"You can, for example, never foretell what any one man will do,

but you can say with precision what an average number will be up to.

Individuals vary, but percentages remain constant.

So says the statistician."

Sherlock Holmes, *The Sign of Four*

"We must understand variation."
W. Edwards Deming, American Statistician, 1900-1993

COHEN CHAP 3. CENTER & SPREAD

For EDUC/PSY 6600

WHAT DO WE MEAN BY **DISTRIBUTION**?

For a **Continuous** variable

- General shape
- Exceptions (outliers)
- Modes (peaks)
- Center & spread (chap 3)
- **Histogram**
- Cumulative polygon or ogive

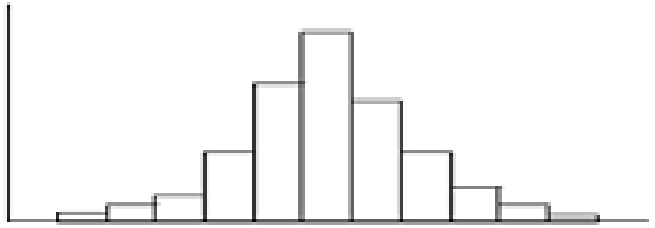
For a **Categorical** variable

- Counts = raw number of ___

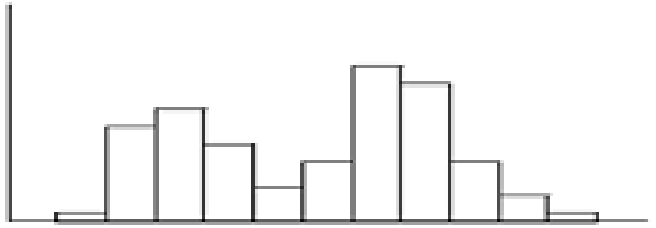
  Percent or Rate - adjusts for an 'out of' to compare
- **Bar chart:** should have space between bars, order?
- Pie chart - avoid!

DISTRIBUTION – EXAMPLES

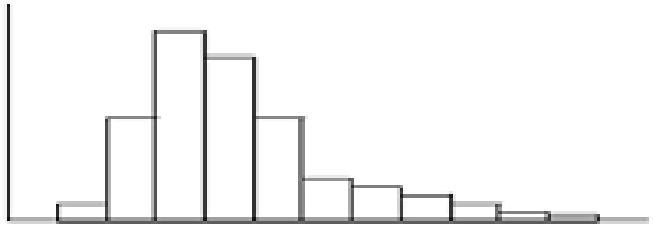
Bell-shaped



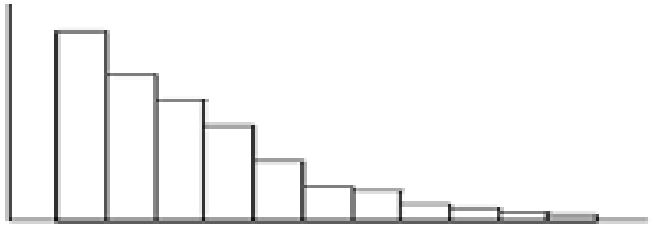
Bimodal



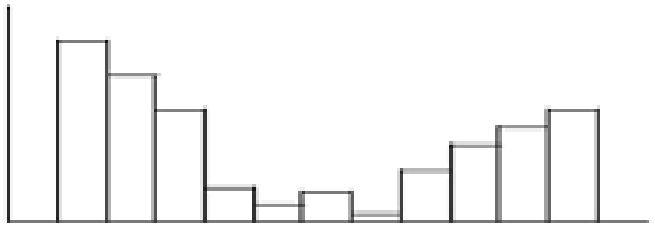
Right-skewed



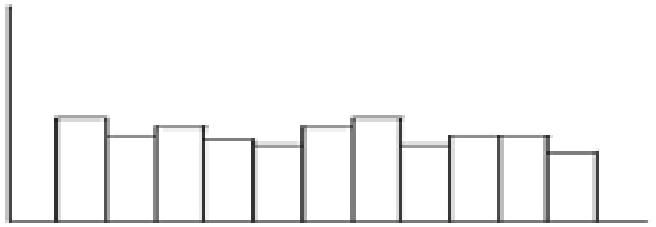
J-shaped



U-shaped



Uniform



THREE MEASURES OF CENTER

Mode

- "Most" common value, largest frequency, highest peak
- Non-uniqueness - can have more than one mode
- Doesn't always represent the 'center'
- Do NOT usually use, other than descriptively

Mean

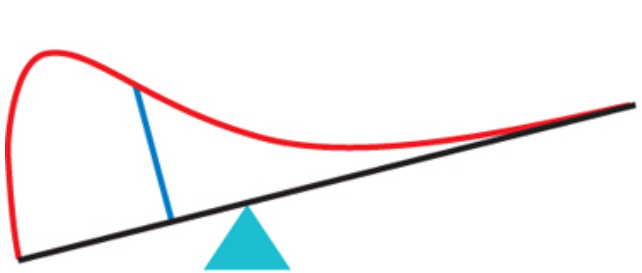
- "*Aritmetic Average*" = add them all up & divide by the count
- Simple to calculate
- Not resistant: easily influenced by extreme values or outliers
- Can do a "trimmed" mean (leave off the most extreme values, like 1% or 5%)
- In a **POPULATION**: "Mu" (μ)
- In a **SAMPLE**: "X-bar" ($\bar{X}$) but APA uses "M" for abbreviation

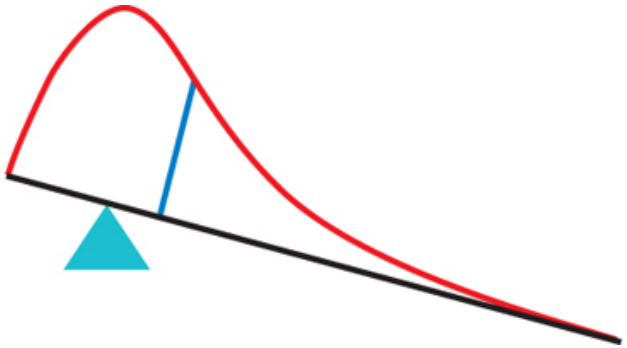
Median

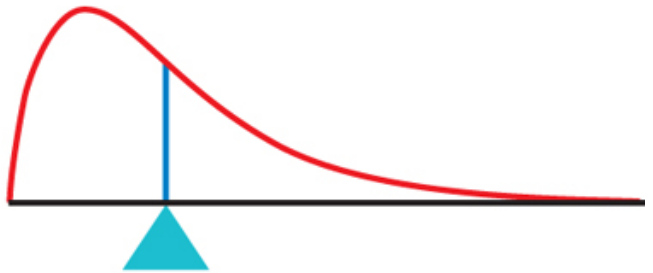
- 50th percentile, APA: "Mdn"
- "Middle" value, when ordered/ranked in increasing order
- ODD #: middle value
- EVEN #: avg of 2 middle
- Half the values are above, and half below
- Easy for a computer to do
- RESISTANT: NOT influenced by a few extreme values or outliers

MEAN VS MEDIAN, WITH PICUTRES

Median = middle value
It's the midpoint
If the histogram was chocolate,
where would you cut it to share?



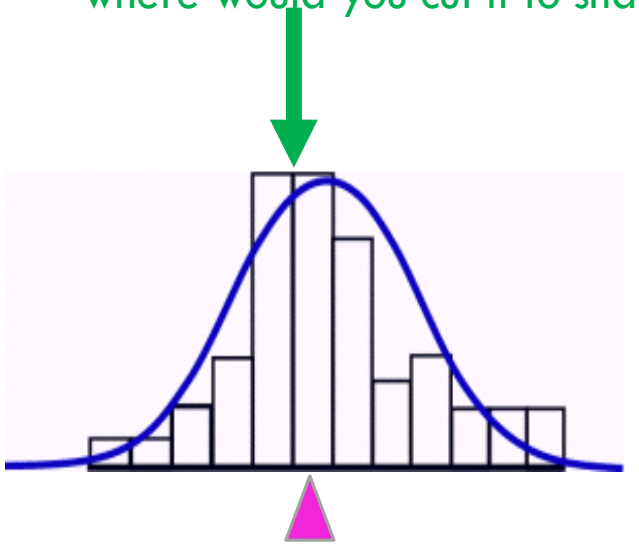




Mean = average, not typical
BUT...$\bar{x}$ is 'pulled' towards extremes
Visually it is the balance point

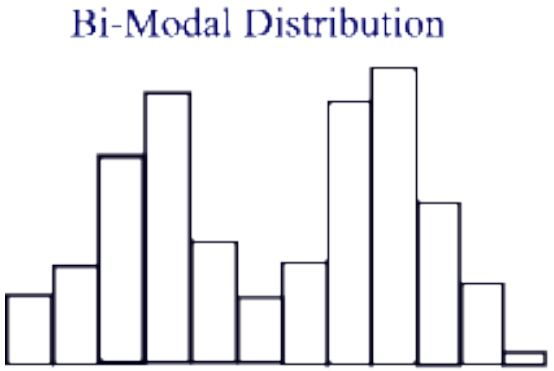
DISPLAYING DISTRIBUTIONS WITH NUMBERS

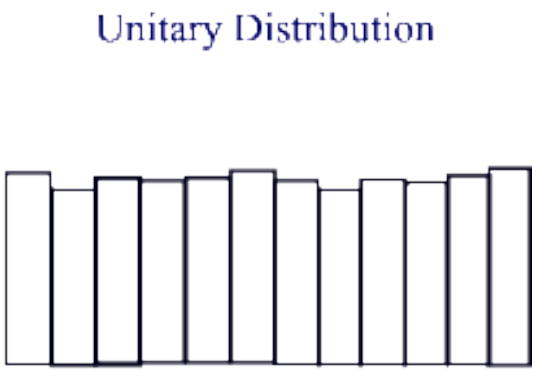
Median = middle value
It's the midpoint
If the histogram was chocolate, where would you cut it to share?

