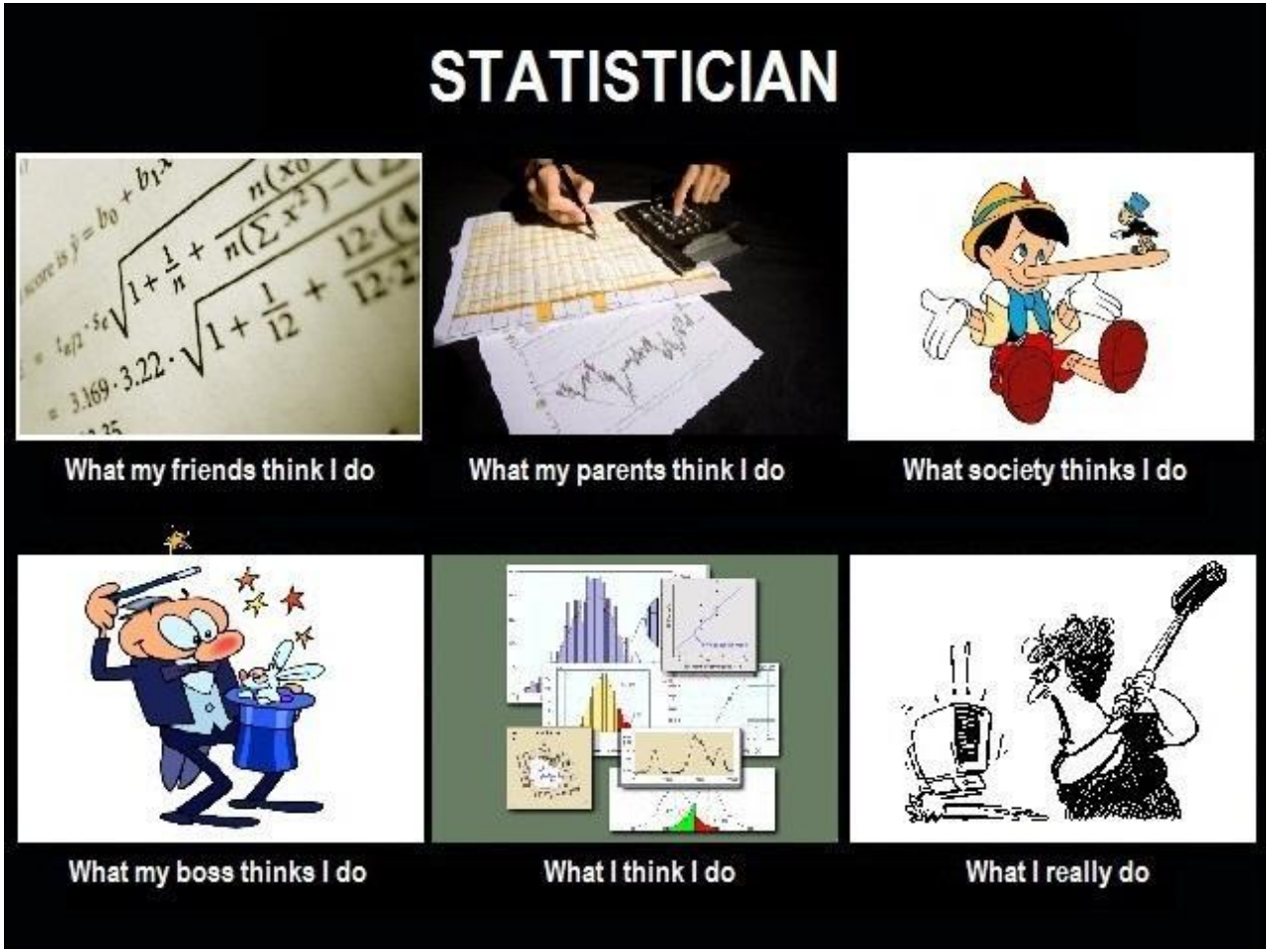
FLORIDA STATE UNIVERSITY
COLLEGE OF MEDICINE

Research Workshop Series #5
Introduction to Biostatistics

Introduction

What is Statistics?



