Week 6: Hypothesis Testing

1. Introduction to Hypothesis Testing

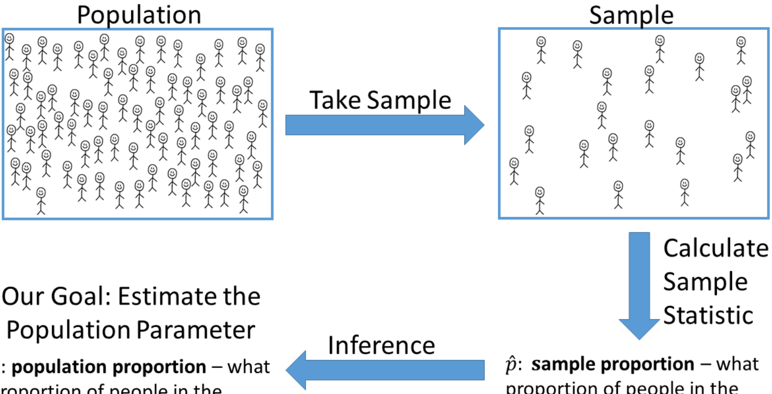
Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Estimation



Population

Take Sample

Sample

Calculate Sample Statistic

Our Goal: Estimate the Population Parameter

Inference

p: **population proportion** – what proportion of people in the **population** support the president?

$\hat{p}$: **sample proportion** – what proportion of people in the **sample** support the president?

A statistical test uses data from a sample to assess **a claim** about a population

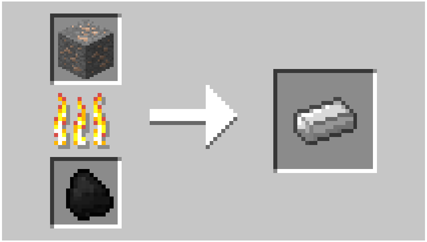
Example: a study on ingot

In factories, about 50% of aluminum ingots need to be recast because of cracking. In an attempt to reduce the cracking proportion, the plant engineers and chemists recently tried out a new method in the casting process.

Has the new method worked?

real ingot

making ingot in minecraft



Poll question

If the new method does not work, what proportion of ingots need to be recast?

a $p = 0\%$

b $p = 25\%$

c $p = 50\%$

d $p = 75\%$

e $p = 100\%$

In a recent production, 400 ingots have been cast and only 180 of them have cracked using the new method (this is 45%).

This provides

1. Strong evidence for the new method
2. Weak evidence for the new method
3. No evidence for the new method
4. Not sure

Randomness or evidence?

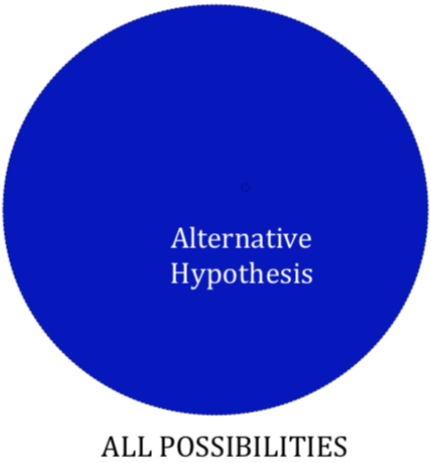
We know that each random sample will have a somewhat different proportion of cracked ingots. Is the 45% we observe merely a result of natural sampling variability, or is this lower cracking rate strong enough evidence to assure management that the true cracking rate now is really below 50%?

Statistical tests are framed formally in terms of two competing hypotheses:

Null Hypothesis (H_0): Claim that there is no effect or difference.

Alternative Hypothesis (H_a): Claim for which we seek evidence.

Statistical Hypothesis



Usually the null is a very specific statement

?
Null
Hypothesis

Can we reject the null hypothesis?

We want to know *if the changes made by the engineers have lowered the cracking rate from 50%*.

We'll make the hypothesis that the cracking rate is still 50% and see if the data can convince us otherwise.

$$H_0 : p = 0.5$$
$$H_a : p < 0.5$$

Helpful hints:

- H_0 usually includes $=$
- H_a usually includes $>, <,$ or $\neq$
- The inequality in H_a depends on the question

Think hypothesis testing as a jury trial

Does the reasoning of hypothesis tests seem backward?

That could be because we usually prefer to think about getting things right rather than getting them wrong.

You have seen this reasoning before because it's the logic of jury trials.

Step 1: null hypothesis
Let's suppose a defendant has been accused of robbery. The null hypothesis is that the defendant is innocent. Instructions to juries are quite explicit about this.

Step 2: collect data
How is the null hypothesis tested? The prosecution first collects evidence. ("If the defendant were innocent, wouldn't it be remarkable that the police found him at the scene of the crime with a bag full of money in his hand, a mask on his face, and a getaway car parked outside?") For us, the data is the evidence.

