Module 5D: Unsupervised Learning

A smattering of options: PCA, permutations, bootstrap

Review of traditional Methods

Hypothesis testing

Possible Null distributions:

- Binomial
- X^2
- Normal
- Poisson
- F
- student’s t

t-test
  One sample
  Paired
  Two Sample

ANOVA

Regression

Correlation

X^2 GOF

X^2 Contingency

Sign test

Mann-Whitney U

Kruskal-Wallis test

Spearman

Why use nonparametric tests at all?

Nonparametric tests **always** have less power than their parametric counterpart because you **always throw out information** by using only rank (and not magnitude): **type II error > β**

Why use nonparametric tests at all then?

- When used correctly, a nonparametric test should give a real Type I error rate = α

This seems kinda lame, right?

- But if you used a parametric test in its place (which would be using the parametric inappropriately since it doesn't meet the requirements), the parametric test will give a **type I error > α**

	Parametric	Nonparametric
Assumptions not met	Type I $> \alpha$	Type I $= \alpha$
Assumptions met	Type II $= \beta$	Type II $> \beta$

ACTUAL: indicated by Type I, Type II

STATED: indicated by α, β

Other “Modern” Statistics Methods

But there are many biologically interesting phenomenon that are not easily described by the tools we have examined so far….

Sometimes there is no standard method

Computers have dramatically expanded the toolkit of statistics/research

Computational methods:

When assumptions of best method available can’t be met

Random sampling is still assumed

No standard method exists

Massive amount of calculations

When we don’t know the null distribution

Two major categories of computational methods

Null sampling distributions:

1. Simulation – hypothesis testing

2. Randomization/Permutation

Precision of estimates:

3. Bootstrapping – sampling distribution of estimate; the values for the parameter estimates that we might obtain and their probabilities.

Bootstrapping:

- ‘re-sampling’ the actual data
  - **Sampling with replacement**
  - Pick the original number of points for each group

