Unit 4: Sampling

5. Simulation based hypothesis testing

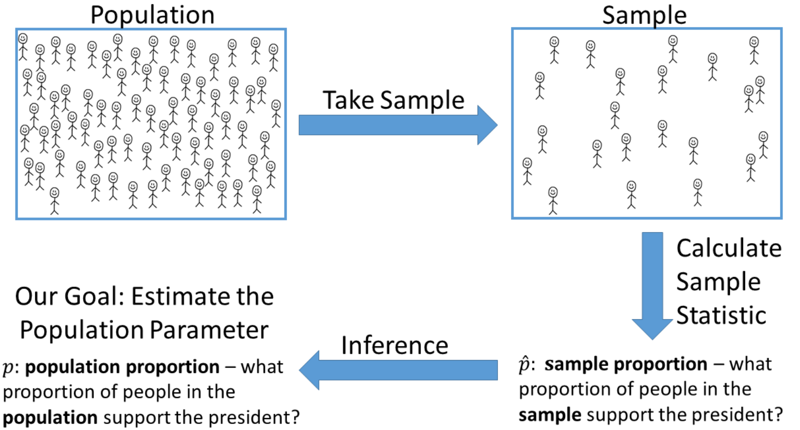
Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

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

- Please post your group information in the chat now.
- Thanks to Ariel, Jamie, Miki and Bianna's group for posting their team information on Piazza.


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?

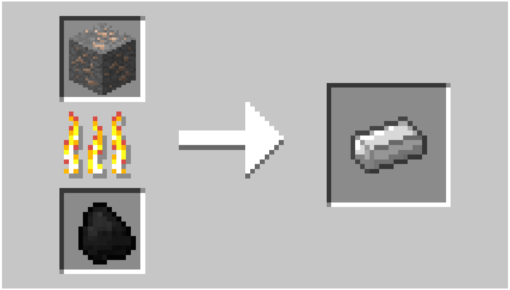
statistical inference is the process of using sample data to make conclusions about the underlying population the sample came from.

Types of inference:

- estimation (confidence interval)
- testing (focus of today)

real ingot

making ingot in minecraft



As the liquid metal cools, cracks can develop on the surface of the ingot, which compromises its integrity. Airplane manufacturers insist that metal for their planes be defect-free, so the ingot must be made over if any cracking is detected.

It's estimated that 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 some changes in the casting process.

Since then, 400 ingots have been cast and only 180 of them have cracked (this is 45%).

Has the new method worked? Has the cracking rate really decreased, or was 45% just due to randomness?

What is ...

- population
- population parameter
- a random sample
- sample statistics

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%?

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.

From your reading yesterday...

1. Start with two hypotheses about the population: the null hypothesis (H_0) and the alternative hypothesis (H_A).
2. Choose a (representative) sample, collect data, and analyze the data.
3. Figure out how likely it is to see data like what we observed, **if the null hypothesis were in fact true**.
4. If our data would have been extremely unlikely under the null hypothesis, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim.

The null hypothesis (often denoted H_0) states that "nothing unusual is happening" or "there is no relationship," etc.

On the other hand, the alternative hypothesis (often denoted H_A) states the opposite: that there is some sort of relationship (usually this is what we want to check or really think is happening).

Remember, in statistical hypothesis testing we always first assume that the null hypothesis is true and then see whether we reject or fail to reject this claim

- **Hypothesis**: a hypotheses is a statement about the value of an unknown population parameter.
  - null hypothesis H_0: the starting hypothesis to be tested
  - alternative hypothesis H_A: contains values of the parameter that we consider plausible if we reject the null hypothesis
- **Hypothesis test**: a hypothesis test consists of a test between two competing hypotheses: (1) a null hypothesis H_0 versus (2) an alternative hypothesis H_A.

What will our null and alternative hypotheses be for this example?

- H_0: the true cracking rate of ingots cast using the new method is 0.50
- H_A: the true cracking rate of ingots cast using the new method is lower than 0.50

Expressed in symbols:

$$H_0 : p = 0.50$$
$$H_A : p < 0.50$$

With these two hypotheses, we now take our sample and summarize the data.

The choice of summary statistic calculated depends on the type of data. In our example, we use the sample proportion,

$$\hat{p} = 180/400 = 0.45 = 45\%$$

- **Test statistic**: a test statistic is a sample statistic used for hypothesis testing
- **Observed test statistic**: the observed test statistic is the value of the test statistic that we observed in real life.

Null distribution: the null distribution is the sampling distribution of the test statistic assuming the null hypothesis H_0 is true.

Back to the ingot example

Assume cracking rate is 0.50, how will the sample proportions vary due to sampling under H_0?

Whether the 45% cracking rate could reasonably occur under the assumption that the crack rate is 50%?

How likely it is to see data like what we observed $\hat{p} = 0.45$, if the null hypothesis were in fact true $p = 0.50$?