What is Statistics?

- "The goal of data analysis is simple – to make the strongest possible conclusions from limited data.
- Statistics help you extrapolate from a particular set of data (sample) to make a more general conclusion (about the population)."

  – Motulsky, "Intuitive Biostatistics", Chapter 3

Parameter

- A **parameter** is a value of interest corresponding to a **population**.
  - Fixed
  - Unknown
- This is the answer we would like to know when we are designing a study
- We usually cannot find the exact value, because we rarely have complete information on all members of the population.

Statistic

- A **statistic** is the estimate of a parameter base on information contained in a **sample**
  - Varies based on the sample taken
  - Can be calculated
- This is the answer we calculate from the study.
- If we did a different study, we would probably get a different statistic.
- We hope that this is a reasonable approximation of the parameter
  - But we have no way to know how close it is in any particular study

Example: Hypertension

The NHANES (2011-2012) study estimated that 29.1% of the U.S. population suffers from hypertension in the U.S.

- This is a statistic based on a sample of members of the U.S. population
  - If we conducted another similar sample, we would have different participants and thus get a different estimate
- The parameter, the true proportion of the U.S. population that suffers from hypertension, is unknown.

Popular Goals in Statistics

Descriptive Statistics

- Reporting values (statistics) from the sample
- Different statistics to report for different types of data

Statistical Inference

- Estimation
  - Point (e.g. sample mean, sample proportion)
  - Interval (e.g. Confidence Intervals)
- Testing
  - Hypothesis Tests
- Modeling/Prediction
- All of these have ***assumptions*** that we should scrutinize!

Descriptive Statistics

Types of Variables

You probably encounter data on a regular basis in your job

Can you think of some examples?

Types of Variables

Categorical

- Binary
- Nominal
- Ordinal

Quantitative

- Count
- Continuous

Categorical Data

Binary

- Two choices: Yes/No, Group A or B, etc.
- Example: Disease status, exposed/unexposed, gender

Nominal

- More than 2 choices with no inherent order
- Example: Blood type (A, B, AB, O)

Ordinal

- More than 2 choices with an inherent order
- Example: Pain level: None, mild, moderate, severe

Quantitative Data

Count Data

- Can take on many non-negative integer values (0, 1, 2, 3,…)
- Example: Number of Dental caries in a routine cleaning

Continuous Data

- Can take on any value within an interval
- Example: Blood Pressure, Height, Body mass index

Measures of Interest

***Depending on the kind of data you have, you would report different statistics to describe your data

Measures for Categorical Data

- For categorical data, the measures of interest are the ***counts*** or ***proportions*** of the observations that fall within a particular group
- We could report these in a short table, called a frequency distribution
  - Example: Number of patients with and without disease

Disease	Healthy
n_d	n_h

Example Table:
(Beyerlein et al 2011)

"Table 1: Study characteristics of the data analyzed (n = 12,383)"

Excerpt: Child's Age (Among Smoking Mothers)

Age	Count	Percentage
3–6 years	580	28.3
7–10 years	607	29.6
11–13 years	415	20.3
14–17 years	446	21.8

Example

- Descriptive Statistics for Quantitative Data in Slade et al. 2011
- Table 5. Quantitative Measures of Symptom Experiences among 185 Cases with Temporomandibular Disorder (TMD)

Measure	Range	N	Mean	SD	Min	5%	25%	50%	75%	95%	Max
Interference in work due to facial pain	10	185	2	2.6	0	0	0	1	3	8	10
Number of days kept from usual activities	180	182	10.7	29.9	0	0	0	0	6	48	180

Descriptive Statistics

Measures

Categorical	Continuous
Proportions	Location (Mean, Median, Mode)
Counts	Spread (Variance, SD, Quantiles, Range, IQR)
	Other (coefficient of variation)

