Decision Errors

Yue Jiang

STA 101 / Duke University / Fall 2023

Statistical inference

A process that converts data into useful information, whereby practitioners

- form a question of interest,
- collect, summarize, and analyze the data,
- and interpret the results

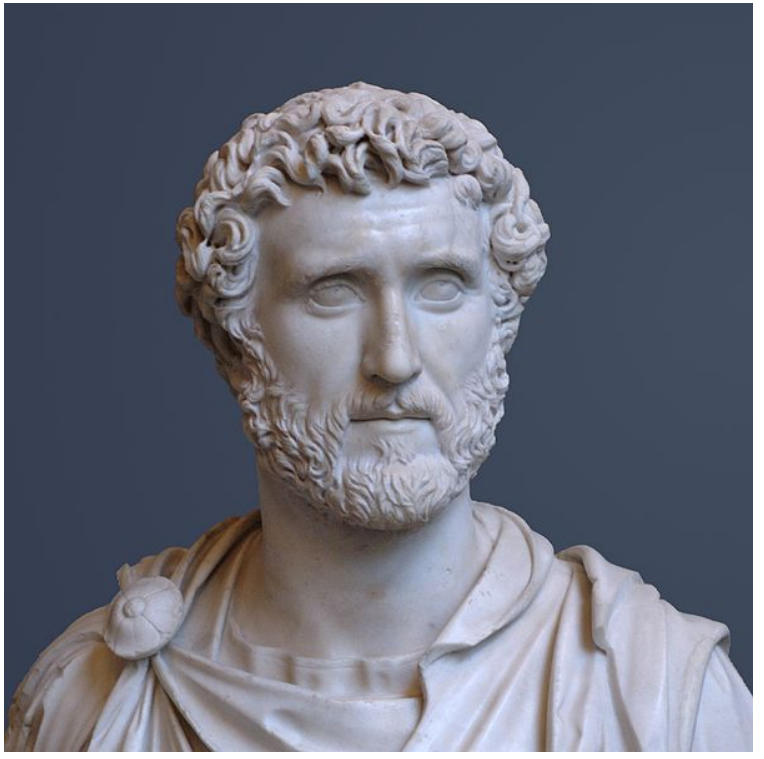
Statistical inference

The **population** is the group we'd like to learn something about. If we had data from every unit in the population, we could just calculate what we wanted and be done!

Unfortunately, we (usually) have to settle with a **sample** from the population. Ideally, the sample is **representative**, allowing us to use **probability and statistical inference** to make conclusions that are **generalizable** to the broader population of interest.

We want to make inferences regarding population **parameters**, which we do with **sample statistics**.

