MHA Statistics Course - Day 1

Pacific University

Dr. Chester Ismay
Email: chester@pacificu.edu

2017/01/27

Slides available at http://bit.ly/mha17-day1

Workshop Agenda - Friday

- Part 1: Introduction
  - Overview of Statistics & Key Terms

These slides available at http://bit.ly/mha17-day1

Workshop Agenda - Friday

- Part 1: Introduction
  - Overview of Statistics & Key Terms
- Part 2: Data Visualization
  - Tables
  - Types of plots and best plots for types of data

These slides available at http://bit.ly/mha17-day1

Workshop Agenda - Friday

- Part 1: Introduction
  - Overview of Statistics & Key Terms
- Part 2: Data Visualization
  - Tables
  - Types of plots and best plots for types of data
- Part 3: Data Summaries
  - Averages
  - Variability
  - Correlation

These slides available at http://bit.ly/mha17-day1

Workshop Agenda - Saturday

- Part 4: Simulating randomness
  - Random sampling & common terms
  - Sampling distribution
  - Simulation

These slides available at http://bit.ly/mha17-day1

Workshop Agenda - Saturday

- Part 4: Simulating randomness
  - Random sampling & common terms
  - Sampling distribution
  - Simulation
- Part 5: Inference
  - Hypothesis testing
  - Confidence intervals

These slides available at http://bit.ly/mha17-day1

Workshop Agenda - Saturday

- Part 4: Simulating randomness
  - Random sampling & common terms
  - Sampling distribution
  - Simulation
- Part 5: Inference
  - Hypothesis testing
  - Confidence intervals
- Part 6: Workshop Review

These slides available at http://bit.ly/mha17-day1

Learning objectives

By completion of the workshop you should understand how to

1. organize data

These slides available at http://bit.ly/mha17-day1

Learning objectives

By completion of the workshop you should understand how to

1. organize data
2. visualize data

These slides available at http://bit.ly/mha17-day1

Learning objectives

By completion of the workshop you should understand how to

1. organize data
2. visualize data
3. summarize data

These slides available at http://bit.ly/mha17-day1

Learning objectives

By completion of the workshop you should understand how to

1. organize data
2. visualize data
3. summarize data
4. simulate sampling of data

These slides available at http://bit.ly/mha17-day1

Learning objectives

By completion of the workshop you should understand how to

1. organize data
2. visualize data
3. summarize data
4. simulate sampling of data
5. infer conclusions about data

These slides available at http://bit.ly/mha17-day1

Learning objectives

By completion of the workshop you should understand how to

1. organize data
2. visualize data
3. summarize data
4. simulate sampling of data
5. infer conclusions about data
6. interpret results about data

These slides available at http://bit.ly/mha17-day1

Learning objectives

By completion of the workshop you should understand how to

1. organize data
2. visualize data
3. summarize data
4. simulate sampling of data
5. infer conclusions about data
6. interpret results about data
7. tell a story effectively with data

These slides available at http://bit.ly/mha17-day1

Ice breaker

- Do you have experience with statistics? Explain, e.g., specific courses, comfort level with the subject.
- What do you expect from this workshop?

These slides available at http://bit.ly/mha17-day1

Arthur Benjamin - Teach Statistics before Calculus!

- Reflect on Arthur Benjamin's TED Talk. What is your response to the question, "why study statistics?"
- Frame your response within the context of healthcare administration.

These slides available at http://bit.ly/mha17-day1

Arthur Benjamin - Teach Statistics before Calculus!

- Reflect on Arthur Benjamin's TED Talk. What is your response to the question, "why study statistics?"
- Frame your response within the context of healthcare administration.
- How does statistics apply to your everyday life? To the world as a whole?

These slides available at http://bit.ly/mha17-day1

First steps

Frequently the first thing you should do when given a dataset is to

- identify the observational unit,
- specify the variables, and
- give the types of variables you are presented with.

These slides available at http://bit.ly/mha17-day1

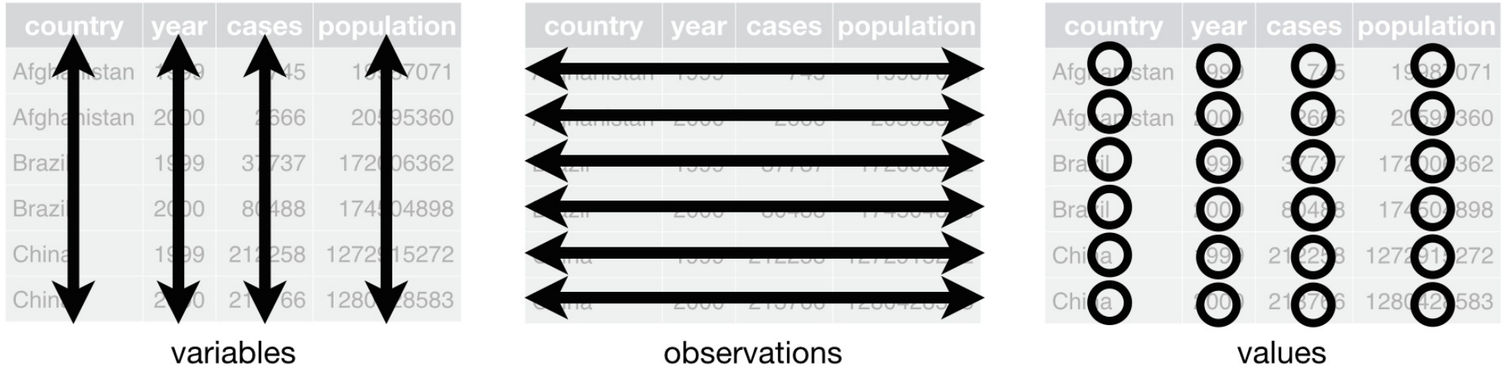
Organizing data

Table 1. Example of a line listing for acute Hepatitis A*

				Signs/Symptoms						Labs		Demographics	
Case #	Report Date	Onset	Physician Diagnosis	N	V	A	F	D	J	HAIgM	Other	Sex	Age
1	10/12/02	10/5/02	Hepatitis A	1	1	1	1	1	1	1	Low SGOT	M	37
2	10/12/02	10/4/02	Hepatitis A	1	0	1	1	1	1	1	Low Alt	M	62
3	10/13/02	10/4/02	Hepatitis A	1	0	1	1	1	1	1	Low SGOT	M	38
4	10/13/02	10/9/02	NA	0	0	1	0	?	0	NA	NA	F	44
5	10/15/02		Hepatitis A	1	1	1	1	1	0	1	Hbs/Ag-	M	17
6	10/16/02	10/6/02	Hepatitis A	0	0	1	1	1	1	1	SGOT=24	F	43

- identify the observational unit
- give the names of the variables
- specify the types of the variables (logical, numerical, categorical)

These slides available at http://bit.ly/mha17-day1



These slides available at http://bit.ly/mha17-day1

So what is Statistics?

Statistics is a set of tools and techniques used for describing, organizing, and interpreting information or data

These slides available at http://bit.ly/mha17-day1

These slides available at http://bit.ly/mha17-day1

Why are tables an effective way to show data?

These slides available at http://bit.ly/mha17-day1

Why are tables an effective way to show data?

- Help clarify exactly what information is most pertinent
- Exact values can be presented easily
- Often simpler to produce than a graphic
- Summarize frequencies and percentages well

These slides available at http://bit.ly/mha17-day1

What are the most important features of this table?

Table 11.1. Community Medical Center analysis showing patients discharged 12/1/20XX

Name	Age	Clinical Service	Length of Stay
Smith	5	Surgical	1
Valdez	22	Obstetrical	1
Chu	26	Obstetrical	2
MacDuff	18	Obstetrical	3
Johnson	10	Surgical	7
O'Brien	80	Surgical	8
Lewandowski	35	Surgical	11
Jones	52	Medical	15
Shultz	69	Medical	37
Martini	49	Medical	42

Source: Community Medical Center.

