The Weight of Evidence (The p-value is the Story)

In our last workshop, you built a mathematical engine to extract a p-value from an observed sample proportion. Now, we must learn to read what that value is trying to tell us. The p-value is a continuous measure of the **unlikelihood**. It answers one question: *Assuming the null hypothesis (H_0) is true, how likely is it that we see data as extreme as our sample?*

Barry the Statypus-in-Training says:

Okay, so I got a final computed p-value of 0.051. Dang it! This website online says that means my experiment completely failed, right? It's over 0.05, so H_0 is instantly, 100% completely true.

Sally the Statypus says: Hold your horses, Barry! You're falling right into the binary trap. A p-value is not a magic truth-meter. Is a p-value of 0.051 really different from a 0.049 in reality? No! The p-value tells you how much lingering belief you can afford to hold in your null assumption.

Instead of looking at hypothesis testing as a strict binary light switch, data scientists evaluate the strength of evidence based on how loudly the sample is screaming against the null hypothesis:

p-value Range	Evidence Tier	Practical Intuition
p-value > 0.10	**Weak Evidence**	Typical random sampling fluke. Nothing surprising.
$0.05 < p$-value ≤ 0.10	**Alright Evidence**	Close to being significant, but not quite.
$0.01 < p$-value ≤ 0.05	**Good Evidence**	Standard scientific detection indicator; clear deviation.
$0.001 < p$-value ≤ 0.01	**Great Evidence**	High-bar exploration; tough to explain away by chance.
p-value ≤ 0.001	**Amazing Evidence**	Exceptional, high-stakes confirmation; stunning anomaly.

Reflection: Pencil Friction: Evaluating the Scream of Data

Classify the objective strength of the evidence in the following independent research studies. Write a brief, plain-English statement summarizing how surprised an investigator should be.

Study 1: Testing a new tracking chip layout. The computed p-value is 0.0034.

Study 2: Testing whether a coin-flipping mechanism is biased. The computed p-value is 0.1240.

Study 3: Evaluating an alternative engine part design. The computed p-value is 0.0310.

The Gatekeeper (α is the Filter)

The concept of a **Level of Significance** is a rough concept, head into your book and read definitions 9.15, 9.16, and 9.17 before continuing.

Statypus Insight: Compressing the Spring

Our belief in H_0 can be thought of as the length of the spring. Prior to investigating the evidence it has a length of 1 and then gets diminished as we run the hypothesis test and the final length of the spring is precisely the p-value. α is our limbo stick. If your spring compresses below the α line, the test "passes" and we abandon our belief in H_0 and declare a change. If it doesn't get under the limbo stick, the evidence wasn't significant and we fail to convict. **You can view a visualization of this by scanning the QR code to the right.**

Barry the Statypus-in-Training says:

So alpha isn't calculated from the data? It's just a rule we choose? The exact same p-value could cause us to reject the null hypothesis in one situation but fail to reject it in another?

Bill the Statypus says: Exactly. In fact, the American Statistical Association says we shouldn't believe in any particular value of α. Go read the Remark below Definition 9.15 in the textbook.

Sally the Statypus says: Exactly, the better way of thinking about the Levels of Significance is to relate them to the "Evidence Tiers" we mentioned previously.

Bill the Statypus says: That's the best way of doing it! If data is sufficient at the $\alpha = 0.05$ level or significance, then it is **"Good Evidence."** If it passes at the $\alpha = 0.01$ level of significance, it is **"Great Evidence"** and if it passes at the $\alpha = 0.001$ level of significance, it is **"Amazing Evidence."** These are not the definitions, but a way for you to understand what we mean by the "Level of Significance."

Barry the Statypus-in-Training says:

So different values of α are more akin to different marks on a ruler and there isn't one "true" threshold that can be crosses.

Sally the Statypus says: That's it! If a sample offers significant evidence at the 0.01 level, it is better than evidence that is only significant with $\alpha = 0.05$.

Reflection: Pencil Friction: Moving the Gate

A bio-agricultural team tests an organic treatment on imported coffee beans to see if it reduces the mutation defect rate below the historical standard. Their data yields an objective evidence metric of **p-value = 0.042**. Evaluate the operational action of this exact same result under two entirely separate policy environments:

1. **Scenario A: Further Inquiry**
   This is a low-stakes scenario where we are simply curious if we should look further into the scenario and we use $\alpha = 0.10$.
   - Does the evidence compress the spring past the gate? (Is p-value $\leq \alpha$?) _____
   - Formal Statypus Action (Circle one): **REJECT** H_0 / **FAIL TO REJECT** H_0

2. **Scenario B: Making a Decision**
   If we decide that we can make a decision, it could cause massive logistical chain shutdowns to address the situation. As such, we want a much stricter threshold and we can use $\alpha = 0.01$.
   - Does the evidence compress the spring past the gate? (Is $p < \alpha$?) _____
   - Formal Statypus Action (Circle one): **REJECT** H_0 / **FAIL TO REJECT** H_0

The Big Takeaway: Did the underlying sample data or empirical evidence change between Scenario A and Scenario B? Explain why or why not, and describe how Barry's realization highlights the true purpose of alpha.

