Biostats Lecture 11: Review

Public Health 783

Ralph Trane
University of Wisconsin–Madison

Fall 2019

Review

Topics covered:

1. Descriptive Statistics
2. Probability and Random Variables
3. Statistical Hypothesis Tests
4. Confidence Intervals
5. Linear Regression

Descriptive Statistics

- Important to describe your sample
    - to avoid generalizing when you shouldn't
- Distinguish between categorical and continuous variables
    - when categorical
        - tables: frequency counts, relative frequencies
        - graphs: bar chart prefered
    - when continuous
        - tables: mean, sd or median, min, max
        - graphs for one variable: histogram, boxplots
        - graphs for two: scatter plot
- Important to consider if data should be stratified
    - often done if main outcome is categorical

Take-aways from descriptive statistics

What you absolutely need to know:

- why we need to describe/summarize our data
- when to pick what kind of summary
- when to pick what kind of plot
- how to read boxplots, histograms, and bar charts
- how to create the simple summary statistics using `SAS`
    - `proc means` and `proc freq`
- how to generate and boxplots using `SAS`
    - `proc sort` and then `proc boxplot`

Probability and Random Variables

- Probability: the proportion of times an event happens if experiment repeated over and over again

- Random Variable: a variable where the outcome cannot be determined before the experiment

- Distribution:

    - for categorical variable, histogram that gives the probabilities of each outcome
    - for continuous, curve
    - for both, total area must be 1!

Probability and Random Variables

- Important distributions:
  - bernoulli
    - discrete
    - binary outcome, one has probability p, the other probability $1-p$
    - think coin toss, sex, disease status (if disease/healthy)
  - binomial
    - discrete
    - a number of bernoulli's added up
    - think number of individuals in a sample with disease
  - normal
    - continuous
    - occurs often in nature
    - turns out, averages (if enough samples are included) are normally distributed (Central Limit Theorem)

Probability and Random Variables

- Other distributions we have used:
  - t-distribution
    - almost normal, and when n large enough, impossible to tell from normal
  - χ^2 distribution
    - used for χ^2 test

Take-aways from Probability and Random Variables

What you absolutely need to know:

- how to interpret a probability
  - Say an event happens with a probability of $.15$. If you were to repeat the experiment many, many times, we expect the event would happen about 15 percent of the times.

What I hope you have learned:

- better intuition about what "probability" is
- statisticians think about experiments as realizations of random variables
- depending on the experiment, a certain type of distribution is appropriate for the corresponding random variable
- a distribution is something that guides the outcome of a random event

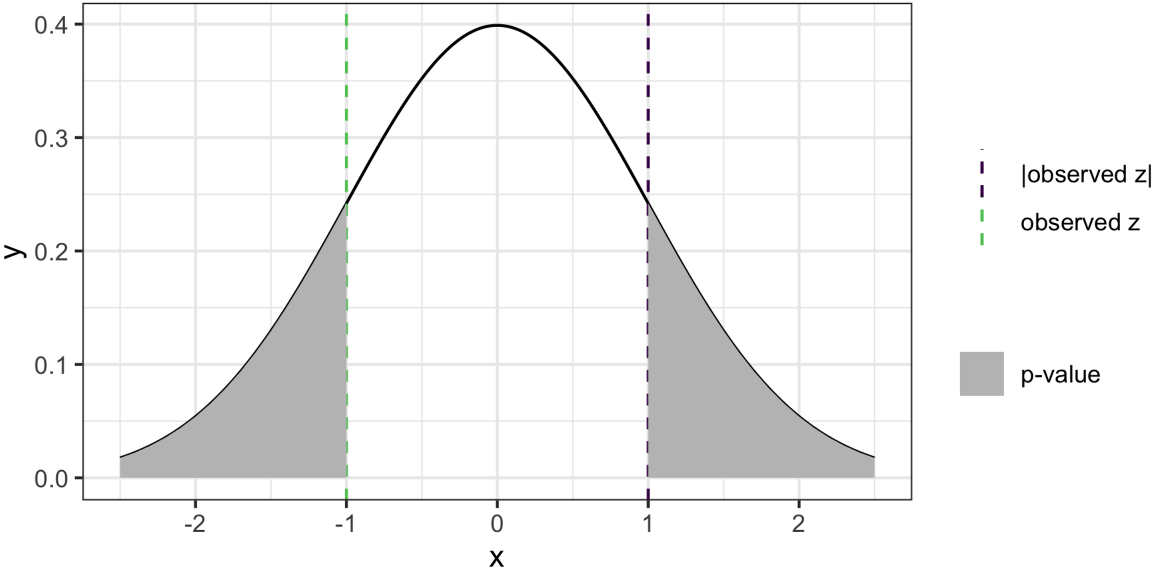
Statistical Hypothesis Tests

General strategy:

- Write down your hypothesis in stats language:
  - if interested if the overall mean is some number, say 2, then $H_0 : \mu = 2$ vs. $H_A : \mu \neq 2$
- Find something that catches the spirit of what you're interested in
  - $\frac{\bar{X}-2}{SD(\bar{X})}$
- Find the distribution of that thing, **IF** H_0 is true:
  - if $\mu = 2$ is actually true, then $\frac{\bar{X}-2}{SD(\bar{X})} \sim N(0, 1)$ since it is of the form "normal random variable minus mean divided by the stanard deviation"
  - remember, average ($\bar{X}$) is normal when n is "large enough"

Statistical Hypothesis Tests

- Find the p-value, i.e. the probability of observing something more extreme **IF** H_0 is true
  - More extreme = data further away from H_0
  - $\bar{X}$ further away from 2
  - $\frac{\bar{X}-2}{SD(\bar{X})}$ further away from 0
  - p-value $= 2 \cdot P\left(Z > \left|\frac{\bar{X}-2}{SD(\bar{X})}\right|\right)$



Statistical Hypothesis Tests

- Went through a bunch (full list here)
- Talked in somewhat detail about:
  - t-test: for means
    - when you want to test if the mean in a group is equal to or different from a number
    - when you want to compare the means in two groups
  - test for proportions:
    - test if the proportion of a population with a certain attribute (disease, for example) is equal to a number (say 0.5)
    - test if two proportions are different
- Also seen:
  - χ^2 test
    - test if two categorical variables are related
  - test for RR

Take-aways from Hypothesis Testing

What you absolutely need to know:

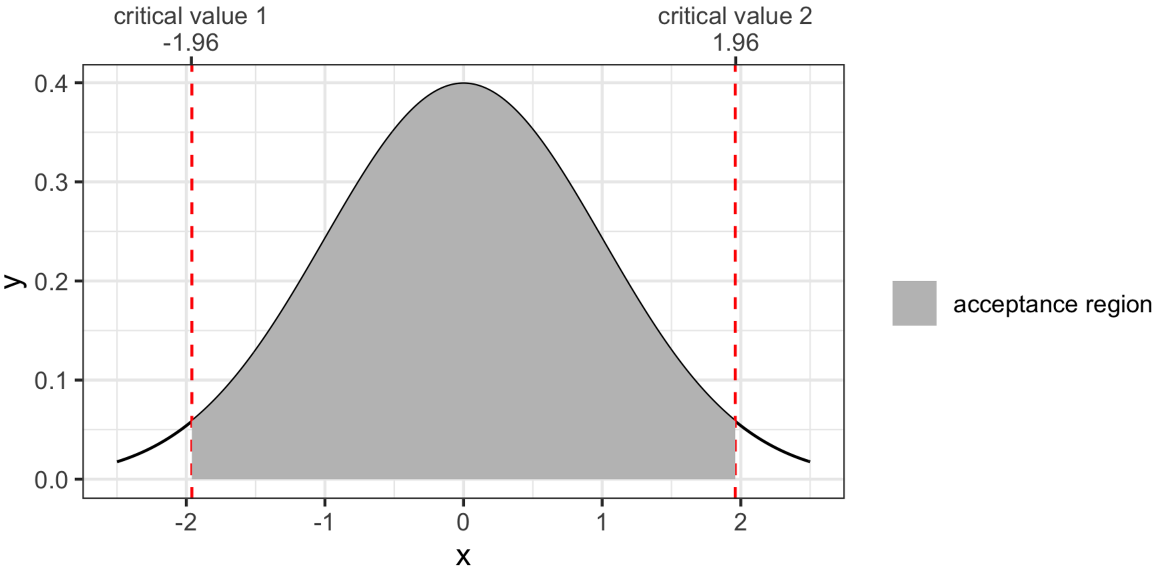
- how to interpret a p-value
    - the probability of something more extreme **IF** H_0 **IS TRUE**
    - if small, evidence against the null, so we reject
    - if large, not evidence against the null, so we do not reject
    - notes:
        - we **NEVER** accept H_0 or H_A. **ALWAYS** either reject of do not reject H_0
        - p-value is **NOT** the probability the null/model is true. Remember, loosely speaking p-value $= P(\text{data}|H_0)$, **NOT** $P(H_0|\text{data})$.
- what test to use when
- how to perform the tests mentioned above in SAS

