Epidemiologic Measures of Disease Burden and Distribution: Part I Descriptive Statistics

OUHSC College of Medicine

Foundations of Biostatistics and Epidemiology

Julie Stoner, PhD

1

Hello, my name is Julie Stoner. I am a professor of Biostatistics at the University of Oklahoma Health Sciences Center.

In this 4-part series, we will discuss epidemiologic measures of disease burden and distribution.

In this first module, I present information related to descriptive statistics.

Learning Objectives

- Utilize descriptive statistics to summarize central tendency and spread of a sample of data

2

After viewing this module, you will be able to utilize descriptive statistics to summarize the central tendency and the spread, or variation, in a sample of data.

Example 1: Duration of Treatment for Postmenopausal Osteoporosis

- **Context:** The optimal duration of treatment of women with postmenopausal osteoporosis is uncertain.
- **Objective:** To compare the effects of discontinuing alendronate treatment after 5 years vs continuing for 10 years among patients receiving low or high dose Alendronate.
- **Design and Setting:** Randomized, double-blind trial conducted at 10 US clinical centers that participated in the Fracture Intervention Trial (FIT)

Reference: *JAMA. 2006;296:2927-2938*

3

I will first introduce several examples of published studies that we will refer to throughout this series.

In the first study, the investigators aimed to determine the optimal duration of treatment of women with postmenopausal osteoporosis. Using a randomized, double-blind clinical trial, the investigators compared the effects of discontinuing alendronate treatment after 5 years vs continuing treatment for 10 years among patients receiving low or high dose Alendronate.

Example 2: Primary Biliary Cirrhosis

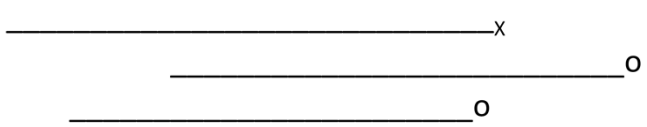
- Aim: Demonstrate that D-penicillamine (DPA) is effective in prolonging the overall survival of patients with primary biliary cirrhosis of the liver (PBC)
  - Mayo Clinic
  - Double-blind, placebo controlled, randomized trial
  - 312 patients
  - Collect clinical and biochemical data on patients
  - Reference: *NEJM.* **312**:1011-1015.1985.

4

In the second example, investigators were interested in determining the effect of D-penicillamine (DPA) in prolonging the overall survival of patients with primary biliary cirrhosis of the liver. This was a double-blind, placebo-controlled randomized trial that was conducted among 312 patients at the Mayo Clinic. In addition to investigating the treatment efficacy question, the investigators also collected detailed clinical, demographic, and biochemical information on patients at baseline to identify factors associated with poorer prognosis.

Example 2 (cont.)

- Patients enrolled over 10 years, between January 1974 and May 1984
- Data were analyzed in July 1986
- Event: death (x)
- Censoring: some patients are still alive at end of study (o)



5

In this clinical trial, patients were enrolled over a 10 year period, beginning in January 1974 and ending in May 1984. The enrollment period was then followed by a 2-year follow-up period and data were analyzed in July 1986. Patient survival was calculated as the time from enrollment (randomization) and death or the time point of last follow-up. Patients who did not die during the follow-up period are said to have a time of death that is censored. The patient is known to have survived up to a certain point in time, but the status is not known after the point of last contact or the end of the study follow-up period.

Let's consider some patient examples. The first patient is enrolled early in the enrollment period (January 1974) and dies in December 1984. The second patient is enrolled later, in January 1979 and remains alive during the entire course of the study follow-up period (through June 1986). The date of death is censored for this patient. We don't know the date of death, but only know that the patient remained alive through June 1986. A similar outcome is seen for the third patient.

The data summary and data analysis methods that we use will need to account for this rolling enrollment period and the censoring of outcomes.

Before moving on to discuss specific types of descriptive statistics, let's first discuss the setting of medical research.

Why is medical research necessary?

- Uncertainty: unknown whether a new drug or therapy is effective and safe
- Variability: not all patients respond in the same way
- Confounding factors: observed associations with a therapy or intervention may be due to confounding factors
  - Example: people who take a certain vitamin may also exercise and eat more healthful foods and therefore have a lower risk of cancer

7

Why is medical research necessary? Why do we conduct medical research? Why are these studies and the results important relative to our clinical practice?

We conduct medical research to answer questions. We don't know is a new drug or therapy is effective and safe. We don't know if a new diagnostic method is accurate. In order to answer such questions, we need to conduct medical research studies.

