Graphing Frequency (count) with Numerical (Quantitative) data

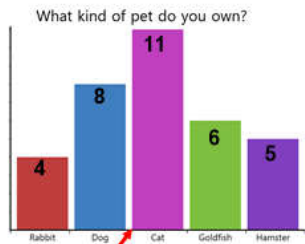In this chapter we are (mainly) still considering the frequency (count), but instead of counts of category amounts, the variable is continuous, so if we want something like a bar chart, we must divide the possible variable values into something like categories - called 'bins':
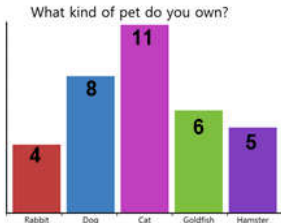
Bar chart
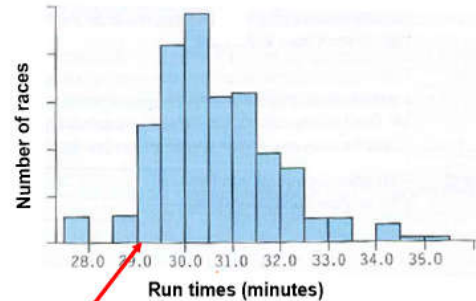**(Categorical data)**
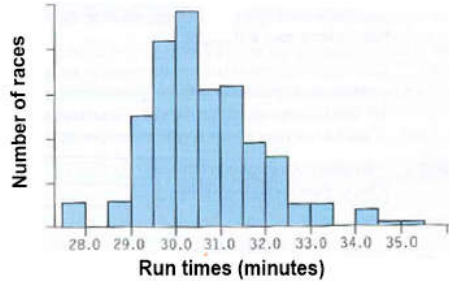
Histogram
(Quantitative data)

Bar chart

gaps between bars
implying the order is arbitrary

no gaps between bars because graph
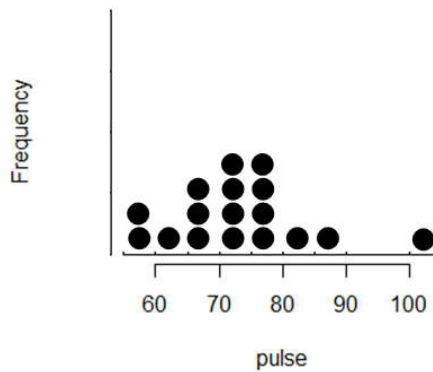accounts for all possible values in a range

order doesn't matter

order matters
(increasing, left to right)

Histogram

Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 105, 72, 84, 72
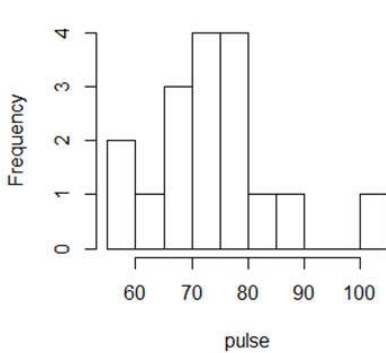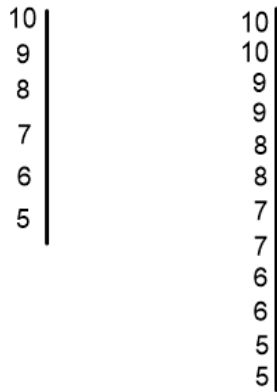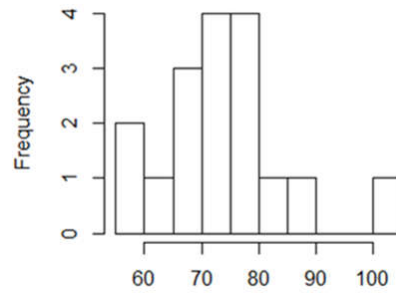
Histogram

Dotplot

(Note: Histograms, Dotplots, and Stem-and-Leaf displays obey the area principle)

Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 72, 84, 72, 105

Stem-and-Leaf

```
10 |          10 |
 9 |          10 |
 8 |           9 |
 7 |           9 |
 6 |           8 |
 5 |           8 |
               7 |
               7 |
               6 |
               6 |
               5 |
               5 |
```

(compare to histogram)

always include a key:

Describing Shape, Center, and Spread

Because the data is numerical and shown in order, the shape of the distribution is meaningful and can be described.

