Unit 1
Summarizing Data

Topics

1. Scales of Measurement …………………………………… 2

2. Descriptives for Nominal and Ordinal Data ……………… 7
   a. The Bar Chart …………………………………….. 9
   b. The Pie Chart …………………………………….. 13

3. Descriptives for Continuous Data ………………………. 14
   a. The Histogram …………………………………….. 15
   b. The Frequency Polygon ………………………… 19
   c. The Cumulative Frequency Polygon ……………. 20
   d. Percentiles (Quantiles) …………………………. 21
   e. Quantile Quantile Plot …………………………… 23
   f. Stem and Leaf Diagram ……………………….… 24
   g. Box and Whisker Plot …………………………... 25

4. The Summation Notation………………………………… 26

5. Measures of Central Tendency …………….…………… 27
   a. The mode …………………………………….. 30
   b. The mean …………………………………….. 31
   c. The median …………………………………... 34

6. Measures of Dispersion ………………………………… 36
   a. Variance ……………………………………… 37
   b. Standard Deviation …………………………….. 38
   c. Median Absolute Deviation from Median …….. 40
   d. Standard Deviation v Standard Error …………. 41
   e. A Feel for Sampling Distributions ……………. 43
   f. The Coefficient of Variation ………………….. 44
   g. The Range …………………………………….. 45

1. Scales of Measurement

The distinction between qualitative versus quantitative is a familiar one. For example, to describe a flower as pretty is a qualitative assessment while to record a child's age as 11 years is a quantitative measurement.

Consider that we can reasonably refer to the child's 22 year old cousin as being twice as old as the child whereas we cannot reasonably describe an orchid as being twice as pretty as a dandelion.

We encounter similar stumbling blocks in statistical work. Depending on the type of the variable, its scale of measurement type, some statistical methods are meaningful while others are not.

QUALITATIVE DATA

- Nominal Scale: Consists of a finite set of possible values or categories which have no particular order.

  Example: Cause of Death
  - Cancer
  - Heart Attack
  - Accident
  - Other

  Example: Gender
  - Male
  - Female

  Example: Race/Ethnicity
  - Black
  - White
  - Latino
  - Other

  Other Examples: Eye Color, Type of Car, University Attended, Occupation

- Ordinal Scale: Consists of a finite set of possible values or categories which DO have an order.

Example: Size of Container
- Small
- Medium
- Large

Example: Pain Level
- None
- Mild
- Moderate
- Severe

For analysis in the computer, both nominal and ordinal data might be stored using numbers rather than text.

Example: Race/Ethnicity
- 1 = Black
- 2 = White
- 3 = Latino
- 4 = Other

The numbers have NO meaning
They are labels ONLY

Example: Pain Level
- 1 = None
- 2 = Mild
- 3 =Moderate
- 4 = Severe

The numbers have LIMITED meaning
4 > 3 > 2 > 1 is all we know
apart from their utility as labels.

QUANTITATIVE DATA

Discrete, Count Data: When we count the number of times some event or outcome occurs, the values we can observe are limited to whole numbers: 0, 1, 2, 3, …

Examples:

Number of children a woman has had
Number of clinic visits made in one year

The numbers are meaningful
We can actually compute with these numbers.

Interval Scale: Interval data are generally measured on a continuum and differences between any two numbers on the scale are of known size.

Example: Temperature in °F on 4 successive days

Day:	A	B	C	D
Temp °F:	50	55	60	65

For these data, not only is day A with 50° cooler than day D with 65°, but it is 15° cooler. Also, day A is cooler than day B by the same amount that day C is cooler than day D (i.e., 5°).

An important property of interval scale data is that there is no true zero point. That is, the value "0" is arbitrary and doesn't reflect absence of the attribute. This is clear in the above example since a temperature of 0° doesn't mean there is no temperature.

Ratio Scale: Ratio scale are also measured on a meaningful continuum. The distinction is that ratio data have a meaningful zero point.

Example: Weight in pounds of 6 individuals
136, 124, 148, 118, 125, 142

Note on meaningfulness of "ratio"-
Someone who weighs 142 pounds is two times as heavy as someone else who weighs 71 pounds. This is true even if weight had been measured in kilograms.

• Examples of meaningful zeros:
  Height in centimeters
  Age in years

- In the sections that follow, we will see that the possibilities for meaningful description (tables, charts, means, variances, etc) are lesser or greater depending on the scale of measurement.

- The following chart gives a sense of this idea.

- For example, we'll see that we can compute relative frequencies for a nominal random variable (eg. Hair color: e.g. "7% of the population has red hair") but we cannot make statements about cumulative relative frequency for a nominal random variable (eg. it would not make sense to say "35% of the population has hair color less than or equal to blonde")

	Random Variable			
	Discrete		Continuous	
	Nominal	Ordinal	Interval	Ratio
Descriptive Methods	Bar chart Pie chart - -	Bar chart Pie chart - -	- - Dot diagram Scatter plot (2 variables) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot	- - Dot diagram Scatter plot (2 vars) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot
Numerical Summaries	Frequency Relative Frequency	Frequency Relative Frequency Cumulative Frequency	- - - means, variances, percentiles	- - - means, variances, percentiles

Note – This table is an illustration only. It is not intended to be complete.

2. Descriptives for Nominal and Ordinal Data

Consider a study of 25 consecutive patients entering the general medical/surgical intensive care unit at a large urban hospital.

- For each patient the following data are collected:

Variable	Code
• Age (in years)	
• Type of Admission:	1= Emergency 0= Elective
• ICU Type:	1= Medical 2= Surgical 3= Cardiac 4= Other
• Systolic Blood Pressure (in millimeters of mercury)	
• Number of Days Spent in ICU	
• Vital Status at Hospital Discharge:	1= Dead 0= Alive

Following are the actual data.

ID	Age	Type_Adm	ICU_Type	SBP	ICU_LOS	Vit_Stat
1	15	1	1	100	4	0
2	31	1	2	120	1	0
3	75	0	1	140	13	1
4	52	0	1	110	1	0
5	84	0	4	80	6	0
6	19	1	1	130	2	0
7	79	0	1	90	7	0
8	74	1	4	60	1	1
9	78	0	1	90	28	0
10	76	1	1	130	7	0
11	29	1	2	90	13	0
12	39	0	2	130	1	0
13	53	1	3	250	11	0
14	76	1	3	80	3	1
15	56	1	3	105	5	1
16	85	1	1	145	4	0
17	65	1	1	70	10	0
18	53	0	2	130	2	0
19	75	0	3	80	34	1
20	77	0	1	130	20	0
21	52	0	2	210	3	0
22	19	0	1	80	1	1
23	34	0	3	90	3	0
24	56	0	1	185	3	1
25	71	0	2	140	1	1

Discrete variables are:

- **Type_Adm**
- **ICU_Type**
- **Vit_Stat**

Continuous variables are:

- **Age**
- **ICU_LOS**
- **SBP**

A tally of the possible outcomes, together with "how often" and "proportionately often" is called a <u>frequency distribution.</u>

- **It is a sensible thing to do for discrete variables only.**
- **For the variable ICU_Type, the frequency distribution is the following:**

ICU_Type	Frequency ("how often")	Relative Frequency ("proportionately often")
Medical	12	0.48
Surgical	6	0.24
Cardiac	5	0.20
Other	2	0.08
TOTAL	25	1.00

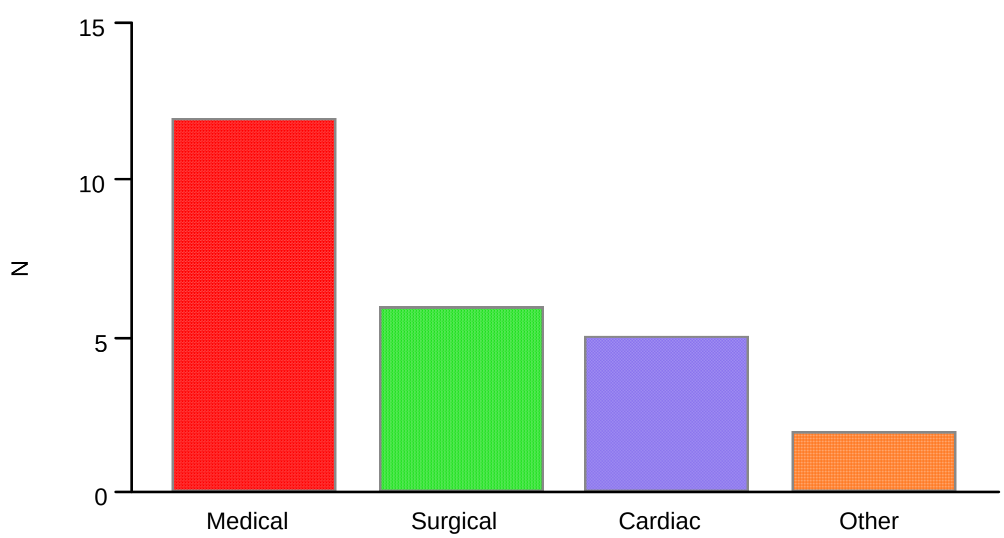
- **This summary will be useful in constructing two graphical displays, the bar chart and the pie chart.**

