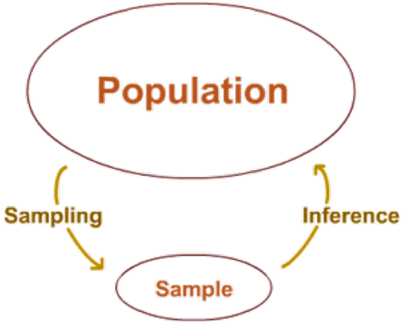
Sampling

Clayton W. Schupp

Galvanize

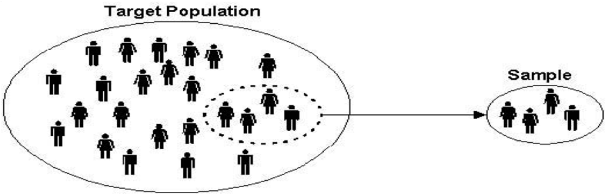
Statistical Inference Process in General



- Start with a question
- Design an experiment
- Collect data
- Analyze the data
- Check the results
- Repeat? Redesign?

Collecting Data

A sample should be representative of the population

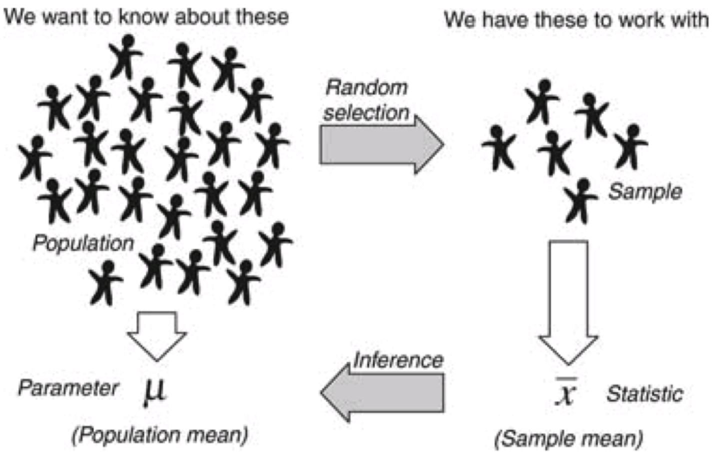


Drawing a random sample from the population is the best way to achieve this

Random Sampling Methods

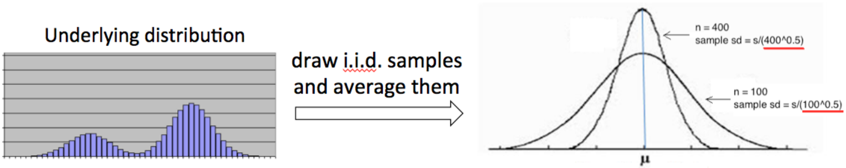
- Simple Random Sampling
  - Each subject has an equal chance of being part of the sample
  - The easiest form of random sampling
- Other random sampling methods
  - Stratified sampling
  - Cluster sampling
  - Systematic sampling

Sampling and Statistical Inference



Central Limit Theorem

The CLT states that given certain conditions, the mean of a sufficiently large number of *i.i.d* random variables will be approximately normal, regardless of the underlying distribution



Central Limit Theorem

- Not only is the sample mean normally distributed, but the variance of the sample mean is smaller

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

- As with any normal variable, we can derive a standard normal Z-score

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

Confidence Interval Estimate

- A confidence interval (CI) is an interval estimate of a population parameter
- The typical level of confidence is 95%, but they can be calculated for any level
- For example, a 95% CI for the population mean is given by

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Confidence Interval Estimates

In reality we don't know the population standard deviation σ

- If sample size is sufficiently large ($n > 30$), we can substitute the sample standard deviation s for it in the previous formula

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

- However if n is small, we need to use the t-distribution with degrees of freedom (df) equal to $n - 1$

$$\bar{x} \pm t_{(\alpha/2, df)} \frac{s}{\sqrt{n}}$$

Bootstrap Sampling

Estimates the sampling distribution of an estimator by sampling with replacement from the original sample

Advantages:

- Completely automatic
- Requires no theoretical calculations
- Not based on asymptotic results
- Available regardless of how complicated the estimator might be

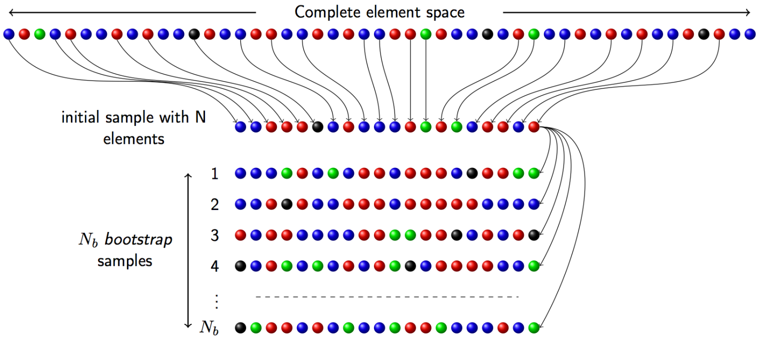
Often used to estimate the standard errors and confidence intervals of a unknown population parameter

Bootstrap Sampling

Method:

- Start with your dataset of size n
- Sample from your dataset with replacement to create 1 bootstrap sample of size n which means many of the observations will be repeated
- Repeat B times
- Each bootstrap sample can then be used as a separate dataset for estimation or model fitting

Bootstrap Sampling



Bootstrap Variance

- Draw a bootstrap sample
$$X_1^*, \ldots, X_n^*$$
- Calculate bootstrap estimate of your parameter:
$$\hat{\theta}^* = t(X_1^*, \ldots, X_n^*)$$
- Repeat steps 1 and 2, B times to get
$$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$$
- Calculate
$$s_{boot}^2 = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\boldsymbol{\theta}}^*)^2$$

where $\hat{\boldsymbol{\theta}}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$

Bootstrap Confidence Intervals

- Percentile Method

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

- Interval assuming approximately normal bootstrap sampling distribution

$$\hat{\boldsymbol{\theta}}^* \pm 1.96 s_{boot}$$