Chapter 2

DESCRIBING DISTRIBUTIONS WITH #s

- **focus on measures of Centre & Spread & choosing Measures of Centre & Spread are best**

A) Measures of Center

- Mean
- Median
- Other (not in text);Midrange/midpoint**, Mode**

B) Measures of Spread

- Range
- Inter Quartile Range(IQR)
- Standard Deviation

C) Numerical Summaries

A. Measures of Center

1. Mean

- Mean = arithmetic average
- most common measure
- For n cases, $x_1, x_2, x_3, \ldots x_n$ the mean is $\bar{x}$, where

$\bar{x}$ **= (sum of values)/(count of values)**

Golf Scores Example (will refer to throughout this note):

10 golf scores; 90 87 95 86 81 102 105 83 88 79

Mean calculations for Golf Scores:

Resistance

- A ***measure is resistant if its value changes only slightly to changes in a few observations, no matter how large those changes are***
- The **mean is not a resistant measure**
  -golf example: replaced 4th observation score of 86 with 126

2. Median

- Middle value(when values are ordered) , i.e. half the values are above, half the values are below
- Median for Golf Score example

- **Median is considered a resistant measure** (golf example)

2. Median(continued)

Finding the Median

- Arrange the values in order
- **For n odd**, median is the middle observation
  - e.g. 7 observations, the 4th is the median(compute (n+1)/2)
- **for n even**, the median is the average of the two middle observations
  - e.g. 8 observations, average the 4^{th} and 5^{th} observations

3. Other Measures used(for center)

(a) Midrange or Midpoint

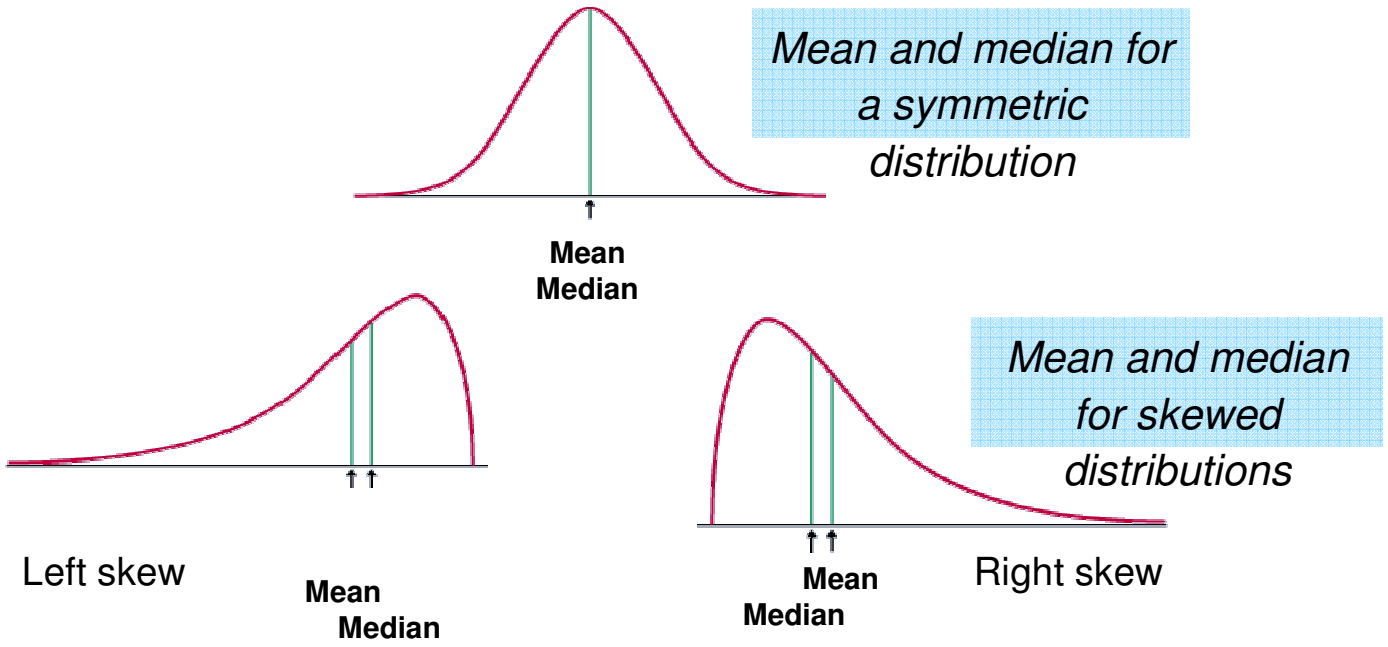
- average of the maximum and minimum values, **midrange=(max+min)/2**
- **Midrange for Golf Score example is 92 =(79+105)/2**
- **Midrange is NOT resistant-golf example if replace 86 with 126 then midrange =(79+126)/2=102.5**

(b) Mode

- most frequently occurring value(peak)
- may be more than one mode
- mode may also be far from the center
- **useful for describing categorical data**

Comparing the Mean and the Median

The mean and the median are the same only if the distribution is symmetrical. The median is a measure of center that is resistant to skew and outliers. The mean is not.