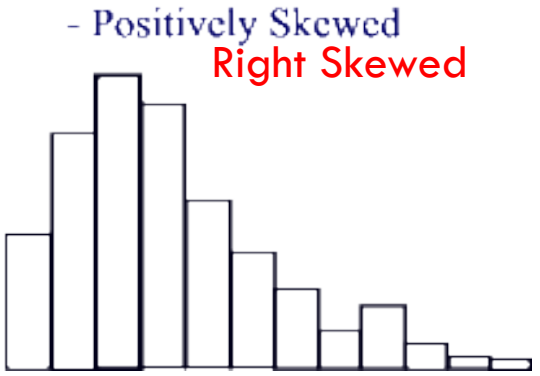


Mean = average, not typical
BUT…$\bar{x}$ is 'pulled' towards extremes
Visually it is the balance point.

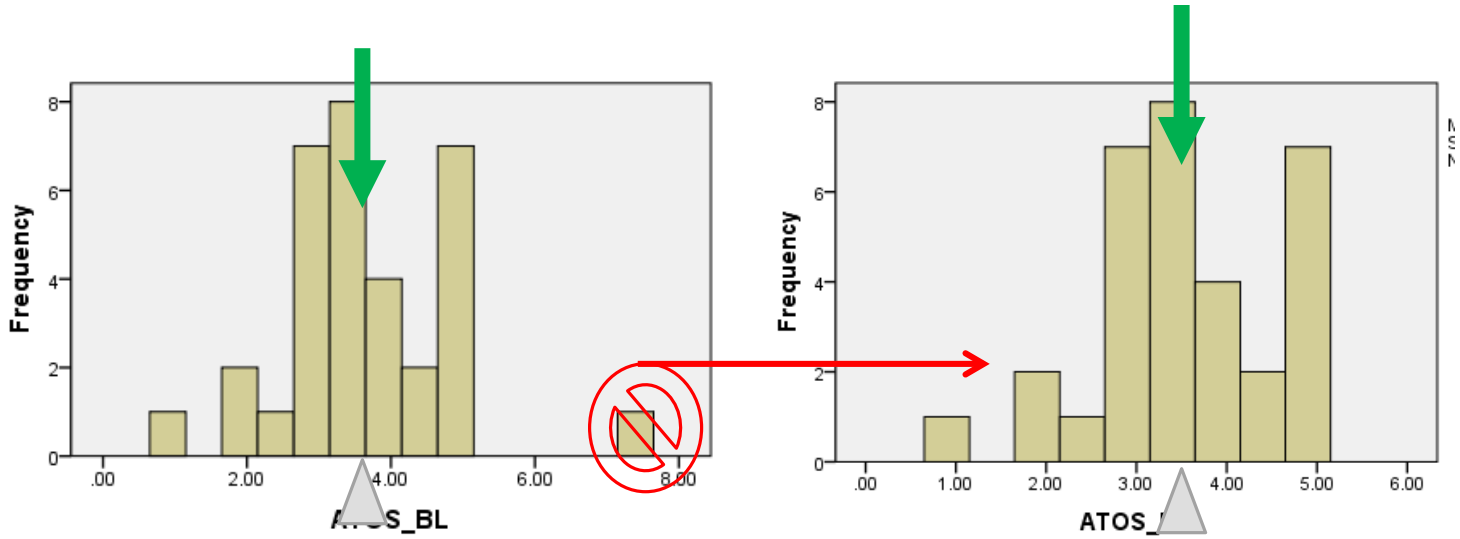
If symetrical → $\bar{x}$ = M







DESCRIBING DISTRIBUTIONS WITH NUMBERS



Median: 3.55 → 3.50

Mean: 3.68 → 3.57

The **Median** is "**resistant**" & doesn't change much

The **Mean** is "**influenced**" & changes more!

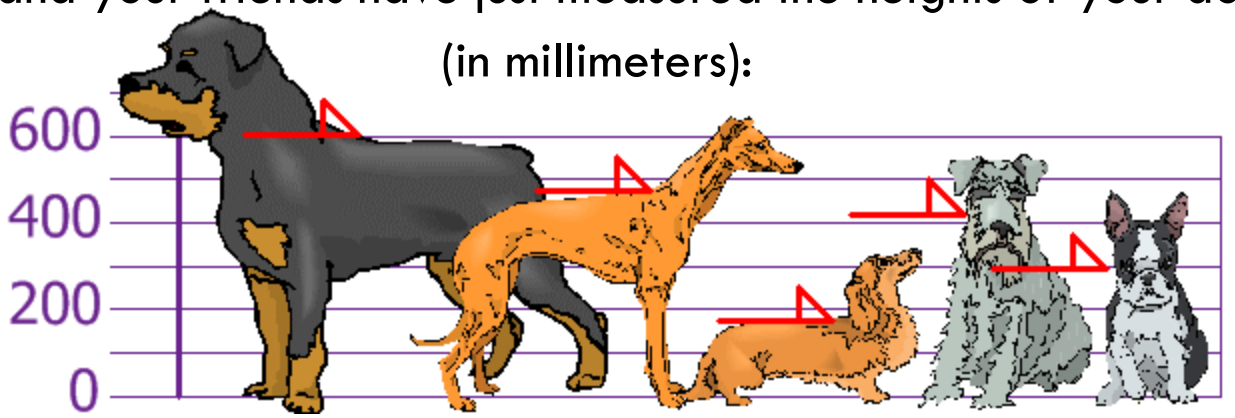
Average does NOT mean typical

DOG EXAMPLE

- Spread : measured by **Variance** and **Standard deviation**

http://www.mathsisfun.com/data/standard-deviation.html

You and your friends have just measured the heights of your dogs (in millimeters):



The heights (at the shoulders) are:

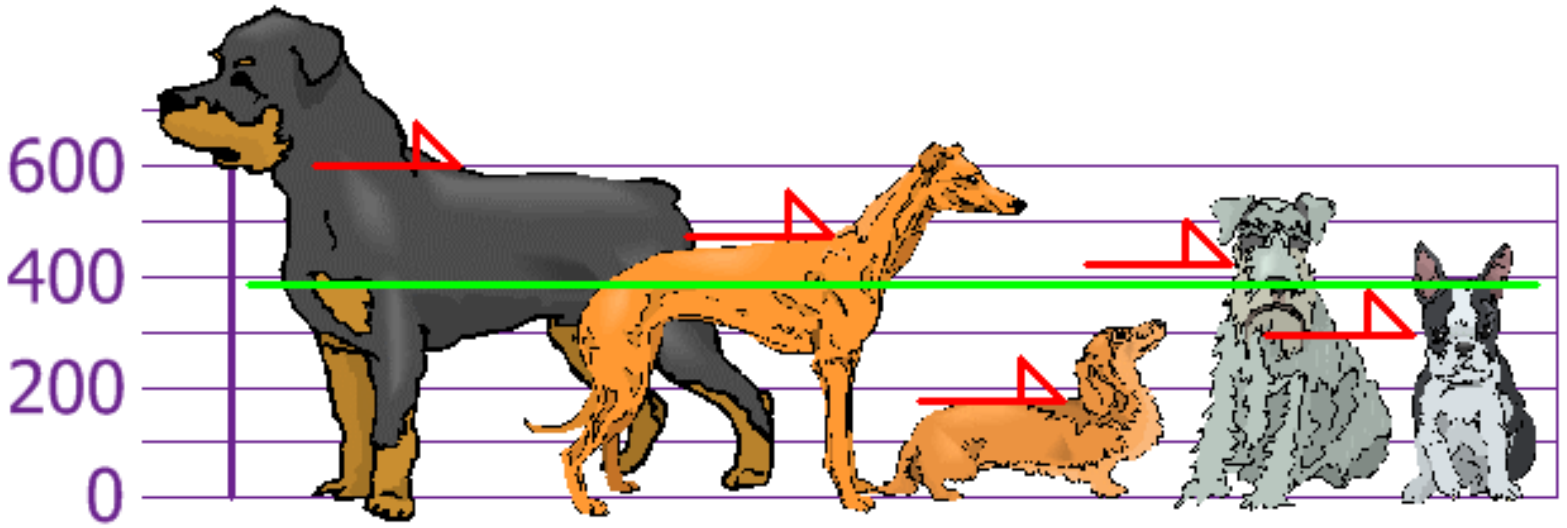
600mm, 470mm, 170mm, 430mm and 300mm.

