STA 570 Spring 2011

Lecture 5 Tues*day, Feb 1*

- Descriptive Statistics
  - Summarizing Univariate Data
    - Standard Deviation, Empirical Rule, IQR
    - Boxplots
  - Summarizing Bivariate Data
    - Contingency Tables
    - Row Proportions, Relative Risk, Odds Ratio

Homework 4: Due next week in lab.

Standard Deviation Interpretation: Empirical Rule

- If the histogram of the data is approximately symmetric and bell-shaped, then
  - About **68%** of the data are within **one** standard deviation from the mean
  - About **95%** of the data are within **two** standard deviations from the mean
  - About **99.7%** of the data are within **three** standard deviations from the mean

More precisely: 68.27% – 95.45% – 99.73%

Yet Another Example

- “Number of people you have known personally who have committed suicide in the last 12 months”

Response	Frequency	Percentage
0	1344	88.8
1	133	8.8
2	25	1.7
3	11	0.7
4	1	0.1

- Mean = 0.15
- Standard Deviation = 0.46
- Are 68% of the observations between –0.31 and 0.61?

Standard Deviation (*s*) vs. Interquartile Range (*IQR*)

- Standard Deviation is affected by outliers
- Interquartile Range is not affected by outliers
- *NB: The Range is most affected by outliers*

- Whenever you use the Median instead of the Mean, you should also use the IQR instead of the Standard Deviation

Example Data Sets

- One Variable Statistical Calculator
- Modify the data sets and see how mean and median, as well as standard deviation and interquartile range change
- Look at the histograms and stem-and-leaf plots – does the empirical rule apply?
- Make yourself familiar with the standard deviation
- Interpreting the standard deviation takes experience

Another Graphical Technique: Boxplots

- A boxplot is intended to be a SIMPLE plot which allows you to quickly see all the features of the distribution.
- Basically a graphical version of the five number summary
- There are marks on a boxplot to show the median, Q_1, Q_3 (together these allow you to see the interquartile range).
- There are "whiskers" on a boxplot to show the tails of the distribution, with circles indicating outliers.

Constructing a boxplot

- Draw an axis which covers the entire range of your data.
- Draw a narrow box extending from Q_1 to Q_3. At the location of the median, draw a line through the box.
- Define “inner fences” as Q_1 – 1.5 IQR and Q_3 + 1.5 IQR. On each side, find the most extreme point within the fences (note the fences are NOT drawn on the plot)

Boxplots, continued

- On each side of the box, draw a line from the end of the box to the most extreme points you found in the previous step. These lines are called “whiskers”.
- Define outer fences (again, NOT drawn) to be Q_1 – 3 IQR and Q_3 + 3 IQR.

Boxplots, continued

- Mark any points appearing between the inner and outer fences with an open circle. These are usually referred to as “mild” outliers.
- Mark any points beyond the outer fences with closed circles or asterisks. These are typically referred to as “extreme outliers”.

Boxplot, continued

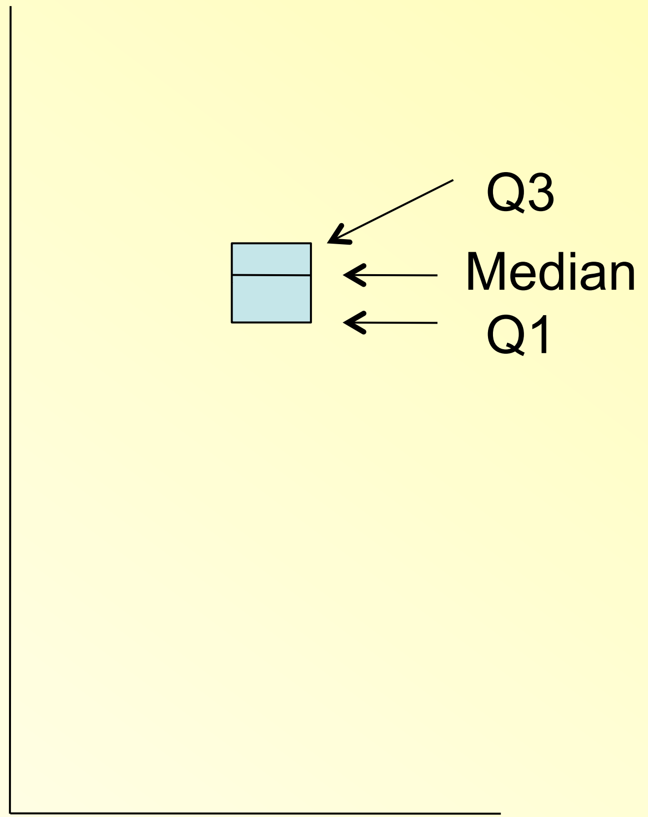
- These rules are the most common, but different people and different computer software may use different symbols, and may have different definitions of mild and extreme outliers.

