JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Public Health Statistics: Continuous Data Measures

John McGready, PhD
Johns Hopkins University

Learning Objectives

- Upon completion of this lecture, you will be able to:
  - Compute a sample mean and standard deviation
  - Interpret the estimated mean, standard deviation, median, and various percentiles computed for a sample of continuous data measures

Summarizing and Describing Continuous Data

- Measures of the center of data
  - Mean
  - Median (50th percentile)
- Measure of data variability
  - Standard deviation
- Other measures of location
  - Percentiles

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)
- Example: mean, small systolic blood pressure (SBP) dataset

Example: Mean, Small SBP Dataset—2

- Five systolic blood pressures (mmHg), n=5: 120 mmHg, 80 mmHg, 90 mmHg, 110 mmHg, 95 mmHg
- Can be represented with math type notation: x_1= 120, x_2=80,...x_5=95
- The sample mean:

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

The Sample Mean, Generally Speaking—1

- Generic formula representation

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \text{where} \sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots\ldots x_n$$

- 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"

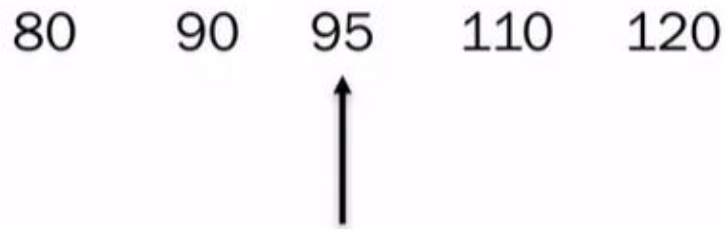
- Also called *sample average* or *arithmetic mean*

- Why is it called the *sample* mean?
  - To distinguish it from population mean (an unknown, unknowable value of interest μ, that can be estimated by $\bar{x}$)

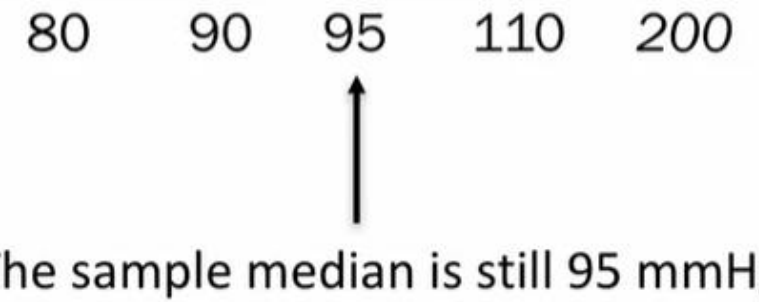
- Sensitive to extreme values (in smaller samples)
  - A change in the value of one data point could make a substantial change in the values of a sample mean

- The median is the middle value in an ordered set of continuous data measures (the median is also called the 50^{th} percentile)

- The median value of the five SBP measurements is 95 mmHg:



- The sample median is not sensitive to the influence of extreme sample values (unlike the sample mean)
  - For example, in the sample of five SBP measurements, if the value 120 was changed to 200, the sample median would remain the same, but the *sample mean would increase from 99 mmHg to 115 mmHg*



Example: The Median, Small SBP Data Set—3

- If the sample size is an even number, then the median is the average of two middle values
- Suppose we add a sixth SBP value to the original SBP values. The (now) six values, in ascending order:

80 90 95 110 120 125

The sample median of these six values is $\frac{95+110}{2} = 102.5$ mmHg

Describing Variability: The Sample Variance

- 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: Sample Standard Deviation (s)

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

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

Example: Standard Deviation, Small SBP Data Set—1

- Systolic blood pressures (mmHg), n=5: 120 mmHg, 80 mmHg, 90 mmHg, 110 mmHg, 95 mmHg. The mean, $\bar{x}$, is 99 mmHg.
- The sample variance computation, numerator:

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 =$$

$$\sum_{i=1}^{5}(x_i - 99)^2 = 21^2 + (-19)^2 + (-9)^2 + 11^2 + (-4)^2 = 1{,}020 \text{ mmHg}^2$$

The Sample Standard Deviation (s), Generally Speaking—1

- The more variability there is in a sample of data, the larger the value of s
- s measures the variability (spread) of the individual sample values around the sample mean
- s can equal 0 only if there is no variability (if all n sample observations have the same value)
- The units of s are the same as the units of the data measurements in the sample (for example, mmHg)
- 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

Example: Standard Deviation, Larger SBP Data Set—1

- SBP measurements taken for a sample of 113 men (n=113)
- The first 50 sample values are shown here:

142	116	137	126	124
123	116	127	115	129
107	103	130	133	116
129	117	131	107	138
114	113	142	120	147
105	122	111	111	129
132	89	134	121	120
128	120	119	112	139
121	124	132	140	120
116	152	123	131	141

Percentiles in Samples With All Unique Values

- Other values that can help us quantify the distribution of continuous data values include the sample percentiles (as estimates of the underlying population percentiles)
- In general, if all sample values are unique, the p^{th} sample percentile is that value in a sample of data such that p percent of the sample values are less than or equal to this value, and $(100-p)$ percent are greater than this value (example: the median is the 50th percentile)
- Percentiles can be computed by hand but are generally done via computer

Example: Percentiles, Larger SBP Data Set—1

- Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population (based on results from a computer)
- The *10th percentile* for these 113 blood pressure measurements is 107 mmHg, meaning that approximately 10% of the men in the sample have SBP ≤ 107 mmHg, and (100−10) = 90% of the men have SBP > 107 mmHg
- The *75th percentile* for these 113 blood pressure measurements is 132 mmHg, meaning that approximately 75% of the men in the sample have SBP ≤ 132 mmHg, and (100−75) = 25% of the men have SBP > 132 mmHg

Summary

- Summary measures that can be computed on a sample of continuous data include the mean, standard deviation, median (50^{th} percentile), and other percentiles

- These sample-based estimates are the best estimates of unknown, underlying population quantities. For example:
  - $\bar{x}$ is the best estimate of the population mean (μ)
  - s is the best estimate of the population standard deviation (σ)

- (Soon) we will discuss how to address the uncertainty in the estimates of certain sample quantities (ex: mean)

Pictures of Data: Continuous Variables

- Histograms and boxplots
  - Means, standard deviations, and percentile values do not tell the whole story of data distributions
  - Differences in shape of the distribution
  - Histograms are a way of displaying the distribution of a set of data by charting the number (or percentage) of observations whose values fall within pre-defined numerical ranges
  - Boxplots are graphics that display key characteristics of a dataset: these are especially nice tools for comparing data from multiple samples visually

Example: Histogram, 113 Systolic Blood Pressures

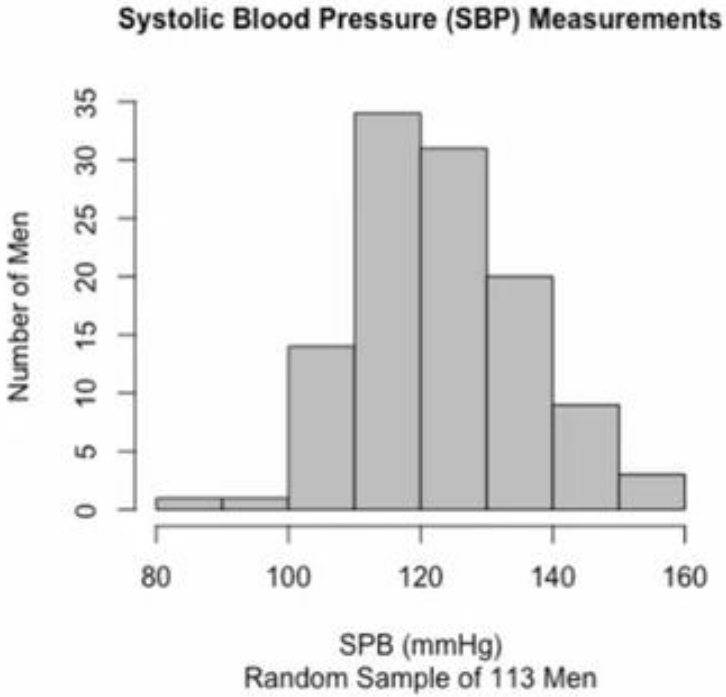
- Data on systolic blood pressures (SBP) from a random clinical sample of 113 men

- A histogram can be created by:
  - Breaking the data (blood pressure) range into bins of equal width
  - Counting the number of the 113 observations whose blood pressure values fall within each bin
  - Plotting the number (or relative frequency) of observations that fall within each bin as a bar graph