Code 1: Sample once

```
outcome <-
c('crack', 'no crack')
samp1 <- sample(outcome, size = 400,
prob = c(0.5, 0.5), replace = TRUE)
```

Code 3: Calculate proportion

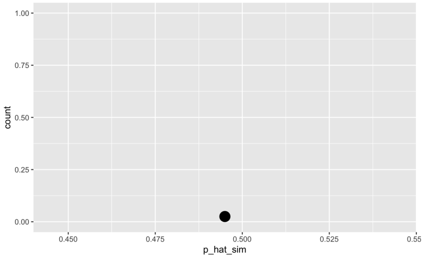
```
p_hat_sim <- sum(samp1 == "crack")/400
```

Output

```
samp1
   crack   no crack
      198          202

p_hat_sim
    0.495
```

Plot



Code 1: Sample once

```
outcome <-
c('crack', 'no crack')
samp2 <- sample(outcome, size = 400,
prob = c(0.5, 0.5), replace = TRUE)
```

Code 3: Calculate proportion

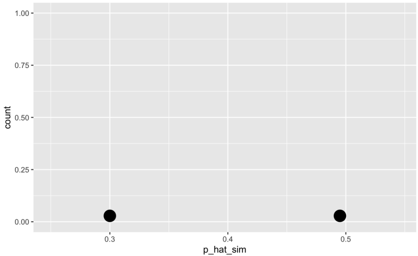
```
p_hat_sim <- sum(samp2 == "crack")/400
```

Output

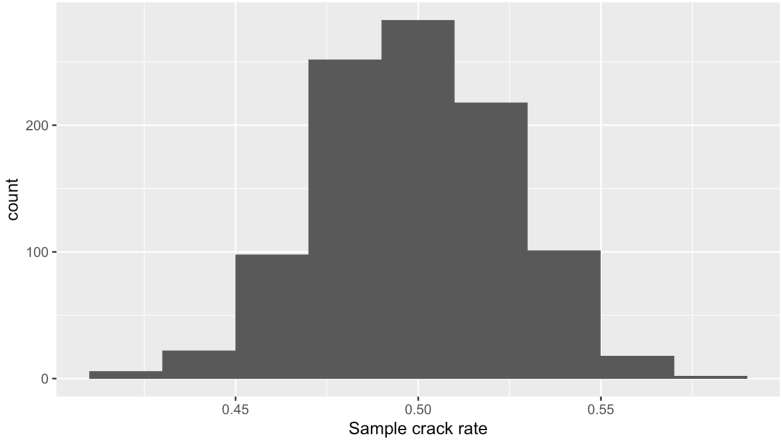
```
samp2
   crack   no crack
      120         280

p_hat_sim
    0.3
```

Plot



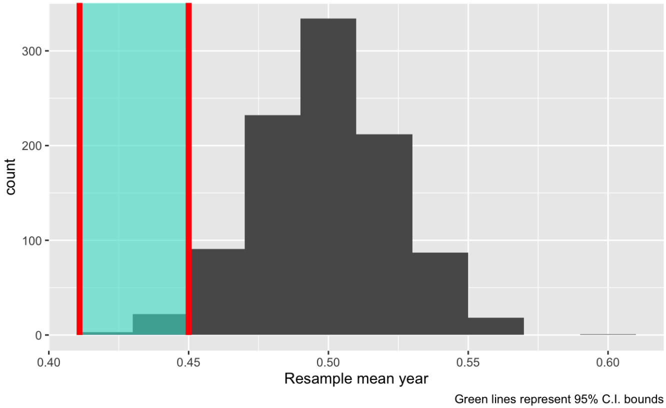
The sampling distribution of the crack rate



How many simulations have a crack rate less than 0.45?

```
p <- sample_1000_prop %>%
  count(sample_crack_rate)
```

$p = 25$
proportion $= 2.5\%$



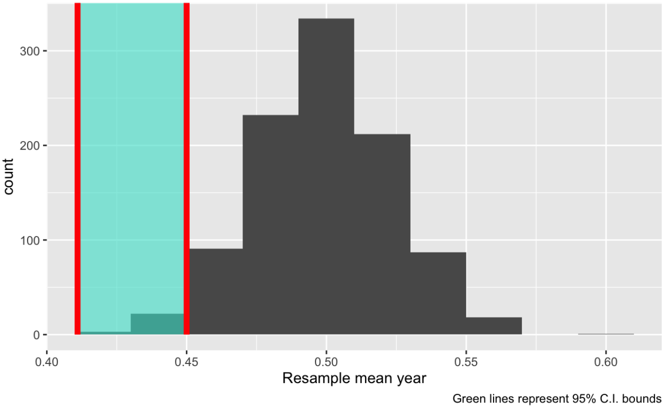
The proportion you calculated above is called an **approximate p-value**. The **p-value** is the probability of getting data like ours, or more extreme, if H_0 were in fact actually true.

For now, think **probability** as the proportion of times that an event occurs in a infinite sequence (or very long run) of separate tries.

So a p-value is the proportion of getting data like ours, or more extreme in the very long run, under the assumption that H_0 is true.

Small p-value means it is very unlikely to observe our data or more extreme if H_0 were true. Therefore, reject H_0!

The p-value can be ***approximated*** *by % of the simulations that has sample proportion at least as extreme as the observed sample proportion*



We often consider a cut point the **significance level** or α defined prior to conducting the analysis. Many analyses use $\alpha = 0.05$.

- If p-value $< \alpha$, reject H_0, data provide evidence for H_A, and the results are statistically significant.
- If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A.

Suppose that the significance level α is set to be 0.05. I have calculated the p-value using the simulation studies, which is 0.025. What is the conclusion of the hypothesis test?

Since the p-value is less than the significance level, we reject the null hypothesis.

How can we answer research questions using statistics?

Statistical hypothesis testing is the procedure that assesses evidence provided by the data in favor of or against some claim about the population (often about a population parameter or potential associations).

From your reading yesterday...

1. Start with two hypotheses about the population: the null hypothesis (H_0) and the alternative hypothesis (H_A).
2. Choose a (representative) sample, compute the test statistic
3. Figure out how likely it is to see data like what we observed, **if the null hypothesis were in fact true**.
4. If our data would have been extremely unlikely under the null hypothesis, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim.

1. Motivation: a study on ingot
2. Hypothesis testing framework