Secondly, there is variability in patient responses, variability in our implementation of protocols, and variability in processes. Therefore, we need to make multiple measures on multiple patients or settings in order to identify signals, or associations or trends, in data amidst the variability.

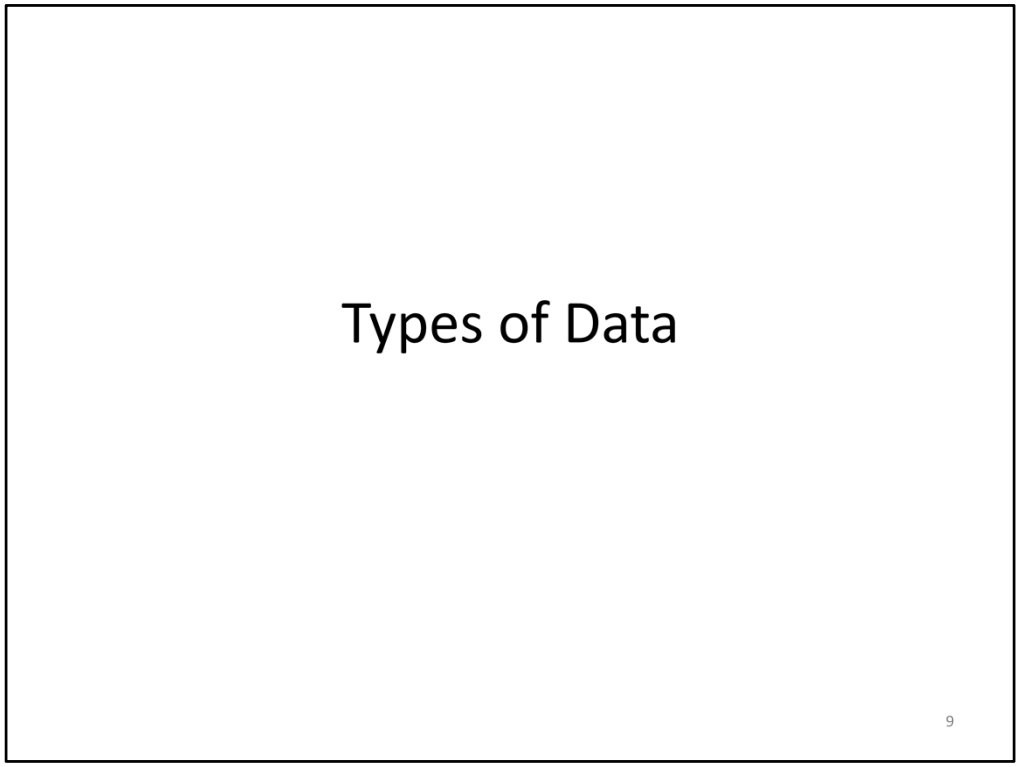
A third motivation for conducting medical research is confounding. In practice, we may observe associations, for example, we may observe that patients who take a certain vitamin appear to have a lower risk of cancer; however, upon closer investigation, we may find that those who take the vitamin also exercise more and eat more healthful foods. Therefore, the effect of vitamin use on the risk of cancer is confounded by other health behaviors. In the presence of such confounding factors, we need to conduct carefully designed medical research studies to better understand the health impacts of particular health behaviors or exposures.

Sample vs. Population

- Population describes the hypothetical (and usually) large number of people to whom you wish to generalize
- Sample describes those individuals who are in the study (fraction of the population)
  - The study is only generalizable to the type of patients that are in the study

8

In practice, we are not able to study an entire population of patients in a given medical research study. Instead, we will study a sample of participants from the target population and will make inference from the sample of participants to the target population at large. Our ability to generalize results from the sample to a given population depends on the sampling selection strategy and the methods of the study. For example, if we only collect information on women in our study sample, we cannot necessarily generalize the results to a broad population of both men and women.



Now, let's discuss the main types of data that we collect in practice. The type of data will drive our data analysis decisions.

Categorical Data

- Provide qualitative description
- Dichotomous or binary data
  - Observations fall into 1 of 2 categories
  - Example: male/female, smoker/non-smoker
- More than 2 categories
  - Nominal: no obvious ordering of the categories
    - Example: blood types A/B/AB/O
  - Ordinal: there is a natural ordering
    - Example: Likert scales,
      Extreme Pain, Moderate Pain, Mild Pain, No Pain

10

First, we will discuss categorical data variables.

