PSY 503: Foundations of Psychological Methods
Lecture 13: Hypothesis Testing I

Robin Gomila

Princeton

October 19, 2020

Research question

Research question

Can Paul the Octopus forecast the results of soccer games?

https://www.youtube.com/watch?v=3ESGpRUMj9E

Hypothesis of journalists and fans

- Journalists and fans think:
  - “Paul is an extraordinary octopus”
- Hypothesis:
  - Paul can see the future

Paul's entire career

Paul's entire career

Opponent	Tournament	Stage	Date	Prediction	Result	Outcome
Poland	Euro 2008	group stage	8 June 2008	Germany	2–0	Correct
Croatia	Euro 2008	group stage	12 June 2008	Germany[3][20]	1–2	Incorrect
Austria	Euro 2008	group stage	16 June 2008	Germany	1–0	Correct
Portugal	Euro 2008	quarter-finals	19 June 2008	Germany	3–2	Correct
Turkey	Euro 2008	semi-finals	25 June 2008	Germany	3–2	Correct
Spain	Euro 2008	final	29 June 2008	Germany[3]	0–1	Incorrect
Australia	World Cup 2010	group stage	13 June 2010	Germany[31]	4–0	Correct
Serbia	World Cup 2010	group stage	18 June 2010	Serbia[31]	0–1	Correct
Ghana	World Cup 2010	group stage	23 June 2010	Germany[31]	1–0	Correct
England	World Cup 2010	round of 16	27 June 2010	Germany[32]	4–1	Correct
Argentina	World Cup 2010	quarter-finals	3 July 2010	Germany[23]	4–0	Correct
Spain	World Cup 2010	semi-finals	7 July 2010	Spain[33]	0–1	Correct
Uruguay	World Cup 2010	3rd place play-off	10 July 2010	Germany	3–2	Correct

Paul's entire career

Opponent	Tournament	Stage	Date	Prediction	Result	Outcome
Poland	Euro 2008	group stage	8 June 2008	Germany	2–0	Correct
Croatia	Euro 2008	group stage	12 June 2008	Germany[3][20]	1–2	Incorrect
Austria	Euro 2008	group stage	16 June 2008	Germany	1–0	Correct
Portugal	Euro 2008	quarter-finals	19 June 2008	Germany	3–2	Correct
Turkey	Euro 2008	semi-finals	25 June 2008	Germany	3–2	Correct
Spain	Euro 2008	final	29 June 2008	Germany[3]	0–1	Incorrect
Australia	World Cup 2010	group stage	13 June 2010	Germany[31]	4–0	Correct
Serbia	World Cup 2010	group stage	18 June 2010	Serbia[31]	0–1	Correct
Ghana	World Cup 2010	group stage	23 June 2010	Germany[31]	1–0	Correct
England	World Cup 2010	round of 16	27 June 2010	Germany[32]	4–1	Correct
Argentina	World Cup 2010	quarter-finals	3 July 2010	Germany[23]	4–0	Correct
Spain	World Cup 2010	semi-finals	7 July 2010	Spain[33]	0–1	Correct
Uruguay	World Cup 2010	3rd place play-off	10 July 2010	Germany	3–2	Correct

11/13 correct results!

Scientific approach: Thinking process

Scientific approach: Thinking process

- What does 11/13 mean in this context?

Scientific approach: Thinking process

- What does 11/13 mean in this context?
- Luck or ability?
  - If luck: Paul is ordinary
  - Otherwise, Paul is extraordinary

Scientific approach: Thinking process

- What does 11/13 mean in this context?
- Luck or ability?
  - If luck: Paul is ordinary
  - Otherwise, Paul is extraordinary
- We know what to expect if Paul's ordinary

Null Hypothesis

H_0:

Null Hypothesis

H_0: Paul is ordinary (i.e., he does not have special ability to predict outcome of soccer games)

Step 1: Suppose the null is true

Step 1: Suppose the null is true

- I REPEAT: SUPPOSE THE NULL IS TRUE

Step 1: Suppose the null is true

- I REPEAT: SUPPOSE THE NULL IS TRUE
  - Keep in mind that we are operating under the assumption that the null is true

Step 1: Suppose the null is true

- I REPEAT: SUPPOSE THE NULL IS TRUE
  - Keep in mind that we are operating under the assumption that the null is true
  - This is perhaps the most important part because if you forget this, it will be hard to understand what's a p-value

Step 2: What do we expect to see under the null

Step 2: What do we expect to see under the null

- Ordinary Paul has a 50% chance to predict the outcome of a game

Pause: Intuition

Pause: Intuition

- Would any of the following results make you doubt that the null is, in fact, true?