Using Boxplots

- Central location may be visualized using the median, on the plot. Remember the box also contains the middle 50% of the data. Some programs will put a separate mark for the mean of the data.
- Spread may be visualized with the IQR, the length of the box. You can also see the range of the data. The standard deviation CANNOT be found with a boxplot.

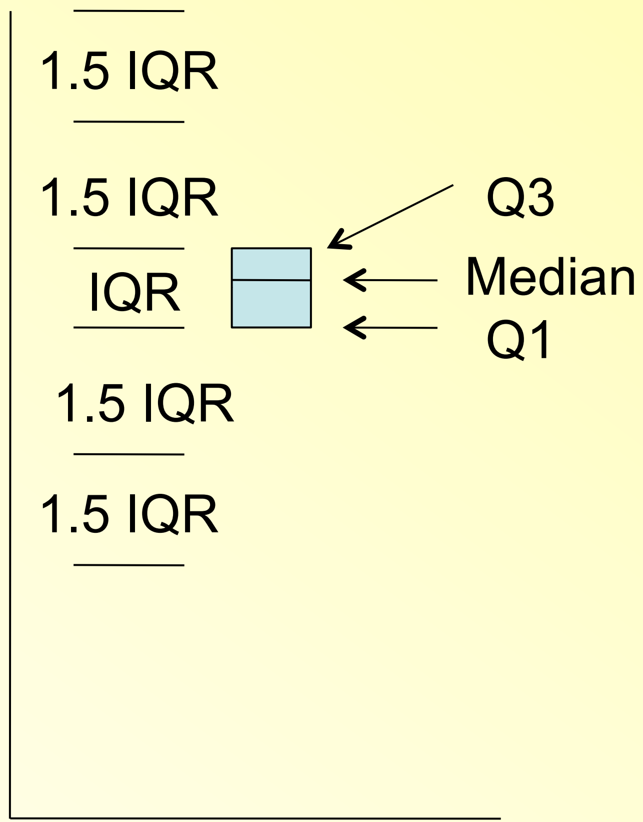
Step 1 for boxplot – The Box

- Box extends from Q1 to Q3, with a line for the median.
- Thus, you can immediately see the median (central location) and the IQR (spread).
- Note the box contains 50% of the data