<u>The Bar Chart</u>

On the horizontal are the possible outcomes. On the vertical is plotted either

- **"how often" - frequency**
- **"how proportionately often" - relative frequency**

Bar chart for Type of ICU Patients



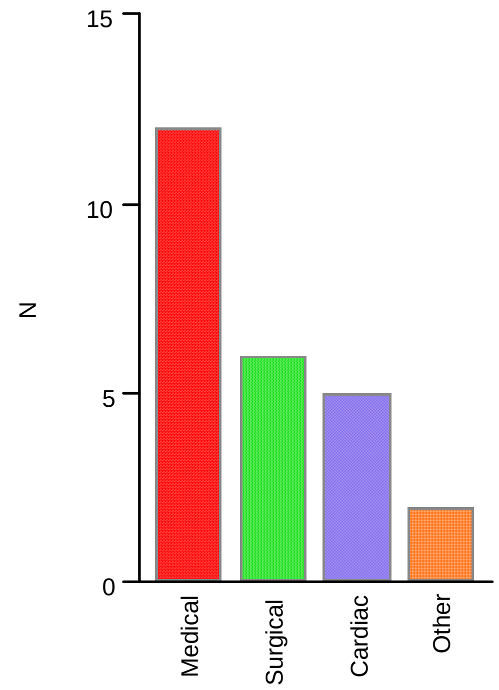
Guidelines for Construction of Bar Charts

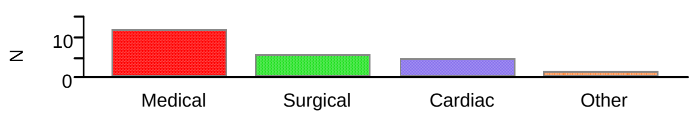
- **Label both axes clearly**
- **Leave space between bars**
- **Leave space between the left-most bar and the vertical axis**
- **When possible, begin the vertical axis at 0**
- **All bars should be the same width**

There's no reason why the bar chart can't be plotted horizontally instead of vertically.

Not surprisingly, if you change the choice of scale, you can communicate to the eye a very different message.

Consider two choices of scale for the vertical axis in the bar chart for ICU_Type:

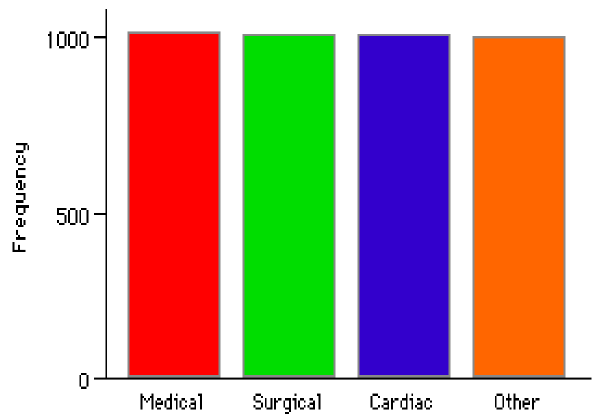
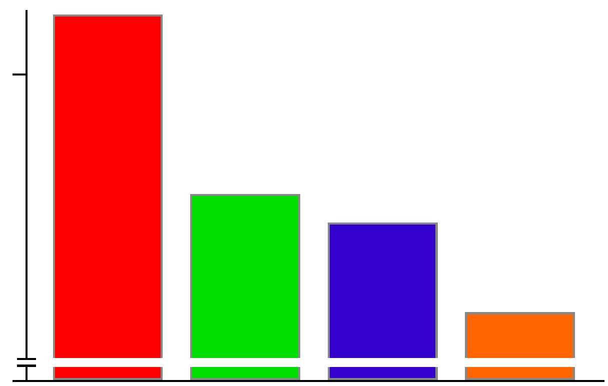




It’s difficult to know how to construct a bar chart when the frequencies are very high.

- **Suppose the frequencies were 1012 medical, 1006 surgical, 1005 cardiac and 1002 other type.**

- **Is it better to have a vertical axis start at zero or to “break” the axis?**

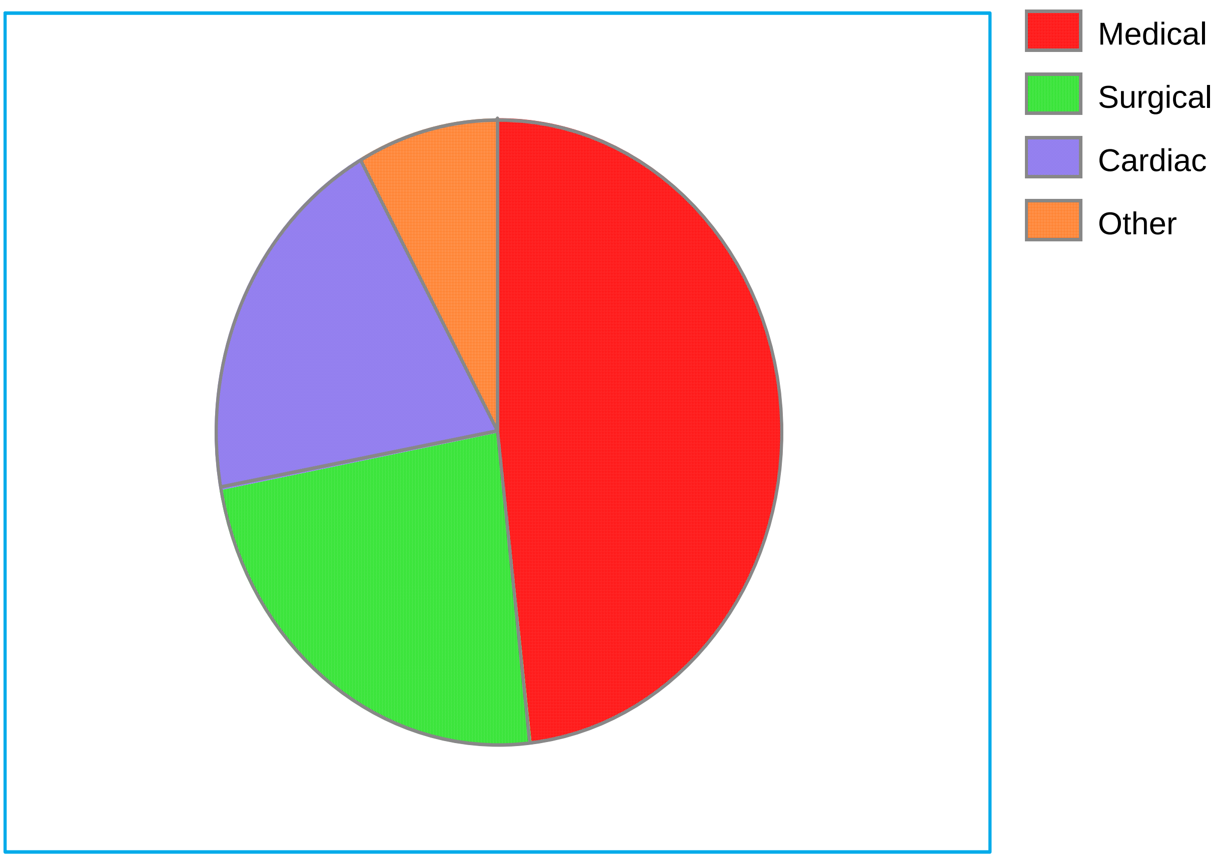
Vertical Axis Starts at Zero	**Vertical Axis has “Break”**
	

The Pie Chart

Instead of "stacks" rising up from the horizontal (bar chart), we could plot instead the shares of a pie.

