

# Center and Spread

## Cohen Chapter 3

EDUC/PSY 6600

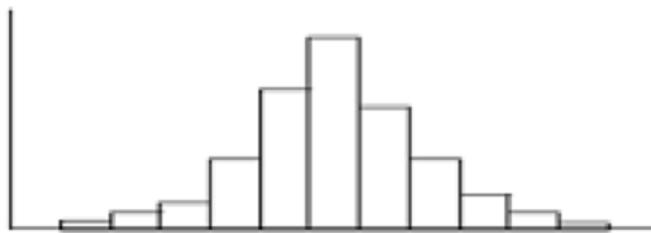
"You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.

*Individuals vary, but percentages remain constant. So says the statistician."*

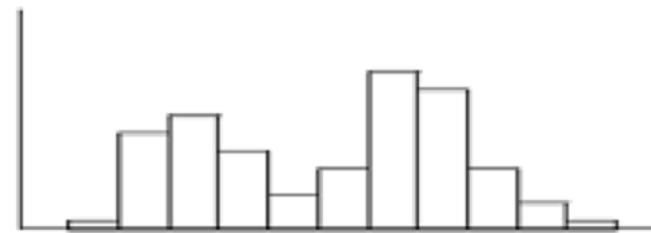
-- Sherlock Holmes, *The Sign of Four*

# Distributions Examples

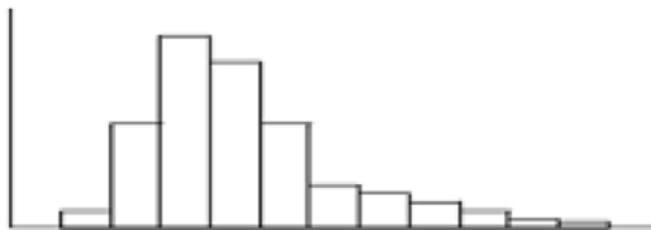
Bell-shaped



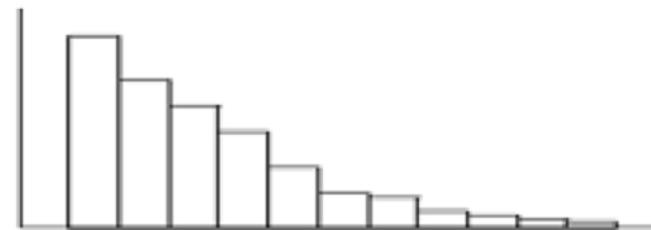
Bimodal



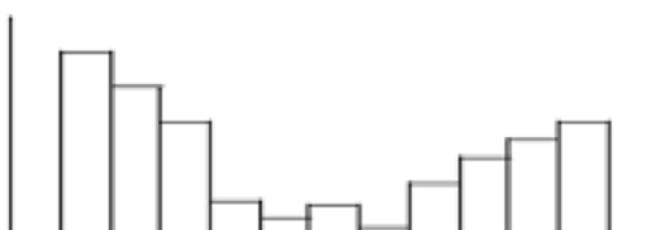
Right-skewed



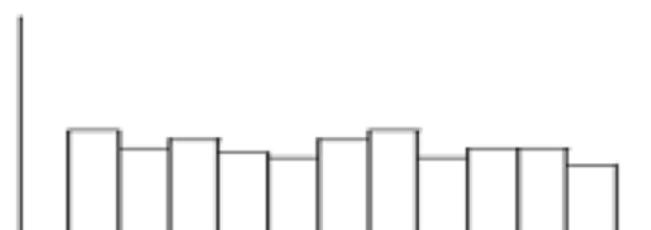
J-shaped



U-shaped



Uniform



# Three Measures of Center

## Mean

“Arithmetic Average” = add them all up and divide by the count

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” ( $\mu$ )

