Decisions and Uncertainty

References:

Chapter 11: Testing Hypotheses

https://www.inferentialthinking.com/chapters/intro

Overview

- Decisions and Uncertainty
- Framework for Decision Making
  - The Hypothesis and The Test Statistic
  - The Distribution of the Test Statistic Decisions and Uncertainty
  - Conclusion of a Test
- Meaning of Consistency
- Conventional Cut-offs and P-Value
- Error Probabilities
- Wrong Conclusions
- Data Snooping and P-Hacking

Decisions and Uncertainty

- We have seen several examples of assessing models that involve chance, by comparing observed data to the predictions made by the models.
- In all of our examples, there has been no doubt about whether the data were consistent with the model's predictions.
- The data were either very far away from the predictions, or very close to them.
- But outcomes are not always so clear cut. How far is "far"? Exactly what does "close" mean? While these questions don't have universal answers, there are guidelines and conventions that you can follow.
- Using statistical tests as a way of making decisions is standard in many fields and has a standard terminology.
- Need for a general framework of decision making, into which all our examples will fit.

General Framework

- Step 1: The Hypotheses
- Step 2: The Test Statistic
- Step 3: The Distribution of the Test Statistic, Under the Null Hypothesis
- Step 4. The Conclusion of the Test

Step 1: The Hypotheses: The null hypothesis.

- All statistical tests attempt to choose between two views of the world. Specifically, the choice is between two views about how the data were generated. These two views are called *hypotheses*.
  - This is a clearly defined model about chances.
  - It says that the data were generated at random under clearly specified assumptions about the randomness.
  - The word "null" reinforces the idea that if the data look different from what the null hypothesis predicts, the difference is due to *nothing* but chance.
  - From a practical perspective, **the null hypothesis is a hypothesis under which you can simulate data.**
- In the example about Mendel's model for the colors of pea plants, the null hypothesis is that the assumptions of his model are good: each plant has a 75% chance of having purple flowers, independent of all other plants.
- Under this hypothesis, we were able to simulate random samples, by using the (929, [0.75, 0.25]). We used a sample size of 929 because that's the number of plants Mendel grew

Step 1: The Hypotheses: The alternate hypothesis.

- This says that some reason other than chance made the data differ from the predictions of the model in the null hypothesis.
- In the example about Mendel's plants, the alternative hypothesis is simply that his model isn't good.

Step 2: The Test Statistic

- In order to decide between the two hypothesis, we must choose a statistic that we can use to make the decision. This is called the **test statistic**.
- Mendel's plants, our statistic |sample percent of purpleflowering plants−75|
- To see how to make the choice in general, look at the alternative hypothesis. What values of the statistic will make you think that the alternative hypothesis is a better choice than the null?
  - If the answer is "big values," you might have a good choice of statistic.
  - So also if the answer is "small values."
  - But if the answer is "both big values and small values," we recommend that you look again at your statistic and see if taking an absolute value can change the answer to just "big values".

Thus this indicates that the statistic should be the absolute value of *distance* between the sample percent and 75 and Big values of the distance will make you lean towards the alternative.

Step 2: Observed value of Test Statistic

- The **observed value of the test statistic** is the value of the statistic you get from the data in the study, **not a simulated value**.
- Among Mendel's 929 plants, 705 had purple flowers. The observed value of the test statistic was therefore

$$= |(100 \times (705/929) - 75)|$$

$$= 0.88805$$

Step 3: The Distribution of the Test Statistic, Under the Null Hypothesis

- The main computational aspect of a test of hypotheses is figuring out *what the values of the test statistic might be if the null hypothesis were true*.
- The test statistic is simulated based on the assumptions of the model in the null hypothesis.
- That model involves chance, so the statistic comes out differently when you simulate it multiple times.
- In other words, we get a good approximation to the probability distribution of the statistic, as predicted by the model in the null hypothesis.
- As with all distributions, it is very useful to visualize this distribution by a histogram

Step 4: The Conclusion of the Test

- The choice between the null and alternative hypotheses depends on the comparison between what you computed in Steps 2 and 3: the observed value of the test statistic and its distribution as predicted by the null hypothesis.
- If the two are consistent with each other, then the observed test statistic is in line with what the null hypothesis predicts.
- In other words, the test does not point towards the alternative hypothesis; the null hypothesis is better supported by the data. This was the case with the assessment of Mendel's model.
- But if the two are not consistent with each other, as is the case in our example about Alameda County jury panels, then the data do not support the null hypothesis (Something other than chance affected their composition.)
- If the data do not support the null hypothesis, we say that the test *rejects* the null hypothesis.