Step 3: judge the evidence

The jury considers the evidence in light of the *presumption* of innocence and judges whether the evidence against the defendant would be plausible *if the defendant were in fact innocent*.

Step 4: make a decision

The standard of "beyond a reasonable doubt" is wonderfully ambiguous because it leaves the jury to decide the degree to which the evidence contradicts the hypothesis of innocence. Juries don't explicitly use probability to help them decide whether to reject that hypothesis. But when you ask the same question of your null hypothesis, you have the advantage of being able to quantify exactly how surprising the evidence would be if the null hypothesis were true.

Example: Sleep versus Caffeine

- Students were given words to memorize, then randomly assigned to take either a 90 min nap, or a caffeine pill. 2 $1/2$ hours later, they were tested on their recall ability.
- Variables
  - Explanatory variable: sleep or caffeine
  - Response variable: number of words recalled
- Is sleep or caffeine better for memory?

Mednick, Cai, Kanady, and Drummond (2008). "Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory," Behavioral Brain Research, 193, 79-86.

Sleep versus Caffeine

- Let μ_s and μ_c be the mean number of words recalled after sleeping and after caffeine.
- Is there a difference in average word recall between sleep and caffeine?

Poll question

What are the null and alternative hypotheses?

- a $H_0 : \mu_s \neq \mu_c, H_A : \mu_s = \mu_c$
- b $H_0 : \mu_s = \mu_c, H_A : \mu_s \neq \mu_c$
- c $H_0 : \mu_s \neq \mu_c, H_A : \mu_s > \mu_c$
- d $H_0 : \mu_s = \mu_c, H_A : \mu_s > \mu_c$
- e $H_0 : \mu_s = \mu_c, H_A : \mu_s < \mu_c$

Difference in Hypotheses

Note: the following two sets of hypotheses are equivalent, and can be used interchangeably:

$$H_0 : \mu_1 = \mu_2 \qquad\qquad H_0 : \mu_1 - \mu_2 = 0$$
$$H_A : \mu_1 \neq \mu_2 \qquad\qquad H_A : \mu_1 - \mu_2 \neq 0$$

Take a minute to write down the hypotheses for each of the following situations:

Does the proportion of people who support gun control differ between males and females?

Is the average hours of sleep per night for college students less than 7?

Key question

How unusual is it to see a sample statistic as extreme as that observed, if H_0 is true?

- If it is very unusual, we have evidence against the null hypothesis H_0, in favor of H_a
- If it is *not* very unusual, our test is inconclusive

Statistical significance

When results as extreme as the observed sample statistic are unlikely to occur by random chance alone (assuming the null hypothesis is true), we say the sample results are **statistically significant**

- If our sample is statistically significant, we have convincing evidence against H_0, in favor of H_a
- If our sample is not statistically significant, our test is inconclusive

Sleep versus Caffeine

Let μ_s and μ_c be the mean number of words recalled after sleeping and after caffeine.

$$H_0 : \mu_s = \mu_c, \quad H_A : \mu_s \neq \mu_c$$

Poll question

The sample difference in means is $\bar{x}_s - \bar{x}_c$ and this is statistically significant. We can conclude...

a. there is a difference between sleep and caffeine for memory
b. there is not a difference between sleep and caffeine for memory
c. nothing

$p =$ Proportion of cracked ingot

$$H_0 : p = 50\%, \qquad H_a : p < 50\%$$

If results are statistically significant...

- the sample proportion of cracked ingots is lower than is likely just by random chance (if the new method did work, and $p < 50\%$)
- we have evidence that the true proportion of cracked ingots really is lower than 50%, and thus have evidence of the new method

Ingot example

$p =$ Proportion of cracked ingot

$$H_0 : p = 50\%, \qquad H_a : p < 50\%$$

If results are NOT statistically significant...

- the sample proportion of cracked ingots could easily happen just by random chance (if the new method does not work, and $p = 50\%$)
- we do not have enough evidence that the true proportion of cracked ingots really is lower than 50% or the new method works

Key question

How unusual is it to see a sample statistic as extreme as that observed, if H_0 is true?

- If it is very unusual, we have statistically significant evidence against the null hypothesis
- How do we *measure* how unusual a sample statistic is, if H_0 is true?

We will learn this tomorrow!

Summary

- Statistical tests use data from a sample to assess a claim about a population
- Statistical tests are usually formalized with competing hypotheses:
  - Null hypothesis (H_0): no effect or no difference
  - Alternative hypothesis (H_a): what we seek evidence for
- If it would be unusual to get results as extreme as that observed, just by random chance, if the null were true, then the data is statistically significant
- If data are statistically significant, we have convincing evidence against the null hypothesis, and in favor of the alternative