

---

# General Biostatistics

Part 1

1

---

---

---

---

---

---

---

---

---

# General Biostatistics

Marie Diener-West, Ph.D.  
Department of Biostatistics  
Johns Hopkins University  
Bloomberg School of Public Health

2

---

---

---

---

---

---

---

---

---

# Course Schedule

- Overview of biostatistics
- Probability and probability distributions
- Methods for statistical inference
  - Estimation
  - Hypothesis testing
- Common statistical tests
  - Comparing means of 2 groups
  - Comparing proportions of 2 groups

3

---

---

---

---

---

---

---

---

## Course Schedule (continued)

---

- Common statistical methods
  - Comparing more than 2 groups
- Sample size and power
- Simple linear regression
- Summary

4

---

---

---

---

---

---

---

---

## Overview of Biostatistics

---

Key Ideas and  
Descriptive Statistics  
Class 1A

5

---

---

---

---

---

---

---

---

## Outline

---

- The scientific method
- The role of biostatistics in science
- The biostatistics paradigm
- Types of variables and measurement scales
- Example of scientific evidence from studies
- Exploratory data methods

6

---

---

---

---

---

---

---

---

## Scientific Method

- Competing hypotheses about nature
  - $H_0, H_1, H_2, H_3, \dots, H_N$
- Design a study and generate data
- Data are evidence in support of some of the hypotheses more than others
- Science is a process of eliminating hypotheses whose predictions are inconsistent with observation (data)

7

---

---

---

---

---

---

---

---

## What is Biostatistics?

- **BIO**statistics- the application of statistical reasoning and methods to the solution of biological, medical and public health problems
- **BioSTATISTICS** - scientific use of quantitative information to describe or draw inferences about natural phenomena

8

---

---

---

---

---

---

---

---

## Role of Biostatistics in Science

- 1 Generate hypotheses: ask questions.
- 2 Design and conduct studies to generate evidence; make observations; collect data
- 3 Descriptive statistics: describe the observations
- 4 Statistical inference: assess strength of evidence for/against a hypothesis; evaluate the data

9

---

---

---

---

---

---

---

---

## Descriptive Statistics

- Exploratory data analysis (EDA)
- Organization and summarization of data
- Tables of summary information
- Graphical display of important patterns and variation
- Hypothesis generating

10

---

---

---

---

---

---

---

---

## Statistical Inference

- Confirmatory data analysis (CDA)
- Draw conclusions about a population from a sample
- Assess strength of evidence for competing hypotheses
- Make comparisons
- Make predictions

11

---

---

---

---

---

---

---

---

## Statistical Inference (continued)

Statistical inference uses data to surmise what is true or likely to be true

12

---

---

---

---

---

---

---

---

## Key Ideas in Statistical Reasoning

- Natural laws do not perfectly predict all phenomena.
- Probability models are useful tools.
- Variation is itself a natural phenomenon.
- Variation leads to uncertainty about an event.
- There are important patterns to be discovered in the midst of variation.

13

---

---

---

---

---

---

---

---

## Sources of Variation in Data

- Natural variation
- Measurement variation (error)
- Bias - difference between the average (expected) value of a measurement and the true value
- Variance - variation among measurements about their average

14

---

---

---

---

---

---

---

---

## Biostatistics Paradigm

Attempt to discover simple explanations for phenomena - the inter-relationships between variables

- **explanations** - hypotheses about mechanisms
- **variable** - a characteristic taking on different values
- **simple** (principle of parsimony)
- **inter-relationships** - associations; causal connections

15

---

---

---

---

---

---

---

---

## Types of Variables

- Random variable - values obtained arise partly as a result of chance factors
- Response variable (Y) - outcome measure (that which is affected or caused)
- Explanatory variables (X) - those which affect or cause the response

16

---

---

---

---

---

---

---

---

## Measurement Scales

- Quantitative: amount; numerical
  - Discrete
  - Continuous
- Qualitative: attribute; categorical
  - Nominal scale
  - Ordinal scale

17

---

---

---

---

---

---

---

---

## Example: Scientific Evidence

- Aceh, Indonesia Vitamin A Trial
- 25,939 preschool children in 450 Indonesian villages in northern Sumatra
- 200,000 IU vitamin A given at 1-3 months after the baseline census and again 6-8 months later
- Consider 23,682 out of 25,939 who were visited on a pre-designed schedule

18

---

---

---

---

---

---

---

---

## Example: Scientific Evidence

- Randomized community trial of children in 450 villages

Vitamin A	Alive at 12 Months- Yes	Alive at 12 Months- No	Total
No	11,514	74	11,588
Yes	12,048	46	12,094
Total	23,562	120	23,682

- Does Vitamin A reduce mortality?