The Meaning of "Consistent"

- Alameda County juries: Our observed test statistic was far from what was predicted by the null hypothesis.
- Mendel’s experiment : Our observed statistic is consistent with the distribution that the null hypothesis predicts.
- But sometimes the decision is not so clear. Whether the observed test statistic is consistent with its predicted distribution under the null hypothesis is a matter of judgment.
- In such cases to make a judgment use the value of the test statistic and a graph of its predicted distribution under the null.
- That will allow your reader to make his or her own judgment about whether the two are consistent

The GSI's Defense : An example requiring judgement

- A Berkeley Statistics class of about 350 students was divided into 12 discussion sections led by Graduate Student Instructors (GSIs).
- After the midterm, students in **Section 3** noticed that their scores were on average lower than the rest of the class.

- In such situations, students tend to grumble about the section's GSI. Surely, they feel, there must have been something wrong with the GSI's teaching. Or else why would their section have done worse than others?
- The GSI, typically more experienced about statistical variation, often has a different perspective: if you simply draw a section of students at random from the whole class, their average score could resemble the score that the students are unhappy about, just by chance.

The GSI's position is a clearly stated chance model. We can simulate data under this model.

Applying the steps of the framework

- **Null Hypothesis.** The average score of the students in Section 3 is like the average score of the same number of students picked at random from the class.

- **Alternative Hypothesis.** No, it's too low.

- A natural statistic here is the average of the scores. Low values of the average will make us lean towards the alternative.

