One-Sample Hypothesis Testing

Part 1 - Sampling from a population

Ryan Miller

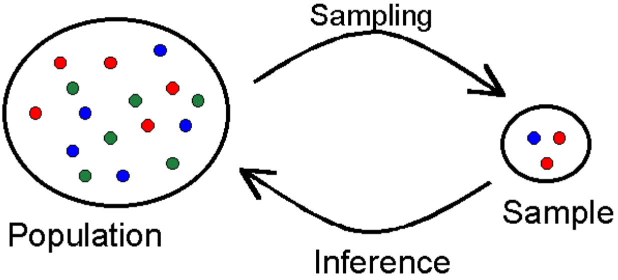
Grinnell College
Statistics

Introduction

- In our infant toy choice example, do you think the researchers were more interested in characteristics of infants in general (ie: all infants) or characteristics of the 16 infants who participated in the study?
- The researchers observed 14 of 16 infants chose the helper, is it possible that this result could change if a different group of 16 infants had participated?

Sampling from a Population

Statisticians adopt the paradigm that the data we are analyzing are a **sample**, or subset of a broader **population**:



- We usually care about characteristics of the population, but we only observe a sample
- The goal of hypothesis testing is to determine whether evidence from a sample is strong enough to support a conclusion about

Sampling from a Population

Statisticians use *statistical notation* to distinguish **population parameters** (things we want to know) from **estimates** (things derived from a sample):

	Population Parameter	Estimate (from sample)
Mean	μ	$\overline{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r

- Parameters are fixed but unknown, while estimates are observable but vary from sample to sample
- In our infant toy choice example, we observed $\hat{p} = 14/16$ and evaluated the population-level hypothesis $p = 0.5$

Sampling Error

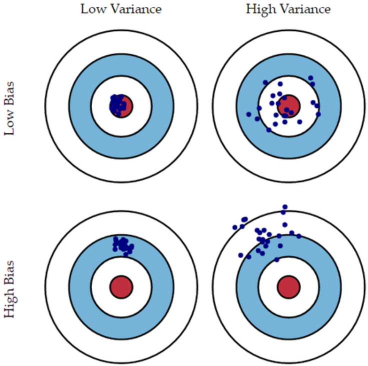
There are 2 main reasons for a sample estimate to differ from what's true of the population:

1) **Sampling Bias** - a systematic flaw in the way cases were selected that leads to certain types of cases being disproportionately represented in the sample data
2) **Sampling Variability** - since a sample doesn't include all of the population, any individual sample might differ from the population due to *random chance* (ie: "the luck of the draw")

Having a larger sized sample will *reduce sampling variability* but *will not fix sampling bias*.

Sampling Error

Each dot represents an estimate (ie: $\hat{p}$) from a *different sample*:



What factors might lead to *high variance*? What might lead to *biased sampling*?

Grinnell College
Statistics

Sampling Strategies and Hypothesis Testing

- Sampling bias can be addressed by taking a **random sample** of cases from the population of interest
  - Random sampling gives each member of the population an equal chance of being selected, so the sample will, on average, match the population
- Other sampling approaches may or may not yield biased samples
  - Surveying students as they leave Noyce for lunch might yield a minimally biased sample of Grinnell science students
  - Surveying students in CSC-301 will yield a biased sample where computer science majors are over-represented

Sampling Strategies and Hypothesis Testing

- Sampling bias can be addressed by taking a **random sample** of cases from the population of interest
  - Random sampling gives each member of the population an equal chance of being selected, so the sample will, on average, match the population
- Other sampling approaches may or may not yield biased samples
  - Surveying students as they leave Noyce for lunch might yield a minimally biased sample of Grinnell science students
  - Surveying students in CSC-301 will yield a biased sample where computer science majors are over-represented
- The other source of error, sampling variability, can be addressed using **hypothesis testing**!!

Hypothesis Testing

Formalizing the steps from our previous lab, hypothesis testing involves two major components:

1) Proposing a **null hypothesis**, H_0, and an **alternative hypothesis**, H_a
   - The null hypothesis is a statement about the population that the researchers seek evidence against
   - The alternative hypothesis represents the conclusion the researchers would like to establish

Hypothesis Testing

Formalizing the steps from our previous lab, hypothesis testing involves two major components:

1) Proposing a **null hypothesis**, H_0, and an **alternative hypothesis**, H_a
    - The null hypothesis is a statement about the population that the researchers seek evidence against
    - The alternative hypothesis represents the conclusion the researchers would like to establish
2) Deciding whether the sample data provide *sufficient evidence* against the null hypothesis
    - A **null distribution** describes the sample outcomes that could have occurred *had the null hypothesis been true*
    - Evidence against H_0 comes from comparing the outcome observed from the *real data* versus the null distribution

