STA201 Week 8: Statistical Inference, Statistics in the News, and Risk

Robert Zimmerman

University of Toronto

February 25, 2019

Last Week

- Basic statistics (mean, standard deviation, quantiles, etc.)
- Histograms
- Normal distribution

This Week

1. Statistical Inference
   - Statistical Hypotheses
   - NHST Framework
   - Statistical Assumptions
   - *p*-values
2. Statistics in the News
   - The Replication Crisis
   - Poor Assumptions
3. Financial Risk
   - Risk
   - The 2008 Financial Crisis

Why Statistics?

Statistical Inference

- We use statistics to extract information from our observations (raw data)
- We use mathematical techniques to ask what scientific claims we can infer from the raw data
- We can't "prove" or "disprove" what we can't fully observe (different from logical statements)
- We can only start with a hypothesis and then ask whether or not our observations support it

Statistical Hypotheses I: H_0 and H_A

- Hypotheses must be *testable* – they have to be expressible in statistical language ← Usually, a mathematical statement
- Example: The average daily snowfall amount in February in Toronto is 10 cm
- Non-example: It's snowing too much today!

Statistical Hypotheses II: H_0 and H_A

"H-naught"

- We start with a *null hypothesis* and an *alternative hypothesis* which we label H_0 and H_A, respectively
- H_0 represents the default scenario: nothing interesting is going on
- H_A represents the alternative situation: we suspect something interesting *is* going on
- H_A and H_0 can never both be true at the same time
- Often, H_A and H_0 are "opposites" – either one is true or the other is true (but this isn't a requirement!)
- For example:

H_0: There tends to be equal amounts of snowfall on Wednesdays and Tuesdays

H_A: There tends to be more snowfall on Wednesdays than on Tuesdays

Statistical Hypotheses III: More Examples

- Coin flip experiment: $H_0: p = 1/2$ (ie, the coin is unbiased)

Let p = probability of heads. $H_A: p \neq 1/2$ (ie, the coin is biased)

- A clinical drug trial compares the effectiveness of a drug with that of a placebo to prevent heart attacks:

H_0: this drug is no different from the placebo at preventing heart attacks

H_A: this drug is better than the placebo at preventing heart attacks

PollEverywhere - Statistical Hypotheses

Which of the following is/are *not* a proper H_0-H_A pair?

- H_0: The average Ontarian is more than 50% likely to vote for a Liberal candidate for Prime Minister
  H_A: The average Ontarian is at most 50% likely to vote for a Liberal candidate for Prime Minister
- H_0: If you buy Glop, you'll love its smooth texture
  H_A: If you buy Glop, you won't love its smooth texture
- H_0: There tend to be more red M&Ms in a pack than yellow M&Ms
  H_A: There tend to be equal amounts of red and yellow M&Ms in each pack
- H_0: The average person in this room is at least 3 feet tall
  H_A: The average person in this room is at least 4 feet tall

NHST Framework I

The *Null Hypothesis Significance Testing (NHTS)* framework consists of the following steps:

"alpha"

1) We state H_0 and H_A, our statistical assumptions, and prespecify a **significance level** α between 0 and 1 (usually close to 0)
2) We collect our data (observations)
3) We ask the **key question**: "If H_0 is true, what is the probability that we would observe data at least as extreme as this?"
4) We calculate this probability = probability of observing data at least as extreme, assuming H_0 is true
5) We apply a rule: If the calculated probability is less than the significance **level** α, we reject H_0 – i.e., the data we see is "too unlikely" to arise merely by chance (otherwise, we fail to reject H_0)

The order of these steps is *crucial* (why?)

NHST Framework II

Statistical Assumptions - Independence I

- In most studies, the observations are assumed to be **independent** (recall that two events are independent when the likelihood of one event occurring has no affect on the likelihood of the other event occurring)
- In terms of our data: none of the observations are affected by one another
- This is a critical assumption: without it, we would have to quantify *how* the observations depend on each other in order to perform statistical calculations

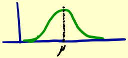
Statistical Assumptions - Independence II

- Is it reasonable to assume the observations are independent?
- Sometimes: for example, estimating the bias of a (possibly biased) coin by flipping it – no flip affects the outcome of any other flip
- Sometimes not: for example, what is the probability of a random man on the street being color-blind? What is the probability of a color-blind man's brother being color-blind?
- Ensuring that an experiment is properly set up (in advance!) to satisfy this assumption is an important process called *experimental design*

