Module 5D: Unsupervised Learning

A smattering of options: PCA, permutations, bootstrap

Important Note:

The remaining modules (5D, 5E, and 5F) are **not** "Unsupervised Methods", they are simply a 'grab bag' of three ubiquitous computational methods from the general realm of "**resampling methods**"

- They are important methods, conceptually and practically
- But I had nowhere else to put, so I appended them to the end of the course!
- We use these methods a lot in biology because:
  - Data are 'high dimensional'
  - n << columns/features
  - True data-generating processes are rarely known.

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

	Parametric	Nonparametric
Assumptions not met	**Type I > α**	**Type I = α**
Assumptions met	**Type II = β**	**Type II > β**

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

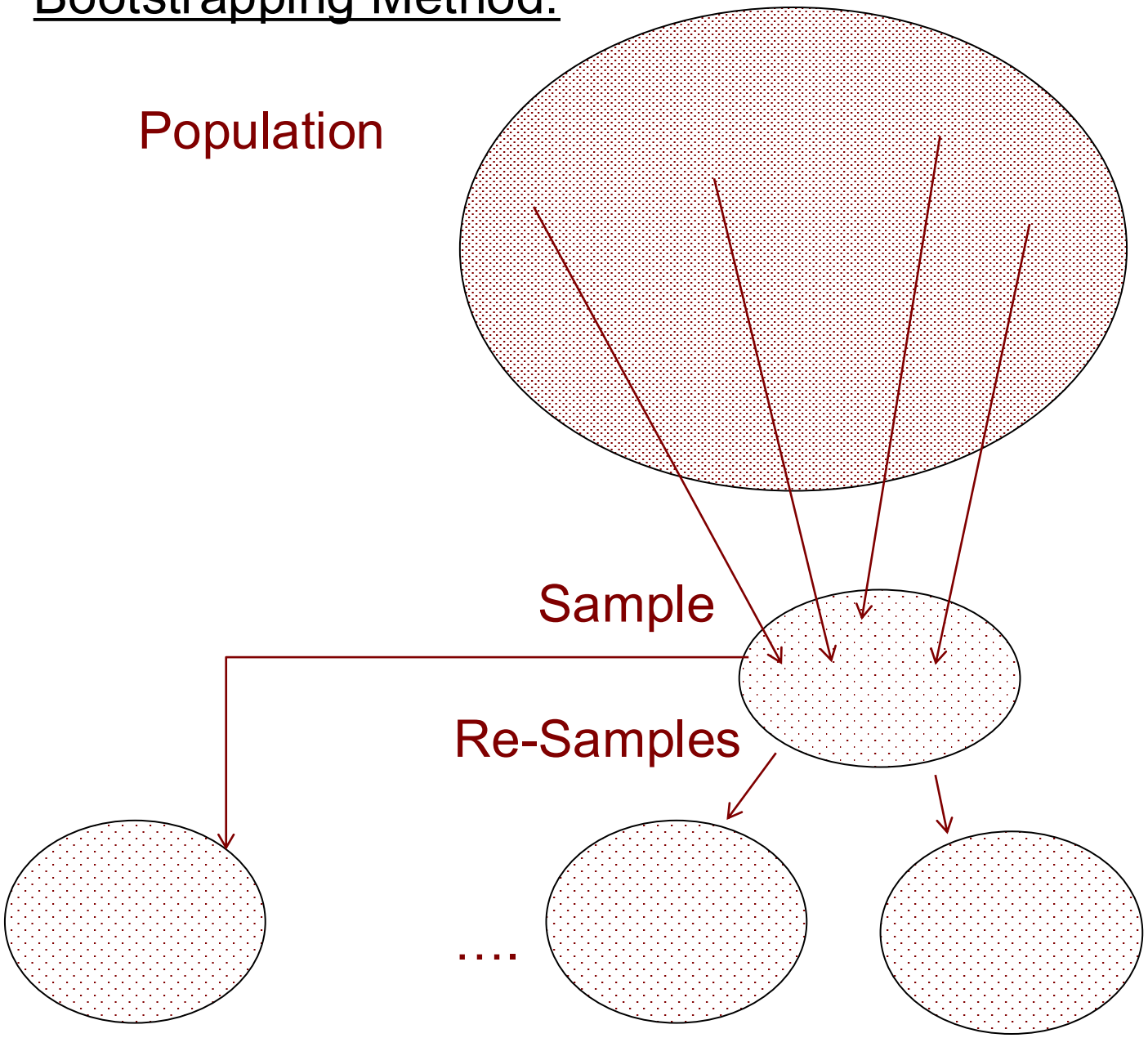
- Approximates the *sampling distribution* of an estimate
  - ***But NOT*** *the null (sampling) distribution as with simulation and randomization*

- 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)