Graphical Displays

Categorical	Continuous
Bar Chart	Histogram
Pie Chart	Boxplot

Statistics Mnemonic

(Numeric Data)

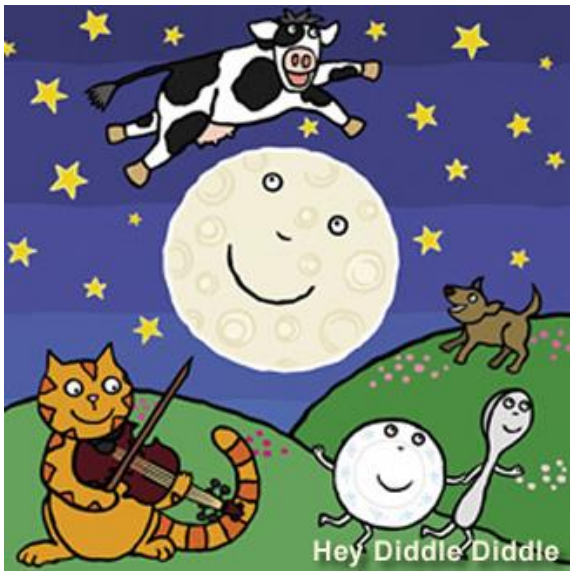
"Hey, diddle diddle,

The **median**'s the **middle**

You add and divide for the **mean**.

The **mode** is the one that appears the **most.**

The **range** is the difference between."