B. Measures of Spread

1. Range

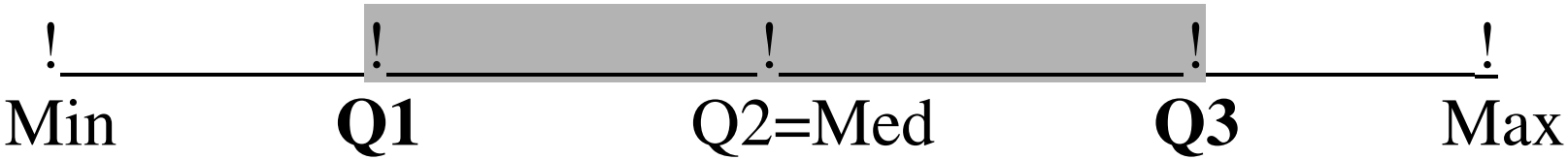
- **full spread of the data(but may include outliers)**
- **Range = max value- min. value**
- Golf example, range calculation
  - **Range is 26=105-79**
- **Range is NOT resistant**
  - **Range=47 =(126-79), if replace 86 with 126**

2. Interquartile Range(IQR)

- **IQR is measure of the spread of the middle half of the data**

2. Interquartile Range(IQR) continued

- **IQR is measure of spread of the middle half of the data**



- Q1 is 1st quartile; i.e. value that is large than ¼ of the observations; **Q1 is the median of the values below Q2**
- Q2 is the Median
- Q3 is 3rd quartile; i.e. value that is larger than ¾ of the observations**; Q3 is the median of the values above Q2**
- **IQR=Q3 – Q1**
- **IQR is** considered to be a **resistant measure**

IQR Examples

(i)Golf Example revisited(again):

(ii) Data: 9 9 22 32 33 39 39 42 49 52 58 70

3. Standard Deviation (s)

- **s measures average distance from the mean**
- understanding standard deviation: in class examples (all with n=3)
- **s is measured in same units as original observations**
- **s is always ≥0; large s value implies data more spread out**
  - **variance (also a measure of spread) denoted as s^2 or s_x^2 and s^2 is the average of the squares of the deviations from the mean**

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$s = \sqrt{s^2}$ = standard deviation (square root of variance)

- **standard deviation is <u>not</u> resistant**
- **standard deviation(s) Calculation of for Golf example**

Extra page for standard deviation(s) calculation- Golf example

Steps(optional)

1) Find the mean
2) subtract mean from each observation
3) square each value(otherwise they'll add to 0)
4) sum squared values
5) divide by (n-1) to get s^2
6) take the square root of s^2 to get s

GOLF EXAMPLE SUMMARY:

DATA: 90 87 95 86 81 102 105 83 88 79

	Original	Revised *	General Comments
Mean	89.6	93.6	Not resistant. Good for symmetric distributions without outliers
Median	87.5	89	Resistant. Good for distributions with outliers/skewness
Range	26	47	Not resistant
IQR	12	19	Resistant. Good for distributions with outliers/skewness
Standard Deviation	8.64	14.24	Not resistant. Good for symmetric distributions without outliers

*** revised: 126 replaces 86(4^{th} data value)**

Spread Measures Example

of Observations (each dataset has n=13)

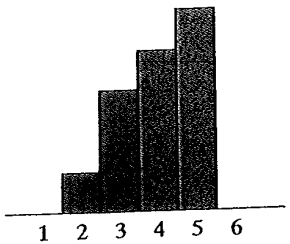
Value	Dataset I	DataSet II	Dataset III	Dataset IV
1				
2	1		1	
3	3	5	2	6
4	4	4	7	1
5	5	3	2	6
6		1	1	

NOTE:

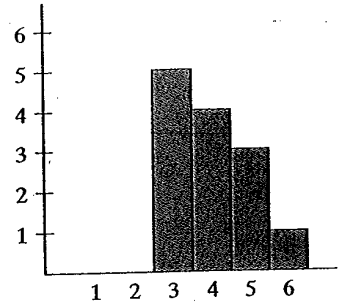
- **mean =median=4 is same for all for datasets**
- **standard deviation=1 is same for all datasets**

but the dataset Distributions are quite different!