Find out the **Mean**, the **Variance**, and the **Standard Deviation**.

DOG EXAMPLE: MEAN

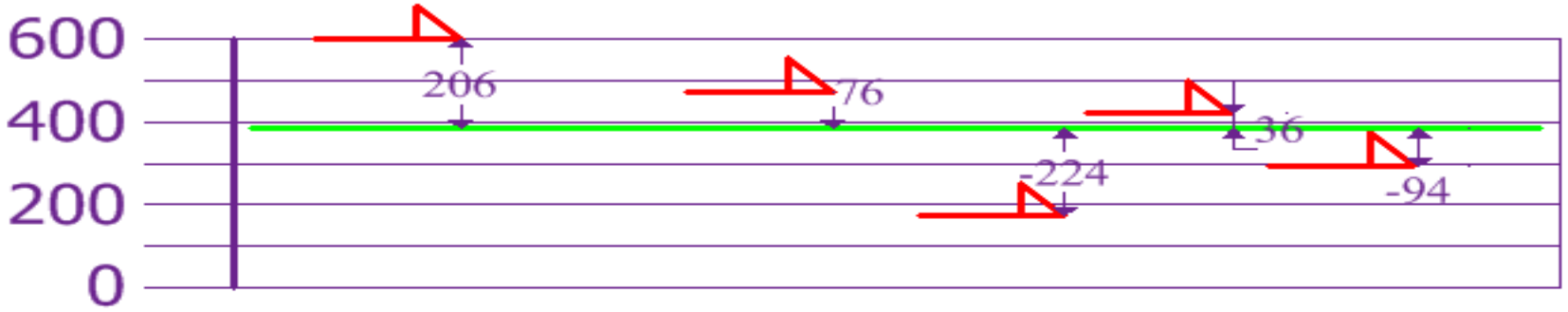
Mean =

so the mean (average) height is mm. Let's plot this on the chart:



DOG EXAMPLE: VARIANCE

Now, we calculate each dogs difference from the Mean:



To calculate the Variance, take each difference, square it, and then average the result:

Variance:

If we didn't square the deviations, the total would always be zero!

So, the Variance is

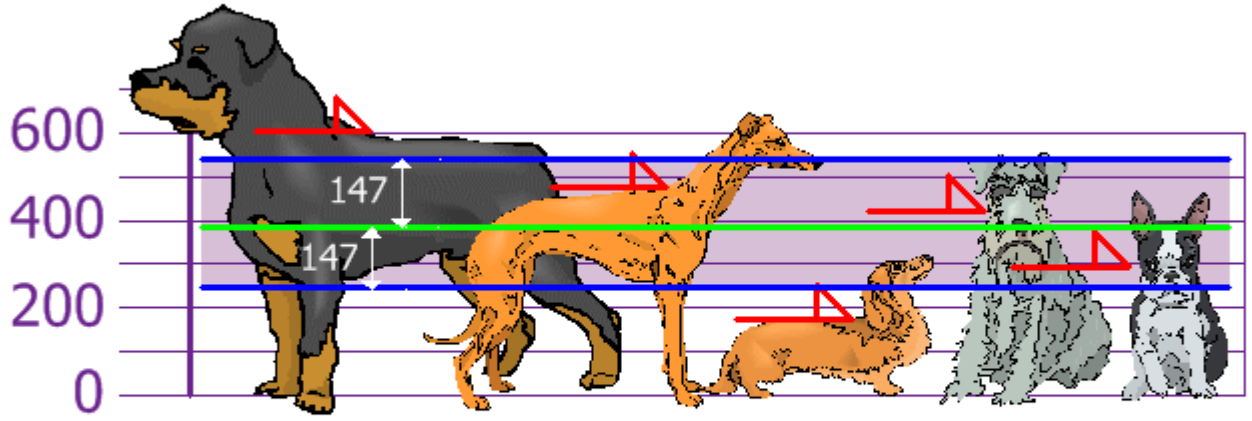
We divide by n-1 because: since we know the deviations all sum up to zero, if we know all but the last one, we can subtract to find it

DOG EXAMPLE: STANDARD DEVIATION

And the Standard Deviation is just the square root of Variance, so:

$$s =$$

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Rottweilers **are** tall dogs. And Dachshunds **are** a bit short ... but don't tell them!

THREE MEASURES OF SPREAD

Sum of Squares

Mean Sum of Squres

Range, IQR, & SIR

- Range = Max – Min
- Interquartile Range
  IQR =
- Semi-Interquartile Range
  SIR =
- Range is super dependent on extreme values or outliers
- IRG & SIR more resistant

Variance

- DEVIANT: how far from the center (mean)
- SQUARE: so + & - don't cancel out to 0
  (units are also squared)
- AVERAGE: summarize with a single value
- In a POPULATION: called "sigma-squared"
- In a SAMPLE: called "s-squared"
- *Degrees of Freedom:*

Standard Deviation

- SQUARE-ROOT VARIANCE to get back to the original units
- In a POPULATION: called "sigma"
- In a SAMPLE: called "s"

BEST SUMMARIES?

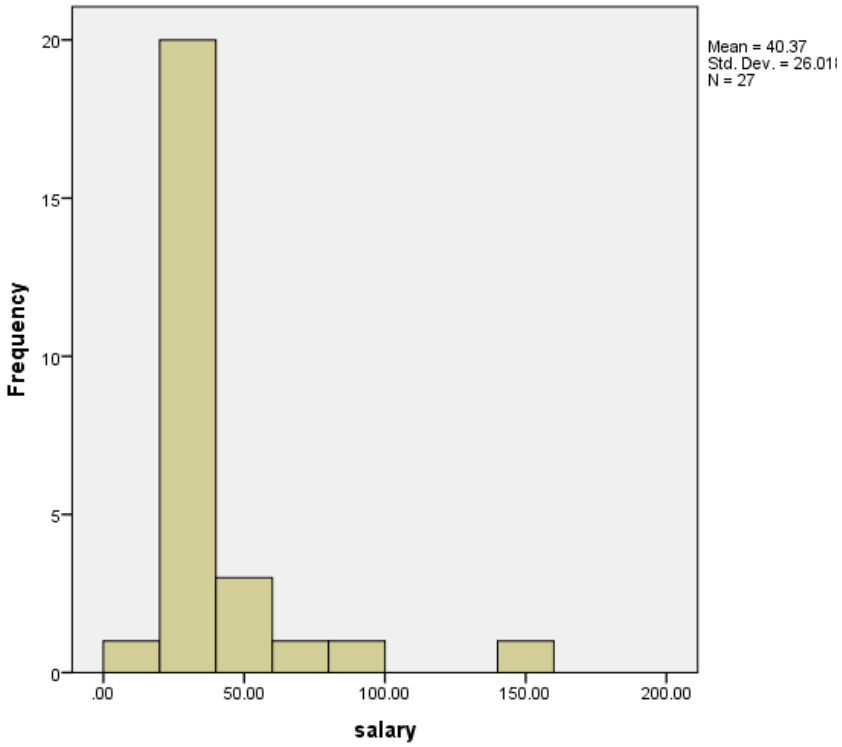
"…the perfect estimator does not exist."
Rand Wilcox, 2001

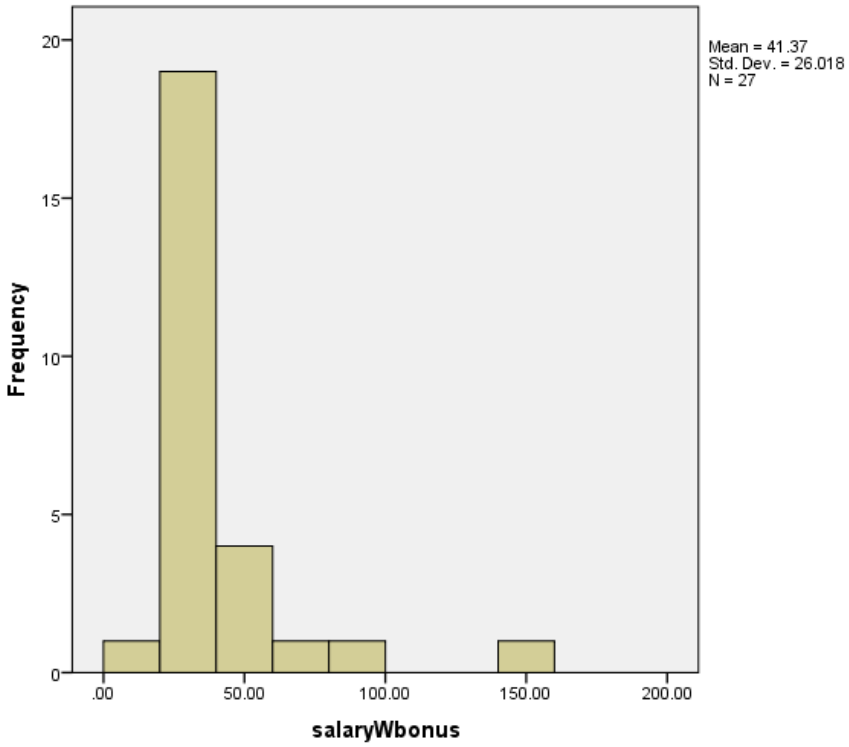
- Which is better to convey a distribution with numbers?
  - Median & SIR
  - Mean & standard deviation
  - !!! A graph gives the best overall picture of a distribution. Number of center/spread convey only some information…

    … ALWAYS PLOT YOUR DATA!!!