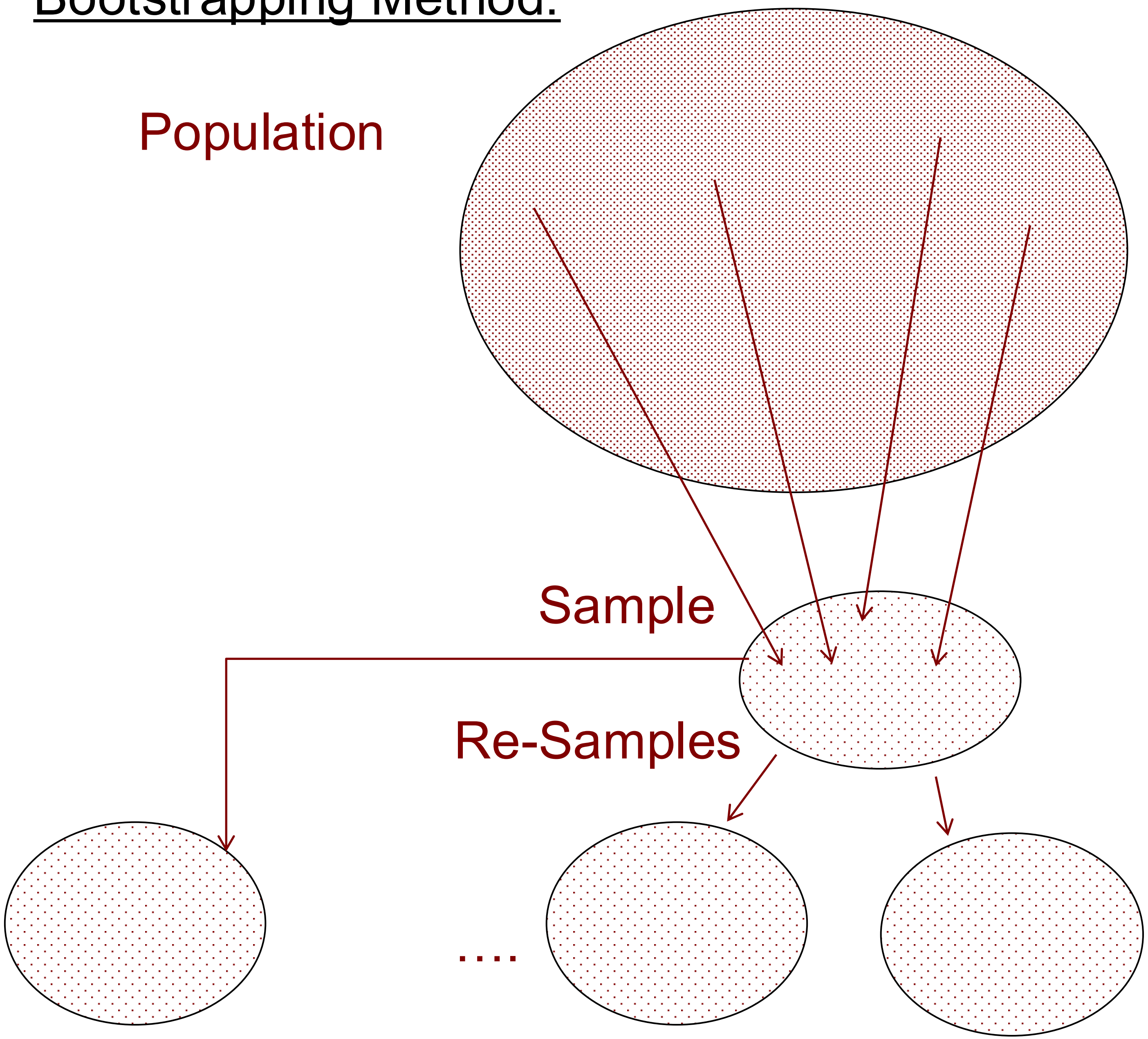
- Approximates the *sampling distribution* of an estimate
  - ***But NOT*** *the null (sampling) distribution as with simulation and randomization*

- Nonparametric and be applied to virtually any parameter – including means, proportions, correlations, linear model coefficients

- Used to find confidence interval and the bootstrap standard error
  - Precision method
  - Particularly useful when there is no ready formula for standard error (median, eigenvalue)

