Lecture 2

Continuous Data Measures

Section A:Continous Data: Useful Summary Statistics

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 data

Summarizing and Describing Continuous Data

- Measures of the center of data
  - Mean
  - Median (50^{th} 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 1: Mean, on Small SBP Dataset

- 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 5: in stat notation the sample mean is frequently represented by a letter with a line over it: for example $\bar{x}$ (pronounced 'x bar')

Mean, Example

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

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

7

Notes on Sample Mean

- Generic Formula Representation

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

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"

8

Notes on Sample Mean

- Also called *sample average* or *arithmetic mean*
- Sensitive to extreme values (in smaller samples)
  - One data point could make a great change in sample 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 $\bar{x}$

9

Example 1: Median, on Small SBP Dataset

- The median is the middle number (also called the *50th percentile)*

80 90 95 110 120

10

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 from 99 mmHg to 115 mmHg*

80 90 95 110 200

11

Sample Median

- If the sample size is an even number

80 90 95 110 120 125

Median

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

12

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

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

13

Describing Variability

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

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

14

Example 1: Standard Deviation, on Small SBP Dataset

- Recall, the 5 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

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (120-99)^2 + (80-99)^2 + (90-99)^2 + (110-99)^2 + (95-99)^2$$

15

Describing Variability

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

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

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

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

16

Describing Variability

- Sample variance

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

- Sample standard deviation (s)

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

s = 15.97 ≈ 16 (mmHg)

17

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

18

Notes on s

- The formula

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

Why divide by *n-1* instead of *n*?

19

Example 2: Blood Pressure Data, 113 Men

- Example 2: Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population

The first 50 measurements :

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

20

Example 2 : Blood Pressure Data, 113 Men

- Example 2: Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population

This, if course, is too much data to compute summary statistics by hand. The results from a computer are as follows:

(Estimate of μ): $\bar{x} = 123.6$ mmHg

(Estimate of σ): $s = 12.9$ mmHg

(Estimate of population median): $\hat{m} = 123.0$ mmHg

21

Percentiles

- 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, the p^{th} sample percentile is that value in sample of data such that *p* percent of the sample values are lesser than or equal to this value, and (100-*p*) percent are greater than this value (example: the median is the 50^{th} percentile)

  Percentiles can be computed "by hand", but are generally done via computer

22

Example 2 : Blood Pressure Data, 113 Men

- Example 2: Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population

Based on the results from computer:

The *10^{th} 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 *75^{h} 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

23

Example 2 : Blood Pressure Data, 113 Men

- Example 2: Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population

Some percentiles of these data, based on the results from computer:

Percentile	Sample Estimate
2.5^{th}	100.7 mmHg
25^{th}	114 mmHg
50^{th}	123 mmHg
75^{th}	132 mmHg
97.5^{th}	151.2 mmHg

24

Example 3: Length of Stay Data

- Example 3: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011[1] (12,928 claims)

The first 50 measurements :

21	1	1	1	1
1	4	3	2	7
12	1	2	2	7
5	8	3	1	1
1	2	6	1	1
1	1	4	3	20
1	1	2	3	3
1	6	7	1	4
2	11	10	10	17
6	1	1	5	4

25

Example 3: Length of Stay Data

- Example 3: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011[1] (12,928 claims)

(Estimate of μ): $\bar{x} = 4.3$ days

(Estimate of σ): $s = 4.9$ days

(Estimate of population median): $\hat{m} = 2$ days

26

Example 3: Length of Stay Data

- Example 3: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011[1] (12,928 claims)

Some percentiles of these data, based on the results from computer:

Percentile	Sample Estimate
2.5^{th}	1 day
25^{th}	1 day
50^{th}	2 days
75^{th}	5 days
97.5^{th}	20 days

27

Summary

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

28

Section B: Continuous Data: Visual Displays

29

Learning Objectives

- Upon completion of this lecture section you will be able to:
  - Utilize histograms and boxplots to visualize the distributions of samples of continuous data
  - Identify key summary statistics on the boxplot
  - Name and describe basic characteristics of some common distribution shapes for continuous data

