Lecture 4: 4.1 Finish Categorical; 4.2 Begin Quantitative Variables (Displays, Begin Summaries)

- Details about Categorical Displays, Summaries
- Quan: Summarize with Shape, Center, Spread
- Displays: Stemplots, Histograms
- Five Number Summary, Outliers, Boxplots

1

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing
    - Single variables: 1 cat. (Lectures 3-4), 1 quantitative
    - Relationships between 2 variables
  - Probability
  - Statistical Inference

2

Two or More Possible Values

Looking Ahead: *In Probability and Inference, most categorical variables discussed have just **two** possibilities.*

Still, we often summarize and display categorical data with **more than two** possibilities.

3

Example: *Proportions in Three Categories*

- **Background**: Student wondered if she should resist changing answers in multiple choice tests. "Ask Marilyn" replied:
  - 50% of changes go from wrong to right
  - 25% of changes go from right to wrong
  - 25% of changes go from wrong to wrong
- **Question:** How to display information?
- **Response:**

5

Definition

- **Bar graph:** shows counts, percents, or proportions in various categories (marked on horizontal axis) with bars of corresponding heights

6

Example: *Bar Graph*

- **Background**: Instructor can survey students to find proportion in each program.
- **Questions:**
  - How can we display the information?
  - What should we look for in the display?
- **Responses:**
  - Look for _____

8

Overlapping Categories

If more than two categorical variables are considered at once, we must note the possibility that categories overlap.

Looking Ahead: *In Probability, we will need to distinguish between situations where categories do and do not overlap.*

9

Example: *Overlapping Categories*

- **Background**: Report by ResumeDoctor.com on over 160,000 resumes:
  - 13% said applicant had "communication skills"
  - 7% said applicant was a "team player"
- **Question:** Can we conclude that 20% claimed communication skills or team player?
- **Response:**

11

Processing Raw Categorical Data

Small categorical data sets are easily handled without software.

12

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing
    - Single variables: 1 cat. (Lectures 3-4), 1 quantitative
    - Relationships between 2 variables
  - Probability
  - Statistical Inference

16

Example: *Proportion from Raw Data*

- **Background**: Harvard study claimed 44% of college students are binge drinkers. Agree on survey design and have students self-report: on one occasion in past month, alcoholic drinks more than 5 (males) or 4 (females)? *Or use these data:*

yes	no	yes	no	no	yes
no	yes	yes	no	yes	no
yes	yes	no	no	yes	yes
yes	no	yes	yes	no	no
yes	no	yes	yes	yes	yes
no	no	yes	no	yes	no
no	yes	no	no	yes	no
no	no	no	no	yes	yes
yes	no	no	no	no	no
no	no	no	no	no	no
no	yes	yes	no	no	yes

- **Question:** Are data consistent with claim of 44%?
- **Response:**

15

Example: *Issues to Consider*

- **Background**: Intro stat students' earnings (in $1000s) previous year: 12, 3, 7, 1, … [survey was anonymous].
- **Questions:**
  - What population do the data represent?
  - Were responses unbiased?
- **Responses:**
  - All students at that university, if sample was representative in terms of _____
  - Probably unbiased because _____

Looking Back: *These are data production issues.*

18

Example: *More Issues to Consider*

- **Background**: Intro stat students' earnings (in $1000s) previous year: 12, 3, 7, 1, … [survey was anonymous].
- **Questions:**
  - How do we summarize the data?
  - Sample average was $3776. Can we conclude population average was less than $5000?
- **Responses:**
  - Mean and other summaries are the focus of this part.
  -

Looking Ahead: *This is an inference question, to be addressed in Part Four.*

Definitions

- **Distribution**: tells all possible values of a variable and how frequently they occur

Summarize distribution of a quantitative variable by telling shape, center, spread.