Recalling that a circle has 360 degrees, that 50% means 180 degrees, 25% means 90 degrees, etc, we can identify "wedges" according to relative frequency

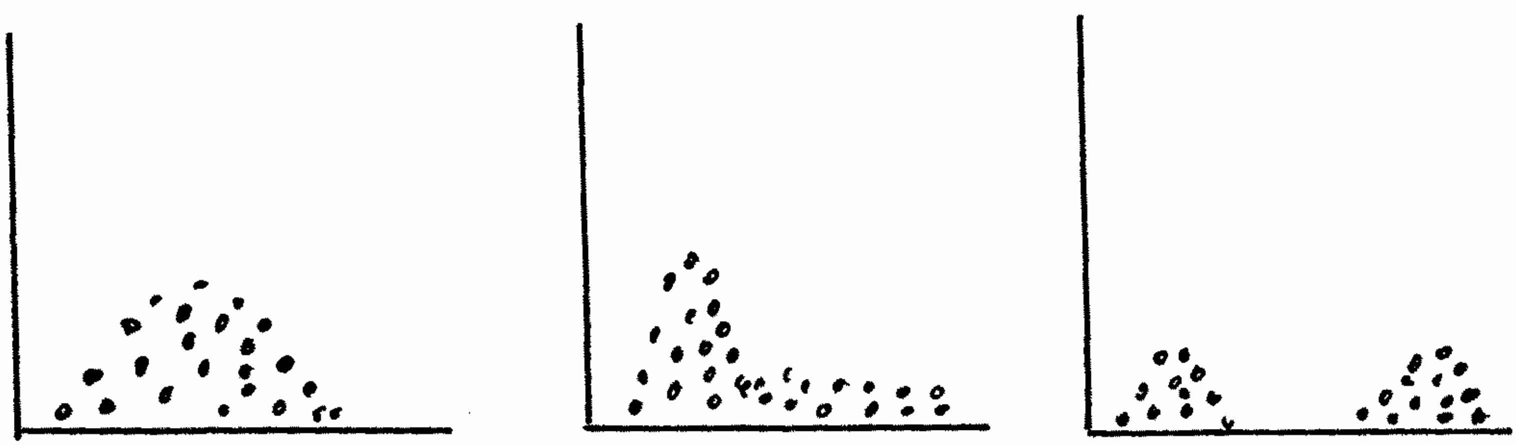
Relative Frequency	Size of Wedge, in degrees
0.50	50% of 360 = 180 degrees
0.25	25% of 360 = 90 degrees
p	(p) x (100%) x 360 degrees



3. Descriptive Statistics for Continuous Data

Tables and graphs for continuous data address many things. Among the most important are:

- **What is typical (location)**
- **What is the scatter (dispersion)**



For a continuous variable (e.g. – age), the <u>frequency distribution of the individual ages</u> is not so interesting.

Age	Frequency
15	1
19	2
29	1
31	1
34	1
39	1
52	2
53	2
45	2
65	1
71	1
74	1
75	2
76	2
77	1
78	1
79	1
84	1
85	1

We "see more" in frequencies of age values in "groupings". Here, 10 year groupings make sense.

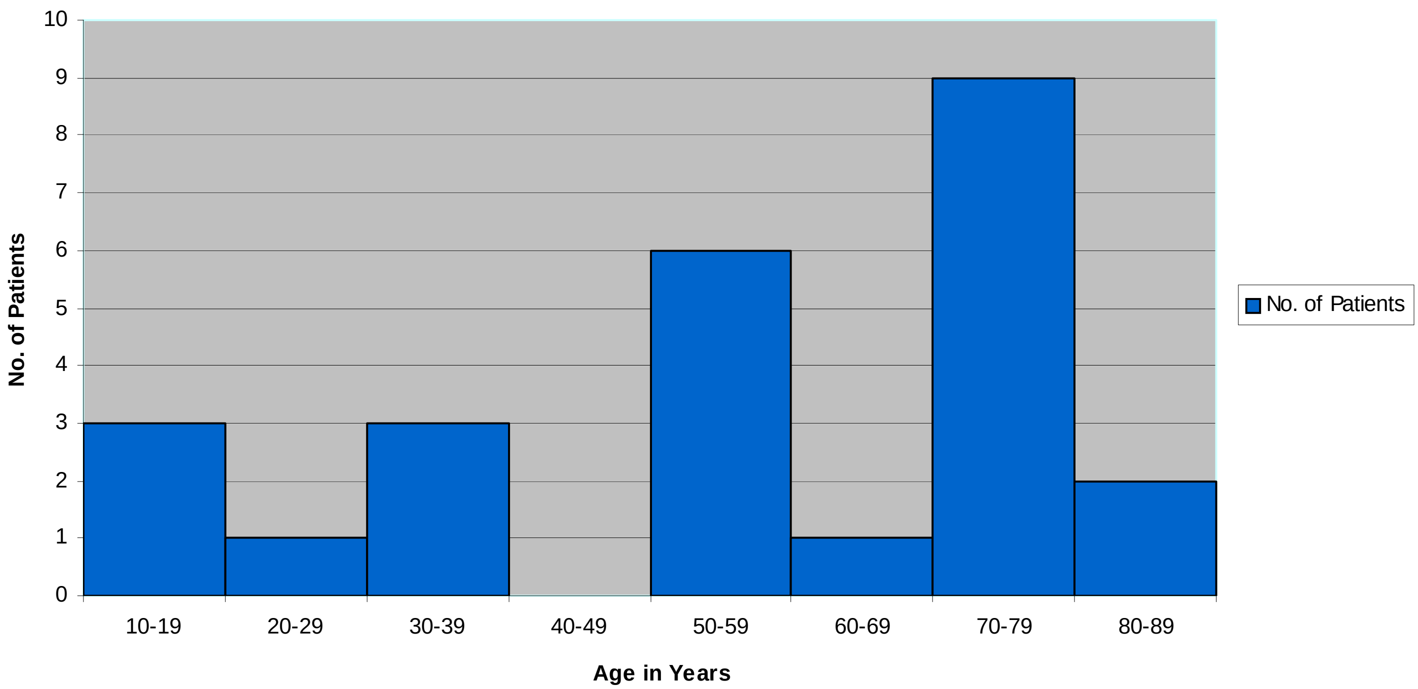
Age Interval	Frequency
10-19	3
20-29	1
30-39	3
40-49	0
50-59	6
60-69	1
70-79	9
80-89	2
TOTAL	25

Histogram

A plot of this "grouped" frequency table gives us a better feel for the pattern of ages with respect to both location and scatter. This plot is called a histogram

- **A histogram is a graphical summary of the pattern of values of a continuous random variable. More formally, it is a graphical summary of the frequency distribution**
- **It is analogous to the bar graph summary for the distribution of a discrete random variable**

Figure 12: Histogram of the Age Distribution of ICU Patients



How to Construct a Histogram

Step 1: Choose the number of groupings ("class intervals"). Call this k.

- **The choice is arbitrary.**
- **K too small over-summarizes. K too large under-summarizes.**
- **Sometimes, the choices of intervals are straightforward – eg 10 year intervals, 7 day intervals, 30 day intervals.**
- **Some text books provide formulae for k. Use these if you like, provided the resulting k makes sense.**
- **<u>Example</u> – For the age data, we use k=8 so that intervals are sensible 10 year spans.**

Step 2: Rules for interval beginning and end values ("boundaries")

- **Boundaries should be such that each observation has exactly one "home".**
- **Equal widths are not necessary. WARNING – If you choose to plot intervals that are of Unequal widths, take care to plot "area proportional to relative frequency". This is explained in Step 3.**

Step 3: In a histogram, always plot "Area Proportional to Relative Frequency"

- **An example is the best explanation.**

Age Interval	Frequency
10-29	4
30-39	3
40-49	0
50-59	6
60-69	1
70-79	9
80-89	2
TOTAL	25

- **For the intervals 30-39, 40-49, 50-59, 60-69, 70-79, 80-89: the widths are all the same and span 10 units of age. Heights plotted are 3, 0, 6, 1, 9, and 2.**

- **But the interval 10-29 spans 20 units of age. A frequency of 4 over 20 units of age corresponds to a frequency of 2 over 10 units of age.**