30

Pictures of Data: Continuous Variables

- Histograms and Boxplots
  - Means , standard deviations and percentile values do not tell 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

31

Example 1: Histogram, 113 SBP Measurements

- Data on systolic blood pressures 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 relatively frequency) of observations that fall within each bin as a bar graph

32

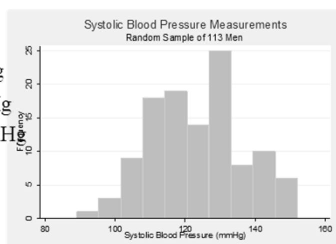
Example 1: Histogram, 113 SBP Measurements

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

$\bar{x} = 123.6$ mmHg

$s = 12.9$ mmHg

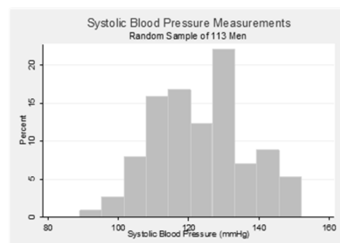
$\hat{m} = 123.0$ mmHg



33

Example 1: Histogram

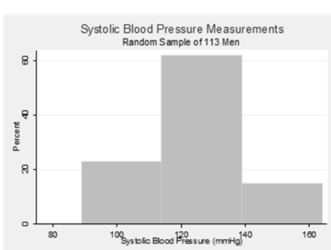
- Data on systolic blood pressures from a random clinical sample of 113 men



34

Example 1: Histogram

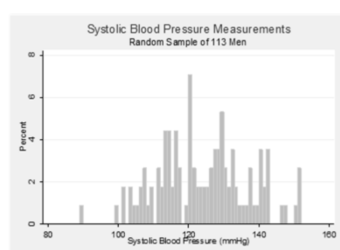
- Data on systolic blood pressures from a random clinical sample of 113 men



35

Example 1: Histogram

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



36

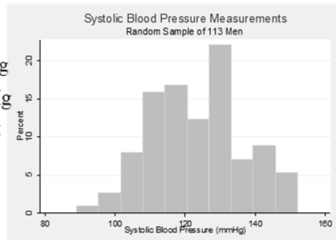
Example 1: Histogram

- How to characterize this distribution?

$\bar{x} = 123.6$ mmHg

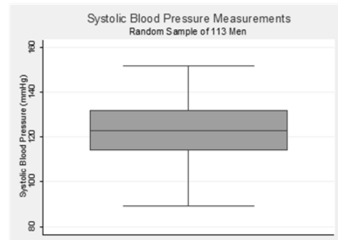
$s = 12.9$ mmHg

$\hat{m} = 123.0$ mmHg



Example 1: Boxplot

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



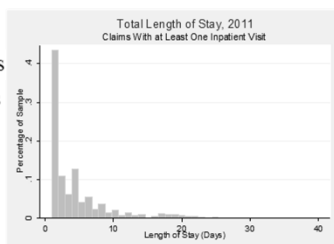
Example 2: Length of Stay Data

- Example 2: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011[1] (12,928 claims)

$\bar{x} = 4.3$ days

$s = 4.9$ days

$\hat{m} = 2$ days



[1] http://inclass.kaggle.com/

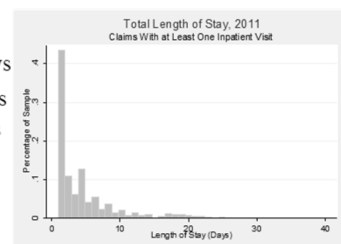
Example 2: Length of Stay Data

- How to characterize this distribution?

