TOPIC: Descriptive Statistics
Single Variable

I. Numerical data – summary measurements

- A. Measures of Location
  - 1. Measures of central tendency
    - Mean; Median; Mode
  - 2. Quantiles - measures of noncentral tendency
    - Quartiles; Percentiles
- B. Measures of Dispersion
  - Range; Interquartile range; Variance; Standard Deviation; Coefficient of Variation
- C. Measures of Shape
  - Skewness
  - 5-number summary
  - Box-and-whisker
  - Stem-and-leaf
- D. Standardizing Data

II. Categorical data

- A. Frequencies (also useful for grouped numerical data)
  - Frequency Distribution
  - Percentage Distribution
  - Cumulative Distribution
  - Histogram
  - Polygon
  - Ogive
- B. Charts
  - Bar chart
  - Pie chart
  - Pareto diagram

Summary Measures

Measures of Location

Measures of Central Tendency

- Mean
- Median
- Mode

The sample mean is the sum of all the observations divided by the number of observations:

$$\overline{X} = \frac{\mathrm{X}_1 + \mathrm{X}_2 + \mathrm{X}_3 + \ldots + \mathrm{X}_\mathrm{n}}{\mathrm{n}}$$

or

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where $\sum \mathrm{X}_\mathrm{i}$ is the same as $\mathrm{X}_1 + \mathrm{X}_2 + \mathrm{X}_3 + \ldots + \mathrm{X}_\mathrm{n}$

Example:

1
2
2
4
5
10

$\overline{X} = 24 / 6 = 4.0$

Example:

1
1
1
1
51

$\overline{X}$ = 55 / 5 = 11.0

Note that the mean is affected by extreme values.

MEDIAN

The median is the middle of the data (after data is arranged in ascending or descending order); half the observations are less than the median and half are more than the median. To get the median, we must first rearrange the data into an ***ordered array***. Generally, we order the data from the lowest value to the highest value.

The median is the data value such that half of the observations are larger than it and half are smaller. It is also the 50th percentile (we will be learning about percentiles).

If n is odd, the median is the middle observation of the ordered array.
If n is even, it is midway between the *two* central observations.

EXAMPLE:

0
2
3
5 ←
20
99
100

Median = 5

n=7 Since n is odd, the median is the (n+1)/2 ordered observation, or the 4th observation.

EXAMPLE n = 6

10
20
30
40
50
60

Median = 35

EXAMPLE: Exam scores

0
0
0 ←
0
100

What is the mean? Suppose the prof lets students know the following grading policy: Anyone who got the mean or better gets an A for the course; anyone who got below the mean fails. Are we happy or unhappy?

Note that the mean and median are UNIQUE for a given set of data.

Advantage: the Median is not affected by extreme values. In the ***previous*** example, if you change the 60 to 6,000, the median will still be 35. The mean, on the other hand will change by a great deal.

Problem: Sometimes it is difficult to order data.

The median has 3 interesting characteristics:

1. The median is not affected by extreme values, only by the number of observations.
2. Any observation selected at random is just as likely to be greater than the median as less than the median.
3. Summation of the absolute value of the differences about the median is a minimum:
   $\sum_{i=1}^{n} \left| X_i - Median \right|$ = minimum

MODE

The <u>mode</u> is the value of the data that occurs with the greatest frequency.

EXAMPLE:

1	1	1	2	3	4	5

The mode is 1 since it occurs three times. The other values only appear once in the data set.

EXAMPLE:

5	5	5	6	8	10	10	10

Mode = 5, 10
The modes for this data set are 5 and 10. This is a ***bi-modal*** dataset.

Problems:
The mode may not exist.
The mode may not be unique.

QUANTILES

Measures of non-central location

Quartiles
Deciles
Percentiles

These are all commonly used quantiles.

QUARTILES split the ordered data into four quarters.

Imagine cutting a chocolate bar into four equal parts… How many cuts would you make? (yes, 3!)

Q_1 – First Quartile – 25% of the observations are smaller than Q_1 and
75% of the observations are larger than Q_1
Q_2 – Second Quartile – 50% of the observations are smaller than Q_2 and
50% of the observations are larger than Q_2. Same as
the Median. It is also the 50th percentile.
Q_3 – Third Quartile – 75% of the observations are smaller than $Q_3$and
25% of the observations are larger than Q_3

The quartiles, like the median, either take the value of one of the observations, or the value halfway between two observations.

If n/4 is an integer, the first quartile (Q1) has the value halfway between the (n/4)th observation and the next observation.

If n/4 is not an integer, the first quartile has the value of the observation whose position corresponds to the next highest integer.

EXAMPLE:

210	
220	
225	Q_1 = 225, the 3rd observation
225	
225	Q_2 = Median - 230
235	
240	
250	Q_3=250, 3rd observation from the bottom
270	
280	

Please note that this technique is an approximation, and much easier than using the formula (algorithm). The Q_1 and Q_3 values you get from MS Excel may be slightly different.

EXAMPLE: Computer Sales

Original data	Ordered data	
3	0	
10	2	
2	3	--Q1
5	4	
9	5	
8	6	--Q2
7	7	
12	8	
10	9	--Q3
0	10	
4	10	
6	12	

Median = 6.5 Mode = 10

Q1 = 3.5 ||||| 25% smaller; 75% larger

Q3 = 9.5 ||||| 75% smaller; 25% larger

$\sum X = 76$

$\overline{X}$ = 76/12 = 6.33

DECILES, PERCENTILES

Similarly,
there are 9 deciles dividing the distribution into 10 equal portions (tenths).
there are 99 percentiles dividing the distribution into 100 equal portions.