19

---

---

---

---

---

---

---

---

---

---

## Example: Scientific Evidence

- Mortality rates per 1,000 child-years
  - No Vitamin A:  $74/11,588 = 6.4$
  - Vitamin A:  $46/12,094 = 3.8$
- Does Vitamin A reduce mortality?
- Calculate a risk ratio or “relative risk”  
$$\frac{\text{Rate with Vitamin A}}{\text{Rate without Vitamin A}} = \frac{3.8}{6.4} = 0.59$$
- 40 percent reduction in mortality!

20

---

---

---

---

---

---

---

---

---

---

## Example: Scientific Evidence

- Does Vitamin A cause this reduction?
- The role of randomization is to balance other observed and unobserved factors.
- The Aceh study is only one small part of the Indonesian population
  - If the study was performed again, would we get the same or similar results?
  - How strong is this evidence that Vitamin A works?

21

---

---

---

---

---

---

---

---

---

---

## Exploratory Data Analysis

---

- LOOK at your data
- Look at detailed distributions before computing summary measures or performing statistical analyses
- Gain insight into the nature of the data set
- Look for errors, anomalies

22

---

---

---

---

---

---

---

---

## Methods for Organizing Data

---

- Ordering data
  - tallies, stem and leaf displays
- Grouping data
  - frequency distributions, percentiles
- Summarizing data
  - measures of central tendency and dispersion
  - box and whiskers plots
- Displaying data in tables and graphs

23

---

---

---

---

---

---

---

---

## Ordering Data

---

- Example: student ages ( $n=10$ )
  - 35, 40, 52, 27, 31, 42, 43, 28, 50, 35
- Could order by hand: tally
  - 20 - 29 //
  - 30 - 39 ///
  - 40 - 49 ///
  - 50 - 59 //

24

---

---

---

---

---

---

---

---

## Ordering Data

- Stem and leaf display: aids in sorting and ordering; more information than a tally

– 35, 40, 52, 27, 31, 42, 43, 28, 50, 35

2 | 78    ⇔    2 | 78

3 | 515    3 | 155

4 | 023    4 | 023

5 | 20    5 | 02

25

---

---

---

---

---

---

---

---

## Grouping Data

- Frequency distribution

20-29    2

30-39    3

40-49    3

50-59    2

Total    10

26

---

---

---

---

---

---

---

---

## Grouping Data

- Relative frequency distribution

20-29    2    0.2

30-39    3    0.3

40-49    3    0.3

50-59    2    0.2

Total    10    1.0

27

---

---

---

---

---

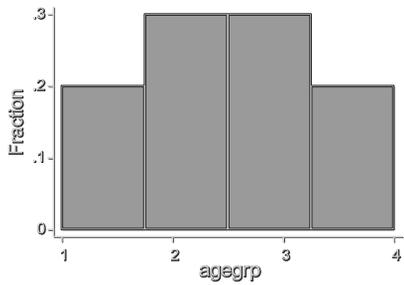
---

---

---

## Grouping Data

---



28

---

---

---

---

---

---

---

---

## Grouping Data

---

- Cumulative frequency distribution

20-29	2	2
30-39	3	5
40-49	3	8
50-59	2	10
Total		

29

---

---

---

---

---

---

---

---

## Grouping Data

---

- Cumulative relative frequency distribution

20-29	2	0.2
30-39	5	0.5
40-49	8	0.8
50-59	10	1.0
Total	10	

30

---

---

---

---

---

---

---

---

## Grouping Data - Percentiles

- The  $r^{\text{th}}$  percentile  $P_r$  is the value that is greater than or equal to  $r$  percent of a sample of  $n$  observations
- Quartiles and percentiles
  - $P_{50} = Q_2 = \text{median} = \text{average of the middle two observations}$
  - $P_{25} = Q_1 = \text{median of the lower half of the data}$
  - $P_{75} = Q_3 = \text{median of the upper half of the data}$