$\bar{x} = 4.3$ days

$s = 4.9$ days

$\hat{m} = 2$ days

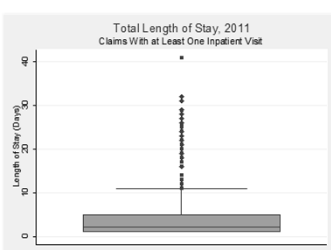


[1] http://inclass.kaggle.com/

Example 2: Length of Stay Data

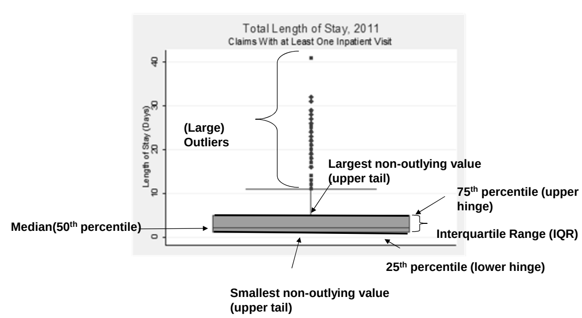
- Example 1: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011[1] (12,928 claims)

/



Example 3: Length of Stay Data

- Anatomy of a boxplot:



Outlier Cutoffs for a Boxplot

- Outlier Cutoffs
  - "Large" outliers: values > upper hinge +1.5*IQR
    - i.e. > 75^{th} percentile +1.5*(75^{th} percentile – 25^{th} percentile)
  - "Small" outliers: values < lower hinge - 1.5*IQR
    - i.e. < 25^{th} percentile -1.5*(75^{th} percentile – 25^{th} percentile)

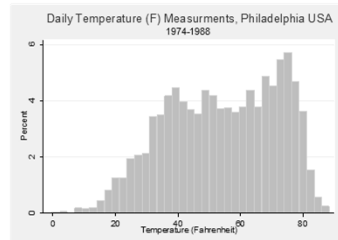
Example 3: Temperature Readings Over Time[2]

- Daily Temperature Measurements (Fahrenheit), Philadelphia

$\bar{x} = 54.3^\circ$

$s = 17.8^\circ$

$\hat{m} = 55.3^\circ$



2 Kelsall JE, Samet J, Zeger SL, Xu J: Air pollution and mortality in Philadelphia, 1974-1988. American Journal of Epidemiology 146(9):750-762, 1997.

44

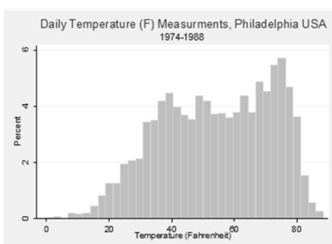
Example 3: Temperature Readings Over Time

- How to characterize this shape?

$\bar{x} = 54.3^\circ$

$s = 17.8^\circ$

$\hat{m} = 55.3^\circ$



45

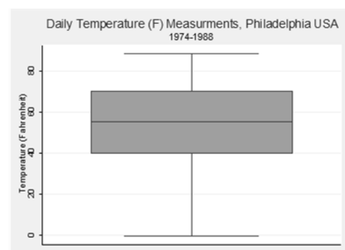
Example 3: Temperature Readings Over Time

- Boxplot

$\bar{x} = 54.3^\circ$

$s = 17.8^\circ$

$\hat{m} = 55.3^\circ$



46

Some "Common" Distribution Shapes

- Symmetric and "bell shaped"

- Right (positively) skewed

47

Some "Common" Distribution Shapes

- Left (negatively) skewed

- Uniform

48

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

49

Section C: Sample Estimates and Sample Size

50

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 sample size does not systematically decrease or decrease sample summary statistic estimates
  - Begin to understand that while sample size does not systematically decrease or decrease sample summary statistic estimates, the estimates become less variable with larger samples

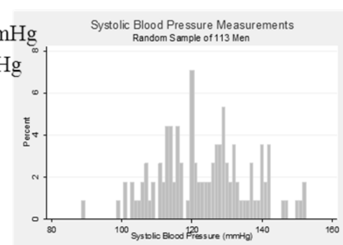
51

Adding More Data to A Random Sample: Example 1

- SBP on 113 Men; Original 113 Measurements

$\bar{x} = 123.6 \text{ mmHg}$

$s = 12.9 \text{ mmHg}$



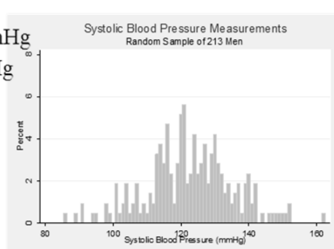
52

Adding More Data to A Random Sample

- SBP on 113 Men; Original 113 Measurements + 100 more measurements, randomly sampled from same population