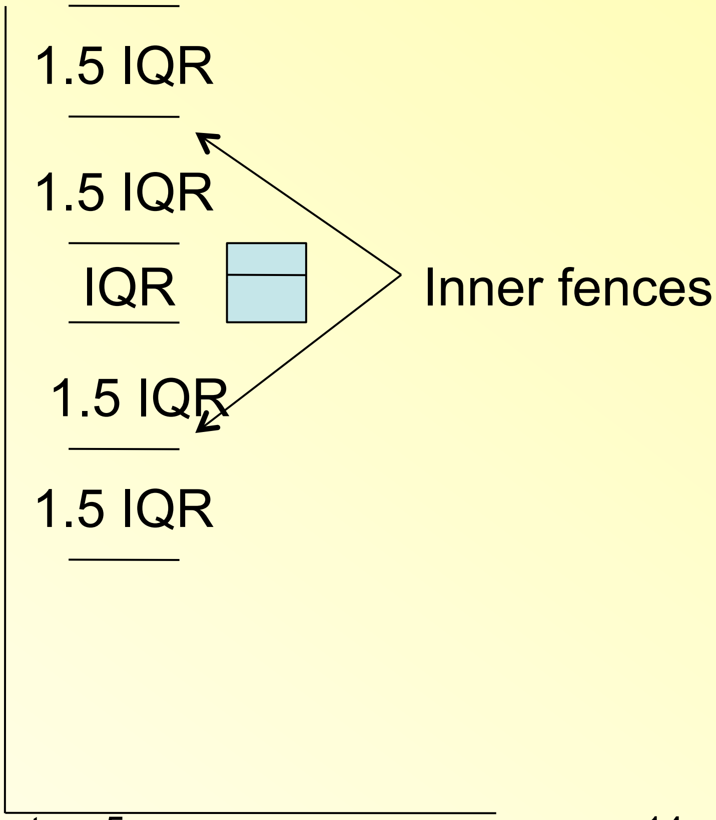
Step 2 for boxplot – The fences

- Construct the “fences”. These are NOT in the final product. They are just used to make decisions on outliers.
- Inner fences are 1.5 IQR from the box, outer fences are 3.0 IQR from the box.



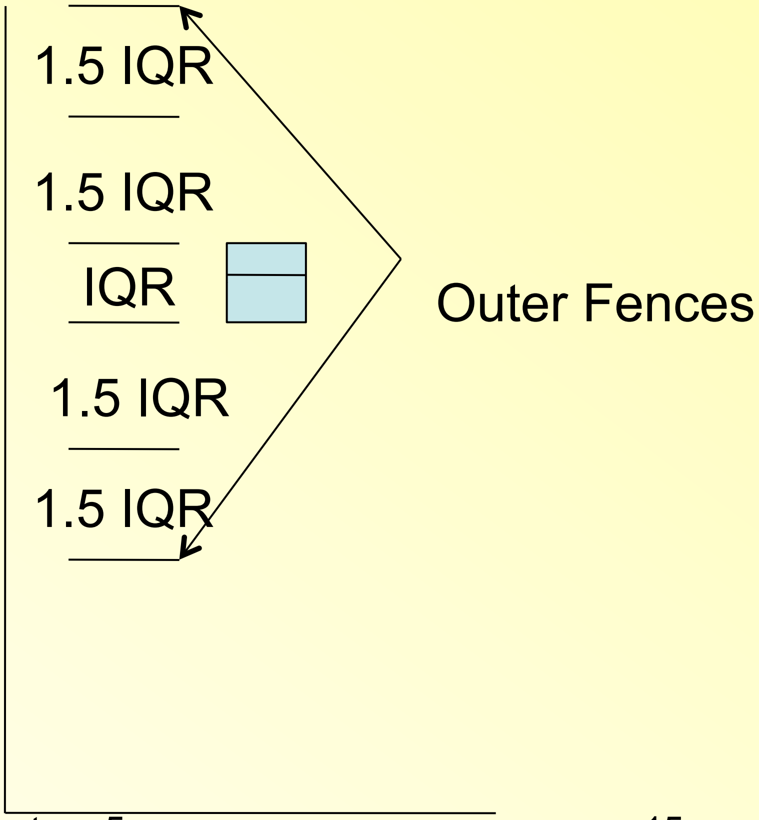
Step 2 for boxplot – Inner Fences

- Construct the "fences". These are NOT in the final product. They are just used to make decisions on outliers.
- Inner fences are 1.5 IQR from the box, outer fences are 3.0 IQR from the box.