Statistical Assumptions - Independence III

Statistical Assumptions - Distributions I

- When we frame H_0 and H_A, we make the **statistical assumption** that the data follow some kind of underlying distribution with defining features (i.e., **parameters**) that we are trying to estimate
- For example: recall that the Normal distribution has two parameters: the mean μ "mu" and standard deviation σ "sigma"



- A study of heights in adult males might conclude that the average (mean) height of the adult male population of Earth is 66 inches
- Assumption: Adult male heights tend to follow a $\overbrace{N(\mu, \sigma)}^{\text{"Normal distribution with mean } \mu \text{ and standard deviation } \sigma\text{"}}$ distribution
- Claim: $\mu = 66$ (inches)

Statistical Assumptions - Distributions II

- Of course, we don't know for sure what distribution the data arises from – there are *lots* more than just the Normal
- So how can we know which distribution to even *assume* the data arises from in the NHST framework?
- Sometimes it's straightforward, as in the case of coin flips (we have a distribution to describe exactly that situation)
- When we know very little about the data, it can be hard to even make a reasonable guess

- Fortunately, **probability theory** gives us very powerful tools to help

Definition

If $X_1, X_2, \ldots, X_n$ represent n numerical observations, then we define the **sample mean** $\bar{X}_n$ to be

$$\bar{X}_n = \frac{1}{n}\overbrace{(X_1 + X_2 + X_3 + \cdots + X_n)}$$

Eg: data:
$X_1 = 9$
$X_2 = 10$
$X_3 = 7$
$X_4 = 12$

Then $\bar{X}_4 = \frac{1}{4}(9+10+7+12)$
$= \frac{1}{4}(38)$
$= 9.5$

Eg: all we know about the data is that $X_1 = -X_2$ and $X_3 = -X_4$. Then $\bar{X}_4 = 0$. Why?

b/c $\bar{X}_4 = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$
$= \frac{1}{4}(\underbrace{X_1 + (-X_1)}_{=0} + \underbrace{X_3 + (-X_3)}_{=0})$
$= \frac{1}{4}(0+0)$
$= 0.$

PollEverywhere - Sample Mean I

Suppose $X_1 = 3$ and $X_3 = 3$ and the sample mean $\bar{X}_3 = 5$. What must X_2 be? X_2 must be 9. Why?

"therefore"

$\bar{X}_3 = \frac{1}{3}(X_1 + X_2 + X_3)$ by definition of sample mean

$\Rightarrow 5 = \frac{1}{3}(3 + X_2 + 3)$ plugging in what we know

$\Rightarrow 15 = 3 + X_2 + 3$ multiplying both sides by 3

$\Rightarrow 15 = 6 + X_2$

$\Rightarrow 9 = X_2$ subtracting 6 from both sides

PollEverywhere - Sample Mean II

Suppose you know that each of $X_1, \ldots, X_9$ are negative numbers, and yet the sample mean $\bar{X}_{10}$ is positive. What's the most you can say about X_{10}?

- X_{10} is positive 44%
- X_{10} is negative 23%
- X_{10} is equal to $-(X_1 + \cdots + X_9)$ 25%
- Nothing 6%

$$\frac{1}{10}(X_1 + \cdots + X_{10}) > 0$$

$$X_1 + \cdots + X_{10} > 0$$

$$[X_1 + \cdots + X_9] + X_{10} > 0$$

$$X_{10} > \underbrace{-\underbrace{[X_1 + \cdots + X_9]}_{<0}}_{>0} > 0.$$

PollEverywhere - Sample Mean III

Suppose all observations X_i are equal to 5, except for $X_{43} = 543$ which is an **outlier**. What happens to the sample mean $\bar{X}_n$ as n gets larger?