- Estimate uncertainty in phylogenies

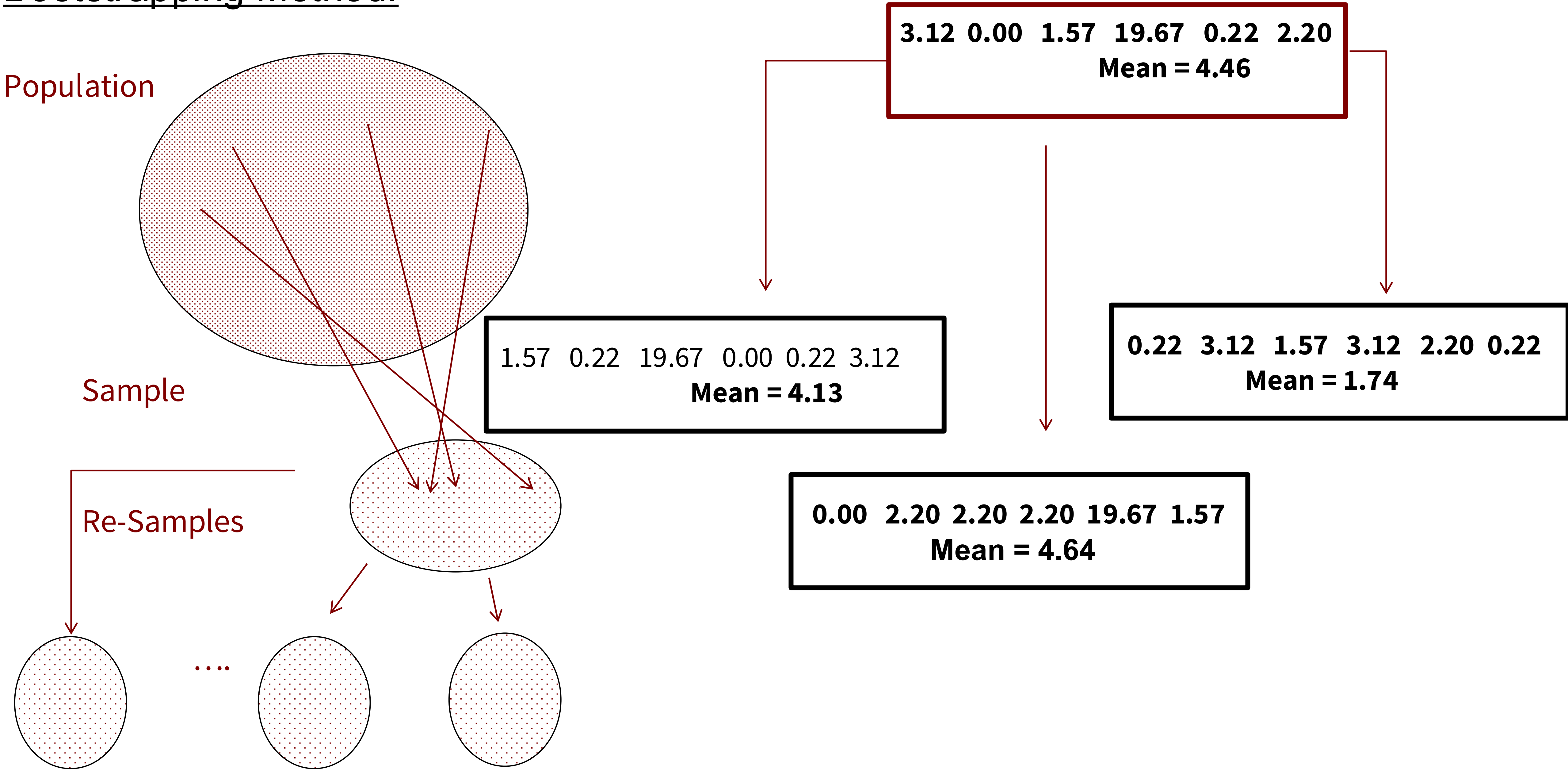
Bootstrapping Method:



Sample size: Large enough so that frequency distribution of sample is reasonable approximation of frequency distribution of population

Too small samples, result in standard errors that are too small and confidence errors are that are too narrow --> overestimate precision

Bootstrapping Method:



Population

Sample

Re-Samples

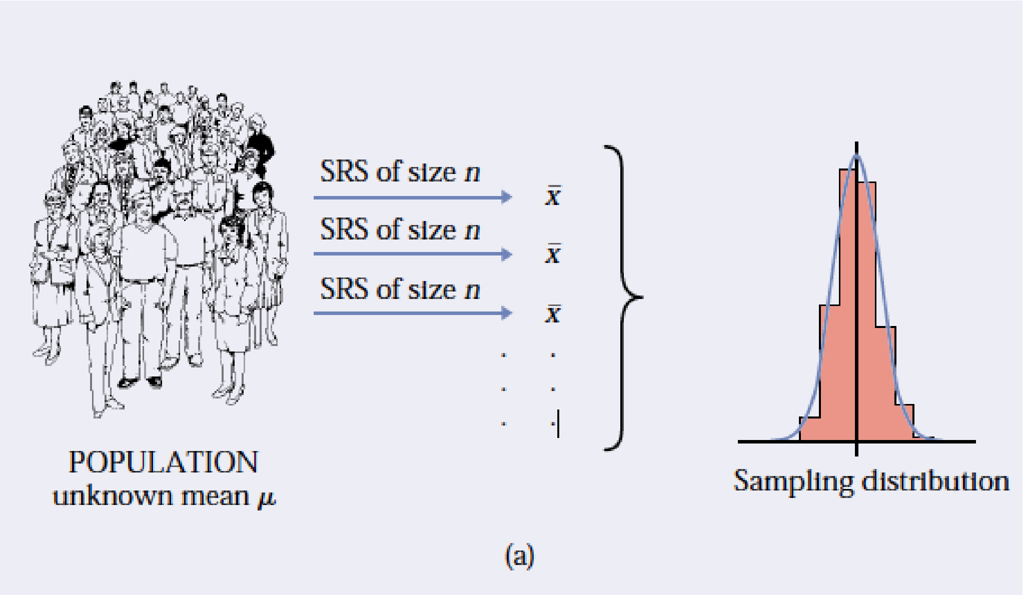
….