$$\mu = \frac{\sum X_i}{N}$$

In a **SAMPLE**: “X-bar” ( $\bar{X}$ ) but APA uses “M” for abbreviation

$$\bar{X} = \frac{\sum X_i}{n}$$

## Median

50<sup>th</sup> 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

## Mode

Most common value, largest frequency, highest peak

Non-uniqueness - can have more than one mode

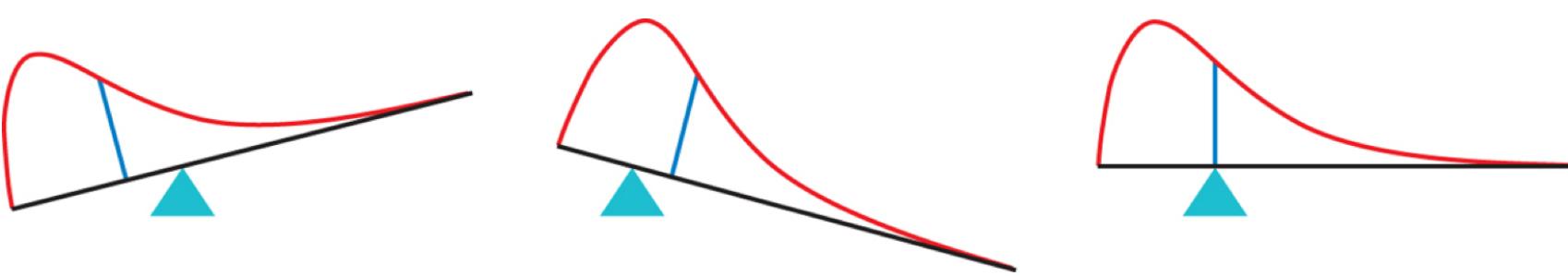
Doesn’t always represent the ‘center’

Do NOT usually use, other than descriptively

# Mean vs. Median

**Median:** the center point, half of values are on each side, not affected by the skew, the "typical value"

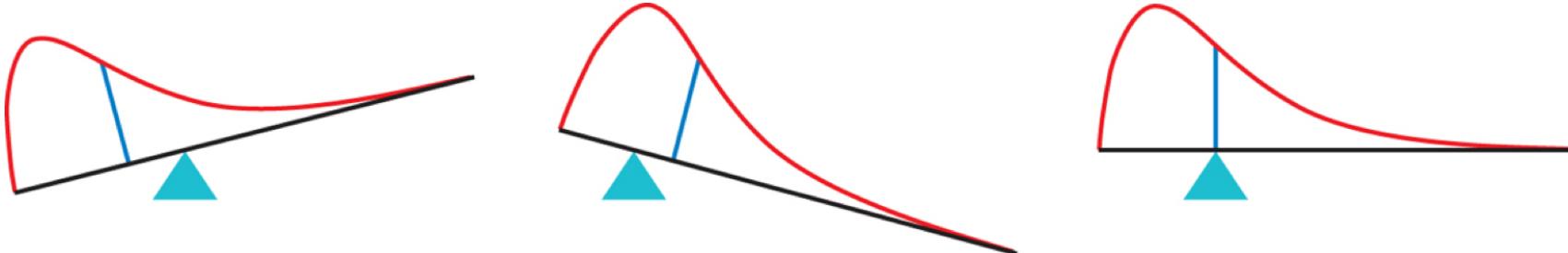
**Mean:** the "balance" point, pulled to the side of the skew, not typical



# Mean vs. Median

**Median:** the center point, half of values are on each side, not affected by the skew, the "typical value"

**Mean:** the "balance" point, pulled to the side of the skew, not typical



If distribution is symmetrical: mean = median

# Distribution of annual household income in the United States

2010 estimate



# Distributions and Numbers

- The MEDIAN is **resistant** & doesn't change much
- The MEAN is **influenced** & changes more!
- Average does NOT mean typical
- Average moves when we remove the high point

# Distributions and Numbers

- The MEDIAN is **resistant** & doesn't change much
- The MEAN is **influenced** & changes more!
- Average does NOT mean typical
- Average moves when we remove the high point



# Distributions and Numbers

- The MEDIAN is **resistant** & doesn't change much
- The MEAN is **influenced** & changes more!
- Average does NOT mean typical
- Average moves when we remove the high point
- Median doesn't move when we remove the high point

# Distributions and Numbers

- The MEDIAN is **resistant** & doesn't change much
- The MEAN is **influenced** & changes more!
- Average does NOT mean typical
- Average moves when we remove the high point
- Median doesn't move when we remove the high point



# Three Measures of Spread

## Range, IQR, & SIR

- Range = Max - Min
- Interquartile Range

$$IQR = Q3 - Q1$$

- Semi-Interquartile Range

$$SIR = (Q3 - Q1) / 2$$

- Range is super dependent on extreme values or outliers
- IQR & SIR more resistant

