STAT 113
Hypothesis Testing II

Colin Reimer Dawson

Oberlin College

October 10, 2017

Two Main Goals of Inference

1. Assessing strength of evidence about "yes/no" questions (hypothesis testing)
2. Estimating unknown quantities in a population using a sample (confidence intervals)

Statistics vs. Parameters

- Summary values (like mean, median, standard deviation) can be computed for populations or for samples.
- In a population, such a summary value is called a **parameter**
- In a sample, these values are called **statistics**, and are used to *estimate* the corresponding parameter

Value	Population Parameter	Sample Statistic
Mean	μ	$\bar{X}$
Proportion	p	$\hat{p}$
Correlation	ρ	r
Slope of a Line	β_1	$\hat{\beta}_1$
Difference in Means	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$
...	...	...

Quantifying H_0 and H_1

Identify the relevant population parameter for each of the following claims and state the null and alternative hypotheses (abbreviated H_0 and H_1), as statements about that parameter.

- Dr. Bristol can tell the difference between cups of tea more often than random guessing. H_0: $p_{correct} = 0.5$, H_1: $p_{correct} > 0.5$, where $p_{correct}$ is her "long run" success rate
- There is a positive linear association between pH and mercury in Florida lakes. H_0: $\rho = 0$, H_1: $\rho > 0$, where ρ is the correlation coefficient between pH and Hg in *all* Florida lakes
- Lab mice eat more on average when the room is light. H_0: $\mu_{\text{light}} - \mu_{\text{dark}} = 0$, H_1: $\mu_{\text{light}} - \mu_{\text{dark}} > 0$, where μ are "long run"/population means for an appropriate measure of amount of food consumed

Outline

Logic of Testing H_0

- Logic: Don't "confirm" H_1; try to reject H_0
- If the data would be very unlikely assuming H_0 were true, *and would be less unlikely if H_1 were true*, we have evidence against H_0 and hence in favor of H_1.

What should we measure the likelihood of?

- Suppose Dr. Bristol gets 9 out of 10 cups of tea right.
- How unlikely is that?
- What should count as "that"?

P-values

- "That" is *all potential outcomes that favor H_1 at least as much as the actual outcome*.
- Sample: 9 of 10 correct. "That" = _____
- The collective probability of all of these outcomes is called the **P-value** for the sample.

P-value Definition

P-value

The probability of obtaining a result at least as "extreme" (i.e., far from what's expected under H_0) as what was actually observed, *assuming H_0 is true* is called the P-value.

Outline

Randomization distribution under H_0

- Often we can simulate the world under H_0 to find a P-value
- Cards
- Computer simulation (e.g., R or StatKey)

Randomization Distribution

A **randomization distribution** is a simulated sampling distribution based on a hypothetical world where H_0 is true.

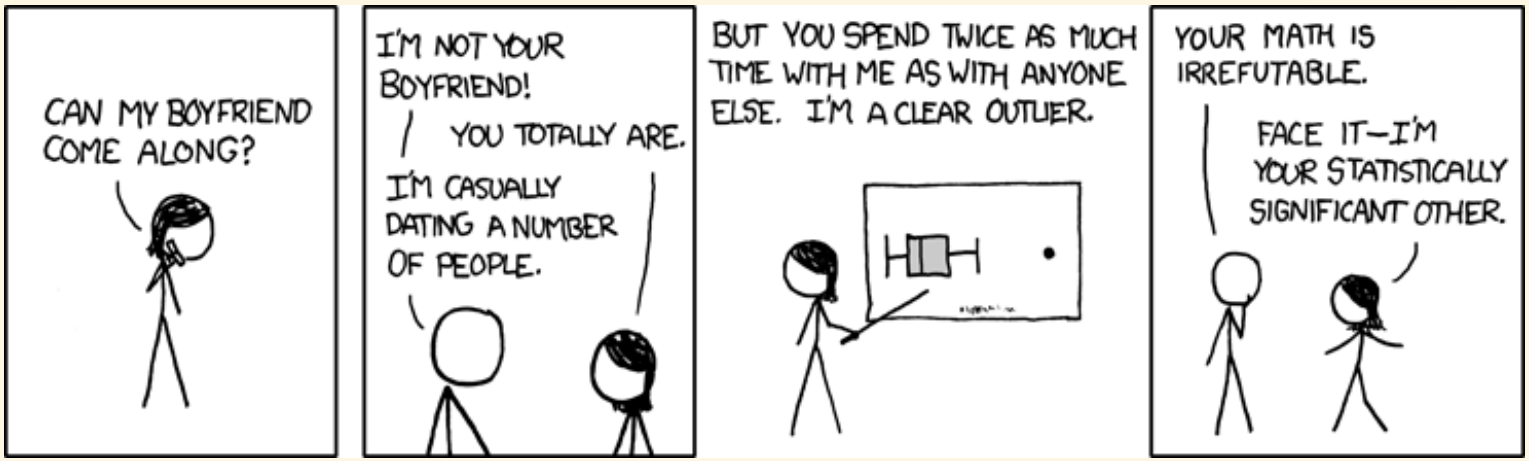
- The randomization distribution shows what types of statistics would be observed, *just by random chance*, if the null hypothesis were true

