Lecture 6: Chapter 4, Section 2 Quantitative Variables (Displays, Begin Summaries)

- □Summarize with Shape, Center, Spread
- □Displays: Stemplots, Histograms
- □Five Number Summary, Outliers, Boxplots

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing
    - Single variables: 1 cat. (Lecture 5), 1 quantitative
    - Relationships between 2 variables
  - Probability
  - Statistical Inference

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.*

Practice: 4.17 p.92

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.*

Practice: 4.29c p.104

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

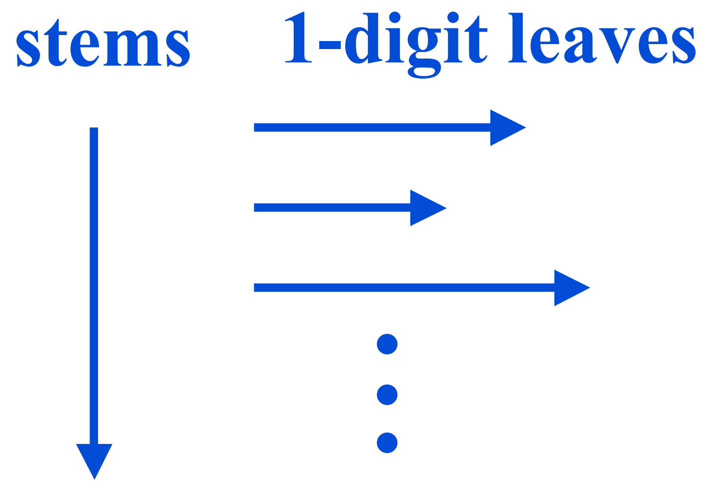
Displays of a Quantitative Variable

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.

Practice: 4.38a p.107

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:

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.

Practice: 4.25b p.92

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).

Practice: 4.27a p.103

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)

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: ______________

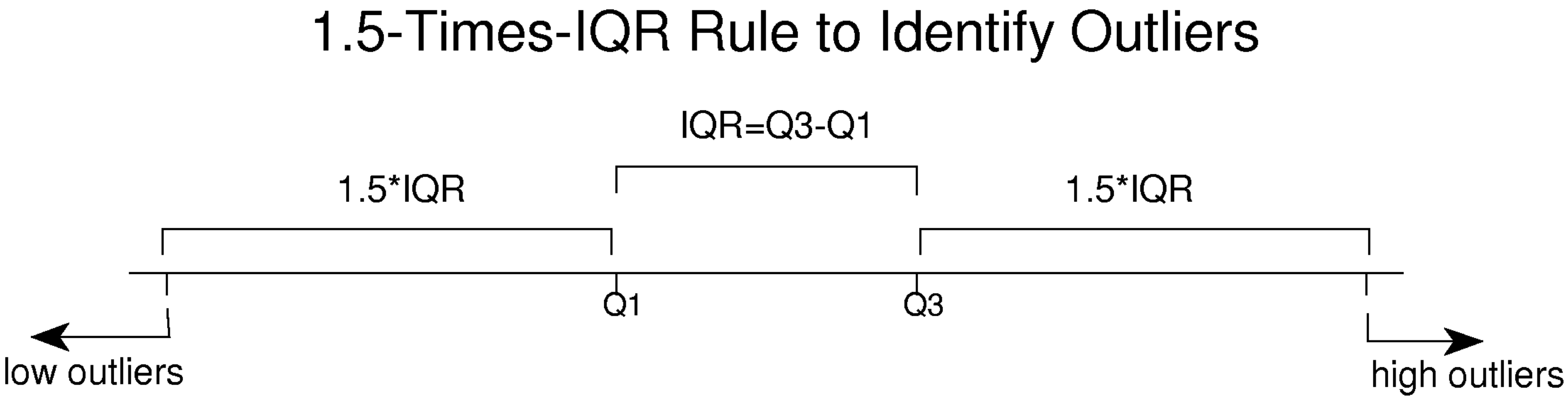
IQR:

Practice: 4.28b p.103

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



Definition

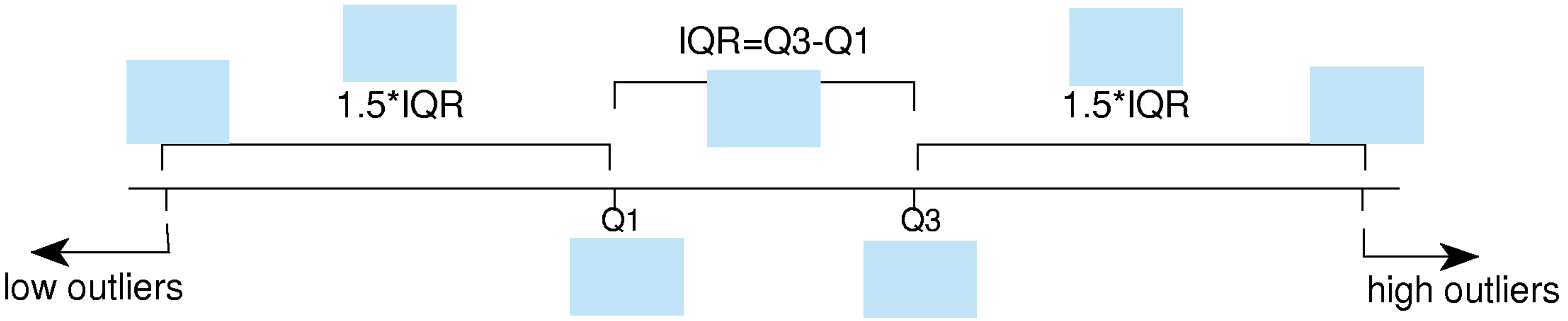
A **boxplot** displays median, quartiles, and extreme values, with special treatment for outliers:

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?______



Practice: 4.28c p.103

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*

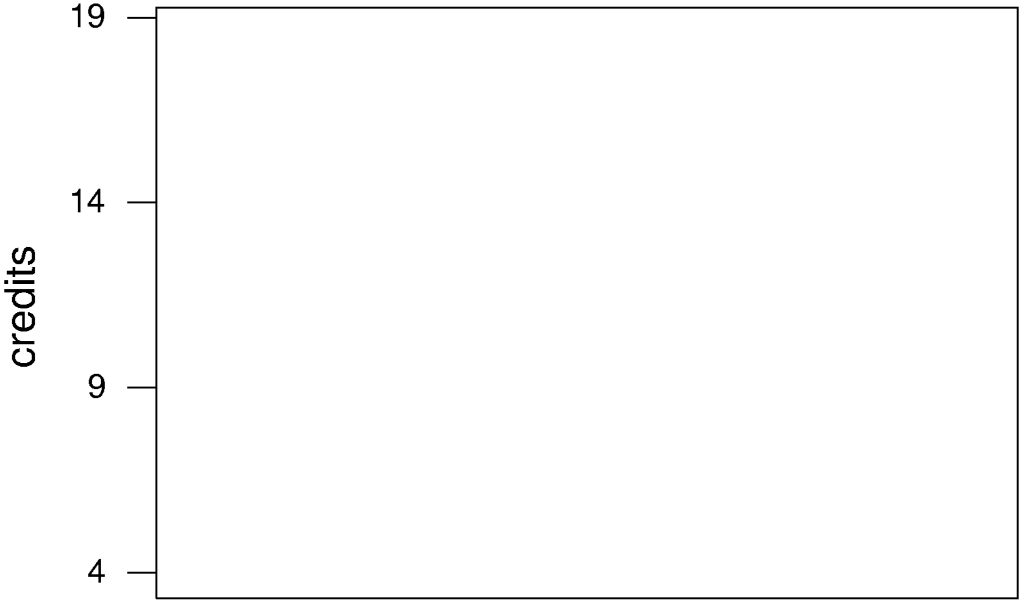
Maximum=18→

Q3=17→

Median=13.5→

Q1=11→

Minimum 4→



Practice: 4.28c p.103

Lecture Summary
(Quantitative Displays, Begin Summaries)

- **Display:** stemplot, histogram
- **Shape:** Symmetric or skewed? Unimodal? Normal?
- **Center and Spread**
  - median and range, IQR
    - identify outliers
    - display with boxplot