- **For the interval 10-29, a height of 2 would be plotted.**

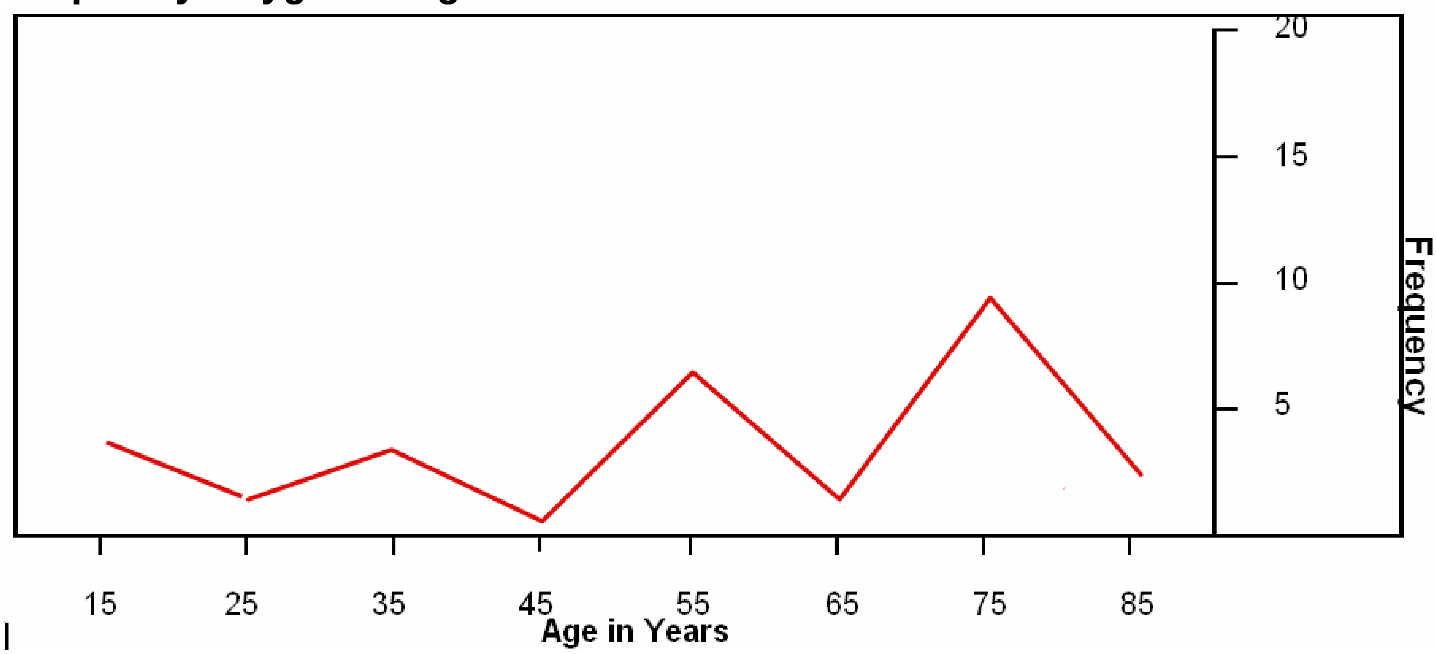
The Frequency Polygon

- The frequency polygon is an alternative to the histogram
- Both the histogram and frequency polygon are graphical summaries of the frequency distribution of a continuous random variable
- Whereas in a histogram …
  - X-axis shows intervals of values
  - Y-axis shows bars of frequencies
- In a frequency Polygon:
  - X-axis shows midpoints of intervals of values
  - Y-axis shows dot instead of bars

Some guidelines:

i. The graph title should be a complete description of the graph
ii. Clearly label both the horizontal and vertical axes
iii. Break axes when necessary
iv. Use equal class widths
v. Be neat and accurate

Frequency Polygon of Age of ICU Patients



The Cumulative Frequency Polygon

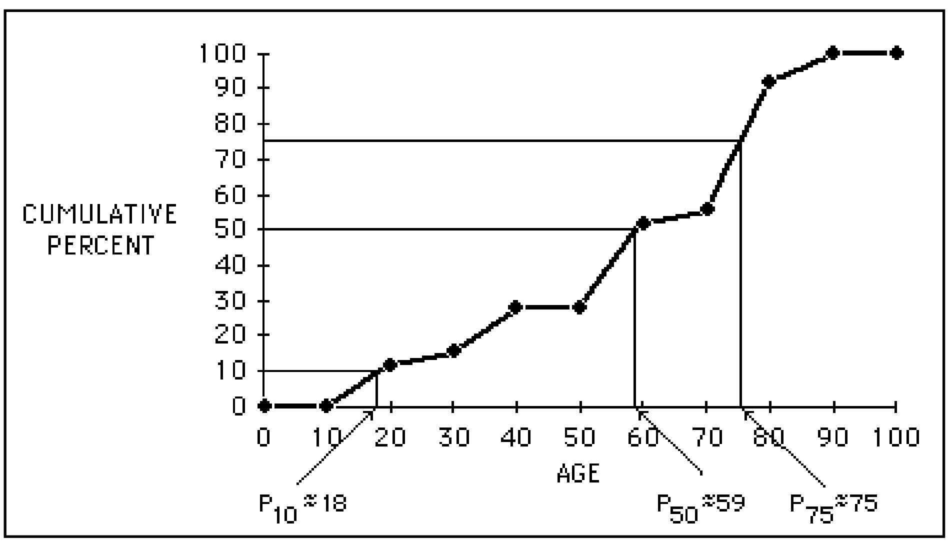
As you might think, a cumulative frequency polygon depicts accumulating data.

This is actually useful in that it lets us see percentiles (eg, median)

Its plot utilizes cumulative frequency information (right hand columns below in blue)

			Cumulative through Interval	
Age Interval	**Frequency**	**Relative Frequency**	**Frequency**	**Relative Frequency**
10-19	**3**	**12**	**3**	**12**
20-29	**1**	**4**	**4**	**16**
30-39	**3**	**12**	**7**	**28**
40-49	**0**	**0**	**7**	**28**
50-59	**6**	**24**	**13**	**52**
60-69	**1**	**4**	**14**	**56**
70-79	**9**	**36**	**23**	**92**
80-89	**2**	**8**	**25**	**100**
TOTAL	**25**	**100**		

Notice that it is the ENDPOINT of the interval that is plotted on the horizontal. This makes sense inasmuch as we are keeping track of the ACCUMULATION of frequencies to the end of the interval.



Percentiles (Quantiles)

Suppose that 50% of a cohort survived at least 4 years.

This also means that 50% survived at most 4 years.

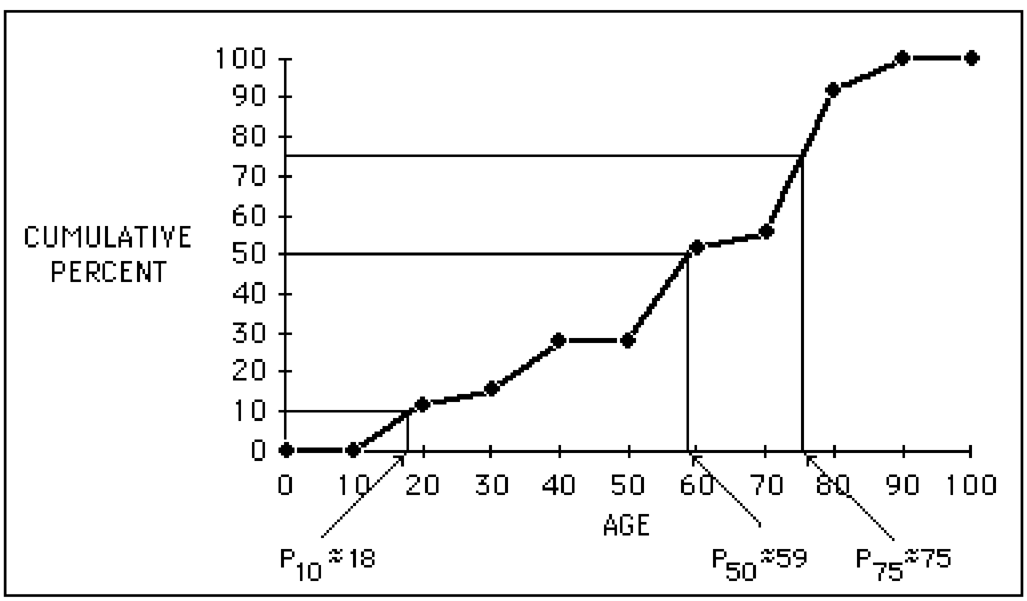
We say 4 years is the median.

The median is also called the 50th percentile, or the 50th quantile. We write P_{50} = 4 years.

Similarly we could speak of other percentiles:

P_{25}: 25% of the sample values are less than or equal to this value.
P_{75}: 75% of the sample values are less than or equal to this value.
P_0: The minimum.
P_{100}: The maximum.

It is possible to estimate the values of percentiles from a cumulative frequency polygon.



How to Determine the Value of the pth Percentile in a Data Set

Consider the following sample of n=40 data values

0	1	1	3	17	32	35	44	48	86
87	103	112	121	123	130	131	149	164	167
173	173	198	208	210	222	227	234	245	250
253	256	266	277	284	289	290	313	477	491

Step 1:
Order the data from smallest to largest

Step 2:

Compute $L = n\left[\frac{p}{100}\right]$ where

n = size of sample (eg; n=40 here)
p = desired percentile (eg p=25^{th})

L is NOT a whole number

Step 3:

Change L to next whole number.