31

---

---

---

---

---

---

---

---

## Grouping Data - Percentiles

- Example: student ages
  - 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
- Quartiles
  - $Q_2 = \text{median} = \text{average of the middle two observations} = (35+40)/2 = 37.5 \text{ years}$
  - $Q_1 = \text{median of the lower half of the data} = 31$
  - $Q_3 = \text{median of the upper half of the data} = 43$

32

---

---

---

---

---

---

---

---

## Summarizing Data

- Measures of central tendency
  - Mean (average)
  - Median = middle observation
  - Mode = most frequent observation
- Measures of dispersion or variability
  - range = largest value - smallest value
  - variance = average of the sum of the squared differences of each observation from the sample mean
  - standard deviation

33

---

---

---

---

---

---

---

---

## Summarizing Data

---

- Example: student ages
  - 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
- Measures of central tendency
  - Mean (average) = 38.3 years
  - Median = middle observation = 37.5 years
  - Mode = most frequent observation = 35 years

34

---

---

---

---

---

---

---

---

## Summarizing Data

---

- Example: student ages
  - 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
- Measures of dispersion or variability
  - range= difference between the largest and smallest values =  $52 - 27 = 25$  years
  - variance = average of the squared differences of each observation from the sample mean =  $s^2 = 74.7$  years<sup>2</sup>
  - standard deviation =  $s = 8.6$  years

35

---

---

---

---

---

---

---

---

## Summarizing Data

---

- Example: student ages
  - 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
- Ratios, proportions, rates
  - A **ratio** is one number divided by another
  - A **proportion** is a ratio in which the numerator is a subset of the denominator
  - A **rate** is a ratio in which calendar time enters the numerator in the same way in which it enters the denominator

36

---

---

---

---

---

---

---

---

## Summarizing Data

- Example: student ages
  - 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
- Ratios, proportions, rates
  - **Ratio** =  $4/2$  = the number of students over age 40 divided by the number of students under age 30
  - **Proportion** =  $4/10 = 0.4 = 40\%$  = the number of students over age 40 divided by the total
  - **Rate** = number of new students divided by total number of students as of July 1

37

---

---

---

---

---

---

---

---

## Displaying Data

- Box and whiskers plot
  - Upper hinge = Q3
  - Median = Q2
  - Lower hinge = Q1
  - Inter-quartile range:  $Q3 - Q1 = IQR$
  - "Fences" identify outliers
    - Upper fence =  $Q3 + 1.5(IQR)$
    - Lower fence =  $Q1 - 1.5(IQR)$
  - "Whiskers"

38

---

---

---

---

---

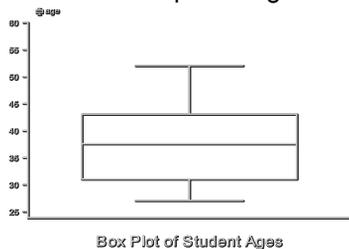
---

---

---

## Displaying Data

- Box and whiskers plot of age



Box Plot of Student Ages

39

---

---

---

---

---

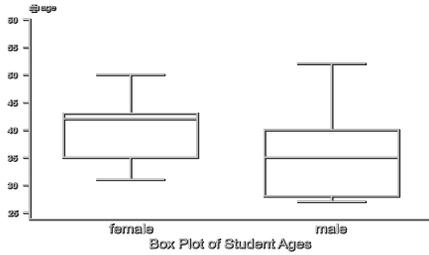
---

---

---

## Displaying Data

- Box and whiskers plot of age by gender



40

---

---

---

---

---

---

---

---

## Summary

- Biostatistics is the science of using quantitative information to deal with uncertainty and variation
- Variables and measurement scales
- Descriptive statistics
- Exploratory data analysis
  - LOOK at your data
  - Identify patterns and outliers

41

---

---

---

---

---

---

---

---