In all these cases, the convention is the same. The point, be it a quartile, decile, or percentile, takes the value of one of the observations or it has a value halfway between two adjacent observations. It is never necessary to split the difference between two observations more finely.

Percentiles are used in analyzing the results of standardized exams. For instance, a score of 40 on a standardized test might seem like a terrible grade, but if it is the 99th percentile, don't worry about telling your parents. ☺

Q1 = which percentile?
Q2 (the median)?
Q3?

EXERCISE:

n=16

1
1
2
2
2
2
3
3
4
4
5
5
6
7
8
10

SOLUTION:

n=16

1
1
2
2
2
2
3
3
4
4
5
5
6
7
8
10

$\overline{X}$ = 65/16 = 4.06
Median = 3.5
Mode = 2
Q1 = 2
Q3 = 5.5

EXERCISE: # absences
n=16

0
5
3
2
1
0
2
4
3
2
1
0
0
0
6
8

SOLUTION: # absences
n=16

0
5
3
2
1
0
2
4
3
2
1
0
0
0
6
8

Ordered list: 0 0 0 0 | 0 1 1 2 | 2 2 3 3 | 4 5 6 8

$\overline{X}$ = 37/16 = 2.31
Median = 2
Mode = 0
Q1 = 0
Q3 = 3.5

EXERCISES – COMPUTE MEAN, MEDIAN, MODE, QUARTILES

EXERCISE: Reading Level of n=16 students in 8th grade

5
6
6
6
5
8
7
7
7
8
10
9
9
9
9
9

ANS 1.
First, order the data, and split into fourths using the three quartiles, Q1, Q2, and Q3:

5 5 6 6 || 6 7 7 7 || 8 8 9 9 || 9 9 9 10
Q1 ***Q2*** ***Q3***

Median = Q2 = 7.5
Q1 = 6, Q3 = 9
Mode = 9
Sum = 120
$\overline{X}$ = 120 / 16 = 7.5 average reading level in the 8th grade class

Alternate method: This data set can also be set up to be analyzed as ***grouped*** data. We don't really have to order the entire data set in this case.

Original data:

5
6
6
6
5
8
7
7
7
8
10
9
9
9
9
9

X_i (reading level)	frequency, f_i
5	2
6	3
7	3
8	2
9	5
10	1

Note that $\sum f_i = n =$ 16

To get the mean, median, etc.:

X_i	f_i	$(X_i)(f_i)$
5	2	10
6	3	18
7	3	21
8	2	16
9	5	45
10	1	10
	16 (= n)	120 (= sum of ungrouped data)

When data is grouped, how do we get the mean? Median? Mode? Quartiles?

Grouped data can also be presented in a ***bar chart*** (later).

EXERCISE: #colds / year
n=16

0 1 2 2 3 3 3 4 4 4 4 5 5 6 8 10

EXERCISE: avg. wait in minutes for a train
n=15

0
4
5
5
6
8
9
10
10
11
12
15
16
17
45

MEASURES OF DISPERSION

Dispersion is the amount of spread, or variability, in a set of data.

Why do we need to look at measures of dispersion?

Example:
A company is buying computer chips. These chips must have an average life of 10 years. The company has a choice of two suppliers. Whose chips should they buy? They take a sample of 10 chips from each of the suppliers.

Supplier A (n = 10)		Supplier B (n = 10)	
11	years	170	years
11		1	
10		1	
10		160	
11		2	
11		150	
11		150	
11		170	
10		2	
12		140	

X_A = 10.8 years	X_B = 94.6 years
Median = 11 years	Median = 145 years
s_A = 0.63 years	s_B = 80.6 years
Range = 2 years	Range = 169 years

Supplier B's chips have a longer average life. However, note that with a 3-year warranty, supplier A will have no returns while supplier B will have 4/10 or 40% returns.

There are 5 major measures of dispersion:

Range
Interquartile Range
Standard Deviation
Variance
Coefficient of Variation

RANGE

Range = Largest Value – Smallest Value

Example:

1 2 3 4 8 Range = 8 – 1 = 7

Problem: The range is influenced by extreme values at either end.

INTERQUARTILE RANGE

IQR = $Q_3 – Q_1$

From previous example:

210	
220	
225	Q1 = 225, the 3rd observation
225	
225	Q2 = Median - 230
235	
240	
250	Q3=250, 3rd observation from bottom
270	
280	

Interquartile Range = 250 – 225 = 25

It is basically the range encompassed by the central 50% of the observations in the distribution.

Problem: The interquartile range does not take into account the variability of the *total* data (only the central 50%). We are "throwing out" half of the data.

STANDARD DEVIATION

The standard deviation measures the "average" deviation about the mean. It is not really the "average" deviation, even though we may think of it that way.

Note: If you take a simple mean, and then add up the deviations about the mean, i.e., $\sum(X-\overline{X})$, this sum is always going to be equal to 0. Therefore, a measure of "average deviation" will not work.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X-\overline{X})^2}{n-1}}$$ "definitional formula"

By squaring the deviation, we have $\sum(X-\overline{X})^2$ which is a minimum (called the "least squares property"). No other value subtracted from X and squared will result in a smaller sum of the deviation squared.