Pause: Intuition

- Would any of the following results make you doubt that the null is, in fact, true?
  - Paul made two correct guesses out of 2 trials

Pause: Intuition

- Would any of the following results make you doubt that the null is, in fact, true?
  - Paul made two correct guesses out of 2 trials
  - Paul made 5 correct guesses out of 13 trials

Pause: Intuition

- Would any of the following results make you doubt that the null is, in fact, true?
  - Paul made two correct guesses out of 2 trials
  - Paul made 5 correct guesses out of 13 trials
  - Paul made 55 correct guesses out of 100 trials

Pause: Intuition

- Would any of the following results make you doubt that the null is, in fact, true?
  - Paul made two correct guesses out of 2 trials
  - Paul made 5 correct guesses out of 13 trials
  - Paul made 55 correct guesses out of 100 trials
  - Paul made 10,000 correct guesses out of 10,000 trials

Step 3: Does the null sound right, conditional on the observed data?

- Have I observed data that make the null sound implausible? Do I have empirical evidence that contradicts the null?

Step 3: Does the null sound right, conditional on the observed data?

- Have I observed data that make the null sound implausible? Do I have empirical evidence that contradicts the null?
- Hypothesis testing will not allow you to prove that the null is or is not true

Step 3: Does the null sound right, conditional on the observed data?

- Have I observed data that make the null sound implausible? Do I have empirical evidence that contradicts the null?
- Hypothesis testing will not allow you to prove that the null is or is not true
- Hypothesis testing results in one of two statements:
    - I haven't observed data that suggest that the null is not true
    - I have observed data that suggest that the null is not true

How unlikely is 11/13 UNDER THE NULL?

- How likely is it that Paul just got super lucky?

How unlikely is 11/13 UNDER THE NULL?

- How likely is it that Paul just got super lucky?
  - 1.1% chance of getting **at least** 11/13 correct guesses

How unlikely is 11/13 UNDER THE NULL?

- How likely is it that Paul just got super lucky?
  - 1.1% chance of getting **at least** 11/13 correct guesses
  - That's the p-value ($p = .011$)

How unlikely is 11/13 UNDER THE NULL?

- How likely is it that Paul just got super lucky?
  - 1.1% chance of getting **at least** 11/13 correct guesses
  - That's the p-value ($p = .011$)
- Do we reject the null?

How unlikely is 11/13 UNDER THE NULL?

- How likely is it that Paul just got super lucky?
  - 1.1% chance of getting **at least** 11/13 correct guesses
  - That's the p-value ($p = .011$)
- Do we reject the null?
  - Sure

How unlikely is 11/13 UNDER THE NULL?

- How likely is it that Paul just got super lucky?
  - 1.1% chance of getting **at least** 11/13 correct guesses
  - That's the p-value ($p = .011$)
- Do we reject the null?
  - Sure
- Does that mean that the null is untrue / Paul is extraordinary?

How unlikely is 11/13 UNDER THE NULL?

- How likely is it that Paul just got super lucky?
  - 1.1% chance of getting **at least** 11/13 correct guesses
  - That's the p-value ($p = .011$)
- Do we reject the null?
  - Sure
- Does that mean that the null is untrue / Paul is extraordinary?
  - No

False positives, file drawer, and publication bias

- During the world cup, if 5,000 animals were filmed trying to make guesses about soccer games,

False positives, file drawer, and publication bias

- During the world cup, if 5,000 animals were filmed trying to make guesses about soccer games, we would expect that around 50 of them would do at least as well as Paul!