Categorical measures provide a qualitative description, such as male or female and level of satisfaction, ranging from very satisfied to very dissatisfied.

When there are only two response categories, we refer to the categorical variable as a dichotomous or binary variable. Examples include male vs. female or if we simply classify smoking status and a current smoker or non-smoker.

When we have more than two categories, we will consider whether the categories are ordered or not. If there is no natural ordering of the categories, such as for blood type, we refer to the variable as a nominal variable. However, if there is a natural ordering, say when evaluating pain on a 5-point pain scale ranging from extreme pain to no pain, we refer to the variable as an ordinal variable.

The data analysis methods that we use will reflect the type of data.

Numerical data (interval/ratio data)

- Provide quantitative description
- Discrete data
  - Observations can only take certain numeric values, counts
  - Example: number of rejection episodes in 12 months
- Continuous data
  - Not restricted to take on certain values
  - Often measurements
  - Example: height, weight, age

11

Numeric data variables provide a quantitative measure and are sometimes referred to as interval or ratio data.

A discrete variable is one that can only take on certain numeric values, and are often time counts, for example, the number of children in a household or the number of rejection episodes in a 12-month period.

A continuous measure, on the other hand, is not restricted to take on only a certain number of values and is often a measurement. Examples include weight, height, and laboratory measures such as total cholesterol. Given any two values, we can (in theory) always find a patient with a value in between the two given values.

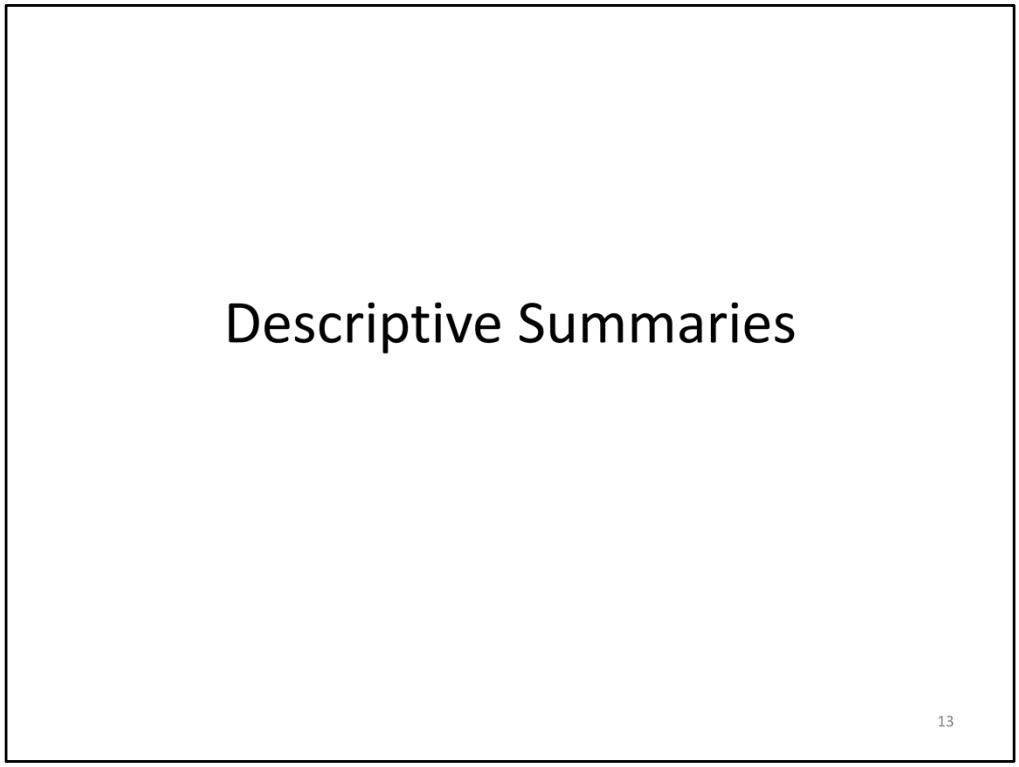
Time to Event Data

- Time between initial and terminal events
  - Initial event: initiation of therapy
  - Terminal event: death, relapse, rejection
- Issues
  - Censoring: Not all subjects experience terminal event
    - Loss-to-follow-up
    - Alive and event-free at end of follow-up
    - Death before event of interest

12

A third type of variable is a time to event measure.

Time to event measures reflect both the timing of an event and the occurrence of the event. For patients who do not develop the event of interest, say death, during the follow-up period, we say that their events are censored.