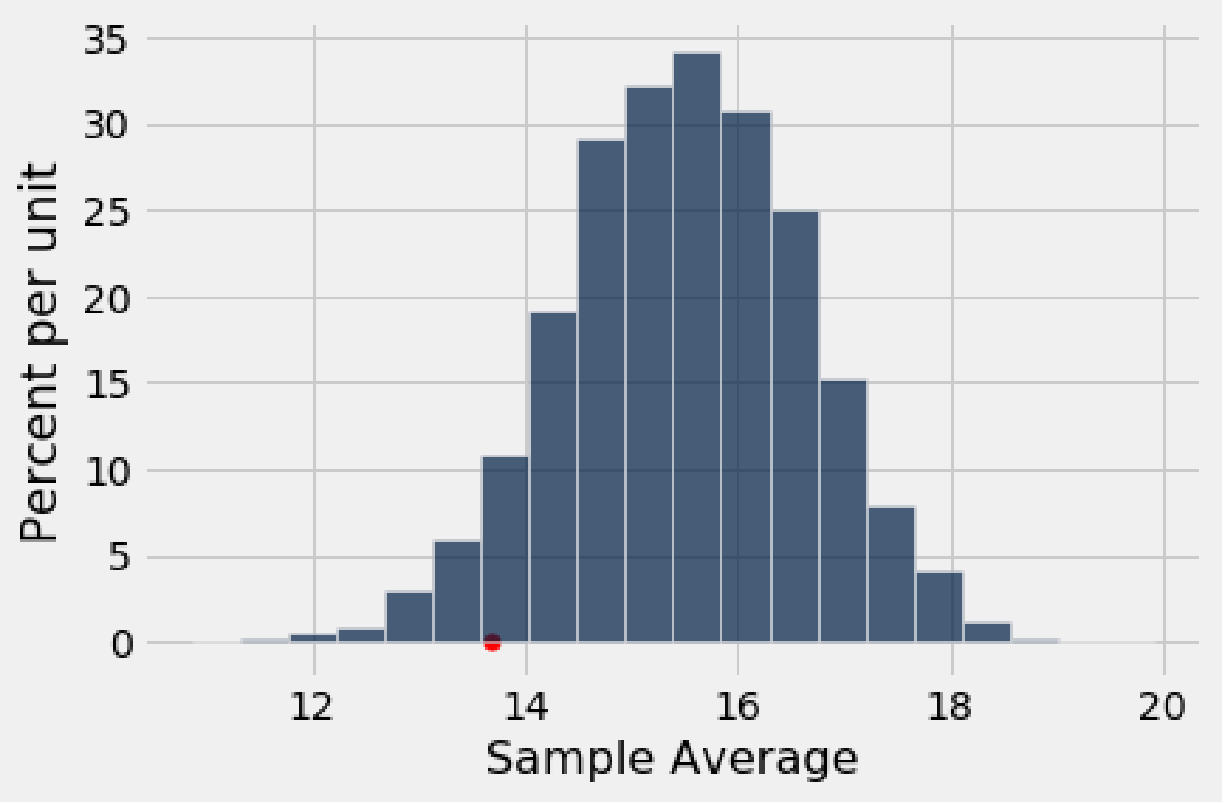
Average score in each section

- The average score of Section 3 is 13.667, which does look low compared to the other section averages. But is it lower than the average of a section of the same size selected at random from the class?
- To answer this, we can select a section at random from the class and find its average. T
- To select a section at random to we need to know how big Section 3 is, which we can by once again using group
- Section 3 had 27 students.

Section	Size	Midterm average
1	32	15.5938
2	32	15.125
3	27	13.6667
4	30	14.7667
5	33	17.4545
6	32	15.0312
7	24	16.625
8	29	16.3103
9	30	14.5667
10	34	15.2353
11	32	15.8077
12	32	15.7333

Empirical Distribution of Test Scores of Group 3

- Now we can figure out how to create one simulated value of our test statistic, the random sample average.
- First we have to select 27 scores at random without replacement and calculate the average
- Simulate the random sample average by repeating the calculation multiple times.
- Plot the histogram
- As we said earlier, small values of the test statistic will make us lean towards the alternative hypothesis, that the average score in the section is too low for it to look like a random sample from the class.
- Is the observed statistic of 13.667 "too low" in relation to this distribution? In other words, is the red far enough out into the left hand tail of the histogram for you to think that it is "too far"?



Conventional Cut-offs and the P-value

- If you don't want to make your own judgment, there are conventions that you can follow. These conventions tell us how far out into the tails is considered "too far".
- The conventions are based on the area in the tail, starting at the observed statistic (the red dot) and looking in the direction that makes us lean toward the alternative (the left side, in this example).
- If the area of the tail is small, the observed statistic is far away from the values most commonly predicted by the null hypothesis.
- In a histogram, area represents percent. To find the area in the tail, we have to find the percent of sample averages that were less than or equal to the average score of Section 3, where the red dot is (for the experiment with an observed statistic is 13.667 the average score of Section 3 to begin with, and had 10,000 repetitions of the random sampling)
- Area works out to **0.0564**
- About 5.7% of the simulated random sample averages were 13.667 or below. If we had drawn the students of Section 3 at random from the whole class, the chance that their average would be 13.667 or lower is about 5.7%.
- This chance has an impressive name. It is called the *observed significance level* of the test. That's a mouthful, and so it is commonly called the *P-value* of the test

P-value

- **Definition:** The P-value of a test is the chance, based on the model in the null hypothesis, that the test statistic will be equal to the observed value in the sample or even further in the direction that supports the alternative.
- If a P-value is small, that means the tail beyond the observed statistic is small and so the observed statistic is far away from what the null predicts.
- This implies that the data support the alternative hypothesis better than they support the null.
- How small is "small"? According to the conventions:
  - If the P-value is less than 5%, it is considered small and the result is called "statistically significant."
  - If the P-value is even smaller – less than 1% – the result is called "highly statistically significant."
- By this convention, our P-value of 5.7% is not considered small. So we have to conclude that the GSI's defense holds good – the average score of Section 3 is like those generated by random chance.
- Formally, the result of the test is not statistically significant.