$\bar{x} = 122.6 \text{ mmHg}$

$s = 13.1 \text{ mmHg}$



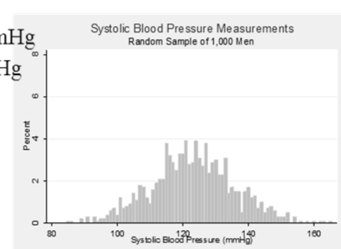
53

Adding More Data to A Random Sample

- SBP on 113 Men; Original 113 Measurements + 887 more measurements, randomly sampled from same population

$\bar{x} = 123.1 \text{ mmHg}$

$s = 12.7 \text{ mmHg}$



54

Summary Statistics and Sample Size

- Sample means from 5 random samples of men, of sizes n=113, 213, and 1,000 respectively

Sample	n=113	n=213	n=1,000
1	124.4	123.5	122.6
2	121.7	123.4	123.6
3	122.6	123.1	122.6
4	123.1	123.4	122.8
5	120.5	123.1	123.1

55

Summary Statistics and Sample Size

- Sample standard deviations from 5 random samples of men, of sizes n=113, 213, and 1,000 respectively

Sample	n=113	n=213	n=1,000
1	12.9	13.1	12.4
2	14.4	11.6	13.3
3	12.0	13.1	13.1
4	14.5	12.9	12.9
5	13.6	13.6	13.1

56

Random Samples of Various Sizes: Example 2

- Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011[1] (12,928 claims)

(Estimate of μ): $\bar{x} = 4.3$ days

(Estimate of σ): $s = 4.9$ days

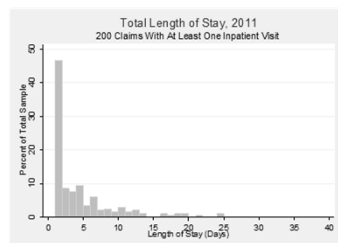
(Estimate of population median): $\hat{m} = 2$ days

57

Random Samples of Various Sizes: Example 2

- Random Sample of 200 Patients, Heritage Health Data

$\bar{x} = 4.1$ days
$s = 4.7$ days

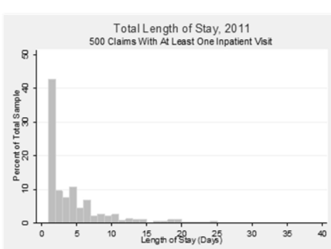


58

Random Samples of Various Sizes: Example 2

- Random Sample of 500 Patients, Heritage Health Data

$\bar{x} = 4.2$ days
$s = 4.6$ days

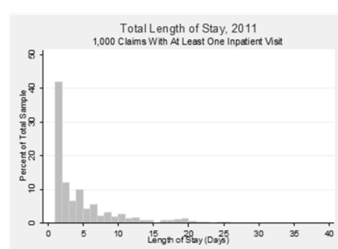


59

Random Samples of Various Sizes: Example 2

- Random Sample of 1,000 Patients, Heritage Health Data

$\bar{x} = 4.3$ days
$s = 4.9$ days



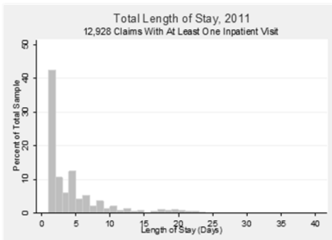
60

Random Samples of Various Sizes: Example 2

- Entire Sample of 12,928 Patients, Heritage Health Data

$\bar{x} = 4.3$ days

$s = 4.8$ days



61

Summary Statistics and Sample Size

- Sample means from 5 random samples of patients, of sizes n=200, 500, and 1,000 respectively

Sample	n=200	n=500	n=1,000
1	4.05	4.19	4.15
2	4.97	4.21	4.12
3	4.06	4.31	4.28
4	4.03	3.99	4.14
5	3.96	4.16	4.24

