We will be looking at graphical and numerical summaries of data, focusing (for today) on just one variable at a time.

Goal:

1 Graphical Summaries of Categorical Variables

The distribution of a categorical variable tells what values it takes and how often it takes those values. This could be summarized in a table giving either the _____ or the _____ of individuals that fall in each category.

Example based on our class: gender, year at Calvin

These tables can be presented graphically:

Bar Graph: Pie Chart:

In each of these pictures, the relative frequency of each level of the categorical variable is represented by _____. In particular, this means that each bar in the bar chart must have the same _____.

Q. Why can graphs charts be used in a broader range of situations than pie charts?

A.

2 Graphical Quantitative Variables

For a summary of the distribution of a quantitative variable we want to know

- the _____ of the data, and
- any _____ .

2.1 Stemplots (stem-and-leaf plots)

A stemplot is a handy graphical display of the distribution of a quantitative variable that works well for _____ .

For a stemplot we separate each numerical value into two pieces: the stem (leading digits) and the leaf (last digit(s)). The possible stems are written down vertically, and the leaves are written to the right of their stems.

Example: Baseline weights of women who dropped out of HWS

unsorted: 204 171 139 179 142 132 129 174 162 120 146 183 141 117 209

sorted: 117 120 129 132 139 141 142 146 162 171 174 179 183 204 209

Stemplots can also be used to compare two groups. Simply place the stems down the middle with leaves for each group placed to either side.

Example: Below are cholesterol levels for the heaviest 2.5% and lightest 2.5% of the women in the HWS.

cholesterol of lightest women: 137 157 164 166 167 181 188 201 203 212 221 225 269

cholesterol of heaviest women: 150 157 159 165 169 171 176 194 195 206 227 243 259

2.2 Histograms

A histogram is similar to a bar graph. First we break the range of values into _____ _____ called _____ . Then we make a bar graph displaying the count or percent (relative frequency) of observations that fall into each bin.A

How to make a histogram

1. Decide on the bins.

   The goal is to select bins that reveal patterns and deviations from those patterns.

   For a standard histogram each bin is the same _____ (but there other possibilities). The selection of bins can have a dramatic difference on the appearance of the histogram. Usually choosing 7 to 12 bins will give you a good starting point, but it may be necessary to make repeated histograms with different bin choices to get the best picture.

2. Make a table of counts or percents in each bin.

3. Draw the histogram. Usually we place the bin boundaries along the horizontal axis and the frequencies along the vertical axis.

EXAMPLE. The Prussian Army kept records of many things, including how many soldiers died from being kicked by horses. Below are the data for fourteen army corps and the number of deaths by horsekick each year from 1875 until 1894.[1]

	Year	G	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XIV	XV	Total
1	75	0	0	0	0	0	0	0	1	1	0	0	0	1	0	3
2	76	2	0	0	0	1	0	0	0	0	0	0	0	1	1	5
3	77	2	0	0	0	0	0	1	1	0	0	1	0	1	1	7
4	78	1	2	2	1	1	0	0	0	0	0	1	0	1	0	9
5	79	0	0	0	1	1	2	2	0	1	0	0	2	1	0	10
6	80	0	3	2	1	1	1	0	0	0	1	1	4	3	0	18
7	81	1	0	0	2	1	0	0	1	0	1	0	0	0	0	6
8	82	1	2	0	0	0	0	1	0	1	1	2	1	4	1	14
9	83	0	0	1	2	0	1	2	1	0	1	0	3	0	0	11
10	84	3	0	1	0	0	0	0	1	0	0	2	0	1	1	9
11	85	0	0	0	0	0	0	1	0	0	2	1	0	0	1	5
12	86	2	1	0	0	1	1	1	0	0	1	0	1	3	0	11
13	87	1	1	2	1	0	0	3	2	1	0	0	1	2	0	15
14	88	0	1	1	0	0	1	1	0	0	0	0	1	1	0	6
15	89	0	0	1	1	0	1	1	0	0	1	2	2	0	2	11
16	90	1	2	0	2	0	1	1	2	0	2	1	1	2	2	17
17	91	0	0	0	1	1	1	0	1	1	0	3	3	1	0	12
18	92	1	3	2	0	1	1	3	0	1	1	0	1	1	0	15
19	93	0	1	0	0	0	1	0	2	0	0	1	3	0	0	8
20	94	1	0	0	0	0	0	0	0	1	0	1	1	0	0	4

[1] Source: Bortkiewicz 1898. As reported at ¡http://www.qc.edu/Biology/fac_stf/marcus/sasman3w.html¿ [{m243}horsekick]

Here is a table summarizing the data about deaths by horsekick. We can use it to make a histogram.

deaths	freq
0	144
1	93
2	30
3	11
4	2

Example. Here are the amounts spent by 50 consecutive shoppers at a supermarket from which we can make a histogram or a stemplot.

```
 3.11  8.88  9.26 10.81 12.69 13.78 15.23 15.62 17.00 17.39
18.36 18.43 19.27 19.50 19.54 20.16 20.59 22.22 23.04 24.47
24.58 25.13 26.24 26.26 27.65 28.06 28.08 28.38 32.03 34.98
36.37 38.64 39.16 41.02 42.97 44.08 44.67 45.40 46.69 48.65
50.39 52.75 54.80 59.07 61.22 70.32 82.70 85.76 86.37 93.34
```

bin	count	percent	Stemplot(s)

Histogram(s)

Q. What are the advantages and disadvantages of stemplots vs. histograms?

A.

2.3 Shapes of Distributions

Once we have a histogram or a stemplot, we can "see" the distribution of a quantitative variable. What shapes do they have?

unimodal bimodal

symmetric right skewed left skewed

low variability hi variability

2.4 Outliers

Outliers are:

2.5 Time Plots (time series plots)

Sometimes data are collected over time and trends can be missed if you look at a summary that ignores when the data were collected (as histograms and stemplots do).

Example: Are there any time-based trends in the horsekick data?

Example: A persons blood pressure over time.

Example: Looking for controlling for "lab drift". A standard sample can be measured periodically to make sure that there has been no change in instrumentation, calibration, research protocols, etc. over time.

3 Numerical Summaries of Quantitative Distributions

Some notation:

- $n =$
- X or $Y =$
- $\{x_1, x_2, x_3, \ldots, x_6\} =$
- $\sum x_i =$
- So if $\{x_1, x_2, x_3, \ldots, x_6\} = \{3, -2, 4, 0, 12\}$, then $\sum x = \sum x_i =$

3.1 Measures of Center: Mean and Median

The mean is what you usually think of as the average. Simply add up all the values and divide by the number of values. Using the notation above, and writing the mean as $\bar{x}$, this can be written as

$$\bar{x} = \sum x/n$$

Example: data=2, 5, 6, 7, 10 $\bar{x} =$

The median is the _____ data value once the data values have been _____ .

Example: data=2, 5, 6, 7, 10 median =

Example: data=2, 5, 6, 7, 10, 15 median =

3.1.1 Lottery Example

Suppose 100 raffle tickets are sold for \$2 each. 85 tickets win nothing, 10 win \$5, 4 win \$10 and 1 wins \$10,0000. What is a raffle ticket worth?

Mean =

Median =

This example shows that the median is a _____ measure of center because

On the other hand, the mean is _____ to _____ because

3.1.2 Relation to Graphs

Where are the mean and median on the graphs of various distributions?

unimodal, symmetric

bimodal, symmetric

right skewed

left skewed

3.2 Measures of Spread: Range, Interquartile Range, Variance, Standard Deviation

The range is a very crude measure of spread.

Range:

A better idea of the spread of the data can be obtained by computing several **percentiles**.

p*th* percentile:

Similarly one can use **quartiles** or **deciles**.

0th quartile =

1st quartile =

2nd quartile =

3rd quartile =

4th quartile =

These five numbers are called the **five-number summary** of a data set.

Example: The number of times in a month (Jan–Dec 2000) that Professor Plantinga's son Peter liked her cooking:

3 3 7 1 8 2 2 5 4 7 8 20 Sorted: 1 2 2 2 3 3 4 5 7 7 8 8 20

3.2.1 Boxplots

A boxplot is a graphical representation of the five-number summary.

Example: Boxplot for Peter liking cooking.

Example: Boxplot for Peter liking cooking with outliers indicated.

Boxplots give a rough indication of the center and spread of a distribution, but usually a stemplot or histogram should also be made to provide a better picture of the overall shape of the distribution.

Side-by-side boxplots can be a convenient way to compare two distributions.

3.2.2 Variance and Standard Deviation

Using our notation, we can express the variance very easily:

$$\text{variance} = s^2 = \sum (x_i - \bar{x})^2/(n-1)$$

The standard deviation is the square root of the variance.

But what does this say?

- $x_i - \bar{x}$:
- if we didn't square before summing:
- dividing by $n - 1$:
- taking the square root at the end (for standard deviation):

Example: 1 3 8 8 10

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$

Properties of standard deviation.

1.
2.
3.

3.3 Picking Measures of Center and Spread