- It gets closer to 543 30%
- It gets closer to 5 57%
- Impossible to say 13%

Rough argument

Take a large n, say n = 1000. Then

$$\bar{X}_{1000} = \frac{1}{1000}(X_1 + \dots + X_{42} + X_{43} + X_{44} + \dots + X_{1000})$$

$$= \frac{1}{1000}(5 + \dots + 5 + 543 + 5 + \dots + 5)$$

$$= \frac{1}{1000}(999 \cdot 5 + 543)$$

$$\simeq \frac{1}{1000}(1000 \cdot 5 + 543)$$

$$= \frac{\cancel{1000} \cdot 5}{\cancel{1000}} + \frac{543}{1000} = 5 + \frac{543}{1000} = 5.543 \simeq 5.$$

Formal argument (basically the same thing, except without choosing a specific value of n)

For any large n,

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_{42} + X_{43} + X_{44} + \dots + X_n)$$

$$= \frac{1}{n}(5 + \dots + 5 + 543 + 5 + \dots + 5)$$

$$= \frac{1}{n}(5(n-1) + 543)$$

$$= \frac{1}{n}(5n - 5 + 543)$$

$$= \frac{1}{n}(5n + 538)$$

$$= 5 + \frac{538}{n}$$

$$\longrightarrow 5 \text{ as } n \text{ gets arbitrarily large}$$

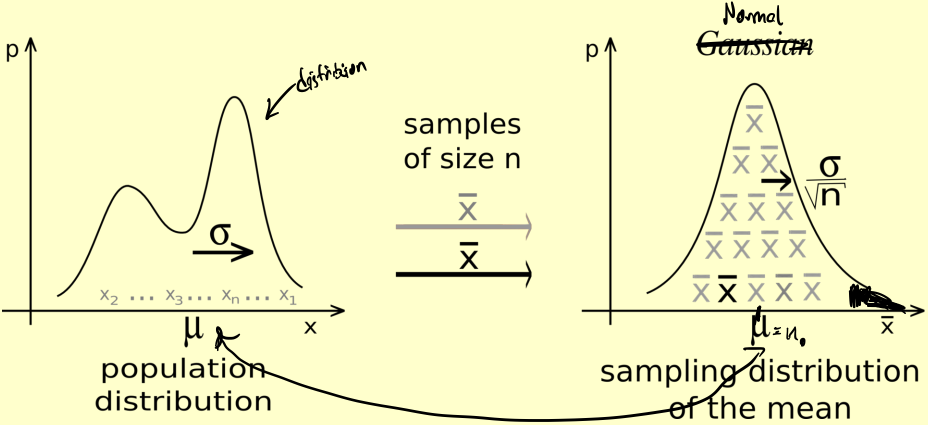
Statistical Assumptions - Distributions IV

- Two extremely important examples of these tools are the Law of Large Numbers and the Central Limit Theorem
- The **Law of Large Numbers** very roughly states that if the data are independent and arise from the same distribution (no matter which one) whose mean is μ, then the value of $\bar{X}_n$ will eventually approach the "true mean" μ as n grows higher and higher
- The **Central Limit Theorem** very roughly states that if the data are independent and arise from the same distribution (no matter which one) whose mean is μ and whose standard deviation is σ, then the distribution of the sample average *itself* eventually looks like a Normal($\mu, \frac{\sigma}{\sqrt{n}}$) distribution n grows higher and higher

Mean of underlying data

approaches 0 as n gets larger, so the average difference between $\bar{X}_n$ and μ gets smaller

Statistical Assumptions - Distributions V



So how do these help?

NHST - Stilt Example I

- Suppose we want to compare the heights of the adult population of planet Stilt with those of adult Earthlings; we know that the average adult height on Earth is 165 cm
  - Stiltians appear to be quite tall...
- We set our significance level at $\alpha = 0.05$ and design our experiment, carefully stating our statistical hypotheses:
  - H_0: There is no statistical difference between the average height of adults on the two planets
  - H_A: The average height of Stiltians is greater than the average height of Earthlings
- We collect 400 independent Stiltian height measurements and calculate the sample mean $\bar{X}_{400}$ =180 cm

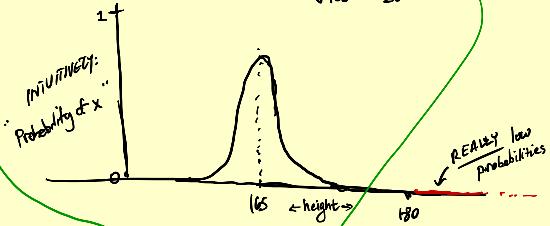
NHST - Stilt Example II

- To avoid technicalities: assume we know for a fact that $\sigma = 20$
- Under H_0, the distribution of $\bar{X}_n$ looks like a Normal distribution with mean 165 and standard deviation $\frac{20}{\sqrt{n}}$ for large n (in our case, $n = 400$ is definitely large enough):

