12.2 Descriptive Statistics

Objectives:

1. I can describe a distribution by its shape, outliers, center, and spread.

2. I can find population percentages of a normal distribution (68-95-99.7 rule).

Vocabulary:

Population: Set of all

Sample: A subset of the population

Parameter: Measures of a population

-Use $\mu = population\ mean$

$\boxed{\sigma} = population\ \underline{standard\ deviation}$ calc

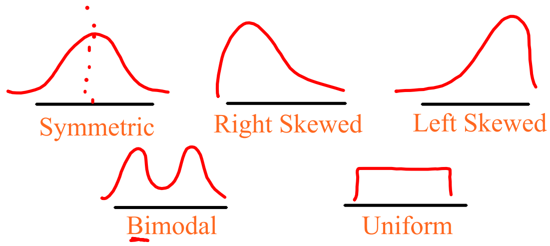
Statistics: Measures of a sample

-Use $\boxed{\bar{x} = sample\ mean}$ calc

$s = sample\ standard\ deviation$ calc

"Remember your S.O.C.S"

1. SHAPE:



2. OUTLIERS: Data far away from the rest of the data. Formula to come ...

80, 92, 76, 101, $\boxed{27}$

3. CENTER: Measures of central tendency:

1. Mean - arithmetic average of the data
2. Median - Middle value when placed in order, or average of the two middle values
3. Mode - Most frequently occurring value(s)

1, 7, 5, 5, 6 5

4. SPREAD: Measure of the variability in the data

Mean - Median - Mode ?

The average on the test was an 84 - Mean

The average test score puts you in the middle of the class - median

The average American student starts college at 18- mode

Mean, Median and Mode

1, 3, 4, 7

$\frac{3+4}{2} = 3.5$

The **mean** of a list of n numbers $\{x_1, x_2, \ldots, x_n\}$ is:

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Summation

n → total

the **mean** is strongly effected by outliers

The **median** of a list of n numbers $\{x_1, x_2, \ldots, x_n\}$ arranged in order (either ascending or descending) is:

- The middle number is n is odd   1, 3, $\boxed{5}$, 10, 11
- The mean of the two middle numbers if n is even   1, 3, $\boxed{5, 10}$, 11, 12 $\frac{5+10}{2} = \boxed{7.5}$

Median is resistant meaning it is not strongly effected by outliers

The **mode** of a list of numbers is the number that appears most frequently.

Find the mean, median, and mode for the following set of data:

12, 14, 10, 1, 9, 13, 17, 14, 16

1, 9, 10, 12, $\boxed{13}$, 14, 14, 16, 17

mean: $\frac{106}{9} = \boxed{11.8}$

median: 13

mode: 14

calculator

Stat → edit → enter data

Stat → calc → 1-var stat

NORMAL FLOAT AUTO REAL RADIAN MP
1-Var Stats
$\bar{x}$=11.77777778   mean
Σx=106
Σx²=1432
Sx=4.790035954
σx=4.516089207
n=9
minX=1
↓Q₁=9.5

NORMAL FLOAT AUTO REAL RADIAN MP
1-Var Stats
↑Sx=4.790035954
σx=4.516089207
n=9
minX=1
Q₁=9.5
Med=13   median
Q₃=15
maxX=17

mode:

1, 3, 4, 7, 8 → none

1, 2, 2, 5, 7, 9, 9 → 2, 9

Why do we have all of these measures?

Example: On a cul-de-sac, you have 5 houses built for:

1, 3, 3, 7

$200,000, $200,000, $\boxed{\$200,000}$, $200,000, $1,200,000

Find the median and the mean? Which one is a better measure?

median: 200,000

mean: 400,000

Spread: When we use the mean to measure center, we use standard deviation

Measures variability

The standard deviation of the numbers $\{x_1, x_2, \ldots, x_n\}$ is

1,3,5,7,10

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \underbrace{\bar{X}}_{5.2})^2}$$

where $\bar{X}$ denotes the mean. The variance is σ^2 the square of the standard deviation.

By hand this can be tedious- luckily we can do this in our calculator.

```
NORMAL FLOAT AUTO REAL RADIAN MP
1-Var Stats
x̄=5.2
Σx=26
Σx²=184
Sx=3.492849839
σx=3.12409987
n=5
minX=1
↓Q1=2
```

standard deviation

Find the standard deviation: Weights in grams of 30 loon chicks

79.5	87.5	88.5	89.2	91.6	84.5	82.1	82.3	85.1	89.8
84.0	84.8	88.2	88.2	82.9	89.8	89.2	94.1	88.0	91.1
91.8	87.0	87.7	88.0	85.4	94.4	91.3	86.3	85.7	86.0

Spread: When we use the median to measure center, we use 5-Number Summary

Range = maximum - minimum

Quartiles split the data into **fourths**

Min

First Quartile (Q_1) = the median of the lower half of the data

Med **Second Quartile** = the median

Third Quartile (Q_3) = the median of the upper half of the data

Max

Interquartile Range (IQR) measures the spread between Q_1 and Q_3

$$IQR = Q_3 - Q_1$$

Five number summary = {minimum, Q_1, median, Q_3, maximum}

$(61.5+66.7)/2$

Find the five number summary for the male and female life expectancies in South American nations and compare.