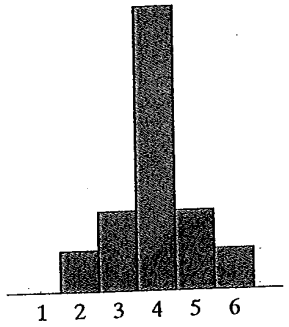
Distribution I



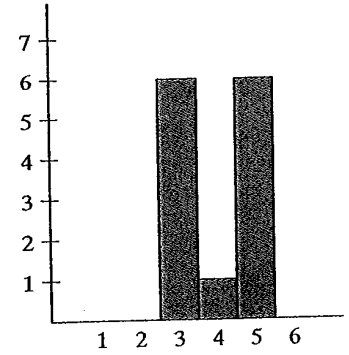
Distribution II



Distribution III



Distribution IV



Mean = Median = 4

S = 1

	I	II	III	IV
Range	3	3	4	2
IQR	2	2	1	2
Std. dev	1	1	1	1

C. NUMERICAL SUMMARIES(of center&spread)

(1) Mean and Standard Deviation

- **Best for symmetric distributions with no outliers**

(2) Five Number summary

Min Q1 M Q3 Max

- 5# summary counts through the data in order
  - reasonably complete description of centre and spread
  - 5#summary can be displayed using boxplots
- **best used for skewed distributions/those with outliers**
- **IQR useful for detecting suspected outliers**
  - Called "1.5xIQR rule for outliers"
  - Illustrate by example (next page)

Ex.-Detecting Suspected Outliers-1.5xIQR Rule

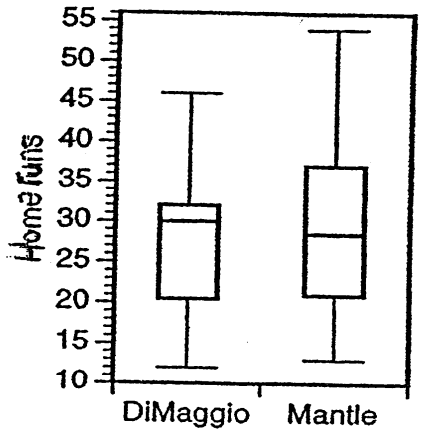
50	105	110	135	175
Min	**Q1**	**M**	**Q3**	**Max**

Note:

1.5xIQR rule is most useful for large volumes of data(should not replace looking at data)

Box Plot Example Home Run Data

	Min	Q1	Med(Q2)	Q3	Max
DiMaggio	12	20.5	30	32	46
Mantle	13	21	28.5	37	54



BOXPLOTS – a graph of the 5 number summary

- Box extends from Q1 to Q3
- Line cuts the box at median
- Lines extend from the box to min. and max. values
  - **note**: *some* boxplot displays show drawn lines to 1.5xIQR limit and show values outside limit (ie.outliers) individually, but this isn't the case with the Moore text
- most useful for comparing several distributions (compare a quantitative variable across categorical groups)
- gives an indication of symmetry of skewness of a distribution
- generally not useful for displaying single distributions since details of shape may be lost(show less detail than histograms/stemplots)

Another Boxplot Example

Babe Ruth data(# yearly home runs(1920-34)

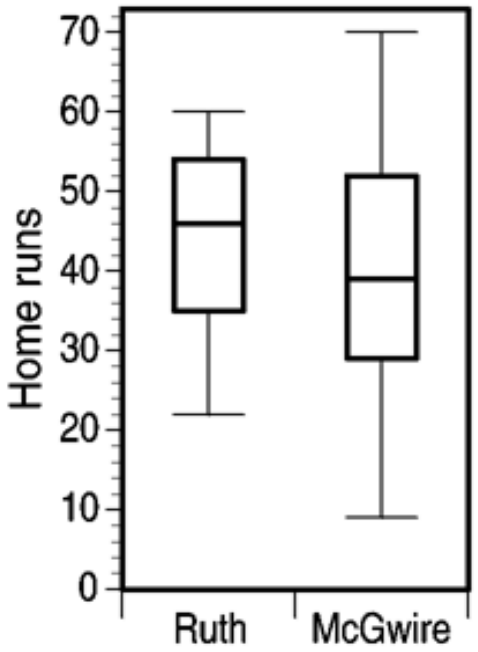
54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

Mark McGwire (# of home runs per season(1987-2001)

49 32 33 39 22 42 9 9 39 52 58 70 65 32 29

Five number Summary

	Min	Q1	M	Q3	Max
Ruth	22	35	46	54	60
McGuire	9	29	39	52	70



Note:

- can't readily assess if there are outliers from this boxplot
  - if using 1.5*IQR rule here, you wouldn't pick up the outliers either! This rule is most useful when looking at large volumes of data
- **'back to back stem & leaf plot would be more useful**

Comparing with Boxplots

- Compare medians(relative difference not absolute difference)
- Compare spread(IQR and range)
- Check for indications of skewness
- If symmetric, median is approximately in center of box, however if median is in center, distribution is not necessarily symmetric(check distance from median to extremes to)
  - **Right skew**-Q3 further above median than Q1 is below it
  - **Left skew**-Q1 further below the median than Q3 is above it
- Check for possible outliers (may use 1.5xIQR rule)

Boxplot comparisons(medians & skewness)

Is the difference in the medians significant?