Now, let's move on to types of descriptive summaries.

Why are descriptive summaries important?

- Identify signals/patterns from noise
- Understand relationships among variables
- Formal hypothesis testing should agree with descriptive results

14

Descriptive statistics are an important first step in any data analysis. Descriptive statistics allow us to detect signals, trends or differences in the data amidst noise and variation. We gain insights into the relation among variables, for example how various exposures may be associated with outcomes. Finally, our formal statistical hypothesis test results should agree with what we see in the descriptive data summaries.

Types of Descriptive Summaries

- Descriptive statistics
  - Measures of location
  - Measures of spread
- Descriptive plots

15

When discussing descriptive statistics, we will focus on measures of location and measures of spread.

In the next module, we will focus on descriptive plots.

Numeric Data: Location

- Measures of location
  - **Mean**: average value
  - **Percentile**: value that is greater than a particular percentage of the data values
  - **Median**: the 50th percentile, 50% of the data values lie below the median

16

Measures of location represent the center of the distribution.

Common measures of the center of the distribution include the mean, or the average value, as well as the 50th percentile or the median. The 50th percentile is the point at which half of the data lie above the cut-point and half of the data lie below. The median is the halfway point among the ordered observations.

Comparison of Mean and Median

- Example: Compare the mean and median from the sample of triglyceride levels

  130, 141, 148, 148, 152, 159, 230

  Mean = 1108/7=158.29, Median = 148

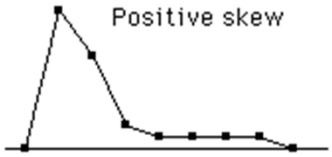
  – The mean may be influenced by extreme data points.

17

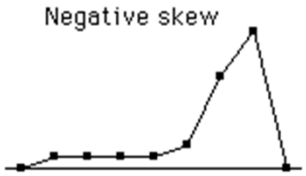
As we can see from this simple example, the mean and median will not always agree. In this example, we have triglyceride measures from 7 patients. We note that six of the measures fall within the range of 130 to 159 mg/dl while the value for one patient is extremely high with a value of 230 mg/dl. The mean, with a value of 158.29 mg/dl, is influenced by this very high value and is higher than the median, with a value of 148 mg/dl. The median, as a midpoint, reflects the rank ordering of the measures, but not the values beyond this rank order.

Skewed Distributions

- Data that are not symmetric and bell-shaped are *skewed*.



Positive skew, or skewed to the right, mean > median



Negative skew, or skewed to the left, mean < median

- Mean may not be a good measure of central tendency. Why?

18

In practice, when there are extreme low or high measures, and the sample size is small, the mean may not be a good measure of the central tendency because it is influenced by the extreme measure. When analyzing variables that are positively skewed (with a long right-hand tail), the mean will be greater than the median (again, the mean is influenced by the extreme positive measures). On the other hand, when analyzing variables that are negatively skewed (with a long left-hand tail), the mean will be lower than the median (the mean is influenced by the extreme low or negative measures). In settings with skewed distributions, in small sample size situations, the median is preferred as a measure of central tendency over the mean.

Numeric Data: Spread

- **Example:**

  1) 2 60 100 $\bar{x}$ =54

  2) 53 54 55 $\bar{x}$ =54

- Both data sets have a mean of 54 but scores in set 1 have a larger range and variation than the scores in set 2.

19

We can see the importance of reporting the central tendency of a numeric measure.

Now, let's consider the two sets of data presented on this slide. In both cases, the mean is 54, but what differs between the data sets?

Measures in the first example are much more variable than measures in the second example.

From this simple example, we can see the importance of reporting both the central tendency and the spread when summarizing the distribution of measures.

Numeric Data: Spread (cont.)

- Measures of spread
  - Variance: average squared deviation from the mean

    For n data points, x_1, x_2, …, x_n the variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - Standard deviation: square root of variance, in same units as original data

20

A common measure of the spread of a continuous numeric variable is the variance.

The variance is the average of the squared difference between each individual observation and the mean. The more spread out the measures are relative to the mean, the larger the variance.

The standard deviation is calculated by taking the square root of the variance. This provides a measure of spread that is in the same units as the original data.

Standard Deviation versus Standard Error of the Mean

- Standard deviation: How much variability can we expect among individual responses?
- Standard error of the mean: How much variability can we expect in the mean response among various samples?

21

In practice, we may see the standard deviation reported, while in other reports, we may see the standard error of the mean reported. How do these measures differ?

