JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section C

Continuous Data: Numerical Summary Measures; Sample Estimates versus Population Measures

Summarizing and Describing Continuous Data

- Measures of the center of data
  - Mean
  - Median

- Measure of data variability
  - Standard deviation (variance)
  - Range

Sample Mean: The Average or Arithmetic Mean

- Add up data, then divide by sample size (n)

- The sample size n is the number of observations (pieces of data)

Mean, Example

- Five systolic blood pressures (mmHg) (n = 5)
  - 120, 80, 90, 110, 95

- Can be represented with math type notation:
  - $x_1 = 120$, $x_2 = 80$, . . . $x_5 = 95$

- The sample mean is easily computed by adding up the five values and dividing by five—in statistical notation the sample mean is frequently represented by a letter with a line over it
  - For example (pronounced "x bar")
  - $\bar{x}$

Mean, Example

- Five systolic blood pressures (mmHg) (n = 5)
  - 120, 80, 90, 110, 95

$$\bar{x} = \frac{120 + 80 + 90 + 110 + 95}{5} = 99 mmHg$$

Notes on Sample Mean

- Generic formula representation

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- In the formula to find the mean, we use the "summation sign"—$\sum$
  - This is just mathematical shorthand for "add up all of the observations"

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots\ldots + x_n$$

Notes on Sample Mean

- Also called *sample average* or *arithmetic mean*
- Sensitive to extreme values
  - One data point could make a great change in sample mean
- Why is it called the *sample* mean?
  - To distinguish it from population mean (will discuss at end of this section)

Sample Median

- The median is the middle number (also called the 50^{th} percentile)
  - Other percentiles can be computed as well, but are not measures of center

80 90 95 110 120

↑

Sample Median

- The sample median is not sensitive to extreme values
  - For example, if 120 became 200, the median would remain the same, but the mean would change to 115

80 90 95 110 200

Sample Median

- If the sample size is an even number

80 90 95 110 120 125

Median

$$\frac{95+110}{2}=102.5mmHg$$

Describing Variability

- Sample variance (s^2)
- Sample standard deviation (s or SD)
- The sample variance is the average of the square of the deviations about the sample mean

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

Describing Variability

- The sample standard deviation is the square root of s^2

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

Describing Variability

- Recall, the five systolic blood pressures (mm Hg) with sample mean ($\bar{x}$) of 99 mmHg

- Five systolic blood pressures (mmHg) (n = 5)
  - 120, 80, 90, 110, 95

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (120 - 99)^2 + (80 - 99)^2 + (90 - 99)^2$$

$$+(110 - 99)^2 + (95 - 99)^2$$

Describing Variability

- Example: n = 5 systolic blood pressures (mm Hg)

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (21)^2 + (-19)^2 + (-9)^2 + (11)^2 + (-4)^2$$

$$= (441) + (361) + (81) + (121) + (16)$$

$$= 1020 \text{mmHg}^2$$

Describing Variability

- Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1020}{4} = 255$$

- Sample standard deviation (s)

$$\sqrt{s^2} = \sqrt{255}$$

s = 15.97 (mmHg)

Notes on *s*

- The bigger *s* is, the more variability there is
- *s* measures the spread about the mean
- *s* can equal 0 only if there is no spread
  - All *n* observations have the same value
- The units of *s* are the same as the units of the data (for example, mm Hg)
- Often abbreviated *SD* or *sd*
- s^2 is the best estimate from the sample of the population variance σ^2; *s* is the best estimate of the population standard deviation σ

Population Versus Sample

- *Sample*: a subset (part) of a larger group (population) from which information is collected to learn about the larger group
  - For example, sample of blood pressures *n* = five 18-year-old male college students in the United States

- *Population*: the entire group for which information is wanted
  - For example, the blood pressure of all 18-year-old male college students in the United States