ADD A CONSTANT AMOUNT TO EVERYONE

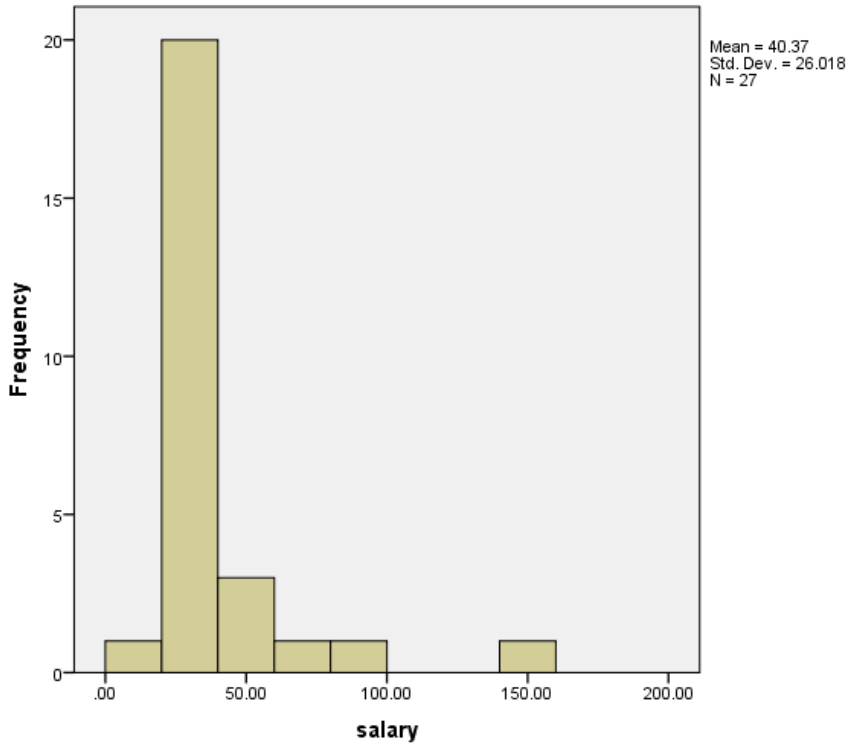
- Give everyone a $1000 bonus (add the same amount)

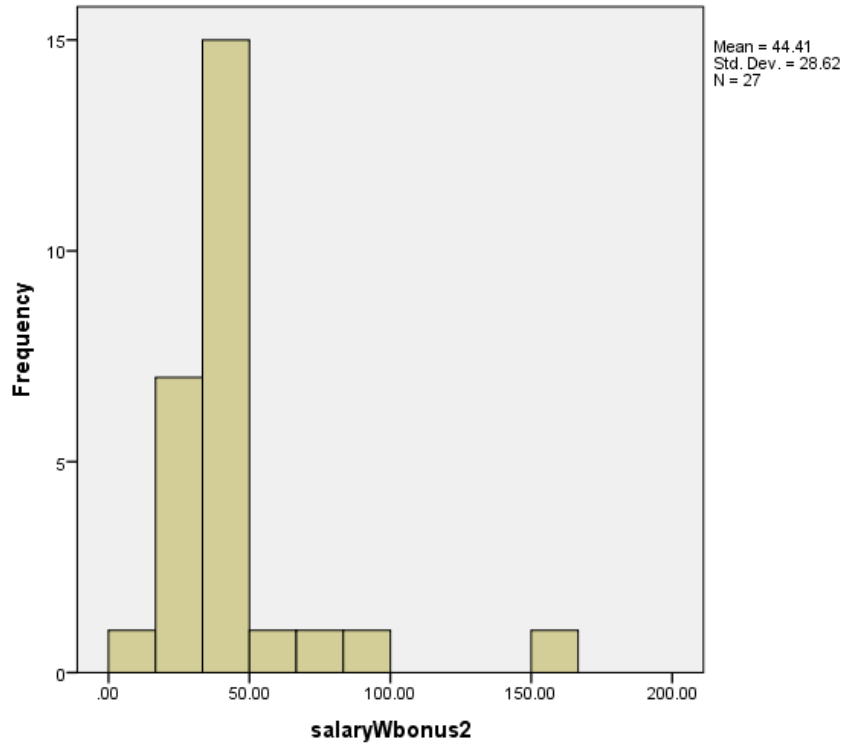




MULTIPLY ALL BY THE SAME CONSTANT

- Give everyone a 10% bonus (multiply by same amount)





PROPERTIES OF THE MEAN & STANDARD DEVIATION

If you _____ the same CONSTANT number Onto every score…	MEAN	STANDARD DEVIATION
ADD (or subtract)		
MULTIPLY (or divide)		

SKEWNESS

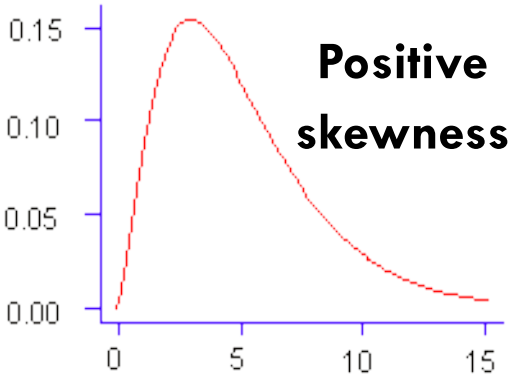
$$\text{Skewness} = \frac{N}{N-2} * \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{(N-1)s^3}$$

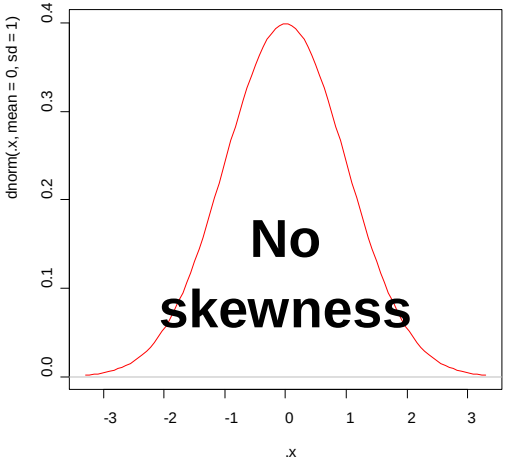
Degree of **symmetry** in distribution

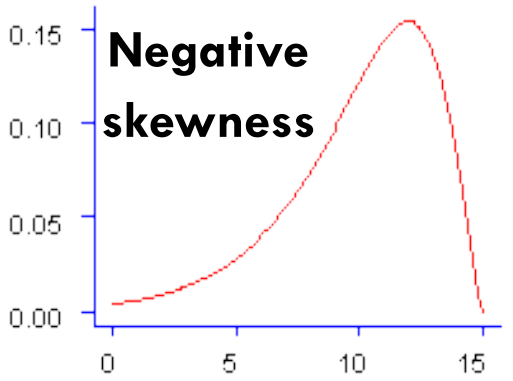
Can detect **visually** (histogram or boxplot)

Skewness statistic
- Based on cubed deviations
  - from *M* (returns statistic to original units)
- Divided by *SE* of skewness
  - > ± 2 indicates likely problem with skewness

Interpreting skewness statistic
- Positive value = Positive (right) skew
- Negative value = Negative (left) skew
- Zero value = No skew







KURTOSIS

$$\text{Kurtosis} = \frac{N(N+1)}{(N-2)(N-3)} * \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{(N-1)s^4} - 3\frac{(N-1)(N-1)}{(N-2)(N-3)}$$

Degree of **flatness** in distribution

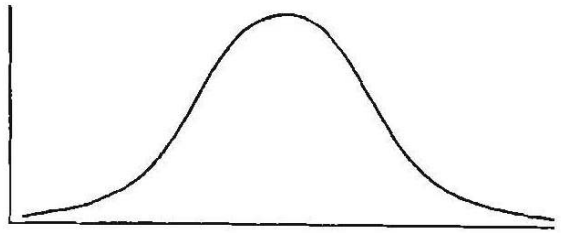
Harder to detect visually

Kurtosis statistic

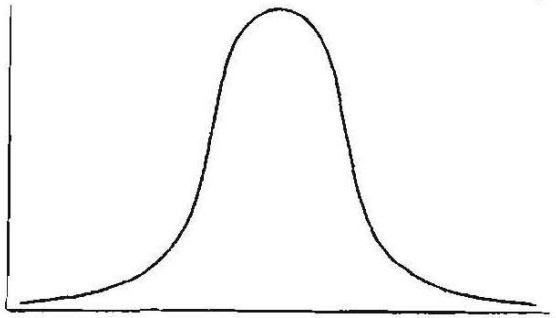
- Based on deviations from *M* raised to 4th power (returns statistic to original units)
- Divided by *SE* of kurtosis
  - > ± 2 indicates likely problem with kurtosis

