Displaying Distributions with Graphs

- **Recall that the *distribution* of a variable indicates two things: (1) What value(s) a variable can take, and (2) how often it takes those values.**
- ***Example 1:* Weights of a sample of statistics students (VCU, 1997).**
- **What kind of variable is *weight*?**
- **How should we display the *pattern* of this data set?**

```
 192 110 195 180 170 215 152 120 170 130 130
125 135 185 120 155 101 194 110 165 185 220 180
128 212 175 140 187 180 119 203 157 148 260 165
185 150 106 170 210 123 172 180 165 186 139 175
127 150 100 106 133 124
```

Histograms

- **Recall we can use a *Data Table* to display the distribution of a *categorical* variable.**
- ***Histograms* are a convenient way to show the distribution of a *quantitative* variable.**
- **We must divide the range of the data set into *measurement classes* of equal width.**
- **Each data value should fall into exactly one measurement class.**
- **The histogram plots the counts of data values in each class according to the heights of vertical bars.**
- **Some histograms plot *relative frequencies* (rather than *frequencies*) for the various measurement classes on the vertical axis.**

Example Histogram

- **Look at *histogram* of the student weight data.**
- **What are the *measurement classes*? (Clicker quiz)**
- **How could we determine the total sample size from this histogram?**
- **How does this differ from a bar graph? (Look at horizontal axis)**
- **We don't want too many or too few measurement classes.**
- **Good software programs usually pick a reasonable number ($\approx$ 6 to 12 classes?)**

Clicker Quiz 1

In the student weight histogram, what are the measurement classes used?

A. [0, 100), [14, 280)

B. [0, 2), [2, 4), [4, 6), [6, 8) [8, 10), [10, 12), [12, 14)

C. [100, 280)

D. [100, 120), [120, 140), [140, 160), [160, 180), [180, 200), [200, 220), [220, 240), [240, 260), [260, 280)

Clicker Quiz 2

In the student weight data set, how many students weigh at least 220 pounds?

A. 0

B. 1

C. 2

D. 6

Interpreting Histograms

- **Computer packages will produce nice graphs.**
- **Our job is to interpret what they tell us about the distribution of data.**
- **Look for *overall pattern* and any *deviations* from that pattern.**

Outliers

- **An *outlier* is a data value that doesn't follow the overall pattern of the bulk of the data.**
- **May be a naturally occurring unusual value**
- **May be due to a *recording error* or *measurement error***
- **May be an observation from a fundamentally different population**
- **When you see an outlier, go back and investigate that observation!**
- **See example with reading data set.**

Noting the Pattern in a Histogram

- **Determine the *overall pattern* based on the bulk of the data (not solitary outliers).**
- **What is the *center* of the distribution of data (average or most typical value)?**
- **How *spread out* is the distribution of data?**
- **What is the basic *shape* of the distribution of data?**

Clicker Quiz 3

The *midpoint* of a distribution is the number having half of the data below it and half of the data above it. In the student weight histogram, what is the approximate *midpoint* of the distribution of weights?

A. 100

B. 170

C. 200

D. 250

Clicker Quiz 4

To describe the *spread* of a distribution, we could give the smallest and largest data values, *ignoring outliers*. In the student weight histogram, what describes the spread of the distribution of weights, not counting solitary outliers?

A. 100 to 280

B. 150

C. 100 to 240

D. 120 to 200

Different Shapes of Distributions

- **A distribution is *symmetric* if the left and right sides of the histogram are approximately mirror images.**
- **Otherwise, the distribution is called *skewed*.**
- **A distribution is *skewed to the right* if the right side of the histogram extends much farther out than the left side ("long right tail")**
- **A distribution is *skewed to the left* if the left side of the histogram extends much farther out than the right side ("long left tail")**
- **Histograms for real data sets hardly ever show "perfect mirror images."**
- **If it's pretty close, we could say the distribution seems symmetric.**

Different Shapes of Distributions (continued)

- **A distribution is *unimodal* if the histogram shows one dominant peak.**
- **A distribution is *bimodal* if the histogram shows two separate peaks.**
- ***Caution*: The same data set could produce somewhat different-looking histograms.**
- **The appearance of the histogram depends somewhat on the choice of the measurement classes.**
- **See examples: Shakespeare data set and elderly residents data set.**

Clicker Quiz 5

Why might the Shakespeare word-length data have a distribution that is skewed to the right?

A. The lengths of words are basically random.

B. Very long words are possible (though uncommon), but no word can be shorter than one letter.

C. Shakespeare liked to impress people by using many words that were very long.

D. The bulk of the sample data is on the right side of the histogram.

Stemplots

- ***Stemplots* are also called *stem-and-leaf plots*.**
- **Similar to histograms, but they show the exact data values (not just measurement classes).**
- **Especially useful for quantitative data when the sample size is *small*.**
- **Each data value divided into a "stem" part and a "leaf" part, and these digits are plotted.**
- **See example stemplot for student weight data (Can you see the possible response/measurement error?)**