62

Summary Statistics and Sample Size

- Sample standard deviations from 5 random samples of patients, of sizes n=200, 500, and 1,000 respectively

Sample	n=200	n=500	n=1,000
1	4.68	4.78	5.02
2	5.67	4.84	4.70
3	4.90	5.10	5.10
4	4.50	4.78	4.74
5	4.45	4.65	4.73

63

Random Samples of Various Sizes: Example 3

- Philadelphia Temperature Data (total of 5,471 measurements over 15 years)

(Estimate of μ): $\bar{x} = 54.3^\circ$

(Estimate of σ): $s = 17.8^\circ$

(Estimate of population median): $\hat{m} = 55.3^\circ$

64

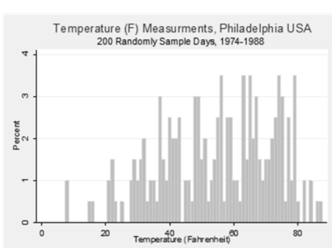
Random Samples of Various Sizes: Example 3

- Philadelphia Temperature Data : random sample of 200 days

$\bar{x} = 54.8^\circ$

$s = 17.5^\circ$

$\hat{m} = 55.6^\circ$



65

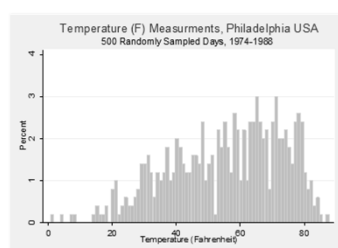
Random Samples of Various Sizes: Example 3

- Philadelphia Temperature Data : random sample of 500 days

$\bar{x} = 54.3^\circ$

$s = 17.7^\circ$

$\hat{m} = 56.5^\circ$



66

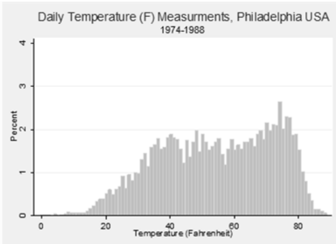
Random Samples of Various Sizes: Example 3

- Philadelphia Temperature Data : entire sample (total of 5,471 measurements over 15 years)

$\bar{x} = 54.5^\circ$

$s = 17.8^\circ$

$\hat{m} = 55.3^\circ$



67

Summary Statistics and Sample Size: Temperature Data

- Sample means from 5 random samples of days, of sizes n=200, 500, and 1,000 respectively

Sample	n=200	n=500	n=1,000
1	56.1	54.3	53.9
2	52.3	53.2	55.0
3	54.0	54.4	54.7
4	52.7	54.3	54.3
5	55.0	53.7	53.7

68

Summary Statistics and Sample Size

- Sample standard deviations from 5 random samples of days, of sizes n=200, 500, and 1,000 respectively

Sample	n=200	n=500	n=1,000
1	18.1	18.1	17.9
2	18.4	17.8	18.0
3	17.8	17.1	17.7
4	17.3	16.9	17.8
5	17.4	18.1	18.0

69

Summary Statistics and Sample Size

- Sample medians from 5 random samples of days, of sizes n=200, 500, and 1,000 respectively

Sample	n=200	n=500	n=1,000
1	56.2	56.5	54.3
2	53.5	52.9	55.5
3	55.6	55.5	56.6
4	51.8	54.0	55.0
5	55.5	54.5	54.1

70

Wrap Up

- The distribution of sample values of continuous value should (imperfectly) mimic the distribution of the values in the population from which the sample was taken
- With regards to the a distribution of sample values, increased sample size:
  - Will not systematically alter the shape of the sample distribution
  - Will result in a more “filled out” distribution

71

Wrap Up

- With regards to sample summary statistics, increased sample size:
  - Will not systematically alter the values of the sample statistic
    - The sample statistic estimates with 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..)

72

Section D: Comparing Distributions of Continuous Data

73

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

74

Motivation

- 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?

