Hypothesis Testing (part 2, p-values)

Ryan Miller

Introduction

- In the last presentation we revisited the infant toy-choice experiment that we explored on the first day of class
  - Because the study effectively used randomization to prevent confounding variables, only two viable explanations remained for 14 of 16 babies choosing the "helper" toy, random chance or a real relationship

Introduction

- In the last presentation we revisited the infant toy-choice experiment that we explored on the first day of class
  - Because the study effectively used randomization to prevent confounding variables, only two viable explanations remained for 14 of 16 babies choosing the "helper" toy, random chance or a real relationship
- The first step in evaluating whether random chance might explain the result was to setup an appropriate null model that described the scenario
  - We used the null model: $H_0 : p = 0.5$, because it implied that each infant's choice was random

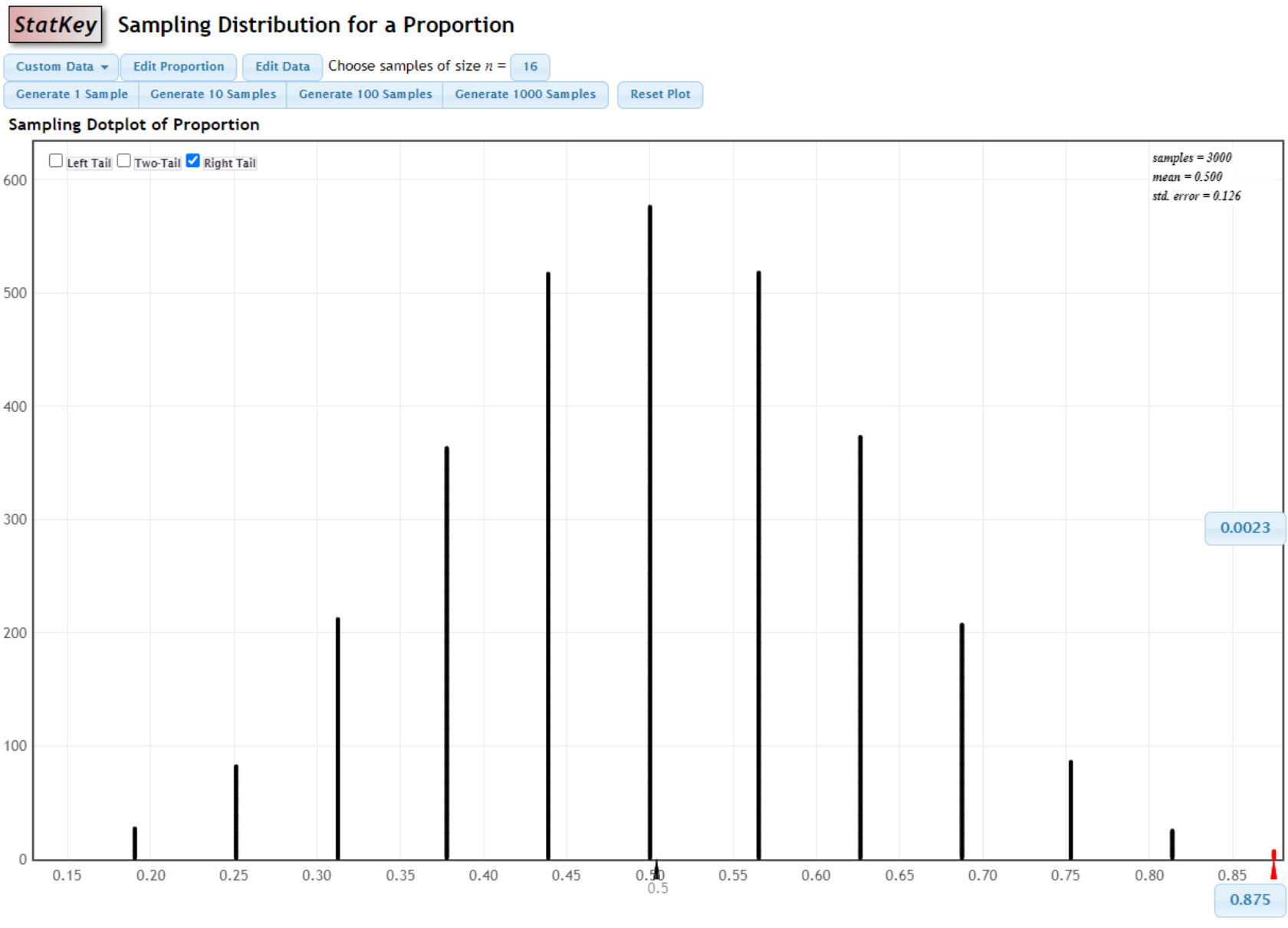
p-values

- Probability theory allows us to quantify how compatible/incompatible the sample data are with a null model
  - The **p-value** is defined as *the probability of seeing an outcome at least as extreme as what was observed in our sample if the null model were true*

