PHP 2510 Principles of Biostatistics & Data Analysis

Week 5: Recap & Exploratory Data Analysis

Motivating Random Variables

- Please fill out this survey
- This is for illustrative purposes
- If you do not want to share real responses, feel free to just invent them

Traffic Light Feedback

Let’s look at the survey data

- Imagine this represents our whole world (no sampling, no inference)
- Imagine your collaborator wants you to describe or summarize each variable
- Imagine your professor wants you to specify what happens when you randomly sample one person (because in real life, we will have a sample)

WHAT DO YOU DO? Let’s discuss each column …

More on Expectations

I like playing the scratchers

A new $5 scratcher comes out. It has the following properties:

- I lose $5 – 90% of the time
- I win $20 – 10% of the time

Is it a sound financial decision to play? How do you decide?

What about this scratcher?

- I lose $5 – 80% of the time
- I break even – 10% of the time
- I win $40 – 10% of the time

An expectation is a theoretical value; a long-run average.
Think of it as a "best guess" for the outcome,
even if not a possible outcome (of a single play/sample)

What about variances?

It's just a "special" expectation

1. Center the data on it's (true) mean
2. Square the value
3. Now take the expected value of that!

It measures the typical spread

"Typical" == in the long run

"Spread" == distance from its own mean, squared

This is also called the centralized second moment (second because we squared). We can also calculate the third moment (after standardization, this is called skewness, and measures symmetry); 4th moment = kurtosis, …

Expectations: true vs estimates

When we use a random variable (and it's distribution) to represent some phenomena, we calculate true theoretical expectations

When we have the entire population as our dataset, we can also calculate true theoretical expectation. Why? Let's look at the formula

When we only have a sample as our dataset … NEXT WEEK

NEXT STEPS

- Expectation/Variance Practice Problems
- Week 4 Handout Answer Review
- Cond/Joint Probabilities
  - Lecture
  - Practice Problems

PHP 2510 Principles of Biostatistics & Data Analysis

Week 5: Recap & Exploratory Data Analysis

Data Detection & Graphics in Public Health

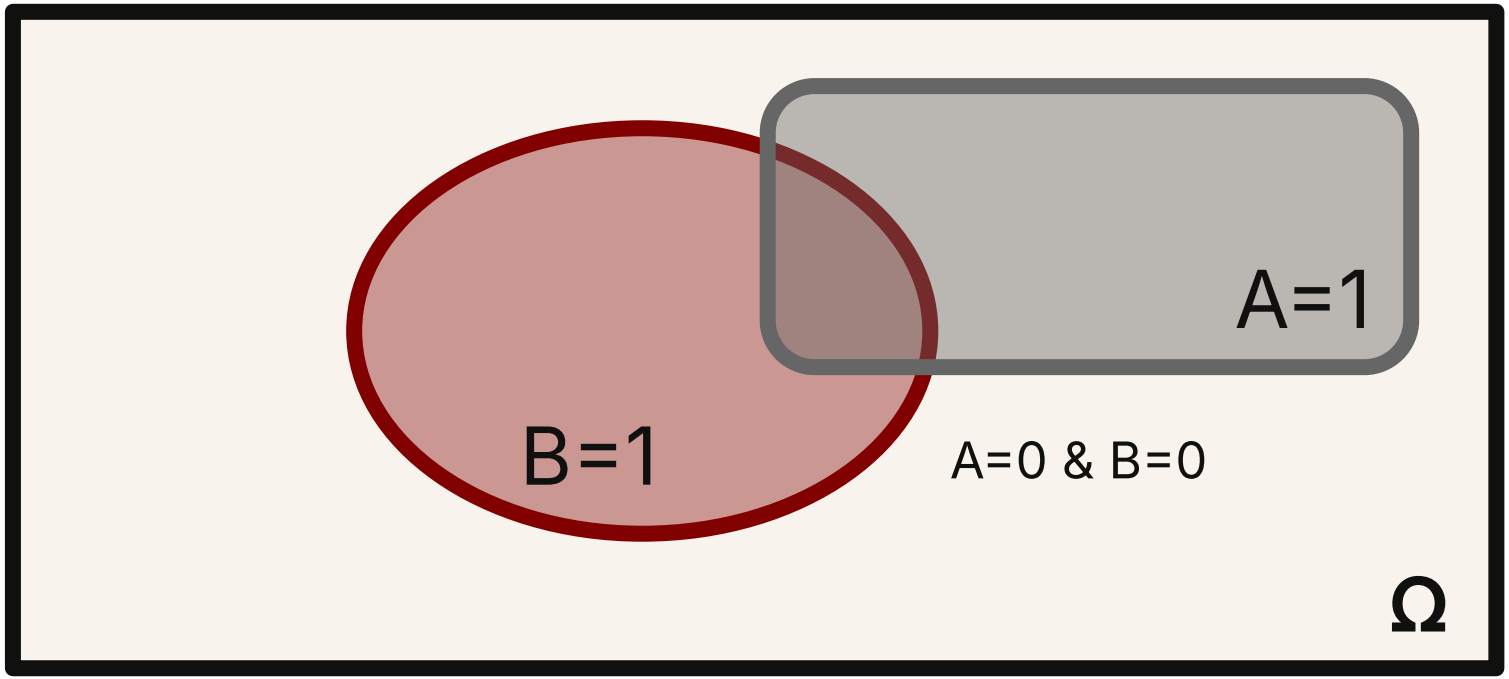
take home activities
due 10/7 @ 9am
lab available for help