Interpretation of kurtosis statistic

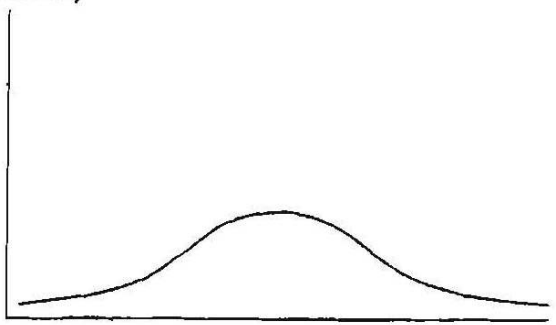
- Positive kurtosis = leptokurtic (peaked)
- Negative kurtosis = platykurtic (flat)
- Zero value = mesokurtic (normal)



Distribution H
(Mesokurtic distribution)



Distribution I
(Leptokurtic distribution)

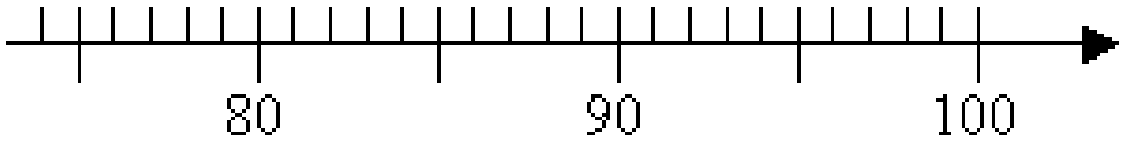


Distribution J
(Platykurtic distribution)

Example of VERY small amount of data (usually lots more)

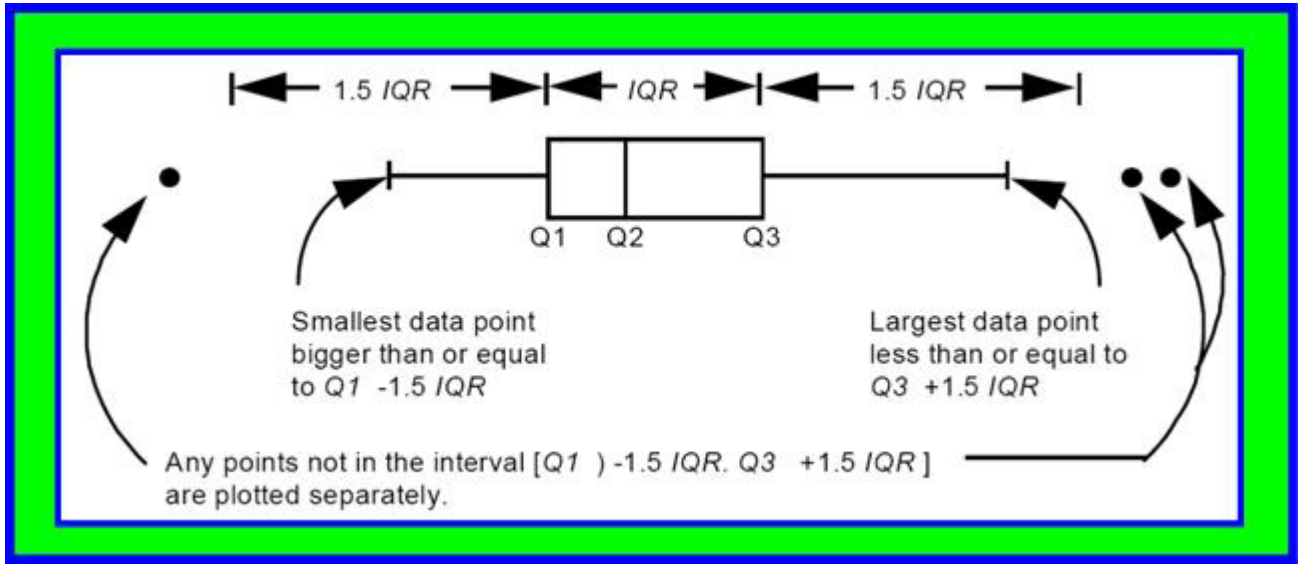
77, 79, 80, 86, 87, 87, 94, 99

Five-number summary =
IQR =
SQR =

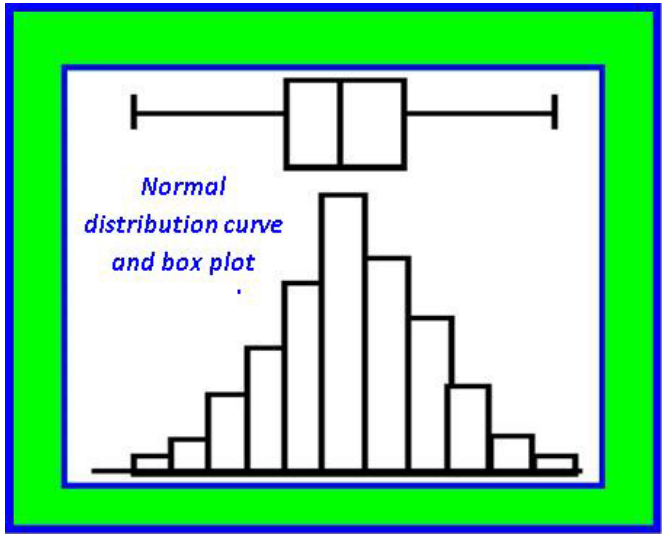


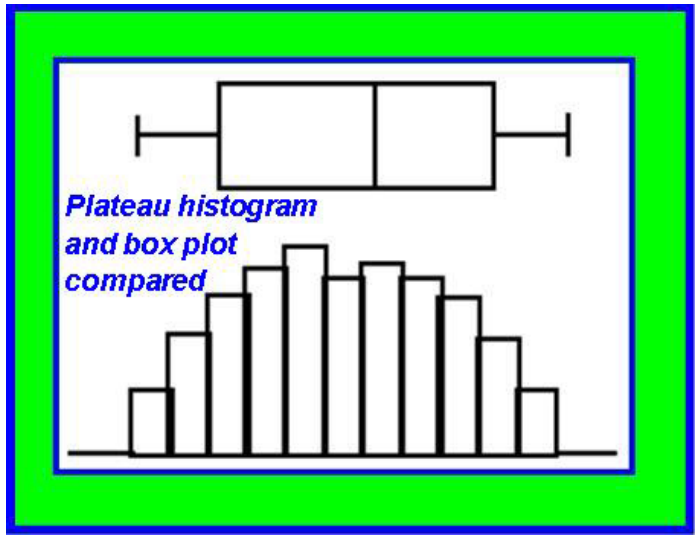
MODIFIED BOX PLOT

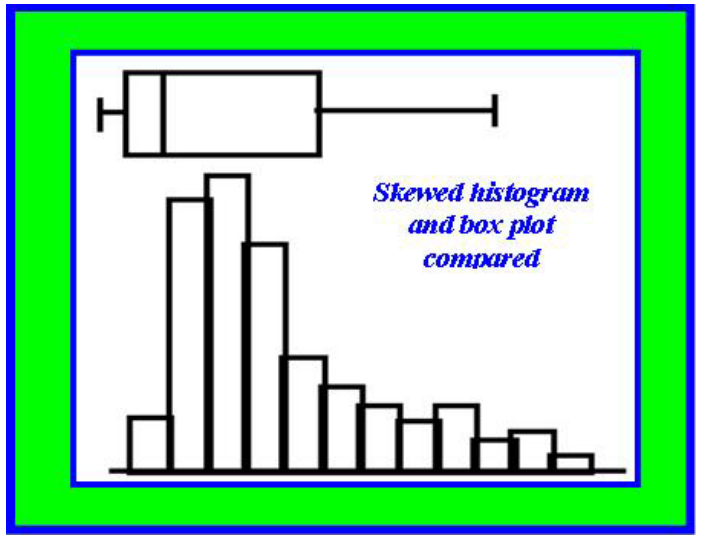
only let the 'whiskers' extend out to 1.5 x IQR and any points beyond that are represented with dots…these are suspected outliers to be investigated