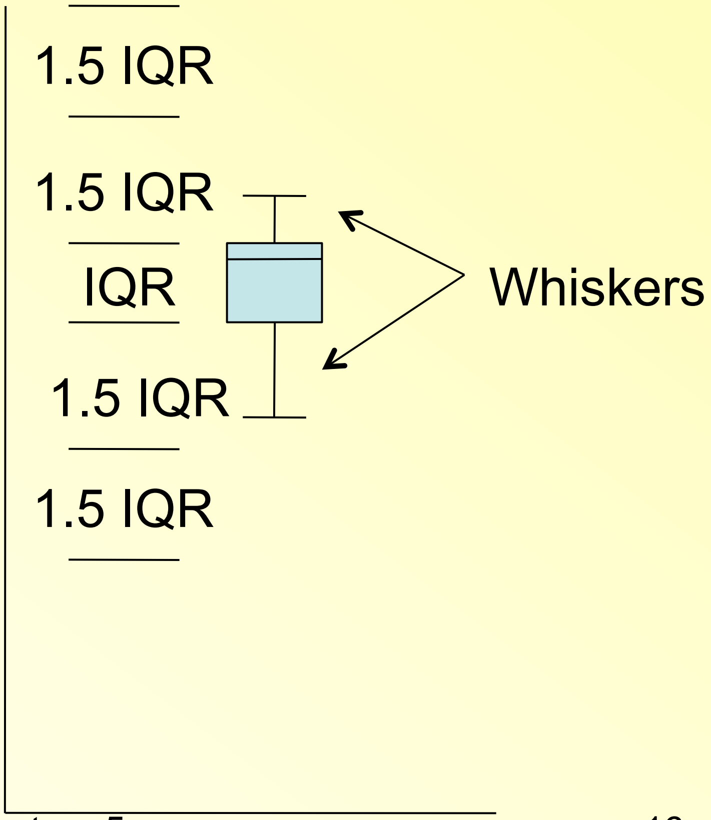
Step 2 for boxplot – Outer fences

- Construct the “fences”. These are NOT in the final product. They are just used to make decisions on outliers.
- Inner fences are 1.5 IQR from the box, outer fences are 3.0 IQR from the box.



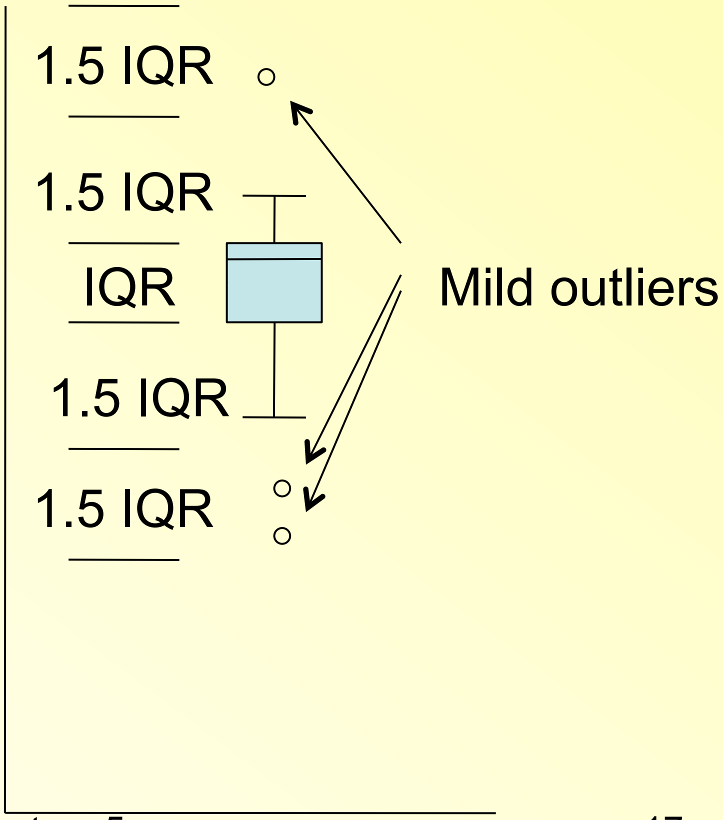
Step 3 for boxplot – Whiskers

- The whiskers extend from the box to the point closest to, but still inside, the inner fence.
- Remember, the whiskers end at a **data point**, not the inner fences.