Random Sampling

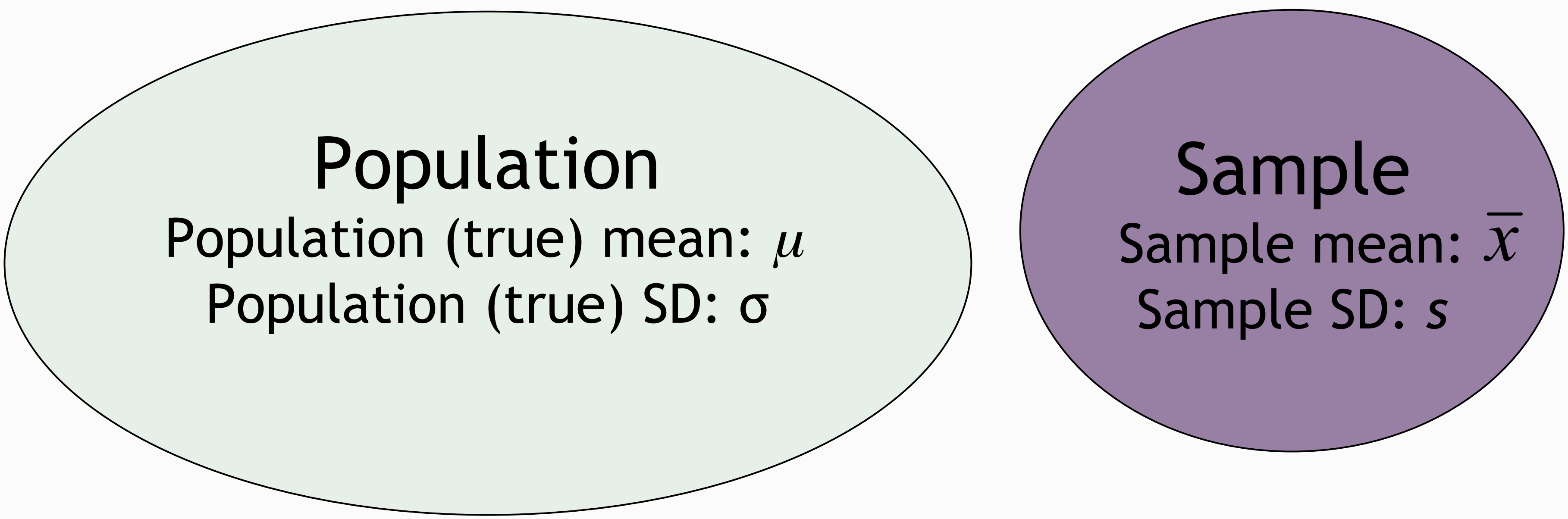
- For studies it is optimal if the sample which provides the data is representative of the population under study
  - Certainly not always possible!

- For this term, we will make this assumption unless otherwise specified

- One way of getting a representative sample: simple random sampling
  - A sampling scheme in which every possible sub-sample of size n from a population is equally likely to be selected
  - How to do it? More detail in second half of term, but think of the “names in a hat” idea

Population Versus Sample

- The sample summary measures (mean, median, sd) are called statistics, and are just estimates of their population (process) counterparts
- Assuming the sample is representative of the population from which it is taken (for example, a randomly drawn sample) these sample estimates should be "good" estimates of true quantities



Population Versus Sample

- For example, we will never know the population mean μ but would like to know it

- We draw a sample from the population

- We calculate the sample mean $\bar{x}$

- How close is $\bar{x}$ to μ?

- Statistical theory allow us to estimate how close $\bar{x}$ is to μ using information computed from the same single sample we use to estimate $\bar{x}$

The Role of Sample Size on Sample Estimates

- Increasing sample size, increases “Goodness” of sample statistics as estimates for their population counterparts
  - Sample mean based on random sample of 1,000 observations is “better estimate” of true (population) mean than sample mean than sample mean based on random sample of 100 from same population
  - Same logic applies to sample standard deviation estimates
  - We’ll define “better estimate” in the third lecture

The Role of Sample Size on Sample Estimates

- Increasing sample size does not dictate how sample estimates from two different representative samples of difference size will compare in value!

- Researcher can not systematically decrease (or increase) value of sample estimates such as mean and standard deviation by taking larger samples!

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

The Role of Sample Size on Sample Estimates

- Extreme values, both larger and smaller, are actually more likely in larger samples
  - The smaller and larger extremes in larger samples they "balance each other out"
  - This "balancing" act tends to keep the mean in a "steady state" as sample size increases—it tends to be about the same

- In addition, "non-extreme" values (values closer to mean) are also more likely in larger samples
  - Hence, sample SD also stays "balanced," i.e., does not systematically increase/decrease with larger samples

SD: Why Do We Divide by n–1 Instead of n?

- We really want to replace $\bar{x}$ with μ in the formula for s^2

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

- Since we don't know μ, we use $\bar{x}$

- But generally, $\sum_{i=1}^{n} (x_i - \bar{x})^2$ tends to be smaller than $\sum_{i=1}^{n} (x_i - \mu)^2$

  – To compensate, we divide by a smaller number: *n-1* instead of *n*

- This will be explored further in an optional component of the third lecture

n-1

- *n-1* is called the *degrees of freedom of the variance* or *SD*

- Why?
  - The sum of the deviations is zero
  - The last deviation can be found once we know the other *n-1*
  - Only *n-1* of the squared deviations can vary freely

- The term *degrees of freedom* arises in other areas of statistics

- It is not always *n-1*, but it is in this case

Why SD as Measure of Variation

- Why note use the range of the data for example?
  - Range = maximum – minimum

- What happens to the sample maximum and minimum as sample size increases?
  - As it turns out, as sample size increases, the maximum tends to increase, and the minimum tends to decrease: Extreme values are more likely with larger samples!
  - This will tend to increase the range systematically with increased sample size