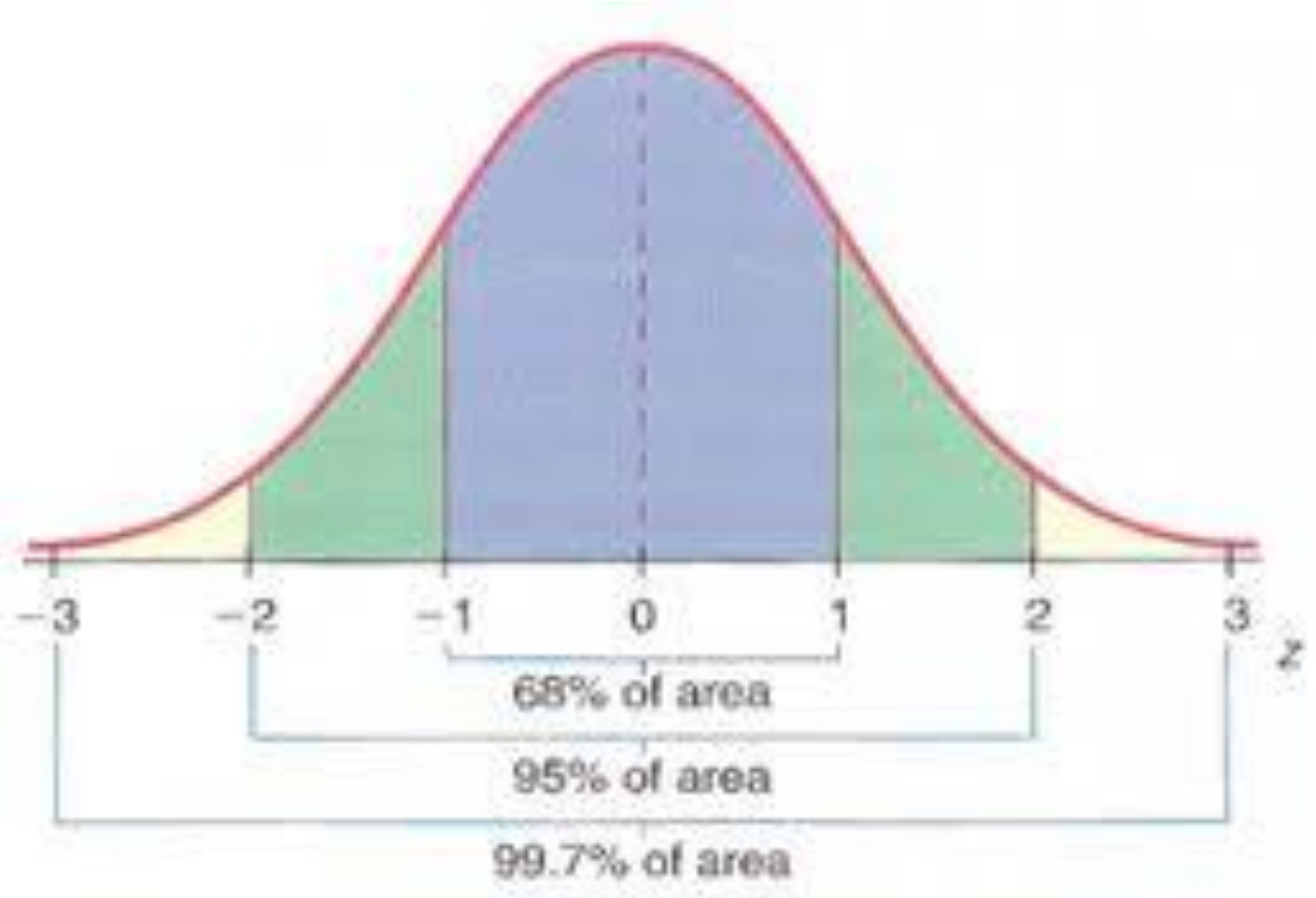
Estimation/Modeling

Background: Normal Distributions

- Normal distributions are everywhere in statistics and data analysis
- They form a special family of continuous distributions
- Properties:
  - Symmetric
  - Bell-shaped
  - Few extreme observations

Normal Distributions

Consider Z ~ N(0,1)



Assumptions

- Before you conduct a statistical test, you MUST consider the assumptions of the test
  - What does the test assume?
  - What happens if the assumptions are violated?
  - Is this test appropriate for my data?
- Same holds if you run a model (and use the corresponding p-values)
- Same holds if you make a confidence interval
  - Connection between tests and confidence intervals

Statistical Tests Revisited

Recall in Presentation 4, you had a handout called “Common Statistical Tests”

We will look at each item and carefully consider assumptions

Example: Statistical Tests

- 1 sample:
  - Is my mean equal to a pre-specified number?
- 2 samples:
  - Are my two means equal?
- 3+ samples:
  - Are all three of my means equal

Review of Common Tests: Quantitative Data

1-Sample T-Test	Indep. T-test	Paired T-test	ANOVA
1 quantitative variable	2 independent quantitative variables	2 dependent quantitative variables	3 or more quantitative variables
Independent observations	**Independent observations and samples**	**Independent pairs** of observations (can be related within a pair)	**Independent observations and samples**
Normally distributed population or a large sample	**Normally distributed** populations or large samples	**Normally distributed** populations (differences) or large samples	**Normally distributed** populations, equal variances

Problem: Small non-normal samples!

Nonparametric Tests

- 1 sample:
  - Is my median equal to a pre-specified number?
- 2 samples:
  - Are my two medians equal?
- 3+ samples:
  - Are all three of my medians equal

Review of Common Nonparametric Tests

Sign test	Wilcoxon rank sum (Mann-Whitney)	Wilcoxon sign-rank	Kruskal Wallis
1 quantitative variable or the differences between paired quantitative variables	2 independent quantitative variables	2 dependent quantitative variables	3 or more quantitative variables
Independent observations	At least 8 **Independent observations, and independent samples**,	At least 6 **Independent observations**, Assumes the difference between the variables is **symmetric**!	**Independent observations and samples**

Correlation Tests

- Pearson:
  - Are my two variables linearly related with a nonzero slope?
    - Yes, this is equivalent to testing the slope in a simple linear regression model (coming up)
- Spearman:
  - Are the ranks of the values of my two variables linearly related?
  - i.e. Are my two variables increasing or decreasing together?