These slides available at http://bit.ly/mha17-day1

What are the most important features of this table?

- Title: The title must explain as simply as possible what is contained in the table. The title should answer the questions:

What are the most important features of this table?

- Title: The title must explain as simply as possible what is contained in the table. The title should answer the questions:
  - What are the data? Are these percentages; frequencies?
  - Who? Who is the table about? For instance, are these males or females; a certain service; a type of disease?
  - Where? For example, is this your hospital; the United States; or your state?
  - When? What is the time period?

What are the most important features of this table?

- Stub heading: The title or heading of the first column
- Column headings: The headings or titles for the columns
- Stubs: The categories (the left-hand column of a table)
- Cells: The information formed by intersecting columns and rows
- Source footnote: The source for any factual data should be identified in a footnote.

Table 11.2. The essential components of a table

Title			
Stub Heading	**Column Heading**	**Column Heading**	**Column Heading**
Stub	Cell	Cell	Cell
Stub	Cell	Cell	Cell
Stub	Cell	Cell	Cell

These slides available at http://bit.ly/mha17-day1

Frequency Distribution Tables

A frequency distribution shows the values that a variable can take and the number of obser-vations associated with each value.

Example: The Utilization Review Committee is interested in knowing the admission days for patients in your hospital. To construct a frequency distribution, you would list the days of the week and then enter the observations or number of patients admitted on the corresponding day of the week.

These slides available at http://bit.ly/mha17-day1

Table 11.4. Report illustrating sample frequency distribution table

Sample Frequency Distribution for Admission Day
June 20XX

Day of the Week	No. of Patients Admitted
Sunday	20
Monday	29
Tuesday	28
Wednesday	12
Thursday	13
Friday	22
Saturday	8
Total	**132**

These slides available at http://bit.ly/mha17-day1

Table 11.5. Report illustrating sample frequency distribution with proportion

Sample Frequency Distribution for Admission Day June 20XX		
Day of the Week	**No. of Patients Admitted**	**Proportion**
Sunday	20	0.15
Monday	29	0.22
Tuesday	28	0.21
Wednesday	12	0.09
Thursday	13	0.10
Friday	22	0.17
Saturday	8	0.06
Total	**132**	**1.00**

These slides available at http://bit.ly/mha17-day1

Rules for building tables

These slides available at http://bit.ly/mha17-day1

Rules for building tables

- Ranges of values should not overlap (1-10, 10-20, etc. for ages is bad)
- Try not to use fewer than four or more than ten categories
- Groupings should be well defined
- Groupings should cover equal ranges (as much as possible)

These slides available at http://bit.ly/mha17-day1

BREAK TIME

The table below lists the patients seen last month at Community Hospital with their age and cholesterol reading. Create a table using common age categories and these ranges for cholesterol:

Desirable ≤ 199
Borderline High 200–239
High ≥ 240

Age	Cholesterol	Age	Cholesterol	Age	Cholesterol	Age	Cholesterol
14	118	44	138	38	165	56	185
80	139	47	204	18	142	20	200
42	187	48	236	62	139	45	241
37	201	25	186	37	202	63	175
23	107	56	201	32	207	70	188
24	109	47	198	17	157	42	239
67	132	20	210	55	238	55	175
55	235	43	248	13	134	61	168
52	185	50	137	44	239	53	173

These slides available at http://bit.ly/mha17-day1

The following table shows a frequency distribution of patients with colon cancer treated at Community Hospital. Compute the proportion of patients in each category.

Community Hospital
Ages of Patients with Colon Cancer
Annual Statistics, 20XX

Age	No. of Patients	Proportion
≤ 30	3	
31–40	12	
41–50	18	
51–60	60	
61–70	65	
71+	48	

These slides available at http://bit.ly/mha17-day1

Unfortunately, tables are... BORING

Hans Rosling's 200 Countries, 200 Years, 4 Minutes

These slides available at http://bit.ly/mha17-day1

What are general guidelines for creating effective plots?

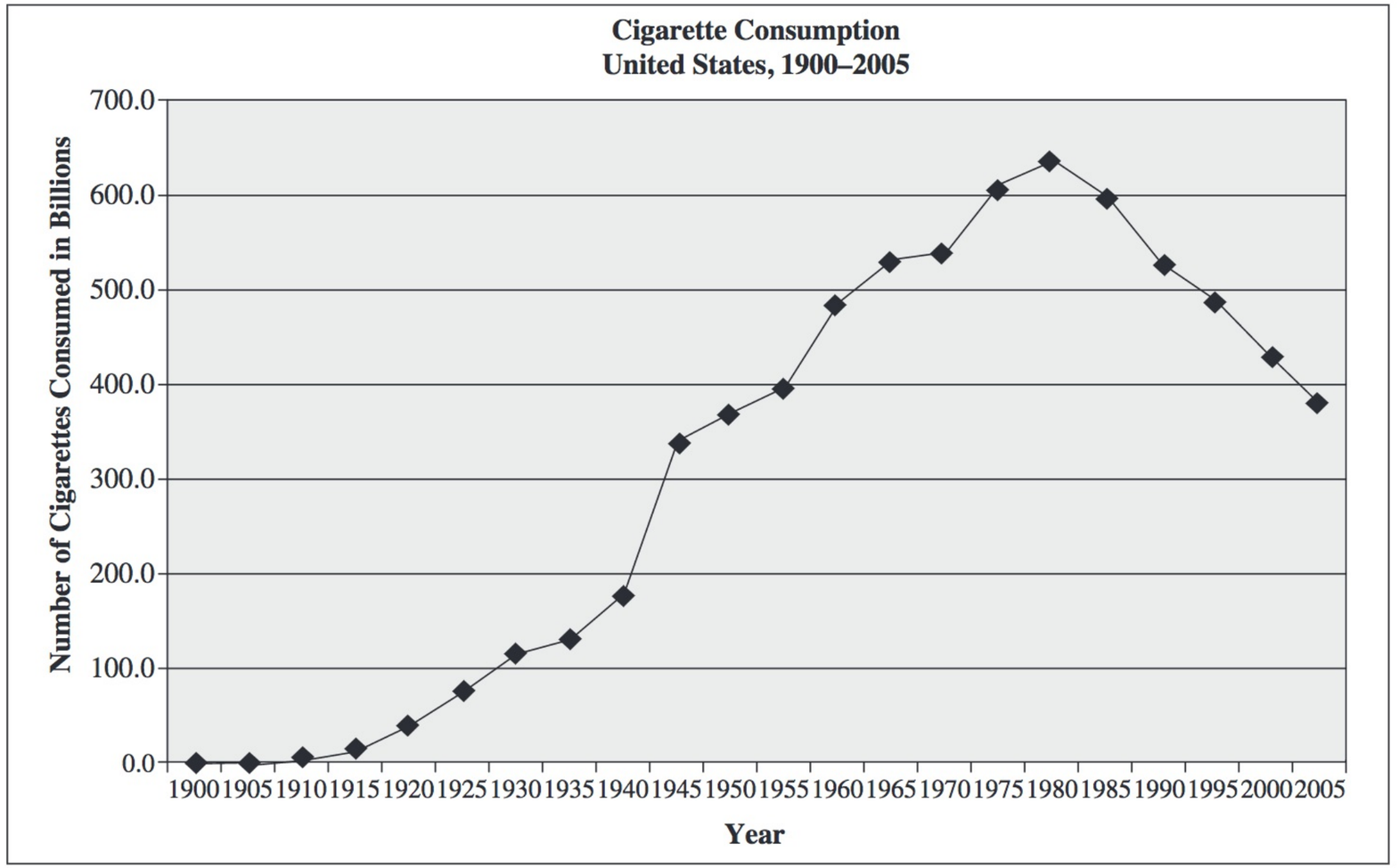
These slides available at http://bit.ly/mha17-day1

What are general guidelines for creating effective plots?

