Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 4 (MWF) Boxplots and standard deviations

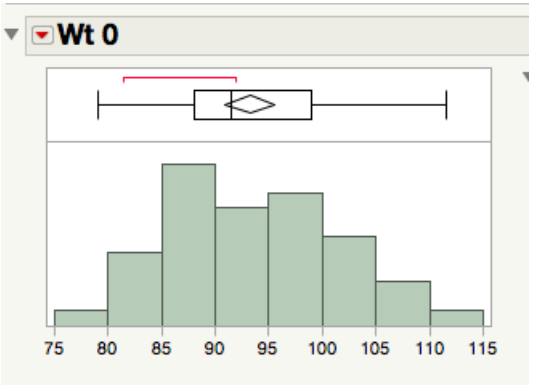
Suhasini Subba Rao

Comparing multiple samples using boxplots

- The quartiles form the basis of the boxplot.

- Boxplots are an excellent graphical tool for comparing multiple samples. It is the "zoomed" out version of the histogram.

- In a boxplot the sides of the box are constructed with the first and third quartiles. The line in the box is the median. Below we give the boxplot of the weights of 44 calves at 0 weeks old.



Remember The true population histogram (density) and boxplot will look different to the one constructed with the sample.

Making comparative boxplots in JMP

- If you want to compare the distribution of several columns with a boxplot, it is easiest if they plotted altogether.

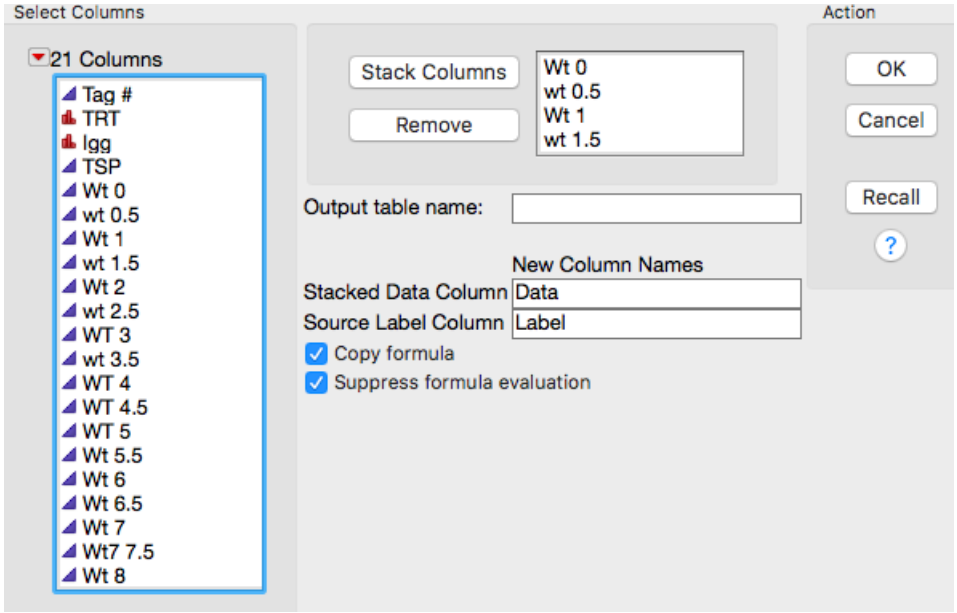
- JMP needs you to turn several columns into one long column, where there is another row containing the factors. As done below

Wt 0	wt 0.5	Wt 1	wt 1.5	Wt 2	wt 2.5	WT 3	wt 3.5	WT 4	WT 4.5
106.5	103	104	100	100	99	102	109.5	112	114
86.5	79	74	69.5	73	71	76	76	79	82.5
79	76	71	71	70	71	79	75	79	82
85.5	82	81	80	83.5	84.5	87	89	92	96
90	83	82	82	80	88	84	87.5	91	95
97	89	89	85	87.5	87	94	94.5	99	97
91	91	89	87.5	86	90	94	103	107.5	107
89.5	87	82	84.5	85	86	94.5	98	100	103
85	82	83	80	81.5	86	90	94	97.5	103
100	95	95	92.5	92.5	96	100	100	105	110
99	95	93	90.5	90	91.5	96.5	96	101	101.5
87.5	80	80	80.5	81	82	86.5	88	93.5	94
111.5	105	106	103	105	103.5	108.5	116	125	125.5
102	100	98	96	100	98	101	106	110	117
99	91	89	82	92	92	95	97.5	105	105
95	89	85	82	86.5	91	98	103.5	110	113.5
90	83	86	85.5	86	86.5	91.5	90	92	95
103.5	98	96	95	100	91	92.5	90	95	96
97	89	97	91	90	92	98	103	108	109