SOCS

Shape Outliers Center Spread

Shape: Symmetric or skewed
Outliers: Gaps, unusual features, IQR rule of thumb
Center: Median or mean
Spread: Range, IQR, Variance, or Standard Deviation

Shape: Symmetric or skewed

Symmetrical

Skewed left
(tail toward lower values)

Skewed right
(tail toward higher values)
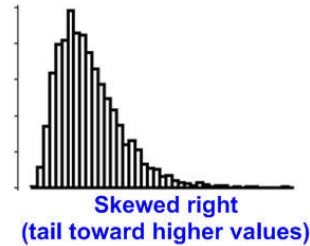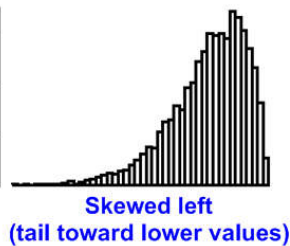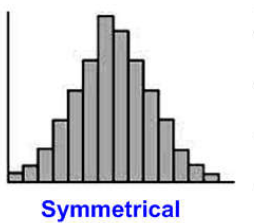
A **'statistic'** is a number that represents a data set in some way.

Center: Median or mean

Median: Middle value (or average of two middle values)

Mean: $\bar{x} = \dfrac{\sum x}{n}$

Mode: Most occurring value(s)

mode

Unimodal

mode

mode

Bimodal

How mean, median, and skew are related

What is the mean and median of this data set?

<div align="center">

6   7   8   9   10

</div>

$$\overline{X} = \frac{6+7+8+9+10}{5} = 8$$

$$Median = 8$$

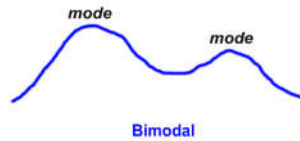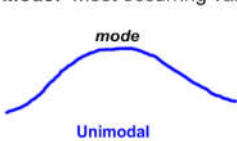**If we move the '10' farther away from the mean, how does mean and median change?**

<div align="center">

6   7   8   9   20

</div>

$$\overline{X} = \frac{6+7+8+9+20}{5} = 10$$

$$Median = 8$$

The median is unaffected, but the outlier 'pulls the mean towards itself'

In a perfectly symmetric distribution, the mean and median are aligned, but in skewed distributions, the mean is pulled in the direction of the tail:

| Symmetrical | Skewed left (mean < median) | Skewed right (mean > median) |
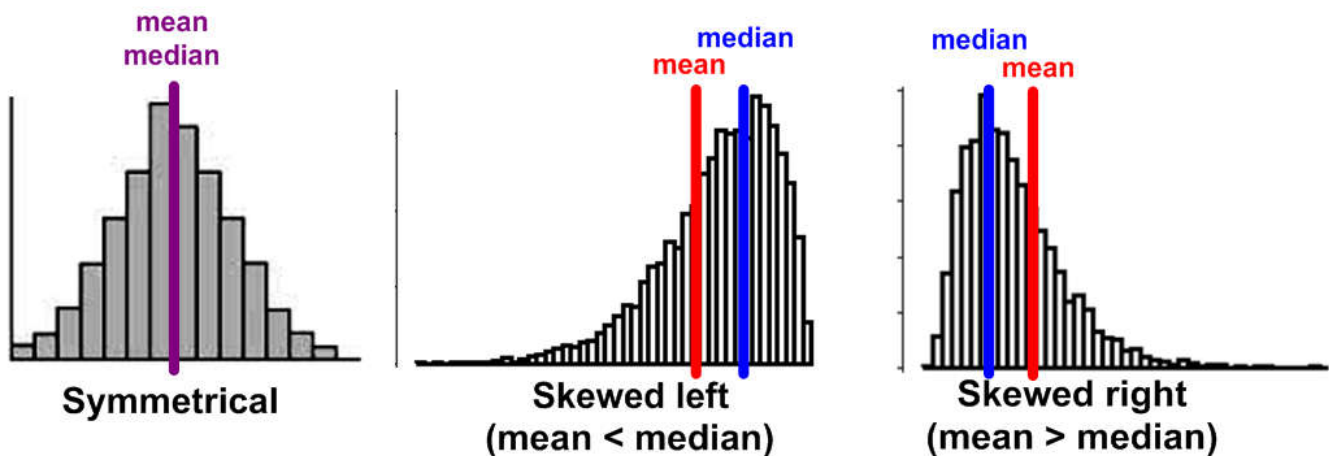
Spread: Range, IQR, Variance, or Standard Deviation