- **Shape:** tells which values tend to be more or less common
- **Center**: measure of what is typical in the distribution of a quantitative variable
- **Spread:** measure of how much the distribution's values vary

Definitions

- **Symmetric distribution**: balanced on either side of center
- **Skewed distribution:** unbalanced (lopsided)
- **Skewed left:** has a few relatively low values
- **Skewed right:** has a few relatively high values
- **Outliers:** values noticeably far from the rest
- **Unimodal:** single-peaked
- **Bimodal:** two-peaked
- **Uniform:** all values equally common (flat shape)
- **Normal:** a particular symmetric bell-shape

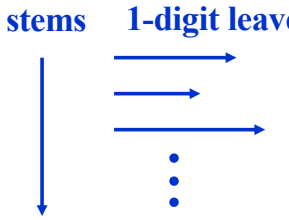
Displays of a Quantitative Variable

Displays help see the shape of the distribution.

- **Stemplot**
  - Advantage: most detail
  - Disadvantage: impractical for large data sets
- **Histogram**
  - Advantage: works well for any size data set
  - Disadvantage: some detail lost
- **Boxplot**
  - Advantage: shows outliers, makes comparisons C➔Q
  - Disadvantage: much detail lost

Definition

- **Stemplot:** vertical list of stems, each followed by horizontal list of one-digit leaves



Example: *Constructing a Stemplot*

- **Background**: Masses (in 1000 kg) of 20 dinosaurs:

0.0 0.0 0.1 0.2 0.4 0.6 0.7 0.7 1.0 1.1 1.1 1.2 1.5 1.7 1.7 1.8 2.9 3.2 5.0 5.6

- **Question:** Display with stemplot; what does it tell us about the shape?

Example: *Constructing a Stemplot*

- **Background**: Masses (in 1000 kg) of 20 dinosaurs:

0.0 0.0 0.1 0.2 0.4 0.6 0.7 0.7 1.0 1.1 1.1 1.2 1.5 1.7 1.7 1.8 2.9 3.2 5.0 5.6

- **Response:**

Do not skip the 4 stem: why?

_____

Long _____ tail→_____-skewed.

1 peak→_____

Most below 2000 kg, a few unusually heavy.

Modifications to Stemplots

- *Too few stems?* **Split…**
  - **Split in 2:** 1st stem gets leaves 0-4, 2nd gets 5-9
  - **Split in 5:** 1st stem gets leaves 0-1, 2nd gets 2-3, etc.
  - **Split in 10:** 1st gets 0, …, 10th gets 9.
- *Too many stems?* **Truncate** last digit(s).

Example: *Splitting Stems*

- **Background**: Credits taken by 14 "other" students:
  4 7 11 11 11 13 13 14 14 15 17 17 17 18
- **Questions:** What shape do we guess for non-traditional (other) students? How to construct stemplot to make shape clear?
- **Responses:**
  - Expect shape _____-skewed due to _____
  - Stemplot: 1st attempt has too few stems

  0 | 4 7

  1 | 1 1 1 3 3 4 4 5 7 7 7 8 so split 2 ways:

Displays of a Quantitative Variable

- **Stemplot**
- **Histogram**
- **Boxplot**

Example: *Truncating Digits*

- **Background**: Minutes spent on computer day before

  0 10 20 30 30 30 30 45 45 60
  60 60 67 90 100 120 200 240 300 420
- **Question:** How to construct stemplot to make shape clear?
- **Response:** Stems 0 to 42 too many: *truncate* last digit, work with 100's (stems) and 10's (leaves):

Skewed _____: most times less than 100 minutes, but a few had unusually long times.

Definition

- **Histogram:** to display quantitative values…
  1. Divide range of data into intervals of equal width.
  2. Find count or percent or proportion in each.
  3. Use horizontal axis for range of data values, vertical axis for count/percent/proportion in each.

Example: *Constructing a Histogram*

- **Background**: Prices of 12 used upright pianos:

  100 450 500 650 695 1100 1200 1200 1600 2100 2200 2300

- **Question:** Construct a histogram for the data; what does it tell us about the shape?
- **Response:**

We opted to put 500 as left endpoint of 2nd interval; be consistent (a price of 1000 would go in 3rd interval, not 2nd).

Definitions *(Review)*

- **Shape:** tells which values tend to be more or less common
- **Center**: measure of what is typical in the distribution of a quantitative variable
- **Spread:** measure of how much the distribution's values vary

Definitions

- **Median:** a measure of center:
  - ***the*** middle for odd number of values
  - average of middle two for even number of values
- **Quartiles:** measures of spread:
  - 1st Quartile (Q1) has one-fourth of data values at or below it (middle of smaller half)
  - 3rd Quartile (Q3) has three-fourths of data values at or below it (middle of larger half)

(By hand, for odd number of values, omit median to find quartiles.)