Barry the Statypus-in-Training says:

I think I get it now! If I just need decent evidence, I may consider $\alpha = 0.05$, but if I need to be very confidence, I would want to consider $\alpha = 0.01$ or $\alpha = 0.001$.

Sally the Statypus says: Exactly! The same evidence may be sufficient to simply take a second look but insufficient to shut down a machine and halt production.

Bill the Statypus says: The p-value is literally the strength of your evidence and the Level of Significance is just telling us how strict of evidence we want.

The Reality Check (Decisions and Errors)

Because α is a human policy boundary rather than a truth machine, our binary verdicts can sometimes clash with objective reality. When we choose how high or low to set the alpha line, we are balancing the risk of two fundamentally different operational failures.

Barry the Statypus-in-Training says:

Wait... if we make α really small because we want to be absolutely sure before shutting down the logistics line, doesn't that make it much harder to trip the spring? What if the coffee beans actually are heavily mutated, but our threshold is so strict that we fail to spot it?

Bill the Statypus says: You've just hit the trade-off of statistical decision-making. Read Section 9.3.4 in your textbook to learn about the different types of error we can make in a hypothesis test.

- A **Type I Error** happens if you reject H_0 when it was actually true. You believed the boy when there was no wofl. You stopped production for an organic batch that was totally normal.
- A **Type II Error** happens if you fail to reject H_0 when it was actually false. You didn't believe the boy and lost your sheep. You approved and shipped a bad, mutated batch.

Sally the Statypus says: It is clear that if you make it harder to make a Type I Error, that you are making it easier to make a Type II Error. This is yet another issue to consider when looking at statistical analyses. In fact, there is a further concept called the **Power of The Test** which measures how often a test will accurately reject H_0 when it is false.

Reflection: Publish or Perish

Unfortunately, there is a major portion of research that simply blindly uses $\alpha = 0.05$ as the threshold for "research level" evidence. A researcher is unlikely to get a paper published if their study results in insufficient evidence.
Question: If a researcher is judged based on how many papers they can publish, do you think they are incentivized to look for really strong evidence or only enough to get a paper published?

Sally the Statypus says: Do not get Bill started on the topic of **Data Dredging** or ***p*-Hacking**!.

Applications

What do you think the proportion of female babies is in the United States amongst all babies born? Most people would assume that the answer is $p = 0.5$, but what if someone claimed that $p < 0.5$? Would you believe them blindly, or would you demand evidence?
Question: What level of significance would it take to convince you that $p < 0.5$?

A sample of 2 girls and 3 boys should not be convincing, so we need more data. Using the CDC website, you can find that in 2024 there were 1,774,965 female babies born out of a total of 3,628,934 births. Open Section 9.3.3 of your textbook and load the Code Template to run a hypothesis test for a population proportion. Now input the values needed to run this hypothesis test. **Note: You cannot leave the commas in the values when you type them into R.**
Question: What is the resulting p-value?

Barry the Statypus-in-Training says:

How is that possible?

Bill the Statypus says: Well the p-value it found wasn't actually zero, but it was so small that R essentially can't tell it apart from 0. Run `?".Machine"` in your R console and look at the information about `double.eps`. Our p-value is really just smaller than this value.

Reflection: Exceptional Evidence

Given the p-value we just found, is there any level of significance that our data would fail to be sufficient at? What is the practical interpretation of this information?

Reflection: The Floating-Point Illusion

`double.eps` is the smallest positive floating-point number x such that $1 + x \neq 1$ and is typically 2.22×10^{-16}. You know that if x is positive, then $1 + x > 1$, but this points to a fact about computing. Theoretically, a decimal has an infinite number of digits. Realistically, we can only keep track of a finite number of digits.

1. Write out 2.22×10^{-16} in decimal notation to get a sense of just how small this value is.

2. Suppose R reports a p-value of exactly `0`. Does this mean the theoretical probability of observing data this extreme under the null hypothesis is truly, absolutely zero?

3. Under a smooth, continuous curve (like a Normal distribution), is it mathematically possible for an entire region or tail area stretching to infinity to have an area of exactly 0?

4. If a computer system cannot distinguish a number smaller than `double.eps` from zero, what is a more precise way to write a p-value that R reports as `0`? (Hint: Think about an inequality using `double.eps` or a very small threshold like 0.0001).

5. Why is it dangerous or misleading to write "$p = 0$" in a formal research report? What would a p-value of literally 0 imply about a sample?

Barry the Statypus-in-Training says:

Thank you so much, BIll and Sally! I think I understand the concept of a hypothesis test now.

Bill the Statypus says: Of course, Agent B... (coughs) I mean Barry.