Introduction to quantitative analysis

GSTTP research mini-course

Corrina Moucheraud, UCLA FSPH
26 April 2021
Learning objectives

• Assess how study design and quantitative data (format, approach) shape quantitative analysis: research questions and hypotheses, choice of analysis methods, formulation and interpretation of results.

• Describe best practices of quantitative data visualization (charts and tables) and critique examples. ➔ We are going to do this later instead
Typology of study designs

- What type of study are you undertaking?
Quick side note: Studies in epidemiology

Epidemiology is about identifying associations between exposures and outcomes. To identify any association, exposures and outcomes must first be measured in a quantitative manner. Then rates of occurrence of events are computed. These measures are called “measures of disease frequency.” Once measured, the association between exposures and outcomes are then evaluated by calculating “measures of association or effect.” Finally, the impact of removal of an exposure on the outcome is evaluated by computing “measures of potential impact.” In general, measures of disease frequency are needed to get measures of association, and both of these are needed to get measures of impact. There is some overlap between these measures, and terminology is poorly standardized.

The superscript numbers refer to the formulae used to compute these measures (formulae shown separately in the following pages).

Modular Pi, McGill University (modular.pi@mcgill.ca), Kristian Fillon, McGill University (Kristian.Fillon@mail.mcgill.ca)
Critically important that you have a clear research question up-front

• Let’s say that I surveyed each of you after every one of these lectures
• What research questions could I answer?

• From your research question, everything else flows.
  • Is it a “how much” question? Or a “who” question? Or a “why” question? Or...
  • This dictates your study design and how you will analyze & present your results

• Also, need to be clear about what outcome you’re measuring
• Ideal scenario! Very clear & easy to interpret. But, very rare.
• This is what randomized control trials emulate
• Asbestos exposure $\rightarrow$ Risk of mesothelioma

• Mediators are alternative pathways between IV & DV
  • Direct effect + indirect effect (operates through mediator) = total effect of IV on DV

• Moderators impact the relationship between IV & DV
  • Aka, effect modifiers: the relationship is not the same for everybody
Quick note on confounding

• “Confounding” is a term you will probably hear all the time
• Affects ability to draw causal inferences from observed relationships
• Because it affects both the IV & DV, creates a spurious association between these if it is left unmeasured

- Warm weather
  - Ice cream sales → Shark attacks
Your job as a researcher

• Be specific, thoughtful, and clear up-front about your research question(s), your study design, what you are measuring as what type of variable (independent, dependent, mediator, moderator, etc.)

• Theoretical frameworks and existing conceptual models can help you specify these

• Also, discussion with your collaborators and experts in the field
Research questions

• Should end in a question mark!

• “Understand the patient population at Hospital ABC”
  • Better: “Who is coming to Hospital ABC”
    • Better x2: “What is the age distribution of people with scheduled outpatient appts at Hospital ABC during January 2020”
      • Note that this is a different question than “What is the average age of people …” or “What is the age distribution of people who attended outpatient appts at Hospital ABC during January 2020”
Improved research questions, ex. 2

• “How diverse is the patient population at Hospital ABC”
  • Better: “How many Latino patients come to Hospital ABC”
    • Better x2: “What percentage of adults attending outpatient appts at Hospital ABC during Jan 2020 identify as Latino?”
      • Different than: “What is the number of adults attending outpatient appts at Hospital ABC during Jan 2020 who identify as Latino”
  • What information does this tell us?
What doesn’t this tell us?

• How many Spanish-speaking staff do we need?
  ➢ Need to know patient language preferences

• How many additional Spanish-speaking staff do we need? (... what types, which service lines)
  ➢ Need to know current staff’s language proficiency

• What care-seeking barriers do Latino patients face?
  ➢ Need additional info about care-seeking patterns + also qualitative data may be informative
Ok, so you set things up well & collected your data. Now what?