False positives, file drawer, and publication bias

- During the world cup, if 5,000 animals were filmed trying to make guesses about soccer games, we would expect that around 50 of them would do at least as well as Paul!
- False positive: Result that indicates that a given condition exists, when it does not

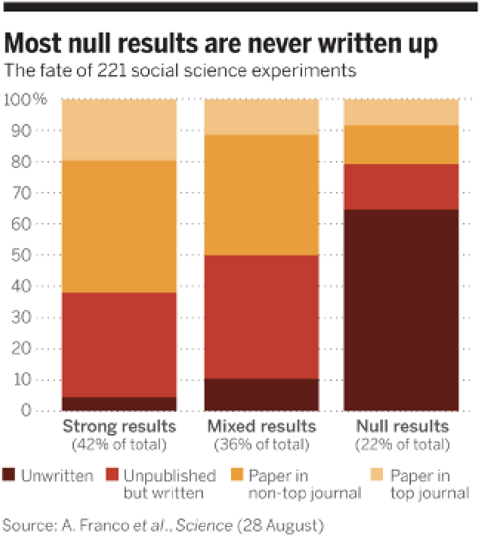
False positives, file drawer, and publication bias

- During the world cup, if 5,000 animals were filmed trying to make guesses about soccer games, we would expect that around 50 of them would do at least as well as Paul!
- False positive: Result that indicates that a given condition exists, when it does not
- Paul is (most likely) a false positive

False positives, file drawer, and publication bias

- During the world cup, if 5,000 animals were filmed trying to make guesses about soccer games, we would expect that around 50 of them would do at least as well as Paul!
- False positive: Result that indicates that a given condition exists, when it does not
- Paul is (most likely) a false positive
- Problem is: Most results remain in the file drawer
    - We only hear about Paul, not about the thousands of others that were also "tested"

Publication Bias



Hypothesis testing: The general framework

Proof by contradiction

General strategy of mathematical proof

Proof by contradiction

General strategy of mathematical proof that consists in demonstrating that assuming the contrary of what we would like to prove leads to a logical contradiction

Proof by contradiction

A way to prove that something is true by showing that if it wasn't true, that would lead to a logical error

Proof by contradiction: Intuition

- Suppose you're accused of robbing a bank

Proof by contradiction: Intuition

- Suppose you're accused of robbing a bank
- Using proof by contradiction, you could argue for your innocence in the following way:

Proof by contradiction: Intuition

- Suppose you're accused of robbing a bank
- Using proof by contradiction, you could argue for your innocence in the following way:
  - Assume the opposite of what you'd like to prove: Assume you actually robbed a bank

Proof by contradiction: Intuition

- Suppose you're accused of robbing a bank
- Using proof by contradiction, you could argue for your innocence in the following way:
  - Assume the opposite of what you'd like to prove: Assume you actually robbed a bank
  - Show that this imply something that is wrong: This imply that you would not be on a plane during the robbery

Proof by contradiction: Intuition

- Suppose you're accused of robbing a bank
- Using proof by contradiction, you could argue for your innocence in the following way:
  - Assume the opposite of what you'd like to prove: Assume you actually robbed a bank
  - Show that this imply something that is wrong: This imply that you would not be on a plane during the robbery
  - Concluding statement: I was on a plane at the time of the robbery, therefore I did not rob a bank

Proof by contraction in mathematics

- Prove by contradiction that **there is no smallest positive rational number**

Proof by contraction in mathematics

- Prove by contradiction that **there is no smallest positive rational number**
  - A rational number is simply a fraction p/q where the numerator and nonzero denominator are whole numbers
  - e.g., $1/2$, $17/25$, $4/1000$

Proof by contraction in mathematics

- Prove by contradiction that **there is no smallest positive rational number**
  - A rational number is simply a fraction p/q where the numerator and nonzero denominator are whole numbers
  - e.g., $1/2$, $17/25$, $4/1000$
- Assume there exist a smallest positive rational number $a = p/q$