Review of Common Correlation Tests

Pearson	Spearman
2 continuous variables	2 continuous or ordinal variables
Independent pairs of **normally distributed** observations, variables are **linearly related**	At least 5 **Independent pairs**

Tests for Categorical Data

- 1 sample:
  - Are the proportions of observations for each possibility consistent with my initial guess?
- 2 samples:
  - Are the proportions of each possibility equal in two different samples?

Review of Common Tests: Categorical Data

Chi Square Goodness of fit	Chi Square Indep.	Fisher's Exact Test
1 Categorical Variable	2 (or more) Categorical variables	2 (or more) categorical variables
Independent observations, at least 5-10 expected in all cells	**Independent observations**, at least 5-10 expected in all cells	**Independent Observations**

Linear Regression

- Simple
  - If I made a scatterplot based on my 2 variables, and drew a line through all the points, would that line be horizontal?
- Multiple
  - Similar interpretation, between outcome and each predictor, holding all others constant

Review of Linear Regression

Simple	Multiple
1 continuous outcome, 1 continuous predictor	1 continuous outcome, 1 or more predictors (they can be continuous, nominal or ordinal)
pairs of observations, variables are **linearly related,** errors (residuals) are **normally distributed and independent**, predictor is measured precisely	**groups** of observations, outcome is **linearly related** to the predictors**,** errors (residuals) are **normally distributed and independent**, predictors are measured precisely and not linearly related to each other

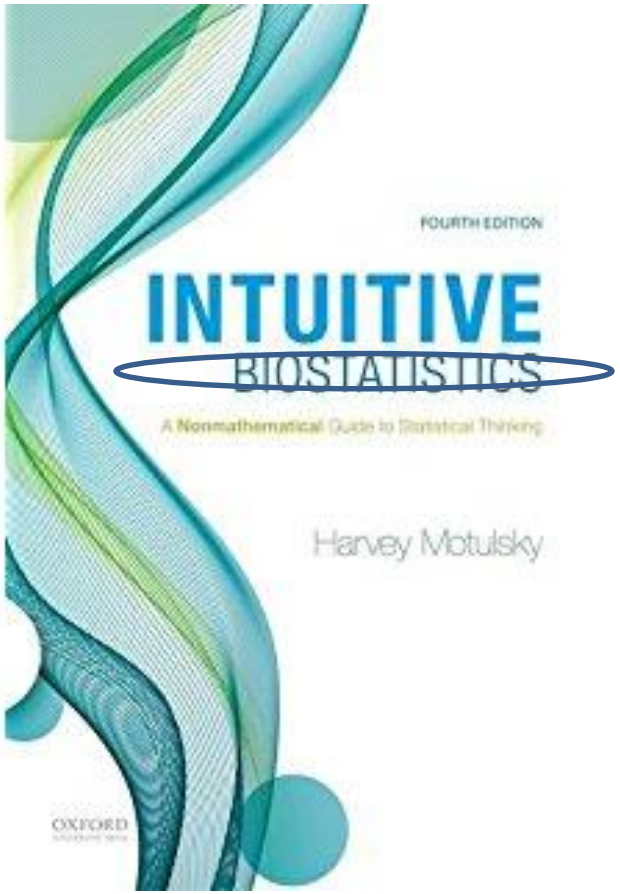
The Importance of Proper Sampling and Analysis

- The way the sample is selected (i.e. the study design), determines if the results are valid!
- Bad study designs yield bad results, may give misleading conclusions, and results from the sample are not generalizable to the population
- Similarly, inappropriate statistical analyses yield invalid conclusions.
- "Garbage in, garbage out!"

Choose Statistics Wisely

- Many analyses are possible, but only a few make sense
- Always look at graphs to visualize your data!
- Always critically evaluate assumptions
- Always consider the broader picture to make sure you are doing analyses that make sense

Recommended Reading



- Excellent reference
- Not a traditional stats text book
- Paperback
- Words, not numbers
- Just concepts

Thank you!

Questions & Discussion