After this week's classes, along with the required readings (SPEEGLE Chapter 7*; CHIHARA Chapter 2), you should be able to:

- Create informative univariate and bivariate visualizations (e.g., histograms, boxplots, scatterplots) in R to explore data patterns
- Interpret graphical and numerical summaries to describe the distribution of a variable and the relationship between two variables

*exposure only; you will not be assessed on ggplot2 commands

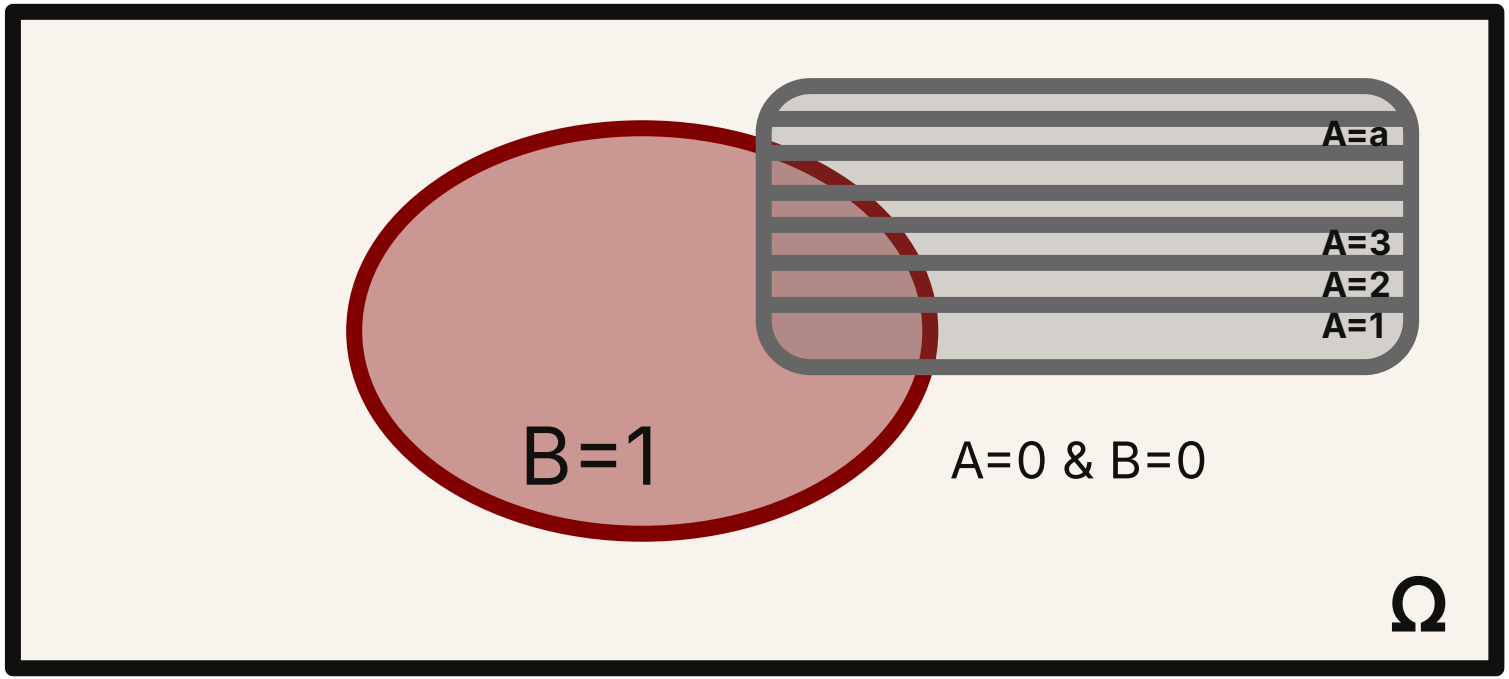
Marginal, Conditional, and Joint Probabilities



A = 1 for people that got hospitalized; B = 1 for people who are uninsured

Ω = Rhode Islanders

Marginal, Conditional, and Joint Probabilities



A: **<u>count</u>** of hospitalizations this year, B =1 for people who are uninsured at any hospitalizations
Ω = Rhode Islanders

Practice Problems

See handout

Challenge Problem

In a community, residents are potentially exposed to two different toxins. The probability of exposure to toxin A is 20%, toxin B is 30%, and both is 15%. The risk of developing a specific cancer is:

- 10% if only exposed to A
- 12% if only exposed to B
- 25% if exposed to both A and B
- 1% if exposed to neither

What is the probability that a randomly selected resident who has the specific cancer was indeed exposed to at least 1 toxin? Use R simulations then confirm, the calculate directly.

Before getting started, let's think about:

- Are the exposure events independent?
- Qualitatively, what should the answer be – i.e. is this event rare or common (or in the middle)?
- Are there any hidden assumptions in this problem?
- What does the pseudocode look like?