Ei incumbit probatio qui dicit



The hypothesis testing framework

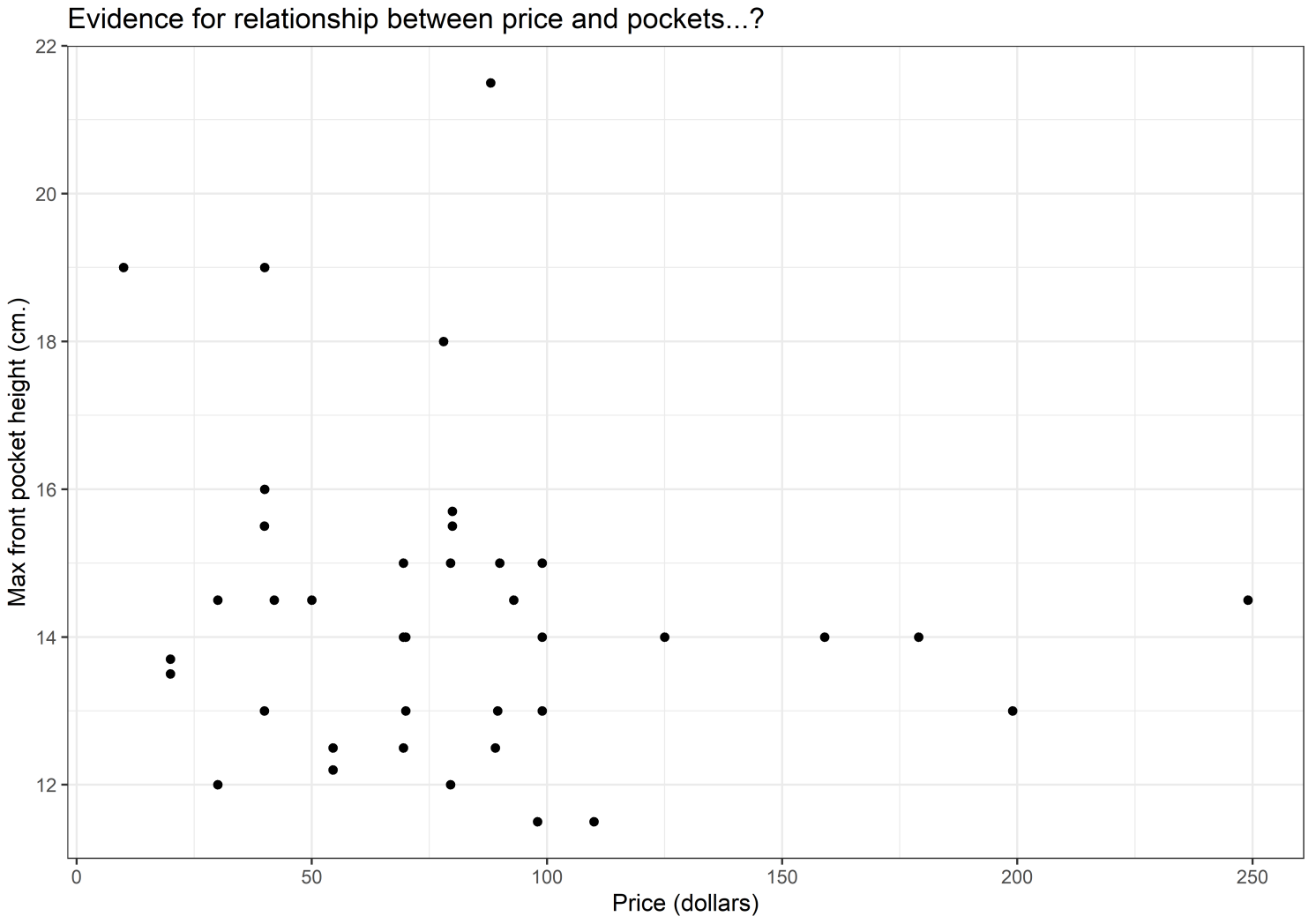
1. Start with two hypotheses about the population: the null hypothesis and the alternative hypothesis
2. Choose a sample, collect data, and analyze the data
3. Figure out how likely it is to see data like what we got/observed, IF the null hypothesis were true
4. If our data would have been extremely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

Do pricy jeans have differentially sized pockets?

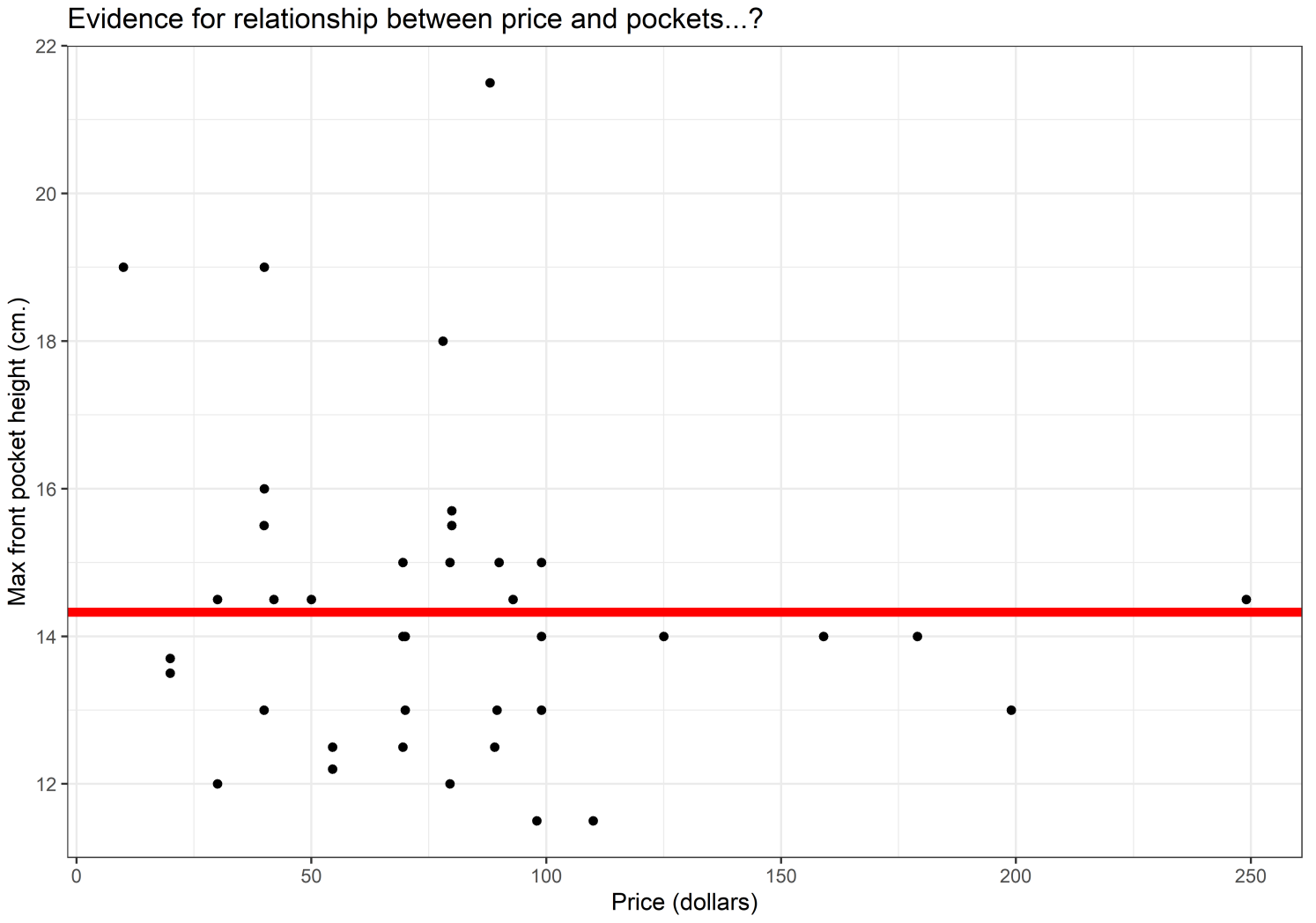
First, what do we *actually* care about here?