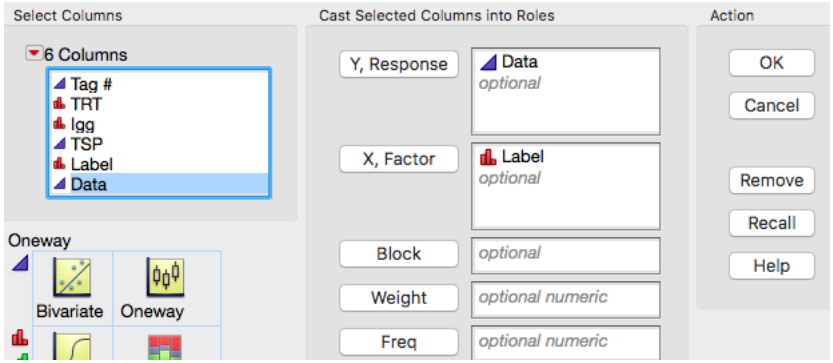
>

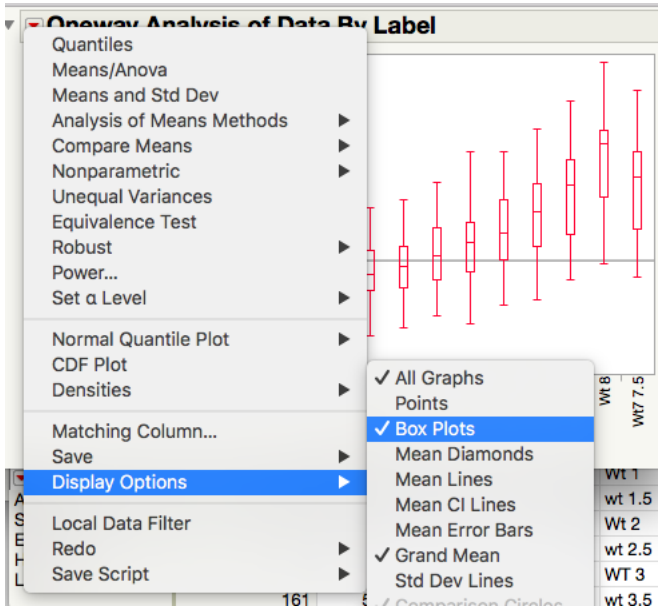
TRT	Igg	TSP	Label	Data
A	F	4.5	Wt 6	113
A	F	4.5	Wt 6.5	117
A	F	4.5	Wt 7	119
A	F	4.5	Wt7 7.5	128
A	F	4.5	Wt 8	130
C	F	4.4	Wt 0	85
C	F	4.4	wt 0.5	82
C	F	4.4	Wt 1	83
C	F	4.4	wt 1.5	80
C	F	4.4	Wt 2	81.5
C	F	4.4	wt 2.5	86
C	F	4.4	WT 3	90
C	F	4.4	wt 3.5	94
C	F	4.4	WT 4	97.5
C	F	4.4	WT 4.5	103
C	F	4.4	WT 5	110
C	F	4.4	Wt 5.5	114

- To turn several columns into one long column go Table > Stack. Put all the columns you would like to stack into Stack column and press okay.