males: {59.0, 60.5, 61.5, 66.7, 67.9, 68.5, 69.0, 70.3, 71.4, 71.9, 72.1, 72.6}

Q1

females: {66.2, 66.7, 67.7, 72.8, 74.3, 74.4, 74.6, 76.5, 76.6, 78.8, 79.0, 79.4}

males: {59, 64.1, 68.75, 71.65, 72.6}

min Q1 med Q3 max

→ calculator

```
NORMAL FLOAT AUTO REAL RADIAN MP
        1-Var Stats
↑Sx=4.728988616
 σx=4.527661894
 n=12
 minX=66.2
 Q1=70.25
 Med=74.5
 Q3=77.7
 maxX=79.4
```

5# Summary

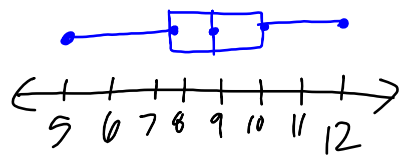
A **box plot** (sometimes called box and whisker plot) is a graph that depicts the five number summary of a data set.

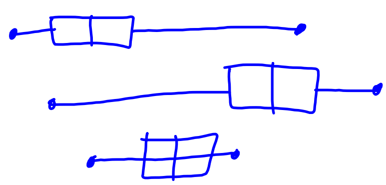
To Construct:

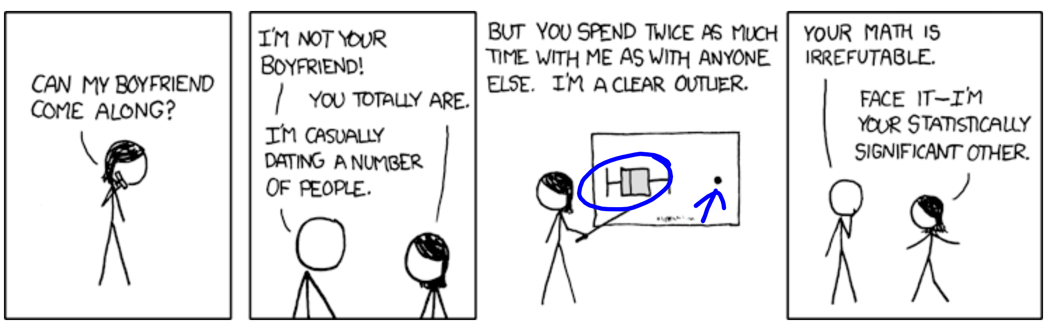
1. Draw a rectangular box from Q_1 to Q_3 with a vertical line for the median
2. Draw line segments (whiskers) that extend from the end of the box to the max and mins respectively

Shoe sizes: $\boxed{5}$, 7, 7.5, 8, 8, $\boxed{8.5}$, 9, 10, 10, 10.5, 12

min: 5
Q1: 7.5
med: 8.5
Q3: 10
max: 12







Box and Whisker plots allow us to get a good visual of outliers: a number that makes one of the whiskers noticeably longer than the box:

RULE OF THUMB: a number is considered an outlier if it is more than $\boxed{1.5 \times IQR}$ below Q_1 or above Q_3

$Q_3 - Q_1$

Is 61 an outlier in Roger Maris's home run data?

Five number summary = {5, 11, 19.5, 30.5, 61}

\# Q_1 Q_3

outlier yes!

$1.5(IQR)$

$IQR = 30.5 - 11 = 19.5$

$1.5 \cdot 19.5 = 29.25$

$30.5 + 29.25 = 59.75$

$Q_3 + 1.5 \cdot IQR =$

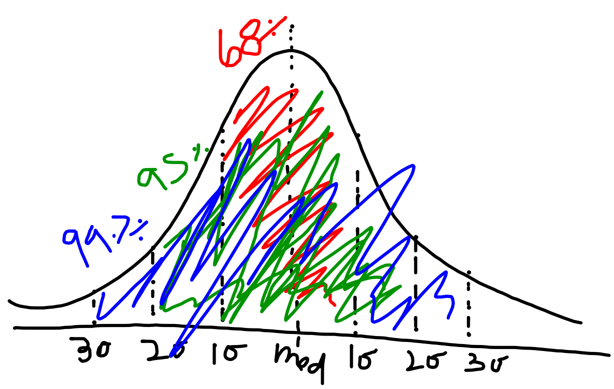
68-95-99.7 Rule

If the data for a population are normally distributed with mean μ and standard deviation σ then,

68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$

95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$

99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$



Would a loon chick weighing 95 grams be in the top 2.5%?

79.5	87.5	88.5	89.2	91.6	84.5	82.1	82.3	85.1	89.8
84.0	84.8	88.2	88.2	82.9	89.8	89.2	94.1	88.0	91.1
91.8	87.0	87.7	88.0	85.4	94.4	91.3	86.3	85.7	86.0

Survey Design: the goal of a survey is to get a sample which accurately reflects the entire population

Bias is a systematic favoritism for a certain outcome

We avoid bias by getting a simple random sample - all subjects have the same chance of being selected to be surveyed

Other sources of bias:

1. Nonresponse: subjects to not respond to the survey

2. Undercoverage: a portion of the population with some commonality is excluded from the survey

3. Voluntary response: the sample chooses itself by responding to a general appeal

4. Response bias: systematic difference between subject's response and the "truth" (i.e. lying)

Observational Study: a study that observes individuals and measures variables of interest, but does not attempt to influence responses. Cause - and - Effect cannot be proven from an observational study, only from a:

Controlled experiment: has 3 parts

1. Random assignment of subjects

2. Treatment groups where treatments are applied

3. Comparison of the outcomes