• Remember data types:
  • Dichotomous
  • Nominal with >2 categories
  • Ordinal
  • Continuous

• Depending on type of data, can undertake different types of analyses
Lots of nice “flow charts” out there to help guide you.
Types of analyses

- **Descriptive**
  - Solely about the sample we have observed

- **Comparative**

- **Inferential**
  - Using data from your sample to infer about population
    
    *Any time you see a p-value or confidence interval, it’s from an inferential analysis*
Description vs. inference

- **Descriptive statistics**: using data to show or summarize (describe) characteristics of a sample

- **Inferential statistics**: using data from a sample to make predictions about (infer) characteristics of a population

- **Predictive methods**: based on what we’ve already seen, what do we expect to see in the future? How might this change if we alter X, Y, Z?
What can descriptive statistics do?

• Raw data are too numerous and can be hard to detect meaningful patterns

• Descriptive statistics enable interpretation – but cannot draw conclusions beyond the sample studied
What can inferential statistics do?

• Might be impossible to survey a full population ("census")

• Robustness of inference depends on how well your sample represents the population
  • Inherent uncertainty (sampling error) + bias
Inferential statistics

- You are a sample of:
  - Medical students at UCLA
  - Graduate students at UCLA
  - Students at UCLA
  - People who live in LA
  - Graduate students in the US
  - People who know about statistics
  - Humans
  - etc.

- To which of these group(s) would you be most comfortable inferring? i.e., for which population(s) are we a good sample?

→ “confidence” is a function of sampling
Hypothesis testing

• Null hypothesis (aka, $H_0$): there is no difference between population A & population B
  • By comparing sample$_A$ to sample$_B$
  • p-value: probability that the null hypothesis is true

• Note: Continuous data exist in a distribution; the larger your sample size, the more stable/apparent the distribution of your data will be
  • We often assume normal distribution **BUT this can be an incorrect** assumption!! Be very careful about this.
Descriptive data with categorical variables

- Frequencies of occurrence: #s or %s
- Consider treating individually vs. combining... Depends on research q!

What are some ways we could analyze these data?

- What is most frequently positive source? Negative source?
- Overall do people hear positive or negative information?
- Are different sources, or types of sources more likely to be pos or neg?
- How many different sources of info do people hear?

Have you ever heard information and/or opinions about any COVID vaccine from any of the following sources? If so, would you characterize these as overall positive (such as: the COVID vaccine is important, safe and effective), overall negative (such as: the COVID vaccine is not important, safe or effective), or neutral (neither positive nor negative, i.e. factual information only without opinion or judgment)?

<table>
<thead>
<tr>
<th>Source</th>
<th>No</th>
<th>Yes, overall positive</th>
<th>Yes, neutral/factual</th>
<th>Yes, overall negative</th>
<th>Yes, but don’t know “tone”</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Ministry of Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Facility in-charge/Immediate supervisor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Colleagues or other health care workers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Clients/patients attending your facility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Traditional practitioner</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Newspapers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Radio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Television</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Internet (website)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. A post on social media from a person I know (friend, family, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A note on analyzing continuous data

• If discrete and small range (e.g., number of pets that people have) can treat as categorical variable
  • If you took an average, would “2.25 cats” be meaningful?
  • So instead, think about each value as a response option of a categorical question: # and/or % of each
  • Sometimes we treat as continuous anyway (examples: total fertility rate [1.73 births per woman in the U.S.], average household size [2.53 people in the U.S.])
Descriptive statistics with continuous data

• If want to treat as continuous, decide: do I care about the middle or do I care about the spread?

• If the middle: median, mean, mode...
  **** Mean is sensitive to the distribution ****

• If the spread: range, variance/standard deviation, interquartile range...

• Often we care about both and may report both a middle-based statistic and a spread-based statistic, like the median and the SD
Inferential methods: a partial list

• t-tests: difference between 2 means (continuous variables)
• chi-square tests: difference between 2+ categorical variables
• ANOVA: difference between 3+ means
• Regression models (linear, logistic, etc.): predicted relationship between variables
A very simple (& admittedly not very good) example for us to walk through

• Research question: Are people who attend medical school taller than people who do not attend medical school?