## Variance

- DEVIANTE: 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"

$$SS = \sum (X_i - \text{mean})^2$$

$$MS = \frac{SS}{df}$$

$$\sigma^2 = \frac{\sum (X_i - \mu)}{N} = \frac{SS}{N} = MS$$

- In a SAMPLE: called "s-squared"

$$s^2 = \frac{\sum (X_i - \bar{X})}{n-1} = \frac{SS}{n-1} = \frac{SS}{df} = MS$$

- Degrees of Freedom:  $df = n - 1$

## Standard Deviation

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

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X_i - \mu)}{N}} = \sqrt{\frac{SS}{N}}$$
$$= \sqrt{MS}$$

- In a SAMPLE: called "s"

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (X_i - \bar{X})}{n-1}} = \sqrt{\frac{SS}{n-1}}$$
$$= \sqrt{MS}$$

# Best Summary of the Data?

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

# Best Summary of the Data?

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

## Median and SIR

Skewed data or outliers

## Mean and SD

Symmetrical and no outliers

# Best Summary of the Data?

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

## Median and SIR

Skewed data or outliers

## Mean and SD

Symmetrical and no outliers

A graph gives the best overall picture of a distribution

# Properties of the Mean and SD

If you ____ the same CONSTANT number onto every score...	MEAN	STANDARD DEVIATION
ADD (or subtract)	ADD (or subtract) SAME amount	unchanged
MULTIPLY (or divide)	MULTIPLY (or divide) by the SAME amount	MULTIPLY (or divide) by the SAME amount

# Skewness

- Degree of **symmetry** in distribution
- Can detect **visually** (histogram, boxplot)
- Skewness statistic
  - Based on cubed deviations from the mean
  - Divided by SE of skewness
  - $> \pm 2$  is a sign of skewed data

# Skewness

- Degree of **symmetry** in distribution
- Can detect **visually** (histogram, boxplot)
- Skewness statistic
  - Based on cubed deviations from the mean
  - Divided by SE of skewness
  - $> \pm 2$  is a sign of skewed data

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

- Interpreting skewness statistic
  - positive value = positive (right) skew
  - negative value = negative (left) skew
  - zero value = no skew

# Skewness

- Degree of **symmetry** in distribution
- Can detect **visually** (histogram, boxplot)
- Skewness statistic
  - Based on cubed deviations from the mean
  - Divided by SE of skewness
  - $> \pm 2$  is a sign of skewed data

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

- Interpreting skewness statistic
  - positive value = positive (right) skew
  - negative value = negative (left) skew
  - zero value = no skew



# Kurtosis

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

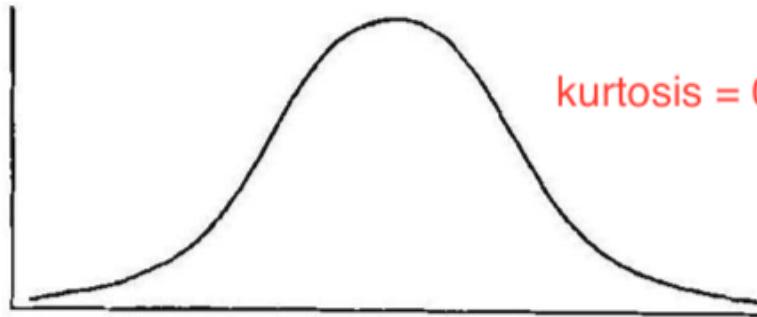
- Degree of **flatness** in distribution
- Harder to detect visually
- Kurtosis statistic
  - Based on deviations from the mean (raised to 4th power)
  - Divided by SE of kurtosis
  - $> \pm 2$  is a sign of problems with kurtosis

