Module 2

Rafael Garcia

Outline

- Descriptive Statistics as Basic Models
- Frequency distributions
  - What they are
  - How to get them
- Basic models
  - Central tendencies
  - Spread and deviation
  - Skew and Kurtosis

DESCRIPTIVE STATISTICS AS BASIC MODELS

Descriptive statistics

- Statistical methods and equations that SOS.
  - Summarize
  - Organize
  - Simplify
- Takes raw data and generates basic models of the data
  - Graphically
  - Statistically

FREQUENCY DISTRIBUTIONS

What are they

- Listing of possible values for a variable, together with the number of observations at each value
  - i.e., category labels with the number of occurrences (frequency) in each category
- The distribution itself can be shown
  - Graphically
    - Tables, Histograms, Stem-and-Leaf plots, and Boxplots (typically)
  - Statistically
    - Measures of central tendency and spread

78.41007	97.46329	93.94233	100.8617	84.04854	97.56548
94.8594	96.78311	120.1307	93.93054	81.1404	95.25304
116.2395	77.28616	119.5105	106.7899	118.3557	108.107
104.6798	103.1343	107.3367	96.39819	89.99162	93.22944
98.59336	113.1344	93.21332	102.8965	98.06298	84.85477
113.9224	119.5482	90.43924	83.27477	103.4965	110.5043
92.32983	96.61805	82.35515	97.55055	97.20157	107.2146
103.3935	94.21196	85.73752	117.6113	96.0238	112.6415
108.1431	99.00023	104.5972	119.7462	96.38129	96.77553
96.67677	99.406	106.4173	96.56493	103.8133	89.62429
91.10055	106.9746	102.0897	97.10971	91.23302	107.8133
101.9511	112.0304	88.84668	90.1707	112.973	88.93628
104.0345	100.3541	88.86743	89.46928	117.6127	93.06964
97.66504	97.90227	103.9001	84.80487	90.53584	112.045
89.39691	110.7899	104.5614	92.41307	106.9121	80.5312
92.98936	105.2603	94.04338	111.6796	86.8331	109.1603
111.7798	98.72292	95.02044	92.49184	160	

X
150-160
140-150
130-140
120-130
110-120
100-110
90-100
80-90
70-80

78.41007	97.46329	93.94233	100.8617	84.04854	97.56548
94.8594	96.78311	120.1307	93.93054	81.1404	95.25304
116.2395	77.28616	119.5105	106.7899	118.3557	108.107
104.6798	103.1343	107.3367	96.39819	89.99162	93.22944
98.59336	113.1344	93.21332	102.8965	98.06298	84.85477
113.9224	119.5482	90.43924	83.27477	103.4965	110.5043
92.32983	96.61805	82.35515	97.55055	97.20157	107.2146
103.3935	94.21196	85.73752	117.6113	96.0238	112.6415
108.1431	99.00023	104.5972	119.7462	96.38129	96.77553
96.67677	99.406	106.4173	96.56493	103.8133	89.62429
91.10055	106.9746	102.0897	97.10971	91.23302	107.8133
101.9511	112.0304	88.84668	90.1707	112.973	88.93628
104.0345	100.3541	88.86743	89.46928	117.6127	93.06964
97.66504	97.90227	103.9001	84.80487	90.53584	112.045
89.39691	110.7899	104.5614	92.41307	106.9121	80.5312
92.98936	105.2603	94.04338	111.6796	86.8331	109.1603
111.7798	98.72292	95.02044	92.49184	160	

X	f
150-160	1
140-150	0
130-140	0
120-130	1
110-120	17
100-110	25
90-100	39
80-90	16
70-80	2

X	f	cf
150-160	1	101
140-150	0	100
130-140	0	100
120-130	1	100
110-120	17	99
100-110	25	82
90-100	39	57
80-90	16	18
70-80	2	2