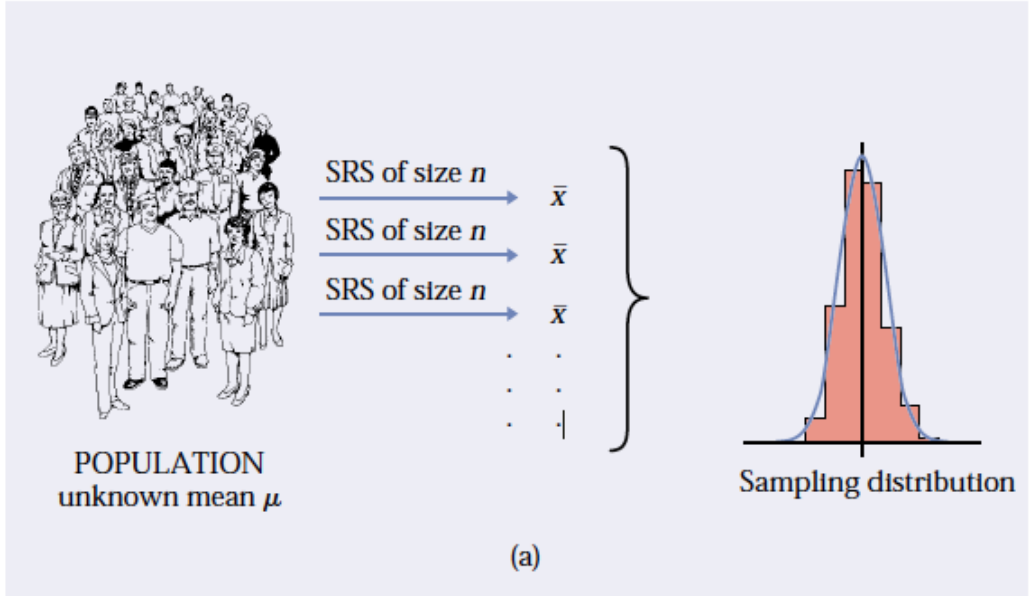
- Estimate uncertainty in phylogenies, in single-cell biology for cluster stability analysis
- Used in estimating “diversity indices” in ecology

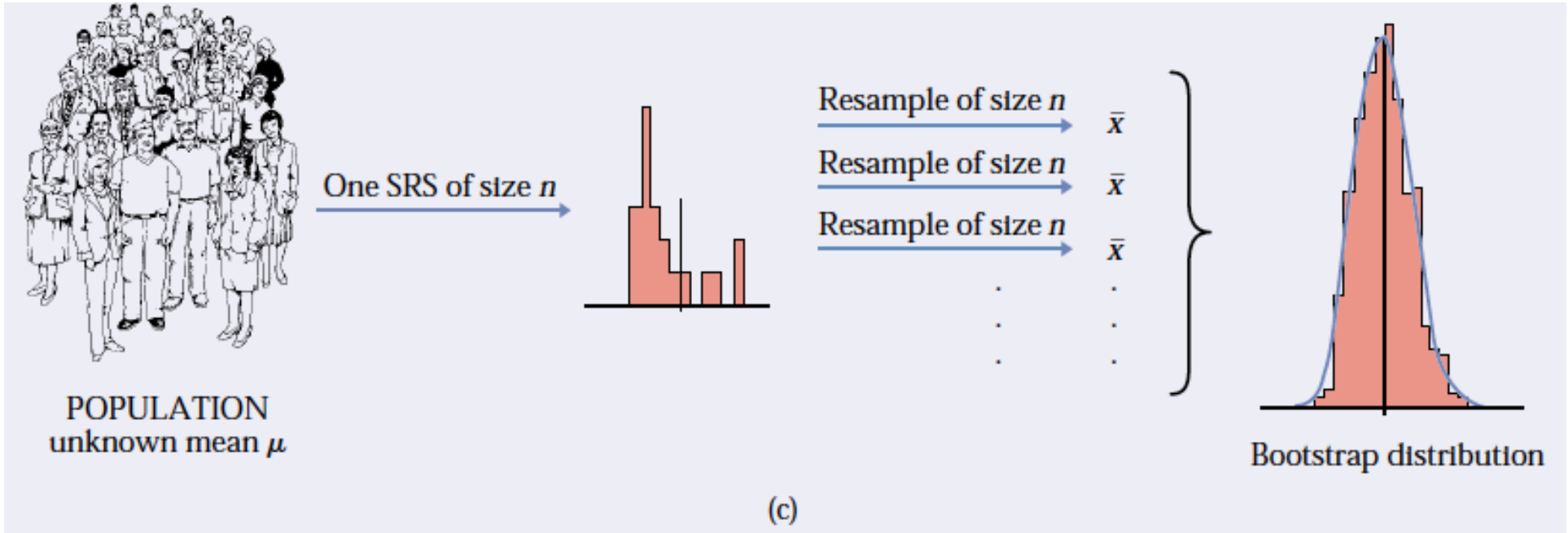
Bootstrapping Method:



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

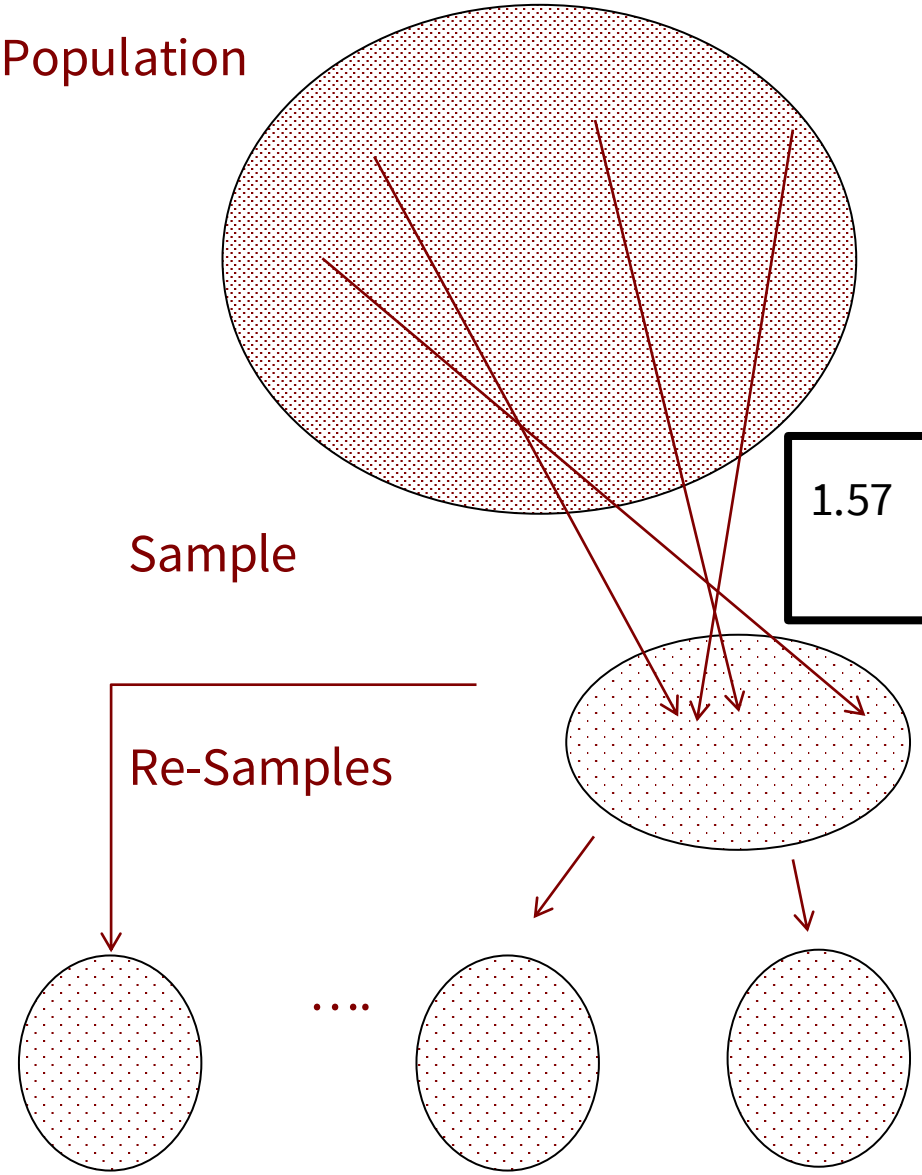
Too small samples, result in standard errors that are too small and confidence errors are that are too narrow --> overestimate precision





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

Bootstrapping Method:



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

0.22 3.12 1.57 3.12 2.20 0.22
Mean = 1.74

0.00 2.20 2.20 2.20 19.67 1.57
Mean = 4.64

Bootstrap Example:

Suppose you have aligned sequence data for 1000 sites for 5 species. Each column is an independent observation of evolutionary history. You use a method to build a phylogenetic tree. If you want to put confidence intervals on each branch of the tree --- you use **bootstrapping!**

Step 1: Resample sites with replacement

- Randomly draw 1000 columns from the original 1000 columns but ***with replacement*** so some columns are represented ≥1 and some columns are represented 0 times.
- This is the bootstrap alignment

Step 2: Rebuild the tree

- Construct a new tree

Step 3: Repeat many times

- Do this 1000 times -→ resulting in 1000 trees

You can then measure *how often each particular branching group/cluster appears to give a probability*

- Out of 1000 trees, how many contain the branch of interest given the trees were created with resampling?
- A measure of *persistence* of the relationship

Site	Species A	Species B	Species C	Species D	Species E
1	A	A	G	A	G
2	T	T	T	C	T
3	G	G	G	G	A
…	…	…	…	…	…
1000	C	C	T	C	T

Bootstrap Example:

Suppose you have aligned sequence data for 1000 sites for 5 species. Each column is an independent observation of evolutionary history. You use a method to build a phylogenetic tree. If you want to put confidence intervals on each branch of the tree --- you use **bootstrapping!**