# Kurtosis

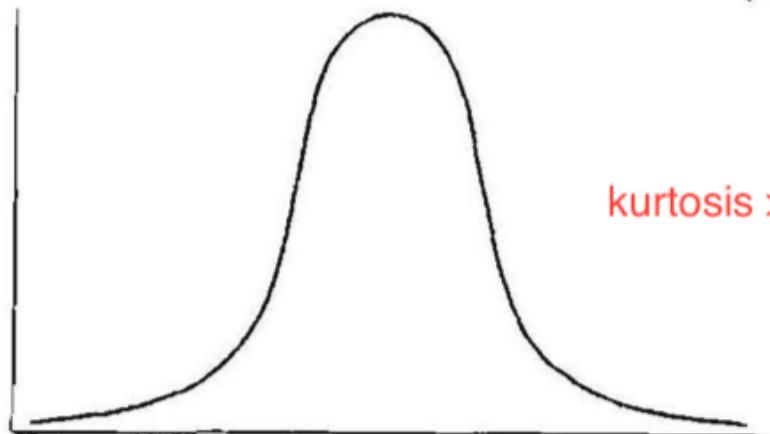
$$Kurtosis = \frac{N(N+1)}{(N-2)(N-3)} \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{(N-1)s^4} - 3 \frac{(N-1)(N-1)}{(N-2)(N-3)}$$

- Degree of **flatness** in distribution
- Harder to detect visually
- Kurtosis statistic
  - Based on deviations from the mean (raised to 4th power)
  - Divided by SE of kurtosis
  - $> \pm 2$  is a sign of problems with kurtosis
- Interpreting kurtosis statistic
  - positive value = leptokurtic (peaked)
  - negative value = platykurtic (flat)
  - zero value = mesokurtic (normal)

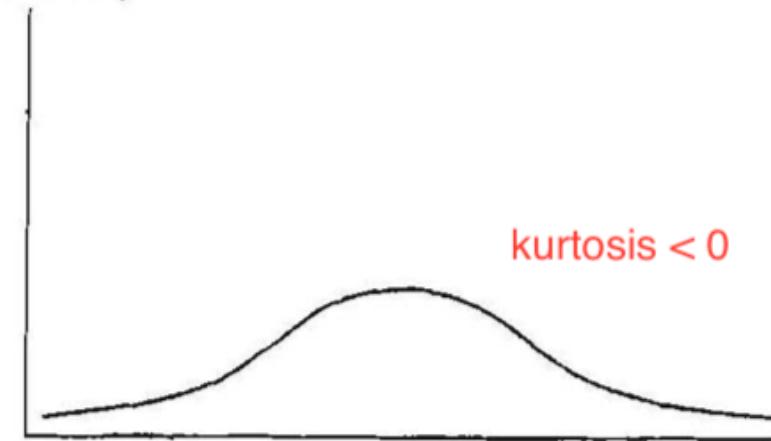
# Kurtosis



**Distribution H**  
**(Mesokurtic distribution)**



**Distribution I**  
**(Leptokurtic distribution)**

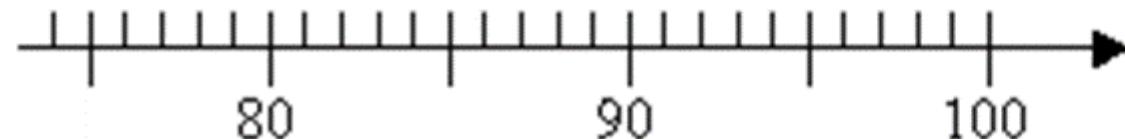


**Distribution J**  
**(Platykurtic distribution)**

# Five-Number Summary

Example of VERY small amount of data

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



# Five-Number Summary - Median

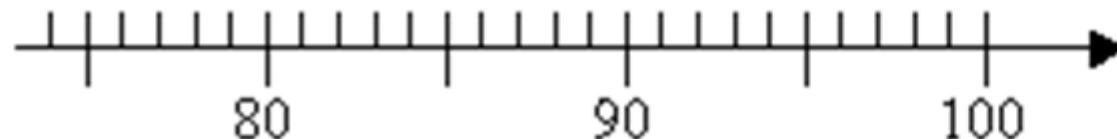
Example of VERY small amount of data  
77, 79, 80, 86, 87, 87, 94, 99