- The title must relate to what the graph is displaying.
- When a variable has multiple levels included on the same graph, each should be identified by using a legend or key.
- The plot should be oriented in the way we read
- Axes are labeled clearly (with units)

These slides available at http://bit.ly/mha17-day1

Figure 11.1. Scale captions on a graph



These slides available at http://bit.ly/mha17-day1

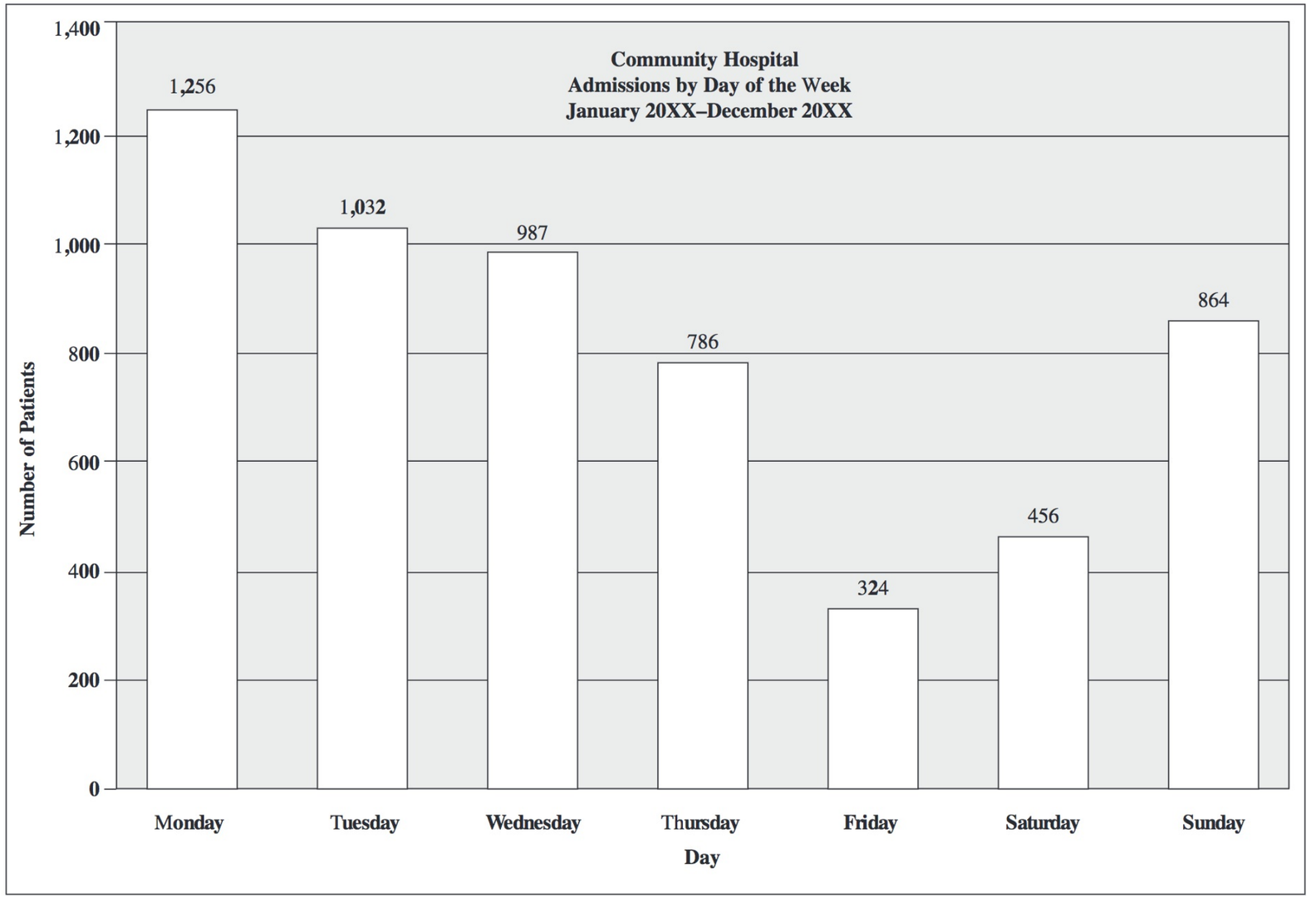
Most common types of plots

- Bar graphs
- Histograms
- Boxplots
- Scatter plots
- Line graphs

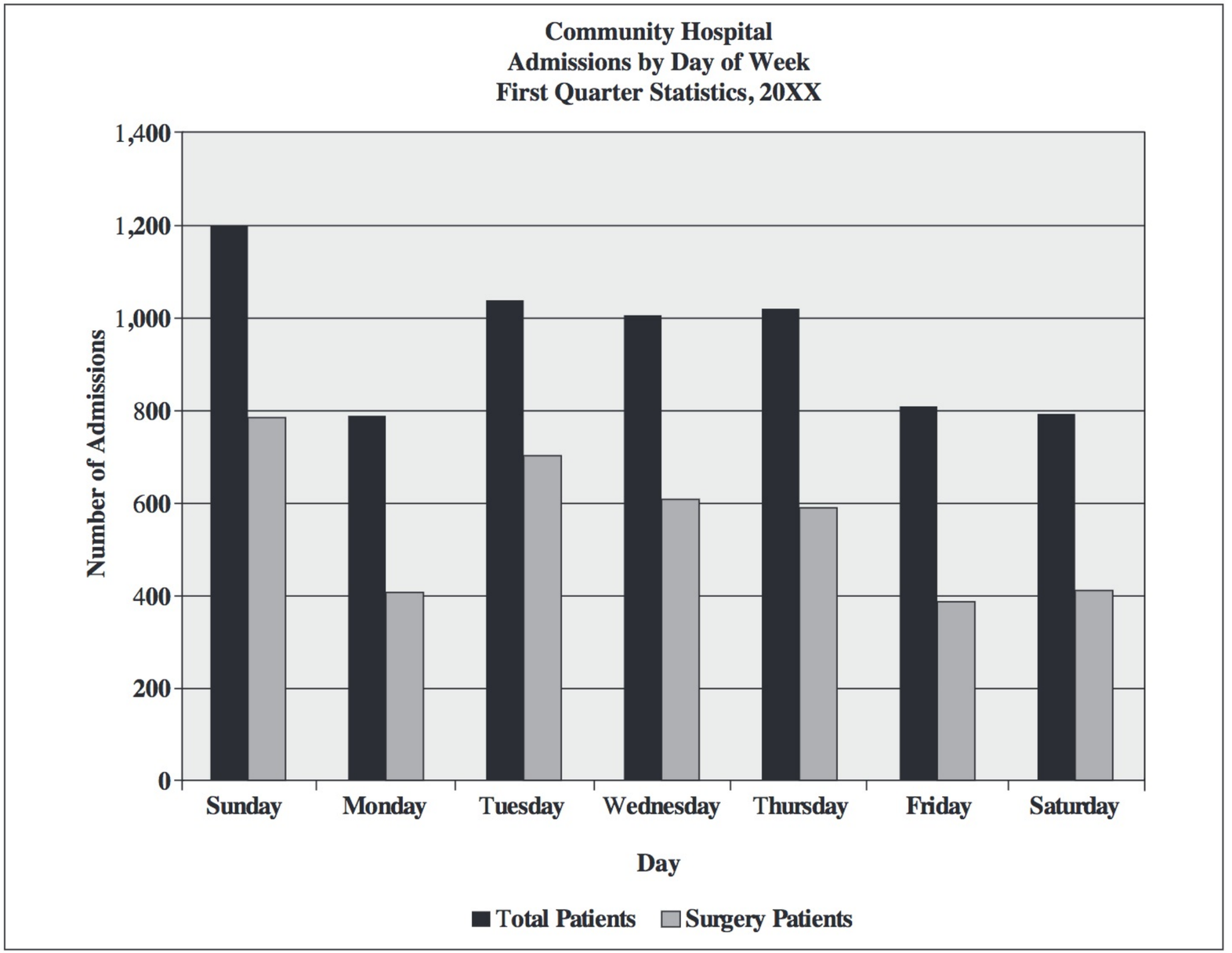
These slides available at http://bit.ly/mha17-day1

