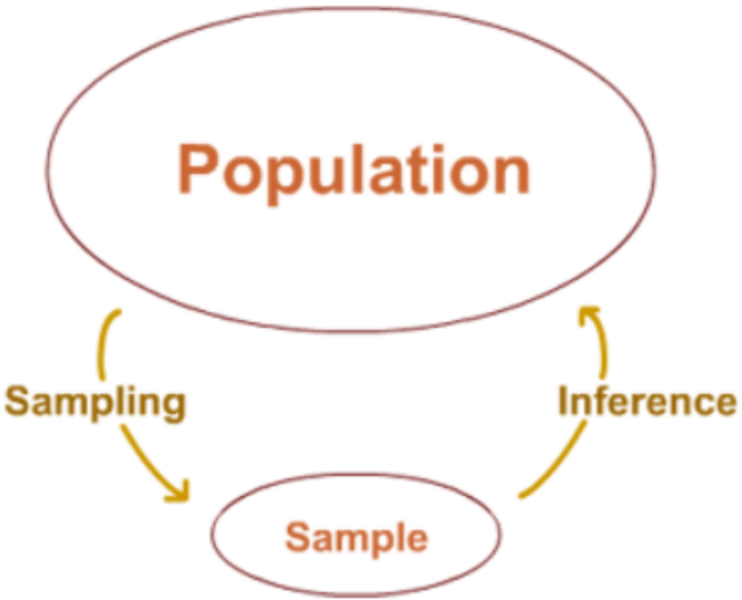
Sampling

By: Jeferson Bisconde

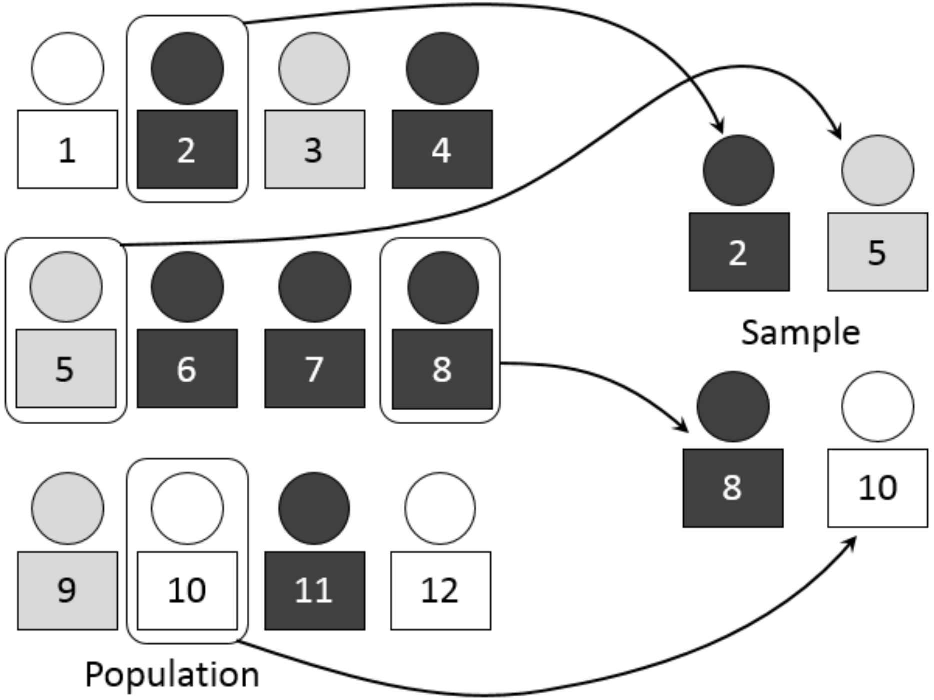
Data Discovery

- Question / Hypothesis
- Experiment
- Collect and Analyze Data
- Check Results
- Repeat / Redesign