$\approx \frac{20}{\sqrt{400}} = \frac{20}{20} = 1$

$H_0: \mu = 165$

$H_A: \mu > 165$



- Then under H_0, $\bar{X}_{400} = 180$ cm is a random draw from the above distribution – that is, a Normal$(165, 1)$ distribution

NHST - Stilt Example III

- Now we ask ourselves: what's the probability of drawing a number at least as high as 180 from the Normal$(165, 1)$ distribution?
- Because the Normal distribution is so well understood, we can calculate this exactly: it's approximately 3.670966×10^{-51}

$$\approx 0.\overbrace{0000 \cdots\cdots 000}^{50 \text{ zeros!}}3671 < 0.05$$

- Thus, we **reject** H_0 at the 5% significance level and conclude that Stiltians are, in fact, higher than Earthlings on average

NHST - Stilt Example IV

NHST - p-values I

- The probability we calculated on the previous slide is called a ***p*-value**
- Formally, the ***p*-value** is the probability that the underlying statistical model would generate data at least as extreme as our observations, *given that the null hypothesis is true*
- When our calculated p-value is below our pre-set ~~confidence~~ significance level, we conclude that the data is too unlikely to have occurred merely by chance under H_0
- Hence, we reject the null hypothesis:
- It is "very unlikely" that the sample of Stiltians whose height we measured just happened by chance to be atypically tall

NHST - *p*-values II

- The study's conclusion might read "The Stiltish population is significantly taller than the Earthian population"
- **IMPORTANT:** We didn't **prove** that Stiltians actually *are* taller than Earthlings on average – to do so would require measuring every Stiltian, computing the average, and comparing it to Earth's average
- If we could do that, we wouldn't need to use statistics!
- Instead, we concluded that it's (extremely) unlikely that Stiltians are not taller than Earthlings

NHST - p-values III

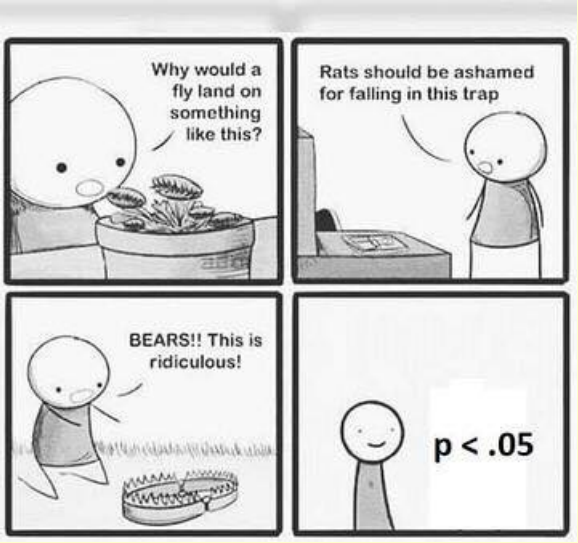
- What significance level is low enough?
- Ronald Fisher, a pioneer of statistics and experimental design, suggested 5%
- Much of scientific literature/academia still uses this
- But it's completely arbitrary!
- "...statistical significance, an odd religion that researchers worship almost blindly"

Problems With p-values - Misinterpretation I

$0.05 = 5\% = 1/20$

- p-values are *probabilities*, not *certainties*
- Suppose many published studies calculate p-values of 0.05
- Then on average, 1 out of 20 studies will incorrectly reject H_0
- How many studies are published every day? Hundreds!