Pth percentile = Lth ordered value in the data set.

L is a whole number

Step 3:

Pth percentile = average of the Lth and (L+1)st ordered value in the data set.

Solution for 25^{th} Percentile

L=(40)(25/100) = 10 → 25^{th} percentile is average of 10^{th} and 11^{th} → = (1/2)(86 + 87) = 86.5

Quantile-Quantile (QQ) and Percentile-Percentile (PP) Plots

- **We might ask:** ***"Is the distribution of my data normal?"***
  QQ and PP plots are useful when we want to compare the percentiles of our data with the percentiles of some reference distribution (eg- reference is normal)

- **QQ Plot: X= quantile in sample Y=quantile in reference**

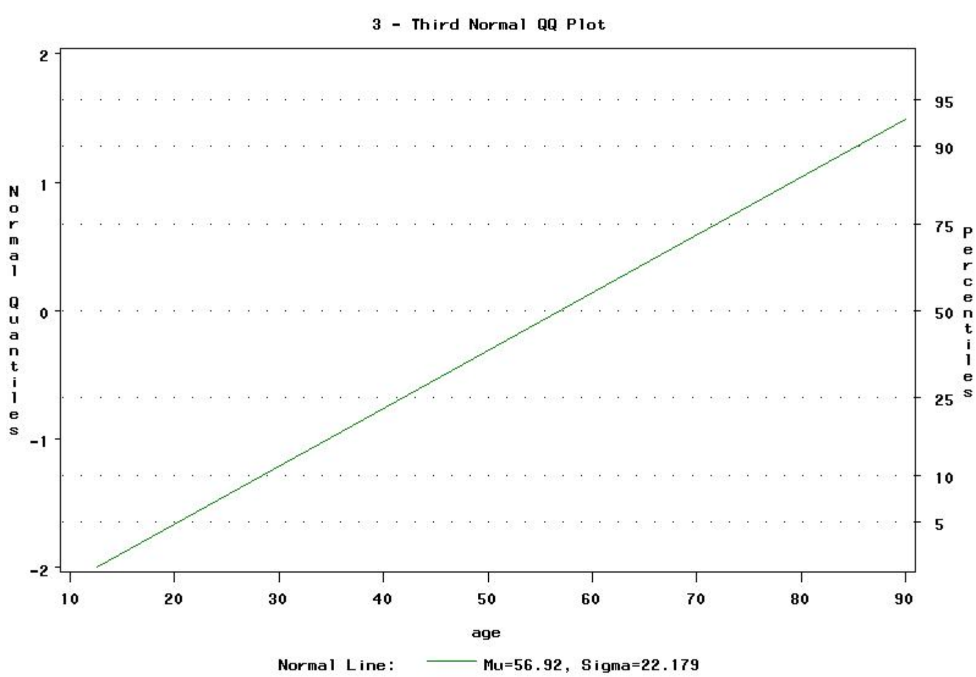
Desired Quantile	**10th**	**20th**	**…**	**99th**
X = Value in Sample of Data				
Y = Value in Reference Distribution				

- **PP Plot: X= percentile rank in sample Y= percentile rank in reference**

Data value	**##**	**##**	**…**	**##**
X = Rank (percentile) in Sample of Data				
Y = Rank (percentile) in Reference Distribution				

- **What to look for: A straight line suggests that the two distributions are the same**

Example – Normal Quantile-Quantile Plot of Age for data on page 8



Note on the Y-axis values: These have been standardized (more on this later)

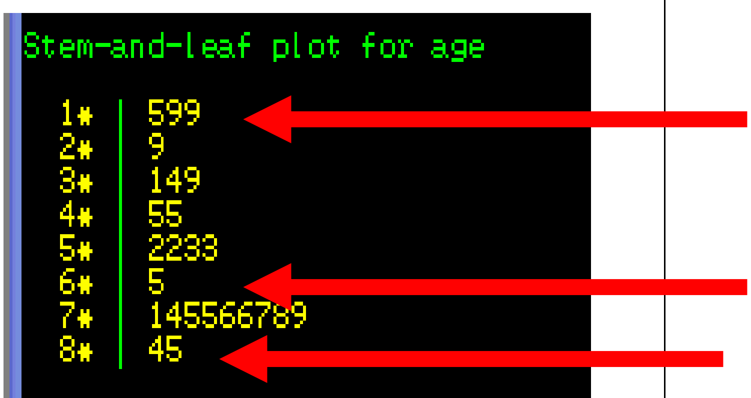
Stem and Leaf Diagram

- A stem and leaf diagram is a “quick and dirty” histogram
- It is also a quick easy and easy way to sort data.

Example - Stem and Leaf Plot of Age of 25 ICU Patients:

- The stems are to the left of the vertical. In this example, each value of stem represents a multiple of 10.
    - *Step 1*: List your stems
- The leaves are to the right of the vertical.

Step 2: “Plot” your leaves.



- **15 is stem=1 + leaf=5**
- **65 is stem=6 + leaf=5**
- **84 is stem=8 + leaf=4**

Box and Whisker Plot

- **The box and whisker plot (also called box plot) is a wonderful schematic summary of the distribution of values in a data set.**

- **It shows you a number of features, including: extremes, 25th and 75th percentile, median and sometimes the mean.**

- **Side-by-side box and whisker plots are a wonderful way to compare multiple distributions.**

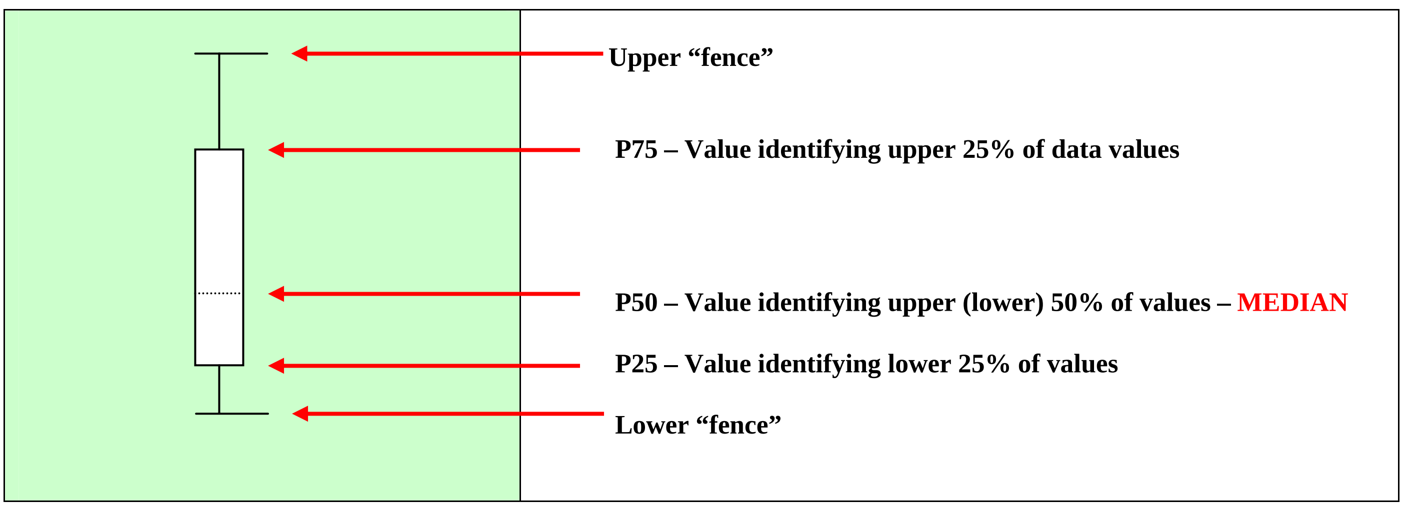
- **Definition**

  - **The central box has P_{25} and P_{75} for its limits. It spans the middle half of the data.**

  - **The line within the box identifies the median, P_{50}. Sometimes, an asterisk within the box is shown. It is the mean. The lines coming out of the box are called the "whiskers". The ends of these "whiskers' are called "fences".**

  - **Upper "fence" = The largest value that is less than or equal to $P_{75} + 1.5*(P_{75} - P_{25})$.**

  - **Lower "fence" = The smallest value that is greater than or equal to $P_{25} - 1.5(P_{75} - P_{25})$.**



Note on the multiplier 1.5

This is a convenient multiplier if we are interested in the comparability of the distribution of our sample values to a normal (Gaussian) distribution in the following way. If the data are from a normal distribution, then 95% of the data values will fall within the range defined by the lower and upper fences.

4. The Summation Notation

The summation notation is nothing more than a <u>secretarial convenience</u>. We use it to avoid having to write out long expressions.

