Sampling

Afternoon Objectives

This afternoon, we will:

- Review population inference and sampling
- Discover the Central Limit Theorem (CLT)
- Apply the Central Limit Theorem to construct Confidence Intervals for the mean of a population
- Use Bootstrapping to construct Confidence Intervals for any population statistic

Population Inference and Sampling

Statistical Discovery in General

1. Start with a question/hypothesis
2. Design an experiment
3. Collect data (Sampling)
4. Analyze data (Estimation/Inference)
5. Repeat? Redesign?

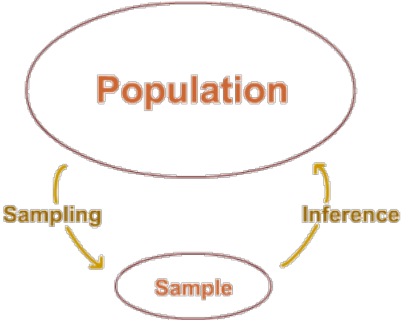

Figure 1:Sampling and Inference

Sampling and Statistical Inference

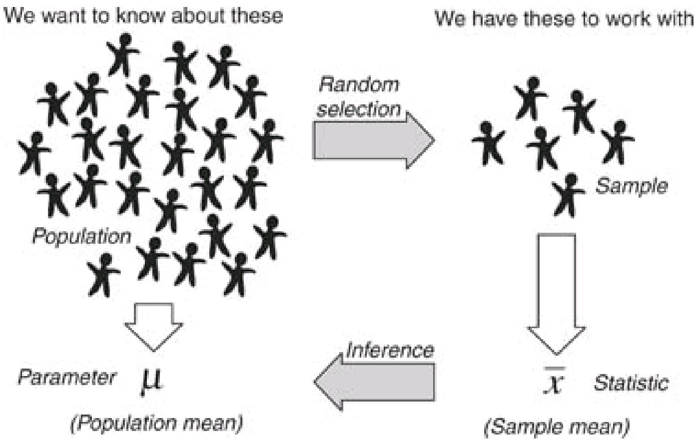

Figure 2:Sampling and Statistical Inference

Collecting Data: Make sure you have good data!

- Your results are only as good as your data. Garbage in, garbage out
- Your sample should be representative of the population

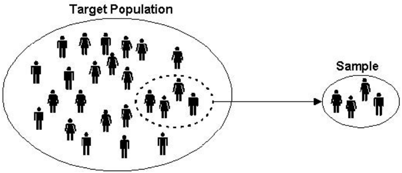

Figure 3:Sampling

- 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

Random sampling is harder than it sounds...

Scenario: You want to estimate the percentage of dog owners in SF.

- Method 1:
  - Go to the nearby dog park and ask random people if they own dogs until you have n responses
- Method 2:
  - Stand on 24th and Mission and ask random people if they own dogs until you have n responses
- Method 3:
  - Repeat n times: Pick a random neighborhood in SF (weighted by census data per neighborhood), go to that neighborhood, ask random people you see if they own dogs until you get 1 response

Random sampling in the digital age...

You might think that random sampling in a digital context is easier, and you're right! But there are still gotchas.

Scenario: Slack is testing a new feature ("channel polling", a way to survey people in a channel). They'd like to test the feature on only a subset of their users (n), then draw inference about their entire userbase.

- Method 1:
  - SELECT user_id FROM users LIMIT n;
- Method 2:
  - SELECT user_id FROM users ORDER BY RAND() LIMIT n;

Random sampling... just do the best you can

Often it's impossible to do perfect random sampling. So:

1. Do the best you can,
2. Call out possible objections, and
3. Make a case for why you think your results are valid

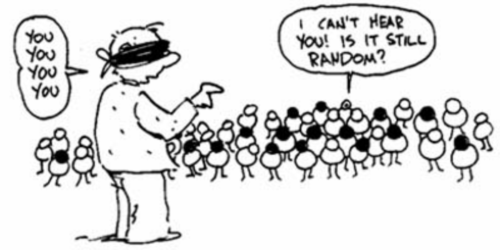

Figure 4:Random Sampling

Central Limit Theorem

Central Limit Theorem (CLT)

One of the most important results in classical statistical inference is the *Central Limit Theorem* (CLT) which says that if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2 then their mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$

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

Central Limit Theorem (CLT)

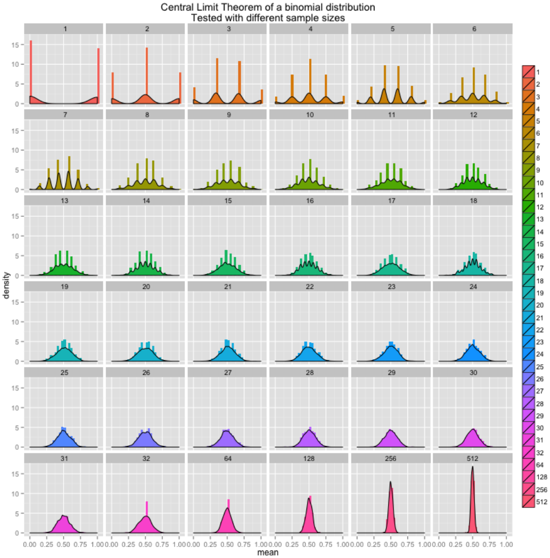

Figure 5:Central Limit Theorem on a Binomial Distribution

Central Limit Theorem (CLT)

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

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

Central Limit Theorem (CLT) - Example

Recall that $B(n, p)$ is the sum of n independent Bernoulli trials with parameter p. This means that

$$B(n, p) \sim N\left(np, np(1-p)\right)$$

Why?