Problems With p-values - Misinterpretation II



Problems With p-values - Misinterpretation III

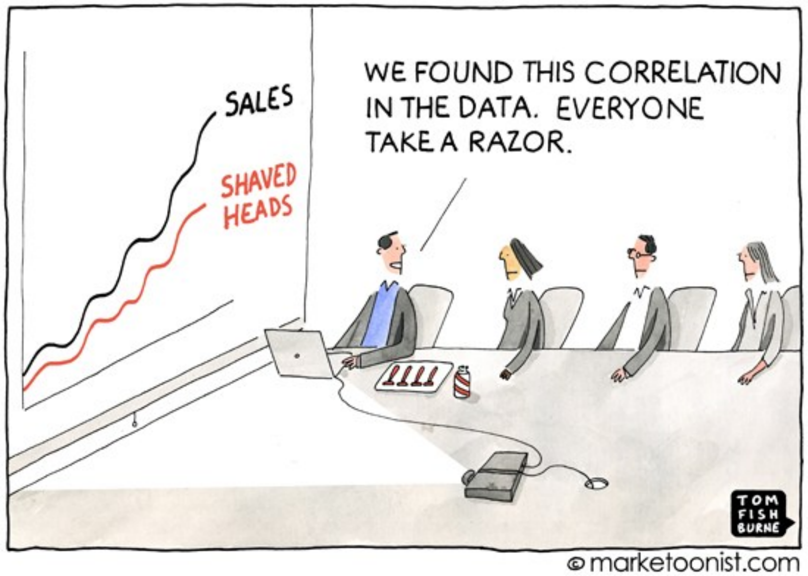
- p-values lead to publication bias
- The $p < 0.05$ is so entrenched that a study result with $p = 0.06$ is considered a "negative" study
- Journals with limited space want to publish new, interesting, "positive" findings – a study with $p > 0.05$ may contain important new findings, but is far less likely to be published

If we calculate $p = 0.1$ ($= 10\%$), then there's a 10% chance that our observations (or more extreme observations) arose due to chance.
Not that high!

- Correlation is not causation!

"Study finds link between annual number of pool drowning deaths and annual number of Nicholas Cage movies ($p < 0.01$)"

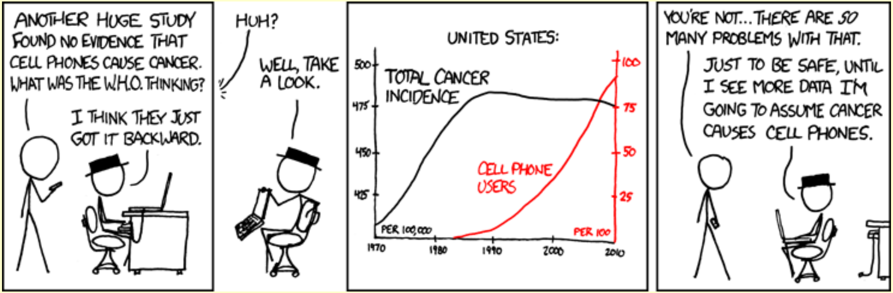
Problems With p-values - Misinterpretation IV



Problems With p-values - Misinterpretation V

- Extrapolating a population-based result to an individual can also be problematic
- For example: your well-conducted study compared the effectiveness of Drug A and Drug B to dangerously high serum rhubarb levels
- You find that Drug A was effective in 90% of patients who received it, while Drug B was effective in 30% of patients who received it
- You conclude that "Drug A works, but Drug B doesn't work"
- This is a disservice to the 30% of patients who responded to Drug B but might not respond to Drug A

Problems With p-values - Misinterpretation VI



Problems With p-values - p-Hacking I

- Changing the prespecified threshold to declare results statistically significant
- Multiple comparisons
    - Multiple comparisons
- Increasing the size of the study population
    - p-value calculations are affected by the sample size
    - A very large medical study might produce a result that's statistically significant, but not *medically* significant
    - The time to achieve a normal temperature was 19.5 hours for Drug A and 19.8 hours for Drug B, a statistically significant difference
    - Drug A advertisement: "Expensive new Drug A reduced fever significantly faster than cheap old Drug B"

Problems With p-values - p-Hacking II

- Post-hoc analysis (testing analysis that were not prespecified)
  - "If you torture the data long enough, it well tell you what you want to hear"
  - Okay as a springboard to discover hypotheses for a new study
- Outright fraud
  - "Editing out" data points that sway the results away from the hoped-for conclusion
  - More difficult these days (still not impossible) now that many reputable journal require study authors to upload their raw data to a repository for scrutiny by other researchers

PollEverywhere - p-values I

Suppose we have (properly) conducted an experiment which yielded a p-value of 0.002. Then H_0 must be false.