Example: Histogram, 113 Systolic Blood Pressures: Number of Observations

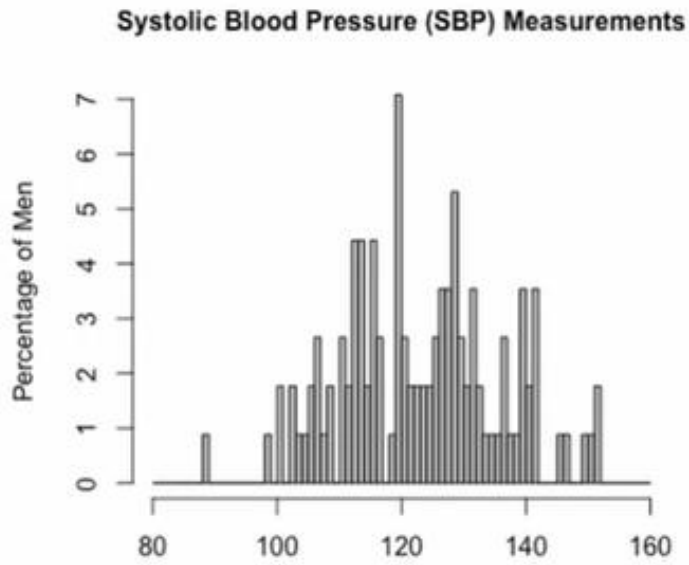
- A basic histogram of these 113 measurements
  - Number of observations on the vertical axis

$$\bar{x} = 123.6 \text{ mmHg}$$
$$s = 12.9 \text{ mmHg}$$
$$\hat{m} = 123.0 \text{ mmHg}$$



Example: Histogram, 113 Systolic Blood Pressures: Narrower Bin Width

- A basic histogram of these 113 measurements
  - Percentage of observations on the vertical axis, narrower bin width



Summary

- Histograms and boxplots are useful visuals tools for characterizing the shape of a data distribution above and beyond the information given by summary statistics
- Relatively common shapes for samples of continuous data measures include symmetric and “bell” shaped, right skewed, left skewed, and uniform

Learning Objectives

- Upon completion of this lecture section, you will be able to:
  - Understand that a random sample taken from a larger population will (imperfectly) mimic the characteristics of the larger population
  - Understand that the distribution of values in a random sample should reflect the distribution of the values in the population from which the sample was taken
  - Understand and explain that increasing sample size does not systematically decrease the value of sample summary statistic estimates
  - Begin to understand that while increasing sample size does not decrease sample summary statistic estimates, the estimates become less variable with larger samples

Summary

- The distribution of sample values of continuous data should (imperfectly) mimic the distribution of the values in the population from which the sample was taken

- With regard to the distribution of sample values and increasing sample size:
  - Will not systematically alter the shape of the sample distribution
  - Will result in a more "filled out" distribution
  - Will not systematically alter the values of the sample statistic
    - The sample statistic estimates will vary from random sample to random sample but will not systematically get larger (or smaller) with increasing sample size
  - Will increase the precision of the summary statistics as estimates of the unknown (population level) true values (more to come shortly)

Learning Objectives

- Upon completion of this lecture section, you will be able to:
  - Suggest graphical approaches to comparing distributions of continuous data between two or more samples
  - Explain why a difference in sample means can be used to quantify, in a single number summary, differences in distributions of continuous data

Motivation: Comparisons

- Frequently, in public health/medicine/science, etc., researchers/practitioners are interested in comparing two (or more) populations via data collected on samples from these populations

- Such comparisons can be used to investigate questions, such as:
  - How does weight change differ between those who are on a low-fat diet compared to those on a low-carbohydrate diet?
  - How do salaries differ between males and females?
  - How do cholesterol levels differ across weight groups?

- While these comparisons can be done visually, it is also useful to have a numerical summary

Motivation: Comparisons

- Frequently, in public health/medicine/science, etc., researchers/practitioners are interested in comparing two (or more) populations via data collected on samples from these populations

- Such comparisons can be used to investigate questions, such as:
  - How does weight change differ between those who are on a low-fat diet compared to those on a low-carbohydrate diet?
  - How do salaries differ between males and females?
  - How do cholesterol levels differ across weight groups?

- While these comparisons can be done visually, it is also useful to have a numerical summary

Motivation: Numerical Summary

- Theoretically, this numerical summary could be many things:
  - Difference in medians
  - Ratio of means
  - Difference in 95^{th} percentiles
  - Ratio of standard deviations
  - Etc.

- However, what is commonly used (for reasons that we will elaborate on shortly) is a difference in sample means
  - When comparing sample distributions, this can be a reasonable measure of the overall differences in these distributions (as an estimate of the underlying difference in the population distributions)

Example: Academic Physician Salaries: Comparing Means between More than Two Groups

- General practice for numerically comparing means between more than two groups
  - Designate one group as the "reference"
  - Report the differences for the other groups compared to this same reference

- For example, if we make "West" the reference region, then the mean differences can be reported as:

$$\bar{x}_{MW} - \bar{x}_W = 198{,}890 - 194{,}474 = \$4{,}416$$
$$\bar{x}_S - \bar{x}_W = 194{,}439 - 194{,}474 = -\$35$$
$$\bar{x}_{NE} - \bar{x}_W = 192{,}152 - 194{,}474 = -\$2{,}322$$