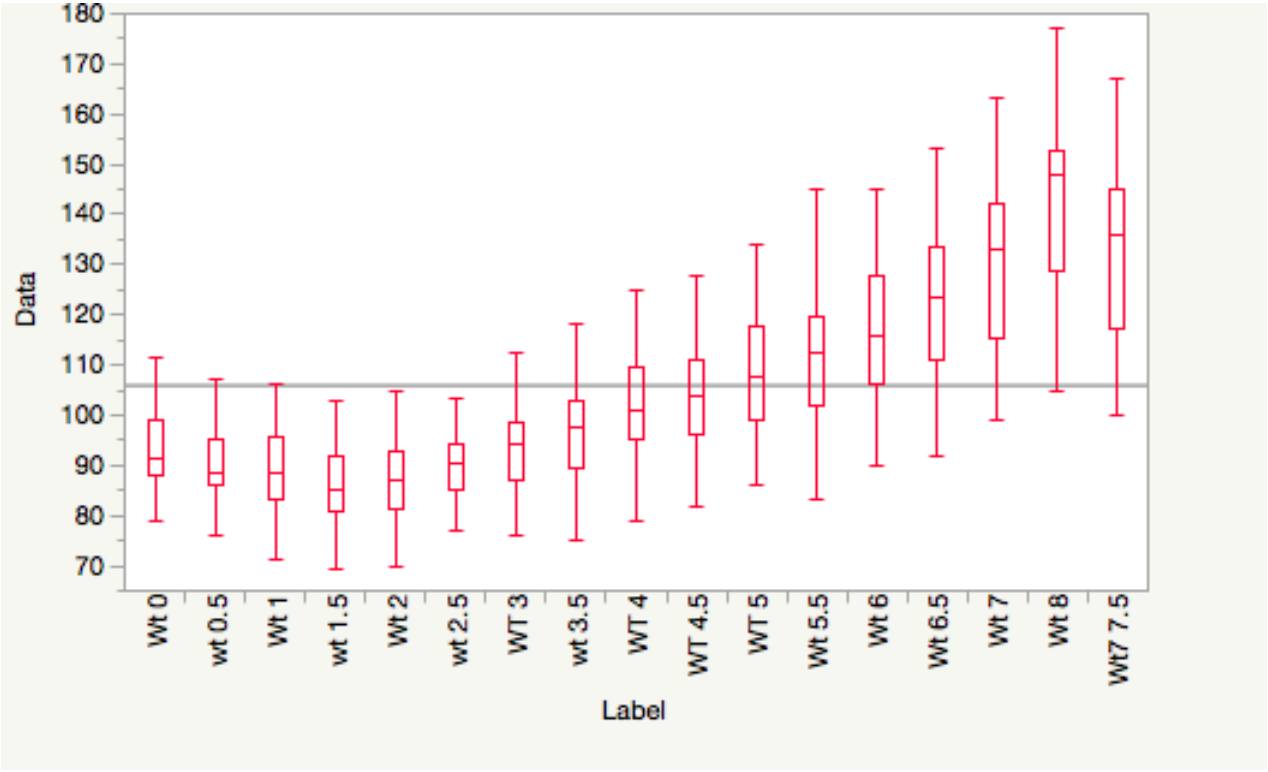
- To make the multiple Boxplots. Go to Analyze > Fit Y by X


Select Columns
Cast Selected Columns into Roles
Action
6 Columns
Tag #
TRT
Igg
TSP
Label
Data
Y, Response
Data
optional
X, Factor
Label
optional
Block
optional
Weight
optional numeric
Freq
optional numeric
OK
Cancel
Remove
Recall
Help
Oneway
Bivariate
Oneway


Oneway Analysis of Data By Label
Quantiles
Means/Anova
Means and Std Dev
Analysis of Means Methods
Compare Means
Nonparametric
Unequal Variances
Equivalence Test
Robust
Power...
Set α Level
Normal Quantile Plot
CDF Plot
Densities
Matching Column...
Save
Display Options
Local Data Filter
Redo
Save Script
All Graphs
Points
Box Plots
Mean Diamonds
Mean Lines
Mean CI Lines
Mean Error Bars
Grand Mean
Std Dev Lines
Wt 8
WT7 7.5
wt 1.5
Wt 2
wt 2.5
WT 3
wt 3.5

A comparison of Calf weights over the weeks

Below we give a boxplot of the 44 calf weight from birth to 8 weeks old. What does the data suggest about how the weights change over time?



Returning to heights and spread

	Sample 1	Sample 2	Sample 3	Sample 4
	68	65	67	69
	74	62	65	62
	68	60	64	71
	61	66	68	72
	61	66	65	66
Average	66.4	63.8	65.8	68

- Recall that the spread of the individual heights is far greater compared with the spread of the sample averages.