The standard deviation has lots of nice properties, which we will study shortly.

Example:

X_i	Y_i
1	0
2	0
3	0
4	5
5	10

Find $\overline{X}$, $\overline{Y}$, s_x, s_y

X	$\overline{X}$	(X-$\overline{X}$)	(X-$\overline{X}$)^2
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
		∑=0	10

$S_X = \sqrt{}10/4 = 1.58$

Y	$\overline{Y}$	(Y-$\overline{Y}$)	(Y-$\overline{Y}$)^2
0	3	-3	9
0	3	-3	9
0	3	-3	9
5	3	2	4
10	3	7	49
		∑=0	80

$S_Y = \sqrt{}80/4 = 4.47$

[CHECK WITH CALCULATOR]

Note that the population standard deviation is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X-\mu)^2}{N}}$$

but it is very rare that we ever take a census of the population and deal with N,

Normally, we work with a sample and calculate the sample measures, like the sample mean and the sample standard deviation, s:

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

The reason we divide by n-1 instead of n is to assure that s is an unbiased estimator of σ. We have taken a shortcut: in the second formula, we are using $\overline{X}$, a statistic, in lieu of μ, a parameter. To correct for this – which has a tendency to understate the true standard deviation – we divide by n-1 which will increase s somewhat and make it an unbiased estimator of σ. Later on in the course we will refer to this as "losing one degree of freedom."

VARIANCE

The variance is

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

or, s = $\sqrt{Variance}$

Computational formula. This is what your calculator uses:

$$s^2 = \frac{\sum_{i=1}^{n} X_i^{\,2} - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2} - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}}$$

EXERCISE: X = # minutes waiting for bus

<u>X</u>
0 5 10 4 8 6 9 0 2 6

EXERCISE: X = hours to complete task

<u>X</u>
6 4 10 4 7 5 9 11

EXERCISE: X = # cups of coffee students drink in a day

<u>X</u>
10 6 9 7 8 0 0 4 5

5. COEFFICIENT OF VARIATION

The problem with s^2 and s is that they are in the “original” units.

This makes it difficult to compare the variability of two data sets, if they are in different units OR if the magnitude of the numbers is very different.

Suppose you wish to compare two stocks and one is in dollars and the other is in yen; if you want to know which one is more volatile, you should use the coefficient of variation. It is also not appropriate to compare two stocks of vastly different prices even if both are in the same units. The standard deviation for a stock that sells for around \$300 is going to be very different than one where the price is around \$0.25. The coefficient of variation will be a better measure of dispersion when comparing the two stocks than the standard deviation (see example below).

CV = $\frac{s}{\overline{X}}$ x 100%

Back to previous example:

X_i	Y_i
1	0
2	0
3	0
4	5
5	10
$\overline{X}$ =3	$\overline{Y}$ =3
s=1.5	s=4.4
8	7

CV_X = $\frac{1.58}{3}$ x 100% = 52.7%

CV_Y = $\frac{4.47}{3}$ x 100% = 149%

These two datasets aren’t really *that* different. Here, s pretty much does the job.

Example:
Which stock price is more volatile?

Closing prices over the last 8 months:

	Stock A	Stock B
JAN	$1.00	$180
FEB	1.50	175
MAR	1.90	182
APR	.60	186
MAY	3.00	188
JUN	.40	190
JUL	5.00	200
AUG	.20	210
Mean	$1.70	$188.88
s^2	2.61	128.41
s	$1.62	$11.33

The standard deviation of B is higher than for A, but A is more volatile:

$$CV_A = \frac{\$1.62}{\$1.70} \text{ x } 100\% = 95.3\%$$

$$CV_B = \frac{\$11.33}{\$188.88} \text{ x } 100\% = 6.0\%$$

Exercise: Test Scores

<u>X</u>

0   0   40   50   50   60   70   90   100   100

Compute the mean, median, mode, quartiles (Q1, Q2, Q3), range, interquartile range, variance, standard deviation, coefficient of variation.

Exercise: Test Scores

<u>X</u>

0   0   40   50   50 || 60   70   90   100   100

Q1 ***Q2*** ***Q3***

Sum = 560

$\bar{X}$ = 560 / 10 = 56.0

median = Q2 = 55

Q1 = 40, Q3 = 90

[Note: Excel gives these as Q1 = 42.5, Q3 = 85]

mode = 0, 50, 100

Range = 100 – 0 = 100

IQ Range = 90 – 40 = 50

s^2 = 11,840 / 9 = 1315.5

s = √1315.5 = 36.27

CV = (36.27 / 56) x 100% = 64.8%

Exercise: # employee absences

<u>X</u>

0   0   1   1   1   2   2   2   2   3   4   6

Exercise: Quiz scores ☹

<u>X</u>

0
0
0
0
0

5
6
7
8
9

10
10
10

C. Measures of Shape

A third important property of data is its shape.

1. Central tendency
2. Dispersion
3. Shape

Shape can be described by degree of asymmetry (i.e., skewness).

mean > median	positive	or right-skewness
mean = median	symmetry	or zero-skewness
mean < median	negative	or left-skewness

Positive skewness arises when the median is increased by some unusually high values. Negative skewness occurs when the mean is decreased by some unusually low values.