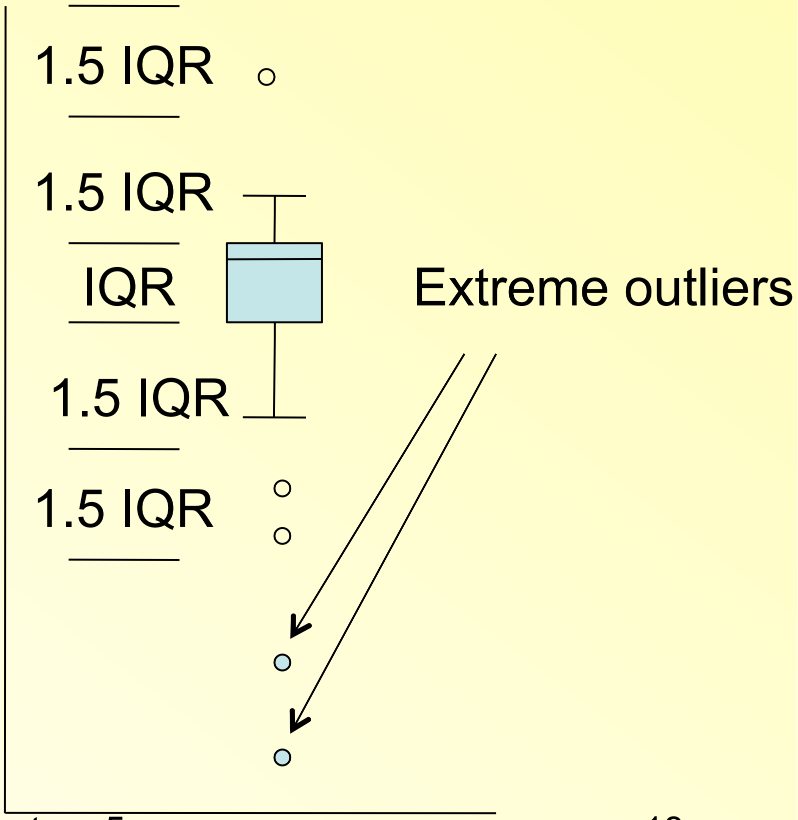
Step 4 for boxplot – Mild outliers

- Mild outliers for a boxplot are defined to be points located between the inner and outer fences.
- They are denoted by open circles.



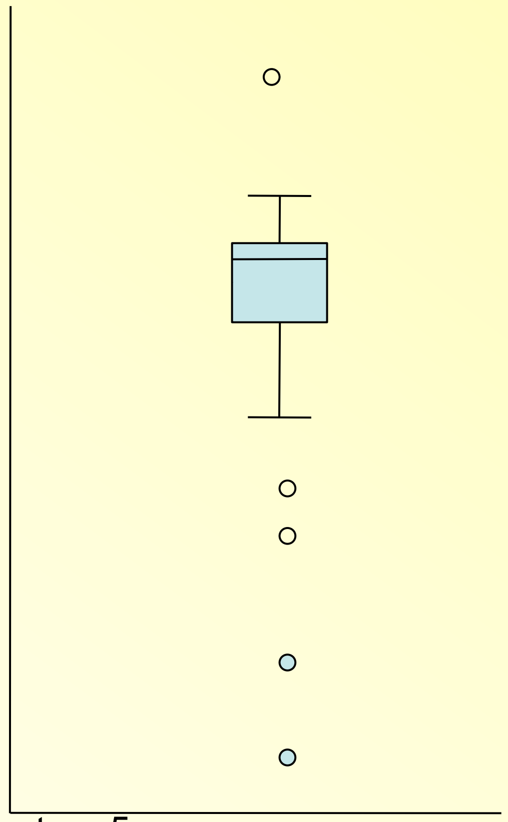
Step 5 for boxplot – Extreme outliers

- Extreme outliers for a boxplot are defined to be points located beyond the outer fences
- They are denoted (here) by filled circles.



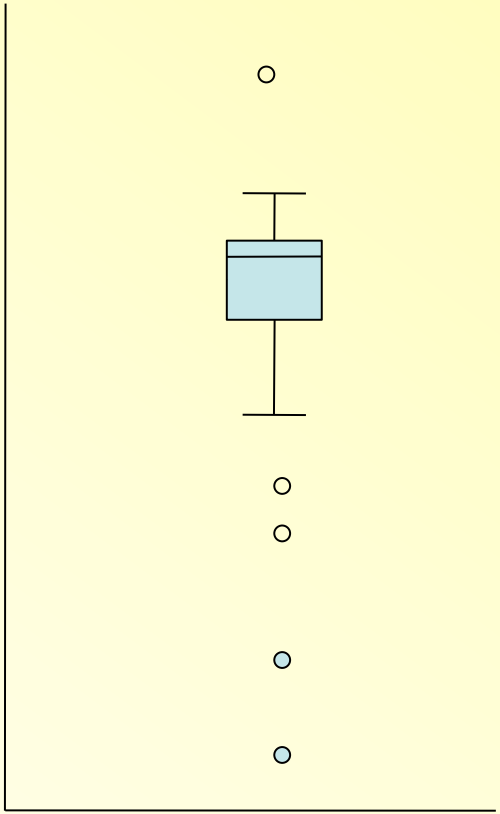
Final boxplot

- Remember, the fences are not actually drawn.
- You can see the four features of distributions easily with a boxplot. Outliers, for example, are explicitly drawn.