Bar graphs

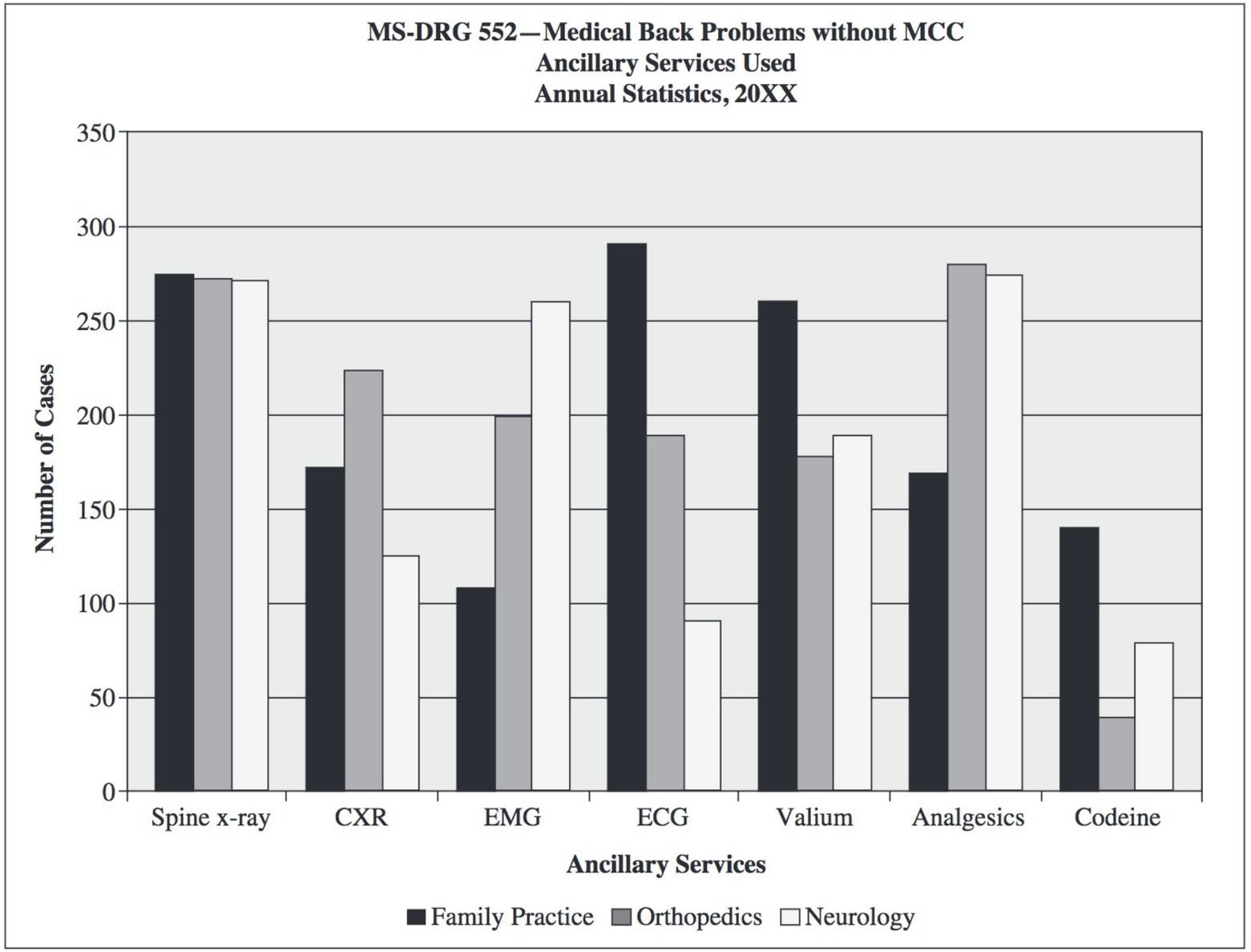
- Appropriate for displaying categorical data
- Usually display either the count or the percentage of each level of one or more categorical variables



These slides available at http://bit.ly/mha17-day1



These slides available at http://bit.ly/mha17-day1



These slides available at http://bit.ly/mha17-day1

Histograms

- Used to display frequency distributions for continuous numerical data

Histograms

- Used to display frequency distributions for continuous numerical data
- created from frequency distribution tables

These slides available at http://bit.ly/mha17-day1

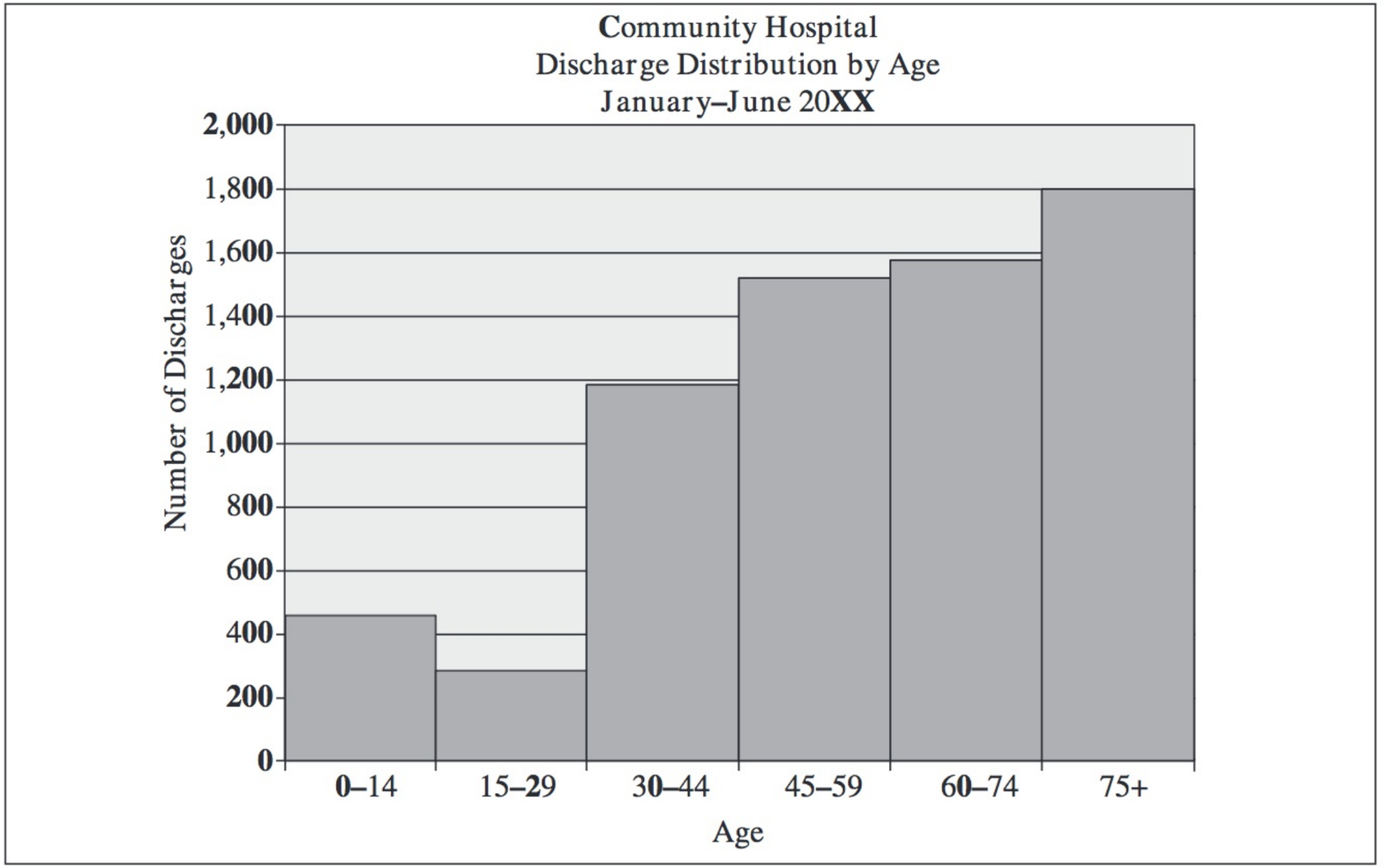
Histograms

- Used to display frequency distributions for continuous numerical data
- created from frequency distribution tables
- look similar to bar graphs except that all the bars in a histogram are touching because they show the continuous nature of the distribution.

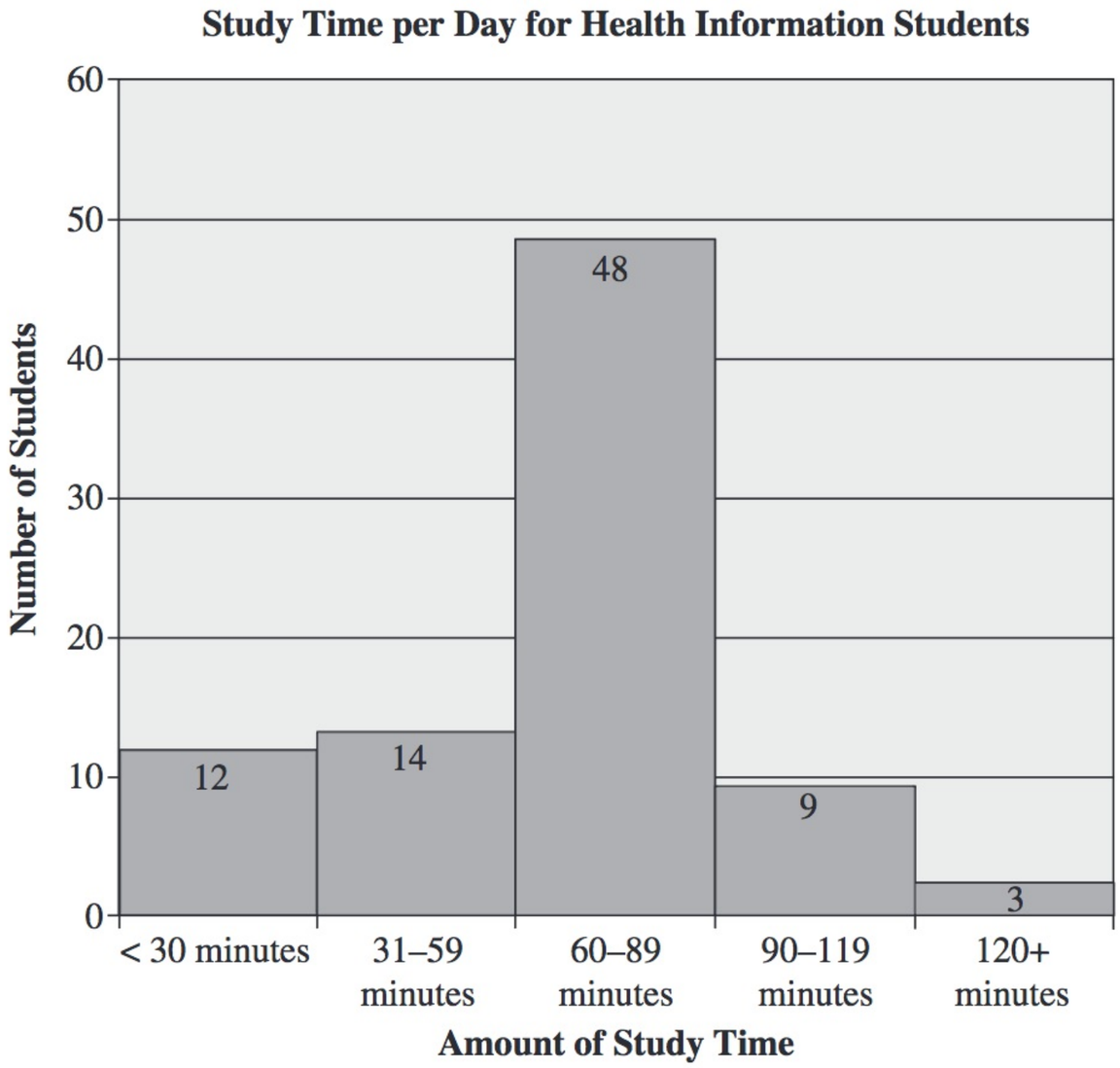
Histograms

- Used to display frequency distributions for continuous numerical data
- created from frequency distribution tables
- look similar to bar graphs except that all the bars in a histogram are touching because they show the continuous nature of the distribution.
- bars should be of equal width

These slides available at http://bit.ly/mha17-day1



These slides available at http://bit.ly/mha17-day1

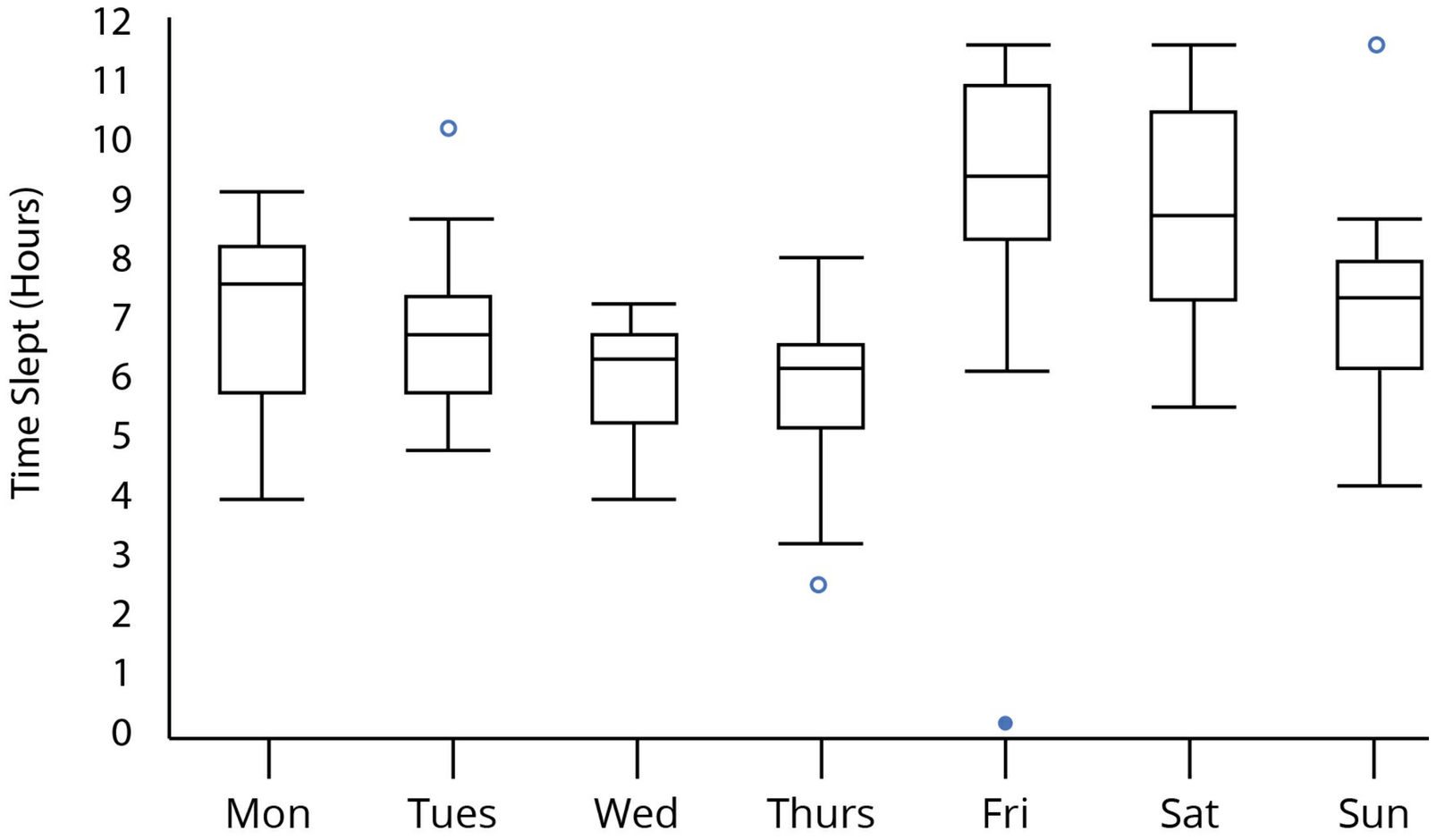


These slides available at http://bit.ly/mha17-day1

Boxplot

- Displays the distribution of a continuous variable based on quantiles
- Can be used to compare the distribution of a continuous variables across the groups of a categorical variable

Time slept versus day of the week for a college professor

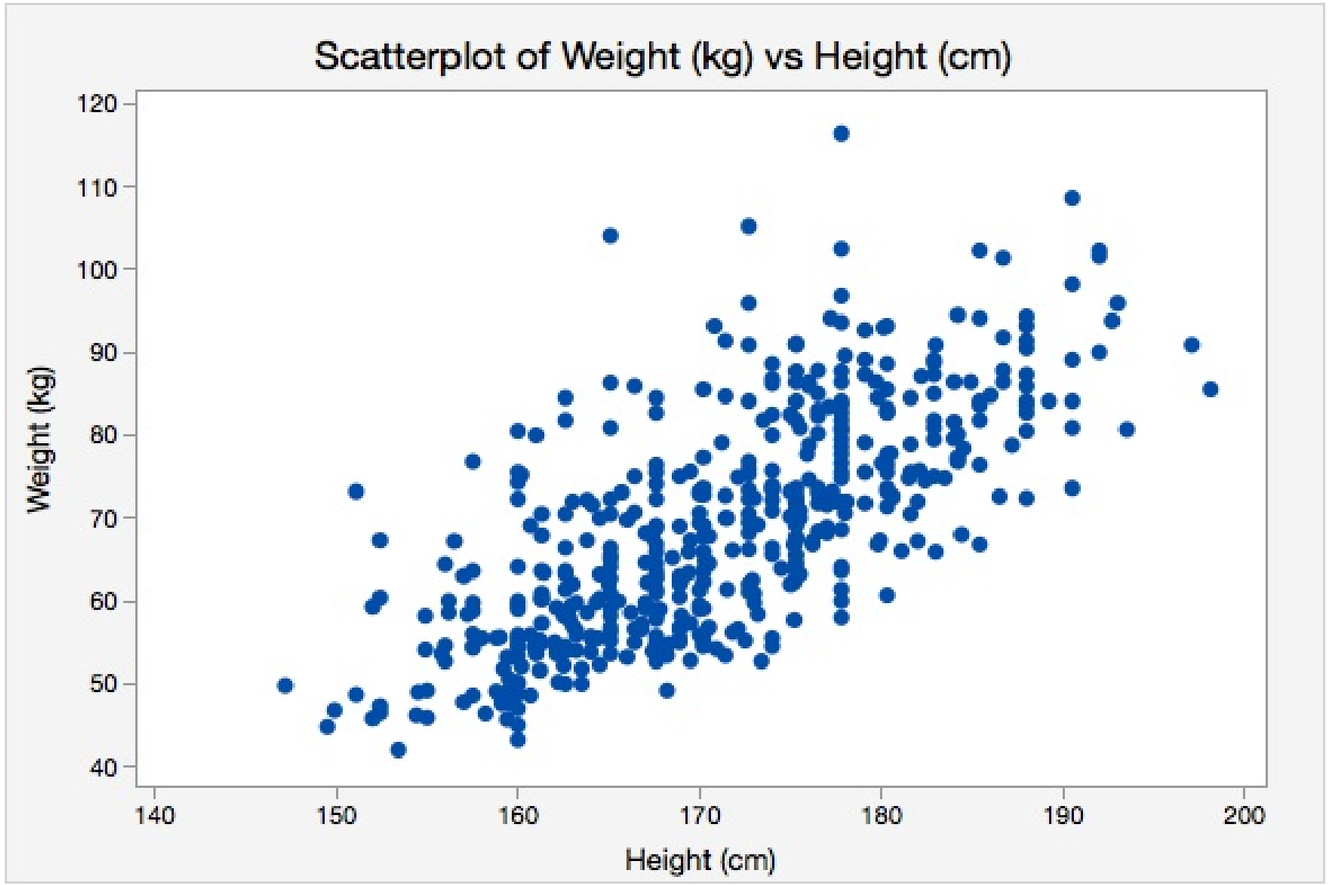


These slides available at http://bit.ly/mha17-day1

Scatter plots

- Shows the relationship between two numerical variables
- Helps to identify whether a linear correlation exists between the variables or not

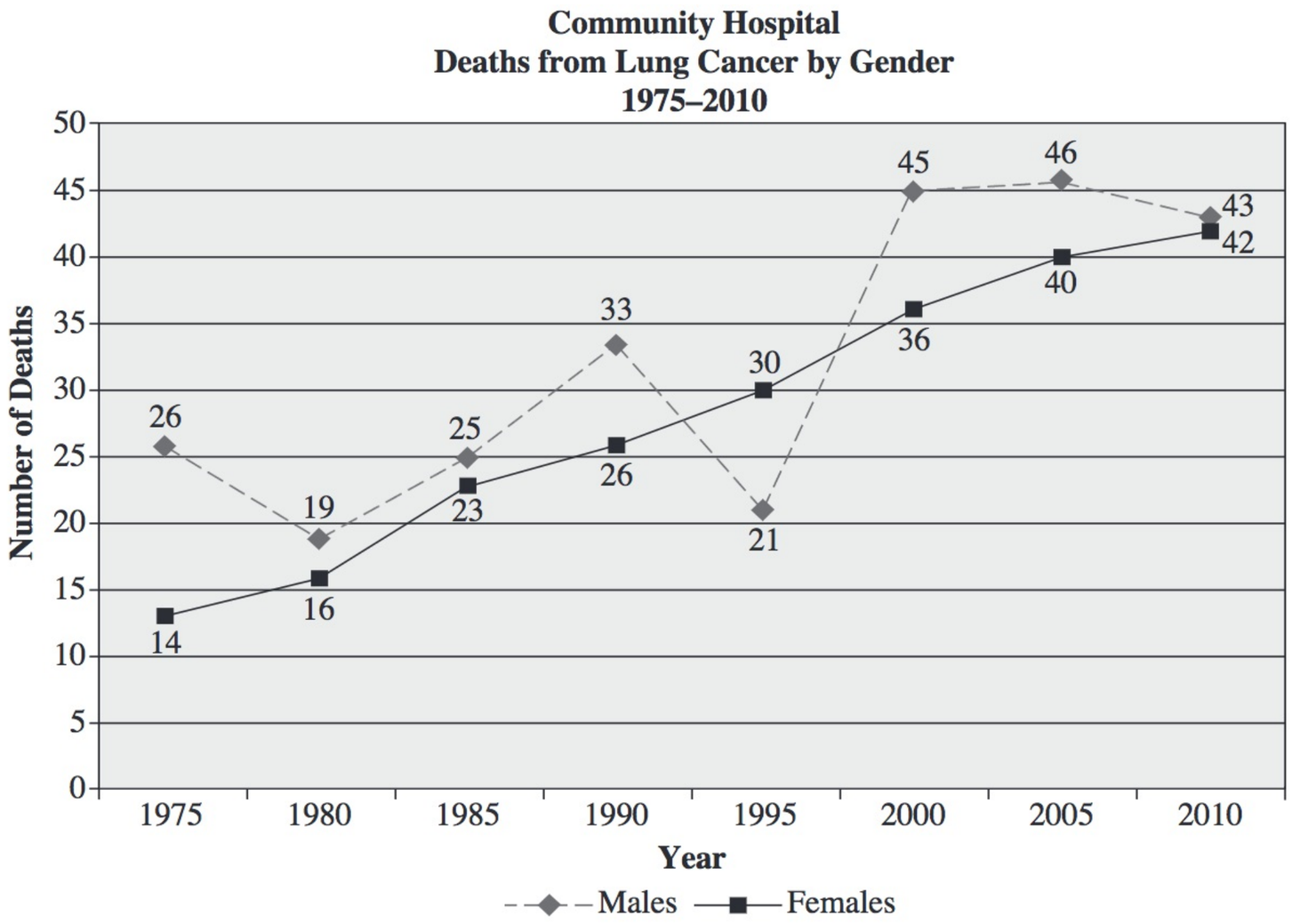
Weight versus height for British adults



These slides available at http://bit.ly/mha17-day1

Line graphs

- Most frequently used when time is the predictor variable
- There must be only one measurement (response value) for each value of time given for each group



These slides available at http://bit.ly/mha17-day1

BREAK TIME

Indicate whether a table or a graph is the preferred method of presentation in the following situations:

a. Distribution by site, sex, race, and time period of all cancers in your healthcare facility
b. Survival trends over time by sex for lung cancer
c. Display of prostate cancer stage of disease for a presentation at a professional conference
d. Detailed treatment distribution of breast cancer for a physician on the staff of your hospital

These slides available at http://bit.ly/mha17-day1

What type of plot is most appropriate for this data?

In September 2011, Community Hospital discharged 120 patients.

- 82 patients were discharged home
- 10 patients were discharged home with follow-up home health
- 6 patients died
- 11 patients were transferred to a skilled nursing facility
- 3 patients were transferred to another acute care facility
- 8 patients were transferred to a rehabilitation hospital

These slides available at http://bit.ly/mha17-day1

What type of plot is most appropriate?

University Hospital
Cancer Registry Data
Lung and Bronchus Cancer
by Year and Gender

Year	Males	Females
1994	112	48
1995	121	47
1996	130	49
1997	123	50
1998	121	54
1999	131	55
2000	150	60
2001	155	65
2002	173	72
2003	171	75
2004	172	80
2005	171	83
2006	170	87
2007	168	93
2008	165	120
2009	169	118
2010	175	121
2011	172	120

These slides available at http://bit.ly/mha17-day1

Shifting gears

Descriptive Statistics

- Used to organize and **describe** the characteristics of a data set

Inferential Statistics

- Used to make **inferences** from a sample to a population

Descriptive statistics include any treatment of data that does not involve generalizations, predictions, or estimations. Once generalizations, estimations, and predictions are involved, the analysis is **inferential**.

These slides available at http://bit.ly/mha17-day1

Summarizing data

Measures of Central Tendency

- The AVERAGE is a single score that best represents a set of scores
- Another name for AVERAGE is *measure of central tendency*

MEASURES OF CENTRAL TENDENCY		
MEAN	MEDIAN	MODE

These slides available at http://bit.ly/mha17-day1

Computing the Mean

1. List the entire set of values in one or more columns
2. Compute the sum or total of all the values
3. Divide the total or sum by the number of values

Example: Use these data to calculate the average wage rate of the employees.

Employee	Hours	Wage
1	20	$15.00
2	40	$18.00
3	35	$12.00
4	30	$20.00
5	37	$14.00
6	25	$23.00

Sum

$15 + $18 + $12 + $20 + $14 + $23 = $102

Divide by *n*

$102/6 = $17 average wage rate

These slides available at http://bit.ly/mha17-day1

Computing the Mean

mean = sum of all scores/number of scores

$$\bar{X} = \frac{\sum X}{n}$$

Symbol	Meaning
$\bar{X}$	(X Bar) is the mean value of the group of scores
Σ	(sigma) tells you to add together whatever follows it
X	is each individual score in the group
n	is the sample size

These slides available at http://bit.ly/mha17-day1

Things to remember about the mean

- The sample mean is a measure of central tendency that best represents the population mean (n = sample size, N = population size)

- The mean is the centermost point where all values on one side of the mean are equal in weight to all values on the other side of the mean

- Mean is VERY sensitive to extreme scores (outliers) that can "skew" or distort findings

These slides available at http://bit.ly/mha17-day1

Median

- Point at which 50% of scores fall below it and 50% fall above it
- Because the median cares about the number of cases, extreme scores (i.e., outliers) do not impact it

Median

- Point at which 50% of scores fall below it and 50% fall above it
- Because the median cares about the number of cases, extreme scores (i.e., outliers) do not impact it
- Steps in finding the median
  1. List the values, in order, either from highest to lowest or lowest to highest.

Median

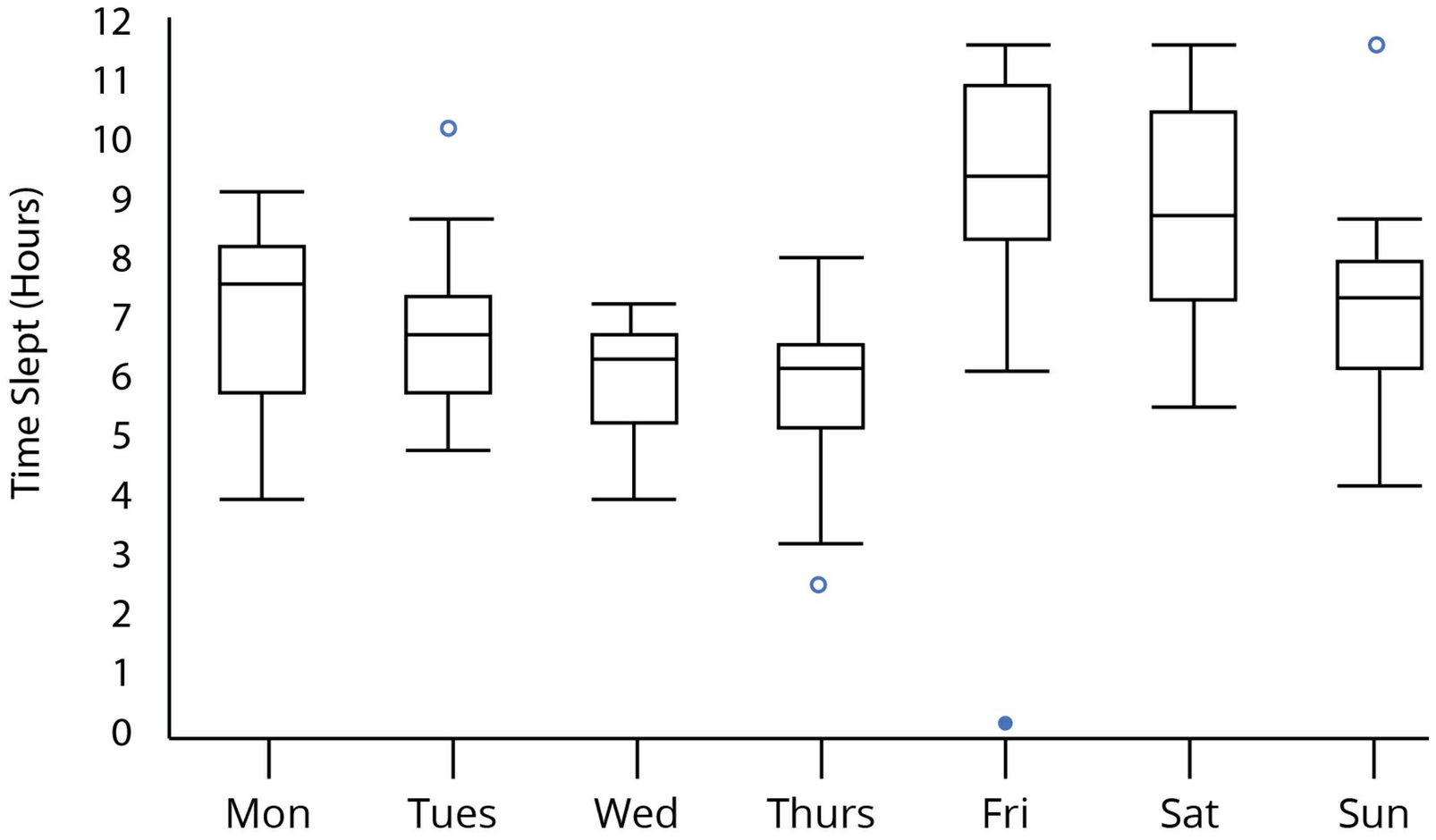
- Point at which 50% of scores fall below it and 50% fall above it
- Because the median cares about the number of cases, extreme scores (i.e., outliers) do not impact it
- Steps in finding the median
  1. List the values, in order, either from highest to lowest or lowest to highest.
  2. Find the middle-most score. That's the median.
     - What if there are two middle scores?
     - What if the two middle scores are the same?

These slides available at http://bit.ly/mha17-day1

Percentiles

- Divide data in 100 equal parts
- Percentile ranks are used to define percent of cases equal to and below a certain point on a distribution
  - 75th percentile – means that the score received is at or above 75% of all other scores in the distribution
  - Median is always at the 50th percentile
- Relates to what is plotted on a boxplot

Time slept versus day of the week for a college professor



These slides available at http://bit.ly/mha17-day1

Mode

- Mode = most frequently occurring score
- When two values occur the same number of times, we have a bimodal distribution
- Steps to finding the mode

1. List all values in the distribution
2. Tally the number of times each value occurs
3. The value occurring the most is the mode

These slides available at http://bit.ly/mha17-day1

When to use what...

- Use mode when the data are qualitative, categorical, or nominal (e.g., eye color, political party) and values can only fit into one category (i.e., mutually-exclusive)
- Use median when you have extreme (numerical) scores
- Use mean when the data does not include extreme scores (i.e., outliers) and are not categorical

Importance of variability

- Variability reflects how scores differ from one another
- Usually measured relative to mean or median

These slides available at http://bit.ly/mha17-day1

Importance of variability

- Variability reflects how scores differ from one another
- Usually measured relative to mean or median
- Measures of variability (aka spread, dispersion)
  - Range
  - Standard deviation

Importance of variability

- Variability reflects how scores differ from one another
- Usually measured relative to mean or median
- Measures of variability (aka spread, dispersion)
  - Range
  - Standard deviation
- Note: You typically report the average and the variability together to describe a distribution

These slides available at http://bit.ly/mha17-day1

Computing the range

- Range is the most general estimate of variability
- $range = h - l$

  where h - highest numerical value and
  l is lowest numerical value

Computing standard deviation

1. Calculate the mean

These slides available at http://bit.ly/mha17-day1

Computing standard deviation

1. Calculate the mean
2. Subtract the mean from each observation, and square the difference

Computing standard deviation

1. Calculate the mean
2. Subtract the mean from each observation, and square the difference
3. Sum the squared difference

Computing standard deviation

1. Calculate the mean
2. Subtract the mean from each observation, and square the difference
3. Sum the squared difference
4. Divide the sum of the squared differences by $n - 1$

These slides available at http://bit.ly/mha17-day1

Computing standard deviation

1. Calculate the mean
2. Subtract the mean from each observation, and square the difference
3. Sum the squared difference
4. Divide the sum of the squared differences by $n - 1$
5. Take the square root of the value obtained in step 4 (the result is the standard deviation)

Computing standard deviation

- Commonly denoted as SD or s for the sample standard deviation. σ for the population standard deviation.

$$s = \sqrt{\frac{\sum(X - \bar{X})}{n - 1}}$$

- Only works well for data that does not have outliers. Why?

Practice

Incubation periods for hepatitis A:

27, 31, 15, 30, 22 days

Calculate the standard deviation

These slides available at http://bit.ly/mha17-day1

Practice

- Sum the squared differences

  Sum = 4 + 36 + 100 + 25 + 9 =174

- Divide the sum of the squared differences by (n – 1). This is the variance.

  Variance = 174 / (5 – 1) = 174 / 4 = 43.5 days squared

- Take the square root of the variance. The result is the standard deviation.

  Standard deviation = square root of 43.5 = 6.6 days

These slides available at http://bit.ly/mha17-day1

Practice

- Sum the squared differences

  Sum = 4 + 36 + 100 + 25 + 9 =174

- Divide the sum of the squared differences by ($n - 1$). This is the variance.

  Variance = 174 / (5 – 1) = 174 / 4 = 43.5 days squared

- Take the square root of the variance. The result is the standard deviation.

  Standard deviation = square root of 43.5 = 6.6 days

How do we interpret this result?

These slides available at http://bit.ly/mha17-day1

Things to remember

- Standard deviation is computed as the average distance from the mean
- The larger the standard deviation the more spread out the values are
- Like the mean, the standard deviation is sensitive to extreme scores
- If $s = 0$, then there is no variability among scores and the scores are identical in value

These slides available at http://bit.ly/mha17-day1

Analogy for Mean and Standard Deviation

- The site for a new school was selected because it provides a central location. An alternative site on the west side of town was considered, but rejected because it would require extensive busing for students living on the east side. The location represents the mean; just as the school is located at the center of town, the mean is located in the center of a distribution of scores.

Analogy for Mean and Standard Deviation

- For each student, it is possible to measure the distance between home and the new school. Some students live only a few blocks from the new school and others lives as much as 3 miles away. Let's say the average distance a student must travel to school was calculated to be 0.8 miles. The average distance from school represents standard deviation, which measures the standard distance from an individual score to the mean.

FINAL BREAK

Use the information below to find your percentile. Your score is 86.

Test Scores out of 100 Points

95	97
99	74
84	91
65	94
54	89
35	88
86	56
77	96
76	27
100	75
92	93

These slides available at http://bit.ly/mha17-day1

More practice

1. Fourteen patients have the following LOS: 1, 4, 4, 2, 5, 16, 3, 3, 1, 6, 4, 5, 7, and 2. Compute the mean, median, and mode.
2. A student's 10 scores on 10-point class quizzes include a 6, a 7, a 4, five 9s, an 8, and a 10. The student claims that her average grade on quizzes is 9 because most of her scores are 9s. Is this correct? Explain.
3. Last month, 10 patients between the ages of 11 and 13 were seen in their pediatrician's clinic. Their heights were recorded as 50, 56, 59, 51, 53, 51, 50, 52, 54 and 51 inches. Determine the mean, median, and mode.

These slides available at http://bit.ly/mha17-day1

More practice

The following sample report from a cancer registry shows the SDs of weights for 20 males with adenocarcinoma of the rectum. Validate the calculations used in the report.

Weights of Males with Adenocarcinoma of Rectum

	Patient	Weight lbs. (X)	$(X - \overline{X})$	$(X - \overline{X})^2$
	1	142	–30	900
	2	148	–24	576
	3	151	–21	441
	4	155	–17	289
	5	155	–17	289
	6	158	–14	196
	7	164	–8	64
	8	165	–7	49
	9	170	–2	4
	10	173	1	1
	11	175	3	9
	12	175	3	9
	13	175	3	9
	14	183	11	121
	15	185	13	169
	16	186	14	196
	17	189	17	289
	18	193	21	441
	19	198	26	676
	20	200	28	784
Total	**20**	**3,440**	**0**	**5,512**

*SD = 17.0; $s^2 = \frac{5,512}{19}$ = 290.1; mean = 172; and $N - 1 = 19$

These slides available at http://bit.ly/mha17-day1

Resources

- ModernDive: An Introduction to Statistical and Data Sciences via R

- Horton, L. A. (2012). Calculating and Reporting Healthcare Statistics. Chicago, Ill: AHIMA Press.

These slides available at http://bit.ly/mha17-day1

Thanks!

- Slides created via the R package `xaringan` by Yihui Xie.
- Email me at chester@pacificu.edu
- Source code for these slides is on GitHub

These slides available at http://bit.ly/mha17-day1