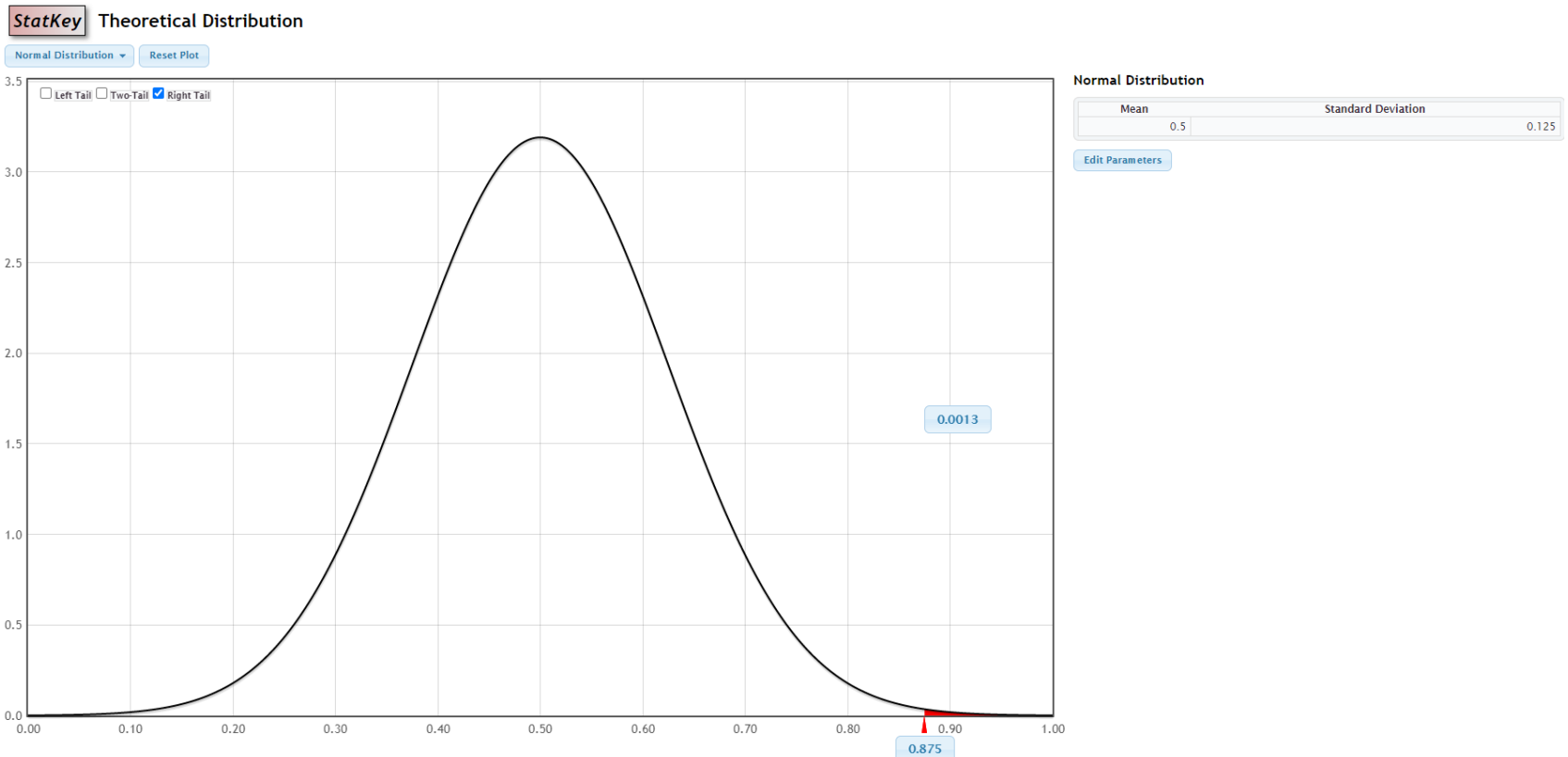
p-values

- Probability theory allows us to quantify how compatible/incompatible the sample data are with a null model
  - The **p-value** is defined as *the probability of seeing an outcome at least as extreme as what was observed in our sample if the null model were true*
- The smaller the *p*-value, the more incompatible the sample data are with the null model, and thus the stronger the evidence is against random chance as a viable explanation
  - For example, a *p*-value of 0.01 indicates a 1/100 chance of seeing results as extreme as the sample data if the null model were true