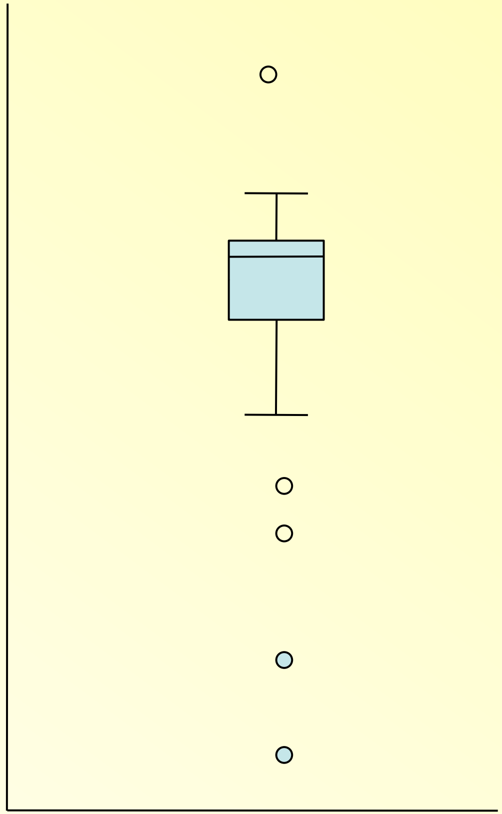
Using Boxplots

- Central location is shown through the median (some boxplots will indicate the mean by a + symbol).



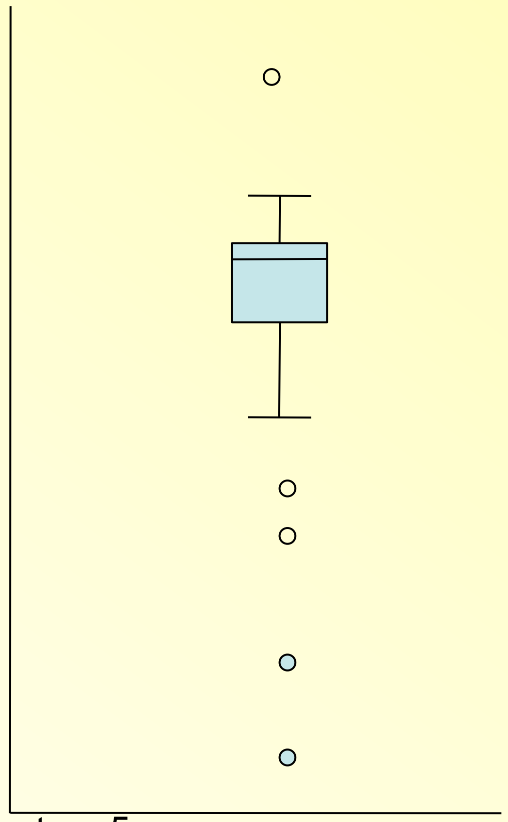
Using Boxplots

- Spread is shown through the IQR (you cannot get the standard deviation from a boxplot).
- You can also see the range of the data, but remember the range is often not that useful.



Using Boxplots

- Shape can be seen through the box and the whiskers. If one side of the box and the corresponding whisker are longer, then the data is skewed in that direction (here left skewed)
- Note: The axis should be labeled



Using boxplots

- Sometime the box “leans” one way and the whiskers the other. Then you can’t tell that much about shape from the boxplot.
- This happens most often in small datasets, where there isn’t much information about shape in the entire dataset anyway.
- In general, symmetric gets the benefit of the doubt, so a slight “lean” isn’t enough to conclude skewness.

Using boxplots

- Outliers are drawn explicitly on the plot
- You don’t have to take the definitions of “mild” and “extreme” as absolute truth, but it is a sometimes useful convention.