What I hope you have learned:

- intuition about what a statistical test is, and how it is build

Confidence Intervals

- hypothesis test: "could this one value be the truth?"
- confidence interval: "what range of values could be the truth?"
- all the values that make the test statistic fall between the critical values:



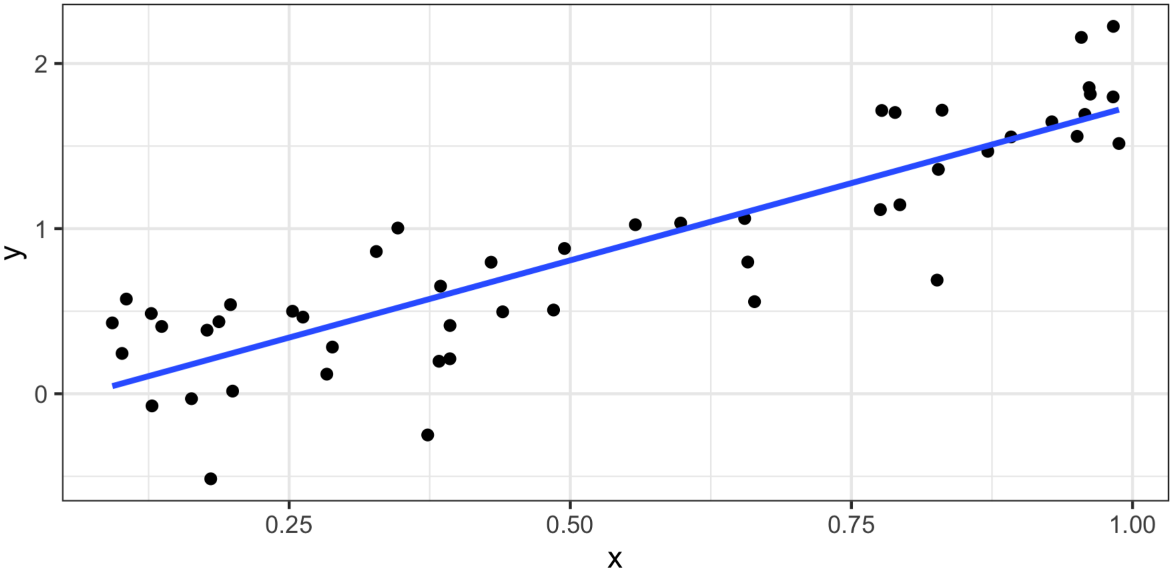
Take-aways from Confidence Intervals

What you absolutely need to know:

- how to interpret a 95% confidence interval:
    - "we are 95% confident that the true value lies in the interval"
    - **NOT** "there's a 95% probability/chance the true value lies in the interval"
        - this makes it seem like the true value varies from experiment to experiment, which is **NOT** how we think about hypothesis testing/confidence intervals/life in general.
- how to compare two confidence intervals:
    - two intervals that do not overlap means there is a statistically significant difference
    - two intervals that do overlap means there is NO statistically significant difference
- how to get confidence intervals for the following quantities in `SAS`:
    - a single mean
    - difference in means
    - a single proportion
    - difference in proportions
    - RR
    - OR

Linear Regression

- Looking for association between two continuous variables
  - find the "best straight line"
  - i.e. find β_0, β_1 such that y_i is as close to $\beta_0 + \beta_1 \cdot x_i$ as possible
  - interpretations:
    - if $\beta_1 = 0$, no association
    - if $\beta_1 < 0$, when x increases, y decreases
    - if $\beta_1 > 0$, when x increases, y increases



Linear Regression

- Also allows us to adjust for other covariates:
  - find $\beta_0, \beta_1, \beta_2$ such that y_i is as close to $\beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \beta_3 \cdot x_{3,i}$ as possible.
  - interpretations
    - if $\beta_1 = 0$: after accounting for everything else, no association
    - if $\beta_1 < 0$: all else being equal, when x increases, y decreases
    - if $\beta_1 > 0$: all else being equal, when x increases, y increases

Take-aways from Linear Regression

What you absolutely need to know:

- ...

What I hope you have learned:

- very basic intuition about linear regression
- some idea of how to interpret coefficients
- hopefully won't be scared when you encounter it in the future (cause you will)