Min = 77

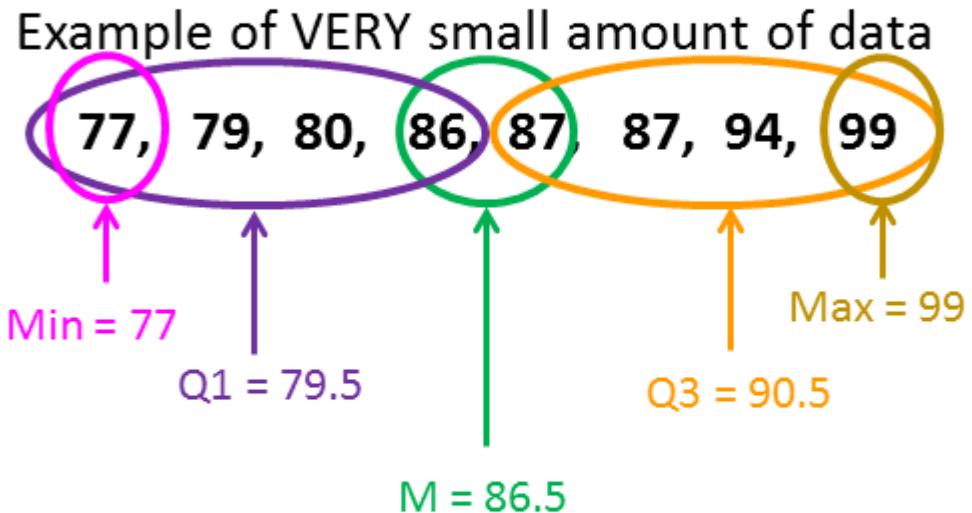
↑

M = 86.5

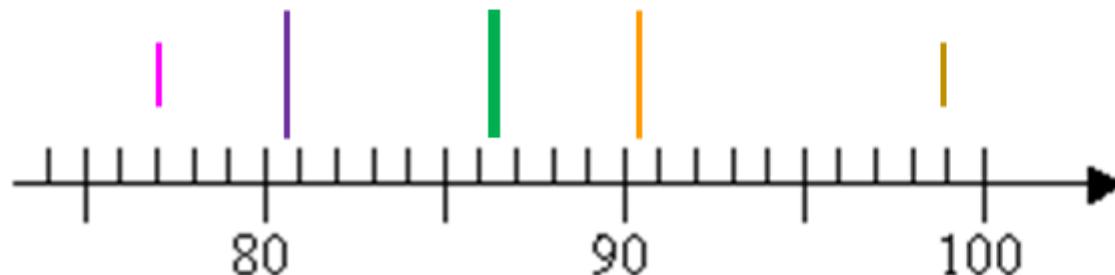
Max = 99



# Five-Number Summary - Quartiles



Five-number summary = 77, 79.5, 86.5, 90.5, 99



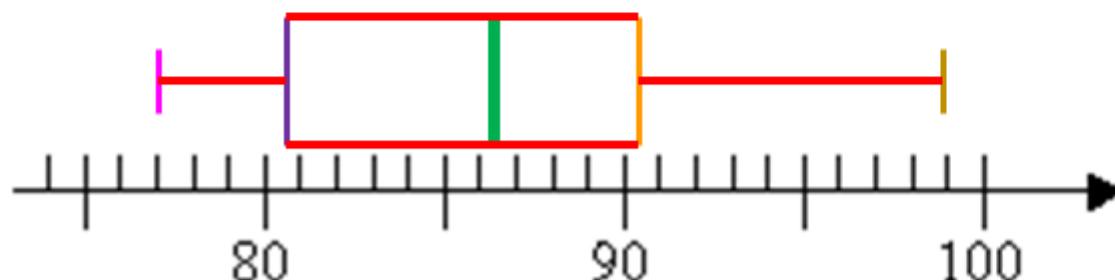
# Boxplots (Modified) - Lines

Example of VERY small amount of data

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

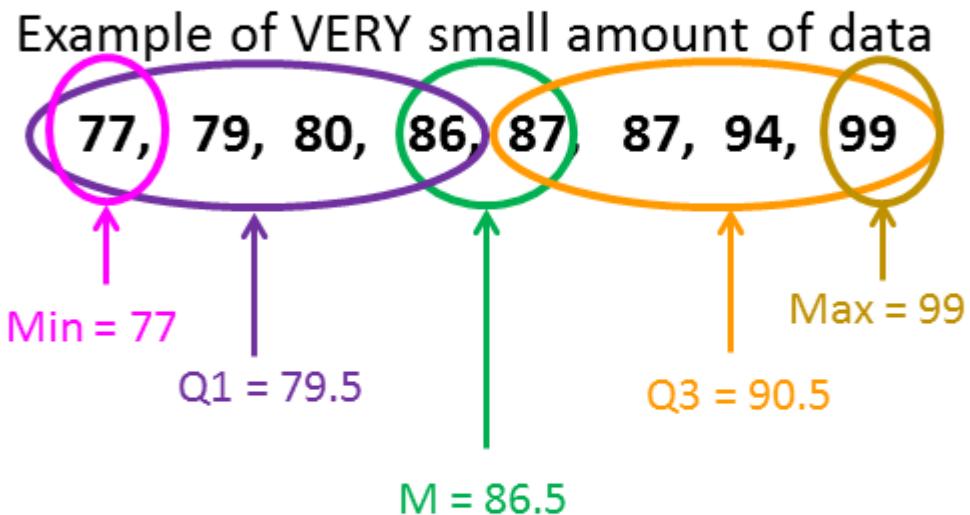
Min = 77      Q1 = 79.5      M = 86.5      Q3 = 90.5      Max = 99

Five-number summary = 77, 79.5, 86.5, 90.5, 99



Box-Plot = connect lines plotted above

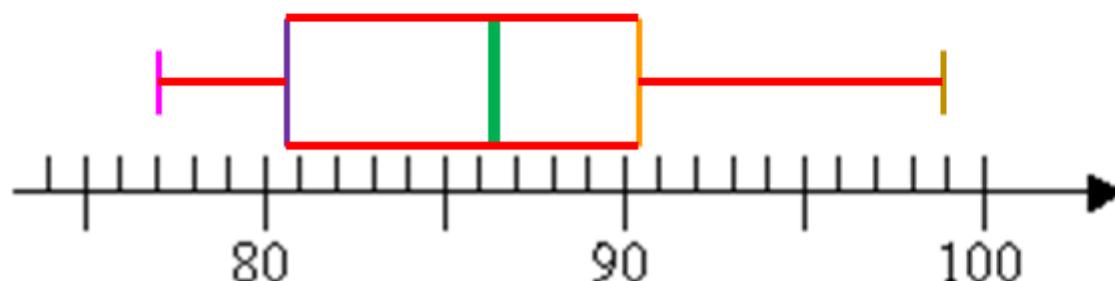
# Boxplots (Modified) - IQR and SIQR



Five-number summary = 77, 79.5, 86.5, 90.5, 99

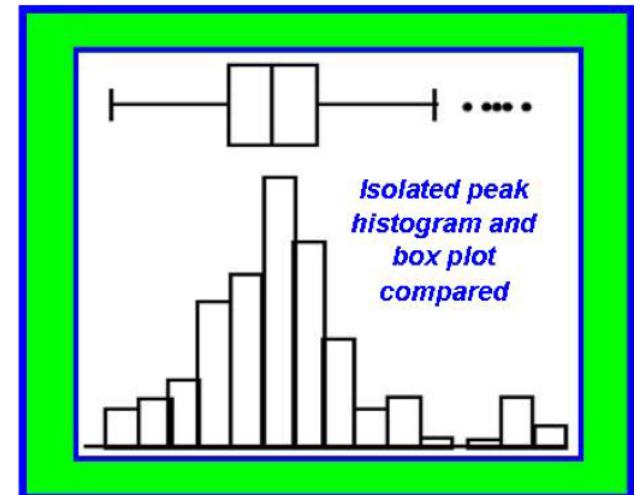
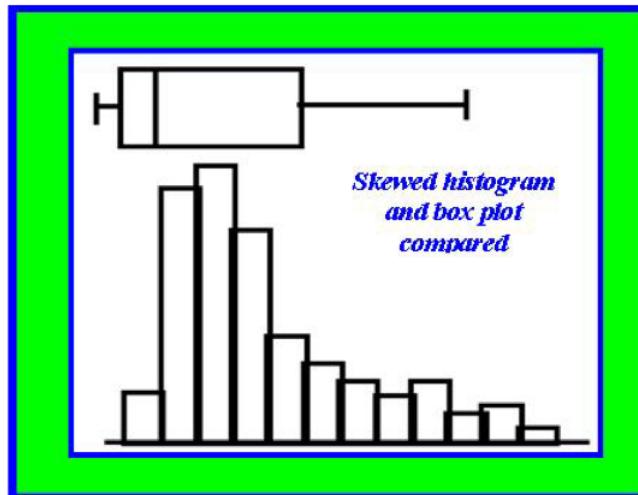
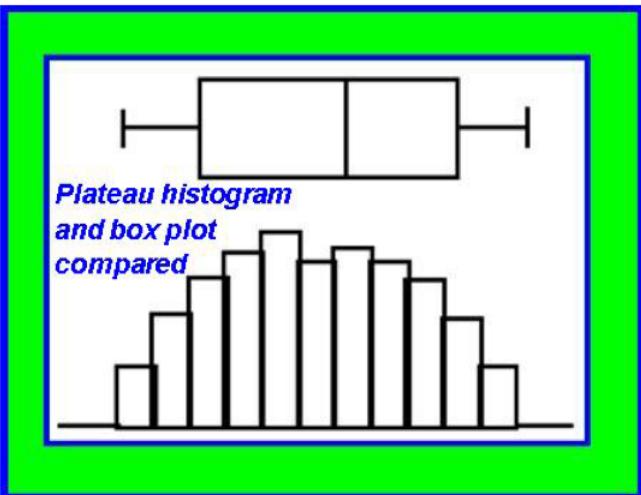
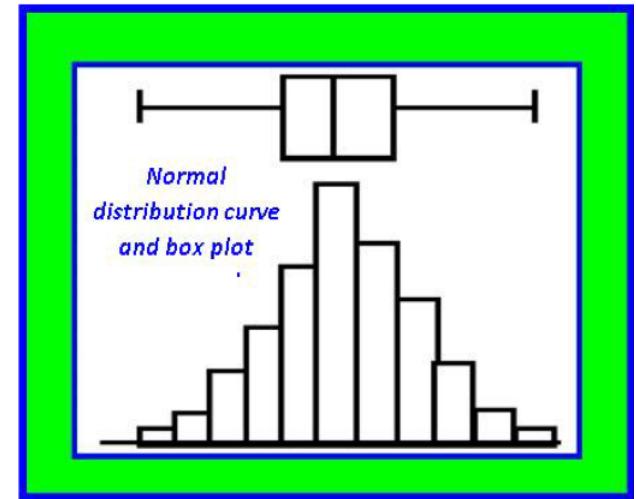
$$\text{IQR} = 90.5 - 79.5 = 11$$

$$\text{SIQR} = 11 / 2 = 5.5$$



Box-Plot = connect lines plotted above

# Boxplot vs. Histogram



# Boxplots by Group



# Density Plots



# Quantile-Quantile (Q-Q) Plot



Let's Apply This To the Cancer Dataset  
(on Canvas)

# Read in the Data

```
library(tidyverse)      # Loads several very helpful 'tidy' packages
library(haven)          # Read in SPSS datasets
library(furniture)       # Nice tables (by our own Tyson Barrett)
library(psych)           # Lots of nice tid-bits

cancer_raw <- haven::read_spss("cancer.sav")
```

# Read in the Data

```
library(tidyverse)      # Loads several very helpful 'tidy' packages
library(haven)          # Read in SPSS datasets
library(furniture)       # Nice tables (by our own Tyson Barrett)
library(psych)           # Lots of nice tid-bits

cancer_raw <- haven::read_spss("cancer.sav")
```

## And Clean It

```
cancer_clean <- cancer_raw %>%
  dplyr::rename_all(tolower) %>%
  dplyr::mutate(id = factor(id)) %>%
  dplyr::mutate(trt = factor(trt,
                             labels = c("Placebo",
                                       "Aloe Juice"))) %>%
  dplyr::mutate(stage = factor(stage))
```

# Frequency Tables with `furniture::tableF()`

```
cancer_clean %>%  
  furniture::tableF(age, n = 8)
```

age	Freq	CumFreq	Percent	CumPerc
27	1	1	4.00%	4.00%
42	1	2	4.00%	8.00%
44	1	3	4.00%	12.00%
46	2	5	8.00%	20.00%
...	...	...	...	...
68	1	20	4.00%	80.00%
69	1	21	4.00%	84.00%
73	1	22	4.00%	88.00%
77	2	24	8.00%	96.00%
86	1	25	4.00%	100.00%

```
cancer_clean %>%  
  furniture::tableF(trt)
```

trt	Freq	CumFreq	Percent	CumPerc
Placebo	14	14	56.00%	56.00%
Aloe Juice	11	25	44.00%	100.00%