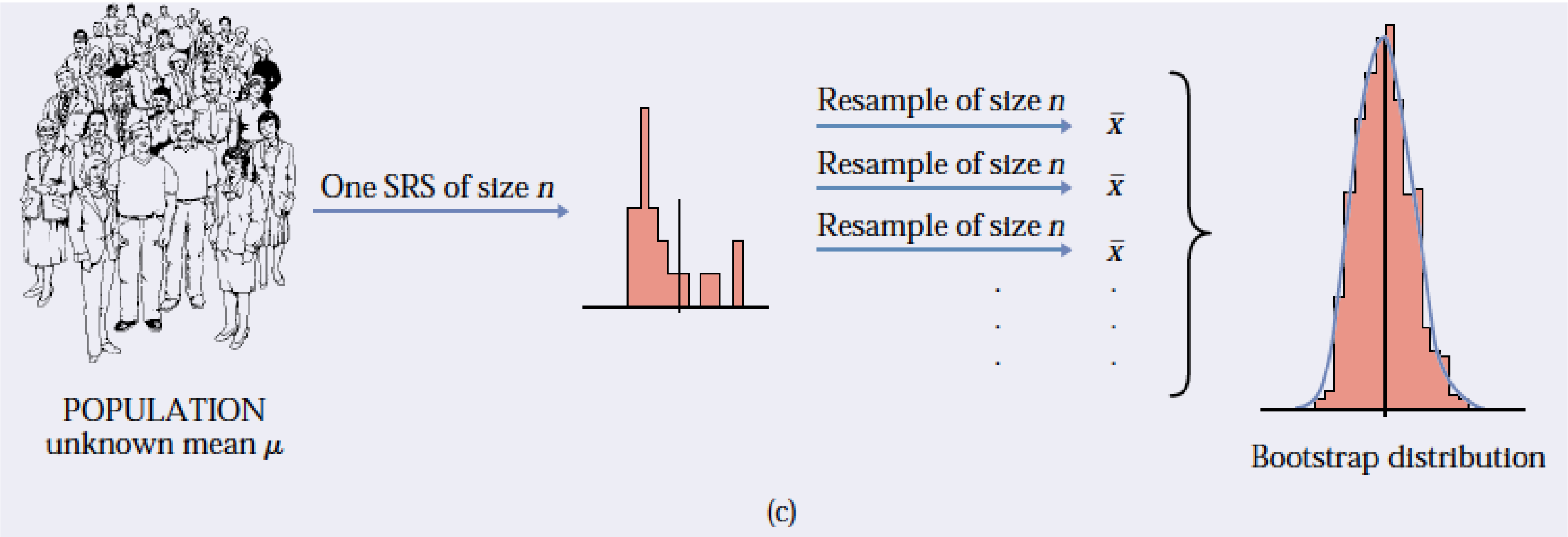
3.12 0.00 1.57 19.67 0.22 2.20
Mean = 4.46

1.57 0.22 19.67 0.00 0.22 3.12
Mean = 4.13

0.00 2.20 2.20 2.20 19.67 1.57
Mean = 4.64

0.22 3.12 1.57 3.12 2.20 0.22
Mean = 1.74





http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore14.pdf