- H_0 : There is no relationship between price and pocket height
- H_1 : There *is* a relationship between price and pocket height

Collecting our data



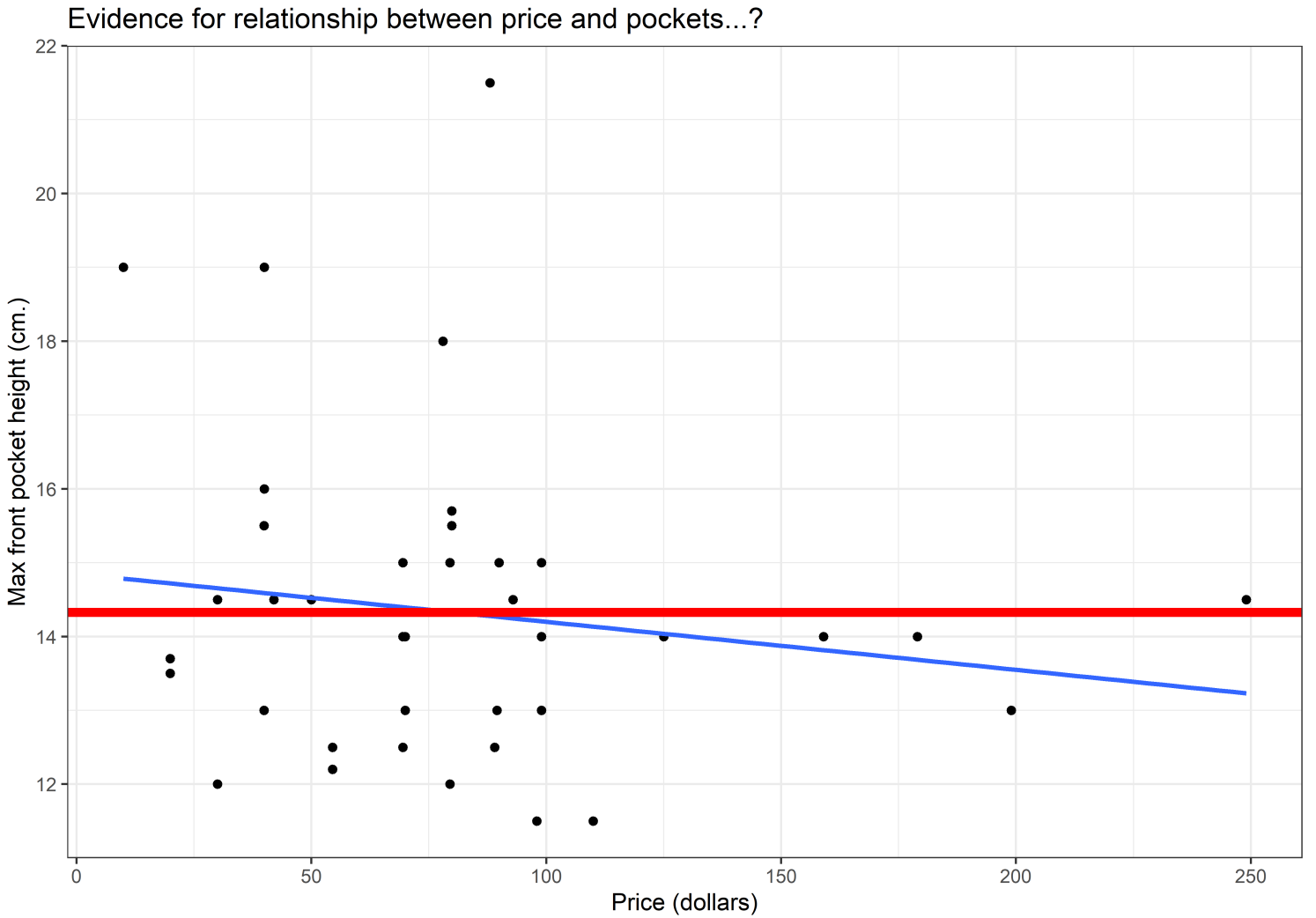
What do we expect if the null is true?