75

Motivation

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

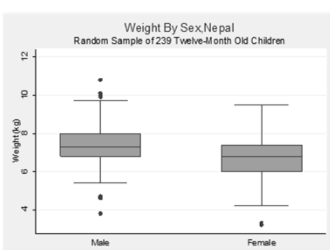
- Theoretically, this numerical summary could be many things:
  - Difference in medians
  - Ratio of means
  - Difference in 95th percentiles
  - Ratio of standard deviations
  - Etc..

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)

76

Example 1:Weight By Sex, Nepali Children

- Boxplot



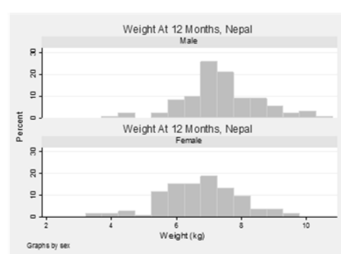
77

Example 1:Weight By Sex, Nepali Children

- Histograms

$\bar{x}_{male} = 7.4$ kg

$\bar{x}_{female} = 6.7$ kg



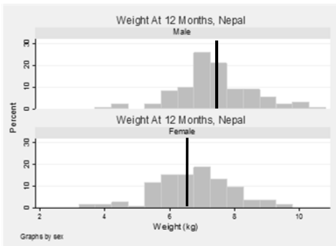
78

Example 1:Weight By Sex, Nepali Children

- Histograms

$\bar{x}_{male} = 7.4\text{ kg}$

$\bar{x}_{female} = 6.7\text{ kg}$



79

Example 1:Weight By Sex, Nepali Children

- Common numerical comparison: difference in means

$$\bar{x}_{male} - \bar{x}_{female} = 7.4\text{ kg} - 6.7\text{ kg} = 0.7\text{ kg}$$

"On average, male children weigh more than female children by 0.7 kg"

80

Example 1:Weight By Sex, Nepali Children

- Common numerical comparison: difference in means: direction is arbitrary, but important to know!

$$\bar{x}_{female} - \bar{x}_{male} = 6.7\text{ kg} - 7.4\text{ kg} = -0.7\text{ kg}$$

"On average, female children weigh less than male children by 0.7 kg"

which is same as stating:

"On average, male children weigh more than female children by 0.7 kg"

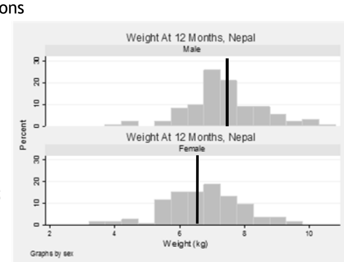
81

Example 1:Weight By Sex, Nepali Children

- Difference in means: why does this characterize difference in distributions

$\bar{x}_{male} = 7.4\text{ kg}$

$\bar{x}_{female} = 6.7\text{ kg}$



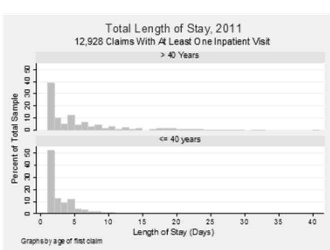
82

Example 2:Length of Stay, By Age of First Claim

- Histograms

$\bar{x}_{>40\text{ years}} = 4.9\text{ days}$

$\bar{x}_{\leq 40\text{ years}} = 2.7\text{ days}$



83

Example 2: Length of Stay, By Age of First Claim

- Mean Difference

$$\bar{x}_{>40\text{ years}} - \bar{x}_{\leq 40\text{ years}} = 4.9\text{ days} - 2.7\text{ days} = 2.2\text{ days}$$

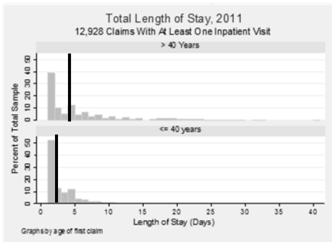
84

Example 2:Length of Stay, By Age of First Claim

- Histograms

$\bar{x}_{>40 \text{ years}} = 4.9$ days

$\bar{x}_{\leq 40 \text{ years}} = 2.7$ days



85

Example 3: Menu Labeling and Calorie Intake[1]

- From Abstract

Objectives. We assessed the impact of restaurant menu calorie labels on food choices and intake.