Proof by contraction in mathematics

- Prove by contradiction that **there is no smallest positive rational number**
  - A rational number is simply a fraction p/q where the numerator and nonzero denominator are whole numbers
  - e.g., $1/2$, $17/25$, $4/1000$
- Assume there exist a smallest positive rational number $a = p/q$
- We can divide a by 2, in which case we get another rational number $b = \frac{p}{2q}$

Proof by contraction in mathematics

- Prove by contradiction that **there is no smallest positive rational number**
  - A rational number is simply a fraction p/q where the numerator and nonzero denominator are whole numbers
  - e.g., $1/2$, $17/25$, $4/1000$
- Assume there exist a smallest positive rational number $a = p/q$
- We can divide a by 2, in which case we get another rational number $b = \frac{p}{2q}$
- b is smaller than a and rational because p and $2q$ are whole numbers

Proof by contraction in mathematics

- Prove by contradiction that **there is no smallest positive rational number**
  - A rational number is simply a fraction p/q where the numerator and nonzero denominator are whole numbers
  - e.g., $1/2$, $17/25$, $4/1000$
- Assume there exist a smallest positive rational number $a = p/q$
- We can divide a by 2, in which case we get another rational number $b = \frac{p}{2q}$
- b is smaller than a and rational because p and $2q$ are whole numbers
- Conclusion: we found a positive rational number smaller than p/q, which contradicts the initial statement. Therefore, assumption that there exist a smallest positive rational number must be false. This means that there is no smallest positive number.

Proof by contraction in statistical hypothesis testing

Proof by contraction in statistical hypothesis testing

- We can never reject a hypothesis with 100% certainty
- We use a probabilistic version of proof by contradiction

Proof by contraction in statistical hypothesis testing

- We can never reject a hypothesis with 100% certainty
- We use a probabilistic version of proof by contradiction
- **Step 1:** We begin by assuming a hypothesis that we would like to eventually refute: The **Null Hypothesis (H_0)**

Proof by contraction in statistical hypothesis testing

- We can never reject a hypothesis with 100% certainty
- We use a probabilistic version of proof by contradiction
- **Step 1:** We begin by assuming a hypothesis that we would like to eventually refute: The **Null Hypothesis (H_0)**
  - Paul the octopus: **sharp null hypothesis.** i.e., all the potential outcomes for each observation are determined under this hypothesis

Proof by contraction in statistical hypothesis testing

- We can never reject a hypothesis with 100% certainty
- We use a probabilistic version of proof by contradiction
- **Step 1:** We begin by assuming a hypothesis that we would like to eventually refute: The **Null Hypothesis (H_0)**
  - Paul the octopus: **sharp null hypothesis.** i.e., all the potential outcomes for each observation are determined under this hypothesis
- **Step2:** We choose a test statistic
  - e.g., number of correct results of soccer games

Proof by contraction in statistical hypothesis testing

- We can never reject a hypothesis with 100% certainty
- We use a probabilistic version of proof by contradiction
- **Step 1:** We begin by assuming a hypothesis that we would like to eventually refute: The **Null Hypothesis (H_0)**
  - Paul the octopus: **sharp null hypothesis.** i.e., all the potential outcomes for each observation are determined under this hypothesis
- **Step2:** We choose a test statistic
  - e.g., number of correct results of soccer games
- **Step 3:** We derive the sampling distribution of the test statistic
  - e.g., the probability of each possible number of correct guesses

Proof by contraction in statistical hypothesis testing

- **Step 4:** We ask whether the observed value OR more more extreme values of the test statistic are likely to occur under the sampling distribution
  - e.g., under the null, how likely is 11/13 or more than 11/13 correct guesses?

Proof by contraction in statistical hypothesis testing

- **Step 4:** We ask whether the observed value OR more more extreme values of the test statistic are likely to occur under the sampling distribution
  - e.g., under the null, how likely is 11/13 or more than 11/13 correct guesses?
- **Step 5:** If it is unlikely, then we "reject the null hypothesis"
  - Otherwise, we "fail to reject the null hypothesis" or "retain the null hypothesis"

Proof by contraction in statistical hypothesis testing

- Rejecting the null does not imply that the null is untrue

Proof by contraction in statistical hypothesis testing

- Rejecting the null does not imply that the null is untrue
- Failing to reject the null does not imply that the null is correct
  - Some degree of consistency between the data and the null
  - Different opinions on this

How should we quantify the degree to which the observed value of the test statistic is unlikely to occur?