Types of Hypothesis Tests

- A **one-sample test** involves treating all of the available data as a single sample from the same population
  - For example, the $n = 16$ infants in our helper-hinderer example were analyzed as a single sample intended to represent the behaviors of "all infants"

Types of Hypothesis Tests

- A **one-sample test** involves treating all of the available data as a single sample from the same population
  - For example, the $n = 16$ infants in our helper-hinderer example were analyzed as a single sample intended to represent the behaviors of "all infants"
- In contrast, a **two-sample test** involves the comparison of two different groups, each of which is taken to represent a different population
  - For example, a drug trial might split participants into treatment and control groups and compare the outcomes of each group
  - The treatment group is taken to represent all users of the drug, and the controls are taken to represent non-users

Types of Hypothesis Tests

- The distinction between one-sample and two-sample tests is important because it will guide the hypotheses we set up as well as the procedure we use to determine the null distribution
  - For *one-sample categorical data*, our null hypothesis has the form $H_0 : p = \_\_\_$
  - For *one-sample quantitative data*, our null hypothesis has the form $H_0 : \mu = \_\_\_$ (we won't test anything other than the mean, at least for now)

P-values

The **p-value** is the *conditional probability* of observing an outcome at least as unusual as the one observed in the real data under the assumption that the null hypothesis is true. We might think of this as $Pr(\text{Our data (or more extreme)}|H_0 \text{ is true})$

- The *p*-value is calculated using either one tail (1-sided test) or both tails (2-sided test) of the *null distribution*
- A small *p*-value can be used to falsify the null hypothesis, but a large *p*-value should be considered "inconclusive"

P-values as Evidence

Below are the original guidelines put forth by Ronald Fisher (creator of the p-value):

p-value	Evidence against the null
0.100	Borderline
0.050	Moderate
0.025	Substantial
0.010	Strong
0.001	Overwhelming

Fisher intended the p-value to be a quantitative measurement describing the strength of the evidence the sample data provide against a null hypothesis.

Decision Thresholds

While p-values measure evidence on a continuum, many scientific fields have adopted $\alpha = 0.05$ as a decision threshold for "statistical significance":

- Data yielding a p-value *smaller* than $\alpha = 0.05$ are seen as *sufficient evidence* for rejecting H_0
- Data yielding a p-value *larger* than $\alpha = 0.05$ are seen as *insufficient evidence* and result in a "failure to reject H_0"

This black and white approach has flaws, but it's still very widely used.

Null Distributions for One-Sample Tests

For now we'll create the *null distribution* for our hypothesis tests using simulations:

- For *one-sample categorical data*, we can simulate proportions that could have occurred in sample if H_0 had been true
  - You might imagine spinning a spinner with the hypothesized probability of the outcome of interest for each case in the sample
- For *one-sample quantitative data*, we can simulate means that could have occurred in sample had H_0 been true by subtracting the hypothesized mean from each data point then sampling the re-centered data with replacement

Practice

Researchers at Johns Hopkins University studying preterm birth in the United States recruited consenting volunteers who had a child born at 25-weeks (15 weeks prematurely) in their hospital system. They found that 31 of 39 babies survived.

- **Part A** - What are the *sample* and the *population of interest* in this study? How would you describe the representativeness of this sample?
- **Part B** - The Wikipedia article on preterm birth lists the survival rate of babies born at 25-weeks at 70%. Is the fact that $31/39 \approx 79.5\%$ enough to conclude Wikipedia is wrong?

Practice (cont.)

- **Part C**: State the null and alternative hypotheses for a two-sided test using statistical notation.
- **Part D** - What is the *sample estimate* of the parameter described in the null hypothesis? Report the estimate using proper statistical notation.
- **Part E** - Use the sample data and hypotheses from earlier to find the two-sided p-value. Do these data provide sufficient evidence to refute Wikipedia's claim? Use StatKey to help you estimate the p-value.
- **Part F** - Suppose we had cheated and set up a one-sided alternative hypothesis of $H_a : p > 0.7$ after noticing the data showed a survival rate above 70%. What would the one-sided p-value be in this scenario?

Wrap-up

Most of our analyses in this class will involve some type of hypothesis test. When reporting the results of a test you are expected to include the following:

1. *Scientific context* - Example: the survival of infants born at 25-weeks
2. *Strength of evidence* - Examples: moderate/strong evidence (typically p-value < 0.05), or a lack of evidence (typically p-value ≥ 0.05)
3. *The type of relationship* - Example: higher survival rate

You should *not* use generic phrases like "reject H_0" or "conclude the alternative due to $p < 0.05$" unless explicitly asked to do so.