- This motivates the definition of variance.

Measures of spread - The Variance

- The variance and standard deviation is a commonly used measure of spread. Given the population $x_1, \ldots, x_N$ we define the population variance as

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_i - \mu)^2, \quad \mu = \text{ population mean.}$$

- In words: this is the sum of all the squared differences between all possible outcomes and the population mean.

- In reality it will never be observed (since the population is unknown) and it has to be estimated instead.

- Given the sample $X_1, \ldots, X_n$ we define the sample variance as

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

given the sample we not know the mean μ, so we estimate it using the sample mean $\bar{X}$ (swop μ with $\bar{X}$).

The reason we divide by $(n-1)$ and not n when obtaining the sample variance is a mathematical quirk to reduce bias. Do not worry about it.

Example - the variance (by hand)

We are given the sample

$5, 4, 3, 0, 3.$

The sample mean $\bar{x}$ is 3. Calculating the sample variance:

x_i	frequency $x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	$5 - 3 = 2$	$(5-3)^2 = 4$
4	$4 - 3 = 1$	$(4-3)^2 = 1$
3	$3 - 3 = 0$	$(3-3)^2 = 0$
0	$0 - 3 = -3$	$(0-3)^2 = 9$
3	$3 - 3 = 0$	$(3-3)^2 = 0$
		sum $= 14$

The sample variance is

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{14}{5-1} = \frac{14}{4}.$$

The variance does not change with shifts

- The variance is invariant to shifts in the data. Suppose we shift the data by 20 (this could be because of a change in units of observations), so we observe $25, 24, 23, 20, 23$. The mean is shifted by 20 (it is $20 + 3 = 23$), but the spread stays the same:

x_i	frequency $x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i	frequency $x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	$5 - 3 = 2$	$(5-3)^2 = 4$	25	$25 - 23 = 2$	$(25-23)^2 = 4$
4	$4 - 3 = 1$	$(4-3)^2 = 1$	24	$24 - 23 = 1$	$(24-23)^2 = 1$
3	$3 - 3 = 0$	$(3-3)^2 = 0$	23	$23 - 23 = 0$	$(23-23)^2 = 0$
0	$0 - 3 = -3$	$(0-3)^2 = 9$	20	$20 - 23 = -3$	$(20-23)^2 = 9$
3	$3 - 3 = 0$	$(3-3)^2 = 0$	23	$23 - 23 = 0$	$(23-23)^2 = 0$
		sum $= 14$			sum $= 14$

- The example illustrates that the variance is simply measuring the **squared** distance from the mean (and is invariant to shift transformations).

The variance does not measure distance..

- Example Consider the observations $-4, -3, -2, -1, 1, 2, 3, 4$. The mean is zero and the sample variance is $\frac{1}{8-1}(4^2+3^2+2^2+1^2+1^2+2^2+3^2+4^2) = 8.57$

- The range of the data is $-4, 4$, which has length 8. The sample variance $= 8.57$, is larger than the range itself!

- This is because the variance **squares** the distances. To overcome this we square root the variance, to give the standard deviation. This is a true measure of average distance.

...but the standard deviation does

- The standard deviation is the square root of the variance. That is

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_i - \mu)^2} \qquad N = \text{size of population}$$

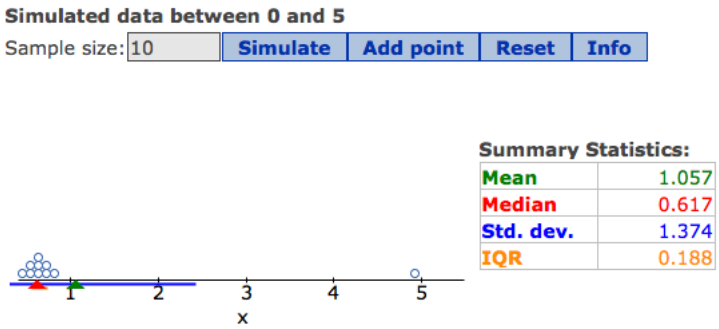
- The sample standard deviation is

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad n = \text{size of sample}$$

- The standard deviation has the advantage that it has the same set of units as the data.