Definitions

- **Percentile:** value at or below which a given percentage of a distribution's values fall

A Closer Look: Q1 is 25th percentile, Q3 is 75th percentile.

- **Range:** difference between maximum and minimum values
- **Interquartile range:** tells spread of middle half of data values, written IQR=Q3-Q1

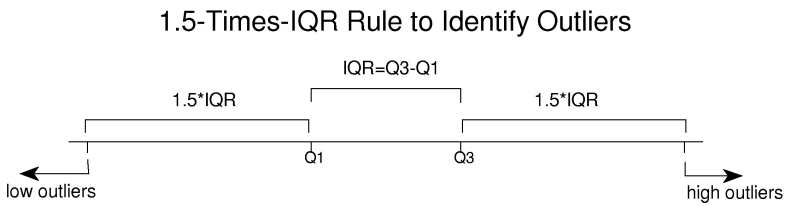
Ways to Measure Center and Spread

- **Five Number Summary:**
  1. Minimum
  2. Q1
  3. Median
  4. Q3
  5. Maximum
- **Mean** and **Standard Deviation**
  (more useful but less straightforward to find)

Definition

The **1.5-Times-IQR Rule** identifies outliers:

- below Q1-1.5(IQR) considered low outlier
- above Q3+1.5(IQR) considered high outlier



Example: *Finding 5 Number Summary and IQR*

- **Background**: Credits taken by 14 non-traditional students:
  4 7 11 11 11 13 13 14 14 15 17 17 17 18
- **Question:** What are Five Number Summary, range, and IQR?
- **Response:**
  1. Minimum: ____
  2. Q1: ____
  3. Median: ____
  4. Q3: ____
  5. Maximum: ____

  Range: _____

  IQR: _____

Displays of a Quantitative Variable

- **Stemplot**
- **Histogram**
- **Boxplot**

Definition

A **boxplot** displays median, quartiles, and extreme values, with special treatment for outliers:

1. Bottom whisker to minimum non-outlier
2. Bottom of box at Q1
3. Line through box at median
4. Top of box at Q3
5. Top whisker to maximum non-outlier

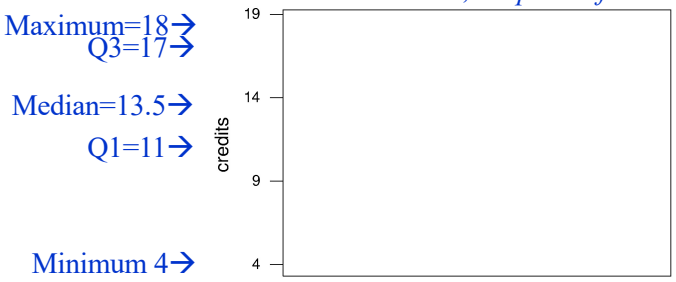
Outliers denoted "*".

Example: *Identifying Outliers*

- **Background**: Credits taken by 14 non-traditional students had 5 No. Summary: 4, 11, 13.5, 17, 18
- **Questions:** Are there outliers?
- **Responses:** Q1=___, Q3=___
  - IQR=_____
  - 1.5×IQR=___
  - Q1-1.5(IQR)=_____: Low outliers? ____.
  - Q3+1.5(IQR)=_____: High outliers?____.

Example: *Constructing Boxplot*

- **Background**: Credits taken by 14 non-traditional students had 5 No. Summary: 4, 11, 13.5, 17, 18
- **Question:** How is the boxplot constructed?
- **Response:** *Typical credits about 13.5, middle half between 11 and 17, shape is left-skewed*



Lecture Summary
(Finish Cat.; Quantitative Displays, Begin Summaries)

- **Issues about Categorical Variables:** Two or more possibilities? Categories overlap? Handle raw data?
- **Quantitative Variables**
  - **Display:** stemplot, histogram
  - **Shape:** Symmetric or skewed? Unimodal? Normal?
  - **Center and Spread**
    - median and range, IQR
      - identify outliers
      - display with boxplot