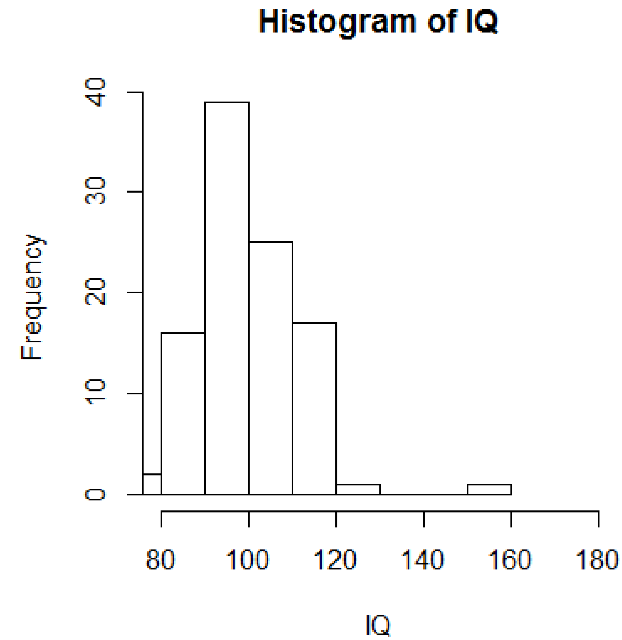
X	f	cf	p
150-160	1	101	0.01
140-150	0	100	0.00
130-140	0	100	0.00
120-130	1	100	0.01
110-120	17	99	0.17
100-110	25	82	0.25
90-100	39	57	0.39
80-90	16	18	0.16
70-80	2	2	0.02

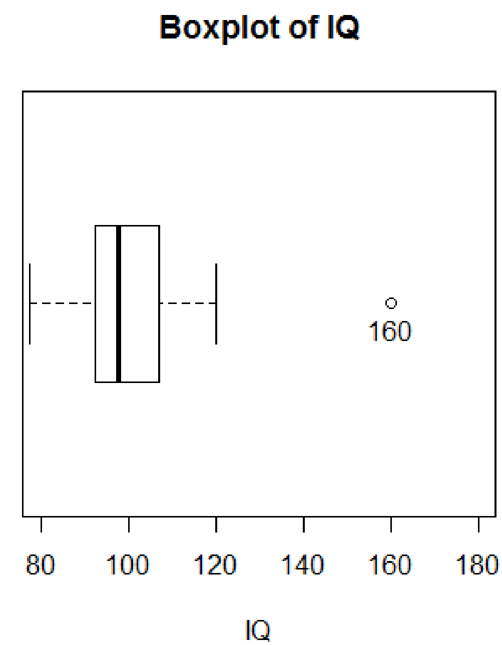
X	f	cf	p	cp
150-160	1	101	0.01	1.00
140-150	0	100	0.00	0.99
130-140	0	100	0.00	0.99
120-130	1	100	0.01	0.99
110-120	17	99	0.17	0.98
100-110	25	82	0.25	0.81
90-100	39	57	0.39	0.56
80-90	16	18	0.16	0.18
70-80	2	2	0.02	0.02

X	f	cf	p	cp	C%
150-160	1	101	0.01	1.00	100
140-150	0	100	0.00	0.99	99
130-140	0	100	0.00	0.99	99
120-130	1	100	0.01	0.99	99
110-120	17	99	0.17	0.98	98
100-110	25	82	0.25	0.81	81
90-100	39	57	0.39	0.56	56
80-90	16	18	0.16	0.18	18
70-80	2	2	0.02	0.02	2

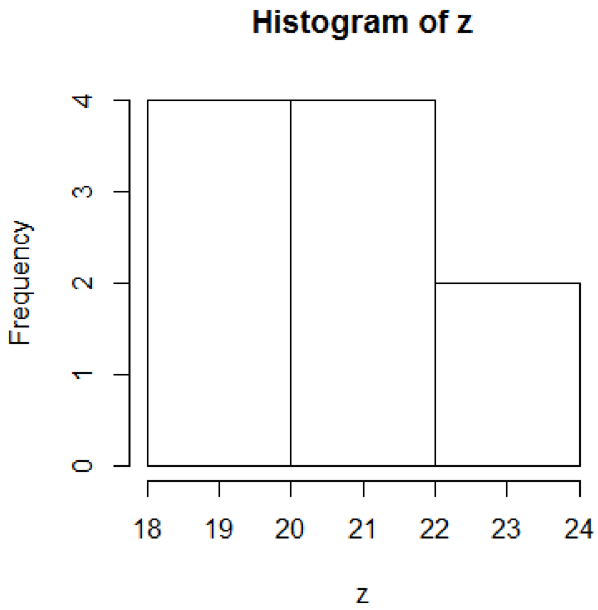
Stem-and-Leaf Plot
IQ scores

```
 7 | 78
 8 | 11234
 8 | 556799999
 9 | 0000111222333334444
 9 | 5556667777777788888889999
10 | 0122333444
10 | 5555677777778889
11 | 112222334
11 | 6888
12 | 0000
12 |
13 |
13 |
14 |
14 |
15 |
15 |
16 | 0
```





