# Probability and Introduction to Statistics

# ALY6010

Tom Breur

Week 4, 17-NOV-2020

# Agenda

- Administrative notes
- Review Discussion board “Simulations”
- Hypothesis testing
- More thoughts on Statistical power
- Preparation week 5

# Administrative notes

- R project submissions require ***two*** components:
  - The worksheet that describes our process
  - An R file
  - *Both* parts of the submission need to adhere to file naming conventions!
- Students who feel disadvantaged due to Quiz typos/errors may file *specific examples* for lenience

# Your TA for ALY6010 CRN 71709

Catherine Richard

**Email:** richard.ca@northeastern.edu

You can reach me via :

- Email
- Post on Canvas
- WhatsApp (781-526-6300)

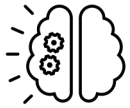
This week TA hours:

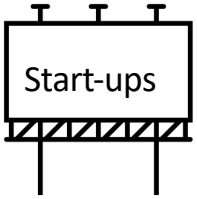
- Thursday 9-10 AM
- Friday 3-5 PM
- Saturday 10:15-11:30 AM

I've worked at and am interested in :

Healthcare

Deep Learning






Tufts University

Northeastern University
College of Professional Studies

MPS Analytics, Statistical Modeling Concentration
I'll graduate at the end of next quarte

# Discussion Board

## Critical review

# Simulations

- There is a principal difference between running simulations and performing probability calculations
- A coin toss with 50/50 odds, no matter how many times you repeat it, will still come up “Heads” about 50% of the time – no need to perform any simulation for that!

So when ***do*** you need to leverage simulations??

- Many real world processes can not (easily) be modeled using probability
- Example last week was ”Snakes & Ladders” – this process can (probably) be modeled analytically, but would require very, very challenging differential equations
  - Other good examples were winning World series (4/7) and complex processes (complex ***is not*** complicated)

# Hypothesis testing

# Three methods to test hypotheses

- The traditional method
- The p-value method
- The confidence interval method

NB.: in all three scenarios, the hypotheses (one or two-sided) must be formulated prior to designing the experiment and running your experiment

- Traditional: state $H_0$ and $H_1$, and announce critical value
- P-value: state $H_0$, then based on results calculate p-value
- Confidence interval: state $H_0$ and $H_1$, and calculate the confidence interval; determine in/out after experiment

# Sample data & error variance

Bluman p. 417

" … sample data are used to determine if a Null hypothesis should be rejected. Because this decision is based on sample data, there is a possibility that an incorrect decision can be made."

- Statistical decision making is based on stochastic distributions (the Law of Large Numbers)
- Test statistics are used to *quantify* the uncertainty, or the odds of making an incorrect decision

# Confusion matrix

| | | Reality | |
|---|---|---|---|
| | | Positive | Negative |
| Study Finding | Positive | **True Positive** (Power) (1-β) | False Positive **Type I Error** (α) |
| | Negative | False Negative **Type II Error** (β) | **True Negative** |

***Optional*** reading:
tombreur.files.wordpress.com/2016/06/statistical-power-analysis-and-the-contemporary-crisis-in-social-sciences_201611.pdf

# Overview tests

- Z test for the difference of two means
- T test for the difference of two means
  - Independent samples
  - Dependent samples (i.e. repeated measures)
- Z test for the difference of two population proportions
- F test for the ratio of two population variances

# Comparing *two* means

Bluman p. 414

"There are two specific statistical tests used for hypotheses concerning means: the *z test* and the *t test*"

- This statement is correct under the explicit provision that this pertains to **comparisons of *two* groups**
  - In Chapter 12 "Analysis of Variance" (p. 645-684) Bluman explains how to test for differences when there are *more than two groups* to compare
  - Multiple groups *can* be tested with (many) pairwise comparisons; however, this results in capitalizing in Type I error (i.e. increased chance of incorrectly rejecting $H_0$)
    - Advanced topic: this effect needs to be mitigated, e.g. using the Bonferroni correction porcedure

# Z or t-test?

Bluman p. 447 (important!!)