For example,

Instead of writing $x_1 + x_2 + x_3 + x_4 + x_5$ **,**

We write $\sum_{i=1}^{5} x_i$

Another example –

Instead of writing $x_1 * x_2 * x_3 * x_4 * x_5$ **,**

We write $\prod_{i=1}^{5} x_i$

This is actually an example of the product notation

<u>The summation notation</u>

$\sum$ **The Greek symbol sigma says "add up some items"**

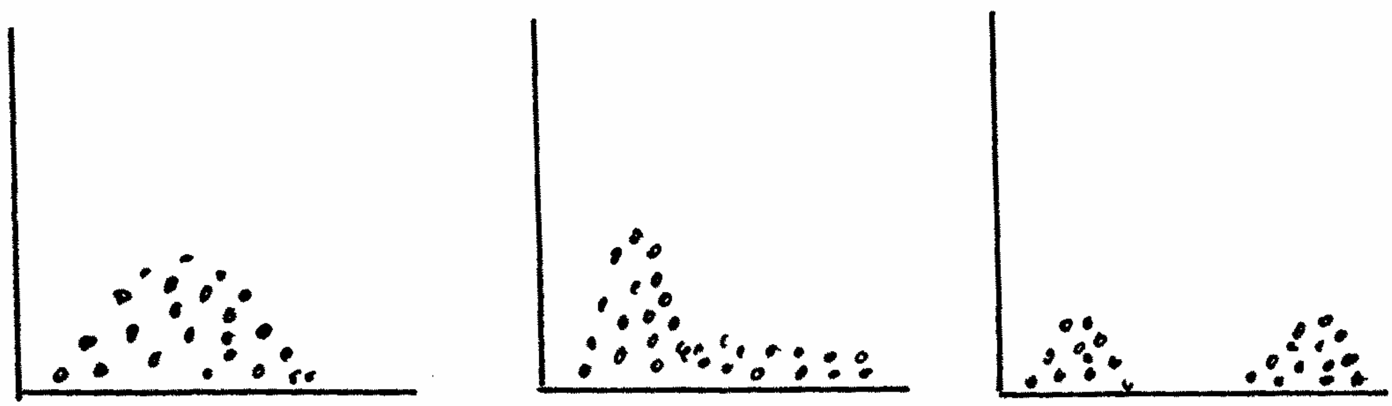
$\sum_{\text{STARTING HERE}}$ **Below the sigma symbol is the starting point**

$\sum^{\text{END}}$ **Up top is the ending point**

5. Measures of Central Tendency

Previously we noted that among the most important tools of description are that address

- **What is typical (location)**
- **What is the scatter (dispersion)**



There are choices for describing *location*

- **Arithmetic average (*mean*)**
- **Middle most (*median*)**
- **“Benchmarks” (*percentiles*)**

<u>Mode.</u> The mode is the most frequently occurring value. It is not influenced by extreme values. Often, it is not a good summary of the majority of the data.

<u>Mean.</u> The mean is the arithmetic average of the values. It is sensitive to extreme values.

$$\textbf{Mean} = \frac{\textbf{sum of values}}{\textbf{sample size}} = \frac{\Sigma\ (\textbf{values})}{\textbf{n}}$$

<u>Median.</u> The median is the middle value when the sample size is odd. For samples of even sample size, it is the average of the two middle values. It is not influenced by extreme values.

We consider each one in a bit more detail …

The Mode

Mode. **The mode is the most frequently occurring value. It is not influenced by extreme values. Often, it is not a good summary of the majority of the data.**

Example

- **Data are: 1, 2, 3, 4, 4, 4, 4, 5, 5, 6**
- **Mode is 4**

Example

- **Data are: 1, 2, 2, 2, 3, 4, 5, 5, 5, 6, 6, 8**
- **There are two modes – value 2 and value 5**
- **This distribution is said to be "bi-modal"**

Modal Class

- **For grouped data, it may be possible to speak of a modal class**
- **The modal class is the class with the largest frequency**

Example

Interval of Values ("Class")	Frequency, f
31-40	1
41-50	2
51-60	5
61-70	15
71-80	25
81-90	20
91-100	12

- **The modal class is the interval of values 71-80**

The Mean

Mean. **The mean is the arithmetic average of the values. It is sensitive to extreme values.**

$$\textbf{Mean} = \frac{\textbf{sum of values}}{\textbf{sample size}} = \frac{\boldsymbol{\Sigma}\ \textbf{(values)}}{\textbf{n}}$$

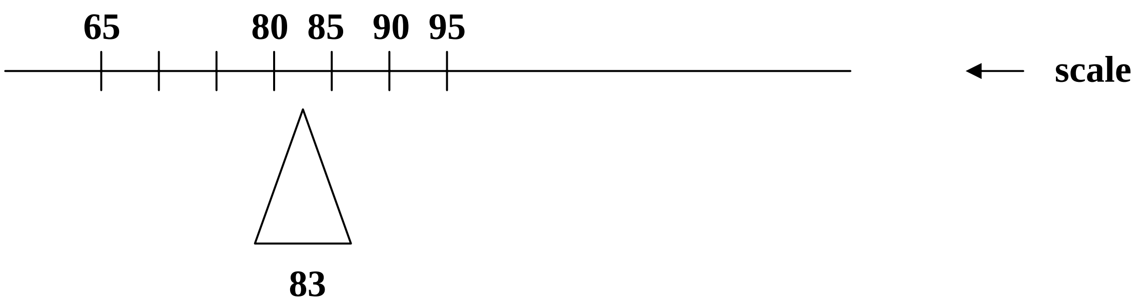
Calculation of a "mean" or "average" is familiar; e.g. -

grade point average
mean annual rainfall
average weight of a catch of fish
average family size for a region

The summation notation is a convenient secretarial shorthand

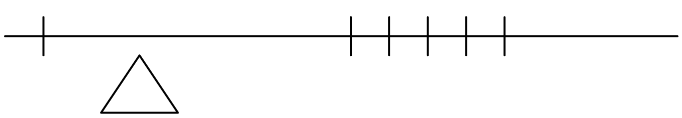
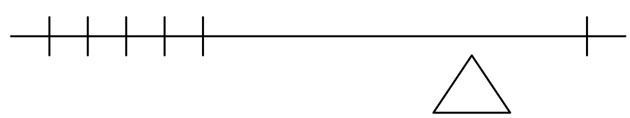
- **Suppose data are: 90, 80, 95, 85, 65**
- sample mean = $\dfrac{90+80+95+85+65}{5} = \dfrac{415}{5} = 83$
- **sample size, n = 5**
- $x_1 = 90, x_2 = 80, x_3 = 95, x_4 = 85, x_5 = 65$
- $\overline{X}$ **= sample mean**
- $\overline{X} = \dfrac{\sum_{i=1}^{5} x_i}{n} = \dfrac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \dfrac{90 + 80 + 95 + 85 + 65}{5} = 83$

<u>The mean can be thought of as a "balancing point", "center of gravity"</u>



- **Net magnitude of scores below balance out the net magnitude of scores above.**
- **In this example, sample mean** $\overline{X} = 83$
- **Often, the value of the sample mean is not one that is actually observed**

<u>When the data are skewed, the mean is "dragged" in the direction of the skewness</u>

Negative Skewness (Left tail)	**Positive Skewness (Right tail)**
 **Mean is dragged left**	 **Mean is dragged right**

The Weighted Mean

- **In a weighted mean, separate outcomes have separate influences.**
- **The influence attached to an outcome is the <u>weight</u>.**
- **Familiar is the calculation of a course grade as a <u>weighted average</u> of scores on separate outcomes.**
- **Example –**

Outcome, X_i is	**Weight (Influence) w_i is**
Homework #1 = $X_1 = 96$	$w_1 = 20$ **(20% of grade)**
Homework #2= $X_2 = 84$	$w_2 = 20$
Homework #3 = $X_3 = 65$	$w_3 = 20$
Midterm Exam = $X_4 = 73$	$w_4 = 10$
Final Exam = $X_5 = 94$	$w_5 = 30$

$$\text{weighted mean} = \frac{\sum (\text{weight})(\text{score})}{\sum (\text{weight})}$$

$$= \frac{\sum_{i=1}^{n} (w_i)(X_i)}{\sum (w_i)}$$

$$= \frac{(20)(96) + (20)(84) + (20)(65) + (10)(73) + (30)(94)}{(20) + (20) + (20) + (10) + (30)}$$

$$= 84.5$$

The Median

Median. **The median is the middle value when the sample size is odd. For samples of even sample size, it is the average of the two middle values. It is not influenced by extreme values.**