Example: # hours to complete a task

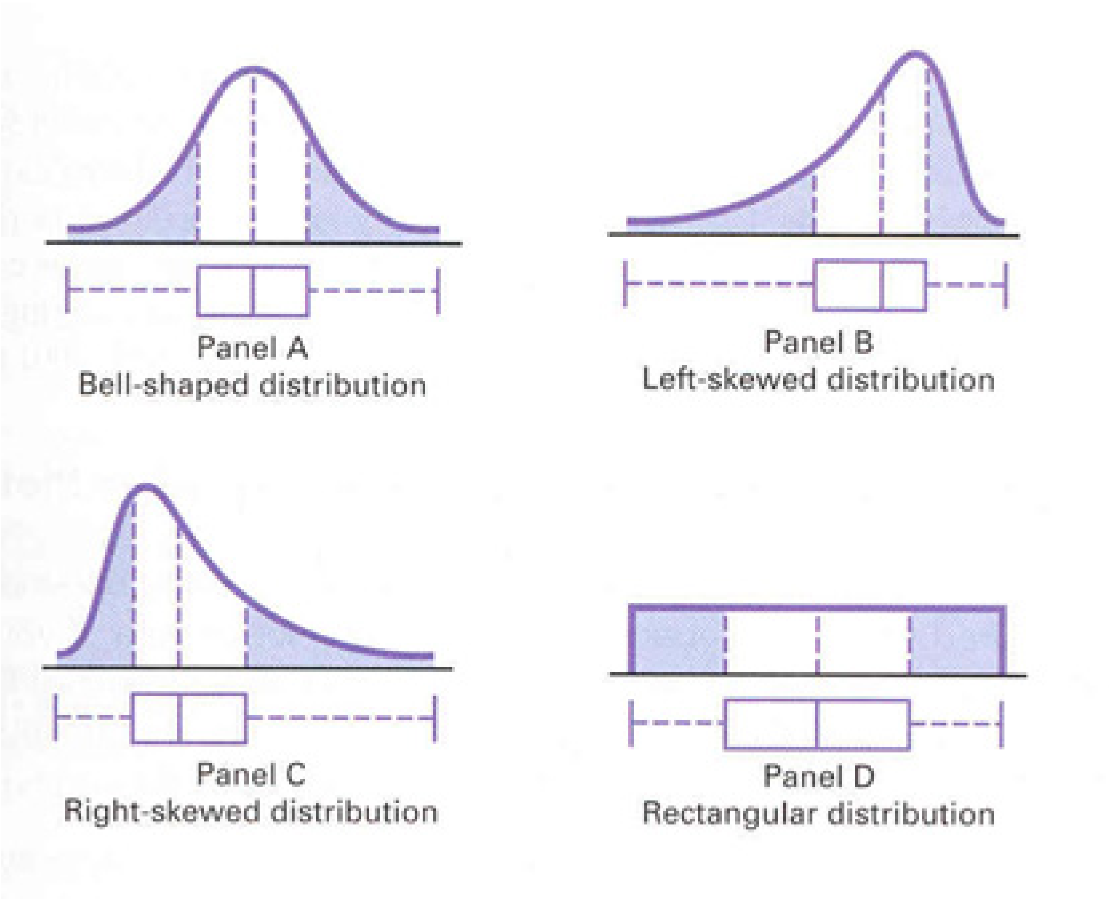
2 3 8 8 9 10 10 12 15 18 22 63

Q_1 (↑ between 8 and 8) Median (↑ between 10 and 10) Q_3 (↑ between 15 and 18)

This guy took a VERY long time!

$\overline{X}$ = 15 – skewed right! Median = 10.

s^2 = 2868 / 11 = 260.79
s = 16.25
CV = 107.7%



Panel A
Bell-shaped distribution

Panel B
Left-skewed distribution

Panel C
Right-skewed distribution

Panel D
Rectangular distribution

To measure skewness, MS Excel uses the Pearson coefficient of skewness:

$$\text{Skewness}_{\text{P}} = \frac{3(\overline{X} - Median)}{s}$$

This formula measures departure from symmetry. If the data is symmetrically distributed, this means that mean=median=mode.

We do not have to remember these formulas. We can get the skewness measures as output from MS Excel.

NOTE: An alternate formula used to measure skewness, if the mode exists, is:

$\text{Skewness}_{\text{P}} = \dfrac{\overline{X} - Mode}{s}$ or, if Mode is not known:

We do not have to know these formulas. We get the skewness measures as output from MS Excel.

EXAMPLE: # defects in a sample of 12 cars
(n=12)

2	3	8	8	9	10	10	12	15	18	22	63

Descriptive Statistics in MS Excel

This is a good point at which to make some time to get more proficient in MS Excel. See the class handout "Obtaining Descriptive Statistics in MS Excel" and follow the instructions to do the problems in the lecture notes and in the Homework Assignment.

STANDARDIZING DATA

Z-scores: We can convert the original scores to new scores with
$\bar{X} = 0$ and $s = 1$.
[Note: you have a pure number with no units of measurement.]

Any score below the mean will now be negative.
Any score at the mean will be 0.
Any score above the mean will now be positive.