- When $\sigma$ is known use the Z-test
  - For $N \geq 30$ can be any distribution
  - For $N < 30$ the variable must be normally distributed
- When $\sigma$ is unknown use the t-test
  - For $N \geq 30$ can be any distribution
  - For $N < 30$ the variable must be normally distributed

# Z test for means

- Derived mean is considered (in Bluman!) to be a “population value” when N>30 ($\sigma$ is known)
  - Note that this assumes (!) random sampling
  - For N<30 the underlying distribution must be (approximately) normal
  - The ”general” Statistics literature hovers around N=20-50; in the context of Bluman/ALY6010 we stick with 30
- Used for both one- and two-sided tests
  - Determine confidence interval, and rejection region(s) accordingly
- The $\alpha = 0.05$ (5%) value is commonly used in social sciences *as a conventional cutoff for significance*
  - However, many industrial/telco applications require (far) greater accuracy, and hence (much) smaller $\alpha$

# T test for two means (1)

- Two possible applications: dependent, or independent samples
  - Essentially the question: do you compare two *groups* (independent samples), or do you compare two measurements taken from the same individual (dependent sampes)
- Can be used for both one- and two-sided tests
  - Calculate the corresponding degrees of freedom, incorporating *both* sample sizes
  - Determine confidence interval, and rejection region(s) accordingly

# T test for two means (2)

- Dependent vs independent design?
- All else being equal, dependent measures t-tests have greater statistical power
  - For the same group sizes, dependent measures are more likely to surface significant findings
  - This “advantage” is because there is less “error variance”: the spread *between* people has been canceled out
- For various reasons, repeated measures designs may not be feasible

# Z test for two proportions

- Proportions are modeled via the Binomial distribution, that is assumed to approach (mimic) the Z-distribution
- For large N, and values close(r) to 0.5 this is a very reasonable assumption
  - Under this assumption, the confidence interval will be symmetric (equally long to the left and right of the mean)
  - Advanced topic: interested readers can find the more precise adjustment in the literature
- After making the adjustments and assumptions, the statistics are a reasonably good approximation (justifying the simplification)
  - Check for heuristics ("assumptions"): $np \geq 5$ *and* $nq \geq 5$

# F test for two variances

- Esoteric application, rarely (if ever) comes up in business settings
  - Single exceptions may be financial trading professionals
- Same F-test (named after Fisher) that is used for ANOVA and Regression
  - Requires calculation of degrees of freedom, based on *both* samples
- Can be used as a one-sided test (a>b?), or two-sided (a=b?)

- Calculation of the rejection area(s), based on one- or two-sided hypethesis

# More thoughts on Statistical power

# Statistical Power and "Big Data"

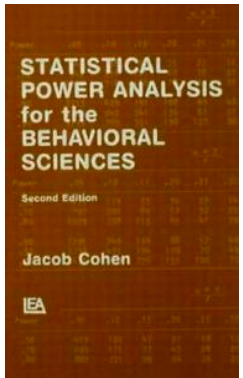
- With the ubiquitous availability of data, there is legitimate concern that data scientists can go "fishing for significance"
- The more statistical tests you perform, the more likely that *some* of them will prove significant, **even when there is no legitimate effect**!
- When the $H_0$ gets incorrectly rejected, you have a "false positive"
- One of the contemporary fields of research for data scientists deals with "FPR's" – False Positive Rates
  - When the $H_0$ *does not get rejected*: research Statistical Power
  - When $H_1$ *gets accepted*: be weary of FPRs
- These are flipsides of the same coin