Random Sampling

- every member is given equal opportunities of being selected



Sampling Methods

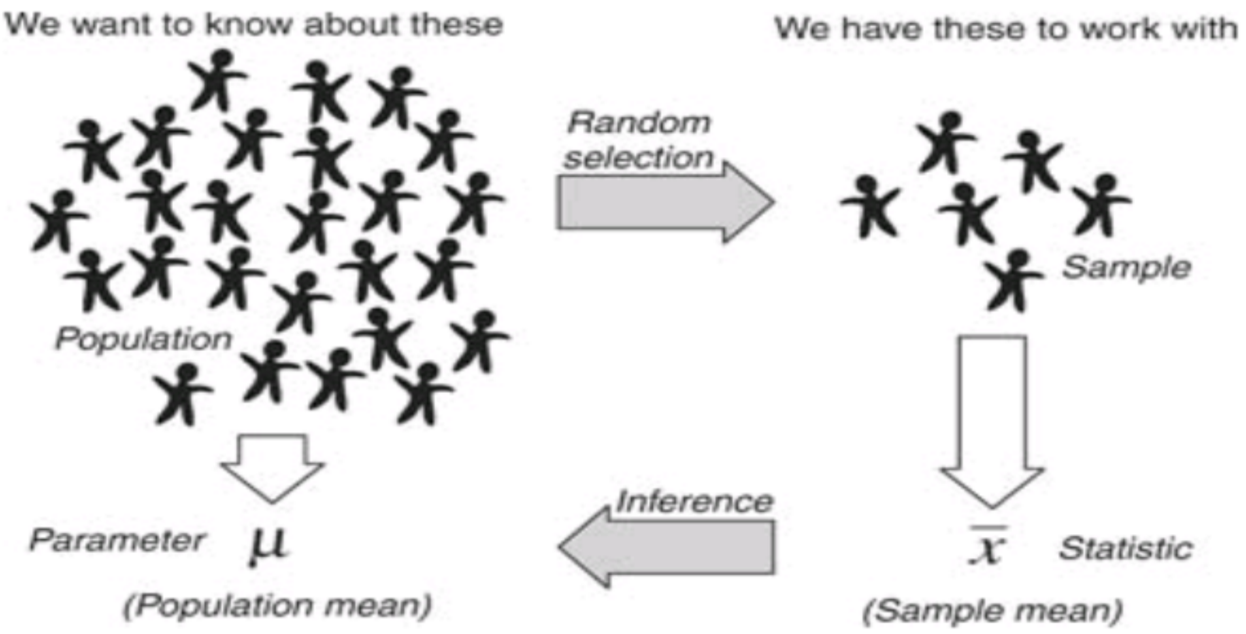
- Simple Random Sampling (SRS)
  - easiest and most widespread

- Other common methods:
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling

Random Sampling & Assignment

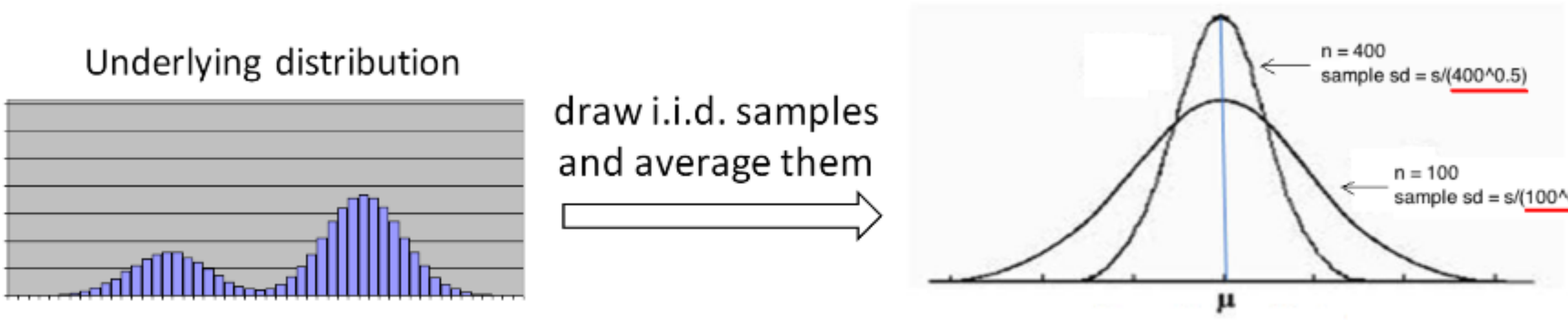
	Random assignment	No random assignment	
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
	Causation	Correlation	

Sampling and Inference



Central Limit Theorem (CLT)

- Given certain conditions
  - the mean will be approximately normal
    - regardless of the underlying distribution