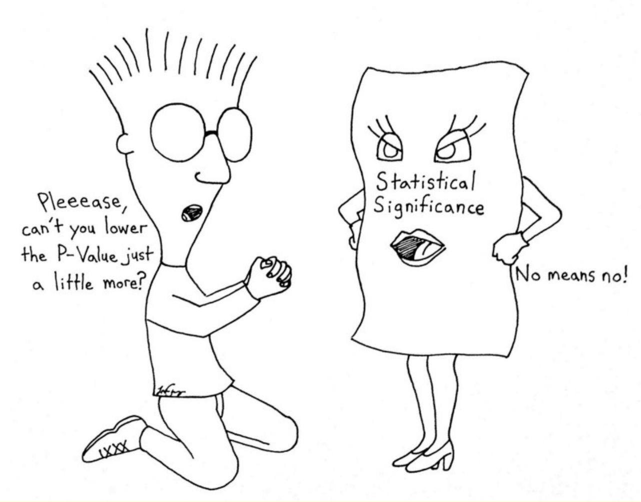
- True 41%
- False 59%

PollEverywhere - p-values II

A hypothesis test is done in which H_A is that more than 10% of the population is left-handed. The p-value for the test is calculated to be 0.25. Which statement is correct?

- 45% We can conclude that more than 10% of the population is left-handed
- 5% We can conclude that more than 25% of the population is left-handed
- 18% We can conclude that exactly 25% of the population is left-handed
- 32% We cannot conclude that more than 10% of the population is left-handed

Statistical Significance



Problems With p-values - What Can We Do

- It's easy to state why the framework is flawed – but much harder to discover how to fix it!
- Suggestions:
  - Make the "standard" significance level (much) less than 0.05 (but many valuable results might be discarded/ignored if the p-value requirement is too stringent)
  - Report confidence intervals
  - Make studies more transparent ~~transparent~~; i.e., *all* data must be submitted to the journal repository (but this won't stop data tampering before the data is submitted)
  - Combat publication bias – all results, whether "negative" or "positive", must be uploaded to a central repository like Health Canada, FDA, APA, etc. (but this won't stop early-stage data tampering and might chill research efforts)

The Replication Crisis

- An *ongoing* crisis, particularly in the field of social psychology
- Nobody questioned the results of many famous scientific studies until just recently, when scholars tried to replicate them without success
- According to a 2016 poll of 1,500 scientists reported in the journal Nature, 70% of them had failed to reproduce at least one other scientist's experiment. 50% had failed to reproduce one of *their own* experiments. (Wikipedia)
- Several well-respected researchers admitted to fabricating their results, and resigned in disgrace
- Optional (but interesting/disturbing) further reading

Polls I

- Poll results are often stated in the form "correct to $\pm3\%$, 19 times out of 20"

$\alpha = 0.05$

- "19 times out of 20" is ~~really a~~ p-value of 0.05
- Even a perfectly conducted poll (free of cognitive biases!) will draw an atypical sample occasionally
- "19 times out of 20" implies that the pollster expects the sample to be atypical 1/20 times
- Framing effect!

Polls II

IT'S OVER.
AFTER TWENTY MONTHS IT'S FINALLY OVER.
I DON'T HAVE TO BE AN ELECTION JUNKIE ANYMORE.
I DON'T HAVE TO CARE ABOUT OPINION POLLS, EXIT POLLS, MARGINS OF ERROR, ATTACK ADS, GAME-CHANGERS, TRACKING POLLS, SWING STATES, SWING VOTES, THE BRADLEY EFFECT, OR <NAME> THE <OCCUPATION>.
I'M FREE.
Google
2012 polling statistics

- Applying wrong assumptions to a statistical model can be disastrous: Sally Clark

$\mathbb{P}(\text{crib death}) \simeq 1/8500$ (for a child from an affluent, nonsmoking family with the mother over 26 years old)

$\mathbb{P}(\text{2 crib deaths}) \stackrel{?}{=} \underbrace{\frac{1}{8500} \times \frac{1}{8500}}_{\text{Assuming \underline{independence}}} = 1 \text{ in } 72{,}250{,}000$

Are the deaths of two siblings from the same health issue independent events?

Of course not!

(and if that wasn't bad enough: the original report showed that Clark's first child died from a respiratory infection in the first place, not crib death)