The Simulated Null Distribution

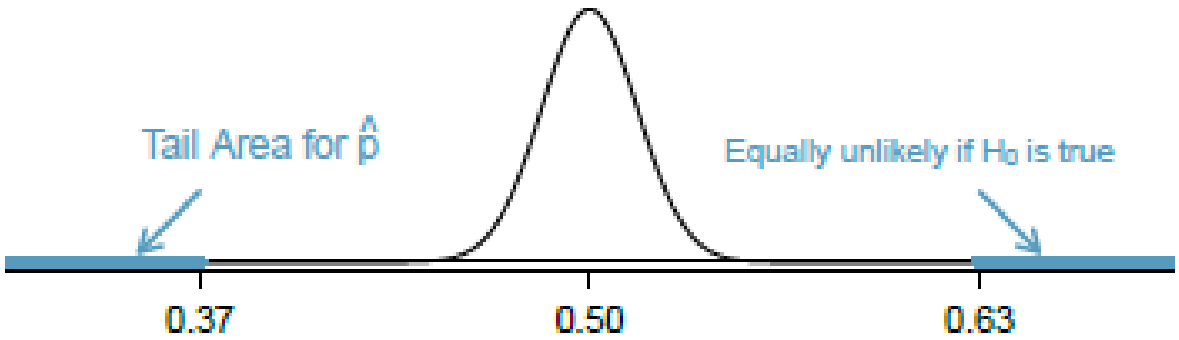


The CLT Null Distribution



Two-sided p-values

- The p-values (0.0023 and 0.0013) we calculated using the simulated/CLT null distributions aren't actually the ones that a researcher would report in an scientific journal
  - Instead, they are a special type of p-value called a *one-sided* p-value that is rarely used
- Instead, statisticians prefer *two-sided* p-values:



- The practical implication is that we must *double* the one-sided tail area to account for *all* areas of the null distribution that are as unlikely as the outcome observed in our sample

Alternative Hypotheses

There are many reason why statisticians prefer two-sided p-values, and one is the notion that any null model must be paired with a *complementary* alternative:

$$H_0 : p = 0.5$$

$$H_A : p \neq 0.5$$

Under this setup, an observed sample proportion that is either very large or very small would provide substantial evidence against the null model

p-values as a Measure of Evidence

Ronald Fisher, creator of the *p*-value, and described by his peers as "a genius who almost single-handedly created the foundations of modern statistical science", suggests the following guidelines:

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

p-values as a Measure of Evidence

Ronald Fisher, creator of the *p*-value, and described by his peers as "a genius who almost single-handedly created the foundations of modern statistical science", suggests the following guidelines:

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

- Many scientific fields use $\alpha = 0.05$ as a "significance threshold" for *rejecting* a null hypothesis
- Thus, *p*-values < 0.05 are described as "statistically significant"

Arguments Against "Statistical Significance"

- $p < 0.05$ is an arbitrary cutoff that shouldn't distract you from the main idea behind p-values
- That is, a p-value of 0.0001 doesn't tell you the same thing as a p-value of 0.04, even though both are "statistically significant"

Arguments Against "Statistical Significance"

- $p < 0.05$ is an arbitrary cutoff that shouldn't distract you from the main idea behind p-values
- That is, a p-value of 0.0001 doesn't tell you the same thing as a p-value of 0.04, even though both are "statistically significant"
- When reporting results you should always include the p-value itself, not just whether it met some arbitrary significance threshold
  - Imagine your weather app only telling you: "it's cold" or "it's not cold"
  - This is bad because "cold" is subjective, it's better to provide the temperature and let you decide for yourself

Summary

- The null distribution is an intermediate step in calculating the p-value, or the probability of observing an outcome at least as unusual that seen in the sample data if the null model were true

Summary

- The null distribution is an intermediate step in calculating the *p*-value, or the probability of observing an outcome at least as unusual that seen in the sample data if the null model were true
  - We will almost always report *two-sided* *p*-values, which involve extreme outcomes on both ends of the null distribution
  - The two-sided *p*-value is found by multiplying the relevant one-sided tail-area by 2

Summary

- The null distribution is an intermediate step in calculating the *p*-value, or the probability of observing an outcome at least as unusual that seen in the sample data if the null model were true
  - We will almost always report *two-sided* *p*-values, which involve extreme outcomes on both ends of the null distribution
  - The two-sided *p*-value is found by multiplying the relevant one-sided tail-area by 2
- You should think of the *p*-value as a measure of incompatibility between the null model and the observed data
  - A smaller *p*-value suggests lower compatibility (ie: it's likely that the null model is wrong)
  - The next presentation will focus on common misinterpretations of the *p*-value