P-value

P-value

The p-value can be understood as the probability that under the null hypothesis, we observe a value of the test statistic at least as extreme as the one we actually observed.

P-value

- Smaller p-value provides stronger evidence against the null hypothesis

P-value

- Smaller p-value provides stronger evidence against the null hypothesis
- The p-value **does NOT** represent the probability that the null is true
  - The probability that the null is true is either 0 or 1
  - That's because the null is either true or false

Rejecting the null based on the p-value

- Specify the level of test α (i.e., same α as confidence interval)
- If p-value less than or equal to α, we reject the H_0

What's in α?

- The probability of false rejection of the null
  - i.e., the null is true but we reject it
  - i.e., the probability of *false positive* (a.k.a, *type I error*)

False positives and false negatives: A trade off

- Minimizing false positives usually increase the risk of false negatives
  - It is not possible to directly control for the probability of *false negatives*

False positives and false negatives: A trade off

- Minimizing false positives usually increase the risk of false negatives
  - It is not possible to directly control for the probability of *false negatives*
- Suppose we set α as **very very very** low to minimize the risk of false positives
  - If the null hypothesis is true:
    - Great! We almost never reject the null, almost no false positive results
  - If the null hypothesis is false:
    - Not so great! We still almost never reject the null, almost always a false negative

One-sided vs. Two-sided alternative hypothesis

- Alternative hypothesis is the complement of the null hypothesis

One-sided vs. Two-sided alternative hypothesis

- Alternative hypothesis is the complement of the null hypothesis
- Paul the octopus: alternative hypothesis is one-sided
  - We ignore extreme values on the other side of the distribution of the test statistic
  - We calculate a *one-sided* (or *one tailed*) p-value

One-sided vs. Two-sided alternative hypothesis

- Alternative hypothesis is the complement of the null hypothesis
- Paul the octopus: alternative hypothesis is one-sided
  - We ignore extreme values on the other side of the distribution of the test statistic
  - We calculate a *one-sided* (or *one tailed*) p-value
- Most often, alternative hypotheses are two-sided and we compute two-sided p-values
  - We take into account extreme values on both sides
  - For a given level α, a two-tailed p-value is equal to the one-tailed p-value multiplied by 2

What do you think about hypothesis testing?

What do you think about hypothesis testing?

- It is an imperfect method and p-values have been criticized

What do you think about hypothesis testing?

- It is an imperfect method and p-values have been criticized
- But hypothesis testing has saved many many many lives and allowed scientists to accumulate knowledge about the world

What do you think about hypothesis testing?

- It is an imperfect method and p-values have been criticized
- But hypothesis testing has saved many many many lives and allowed scientists to accumulate knowledge about the world
- What's important is to (deeply) understand hypothesis testing, its limitations, what it allows you to do and not to do, p-values, what to conclude and not conclude from a study

What do you think about hypothesis testing?

- It is an imperfect method and p-values have been criticized
- But hypothesis testing has saved many many many lives and allowed scientists to accumulate knowledge about the world
- What's important is to (deeply) understand hypothesis testing, its limitations, what it allows you to do and not to do, p-values, what to conclude and not conclude from a study
- Keep in mind: p-values are not substantive quantities of interest
  - In fact, we may believe that the null hypothesis never holds true
  - Importance of effect sizes!

What do you think about hypothesis testing?

- It is an imperfect method and p-values have been criticized
- But hypothesis testing has saved many many many lives and allowed scientists to accumulate knowledge about the world
- What's important is to (deeply) understand hypothesis testing, its limitations, what it allows you to do and not to do, p-values, what to conclude and not conclude from a study
- Keep in mind: p-values are not substantive quantities of interest
  - In fact, we may believe that the null hypothesis never holds true
  - Importance of effect sizes!
- Thoughts?