Example:

X		Z
0	$\frac{0-5}{3.74}$	-1.34
2	$\frac{2-5}{3.74}$	-.80
4	$\frac{4-5}{3.74}$	-.27
6	$\frac{6-5}{3.74}$	.27
8	$\frac{8-5}{3.74}$	.80
10	$\frac{10-5}{3.74}$	1.34

Example:

X	Z			X	Z			X	Z
65	-0.45			65	-0.81			65	-1.40
73	-0.11			73	-0.38			73	-0.79
78	0.10			78	-0.10			78	-0.40
69	-0.28			69	-0.60			69	-1.09
78	0.10			78	-0.10			78	-0.40
7	-2.89	<=		**97**	0.94			97	1.07
23	-2.21			**23**	-3.12	<=		**93**	0.76
98	0.94			98	0.99			98	1.14
99	0.99			99	1.05			99	1.22
99	0.99			99	1.05			99	1.22
97	0.90			97	0.94			97	1.07
99	0.99			99	1.05			99	1.22
75	-0.02			75	-0.27			75	-0.63
79	0.14			79	-0.05			79	-0.32
85	0.40			85	0.28			85	0.14
63	-0.53			63	-0.92			63	-1.56
67	-0.36			67	-0.70			67	-1.25
72	-0.15			72	-0.43			72	-0.86
73	-0.11			73	-0.38			73	-0.79
93	0.73			93	0.72			93	0.76
95	0.82			95	0.83		.	95	0.91
Mean	75.57			Mean	79.86			Mean	83.19
Std. Dev.	23.75			Std. Dev.	18.24			Std. Dev.	12.96

For standardized data, if it is normally distributed, 95% of the data will be between 2 standard deviations about the mean.

If the data follows a normal distribution, 95% of the data will be between -1.96 and +1.96. 99.7% of the data will fall between -3 and +3. 99.99% of the data will fall between -4 and +4.

No matter what you are measuring, a Z-score of more than +5 or less than – 5 would indicate a very, very unusual score.

Worst case scenario: 75% of the data are between 2 standard deviations about the mean. [Chebychev.]

5-NUMBER SUMMARY

[Developed by Tukey]

Median

Q_1 Q_3

Lowest Highest

Example:

2 3 8 8 9 10 10 12 15 18 22 63

↑ ↑ ↑

Q_1 Median Q_3

$\overline{X} = 15$

$s^2 = 2868 / 11 = 260.79$

$s = 16.25$

CV = 107.7%

10

8 16.5

2 63

This data is right-skewed.

In right-skewed distributions, the distance from Q_3 to X_{largest} is significantly greater than the distance from X_{smallest} to Q_1.

Also, midrange > midhinge > median

With left-skewed data,

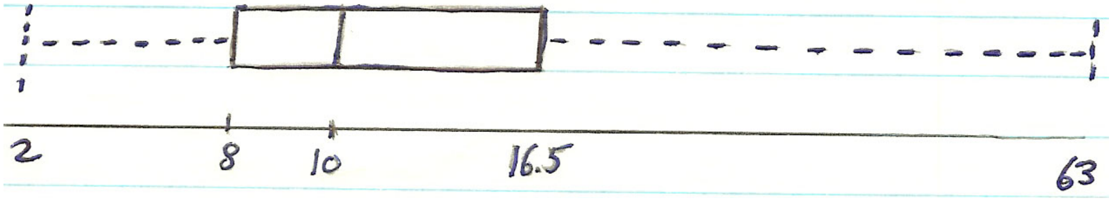
(X_{smallest} to Q_1) > (Q_3 to X_{largest})

Median > midhinge > midrange

BOXPLOT

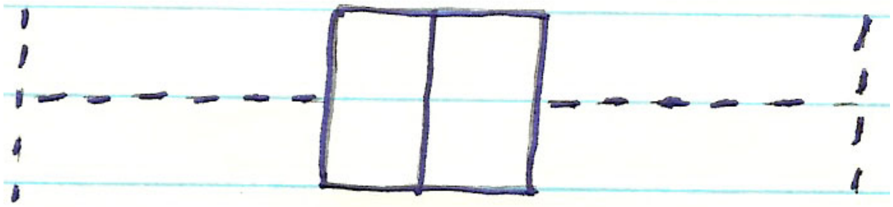
Vertical line drawn within the box = median
Vertical line at the left side of box = Q_1
Vertical line at the right side of box = Q_3
Dashed line on left connects left side of box with $X_{smallest}$ (lower 25% of data)
Dashed line on right connects right side of box with $X_{largest}$ (upper 25% of data)

EXAMPLE:
For the data set above, the Box-and-Whisker Plot would look like this:



(Right-skew)

A "bell-shaped" data distribution would look like this:



(Symmetric)

FREQUENCY DISTRIBUTION

Frequency Distribution: Records data grouped into classes and the number of observations that fell into each class.

A frequency distribution can be used for: categorical data; for numerical data that can be grouped into categories (or, classes); and for numerical data with repeated observations.

A Percentage Distribution records the percent of the observations that fell into each class.

Example: A (fictitious) sample was taken of 200 professors at a local college. Each was asked for his or her weekly salary. The responses ranged from about \$520 to \$590. If we wanted to display the data in, say, 7 equal intervals, we would use an interval width of \$10.