Making a decision

We reject the null hypothesis if the conditional probability of obtaining our test statistic, or more extreme, given it is true, is very small. This probability is called the p-value

Making a decision



Making a decision

What is very small? We often consider a cutpoint (the **significance level** or **discernability level**, commonly denoted by α) defined prior to conducting the analysis

Choosing a discernability level

This should always be done *before* seeing the data

What discernability level might you choose?

Making a decision

If the p-value is less than α, we say the results are statistically significant and we reject the null hypothesis. On the other hand, if the p-value is α or greater, we say the results are not statistically significant and fail to reject H_0.

What might we say if $p \geq \alpha$?

But wait...

We **never** "accept" the null hypothesis - we assumed that H_0 was true to begin with and assessed the probability of obtaining our test statistic (or more extreme) under this assumption

When we fail to reject the null hypothesis, we are stating that there is *insufficient evidence* to assert that it is false

But wait...

Importantly, we have assumed from the start that the null hypothesis is true, and the p-value calculates conditioned on that event

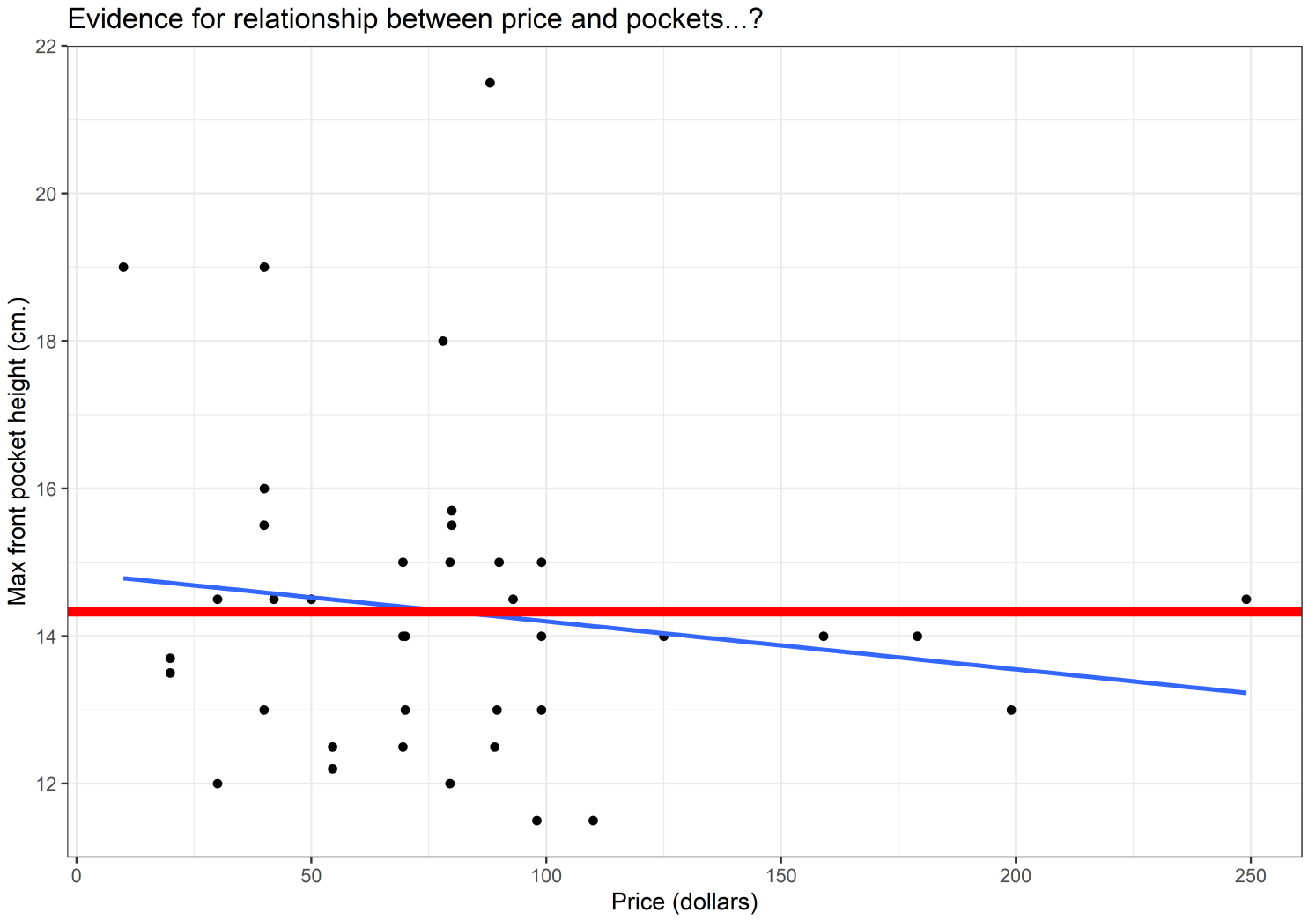
p-values do **NOT** provide information on the probability that the null hypothesis is true given our observed data