Central Limit Theorem (CLT) - Example

- If X follows a Bernoulli distribution with parameter p,

we have: $\begin{cases} \mu = p \\ \sigma^2 = p(1-p) \end{cases}$

- According to the CLT:

$$\bar{X} \sim N\left(\underset{p}{\cancel{\mu}}, \frac{\cancel{\sigma^2}\; p(1-p)}{n}\right)$$

Central Limit Theorem (CLT) - Example

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- $Y = n\bar{X}$ (the sum of n Bernoulli trials)

$$\Leftrightarrow E[Y] = n \underbrace{E[\bar{X}]}_{p}$$

$$Var[Y] = n^{\cancel{2}} \underbrace{Var[\bar{X}]}_{\frac{p(1-p)}{\cancel{n}}}$$

$$\boxed{Y \sim N(np, np(1-p))}$$

Confidence Intervals

Confidence Intervals

A *confidence interval* is an interval estimate of the true parameter of your population

- An α confidence interval is an interval centered around estimated parameter which contains the true value of that parameter with *confidence* α (α is usually 99%, 95%, or 90%)
- In other words, if you resample or rerun the experiment many times, α percent of the time the true value will be in the computed confidence interval
- It is *not* a statement that the true value of the parameter is contained in the interval with a certain probability (the true value is the interval or it isn't)

Confidence Intervals

For example, a 95% confidence interval for the population mean is given by

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

Why?

Confidence Intervals

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

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

$$(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}})$$

- However if n is small, the central limit theorem does not guarantee normality and we need to use instead the t-distribution

$$(\bar{x} - t_{(\alpha/2, n-1)}\frac{s}{\sqrt{n}}, \bar{x} + t_{(\alpha/2, n-1)}\frac{s}{\sqrt{n}})$$

Breakout

Using Python, sample 100 times from a normal distribution.

- Compute the sample mean and a 95% confidence interval
- Is the true mean in your interval?
- Rerun your code several time and see if you find an interval which doesn't contain the true mean

Bootstrapping

Bootstrapping

Also called bootstrap sampling. Another way to generate confidence intervals for a population parameter is through a process called bootstrapping

Simple idea:

- Sample from your observed data *with replacement* B times
- With these B samples, compute the statistic (i.e., mean, median, variance, etc.) of interest and then estimate the sample variance

Bootstrapping

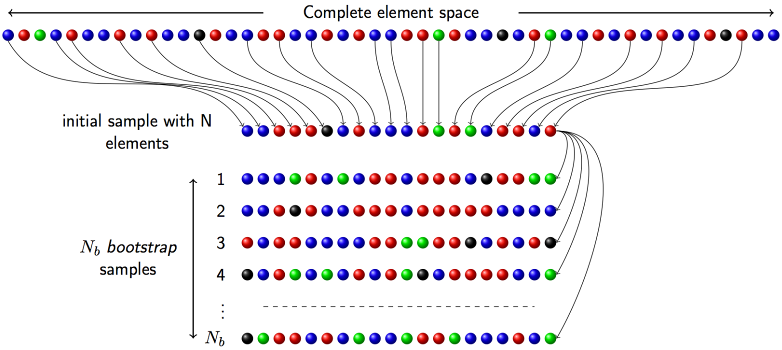

Figure 6:Bootstrapping

Bootstrapping

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

Bootstrapping

Method:

- Start with n i.i.d. samples $X_1, \ldots, X_n$
- For i from 1 to B:
  1. Sample $X_1^*, \ldots, X_n^*$ with replacement from your data
  2. Compute the sample statistic of the parameter you're interested in $\hat{\theta}_i^* = g(X_1^*, \ldots, X_n^*)$
- Then compute the bootstrap variance, the sample variance of your statistic:

$$s^2_{bootstrap} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \bar{\theta}^* \right)^2 \text{ where } \bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

Bootstrap Confidence Intervals (Normal Interval)

There are a few different ways to build bootstrap confidence intervals that rely of differing assumptions. The first is the *normal interval*

- If your parameter is approximately normally distributed (like the mean of a sample with $n > 30$) your interval will be

$$\theta_n \pm z_{\alpha/2} s_{bootstrap}$$

where $\theta_n = g(X_1, \ldots, X_n)$ is your estimate of the parameter, z is standard normal (e.g., 1.96 for 95%)

Bootstrap Confidence Intervals (Percentile Method)

Let θ^*_β be the β sample quantile of your bootstrap sample statistics $(\theta^*_1, \ldots \theta^*_B)$.

Then an $1 - \alpha$ bootstrap percentile interval is

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

Why Bootstrap?

Why would we use bootstrapping over standard confidence intervals?

- The theorical distribution of the statistic is complicated or unknown (e.g., median or correlation)
- The sample size is too small for traditional methods
- Favor accuracy over computational cost

Questions

- What's boostrapping?
- When might I think of using it?
- What are the steps to setting up a boostrap estimate?

Questions

- What's boostrapping?
  - Random sampling with replacement technique
- When might I think of using it?
  - Want sense of accuracy of some sample estimate
- What are the steps to setting up a boostrap estimate?
  - Sample with replacement B times, compute B estimates from these B samples, get standard errors, confidence intervals, etc.