If the sample size n is ODD	$\text{median} = \frac{n+1}{2}\text{th largest value}$
If the sample size n is EVEN	$\text{median} = \text{average of}\left(\left[\frac{n}{2}\right]th, \left[\frac{n+2}{2}\right]th\right)\text{values}$

Example

- **Data, from smallest to largest, are: 1, 1, 2, 3, 7, 8, 11, 12, 14, 19, 20**
- **The sample size, n=11**
- **Median is the** $\frac{n+1}{2}\text{th largest} = \frac{12}{2} = \text{6th largest value}$
- **Thus, median value is = 8**
- **Five values are smaller than 8; five values are larger.**

Example

- **Data, from smallest to largest, are: 2, 5, 5, 6, 7, 10, 15, 21, 22, 23, 23, 25**

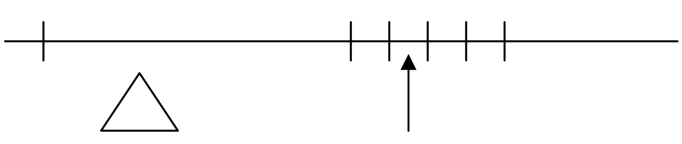
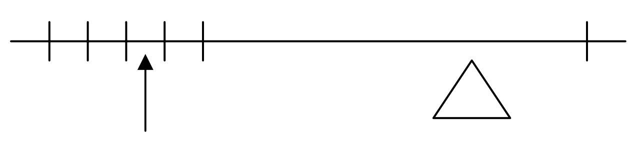
- **The sample size, n=12**
- **Median = average of** $\frac{n}{2}$th largest, $\frac{n+2}{2}$ = average of 6th and 7th largest values

- **Thus, median value is = average (10, 15) = 12.5**

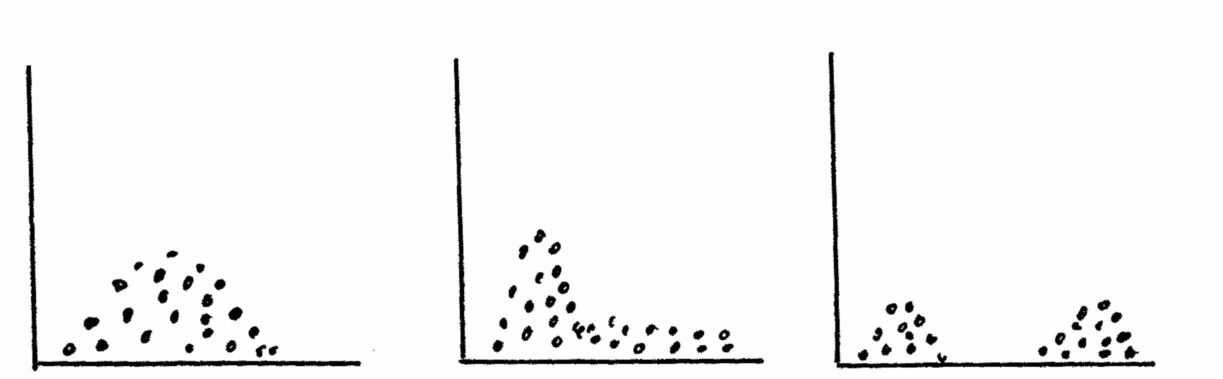
The Median is a Better Description (than the Mean) of the Majority When the Distribution is Skewed

Example

- **Data are: 14, 89, 93, 95, 96**
- **Skewness is reflected in the outlying low value of 14**
- **The sample mean is 77.4**
- **The median is 93**

Negative Skewness (Left tail)	**Positive Skewness (Right tail)**
MEAN < Median	**MEAN > Median**
 **Mean is dragged left** **MEDIAN**	 **MEDIAN** **Mean is dragged right**

6. Measures of Dispersion



There are choices for describing *dispersion*, too.

- ***Variance (S^2)***
- ***Standard Deviation (S or SD)***
- ***Median of absolute deviation from median (MADM)***

A question that is often asked is – *What is the distinction between the standard deviation (SD or S) and the standard error (SE)?*

In what follows, we will see that

the SD or S addresses questions about variability of <u>individuals</u> (imagine a collection of individuals) whereas the SE addresses questions about the variability of a <u>summary statistic</u> (imagine repeating your study multiple times so as to obtain a collection of summaries)

The Variance

Thinking first about individuals ….

Variance. **The variance is a summary measure of the squares of individual departures from the mean.**

In a population, the variance of individual values is written σ^2.

A sample variance is calculated for a sample of individual values and uses the sample mean (e.g. $\overline{X}$) rather than the population mean μ. It is written S^2 and is almost, but not quite, calculated as the arithmetic average of

$$\text{(value – mean)}^2 = (X - \overline{X})^2$$

the divisor is (sample size – 1) = (n-1) instead of (sample size).

$$S^2 = \frac{\text{sum of (value – sample mean)}^2}{\text{sample size – 1}}$$

$$= \frac{\sum \text{(value – sample mean)}^2}{n - 1}$$

$$= \frac{\sum_{i=1}^{n} (X - \overline{X})^2}{n - 1}$$

The Standard Deviation

Standard Deviation. The standard deviation is the square root of the variance. This produces a measure having the same scale as that of the individual values.

In the population, this is σ.

In the sample, this is S or SD. We use the standard deviation when we want to describe typical variability of individual values.

$$\textbf{Standard Deviation (S or SD)} = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$$

Disparity between individual and average …	$(X-\overline{X})$
Disparity between individual and average …	$(X-\overline{X})^2$
The average of these …	$\frac{\sum(X-\overline{X})^2}{n}$
The sample variance S^2 is an "almost" average	$\frac{\sum(X-\overline{X})^2}{n-1}$
The related measure S (or SD) returns measure of dispersion to original scale of observation …	$\sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$

Example of Sample Variance (S^2) and Standard Deviation (S) Calculation –

Following are the survival times of n=11 patients after heart transplant surgery. Interest is to calculate the sample variance and standard deviation.

- **Patients are identified numerically, from 1 to 11.**
- **The survival time for the "ith" patient is represented as X_i for i= 1, …, 11.**

Patient Identifier, "i"	Survival (days), X_I	Mean for sample, $\overline{X}$	Deviation , $(X_i - \overline{X})$	Squared deviation $(X_i - \overline{X})^2$
1	135	161	-26	676
2	43	161	-118	13924
3	379	161	218	47524
4	32	161	-129	16641
5	47	161	-114	12996
6	228	161	67	4489
7	562	161	401	160801
8	49	161	-112	12544
9	59	161	-102	10404
10	147	161	-14	196
11	90	161	-71	5041
TOTAL	1771		0	285236

- $\sum_{i=1}^{11} X_i = 1771$ **days**

- $\overline{X} = \frac{1771}{11} = 161$ **days**

- $s^2 = \frac{\sum_{i=1}^{11}(X_i - \overline{X})^2}{n-1} = \frac{285236}{10} = 28523.6$ **days^2**

- $s = \sqrt{s^2} = \sqrt{28523.6} = 168.89$ **days**

Median Absolute Deviation About the Median (MADM)
Still Thinking About Individuals …

Median Absolute Deviation about the Median (MADM) **- Another measure of variability is helpful when we wish to describe scatter among data that is skewed.**

Recall that the median is a good measure of location for skewed data because it is not sensitive to extreme values.

Distances are measured about the median, not the mean.

We compute deviations rather than squared differences.

Thus

Median Absolute Deviation about the Median (MADM)

MADM = $median[|X_i - median\{X_1, ..., X_n\}|]$

Example.

Original data: { 0.7, 1.6, 2.2, 3.2, 9.8 }

Median = 2.2

X_I	$\lvert X_i$ – median $\rvert$
0.7	1.5
1.6	0.6
2.2	0.0
3.2	1.0
9.8	7.6

MADM = median { 0.0, 0.6, 1.0, 1.5, 7.6 } = 1.0

Standard Deviation (S or SD) versus Standard Error (SE)

**Think not about the variability of individuals. Instead,
Imagine now that the study is repeated many times ….**

Variability of Individuals, variability of statistics and the idea of sampling distributions…

So far, we have described the variability of individuals.

We often want to describe also the variability of a statistic from one conduct of our study to the next. This is what sampling distributions are about.

Standard Error (SE). The standard error (SE) is used to describe the variability among separate sample means.

For example, imagine 5,000 samples, each of the same size n=11. This would produce 5,000 sample means. This new collection has its own pattern of variability. We describe this new pattern of variability using the standard error, not the standard deviation.

Distinction between Standard Deviation and Standard Error .

The typical variability among individual values is described using the standard deviation (SD).