Simulating a Randomization Distribution

Handout

Outline

Statistical Significance



Statistical Significance

A finding in a sample (e.g., a correlation, or a difference between groups) is said to be **statistically significant** if the sample value (or one more extreme) would be very *unlikely* if H_0 is true (i.e., the P-value is low)

What is low enough?

Significance level (α)

We need to decide for ourselves, *in advance of collecting data*, what we will count as a "low enough" P-value to achieve statistical significance. This threshold is called the **significance level** of the test. (Notation: α)

Making a Decision

Reject H_0 or not?

(a) If $P \geq \alpha$: Do not reject H_0. (Data wouldn't be that surprising if H_0 true. H_0 is "presumed innocent".)

(b) If $P < \alpha$: Reject H_0. (Data would be too surprising if H_0 were true. Beyond a "reasonable doubt".)

Caution: We do not "accept H_0". We "fail to reject" it. (Not enough evidence to decide)

What if we're wrong?

Example

H_1 : Drug is better than a placebo
H_0 : Drug no better than a placebo

- We *reject* H_0 if the data (or something even less consistent with H_0) would be *improbable* in a world where H_0 is true.
- But improbable things happen sometimes! This means that we will occasionally reject H_0 incorrectly!
- E.g., we conclude that the drug works when in fact it doesn't: reject H_0 by mistake.

Types of Errors

Example

H_1 : Drug is better than a placebo
H_0 : Drug no better than a placebo

- We could prevent this from ever happening by never rejecting H_0
- Why not do this?

Types of Errors

2×2 table of possibilities. Is H_0 actually false (does the treatment actually work)? Did we reject H_0 (did we conclude that it works)?

		Action	
		H_0 **rejected**	H_0 **not rejected**
Truth	H_0 **is false**	True Discovery	Missed Discovery
	H_0 **is true**	False Discovery	No Error

Table: Possible outcomes of a null hypothesis significance test

Which is worse?

Pairs: What does increasing or decreasing α do to the likelihood of each possibility?