When you make a conclusion in this way, we recommend that you don't just say whether or not the result is statistically significant. Along with your conclusion, provide the observed statistic and the P-value as well, so that readers can use their own judgment.

A note on Conventions

- Whether you use a conventional cutoff or your own judgment, it is important to keep the following points in mind.
  - Always provide the observed value of the test statistic and the P-value, so that readers can decide whether or not they think the P-value is small.
  - Don't look to defy convention only when the conventionally derived result is not to your liking.
  - Even if a test concludes that the data don't support the chance model in the null hypothesis, it typically doesn't explain *why* the model doesn't work

Error Probabilities

- In the process by which we decide which of two hypotheses is better supported by our data, the final step involves a judgment about the consistency of the data and the null hypothesis.

- While this step results in a good decision a vast majority of the time, it can sometimes lead us astray. The reason is chance variation.
  - For example, even when the null hypothesis is true, chance variation might cause the sample to look quite different from what the null hypothesis predicts.

Wrong Conclusions

	Null is True	Alternative is True
Test Favors the Null	Correct result	Error
Test Favors the Alternative	Error	Correct result

- Since the null hypothesis is a completely specified chance model, we can estimate the chance of the first type of error. The answer turns out to be essentially the cutoff that we use for the P-value. Let's see how.

The Chance of an Error

Suppose you want to test whether a coin is fair or not. Then the hypotheses are:

Null: The coin is fair. That is, the results are like draws made at random with replacement from *Heads, Tails*.

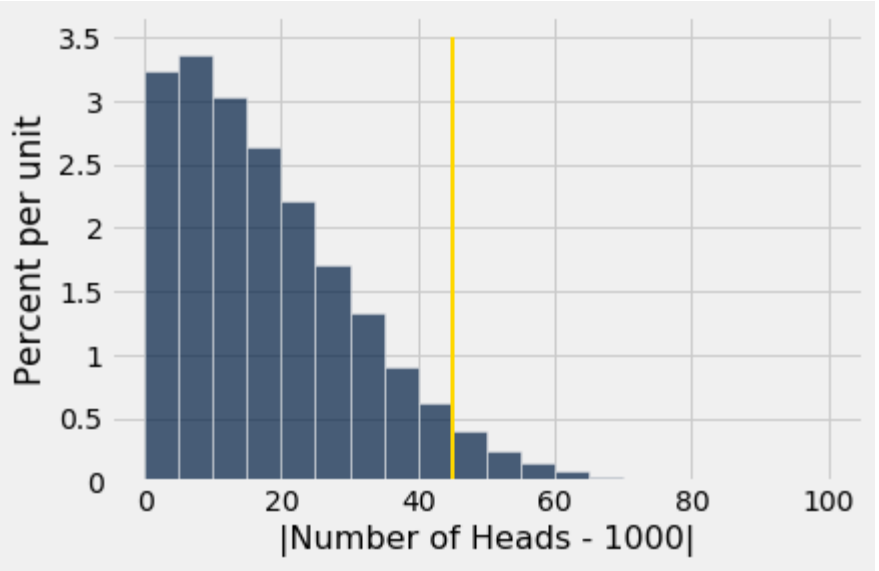
Alternative: The coin is not fair.

Suppose you are going to test this hypothesis based on 2000 tosses of the coin. You would expect a fair coin to land heads 1000 times out of 2000, so a reasonable test statistic to use is

