

Topics List

Material covered

- Lectures 1-7, HW 1-3, Labs 1-3

Coding

- Console vs. Quarto document. When might you code in Quarto document and not in Console? How can using Quarto help with reproducibility?
- `View()` and `glimpse()` and `names()`
- `getwd()`. What does it tell you?
- YAML: what is it, how is it helpful for documentation?
- How to create headers in Quarto
- How to comment out code within a code chunk
- `library()` vs. `install.packages()`
- What is `ggplot()`?
- What is the tidyverse ?
- `geom_point()`, `geom_boxplot()`, `geom_bar()`, `geom_line()`, `labs()`, `size`, `color`, `alpha`, `fill`. What do all these do?
- Which type of ggplot would be appropriate based on different variable types (e.g. scatterplot, boxplot, line plot)
- Detecting common sources of errors: missing parentheses, missing quotation mark, missing comma
- `#| warning: false` and `#| message: false`. What do they do?

- `filter()`, `slice()`, `distinct()`, `select()`, `arrange()`, `summarize()`, `group_by()`, `mutate()`, `sample_n()`, `sample_frac()`, `case_when()`. What are each of these used for?
- pipe, `|>`, What is it used for?
- Logical operators in R
- `%in%`
- `na.rm = T`
- Given a particular `ggplot` code, describe the type of plot that would be created. (Review examples in Lab 3).
- `dbinom()` and `dpois()`, how to use to find probabilities.

Conceptual & Applied

- Population vs. sample vs. one observation within a research question
- What does “representative” sample mean? What does it mean for a study to be “generalizable”?
- Probability sampling vs. non-probability sampling & knowing a few examples of each. Which is more generalizable?
- Experimental vs. observational studies.
- What is bias? Examples: selection, reporting, non-response, attrition, falsified data.
- Why is transparency and data/code sharing important ethically in statistical studies?
- Define Reproducibility vs. replicability
- Exploratory data analysis. What is it and why is it helpful?
- Categorical: Nominal vs. Ordinal data & know a few examples of each
- Numerical: Count/rank vs. continuous data & know a few examples of each
- Identifying data types based on a Table 1 (i.e. demographics or baseline characteristics table) from a paper
- Parameter vs. statistic
- Mean, median, mode, sample minimum and maximum, range, IQR, five-number summary, sample standard deviation
- Symmetric vs. skewed distributions, right vs. left-skew
- Chebyshev’s inequality, why is it useful?

- Frequentist vs. Bayesian schools of thought. When might one be helpful over the other?
- Probability space & its components.
- Intersection, union, complement, difference, disjoint, complement rule, inclusion-exclusion. Drawing out venn-diagrams for each.
- Determining probabilities and conditional probabilities of events given a table (review Coffee Drinking example in Lecture 3)
- Verifying Komolgorov axioms in a simple example
- Conditional probability (formula provided)
- Definition of independence (in terms of probabilities): $P(A \cap B) = P(A) \times P(B)$, or equivalently, $P(A|B) = P(A)$ or $P(B|A) = P(B)$
- Given two simple events A and B , calculating $P(A)$, $P(B)$, $P(A \cap B)$, $P(A \cup B)$, $P(A|B)$, $P(B|A)$ (Review example in Lecture 4.)
- Law of total probability, using this to find probabilities of events given a scenario (Review Coffee example and sleep example in Lecture 4)
- Order matters in conditional probability (Coffee example in Lecture 4)
- Know how to define events and set up Bayes' rule to solve for conditional probability given a scenario (Review Coffee Example Lecture 4)
- Prevalence, sensitivity, specificity, positive predictive value, negative predictive value. **Know the definitions** in words and in symbols.
- Tradeoffs of sensitivity and specificity in a medical diagnostic test
- How to read an ROC curve, what's a "good" ROC curve, what is AUC
- Given a condition/disease and a diagnostic test, define each event and **write out how you would solve** for the Sensitivity, Specificity, PPV, and/or NPV using Bayes' rule. (You do not need to actually solve, just leave it in an expression.) (Review example in Lecture 5 slide 17, particularly steps 1-3)
- How can low prevalence vs. high prevalence affect PPV in a setting with a modest false positive rate?
- Discrete distributions. Know the three rules, verify the three rules in an example
- Bernoulli random variable. State a simple example of a Bernoulli random variable. What's the one parameter associated with this distribution?
- For Binomial and Poisson random variables, Know the assumptions for each. State a simple example, verify the assumptions.
- What is expected value, in words? How about variance?

- Density functions (for continuous variables), two rules of density functions.
- Z-scores, how to calculate, why is it helpful to standardize?