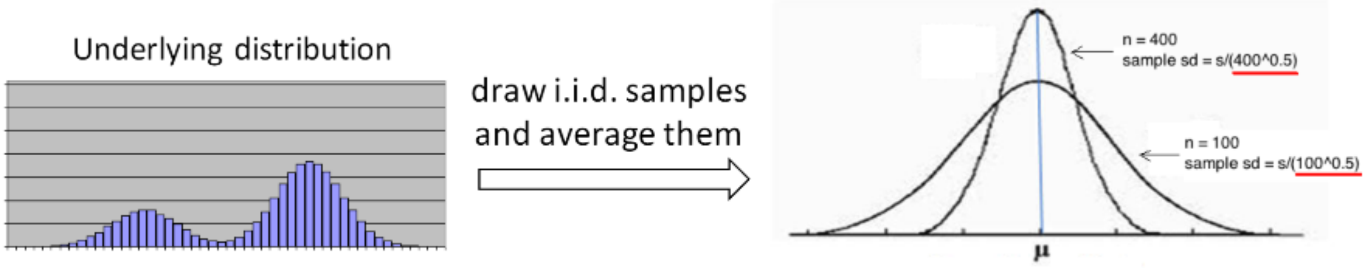
Central Limit Theorem (CLT)

- Not only is the sample mean normally distributed, we have....

$$\boxed{\bar{X} \sim Normal(\mu, \frac{\sigma^2}{n})}$$

- And as usual, from any normally distributed random variable, we can derive a standard normal variable.   In this case...

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$



Confidence Interval

- interval estimate of a population parameter

- stated at 95% CI
  - can be 50%, 90% or 99%

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) \quad \text{or} \quad \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Confidence Interval - cont

- if σ is not known
- and if N > 30, then

$$\overline{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

- When N is small

$$\overline{\mathrm{x}} \pm \mathrm{t}_{(\alpha/2,\mathrm{n}-1)} \frac{s}{\sqrt{n}}$$

Resampling

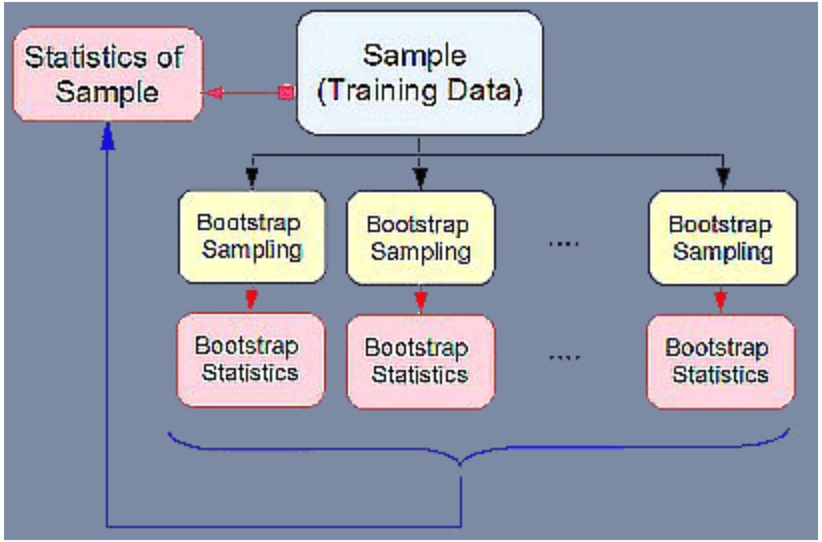
- drawing repeated samples from the data

- Common techniques:
  - Bootstrapping
  - Jackknifing
  - Cross-validation
  - Permutation tests

Bootstrapping

- Estimates the sampling distribution
  - sampling with replacement from original sample

- used to estimate standard errors and confidence intervals of a population parameter



Bootstrap Variance Estimation

Draw $X_1^*, \ldots, X_n^* \sim \hat{F}_n$

Compute $\hat{\theta}^* = t(X_1^*, \ldots, X_n^*)$

Repeat steps 1 and 2, B times, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$

Let $v_{boot} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \frac{1}{B}\sum_{r=1}^{B}\hat{\theta}_r^*)^2$

$(\hat{se}_{boot} = \sqrt{v_{boot}})$

Bootstrap Confidence Interval

- Percentile method

$$C_n = (\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$$

- The Normal interval

$$\hat{\theta} \pm z_{\alpha/2}\hat{se}_{boot}$$

When to Bootstrap?

- Theoretical distribution is complicated or unknown
- sample size is too small
- estimating the variance of a statistic
  - small pilot sample for power calculations