Methods. Participants in a study dinner (n=303) were randomly assigned to either (1) a menu without calorie labels (no calorie labels), (2) a menu with calorie labels (calorie labels), or (3) a menu with calorie labels and a label stating the recommended daily caloric intake for an average adult (calorie labels plus information). Food choices and intake during and after the study dinner were measured.

Results. Participants in both calorie label conditions ordered fewer calories than those in the no calorie labels condition. When calorie label conditions were combined, that group consumed 14% fewer calories than the no calorie labels group. Individuals in the calorie labels condition consumed more calories after the study dinner than those in both other conditions. When calories consumed during and after the study dinner were combined, participants in the calorie labels plus information group consumed an average of 250 fewer calories than those in the other groups.

Conclusions. Calorie labels on restaurant menus impacted food choices and intake; adding a recommended daily caloric requirement label increased this effect, suggesting menu label legislation should require such a label. Future research should evaluate menu labeling's impact on children's food choices and consumption. (*Am J Public Health.* 2010;100:312–318. doi:10.2105/AJPH.2009.160226)

[1] Roberto C, et al. Evaluating the Impact of Menu Labeling on Food Choices and Intake. *American Journal of Public Health* (2010); 100(2); 313-318.

86

Example 3: Menu Labeling and Calorie Intake[1]

- From Abstract

Objectives. We assessed the impact of restaurant menu calorie labels on food choices and intake.

Methods. Participants in a study dinner (n=303) were randomly assigned to either (1) a menu without calorie labels (no calorie labels), (2) a menu with calorie labels (calorie labels), or (3) a menu with calorie labels and a label stating the recommended daily caloric intake for an average adult (calorie labels plus information). Food choices and intake during and after the study dinner were measured.

Results. Participants in both calorie label conditions ordered fewer calories than those in the no calorie labels condition. When calorie label conditions were combined, that group consumed 14% fewer calories than the no calorie labels group. Individuals in the calorie labels condition consumed more calories after the study dinner than those in both other conditions. When calories consumed during and after the study dinner were combined, participants in the calorie labels plus information group consumed an average of 250 fewer calories than those in the other groups.

Conclusions. Calorie labels on restaurant menus impacted food choices and intake; adding a recommended daily caloric requirement label increased this effect, suggesting menu label legislation should require such a label. Future research should evaluate menu labeling's impact on children's food choices and consumption. (*Am J Public Health.* 2010;100:312–318. doi:10.2105/AJPH.2009.160226)

[1] Roberto C, et al. Evaluating the Impact of Menu Labeling on Food Choices and Intake. *American Journal of Public Health* (2010); 100(2); 313-318.

87

Example 3: Menu Labeling and Calorie Intake[1]

- Article Graphic

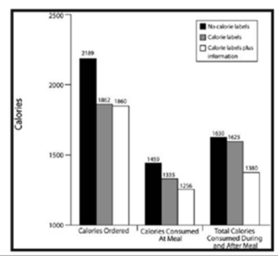


FIGURE 2—Calories ordered and consumed, by menu type: New Haven, CT, 2007–2008.

[1] Roberto C, et al. Evaluating the Impact of Menu Labeling on Food Choices and Intake. *American Journal of Public Health* (2010); 100(2); 313-318.

88

Example 4: Academic Physician Salaries[2]

- From abstract

Results The mean salary within our cohort was $167 669 (95% CI, $158 417-$176 922) for women and $200 433 (95% CI, $194 249-$206 617) for men. Male gender was associated with higher salary (+$13 399; *P*=.001) even after adjustment in the final model for specialty, academic rank, leadership positions, publications, and research time. Peters-Belson analysis (use of coefficients derived from regression model for men applied to women) indicated that the expected mean salary for women, if they retained their other measured characteristics but their gender was male, would be $12 194 higher than observed.