$$\text{test statistic} = |\ \text{number of heads} - 1000\ |$$

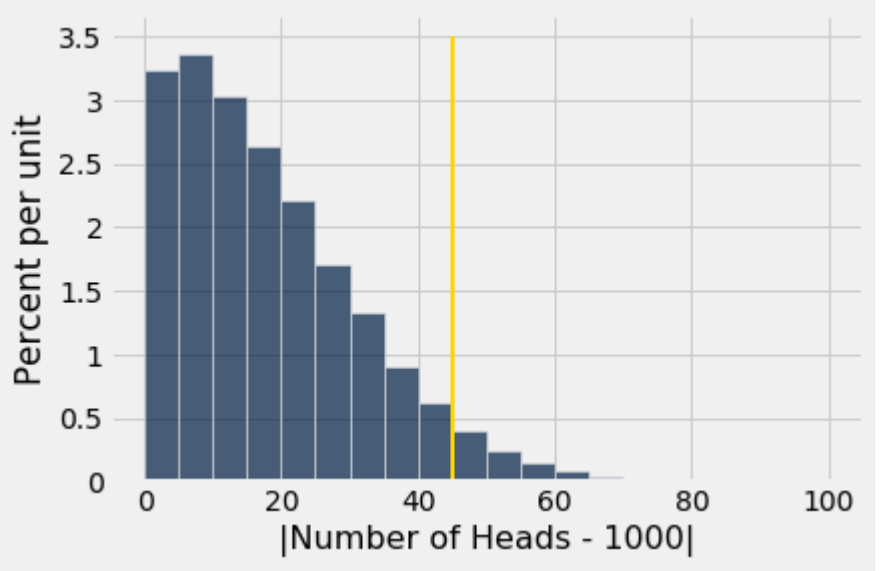
Small values of this statistic favor the null hypothesis, and large values favor the alternative.

We have simulated this statistic under the null hypothesis many times, and drawn its empirical distribution.



The Chance of an Error

- The area to the right of 45 (where the gold line is) is about 5%.
- Large values of the test statistic favor the alternative.
- So if the test statistic comes out to be 45 or more, the test will conclude that the coin is unfair.
- However, as the figure shows, a fair coin can produce test statistics with values 45 or more. In fact it does so with chance about 5%.
- So *if the coin is fair* and our test uses a 5% cutoff for deciding whether it is fair or not, then there is about a 5% chance that the test will wrongly conclude that the coin is unfair.



The Cutoff for the P-value is an Error Probability

- If you use a **p % cutoff** for the P-value, and the null hypothesis happens to be true, then there is about a **p % chance** that your test will conclude that the alternative is true.
- The 1% cutoff is therefore more conservative than 5%. There is less chance of concluding "alternative" if the null happens to be true. **For this reason, randomized controlled trials of medical treatments usually use 1% as the cutoff** for deciding between the following two hypotheses:

  Null Hypothesis. The treatment has no effect; observed differences between the outcomes of the treatment and control groups of patients are due to randomization.

  Alternative Hypothesis. The treatment has an effect.
- The idea is to control the chance of concluding the treatment does something if in fact it does nothing. This reduces the risk of giving patients a useless treatment.
- Still, even if you set the cutoff to be as low as 1%, and the treatment does nothing, there is about a 1% chance of concluding that the treatment does something. This is due to chance variation.
- There is a small chance that data from random samples end up leading you astray.

Data Snooping and P-Hacking

- Thus if each of 100 different research groups runs a separate randomized controlled experiment about the effect of a treatment that in fact has no effect, and each experiment uses a 1% cutoff for the P-value, then by chance variation, one of the experiments is expected to wrongly conclude that the treatment does have an effect.
- Unfortunately, that could be the one that gets published. This is why it is important that experiments be *replicated*. That is, other researchers ought to be able to carry out the experiment and see if they get similar results.
- It is not uncommon for researchers to test multiple hypotheses using the same data. For example, in a randomized controlled trial about the effect of a drug, researchers might test whether the drug has an effect on various different diseases.

Data Snooping and P-Hacking

- So, when you read a study that uses tests of hypotheses and concludes that a treatment has an effect, always ask how many different effects were tested before the researchers found the one that was reported
- If the researchers ran multiple different tests before finding one that gave a "highly statistically significant" result, use the result with caution. The study could be marred by *data snooping*, which essentially means torturing the data into making a false confession. This is sometimes also called *p-hacking*.
- In such a situation, one way to validate the reported result is by replicating the experiment and testing for that particular effect alone. If it comes out significant again, that will validate the original conclusion.