The standard deviation indicates how much variability there is among individual observations while the standard error of the mean indicates how much variability we can expect among mean responses from various samples of data.

Standard Deviation versus Standard Error of the Mean (cont.)

- The standard error of the mean is estimated as

$$s.e.m. = \frac{s.d.}{\sqrt{n}}$$

where s.d. is the estimated standard deviation

- Based on the formula, will the standard error of the mean will always be smaller or larger than the standard deviation of the data?
  - Answer: smaller

22

The standard error of the mean is calculated as the standard deviation divided by the square root of the sample size.

Based on the formula, assuming a sample size greater than one, will the standard error of the mean will always be smaller or larger than the standard deviation of the data?

The standard error of the mean will always be smaller than the standard deviation.

This makes sense. The mean is a smoothed measure, averaging values among observations within a sample, and will be less variable than the individual observations.

Numeric Data: Spread (cont.)

- Measures of spread
  - Minimum, maximum
  - Range: maximum-minimum
  - Interquartile range: difference between 25^{th} and 75^{th} percentile, values that encompass middle 50% of data

23

Other measures of spread include the minimum, the maximum and the difference between these measures, the range.

In addition, we can report certain percentiles of the distribution to summarize the spread of the data. As an example, the interquartile range, calculated as the difference between 25^{th} and 75^{th} percentile, or the values that encompass the middle 50% of the data, are also commonly presented as a measure of the spread of the data.

Descriptive Statistics: Numeric Data

- Helpful to describe both location and spread of data
  - Location: mean
    Spread: standard deviation
  - Location: median
    Spread: min, max, range
    interquartile range
    quartiles

24

In summary, we will couple measures of location with measures of spread when summarizing the distribution of numeric measures. We often report the mean as a measure of location and the standard deviation as a measure of spread; however, when the distribution is skewed or sample sizes are small, we may instead report percentiles including the median (50^{th} percentile) and the interquartile range (difference between the 25^{th} percentile and the 75^{th} percentile).

Descriptive Statistics: Categorical Data

- Measures of distribution
  - Proportion:

    $$\frac{\text{Number of subjects with characteristics}}{\text{Total number subjects}}$$

  - Percentage:

    Proportion * 100%

25

When summarizing the distribution of categorical measures, we report proportions, the number of individuals with a particular characteristic out of the total number of individuals, or the percentage, which is the proportion multiplied by 100%.

Example 1: Duration of Treatment for Postmenopausal Osteoporosis

Table 1. Characteristics of the Study Participants at FLEX Baseline*

Characteristics	Placebo (n = 437)	Alendronate 5 mg/d (n = 329)	Alendronate 10 mg/d (n = 333)	*P* Value
Age, mean (SD), y	73.7 (5.9)	72.7 (5.7)	72.9 (5.5)	.05
Body mass index, mean (SD)†	25.8 (4.3)	25.7 (4.2)	25.9 (4.5)	.73
Race				
White	421 (96.3)	322 (97.9)	327 (98.2)	.22
Other	16 (3.7)	7 (2.1)	6 (1.8)	
General health, self-reported				
Very good or excellent	252 (57.8)	204 (62.2)	210 (63.3)	.08
Good	157 (36.0)	109 (33.2)	100 (30.1)	
Fair or poor	27 (6.2)	12 (3.7)	22 (6.6)	

26

As an example, we can focus on the descriptive summary of the baseline information of the patients who were enrolled in the clinical trial investigating the effect of Alendronate on bone mineral density changes among patients with postmenopausal osteoporosis. We see that the investigators reported means and standard deviation for continuous measures including age and body mass index, while they reported counts and percentages for categorical measures, such as race and self-reported general health status.

From this table, we have an understanding of the sample characteristics and therefore, an idea of the relevant target population (i.e., older women with a mean age of roughly 73 years who were primarily White and of fairly good health with roughly 94% rating their general health as good, very good, or excellent).

We are also interested in reviewing these descriptive statistics to understand whether the randomization was effective in balancing the intervention groups according to baseline characteristics to address concerns regarding confounding of factors by intervention group. In this case, the groups are fairly well balance in terms of baseline characteristics.

Summary

- Descriptive statistics – essential first step in data analysis
- Important to summarize
  - Central tendency
  - Spread or variation
- Choice of summary statistics driven by type of data

In summary, we see that descriptive statistics are an essential first step in data analysis.

It is important to summarize both the central tendency and the spread of the data.

Finally, the choice of summary statistics is driven by the type of data.

In the next module, we will learn more about different types of descriptive plots.