# Extensive Descriptive Stats psych::describe()

```
cancer_clean %>%
  dplyr::select(age, weighin, totalcin, totalcw2, totalcw4, totalcw6) %>%
  psych::describe()
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
age	1	25	59.64	12.93	60.0	59.95	11.86	27	86.0	59.0	-0.31
weighin	2	25	178.28	31.98	172.8	176.57	21.05	124	261.4	137.4	0.73
totalcin	3	25	6.52	1.53	6.0	6.33	0.00	4	12.0	8.0	1.80
totalcw2	4	25	8.28	2.54	8.0	8.10	2.97	4	16.0	12.0	1.01
totalcw4	5	25	10.36	3.47	10.0	10.19	2.97	6	17.0	11.0	0.49
totalcw6	6	23	9.48	3.49	9.0	9.21	2.97	3	19.0	16.0	0.77
			kurtosis	se							
age			-0.01	2.59							
weighin			0.07	6.40							
totalcin			4.30	0.31							
totalcw2			1.14	0.51							
totalcw4			-1.00	0.69							
totalcw6			0.53	0.73							

# Smaller Set with furniture::table1()

For the Entire Sample

```
cancer_clean %>%  
  furniture::table1(trt, age, weighin)
```

-----

	Mean/Count (SD/%)
n	= 25
trt	
Placebo	14 (56%)
Aloe Juice	11 (44%)
age	59.6 (12.9)
weighin	178.3 (32.0)

-----

Breaking the Sample by a Factor

```
cancer_clean %>%  
  dplyr::group_by(trt) %>%  
  furniture::table1(age, weighin)
```

-----

trt	Placebo	Aloe Juice
n	= 14	= 11
age	59.8 (9.0)	59.5 (17.2)
weighin	167.5 (23.0)	192.0 (37.4)

-----

# Boxplot, one one `geom_boxplot()`

```
cancer_clean %>%
  ggplot(aes(x = "Full Sample",      # x = "quoted text"
             y = age)) +
  geom_boxplot()
```



# Boxplots, by groups - (1) fill color

```
cancer_clean %>%
  ggplot(aes(x = "Full Sample",
              y = age,
              fill = trt)) +
  geom_boxplot()
```



# Boxplots, by groups - (2) x-axis breaks

```
cancer_clean %>%  
  ggplot(aes(x = trt,                      # x = group_var (no quotes)  
             y = age)) +                  # y = contin_var (no quotes)  
  geom_boxplot()
```



# Boxplots, by groups - (3) seperate panels

```
cancer_clean %>%  
  ggplot(aes(x = "Full Sample",      # x = "quoted text"  
             y = age)) +           # y = contin_var (no quotes)  
  geom_boxplot() +  
  facet_grid(. ~ trt)          # . ~ group_var (no quotes)
```



# Boxplot for a Subset - 1 requirement

```
cancer_clean %>%                      # Less than 172 Pound at baseline  
  dplyr::filter(weighin < 172) %>%  
  ggplot(aes(x = "Weigh At Baseline < 172",  
             y = age)) +  
  geom_boxplot()
```



# Boxplot for a Subset - 2 requirements

```
cancer_clean %>%          # At least 150 pounds AND not in Aloe group
  dplyr::filter(weighin >= 150 & trt == "Placebo") %>%
  ggplot(aes(x = "Placebo and at least 150 Pounds",
             y = age)) +
  geom_boxplot()
```



# Boxplot for a Subset - 2 requirements (%in%)

```
cancer_clean %>%          # In Aloe group, but only stages 2-4
  dplyr::filter(trt == "Aloe Juice" & stage %in% c(2, 3, 4)) %>%
  ggplot(aes(x = "On Aloe Juice and Stage 2-4",
             y = weighin)) +
  geom_boxplot()
```



# Boxplot for Repeated Measures

```
cancer_clean %>%
  tidyr::gather(key = "time",           # stack the repeated measures
                value = "value",
                totalcin, totalcw2, totalcw4, totalcw6) %>%
  ggplot(aes(x = time,
             y = value)) +
  geom_boxplot()
```



# Boxplot: COMPLICATED!

```
cancer_clean %>%
  dplyr::filter(weighin > 130 & stage %in% c(2, 4)) %>%
  tidyr::gather(key = "time", value = "value",
                totalcin, totalcw2, totalcw4, totalcw6) %>%
  ggplot(aes(x = time, y = value, fill = stage)) +
  geom_boxplot() +
  facet_grid(. ~ trt)
```



# Questions?

# Next Topic

Standard and Normal