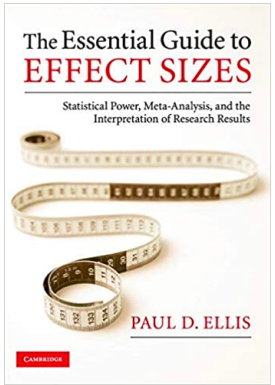
# Finding β

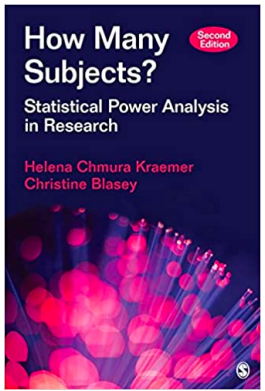
Bluman p. 419

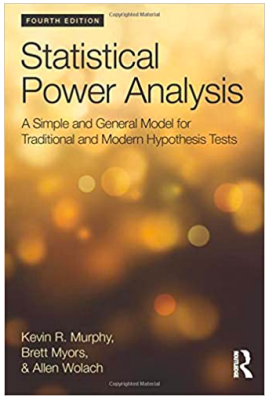
"In most hypothesis testing situations, β cannot be easily computed; However, α and β are related in that decreasing one increases the other

- The first part of this statement is questionable (since 1977)
- The second part of this statement *is* correct
- Since Cohen (1977), (most) tables for Power Analysis have been available:



# Preparation week #5

# Requirements

- Discussion board:
  - Post contributions on *successive* (distinct!) days
  - *Minimum* of three posts, but this need not limit you
  - *First* post your primary contribution, *only then* will get access to other peoples' contributions
- Quizzes:
  - Bluman 9-1 to 9-5 quizzes
  - R project M 3
- Reading preparation week 5 (Chapter 9 Bluman)
- Reading preparation week 5 (Chapter 6-7 Kabacoff)

# Discussion board: requirements

- You ***first*** (!) need to post an original contribution ("primary post") first, with a minimum (!) of 250 words
- This post needs to contain an academic reference to a reliable (!) and relevant source
  - The reference needs to be set in APA standard
- A minimum (!) of two responses are required, each 80+ words, and posted on distinct, successive dates
- All contributions need to be substantive
  - For clarity: "I agree", "I like your post", etc. do not count as substantive replies. *Instead* reason why you agree or disagree, and refer to outside soures to justify your position
- Referring to other sources or posts, or previous classes, earns "brownie points" towards top grades (100 points) for integrative learning

# Discussion board: substance (1)

- #1
  When constructing and implementing hypothesis tests, what reasoning is used behind the statement of the Null and Alternative hypotheses? Why are hypothesis tests set up in this way? Can a confidence interval obtained for estimating a population parameter be used to reject the null hypothesis? If your answer is yes, explain how. If your answer is no, explain why.
  - There are principled reasons why the pairs of Hypotheses, $H_0$ and $H_1$, are framed, and not in reverse, for instance. Which statement is used for $H_0$ and which one for $H_1$ is chosen deliberately. Explain what that reason is. Hint: the answer to this question does not follow (easily) from Bluman's text, hence needs to be researched elsewhere

# Discussion board: substance (2)

- #1
  Can a confidence interval obtained for estimating a population parameter be used to reject the null hypothesis? If your answer is yes, explain how. If your answer is no, explain why.
  - The relationship between confidence interval and hypothesis testing *is* available in Bluman, but needs to be ferreted out across multiple chapters
  - Sometimes the correct answer is “it depends”, and ***if*** this is your conclusion, then clarify the critical conditions conditions that need to be regarded

# Discussion board: substance (3)

- #2
  When performing a hypothesis testing, two types of errors can be made: Type I and Type II. Explain in your opinion which of these errors would be a more serious error. Use specific examples to support your argument and reasoning
  - For this item, you want to think about the relative "cost" (consequences) of making ***either*** error: either Type I or Type II
  - Give examples of ***both*** types of errors where the cost differential plays out different across Type I and Type II errors
  - How can Statistics be leveraged to to clarify the trade-offs between Type I and Type II errors?
  - What can you do to reduce the incidence of Type I and Type II errors?