• $H_0$: Medical students are the same height as people not in medical school  
  (based on the sample of first-year DGSOM students who I’ve measured)  
  *Usually we don’t say this part “out loud”*

• $H_A$ (aka $H_1$): Medical students are taller than people not in medical school  
  (based on my sample)  
  $\rightarrow$ NOTE: this is a one-sided hypothesis; I could have said “Medical students are a different height” and this would be two-sided (could be taller, could be shorter)
Descriptive statistics

• Could start with the mean & the median
  • Mean will be very sensitive (likely to get “pulled”) if you have some really tall, or some really short, classmates

• But first-year DGSOM students are just a sample... So we need to make some inference if we want to address our research question

• The average height of first-year DGSOM students is 71 inches
• If I estimate that this has a 95% confidence interval of 68-74 inches, I am saying “I am 95% sure that the true average height of the population [remember, think about what population you are a sample of!] lies between 68 and 74 inches”
Inference & beware confounding!

• To really answer my research question, I would need to compare the height of DGSOM students to the height of similar people who are not in medical school (want to avoid mediators, confounders, etc.)
  • Maybe really tall people were more likely to be recruited for varsity sports in high school & college, so less likely to take all the premed classes etc., so less likely to go to medical school → this is OK, it is a “common cause”
    • If we hypothesize this operates differently for boys & girls, however, we would need to include this as a moderator
  • Maybe the California sunshine was really helpful for kids’ growth, and also DGSOM students are more likely to be from California → so our relationship between height & DGSOM status is confounded by growing up in California
• Let’s assume you found a good comparison population
Inference & p-values & what it all means

• Could do a t-test of independent sample to compare the average height in sample\textsubscript{A} (DGSOM students) to average height in sample\textsubscript{B} (comparison group)
  • Will generate what is called a “test statistic” (a t-value in this case)
    • Compares our sample average value (71”) to the average in the other sample
      • Null hypothesis says: the average height in sample\textsubscript{B} is also 71”
      • Big test statistic means big difference between the groups
  • AND, a p-value on this test statistic: what is the probability that this test statistic could be obtained if \( H_0 \) is true
    • We reject the null based on this probability
      • We never accept the null! We just reject, or fail to reject, it

• If our t-statistic has a p-value of 0.03 this can be read as:
  There is a 3% (or less) chance that the observed difference in average height between sample\textsubscript{A} (DGSOM students) & sample\textsubscript{B} (comparison group) and the true population difference would have occurred if medical students are the same height as other people (if \( H_0 \) were true).
Regression models

• Two continuous variables – car weight, and mileage (mpg)

• We can see there is a negative association but, how much? For each additional pound of weight, how much worse does mileage get?

• This is a best-fit line: it minimizes the vertical distance from the line to each point

• You may recall that straight lines are expressed as: $Y = mx+b$

• Same for (linear) regression! At 0 lbs (intercept), mileage (Y) would be 39.4 mpg; each additional unit of X (pound of weight) subtracts 0.006 mpg
  • There is also an error term because the line is not fit perfectly: $Y=\alpha+\beta x+\varepsilon$
  • If there are multiple explanatory variables, equation expands: $Y=\alpha+\beta_1 x_1 + \beta_2 x_2 +...+\varepsilon$
Resources

• Many (many many!) books, including:
  • Rosner, “Fundamentals of biostatistics”
  • Van Emden, “Statistics for terrified biologists”
  • Wheelan, “Naked statistics”
  • Coolidge, “Statistics: A gentle introduction”

• Stata has informative help files for programming all of this!
• Institute for Digital Research and Education at UCLA is amazing (https://idre.ucla.edu/)
  • Resources on their website (FAQ etc) + also consultations for UCLA community