24	19	21	21	21
22	23	20	19	20



X	f	cf	p	cp	C%
23-24	2	10	0.20	1.00	100
21-22	4	8	0.40	0.80	80
19-20	4	4	0.40	0.40	40

BASIC MODELS

Central tendency

- Because the data ‘centers’ around them
- These are the simplest statistical models that I know of
  - **Median-** the score that falls in the middle of an ordered list
  - **Mode-** the most frequently occurring score
  - **Mean-** the arithmetic average score

- Let’s take each now, in turn

Median

- Center score of the numerically ordered scores(50% > median >50%)

- If n is odd, median is the [(n+1)/2]th term
  – Ex: 10,11,12,13,14   n=5   [(5+1)/2]=3rd term = 12

- If n is even, median is in between the two middle scores (still the [(n+1)/2]th term!)
  – Ex: 10, 11, 13, 14   n=4   [(4+1)/2]= 2.5th term

    Between 2nd and 3rd terms → (11+13)/2 = 12

Mode

- Most frequent value
  - Count and look
  - Can be more than one value
    - Unimodal if one, bimodal if two, etc.
  - Frequency distribution tables very helpful here!
  - Ex: 10, 11, 12, 12, 14

10-1

11-1

<u>12-2</u>

14-1

Mean

(most commonly used central tendency)

- (Sum of observations)/(# of observations)
- For given observations: $x_1, x_2, x_3, \ldots, x_n$
  - Where n = sample size
- The mean is $\bar{x}$ also called x-bar
- Equations
  - $\text{x-bar} = \dfrac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$ *OR* $\text{x-bar} = \dfrac{\Sigma x_i}{n}$

Mean (continued)

- Influenced by outliers

- <u>Weighted average</u>
  - The mean of two or more groups (n_1= size for group1, n_2= size for group2)

$$m = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}$$

- Numerator is the sum of all observations, and the denominator is the total sample size

How is the mean a model?

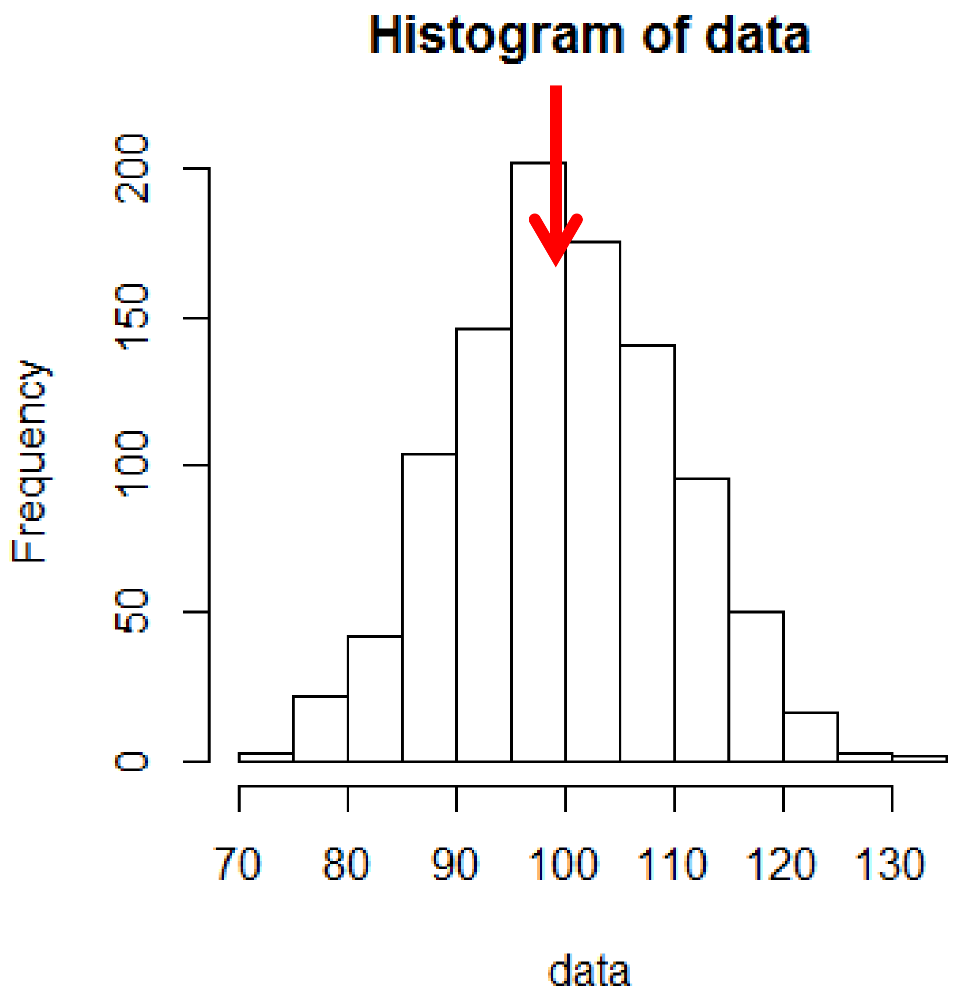
- For each score in the dataset (x_i):
  - $x_i = M + \varepsilon_i$
  - Ex: $x_1, x_2, x_3, x_4, x_5 \Leftrightarrow 2, 3, 4, 5, 6$ n=5

$$(2+3+4+5+6)/5 = 4 = M$$

$x_1 = 2 = 4 + \varepsilon_i$ where $\varepsilon_i = (-2)$
$x_2 = 3 = 4 + \varepsilon_i$ where $\varepsilon_i = (-1)$
$x_3 = 4 = 4 + \varepsilon_i$ where $\varepsilon_i = (0)$
$x_4 = 5 = 4 + \varepsilon_i$ where $\varepsilon_i = (1)$
$x_5 = 6 = 4 + \varepsilon_i$ where $\varepsilon_i = (2)$

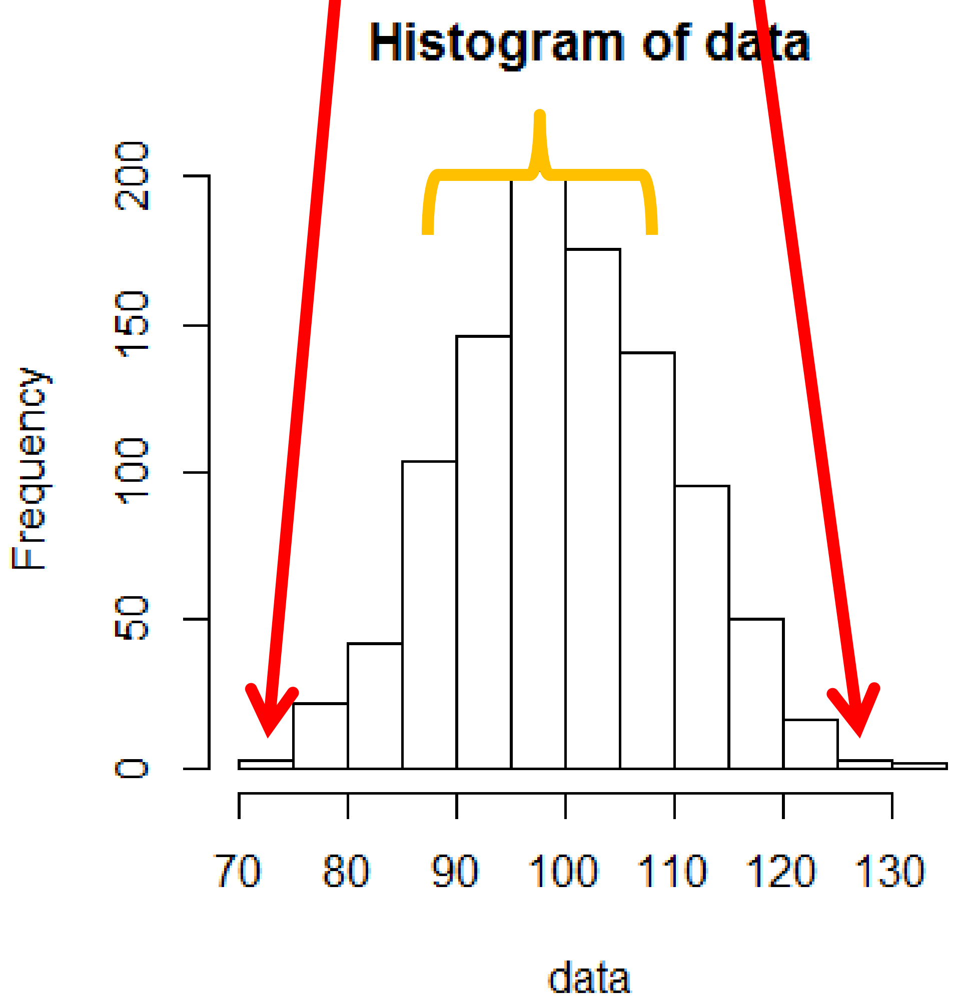
Now what about these ε_i *s?*

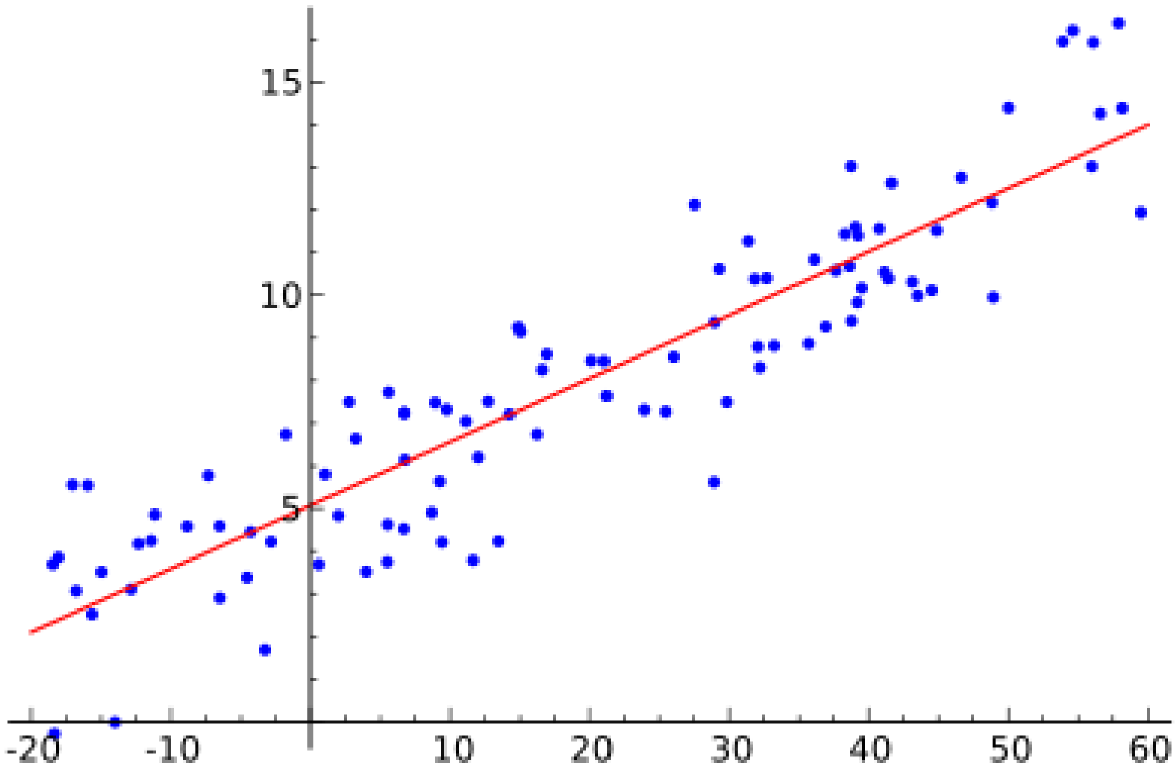
Median	**Mean**
99.60	99.80



But is that all?

Min.	**Max.**
73.41	134.20



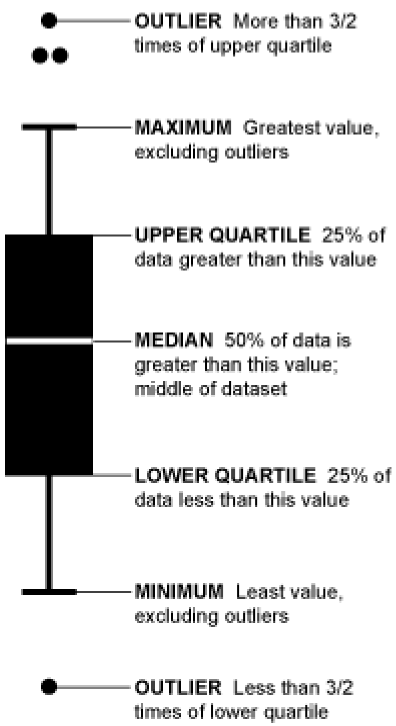


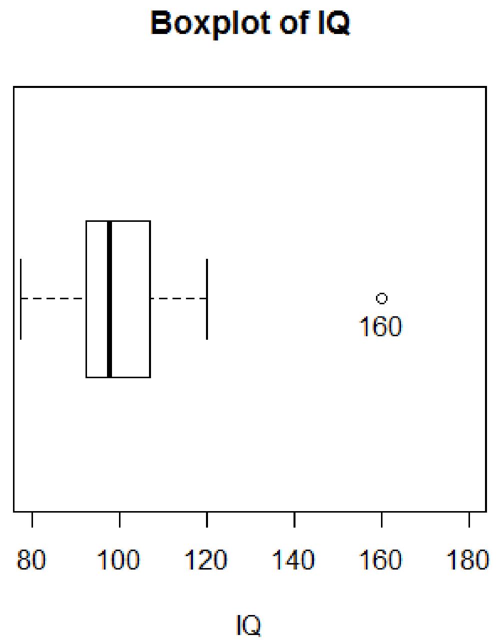
http://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear_regression.svg/400px-Linear_regression.svg.png

Spread and deviation

- Spread
  - This is a vague term I use to talk about a lot of different things all at once
  - **Range** = (Max score – Min score)
  - **Quartiles** = the scores that denote 25%, 50%, 75%, and 100% of scores
  - **Deviation** = $(x_i – M)$
    - Didn't we see this before?
      - $x_i = M + \varepsilon_i \Leftrightarrow \varepsilon_i = x_i – M$

Box plot (revisited)





http://flowingdata.com/wp-content/uploads/2008/02/box-plot-explained.gif

Bring back our example

- We still get M=4
- Σ(Deviation) = 0
  - Does that tell us anything?
    - For the individual scores, sure
    - For the total deviation, not really
- It would be nice to get a sort of 'mean' deviation

Part.	X	X-M
P1	2	-2
P2	3	-1
P3	4	0
P4	5	1
P5	6	2
Σ	**20**	**0?**
n=5	M=4	

Finding a ‘mean’ deviation

- We need a sort of ‘mean’ deviation
- Deviation = $(x_i - M) = (X\text{-}M)$
- Mean Deviation = $\Sigma(X\text{-}M)/(n\text{-}1) = 0/(n\text{-}1)$
- Squared Deviation = $(X\text{-}M)^2$
- Mean Squared Deviation = $\Sigma(X\text{-}M)^2/(n\text{-}1) \neq 0$
  - But this is squared… (messes up units)

Standard deviation

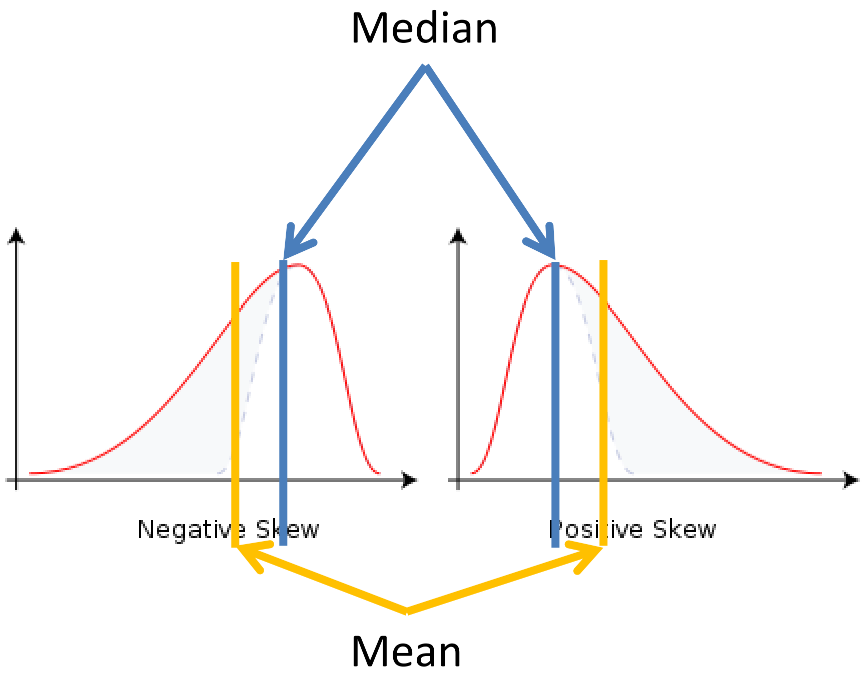
- So take the square root
  - $\sqrt{[\Sigma(X-M)^2/(n-1)]}$
- This is our 'mean' deviation
  - We call it the standard deviation
    - Population parameter= σ
      - $\sqrt{[\Sigma(X-\mu)^2/(N)]}$
    - Sample statistic = s
      - $\sqrt{[\Sigma(X-M)^2/(n-1)]}$

Variance

- If we take the Mean Squared Deviation, we have what we call the variance
  - Population parameter= σ^2
    - $\Sigma(X-\mu)^2/(N)$
  - Sample statistic = s^2
    - $\Sigma(X-M)^2/(n-1)$
  - The numerator is referred to as the Sum of Squares (SS) so another way to write the equation is:
    - SS/N for the population
    - SS/(n-1) for the sample
- Go ahead and ask… “what’s with the (n-1)?”

Skew

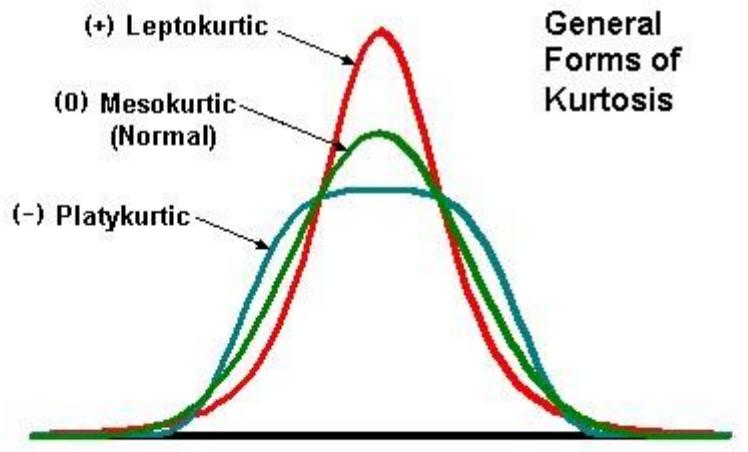
- Three types
  - **No skew**: the mean is on top of the median
  - **Positive**: the mean is to the right of the median
  - **Negative**: the mean is to the left of the median



http://upload.wikimedia.org/wikipedia/commons/thumb/b/b3/Skewness_Statistics.svg/446px-Skewness_Statistics.svg.png

Kurtosis

- Three types
  - **No kurtosis:** aka mesokurtic
  - **Leptokurtic:** tall and thin
  - **Platykutic:** short and flat



http://mvpprograms.com/help/images/KurtosisPict.jpg

Additional resources

- The descriptive statistics chapter of any introductory statistics text
- Descriptive Statistics
  - http://mste.illinois.edu/hill/dstat/dstat.html
  - Podcasts for some of the topics: http://www.discoveringstatistics.com/html/limbo.html
  - Formulas: http://psystats.wikispaces.com/Formulas
- Why (n-1)?
  - http://duramecho.com/Misc/WhyMinusOneInSd.html