In SAS Code

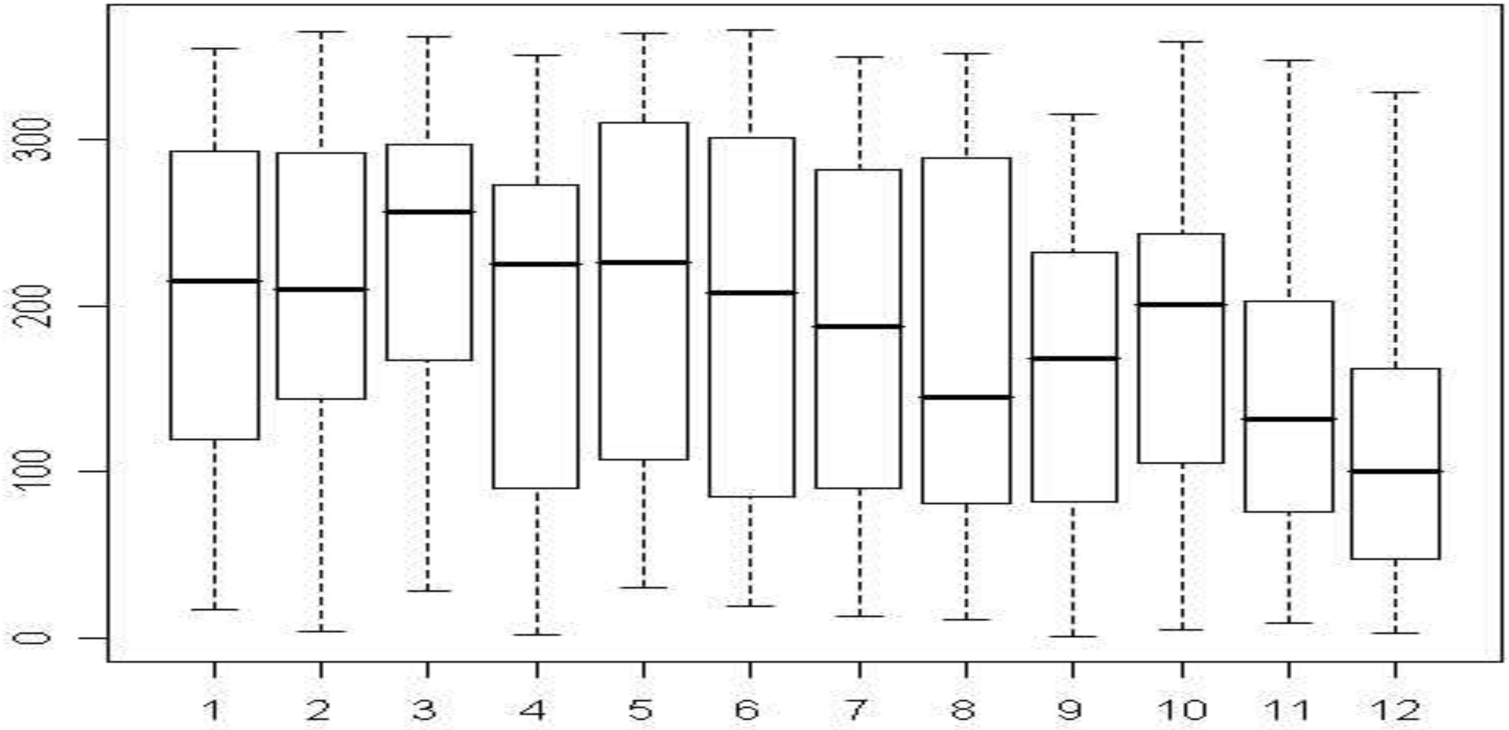
```
proc univariate data=[name] plot;
var [variable];
run;
```

- Calculates
  - – Mean, Median, Quartiles, Minimum, Maximum
  - – Range, Interquartile Range, Variance, Standard Deviation
  - – …and much more
- Creates
  - – Box Plot, Stem and Leaf Plot

Side by side boxplots

- When comparing multiple groups of people (or anything else), boxplots provide a handy method for comparison.
- By placing the boxplots side by side, you can immediately see similarities and differences in central location, spread, and shape.

1970 Draft Lottery – months on x axis, draft number on y axis.



Review: Summarizing Univariate Data Numerically

- Center of the data
  - Mean: Arithmetic average *(Interval)*
  - Median: Midpoint of the observations when they are arranged in increasing order *(Interval, Ordinal)*
  - Mode: Most frequent value *(Interval, Ordinal, Nominal)*
- Dispersion of the data
  - Variance, Standard deviation
  - Interquartile range
  - Range
- Skewness

Sample and Population Measures of Variation

- Range: maximum - minimum
- Variance (sample / population):

$$s^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-1} \qquad \sigma^2 = \frac{\sum (Y_i - \mu)^2}{N}$$