Type I vs. Type II Errors

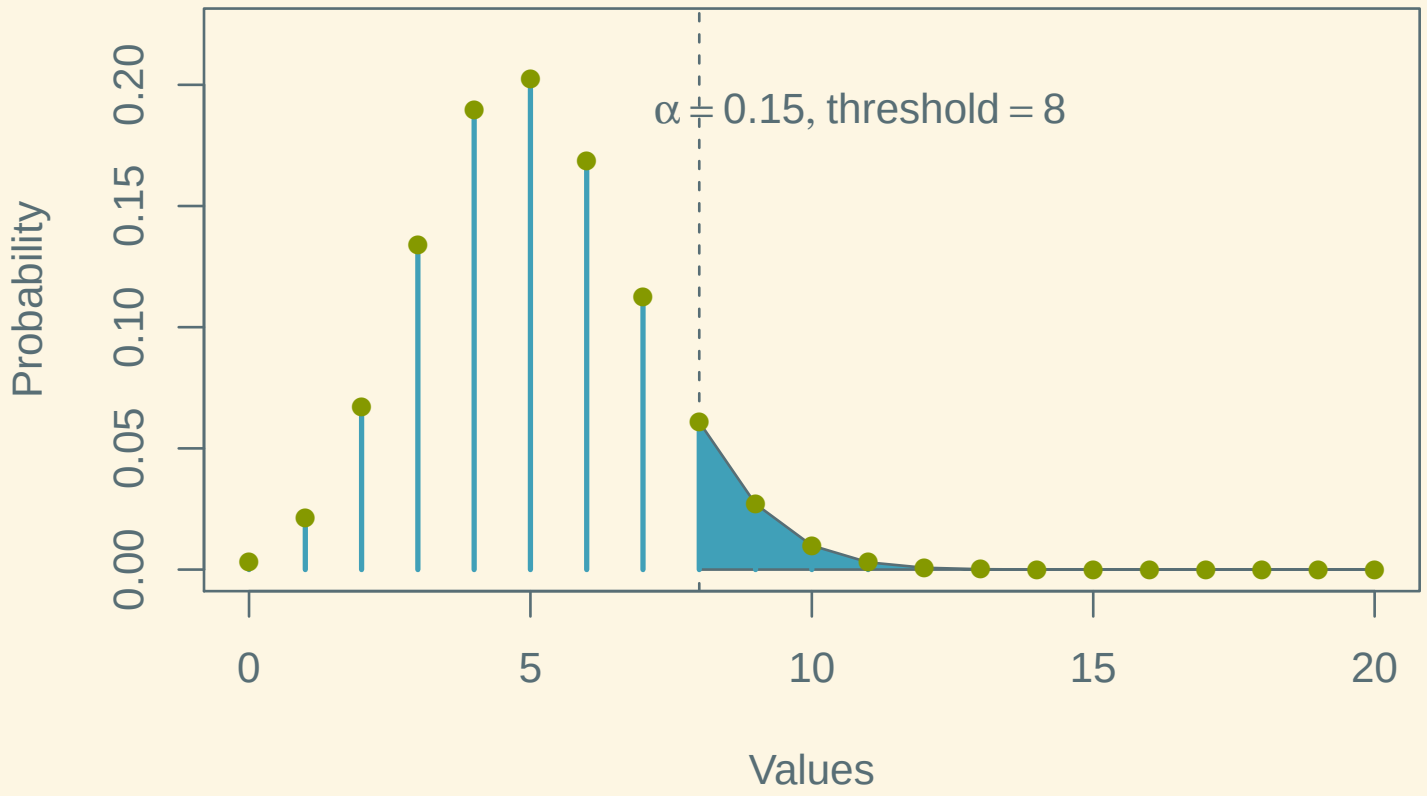
- We can set α to whatever we want. The lower it is, the less often we make false discoveries (also called "Type I" Errors).
- So why not make it really small?
- Tradeoff: Fewer false discoveries (Type I Errors) $\rightarrow$ More missed discoveries (Type II Errors).

Multiple Choice Test

- A professor writes a multiple choice "pretest" to assess whether students already know some of the course material when the semester starts.
- There are 20 questions, each with 4 options.
- For a particular student, we can ask "Do they know anything about this material?"
- H_0: $p_{\text{correct}} = 0.25$, H_1: $p_{\text{correct}} > 0.25$

Type I vs. Type II Errors

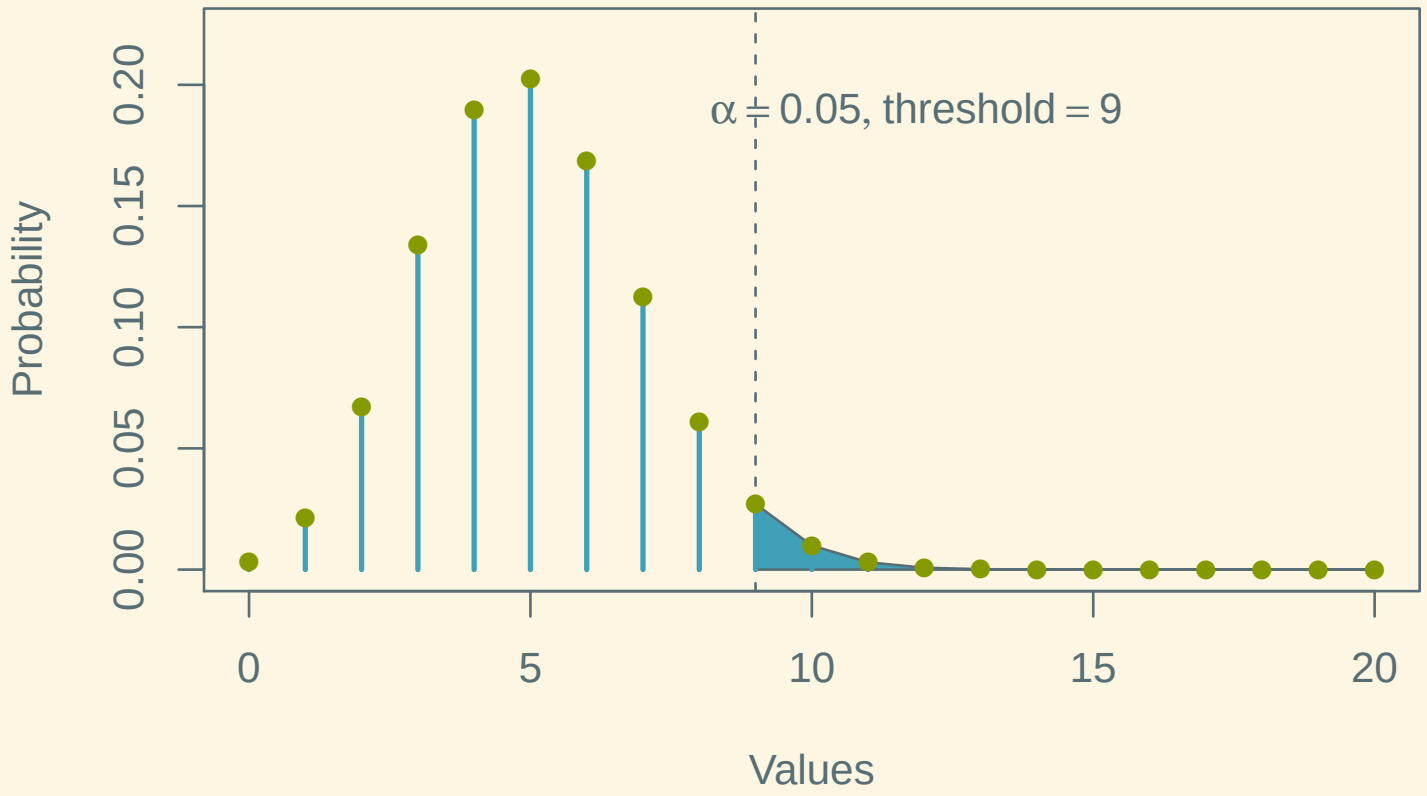
Decreasing α moves the rejection threshold out toward the tail of the H_0 distribution.



Blue spikes: Distribution of outcomes if H_0 is true

Type I vs. Type II Errors

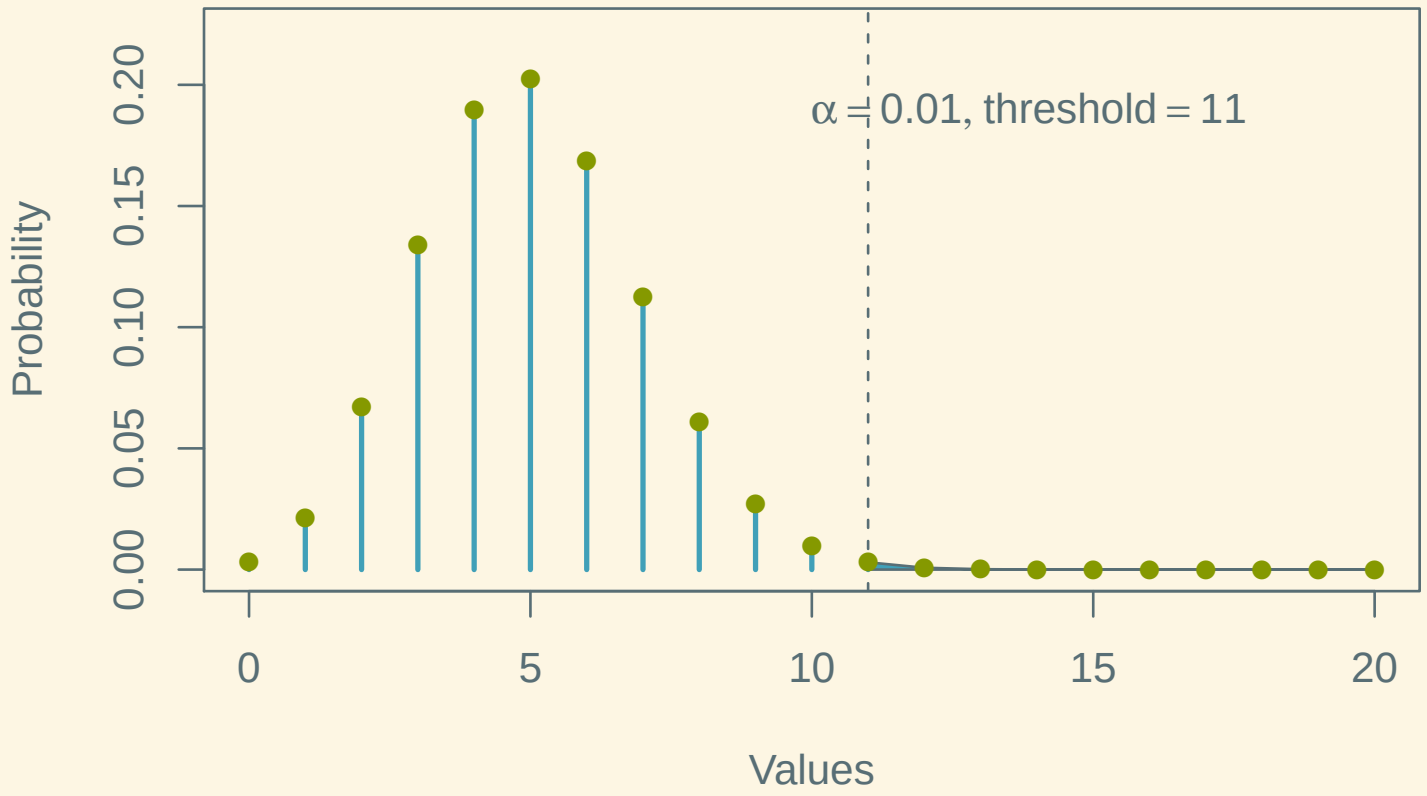
Decreasing α moves the rejection threshold out toward the tail of the H_0 distribution.



Blue spikes: Distribution of outcomes if H_0 is true

Type I vs. Type II Errors

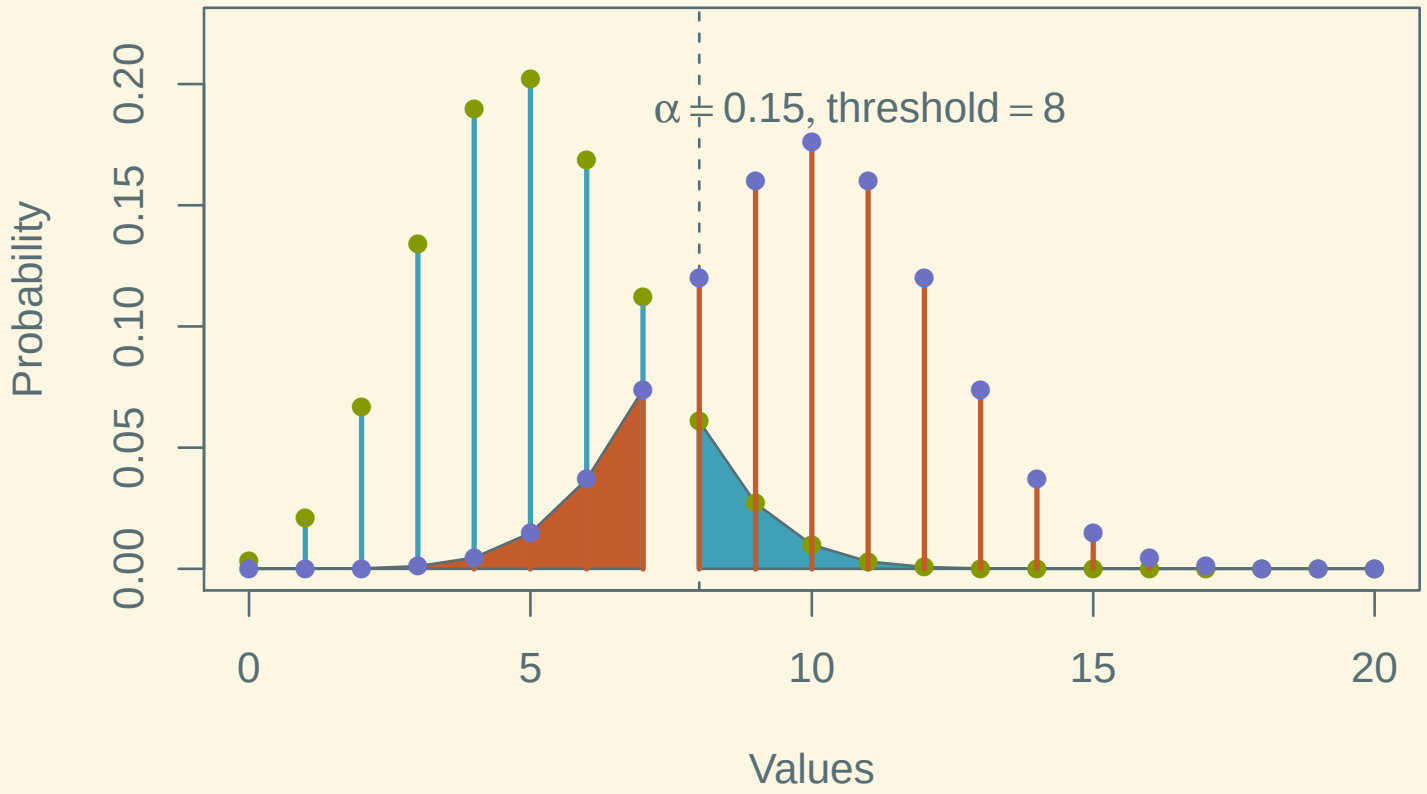
Decreasing α moves the rejection threshold out toward the tail of the H_0 distribution.



Blue spikes: Distribution of outcomes if H_0 is true

Type I vs. Type II Errors

We retain H_0 when we do not exceed the threshold. But if H_1 is correct, this is a Type II Error. More stringent threshold $\rightarrow$ more missed discoveries.

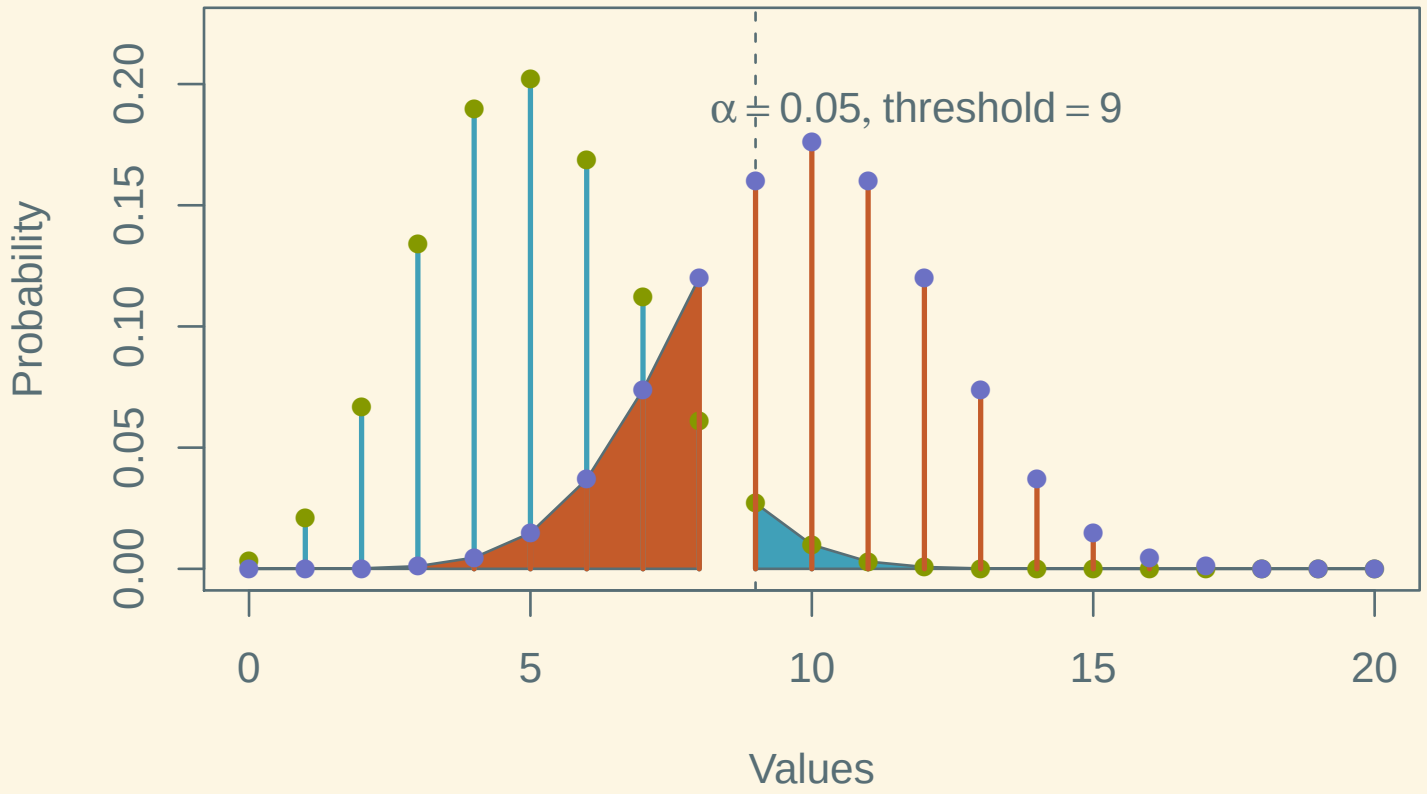


Blue spikes: Distribution of outcomes if H_0 is true

Orange spikes: Distribution of outcomes for one possible parameter

Type I vs. Type II Errors

We retain H_0 when we do not exceed the threshold. But if H_1 is correct, this is a Type II Error. More stringent threshold $\rightarrow$ more missed discoveries.

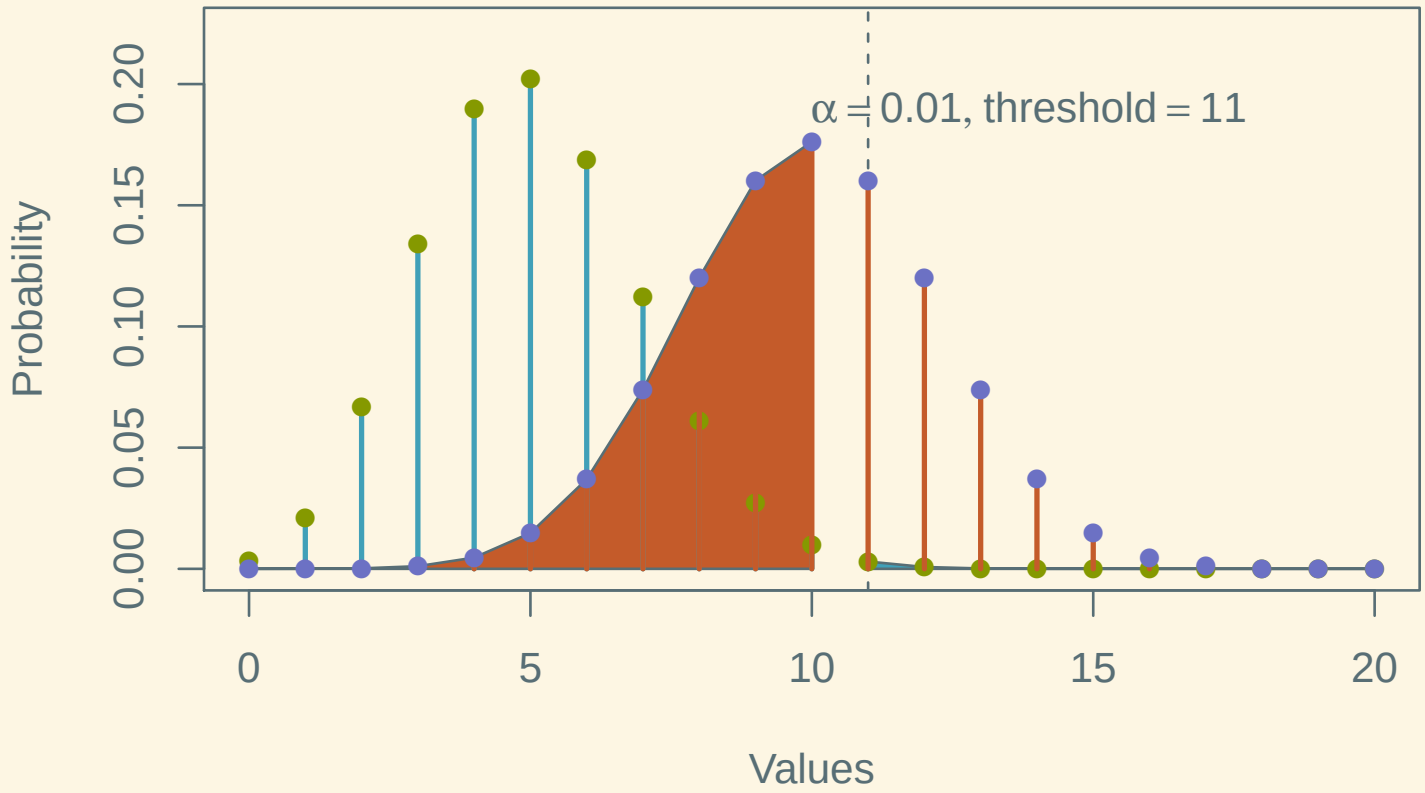


Blue spikes: Distribution of outcomes if H_0 is true

Orange spikes: Distribution of outcomes for one possible parameter

Type I vs. Type II Errors

We retain H_0 when we do not exceed the threshold. But if H_1 is correct, this is a Type II Error. More stringent threshold $\rightarrow$ more missed discoveries.



Blue spikes: Distribution of outcomes if H_0 is true

Orange spikes: Distribution of outcomes for one possible parameter

Outline

Testing Fairness of a Coin

- I suspect a coin is biased, but I don't know in what direction.
  - What is my parameter of interest?
  - What are my H_0 and H_1?
- I test it by flipping 100 times. What kinds of outcomes provide evidence against H_0?
- I get 63 heads. What's the "that" that I should counts toward the P-value? (StatKey)

Two-Tailed Tests

Two-Tailed Test

In a **Two-Tailed Test**, H_1 does not specify the direction (sign) of a difference/correlation/slope. So outcomes at either extreme count in its favor. The P-value therefore uses outcomes at or past the observed one, *but also the symmetric outcomes on the other "tail"*

We should prefer two-tailed tests, unless only one side of the alternative is plausible *a priori*.

Dangers of Directional Tests

Cardiac Arrhythmia Suppression Trial

Three drugs compared to a placebo, with the hope that they reduce deaths.

But some of them led to *more* deaths. We'd better be prepared to detect that.

Benefits of Non-directional Tests

Superconductors

In 1986, Alex Müller and Georg Bednorz created a brittle ceramic compound that superconducted at the highest temperature then known. What made this discovery so remarkable was that ceramics are normally insulators.

Unexpected benefit of finding the opposite effect from what was expected!