Conclusion Gender differences in salary exist in this select, homogeneous cohort of mid-career academic physicians, even after adjustment for differences in specialty, institutional characteristics, academic productivity, academic rank, work hours, and other factors.

JAMA. 2012;307(22):2410-2417 www.jama.com

[2] Jagsi R, et al. Gender Differences in the Salaries of Physician Researchers. *Journal of the American Medical Association* (2012); 307(22); 2410-2417.

89

Example 4: Academic Physician Salaries

- From abstract

Results The mean salary within our cohort was $167 669 (95% CI, $158 417-$176 922) for women and $200 433 (95% CI, $194 249-$206 617) for men. Male gender was associated with higher salary (+$13 399; *P*=.001) even after adjustment in the final model for specialty, academic rank, leadership positions, publications, and research time. Peters-Belson analysis (use of coefficients derived from regression model for men applied to women) indicated that the expected mean salary for women, if they retained their other measured characteristics but their gender was male, would be $12 194 higher than observed.

Conclusion Gender differences in salary exist in this select, homogeneous cohort of mid-career academic physicians, even after adjustment for differences in specialty, institutional characteristics, academic productivity, academic rank, work hours, and other factors.

JAMA. 2012;307(22):2410-2417 www.jama.com

90

Example 4: Academic Physician Salaries

- Article Table

Table 2. Bivariable Associations Between Salary and Measured Characteristics (continued)

Characteristics	Salary Estimate, $ (95% CI)	*P* Value
Current institution National Institutes of Health funding rank group		
First	177 590 (166 584 to 188 596)	<.001
Second	184 771 (173 608 to 195 934)	
Third	188 104 (177 991 to 198 217)	
Fourth	206 621 (196 953 to 216 288)	
Current institution region		
West	194 474 (183 876 to 205 073)	.09
Midwest	198 890 (186 426 to 211 354)	
South	194 439 (182 105 to 206 773)	
Northeast	192 152 (173 878 to 190 425)	
Still at K institution at time of survey		
Yes	186 985 (180 681 to 193 289)	.06
No	197 734 (188 330 to 207 139)	
K award type		
K08	192 786 (185 885 to 199 686)	.36
K23	186 943 (178 885 to 195 011)	
Years since K award		
7	184 672 (174 741 to 194 603)	.30
8	194 387 (186 740 to 202 035)	
9	188 966 (178 506 to 199 427)	
Funding institute tier		
First	201 873 (192 185 to 211 561)	<.001
Second	181 350 (173 359 to 189 341)	
Third	191 978 (182 184 to 201 771)	
Attained R01 grant or >$1 million in grant funding		
Yes	187 749 (180 670 to 194 828)	.35
No	193 440 (185 634 to 201 247)	
Publications, per 1-publication increase	871 (660 to 1081)	<.001
Leadership position		
Yes	255 726 (242 720 to 268 733)	<.001
No	179 448 (174 146 to 184 750)	
Work hours, per 1-h increase	1525 (1122 to 1927)	<.001
Research time, per 1% increase	−937 (−1128 to −746)	<.001

91

Example 4: Academic Physician Salaries

- Article Table

Table 2. Bivariable Associations Between Salary and Measured Characteristics (continued)

Characteristics	Salary Estimate, $ (95% CI)	*P* Value
Current institution National Institutes of Health funding rank group		
First	177 590 (166 584 to 188 596)	<.001
Second	184 771 (173 608 to 195 934)	
Third	188 104 (177 991 to 198 217)	
Fourth	206 621 (196 953 to 216 288)	
Current institution region		
West	194 474 (183 876 to 205 073)	.09
Midwest	198 890 (186 426 to 211 354)	
South	194 439 (182 105 to 206 773)	
Northeast	192 152 (173 878 to 190 425)	
Still at K institution at time of survey		

92

Example 4: Academic Physician Salaries

- Mean differences that can be reported, for example:

93

Summary

- While the distributions of continuous data can be compared between samples in many ways, some key approaches include:
  - Visual comparisons, such as side-by-side boxplots
  - Numerical comparisons, mainly the mean difference between any group of two samples

94