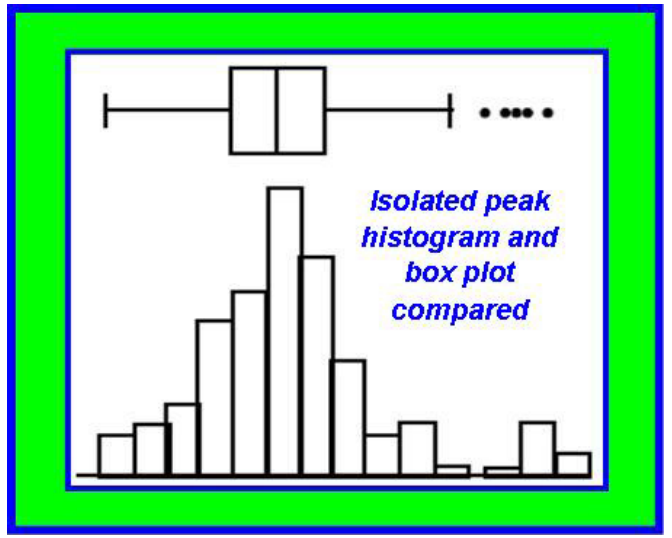


BOXPLOT VS. HISTOGRAM

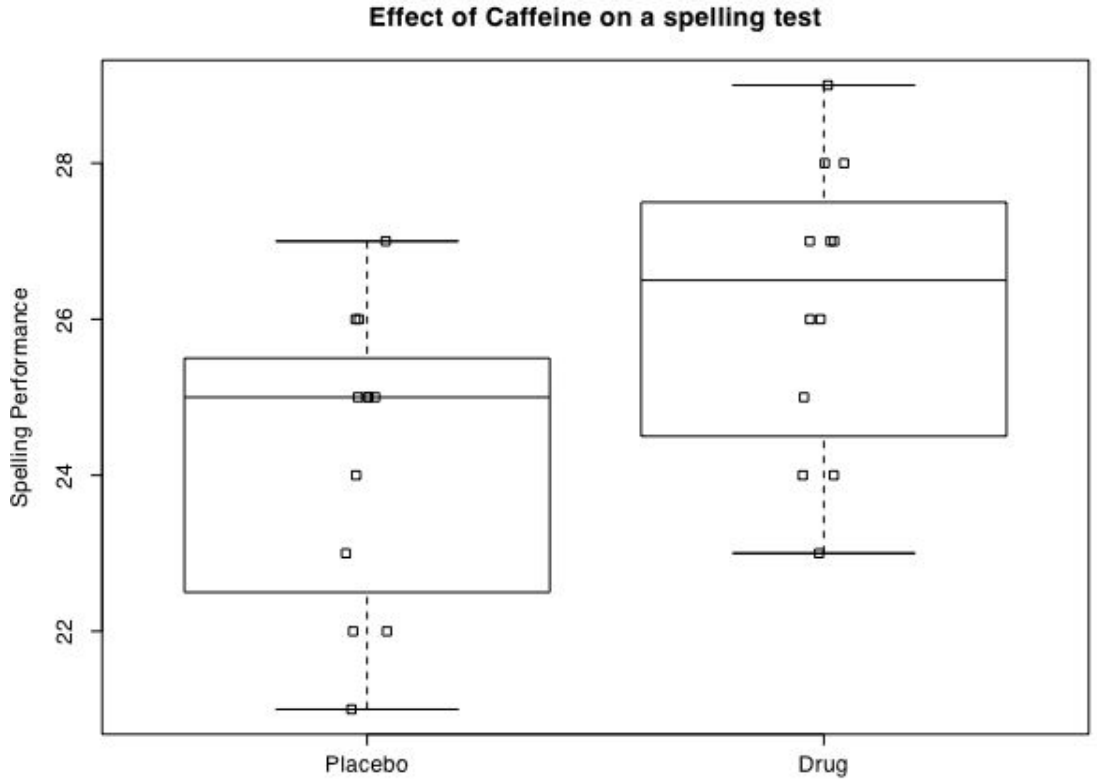








BOX PLOTS: COMPARING GROUPS EXAMPLES



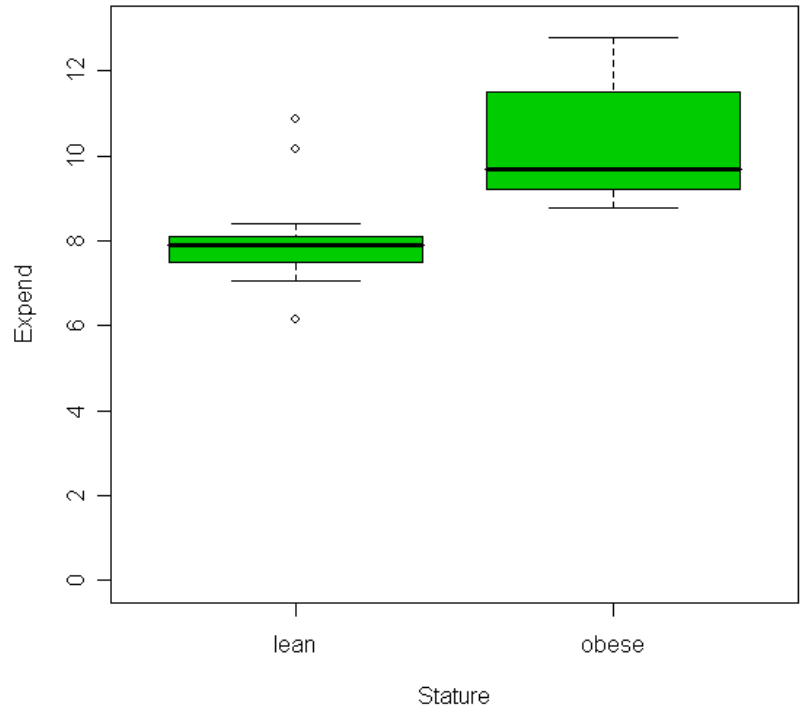


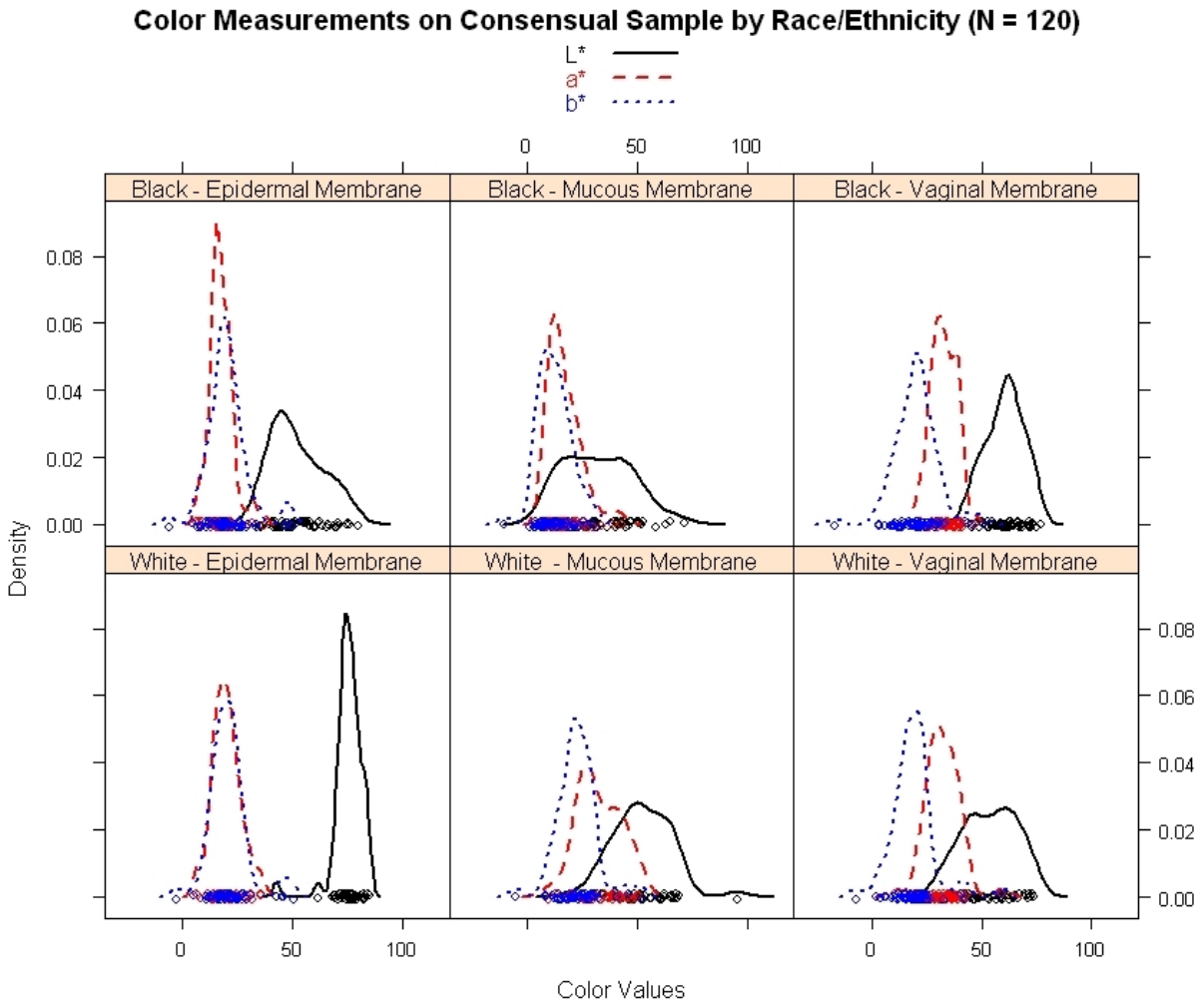
Figure 6. Boxplots of energy expenditure during 20 minutes of physical activity in lean (*n* = 11) and obese *(n* = 11) males.

(KERNEL) DENSITY PLOTS

- Smoothed histogram

- Non-parametric plotting technique that estimates a variable's probability distribution
- Specify…
  - Smoothing window: Symmetric range of values about each value
  - Bandwidth (*bw*) distance: Distance from edge of window to center ($2d$ = smoothing window width)
- Adjacent values within window of frequency distribution are averaged, resulting in a smoothed plot

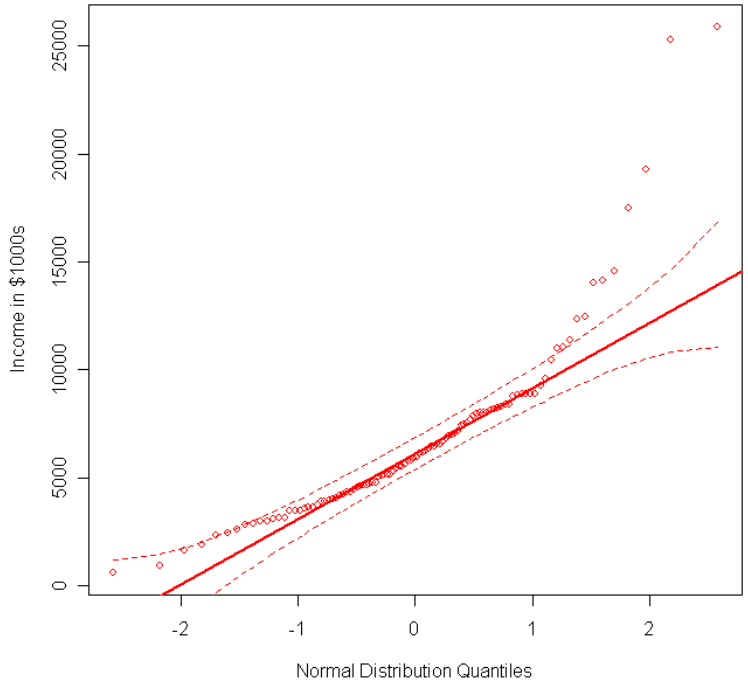
(KERNEL) DENSITY PLOTS - EXAMPLE



QUANTILE-QUANTILE (Q-Q) COMPARISON PLOT

- Scatterplot: Observed vs. theoretical distribution
- Theoretical can be any type: Normal, Poisson, etc.

- If observed variable follows chosen distribution, coordinate points will fall along 45° line w/in 95% confidence envelope
- Best method for evaluating normality; other methods better for evaluating symmetry and outliers



EXAMPLE: CANCER DATASET

```
GET FILE='C:\Users\A00315273\Box Sync\Teaching\Educ6600\Dataset\Cancer.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

VARIABLE LABELS
ID "Patient identification number"
TRT "Treatment Group"
AGE "Patient's Incoming Age"
WEIGHIN "Patient's Incoming Weight in pounds"
STAGE "Patient's Stage of Cancer".

VALUE LABELS TRT 0 "control" 1 "aleo treatment".

RECODE
TRT AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   (SYSMIS = 999).
EXECUTE.

VALUE LABELS
TRT
   0 "control" 1 "aleo treatment" 999 "missing"/
AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   999 "missing".

MISSING VALUES
TRT AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   (999).
```

Available on Canvas
Save to your computer
Edit the path to match

SPSS: SUMMARY STATISTICS W/FREQ

```
* you can ask for everything.

FREQUENCIES AGE
    /FORMAT NOTABLE
    /STATISTICS all.
```

Statistics

AGE Patient's Incoming Age

N	Valid	25
	Missing	0
Mean		59.64
Std. Error of Mean		2.586
Median		60.00
Mode		67
Std. Deviation		12.932
Variance		167.240
Skewness		-.348
Std. Error of Skewness		.464
Kurtosis		.584
Std. Error of Kurtosis		.902
Range		59
Minimum		27
Maximum		86
Sum		1491

```
* or just list out the ones you want.

FREQUENCIES AGE
    /FORMAT NOTABLE
    /STATISTICS MINIMUM MAXIMUM MEAN MEDIAN
                STDDEV RANGE SKEWNESS KURTOSIS.
```

Statistics

AGE Patient's Incoming Age

N	Valid	25
	Missing	0
Mean		59.64
Median		60.00
Std. Deviation		12.932
Skewness		-.348
Kurtosis		.584
Range		59
Minimum		27
Maximum		86

SPSS: SUMMARY STATISTICS W/FREQ

```
* to get IQR or SIR, you need to ask for the quartiles and then do the math.

FREQUENCIES AGE
    /NTILES(4)
    /FORMAT NOTABLE.
```

Statistics

AGE Patient's Incoming Age

N	Valid	25
	Missing	0
Percentiles	25	51.50
	50	60.00
	75	67.50

IQR = Q3 – Q1 = 67.50 – 51.50 = 16.00

SIR = IQR / 2 = 16.00 / 2 = 8.00

SPSS: SUMMARY STATISTICS FOR EACH GROUP

```
*** USING EXAMINE COMMAND.

EXAMINE VARIABLES=AGE BY STAGE
  /PLOT NONE
  /STATISTICS=DESCRIPTIVES
  /NOTOTAL.
```

Descriptives[a,b]

	STAGE Patient's Stage of Cancer		Statistic	Std. Error
AGE Patient's Incoming Age	1	Mean	61.67	4.506
		95% Confidence Interval for Mean — Lower Bound	51.75	
		95% Confidence Interval for Mean — Upper Bound	71.59	
		5% Trimmed Mean	62.24	
		Median	66.00	
		Variance	243.697	
		Std. Deviation	15.611	
		Minimum	27	
		Maximum	86	
		Range	59	
		Interquartile Range	15	
		Skewness	-.906	.637
		Kurtosis	1.321	1.232
	2	Mean	56.33	4.917
		95% Confidence Interval for Mean — Lower Bound	43.69	
		95% Confidence Interval for Mean — Upper Bound	68.97	
		5% Trimmed Mean	55.87	
		Median	55.50	
		Variance	145.067	
		Std. Deviation	12.044	
		Minimum	44	
		Maximum	77	
		Range	33	
		Interquartile Range	19	
		Skewness	1.002	.845
		Kurtosis	.986	1.741
	4	Mean	56.80	3.611
		95% Confidence Interval for Mean — Lower Bound	46.77	
		95% Confidence Interval for Mean — Upper Bound	66.83	
		5% Trimmed Mean	56.78	
		Median	56.00	
		Variance	65.200	
		Std. Deviation	8.075	
		Minimum	46	
		Maximum	68	
		Range	22	
		Interquartile Range	14	
		Skewness	.123	.913
		Kurtosis	.676	2.000

a. AGE Patient's Incoming Age is constant when STAGE Patient's Stage of Cancer = 0. It has been omitted.

b. AGE Patient's Incoming Age is constant when STAGE Patient's Stage of Cancer = 3. It has been omitted.

SPSS: SUMMARY STATISTICS FOR EACH GROUP

```
* FIRST:  you have to SORT by the variable you are going to split on.
* SECOND: make sure you use a 'temporary.' command so its not permanent.
* THIRD:  make sure you turn the split off at the end.
* FOURTH: its nice to go back to the original sorting.

SORT CASES by STAGE.
TEMPORARY.
SPLIT FILE by STAGE.
FREQUENCIES AGE
   /FORMAT NOTABLE
   /STATISTICS MEAN MEDIAN STDDEV.
SPLIT FILE off.
SORT CASES by id.
```

Statistics

AGE Patient's Incoming Age

0	N	Valid	1
		Missing	0
	Mean		73.00
	Median		73.00
1	N	Valid	12
		Missing	0
	Mean		61.67
	Median		66.00
	Std. Deviation		15.611
2	N	Valid	6
		Missing	0
	Mean		56.33
	Median		55.50
	Std. Deviation		12.044
3	N	Valid	1
		Missing	0
	Mean		56.00
	Median		56.00
4	N	Valid	5
		Missing	0
	Mean		56.80
	Median		56.00
	Std. Deviation		8.075

SPSS: CREATE BOXPLOTS

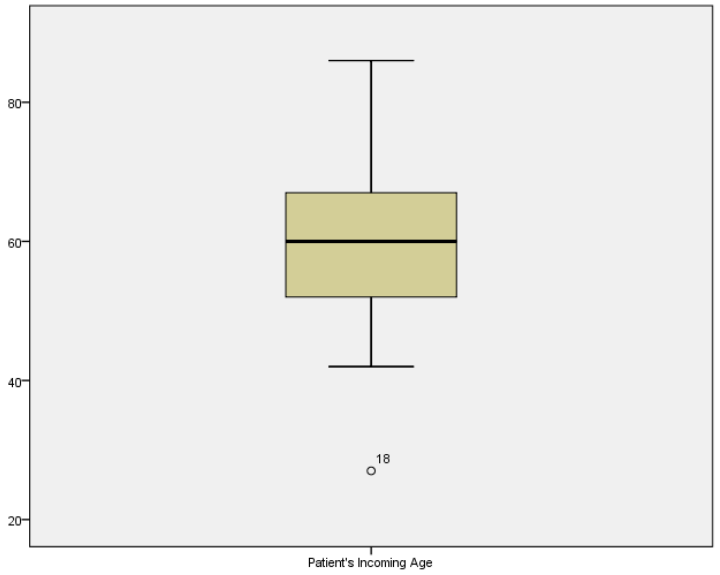
```
* For all subjects.

EXAMINE AGE /PLOT BOXPLOT /STATISTICS NONE.
```

Total Sample

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
AGE Patient's Incoming Age	25	100.0%	0	0.0%	25	100.0%

AGE Patient's Incoming Age




```
* seperately by groupings.

EXAMINE VARIABLES=AGE BY STAGE
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```

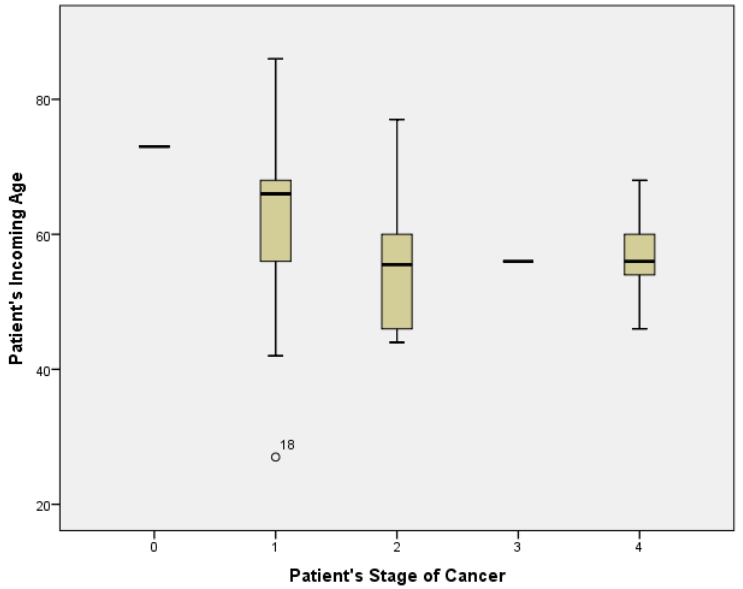
AGE Patient's Incoming Age is constant when STAGE Patient's Stage of Cancer = 0. It will be included in any boxplots produced but other output will be omitted.
AGE Patient's Incoming Age is constant when STAGE Patient's Stage of Cancer = 3. It will be included in any boxplots produced but other output will be omitted.

STAGE Patient's Stage of Cancer

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	STAGE Patient's Stage of Cancer	N	Percent	N	Percent	N	Percent
AGE Patient's Incoming Age	0	1	100.0%	0	0.0%	1	100.0%
	1	12	100.0%	0	0.0%	12	100.0%
	2	6	100.0%	0	0.0%	6	100.0%
	3	1	100.0%	0	0.0%	1	100.0%
	4	5	100.0%	0	0.0%	5	100.0%

AGE Patient's Incoming Age



SPSS: CREATE BOXPLOTS

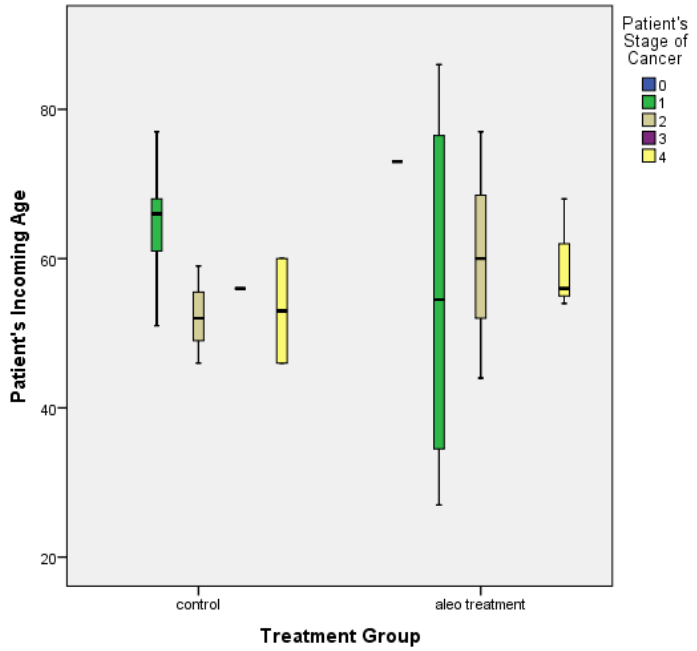
```
* seperately by two grouping variables.

EXAMINE VARIABLES=AGE BY TRT*STAGE
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```

TRT Treatment Group*STAGE Patient's Stage of Cancer

Case Processing Summary

			Cases					
			Valid		Missing		Total	
	TRT Treatment Group	STAGE Patient's Stage of Cancer	N	Percent	N	Percent	N	Percent
AGE Patient's Incoming Age	0 control	1	8	100.0%	0	0.0%	8	100.0%
		2	3	100.0%	0	0.0%	3	100.0%
		3	1	100.0%	0	0.0%	1	100.0%
		4	2	100.0%	0	0.0%	2	100.0%
	1 aleo treatment	1	4	100.0%	0	0.0%	4	100.0%
		2	3	100.0%	0	0.0%	3	100.0%
		4	3	100.0%	0	0.0%	3	100.0%
		0	1	100.0%	0	0.0%	1	100.0%

AGE Patient's Incoming Age



SPSS: RESTRICTING CASES

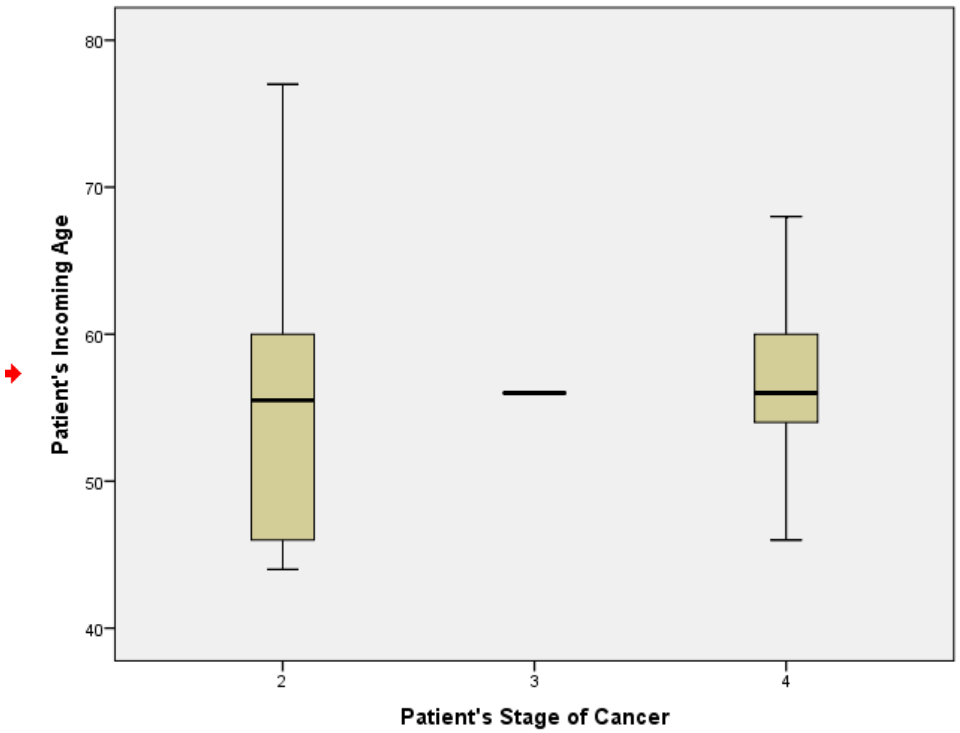
```
* Create a box plot
   ONLY for cases with stage greater than 1.

TEMPORARY.
SELECT IF  STAGE > 1.
EXAMINE VARIABLES=AGE BY STAGE
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```

STAGE Patient's Stage of Cancer

Case Processing Summary

	STAGE Patient's Stage of Cancer	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
AGE Patient's Incoming Age	2	6	100.0%	0	0.0%	6	100.0%
	3	1	100.0%	0	0.0%	1	100.0%
	4	5	100.0%	0	0.0%	5	100.0%

AGE Patient's Incoming Age



SPSS: RESTRICTING CASES

```
* Create a box plot ONLY for cases
   with stage 1 & in the treatment group.

TEMPORARY.
SELECT IF  STAGE = 1 & TRT = 1.
EXAMINE VARIABLES=AGE
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
AGE Patient's Incoming Age	4	100.0%	0	0.0%	4	100.0%

AGE Patient's Incoming Age