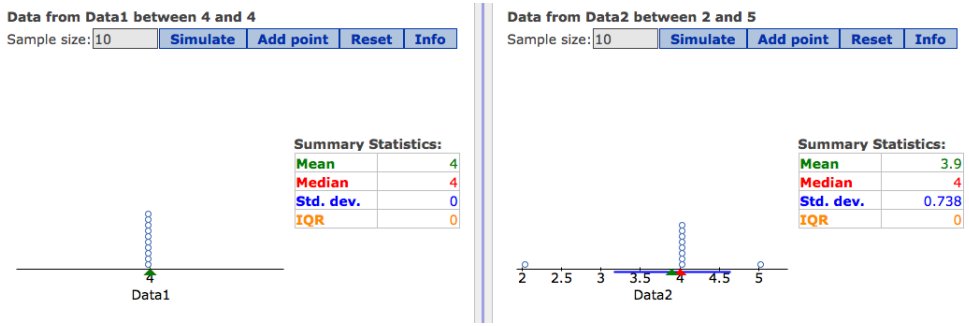
Comparing standard deviations and IQR

- For the majority of the analysis we do in this class, we be using the standard deviation, but it is useful to understand how outliers may influence the standard deviation:



- Observe that the majority of the points are about zero but one value is 5. This pulls the mean and the standard deviation to the right. However, it does not effect the median or IQR.

Standard deviation and IQR



- Here we illustrate the differences between the two measures of spread; IQR and standard deviations.
  - We observe if all the data takes the same value, the IQR and standard deviation are both zero. The standard deviation is only zero if all the data is the same.
  - However, we see on the second plot that the IQR can still be zero if most of the values (but not all) are the same.

Summary: Differences in the IQR and the standard deviation

- The standard deviation is only zero if all observations (numbers) take the same value.

- The IQR is zero when the first and third quartiles are the same. This does not mean all the observations are zero.

- Both the IQR and standard deviation are always non-negative.

The empirical rule (this just a rule of thumb)

A general rule of thumb of data (which is not always true) is that

- Approximately 68% of the observations lie inside the interval $[\bar{x}-s, \bar{x}+s]$.
- Approximately 95% of the observations lie inside the interval $[\bar{x}-2s, \bar{x}+2s]$.
- Approximately 99.7% of the observations lie inside the interval $[\bar{x}-3s, \bar{x}+3s]$.

This is only a rule of thumb and it only applies to data, which is 'normally' distributed. Don't take it seriously. The take home message is that that the majority of the data will be within three standard deviations of the mean.

Interpretation of the empirical rule

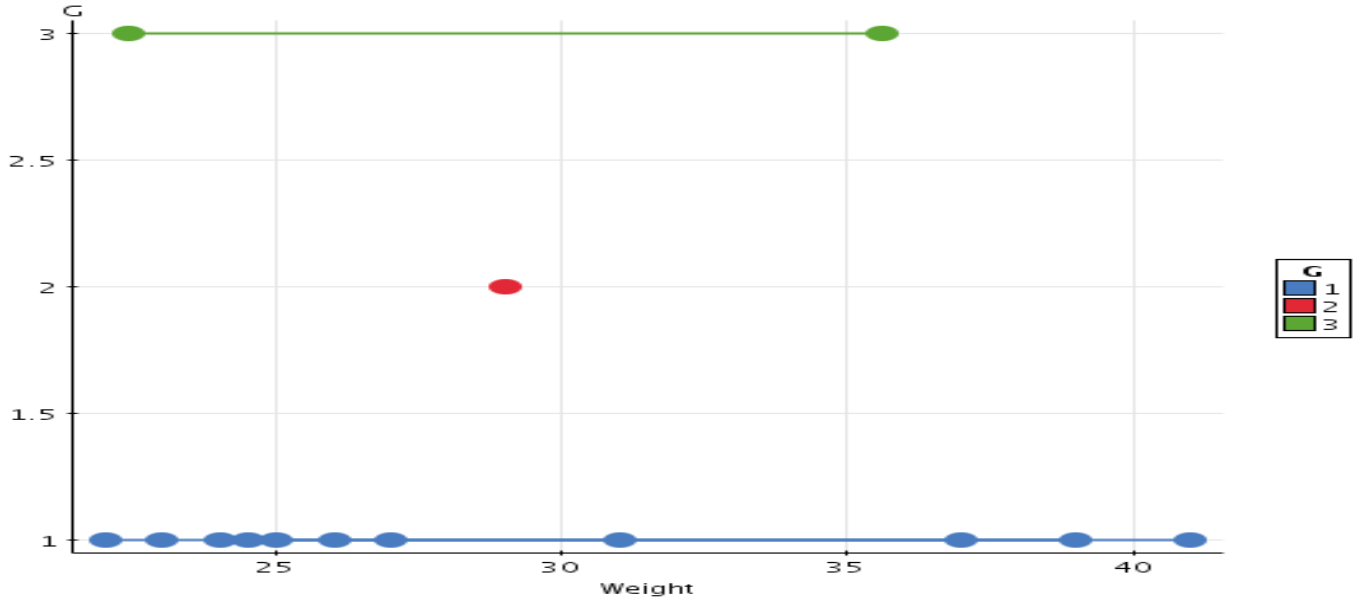
Some maths background:

- What do we mean by the interval $[a, b]$? These are two points of the 'time line'. Starting with a and ending at b. Example, what does the interval $[3, 4.5]$ mean?

- What we mean by ' Approximately 68% of the observations lie inside the interval $[\bar{x} - s, \bar{x} + s]$'? We calculate s and $\bar{X}$ from the data. For example, it could be $\bar{X} = 3$ and $s = 1$. For this example, the interval $[\bar{x} - s, \bar{x} + s]$ is $[3 - 1, 3 + 1]$, and we count the number of observations in this interval. The empirical rule states that for many data sets 68% of the observations lie in this interval.

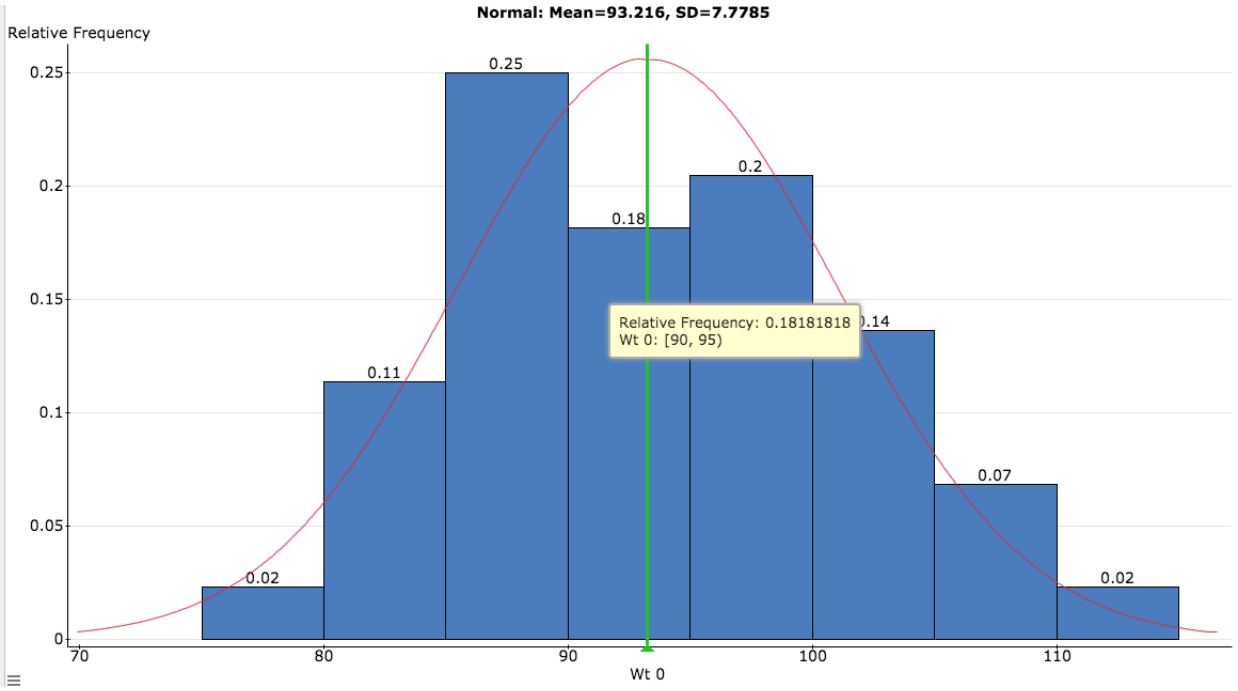
  Similarly $[\bar{x} - 2s, \bar{x} + 2s] = [3 - 2, 3 + 2]$ and $[\bar{x} - 3s, \bar{x} + 3s] = [3 - 3, 3 + 3]$.