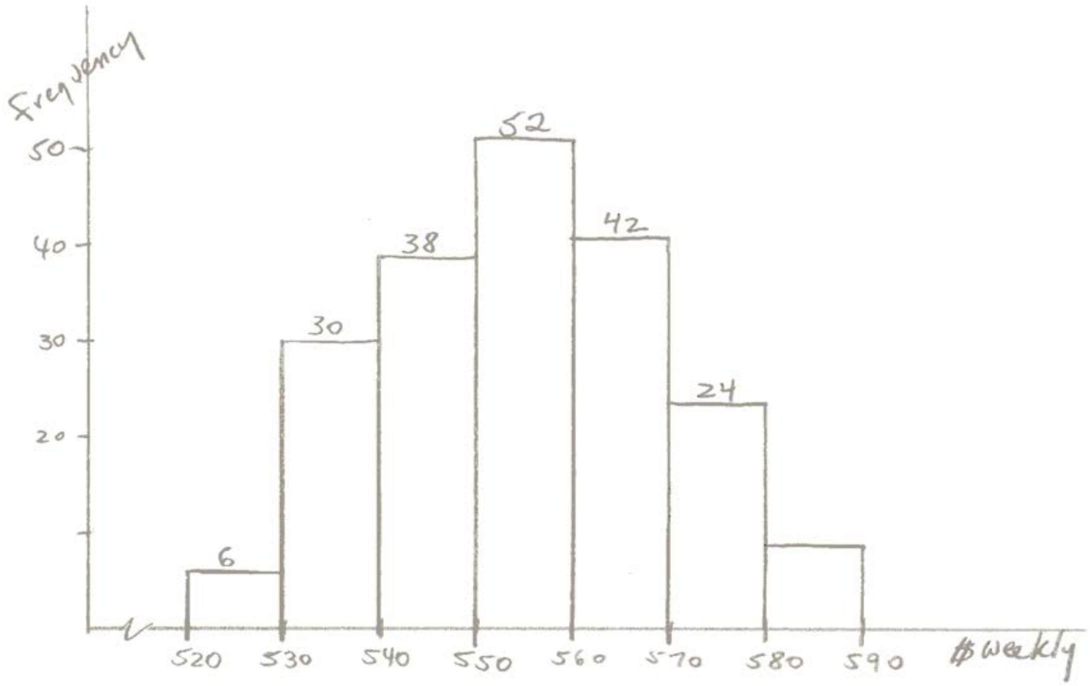
width of interval = range/number of classes
= \$70/7 = \$10/class.

FREQUENCY DISTRIBUTION

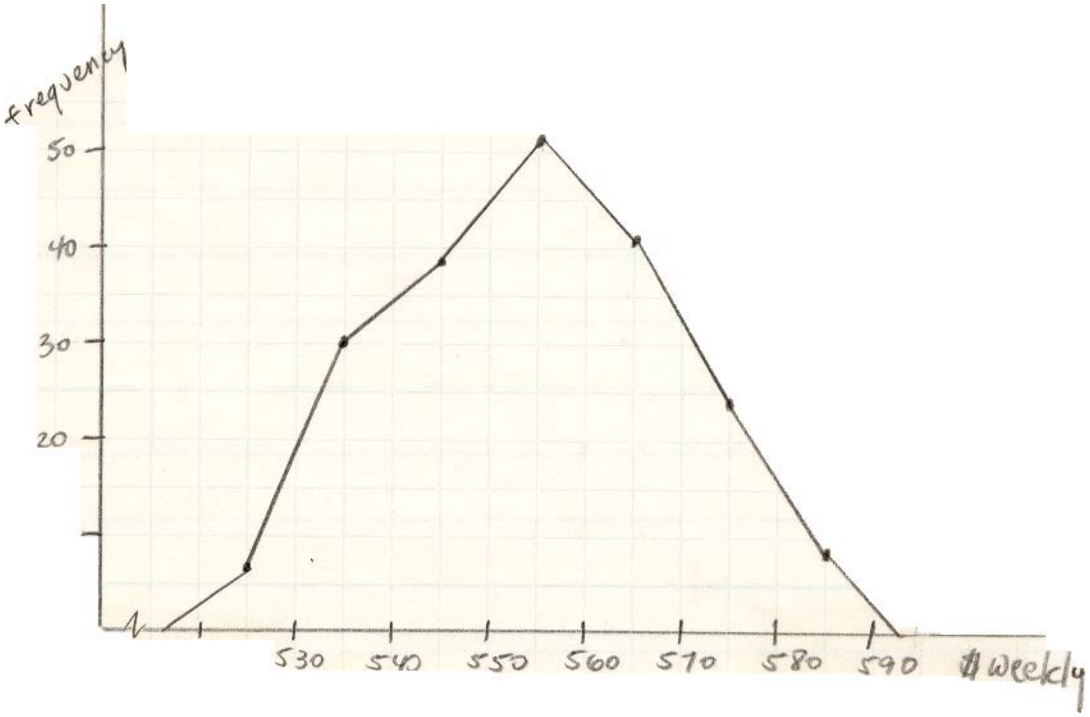
Weekly earnings	Frequency	Percentage
520 and under 530	6	3 %
530 " " 540	30	15
540 " " 550	38	19
550 " " 560	52	26
560 " " 570	42	21
570 " " 580	24	12
580 to 590	8	4
	n = 200	100 %

A <u>Cumulative Distribution</u> focuses on the number or percentage of cases that lie below or above specified values rather than within intervals.