- Standard Deviation (sample / population):

$$s = \sqrt{\frac{\sum (Y_i - \overline{Y})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\sum (Y_i - \mu)^2}{N}}$$

- Interquartile Range: Q3-Q1

Sample Variance

$$s^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-1}$$

The variance of *n* observations is the sum of the squared deviations, divided by *n-1.*

Sample Variance Step By Step

1. Calculate the mean
2. For each observation, calculate the deviation
3. For each observation, calculate the squared deviation
4. Add up all the squared deviations
5. Divide the result by (n-1)

(To get the standard deviation,
take the square root of the result)

Sample Standard Deviation

- The sample standard deviation *s* is the positive square root of the sample variance

$$s = \sqrt{\frac{\sum (Y_i - \overline{Y})^2}{n-1}}$$

- Generally, interpreting the standard deviation takes experience. The *empirical rule* helps *sometimes*.

Use of the Standard Deviation

- Empirical rule: If the histogram of the data is approximately symmetric and bell-shaped, then
  - About **68%** of the data are within **one** standard deviation from the mean
  - About **95%** of the data are within **two** standard deviations from the mean
  - About **99.7%** of the data are within **three** standard deviations from the mean
- Chebysheff's inequality:
  - The proportion of observations that lie within ***k*** standard deviations of the mean is **at least $1-1/k^2$** for $k>1$

Stem and Leaf Plot, Box Plot, Empirical Rule

Example: Highway Gas Mileage for 24 Cars

```
Stem Leaf            #      Boxplot
5 0                  1         |
4                              |
4 00                 2         |
3 5678               4      +-----+
3 01112              5      *--+--*
2 557889             6      |     |
2 0134               4      +-----+
1 7                  1         |
1 3                  1         |
----+----+----+----+
Multiply Stem.Leaf by 10**+1
```

Mean=29.625　　　　Standard Deviation=8.261

Q1=24.5　　　　Median=29.5　　　　Q3=35.5

Application of the Empirical Rule

- Gas Mileage Data
  - Mean = 29.6
  - Standard Deviation = 8.3
  - 68% of the data (______observations) are supposed to be between _______ and ______
  - How many observations are actually within one standard deviation from the mean?
  - 95% of the data (______observations) are supposed to be between _______ and ______
  - How many observations are actually within two standard deviations from the mean?

Comparison to Chebycheff's Inequality

- Gas Mileage Data
  - Mean = 29.6, Standard Deviation = 8.3
  - Within 2 standard deviation of the mean are ***at least*** 1-1/4=3/4=75% of the observations

    (empirical rule: ***about 95%*** when the data is approximately bell-shaped)
  - Within 3 standard deviations are ***at least***

    1-1/9=8/9=88.9%
  - Within 4 standard deviations are ***at least***

    1-1/16=15/16=93.75%

Chebycheff’s Inequality *Always* Works

- “Number of people you have known personally who have committed suicide in the last 12 months”

Response	Frequency	Percentage	Cumulative Percentage
0	1344	88.8	88.8
1	133	8.8	97.6
2	25	1.7	99.2
3	11	0.7	99.9
4	1	0.1	100.0

•Mean = 0.15, Standard Deviation = 0.46

•***At least*** 75% of the observations are within 2 standard deviations from the means, that is, between –0.77 and 1.07.

•***At least*** 88.9% are between -1.23 and 1.53

•***At least*** 93.75% are between -1.7 and 2.0

There is also: Coefficient of Variation

- Standardized measure of variation
- Idea: A standard deviation of 10 may indicate great variability or small variability, depending on the magnitude of the observations in the data set
- CV = Ratio of standard deviation divided by mean
- Population and sample version

So far: Summarizing Univariate Data
Next: Summarizing Bivariate Data

- Two categorical variables
  - *Contingency Table*
  - *Row/Column Relative Frequencies*
  - *Relative Risk, Odds Ratio*
- Two quantitative variables
  - *Scatter Diagram*
  - *Regression Line*
  - *Correlation Coefficient, Coefficient of Determination, Slope and Intercept of Regression Line*

Quiz!

- Take a piece of paper and write your name and section number on top of it
- Please write your answers to the questions legibly.
- When you are done,
  - quietly leave your seat,
  - turn in the paper,
  - and quietly leave the room
- Question: