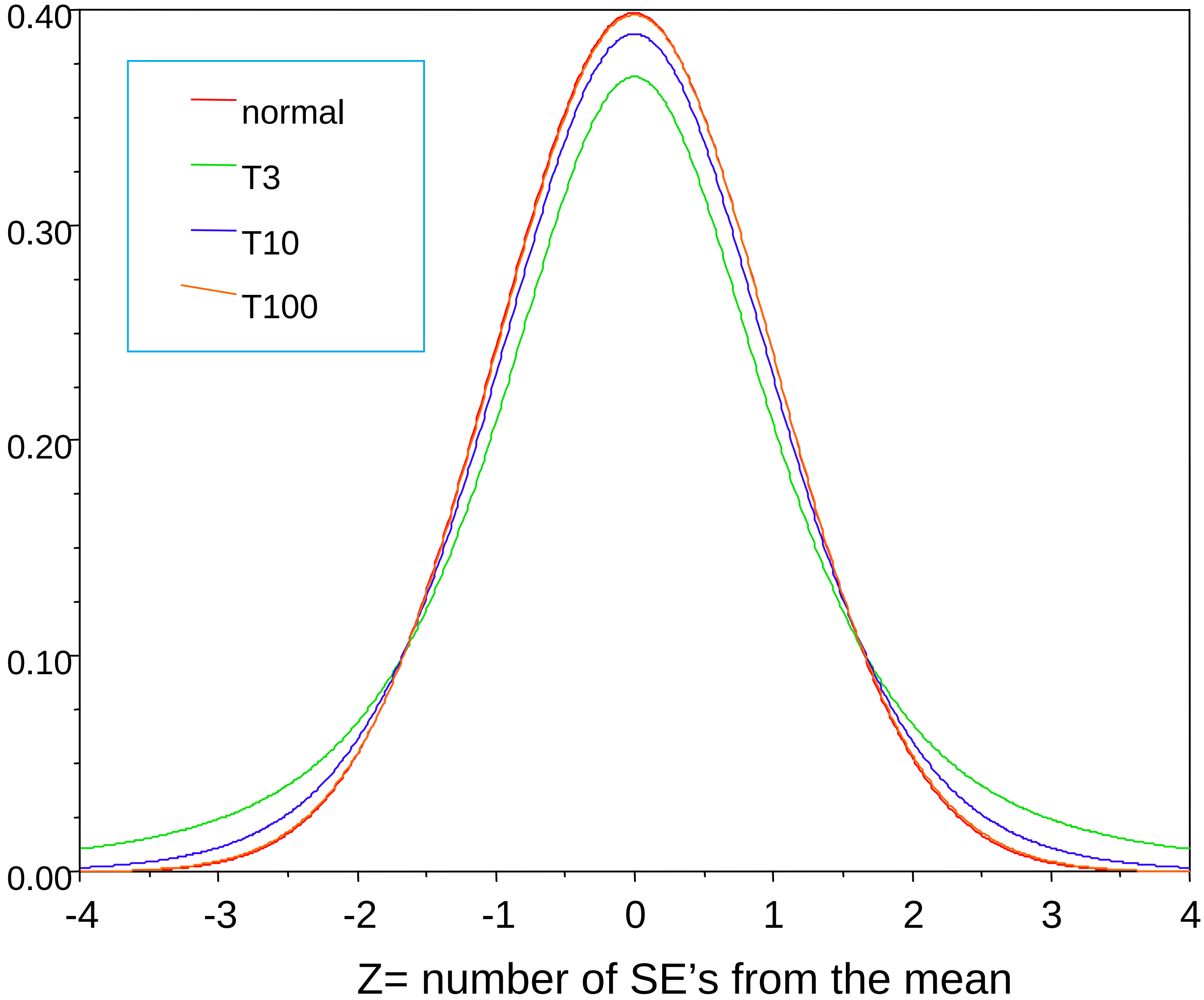
- For $\bar{X}$ from sample of size n from a population with mean μ,

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

  has a *t* distribution with n-1 "degrees of freedom"
  - Precisely if population has normal distribution
  - Approximately for sample mean via CLT
- Use the t distribution to build a CI for the mean:

$$\Pr\left(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

Review: the *t* Distribution



Confidence Interval for μ (2)

- Flip the probability statement around to get a confidence interval:

$$(1)\quad \Pr\left(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\right) = 1-\alpha$$

$$(2)\quad \Pr\left(-t_{\alpha/2,n-1} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1-\alpha$$

$$(3)\quad \Pr\left(\bar{X}-t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{X}+t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

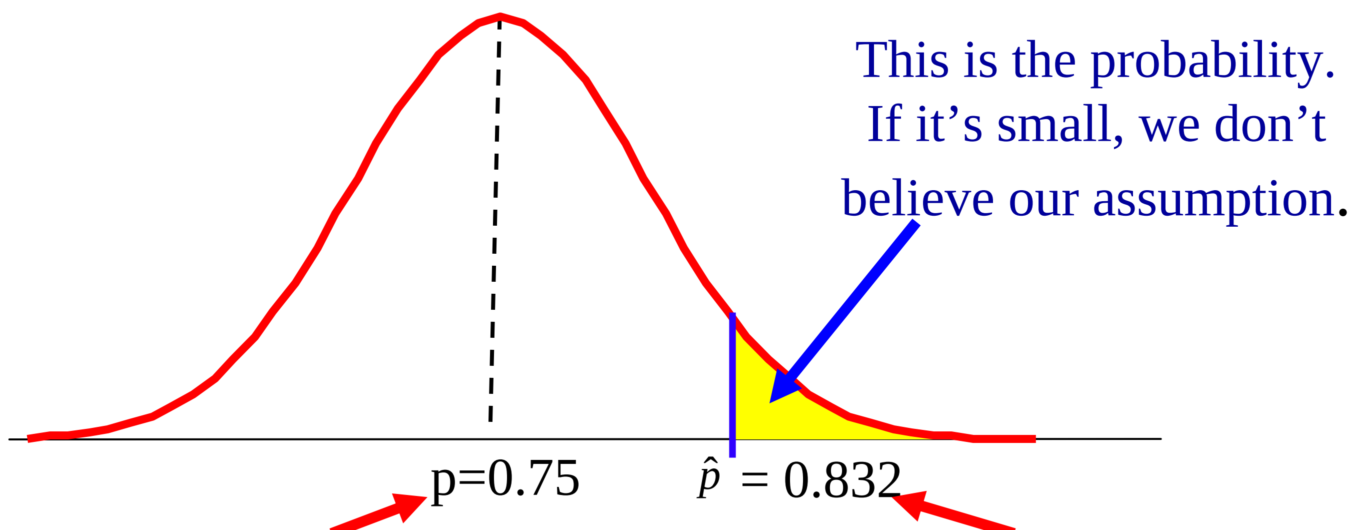
Example: Constructing a 95% Confidence Interval for μ

- Choose the confidence level: 1-α
- Remember the degrees of freedom (ν) = n -1
- Find $t_{\alpha/2,n-1}$
  - Example: if α = 0.05, df=7 then $t_{0.025,7}$ = 2.365
- Calculate $\bar{X}$ and $s/\sqrt{n}$
- Then

$$\Pr\left(\bar{X} - 2.365\frac{s}{\sqrt{n}} < \mu < \bar{X} + 2.365\frac{s}{\sqrt{n}}\right) = 0.95$$

Hypothesis Tests

- Basic idea is to test a hypothesis / theory on empirical evidence from a sample
  - E.g., “The fraction of new students aware of the school discrimination policy is less than 75%.”
  - Does the data support or refute the assertion?



If we assume this is true, how likely are we to see this (or something more extreme)?

One-Sample, Two-sided t-Test

- Hypothesis:

  H$_0$: $\mu = \mu_0$

  H$_a$: $\mu \neq \mu_0$

- Standardized test statistic: $t = \dfrac{\bar{X} - \mu_0}{\sqrt{s^2 / n}}$

- p-value = Pr(T<-t and T>t) = Pr(|T|>t), where T follows a t distribution with n-1 degrees of freedom

- Reject H$_0$ if $p < \alpha$, where α is the predetermined significance level

One-Sample, One-sided t-Tests

- Hypotheses:

  $H_0: \mu = \mu_0$ or $H_0: \mu = \mu_0$

  $H_a: \mu < \mu_0$ $\quad$ $H_a: \mu > \mu_0$

- Standardized test statistic: $t = \dfrac{\bar{X} - \mu_0}{\sqrt{s^2 / n}}$

- p-value = $\Pr(T < t)$ or p-value = $\Pr(T > t)$, depending on H_a, where T follows a t distribution with n-1 degrees of freedom

- Reject H_0 if $p < \alpha$

Applying Continuous Methods to Binary Survey Questions

- In surveys, often have binary questions, where desire to infer proportion of population in one category or the other
- Code binary question responses as 1/0 variable and for large n appeal to the CLT
  - Confidence interval for the mean is a CI on the proportion of “1”s
  - T-test for the mean is a hypothesis test on the proportion of “1”s

Applying Continuous Methods to Likert Scale Survey Data

- Likert scale data is inherently categorical
- If willing to make assumption that "distance" between categories is equal, then can code with integers and appeal to CLT

☐ Strongly agree	→	1
☐ Agree	→	2
☐ Neutral	→	3
☐ Disagree	→	4
☐ Strongly disagree	→	5

Adjusting Standard Errors (for Basic Survey Sample Designs)

- Sample > 5% of population: finite population correction
  – Multiply the standard error by $\sqrt{(N-n)/N}$
  – E.g. $$s.e.(\bar{x}) = \sqrt{(N-n)/N} \times s/\sqrt{n}$$

- Stratified sample, weighted sum of the strata variances:

$$s.e.(\bar{x}) = \sqrt{\sum_{h=1}^{H} \left(N_h/\mathrm{N}\right) \mathrm{Var}(\bar{x}_h)}$$

What We Have Just Reviewed

- Review of descriptive statistics
- Review of basic statistical inference
  - Point estimation
  - Sampling distributions and the standard error
  - Confidence intervals for the mean
  - Hypothesis tests for the mean
- Compare and contrast classical statistical assumptions to survey data requirements
- Discuss how to adapt methods to survey data with basic sample designs