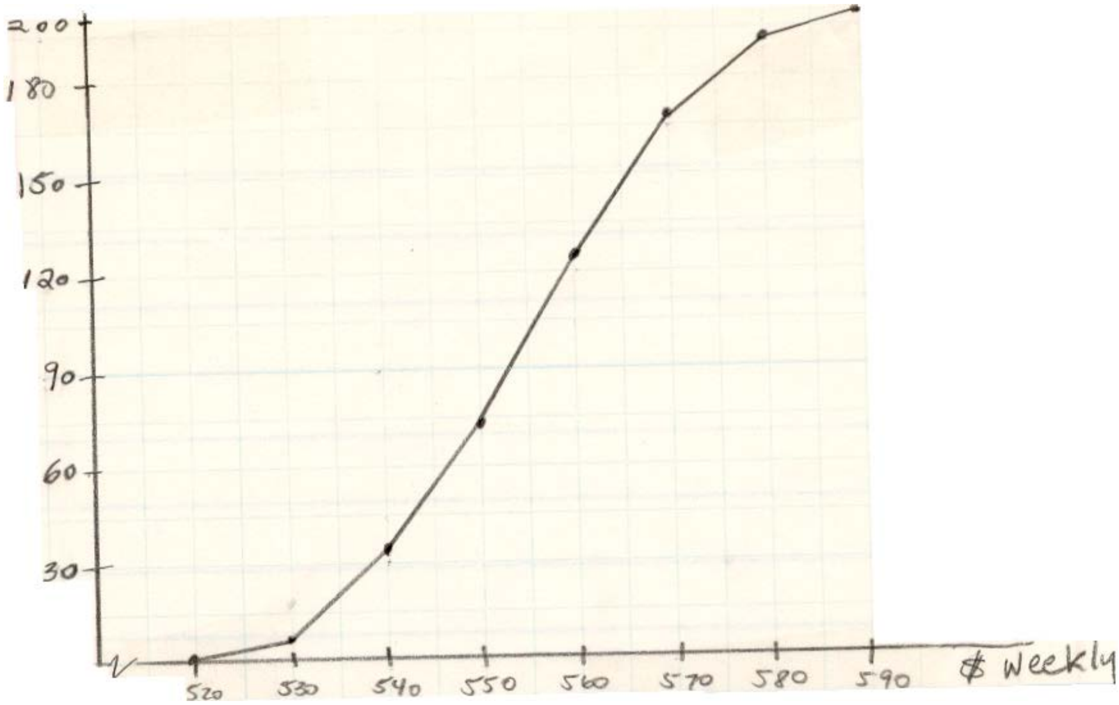
			Frequency	Percentage
less	than	520	0	0 %
"	"	530	6	3
"	"	540	36	18
"	"	550	74	37
"	"	560	126	63
"	"	570	168	84
"	"	580	192	96
"	"	590	200	100



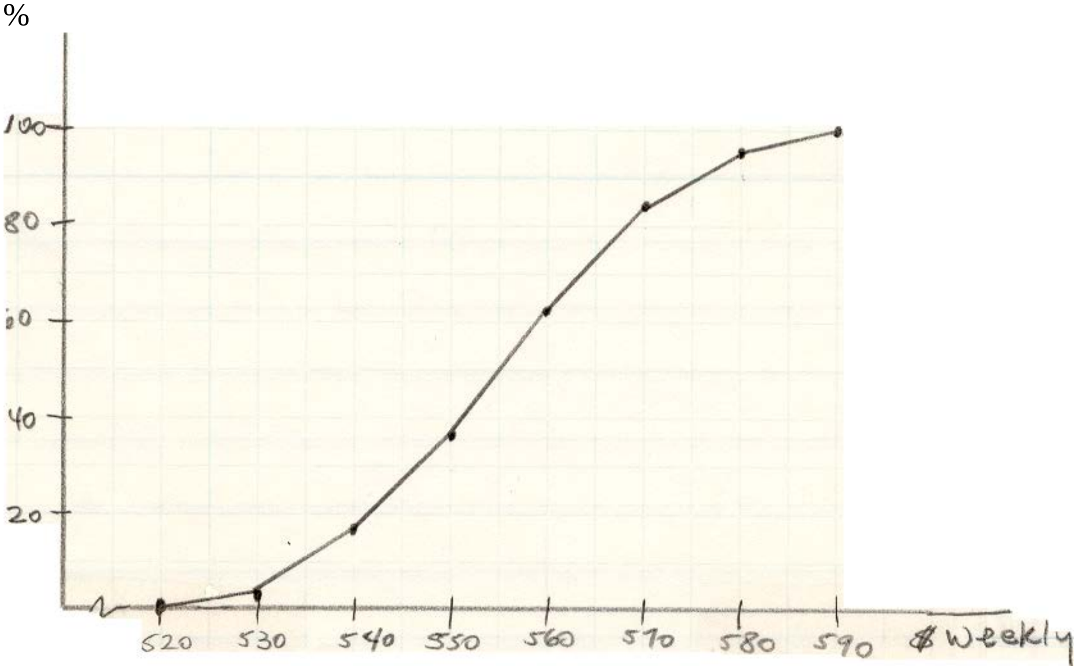
FREQUENCY HISTOGRAM



FREQUENCY POLYGON



CUMULATIVE FREQUENCY DISTRIBUTION



CUMULATIVE PERCENTAGE DISTRIBUTION

Other types of charts for data presentation, categorical variable:

Bar chart

Pie chart

Pareto Diagram

TOPIC: Descriptive Statistics
Two Variables

I. Categorical data

A. Contingency Tables

B. Side-by-Side Bar Chart

II. Numerical data – i.e., looking for relationships in bivariate data

A. Scatter Plot (Ch. 2)

B. Correlation (Ch. 3)

C. The Regression Line (Ch. 12)

A. The Contingency Table

Two categorical variables are most easily displayed in a contingency table. This is a table of two-way frequencies.