Assessing the evidence

Suppose our p-value was 0.392. What might you conclude?

What if it were 0.001?

Statistical significance vs. practical significance



Ultra-low dose contraception

Oral contraceptive pills work well, but must have a precise dose of estrogen.

If a pill has too high a dose, then women may risk side effects such as headaches, nausea, and rare but potentially fatal blood clots.

If a pill has too low a dose, then women may get pregnant.

Ultra-low dose contraception

A certain contraceptive pill is supposed to contain precisely 0.020 μg of estrogen. During QC, 50 randomly selected pills are tested, with sample mean dose 0.017 μg and sample SD 0.008 μg.

Do you think this is cause for concern? Why or why not? (Don't worry about calculations for now)

Defining the null and alternative hypotheses

Stated in words:

- H_0: The pills are consistent with a population that has a mean of 0.020 μg estrogen
- H_1: The pills are not consistent with a population that has a mean of 0.020 μg estrogen

Stated in symbols:

- $H_0 : \mu = 0.020$
- $H_1 : \mu \neq 0.020$,

where μ is the mean estrogen level of the manufactured pills, in μg

Collecting and summarizing the data

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

In our example, quality control technicians randomly selected a sample of 50 pills and calculated the sample mean $\bar{x} = 0.017$ μg and sample standard deviation $s = 0.008$ μg

Assessing the evidence observed

Next, we calculate the probability of getting data like ours, or more extreme, if H_0 were actually true

This is a conditional probability: *if* H_0 *were true* (i.e., if μ were truly 0.020), what would be the probability of observing $\bar{x} = 0.017$ and $s = 0.008$?

Again, this probability is the p-value

Making a decision

As it turns out, the probability of observing a sample mean of 0.017 and sample SD of 0.008 in 50 pills if H_0 were actually true is approximately 0.01.

What might we conclude?

What could go wrong?

Suppose we test the null hypothesis $H_0 : \mu = \mu_0$. We could potentially make two types of errors:

Truth	$\mu = \mu_0$	$\mu \neq \mu_0$
Fail to reject H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

- Type I Error: rejecting H_0 when it is actually true (falsely rejecting the null hypothesis)
- Type II Error: not rejecting H_0 when it is false (falsely failing to reject the null hypothesis)

While we of course want to know if any one study is showing us something real or a Type I or Type II error, hypothesis testing does NOT give us the tools to determine this

More on discernability levels...

What are the consequences for making each type of error here? What discernability level might you choose in each situation?

Power

Power is the probability of rejecting the null hypothesis when it is false (i.e., of avoiding a Type II error)

$$\text{Power} = P(\text{reject } H_0 | H_0 \text{ is false})$$

and can be also thought of as the likelihood a planned study will detect a deviation from the null hypothesis if one really exists. Power is a function of

- Sample size
- Deviation from the null one hopes to detect
- Variability in your data
- The discernability level you choose

Takeaways