The typical variability of a statistic from one sample to another (e.g. the sample mean) is described using the standard error (SE). Under reasonable assumptions,

$$SE\left(\overline{X}\right) = \frac{SD}{\sqrt{n}}$$

Note the limitation of the SE is that it is a function of both the natural variation (SD in the numerator) and the study design (n in the denominator).

Example of Investigation of Heart Transplant Surgery – continued -

Previously, we summarized the results of one study that enrolled n=11 patients after heart transplant surgery. For that one study, we obtained an average survival time of $\overline{X} = 161$ days.

What happens if we repeat the study? What will our next $\overline{X}$ be? Will it be close? How different will it be? Focus here is on the generalizability of study findings.

> **The behavior of $\overline{X}$ from one replication of the study to the next replication of the study is referred to as the sampling distribution of $\overline{X}$.**
>
> **(We could just as well have asked about the behavior of the median from one replication to the next (sampling distribution of the median) or the behavior of the SD from one replication to the next (sampling distribution of SD).)**

Thus, interest is in a measure of the "noise" that accompanies $\overline{X} = 161$ days. **The measure we use is the standard error measure. This is denoted SE. For this example, in the heart transplant study**

$$SE(\overline{X}) = \frac{SD}{\sqrt{n}} = \frac{168.89}{\sqrt{11}} = 50.9$$

We interpret this to mean that a similarly conducted study might produce an average survival time that is near 161 days, give or take 50.9 days.

A Feel for Sampling Distributions

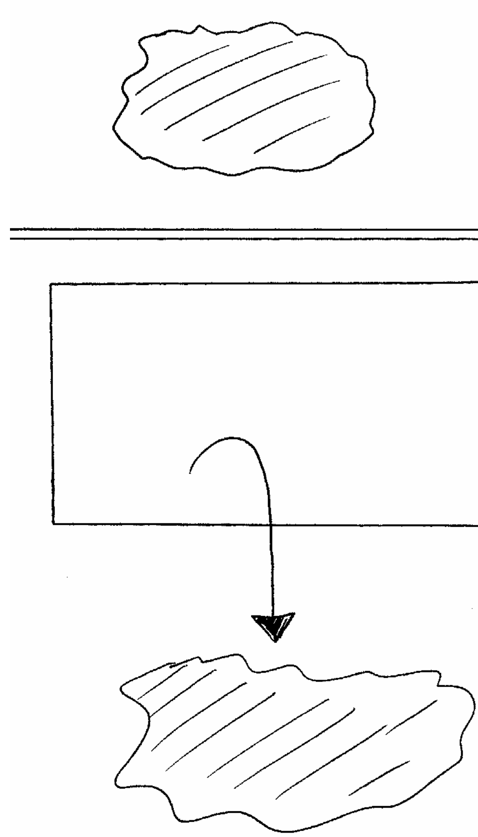
So far, Interest has been Restricted to a Description of the Sample

Our goal has been communication of the "forest" of the trees. Our "lense" did not extend beyond the sample.

Quite naturally, we will extend our "lense" to the larger goal of guesses of location and scatter that operate in a population.

What we calculate from a sample	Is a Guess of the Population
The sample mean = $\overline{X}$ **also** $\hat{\mu}$	**Population mean =** μ
The sample variance = S^2 **also** $\hat{\sigma}^2$	**Population variance =** σ^2

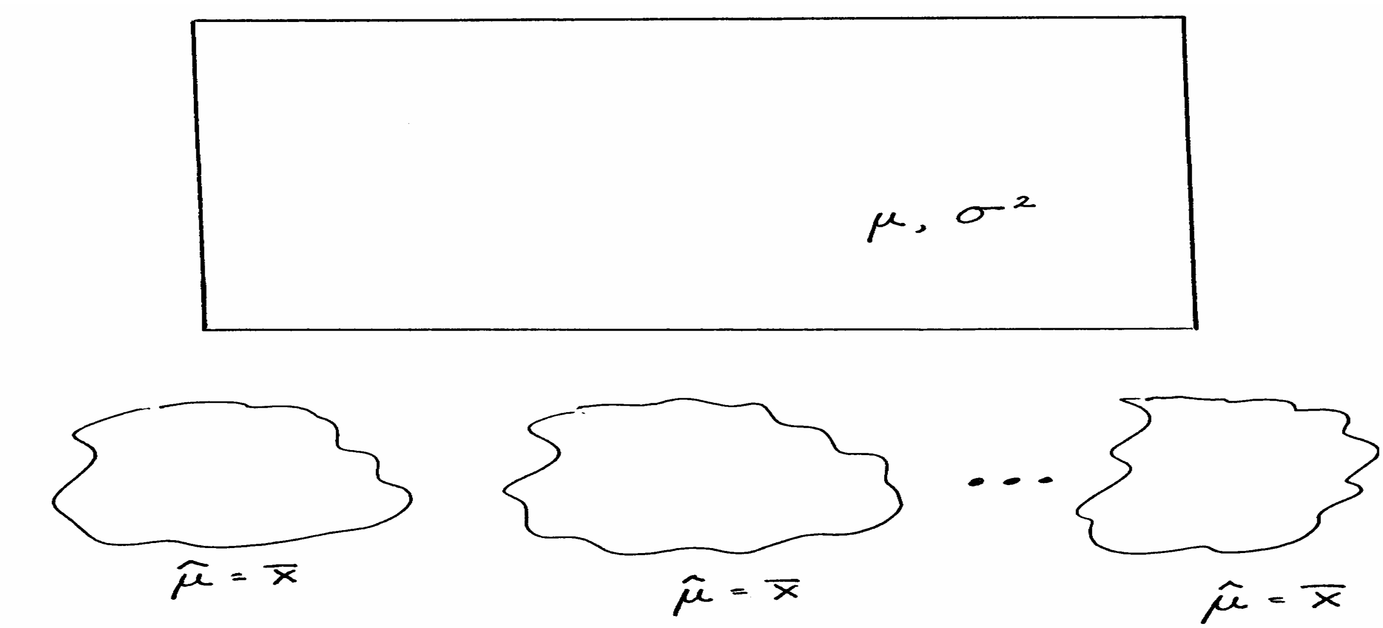
Note – The "caret" or "hat" is a convention used to represent a guess based on a sample.



<u>If it's of relevance, then others will be interested.</u>

What is the behavior of our "guess" from one conduct of study to the next? (Again, this is the idea of *sampling distributions*)

How good is the "guess" in our one study?



Standard Deviation (S or SD) v Standard Error (SE or SEM)

Now we can clarify the distinction with some rationale …

Standard Deviation	***Standard Error***
• **Describes variation in values of *individuals.***	• **Describes variation in values of a *statistic* from one conduct of study to the next.** • **Often, it is the variation in the *sample mean* that interests us.**
• **In the population of *individuals*:** σ	• **In the population of all possible *sample means* ("sampling distribution of mean"):** $\sigma / \sqrt{n}$
• **Our "guess" is S**	• **Our "guess" is** $S / \sqrt{n}$

The Coefficient of Variation

The coefficient of variation is the ratio of the standard deviation to the mean of a distribution.

- **It is a measure of the spread of the distribution relative to the mean of the distribution**

- **In the population, coefficient of variation is denoted ξ and is defined**

$$\xi = \frac{\sigma}{\mu}$$

- **The coefficient of variation ξ can be estimated from a sample. Using the hat notation to indicate "guess". It is also denoted CV**

$$\text{cv} = \hat{\xi} = \frac{S}{\overline{X}}$$

- **Example – "Cholesterol is more variable than systolic blood pressure"**

	S	$\overline{X}$	$\text{cv} = \hat{\xi} = s/\overline{x}$
Systolic Blood Pressure	**15 mm**	**130 mm**	**.115**
Cholesterol	**40 mg/dl**	**200 mg/dl**	**.200**

- **Example – "Diastolic is relatively more variable than systolic blood pressure"**

	S	$\overline{X}$	$\text{cv} = \hat{\xi} = s/\overline{x}$
Systolic Blood Pressure	**15 mm**	**130 mm**	**.115**
Diastolic Blood Pressure	**8 mm**	**60 mm**	**.133**

The Range

<u>The range</u> is the difference between the largest and smallest values in a data set.

- **It is a quick measure of scatter but not a very good one.**

- **Calculation utilizes only two of the available observations.**

- **As n increases, the range can only increase. Thus, the range is sensitive to sample size.**

- **The range is an unstable measure of scatter compared to alternative summaries of scatter (e.g. S or MADM)**

- **HOWEVER – when the sample size is very small, it may be a better measure of scatter than the standard deviation S.**

<u>Example –</u>

- **Data values are 5, 9, 12, 16, 23, 34, 37, 42**

- **range = 42-5=37**