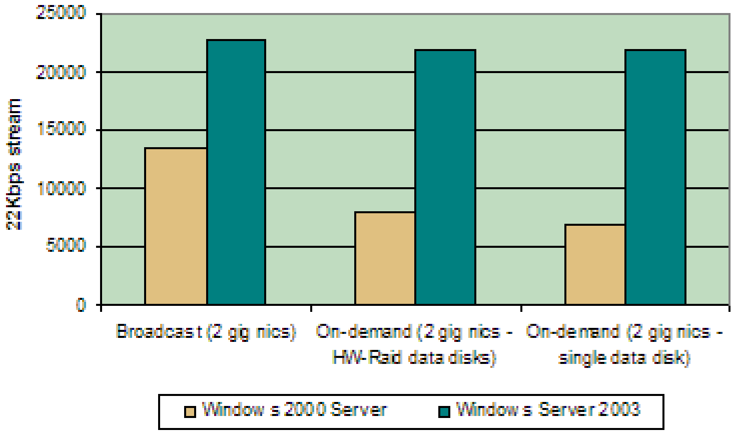
Example: “Who would you vote for in the next election?”

	Male	Female	
Republican Candidate	250	250	500
Democrat Candidate	150	350	500
	400	600	1000

Also for two-way percentages:

	Male	Female	
Republican Candidate	25%	25%	50%
Democrat Candidate	15%	35%	50%
	40%	60%	100%

B. The Side – by – Side Bar Chart



Relative Performance (Source: Microsoft.com)

What do we do with 2 numerical variables? We can graph them:

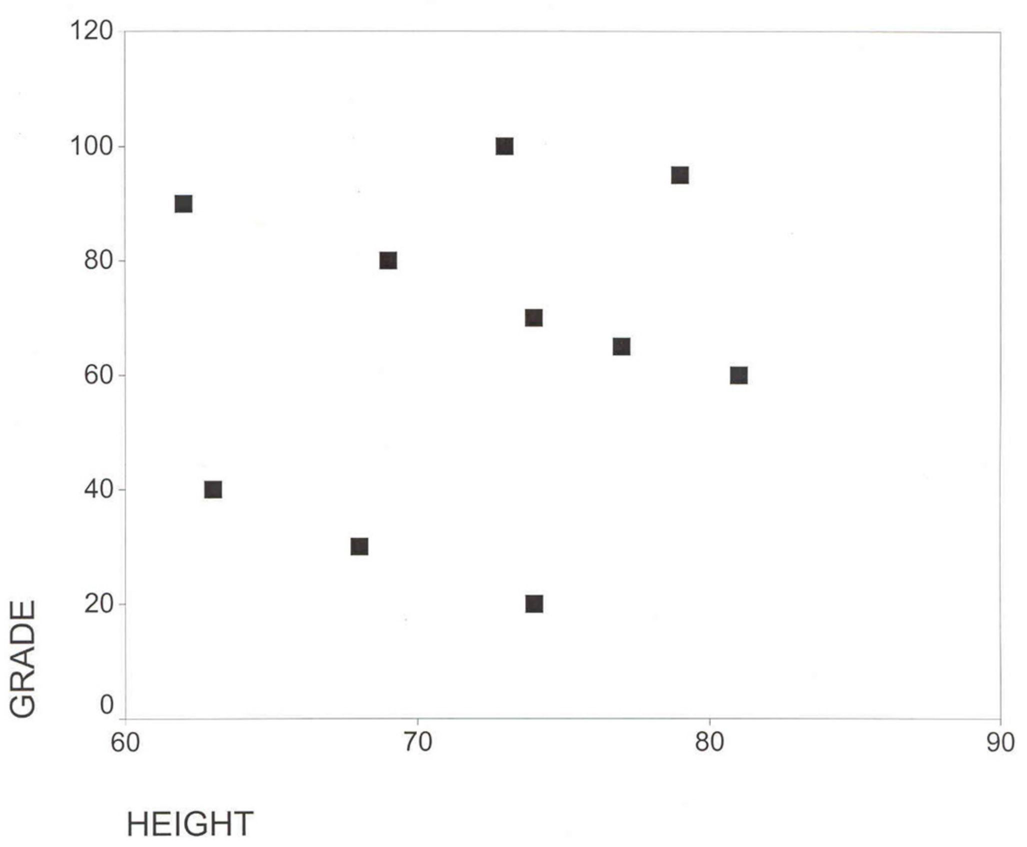
Scatter Plot

Example:

Y (Grade)	100	95	90	80	70	65	60	40	30	20
X (Height)	73	79	62	69	74	77	81	63	68	74

in inches

$r = .12$
$r^2 = .01$



For more about Scatter Plots, see the ***Correlation*** lecture notes.