Standard deviation: Graphical illustration 1



Mean is 29 and standard deviation is 6.6. The green line $[29 - 6.6, 29 + 6.6] = [22.4, 35.6]$ is one standard deviation from the mean. Observe that $7/11 = 63.3\%$ of the points are within one standard deviation of the mean.

Standard deviation: Graphical illustration 2



Locate, on the histogram above, the sample mean and the interval one sample standard deviation from the sample mean.

Where you may have previously encountered the standard deviation

- When you go to the doctors office, the technicians take several measurements, such as weight, height, blood pressure and blood samples.

- How do they determine whether it is 'normal' (this does not refer to the distribution).

- **Consider the following example**:

- Suppose that you have a blood sample taken. The mean for the reading is 20 and you have 14, is that normal?

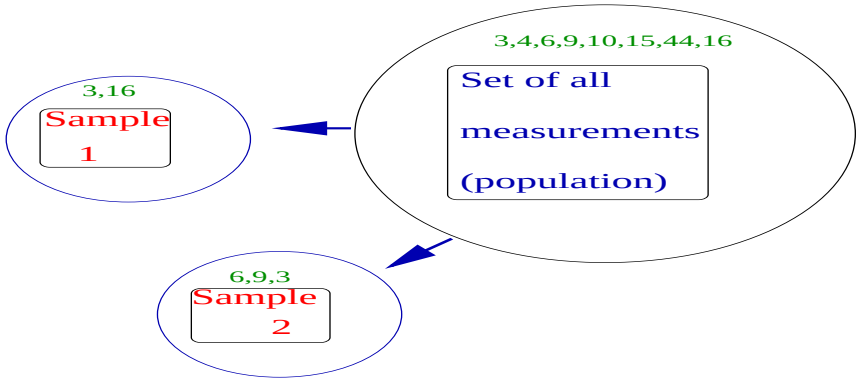
- The *raw* difference of $14 - 20$ is unformative unless you know the general spread of the readings.

- Suppose, the standard deviation is 8. Then you are within $(14-20)/8 = -6/8$ standard deviations of the mean. The reading is "relatively close" to the mean.

- In addition if blood samples followed the empirical rule, since your reading is within one standard deviation of the mean then your reading is 'within' 68% of the mean

- Another example:
  - A friend has their blood sample taken, he has the reading 45. Is that normal?
  - His reading is 25 from the mean and $(45-20)/8 = 3.125$ standard deviations from the mean.
  - His reading is in the far right tail of the distribution. The empirical rule tells us that for many data sets 99.7% of observations are within 3 standard deviations of the mean).

Reminder: population, sample and standard deviation



- Population standard deviation In this case it is

$$\sigma = \sqrt{\tfrac{1}{8}[(3-13.375)^2 + (4-13.375)^2 + \ldots + (16-13.375)^2]} = 12.4.$$

Observations

- Like the sample mean the sample variance is random and varies from sample to sample.

- Sample standard deviation

  For Sample 1: $s_1 = \sqrt{\frac{1}{3-1}[(6-6)^2 + (9-6)^2 + (3-6)^2]} = 3.$

  For Sample 2: $s_2 = \sqrt{\frac{1}{2-1}[(3-9.5)^2 + (16-9.5)^2]} = 9.12.$

- Similar to the sample mean, the sample standard deviation is also random (changes according to the sample).

- The sample standard deviation often underestimates the population standard deviation.

- **Rule for later in the course**

- If prior information means that the population variance is *known* (a rare case). Then inference about the population mean uses the normal distribution (the reason for normal distribution will come later).

- If the population variance is unknown and we use the same data to estimate both the mean and variance. Then we use the t-distribution.

  Changing from a normal distribution to a t-distribution corrects from the underestimation of the sample mean.