Range: Maximum value - Minimum value

Quartiles: 56, 60, 64, 68, 68, 68, 72, 72, 72, 72, 76, 76, 80, 80, 84, 88, 105

Min   Q1   Median (IQR)   Q3   Max

Interquartile Range (IQR):   Q3 - Q1   (50% of the data lies in the IQR)

Variance: $\quad s^2 = \dfrac{\sum\left(x-\overline{x}\right)^2}{n-1}$

Standard Deviation: $\quad s = \sqrt{s^2} = \sqrt{\dfrac{\sum\left(x-\overline{x}\right)^2}{n-1}}$

Conceptual meaning of variance and standard deviation

women's pulse (beats per minute)

56, 60, 64, 68, 68, 68, 72, 72, 72, 72, 76, 76, 80, 80, 84, 88, 105

How far away from mean is this value?

68 - 74.1 = -6.1 bpm

What is the 'typical' or 'average' value of the distances away from the mean?

If we just added the distances for each, some would be positive, some negative, and they would cancel each other out. So we square the difference to make them all positive before we add. Then we divide by the number of distances^2 to get average distance^2.
This is **variance**:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$
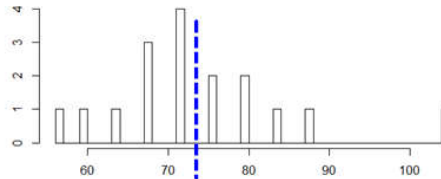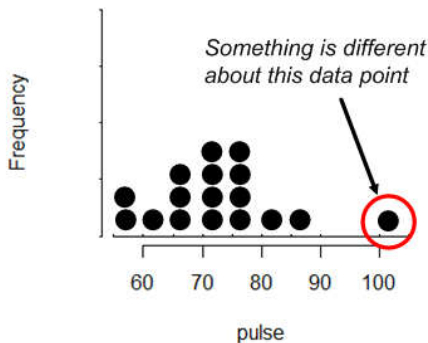

$\bar{x}=74.1$

But we really want distances (spread) in terms of the original units (bpm, not bpm^2) so we take the square root of the result. This the **standard deviation**:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Standard deviation is 'typical' or 'average' distance of all data points from the mean.

Outliers: Gaps, unusual features, IQR rule of thumb

Something is different about this data point

Outlier rule-of-thumb

A data point is an outlier if its value is:

$$< Q1 - 1.5IQR$$

or

$$> Q3 + 1.5IQR$$

Summarizing information about a dataset

Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 72, 84, 72, 105

5 number summary

Min: 56
Q1: 68
Median: 72
Q3: 80
Max: 105

Box & Whisker Plot
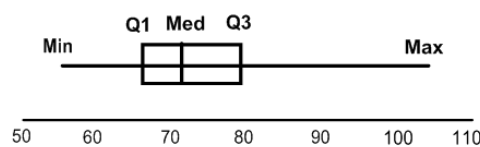
Sometimes, outliers are excluded and plotted separately:
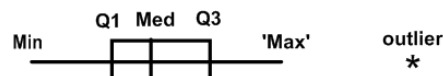
Is 105 an outlier?

$$IQR = Q3 - Q1 = 80 - 68 = 12$$

$$upper\ fence = Q3 + 1.5IQR = 80 + 1.5(12) = 98$$

105 > 98, so, <u>yes</u>, 105 is an outlier.

Calculator statistics and plots

If you have complete data set: Use 1 list and 1-Var-Stats
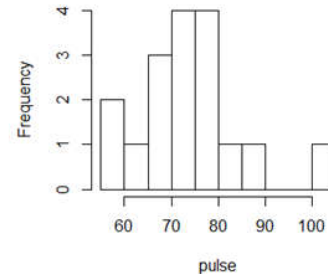
Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 72, 84, 72, 105

- Enter data in L1 (Stats, Edit)
- Stats, Calc, 1-Var Stats

If you have a histogram: Use 2 lists and 1-Var-Stats

- Enter representative bar values data values in L1 (Stats, Edit)
- Enter data counts (height of bars) in L2
- Stats, Calc, 1-Var Stats w/FreqList set to L2

To display a histogram:

- Enter data in L1 (Stats, Edit) or L1 and frequency in L2
- "Y=" and clear out any equations
- 2nd "Y=" to enter set up for Statistics Plot 1.
- Set to match the screen on the right:
- Zoom: 9 (ZoomStat)

'WINDOW'
change Xscl
'GRAPH'
to change
bin width

To display a boxplot:

- Enter data in L1 (Stats, Edit) or L1 and frequency in L2
- "Y=" and clear out any equations
- 2nd "Y=" to enter set up for Statistics Plot 1.
- Set to match the screen on the right:
- Zoom: 9 (ZoomStat)

always use the
one with the dots
so you'll see
outliers

Boxplot to Histogram (approximate) - the 'Aquarium'

The distribution is broken into quarters at the Q1, Median, and Q3 marks and
each quarter contains about 25% of the 'water' (distribution):
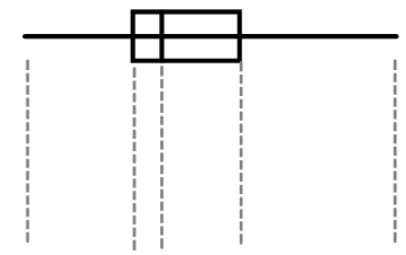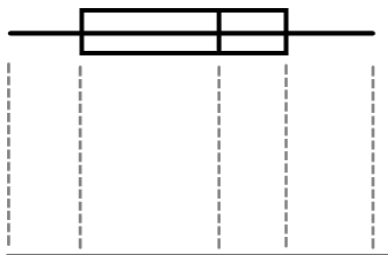
'Resistant' Statistics

Statistics are **resistant** if they are not overly affected by the presence of outliers.

<u>Measures of Center</u>

Median: resistant
Mean: not resistant

<u>Measures of Spread</u>

IQR: resistant
Variance: not resistant
Std deviation: not resistant

Conclusions: If the plot (box, histogram, stem & leaf, dot) of the distribution shows that it is...

...symmetric:

- Can use mean or median to measure center.
- Can use standard deviation or IQR to measure spread.

...skewed:

- Must use median to measure center.
- Must use IQR to measure spread.

Re-expressing Skewed Data (Transforms)

Sometimes, data is so skewed that it is difficult to conclude things even with a picture:
Example: CEOs asked, 'what is your annual income?'

Data can made more symmetric by re-expressing, or transforming each data value. Usually this is done by taking a logarithm (base 10) or square root.

original income value	transformed income value $\log_{10}$
1,500,000	6.2
15,000,000	7.2
150,000,000	8.2
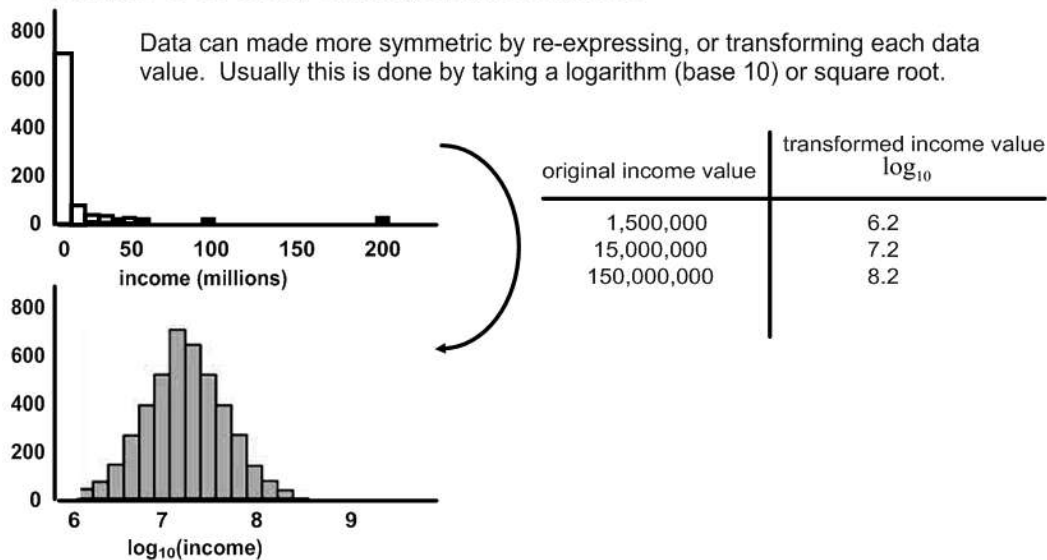
This chapter is mostly about _single variable_ data

For most of the data in this chapter, there is only **one** variable. If we are looking at a histogram, the values of the variable are shown on the horizontal axis and the vertical axis isn't another variable...it is the frequency or count of the number of data with a particular value:

15

Frequency
(count)   10

5

0

          50 60 70 80 90 100

Test scores

This is the variable
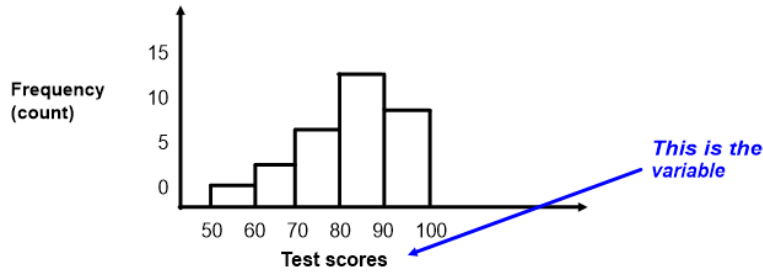
Timeplots

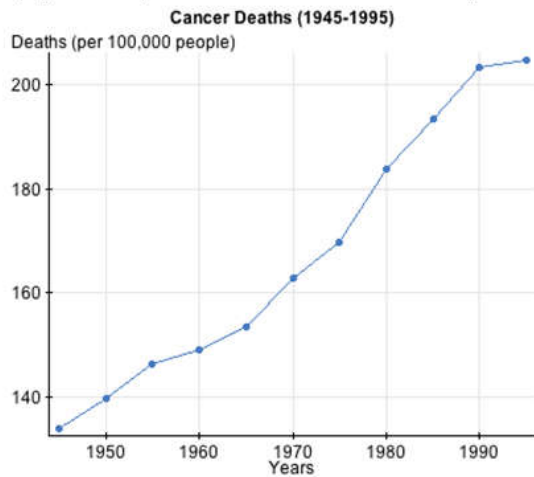This chapter is mostly about displaying information about the frequency (counts) for various 'bin' values of a variable. The book mentions one other type of plot called a 'Timeplot' which is fundamentally different (more like later chapters).

A *timeplot* is simply a 2-D plot of actual data values (not counts) vs time.

Cancer Deaths (1945-1995)

Deaths (per 100,000 people)

200

180

160

140

          1950   1960   1970   1980   1990
                          Years

'Variable' vs. 'Count'

	Categorical (Qualitative)	*Numerical (Quantitative)*

1 variable

count (freq) 30 20 10

male   female
gender

possible values

count (freq) 30 20 10

10 20 30 40 50 60
age

variable

2 variables

variables → gender

possible values → male   female

favorite flavor

	male	female
chocolate	22	17
vanilla	31	24
other	6	9

income $120k
        $80k
        $40k

        10 20 30 40 50 60
        age

variables → *age*