Statics

Descriptive statistics (Analysis)

1. Organizing and Summarizing Data. . (الصورة بألف كلمة) data Visualization or tables.
2. Numerically summarizing Data
3. Describe the relation between two variables.

Organizing and Summarizing Data

Describe data sets:

Describe data sets

Graphical (visualizations)

Tabular (Tables)

Must check Frequency To Analyze

Qualitative
Ideal Graphs for (categorical) Data
EX: m/f – excellent/good

1-Bar graph
2- pie chart(5 sectors)
3-Line Graph
4- Frequency polygon

(Frequency Table) detects frequent relation between values

(Relative Frequency table)
1var /Total data = Frequency
Grouping Data

Quantitative Data

Dot plots chart
Low number of data set

Histograms chart
Avoiding (Bill Chart) if it shows you a Normal Distribution of (Interval Measurement)

1. Organizing and Summarizing Data

Describing data sets

The numerical findings of a study should be presented clearly, concisely, and in such a manner that an observer can quickly obtain a feel for the essential characteristics of the data. Over the years it has been found that tables and graphs are particularly useful ways of presenting data, often revealing important features such as the range, the degree of concentration, and the symmetry of the data. In this section we present some common *graphical* and tabular ways for presenting data.

Frequency tables and graphs

Frequency tables

A data set having a relatively small number of distinct values can be conveniently presented in a *frequency table*.

Starting Yearly Salaries.

Starting Salary	Frequency
57	4
58	1
59	3
60	5
61	8
62	10
63	0
64	5
66	2
67	3
70	1

relative frequency

The **relative frequency** is the proportion (or percent) of observations within a category and is found using the formula

$$\textit{Relative frequency} = \frac{\textit{frequency}}{\textit{sum of all frequencies}}$$

A **relative frequency distribution** lists each category of data together with the relative frequency.

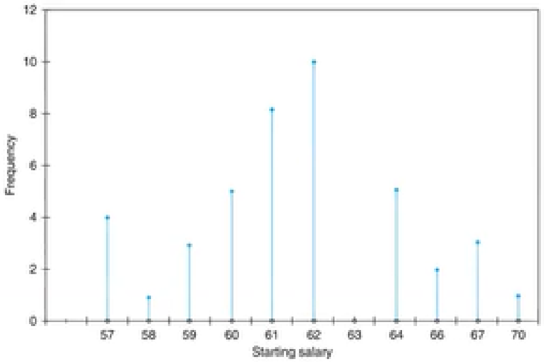
Body Part	Frequency	Relative Frequency
Back	12	$\frac{12}{30} = 0.4$
Wrist	2	$\frac{2}{30} = 0.0667$
Elbow	1	0.0333
Hip	2	0.0667
Shoulder	4	0.1333
Knee	5	0.1667
Hand	2	0.0667
Groin	1	0.0333
Neck	1	0.0333
Total	**30**	**1**

Describing data sets

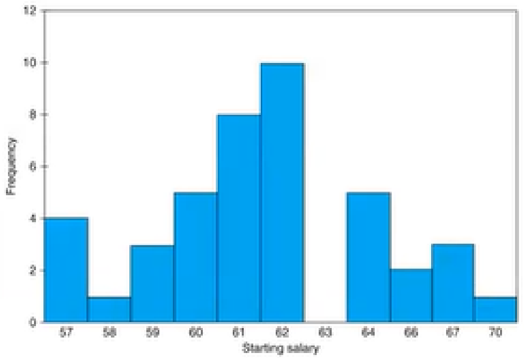
Frequency tables and graphs

graphs

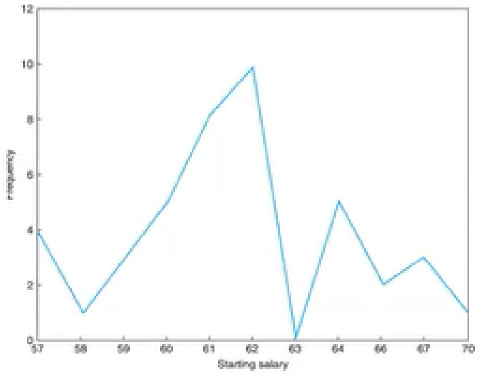
- *line graph*
- *bar graph.*
- *frequency polygon*



line graph



bar graph



polygon

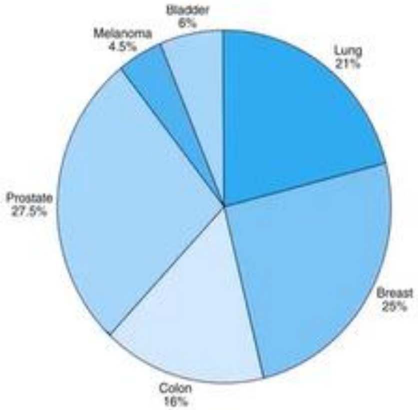
Describing data sets

Relative frequency tables and graphs

Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio f/n is called its *relative frequency*.

The following data relate to the different types of cancers affecting the 200 most recent patients to enroll at a clinic specializing in cancer.

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06



pie chart

Describing data sets

Dot plot

1	7	4	1
2	4	3	48
3	5	3	6

outlier

Autosaved].pptx - PowerPoint

ep your files safe with genuine Office today. Get genuine Office Learn more

Ideal Graphs for Quantitative data sets
(Histogram and Dot plot)

Histogram deals with (interval measure) a large amount of data up to 30.000 row

Describing data sets

Grouped data, histograms

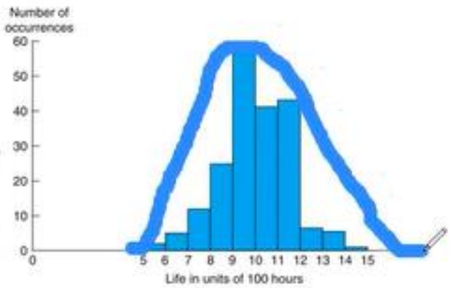
A **histogram** is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

Life in Hours of 200 Incandescent Lamps.

Item Lifetimes									
1067	919	1196	785	1126	936	918	1156	920	948
855	1092	1162	1170	929	950	905	972	1035	1045
1157	1195	1195	1340	1122	938	970	1237	956	1102
1022	978	832	1009	1157	1151	1009	765	958	902
923	1333	811	1217	1085	896	958	1311	1037	702
521	933	928	1153	946	858	1071	1069	830	1063
930	807	954	1063	1002	909	1077	1021	1062	1157
999	932	1035	944	1049	940	1122	1115	833	1320
901	1324	818	1250	1203	1078	890	1303	1011	1102
996	780	900	1106	704	621	854	1178	1138	951
1187	1067	1118	1037	958	760	1101	949	992	966
824	653	980	935	878	934	910	1058	730	980
844	814	1103	1000	788	1143	935	1069	1170	1067
1037	1151	863	990	1035	1112	931	970	932	904
1026	1147	883	867	990	1258	1192	922	1150	1091
1039	1083	1040	1289	699	1083	880	1029	658	912
1023	984	856	924	801	1122	1292	1116	880	1173
1134	932	938	1078	1180	1106	1184	954	824	529
998	996	1133	765	775	1105	1081	1171	705	1425
610	916	1001	896	709	860	1110	1149	972	1002

Table 2.4 A Class Frequency Table.

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1



Normal Bell distribution For Data

Slide 22 of 188 English (United States) Accessibility: Investigate

Describing data sets

Quantitative data set
Little data

Dot plot

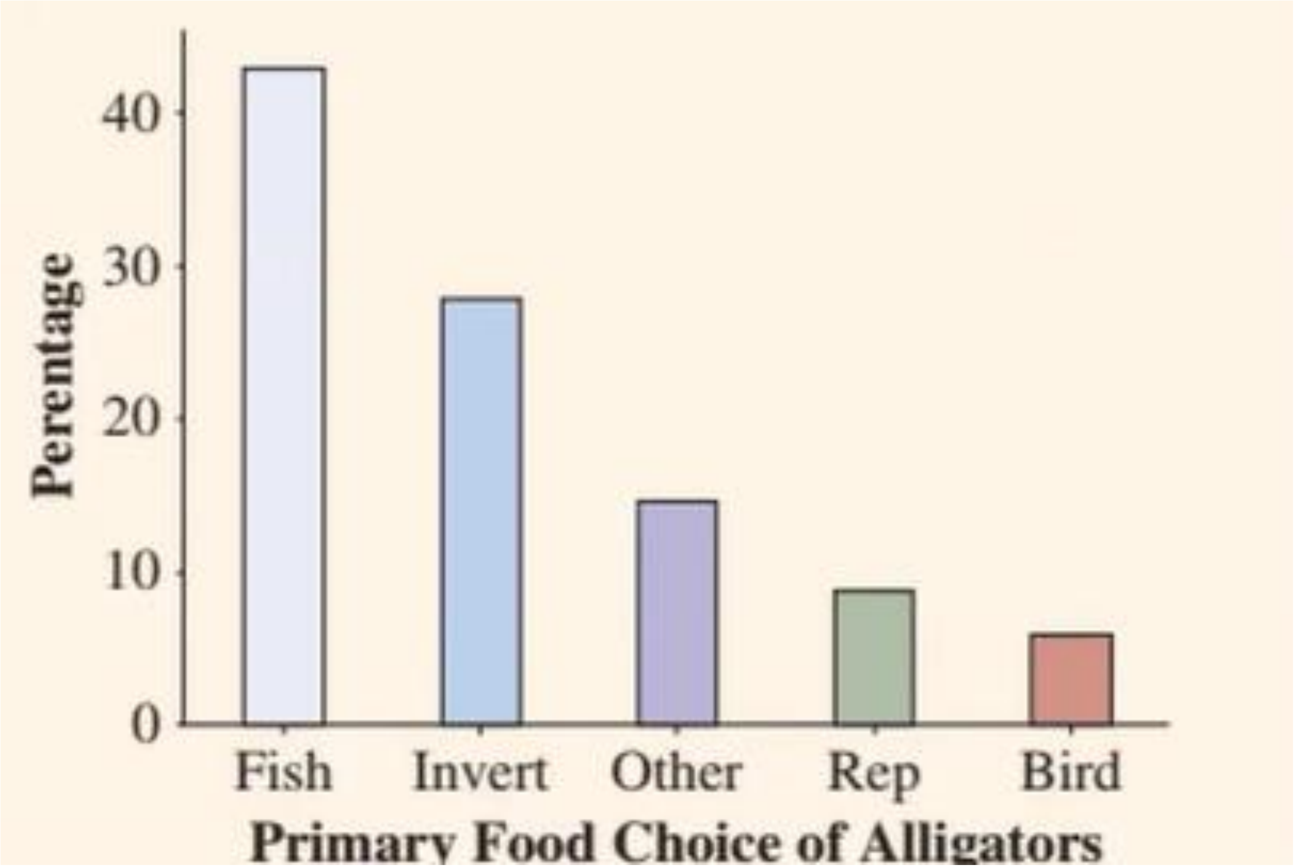
1	7	4	1
2	4	3	48
3	5	3	6

outlier

From Python Library (map plot lib)
Put the data into library and it will be done

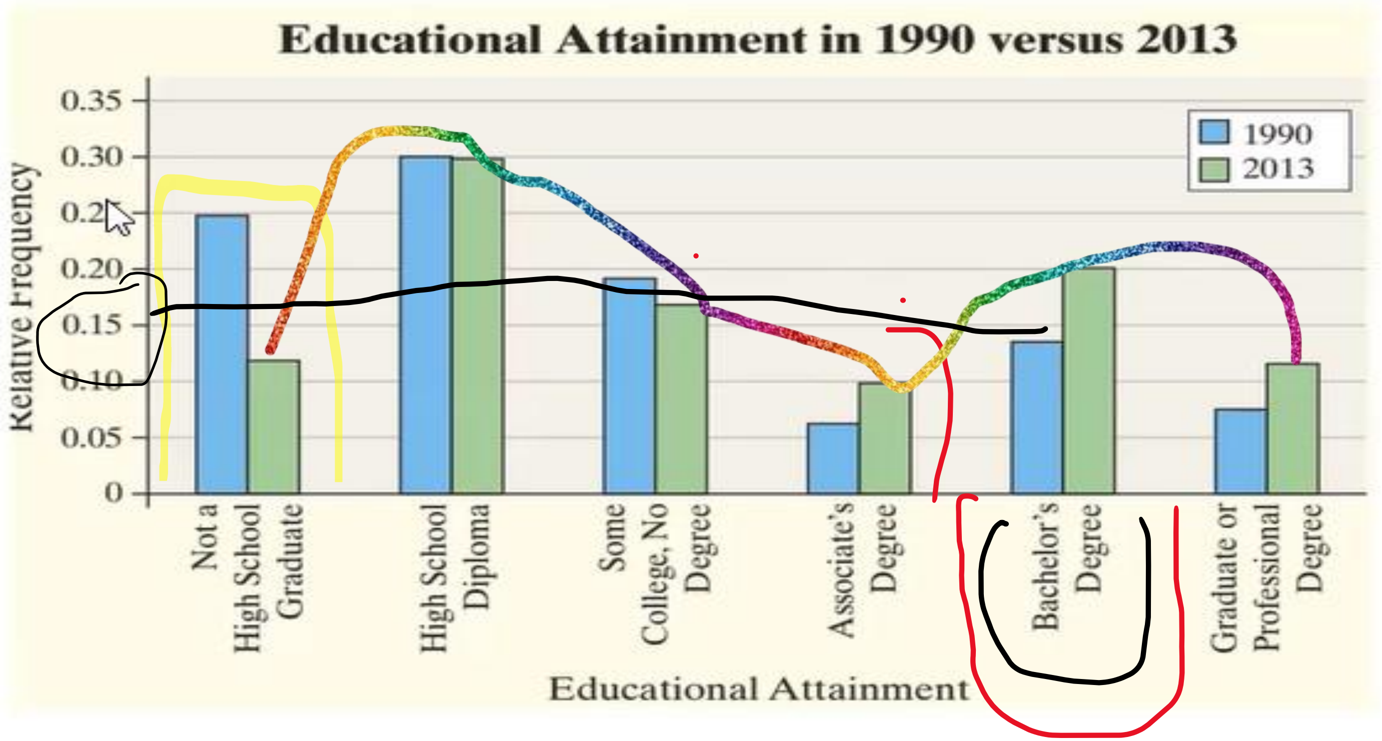
- **Example 1**

- What do alligator eat?
- 1- is primary food choice categorical or quantitive? (categorical) (Bar chart).
- 2- Which is the model category for primary food? (fish)
- 3- About what percentage alligators had fish as the primary food choice? 42%



Example 2
Comparing two data sets

Educational Attainment	1990	2013
Not a high school graduate	0.2476	0.1185
High school diploma	0.2999	0.2982
Some college, no degree	0.1874	0.1682
Associate's degree	0.0616	0.0984
Bachelor's degree	0.1311	0.2009
Graduate or professional degree	0.0722	0.1157



a) Draw side by side relative frequency bar graph of data. (quantitative data histogram graph)

b) Make some general conclusion based on the graph.

Analysis:

1-Relative frequency of adult who are not high school graduates in 2013 is less than half that of 1990.(yellow marker)

2 – A much higher percentage of the adult population has at least a bachelor's degree. (Black Marker)

3- The percentage of the population with a bachelor's degree has not doubled(ad the frequencies in dataset Table might suggest. (red marker in the table)

- An overall conclusion is that adults American are more educated in 2013 than they were in 1990. (Red marker).

Example 3: Walt Disney Stock

The table shows the movement of Walt Disney stock for 30 randomly selected trading days. “Up” means the stock price increased in value for the day, “Down” means the stock price decreased in value for the day, and “No Change” means the stock price closed at the same price it closed for the previous day.

Down	Up	Up	Down	Down	Up
Down	Up	Down	Up	Down	Up
Down	Down	Up	Up	Up	Up
Down	Down	Down	Up	Down	Up
No Change	Up	Down	Down	No Change	Down

Source: Yahoo!Finance

Example 3: Walt Disney Stock

Categorical Data
Best Graph: pie chart

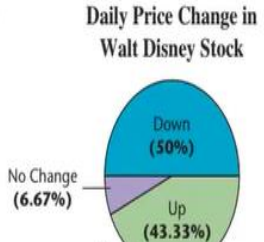
(a), (b)

Price Change	Frequency	Relative Frequency
Down	15	0.5
No Change	2	0.067
Up	13	0.433

(c)

(d)

(e)



(a) Construct a frequency distribution.
(b) Construct a relative frequency distribution
(c) Construct a frequency bar graph.
(d) Construct a relative frequency bar graph.
(e) Construct a pie chart.

a) Frequency table

b)Up/30 all data it means 1/3 form data

c) visualize a bar graph

d) visualize a frequency bar graph

e) Visualize a pie chart(best one)

2. Numerically Summarizing Data

Numeric Summarization
Built in function in python .(describe)
Gives you a feed back about your data
To enables you to make analytics or statistics

2.1 Measures of Central Tendency
2.2 Measures of Dispersion
2.3 Measures of Central Tendency and Dispersion from Grouped Data
2.4 Measures of Position and Outliers
2.5 The Five-Number Summary and Boxplots

9867 rows × 11 columns

[4]:
```
1  df.describe()
```

[4]:

	wind_speed	tmin	tmax	t	dayLenght	rh	GDD	PTU	HYTU	Prod
count	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000
mean	4.505868	15.492754	28.065555	21.330225	11.997602	55.031161	16.830225	207.498762	916.131627	2.820439
std	1.203268	5.113966	6.379263	5.446211	1.400406	12.685606	5.446211	83.476255	359.506825	0.098942
min	1.382345	2.152200	10.856100	7.639600	9.985629	3.345892	3.139600	31.415192	59.481575	2.617500
25%	3.718388	10.985500	22.597450	16.319900	10.616327	50.642187	11.819900	127.289356	629.516456	2.778900
50%	4.406298	15.532700	28.797400	21.642400	12.000000	57.460712	17.142400	208.285750	880.389701	2.781000
75%	5.128551	20.198750	33.707900	26.558200	13.371042	62.870843	22.058200	289.578827	1223.286253	2.895400
max	12.898120	25.362900	44.825700	34.063200	14.014371	93.192306	29.563200	404.932133	1729.098153	3.052500

[5]:
```
1  df.info()
```

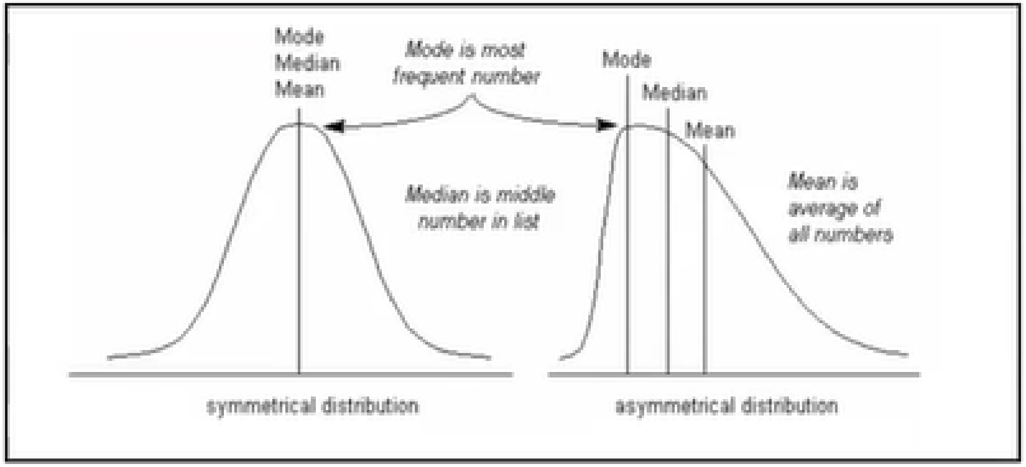
2.1 Measures of Central Tendency

1. Determine the arithmetic mean of a variable from raw data
2. Determine the median of a variable from raw data
3. Explain what it means for a statistic to be resistant
4. Determine the mode of a variable from raw data

2.1Measures of Central Tendency

Summarizing data sets

Mean: This is a simple arithmetic average, which is computed by taking the aggregated sum of values divided by a count of those values. The mean is sensitive to outliers in the data. An outlier is the value of a set or column that is highly deviant from the many other values in the same data; it usually has very high or low values.
Median: This is the midpoint of the data, and is calculated by either arranging it in ascending or descending order. If there are N observations.
Mode: This is the most repetitive data point in the data:



2.1 Measures of Central Tendency

Summarizing data set:

The OUTLIER: القيم الشاذة فى البيانات

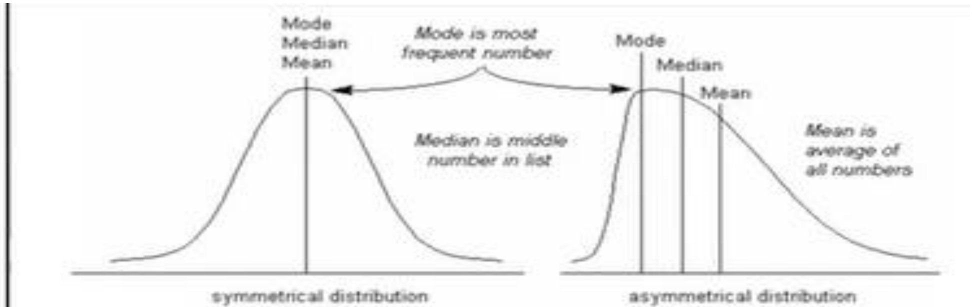
1-We must check the data if the (Mean is more than Median) it means there should be an (OUT LAYER) NOT Symmetrical Distribution. (we must drop the variable(delete from the data).
Age:25,43,30 Age:65 (OUTLIER) must be dropped from data.(eliminated)

2- If(Mean not close to Median) it make a conflict with decision making .
a) Median : ideal use with the **OUTLIER**. **It** organize the data.

Kinds of Out layer:

1- Anomaly Detection: قيم غير متوقعة (Ex: bank withdraw: 1000 $. Suddenly withdraw from visa 100000$) so we need this data we can't drop it. (visa bank). Or geographical (fraud detection)
2- wrong data entry we edit it in the data if, we are sure.
3- Prediction:

Biased Data not symmetrical data can't work with **ML**
And prediction **modules**.



Dealing with missing data:

Generally, We can fill it by (Mean)

Case one: few data sets we can drop the missing data. 20000: missing 10 drop, but If it is a Big data and a lot of missing so we must figure out the missing, we can fill it by (Mean) as if it were 200 missing ,so it should at least 100 correct of them so it will not affect the result.

Case two: if we have a lot of outlier values so we must use (median).

Case three: analyze a quick changing values in the market like (Gold or forex) so we can't use (Mean and Median) to fill the missing value in the data it will mislead us.

So, we can take (Near Mean and Near Median) from the raw (original data source) data from the above 3 rows of blank ,and the 3 bottom rows under the missing or blank cells.

Machine Learning doesn't work with missing values.###

D – Case four: We can figure out the normal data from numeric . (describe) If mean = median = mod so it's symmetric data set.

Case Five: Relation between mean and median types of Bell Graph:

1- Left skewed (mean > median)

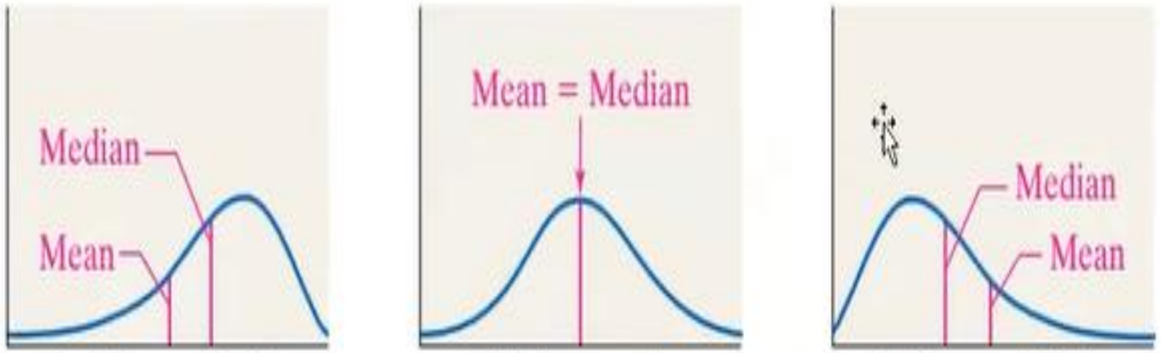
2- Right skewed(mean > mod)

Handle data:

a) Missing data :complete missing values.

b) Bias data: make a(LOG Trick)to make temporarily normal then return to it's original.(Will take it soon).

Regression analysis means تحليل الانحدار



(a) Skewed Left Mean < Median

(b) Symmetric Mean = Median

(c) Skewed Right Mean > Median

Mean

$$\frac{\Sigma x}{n}$$

Add all the numbers together...

...then divide by however many there are.

Sharpen your pencil Solution

Have a go at calculating the mean age of the Power Workout class? Here are their ages.

Age	19	20	21
Frequency	1	3	1

To find μ, we need to add all the people's ages, and divide by how many there are.

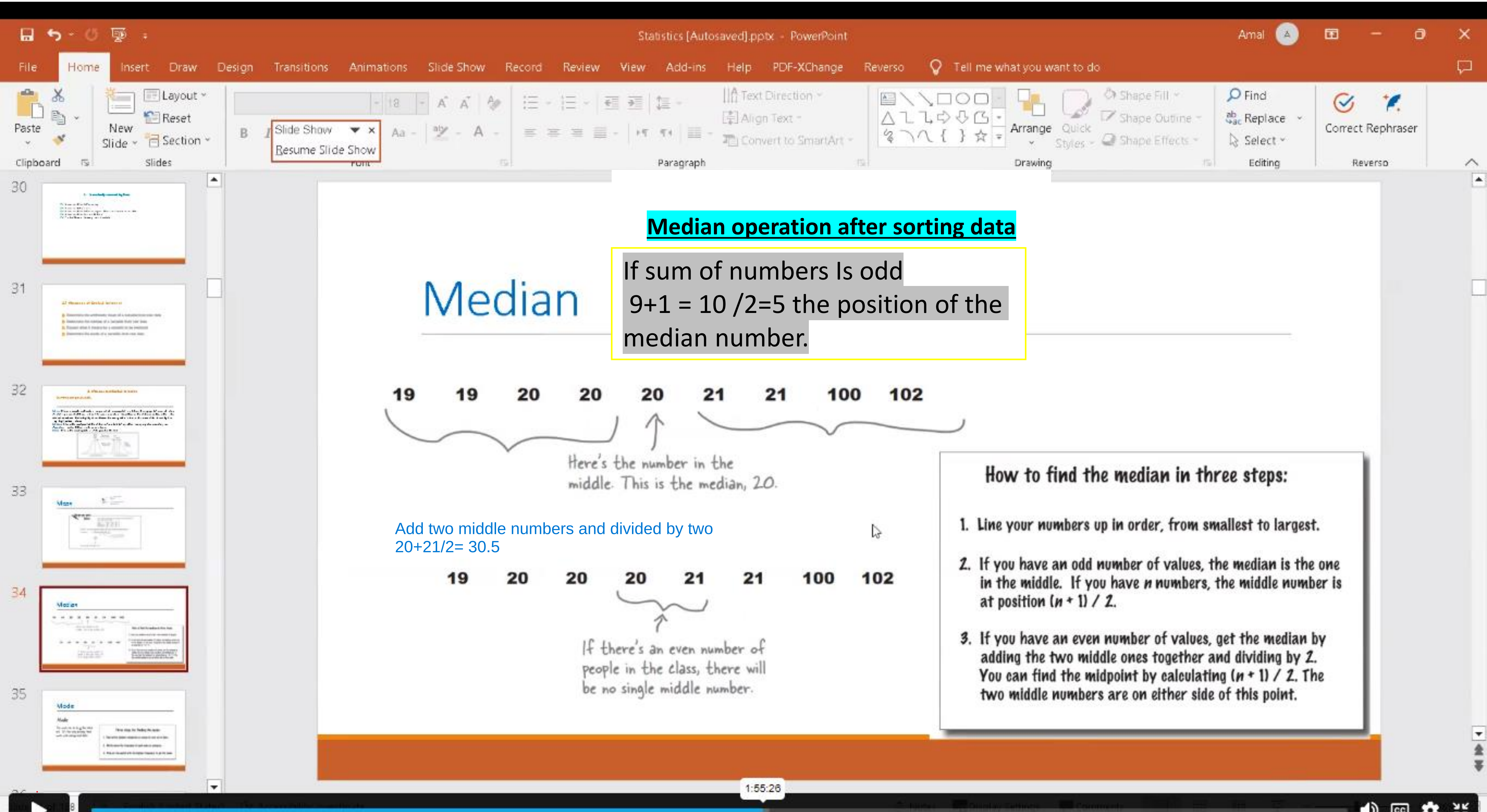
This gives us $\mu = \frac{19 + 20 + 20 + 20 + 21}{5}$

Remember that there are 3 people of age 20.

$= \frac{100}{5}$

$= 20$

The mean age of the class is 20.


Statistics [Autosaved].pptx - PowerPoint
Median operation after sorting data
Median
If sum of numbers Is odd
9+1 = 10 /2=5 the position of the median number.
19 19 20 20 20 21 21 100 102
Here's the number in the middle. This is the median, 20.
Add two middle numbers and divided by two
20+21/2= 30.5
19 20 20 20 21 21 100 102
If there's an even number of people in the class, there will be no single middle number.
How to find the median in three steps:
1. Line your numbers up in order, from smallest to largest.
2. If you have an odd number of values, the median is the one in the middle. If you have n numbers, the middle number is at position (n + 1) / 2.
3. If you have an even number of values, get the median by adding the two middle ones together and dividing by 2. You can find the midpoint by calculating (n + 1) / 2. The two middle numbers are on either side of this point.
1:55:26

Mode

Mode

Only average work with categorical data

The mode has to be in the data set. It's the only average that works with categorical data.

Three steps for finding the mode:

1. Find all the distinct categories or values in your set of data.
2. Write down the frequency of each value or category.
3. Pick out the one(s) with the highest frequency to get the mode.

Example 1: M&Ms

The following data represent the weights (in grams) of a simple random sample of 50 M&M plain candies.

0.87	0.88	0.82	0.90	0.90	0.84	0.84
0.91	0.94	0.86	0.86	0.86	0.88	0.87
0.89	0.91	0.86	0.87	0.93	0.88	
0.83	0.95	0.87	0.93	0.91	0.85	
0.91	0.91	0.86	0.89	0.87	0.84	
0.88	0.88	0.89	0.79	0.82	0.83	
0.90	0.88	0.84	0.93	0.81	0.90	
0.88	0.92	0.85	0.84	0.84	0.86	

Source: Michael Sullivan

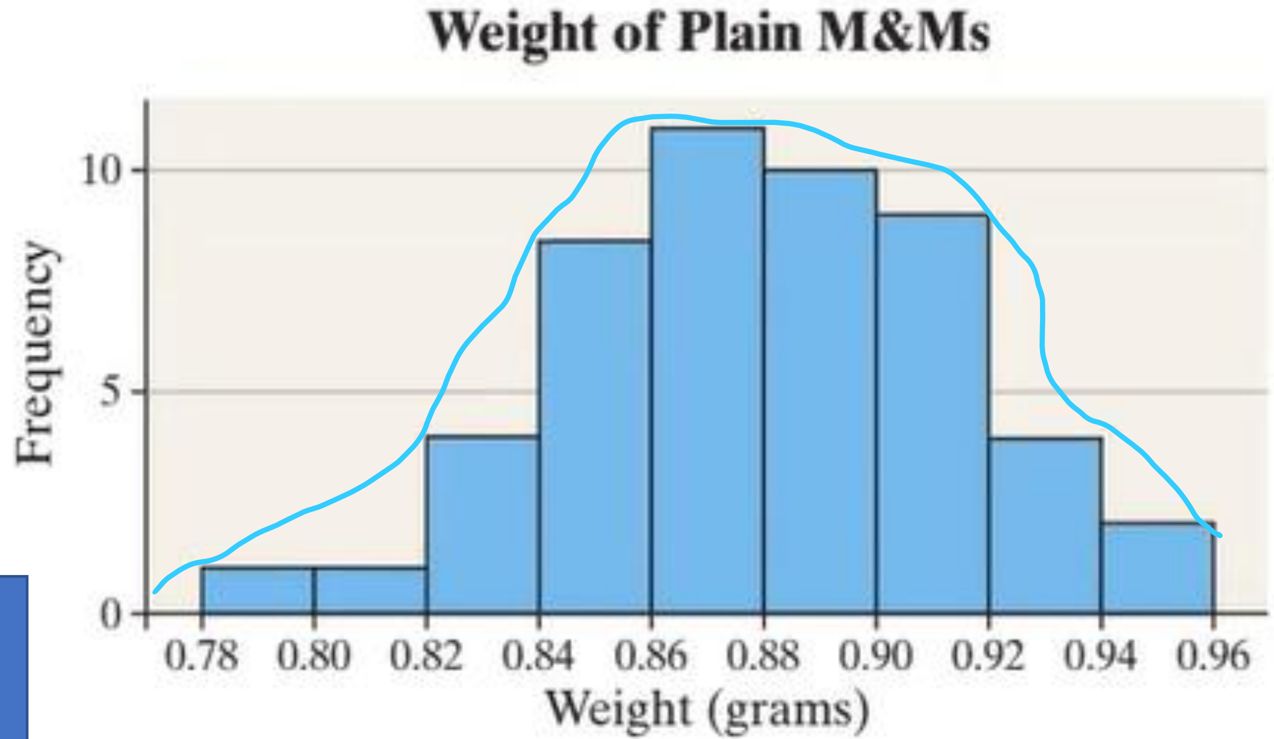
Determine the shape of the distribution of weights of M&Ms by drawing a frequency histogram. Find the mean and median. Which measure of central tendency better describes the weight of a plain M&M?

Example 1: M&Ms

x = 0.875 gram; M = 0.875 gram. The distribution is symmetric, so the mean is the better measure of central tendency

μ

$\bar{x}$



The two means the MEAN
MEAN (MU) for population
X or x bar for Sample

Conclusion :
1- Symmetric data set
2- NO OUTLIER.

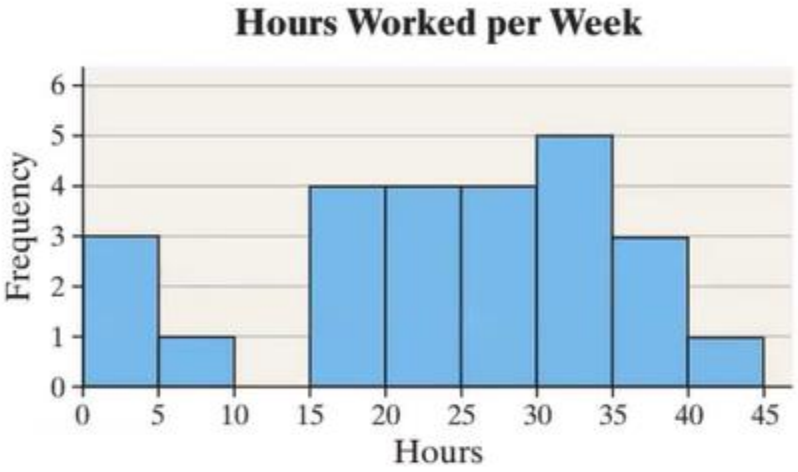
Example 2: Hours Working

A random sample of 25 college students was asked, "How many hours per week typically do you work outside the home?" Their responses were as follows:

0	0	15	20	30
40	30	20	35	35
28	15	20	25	25
30	5	0	30	24
28	30	35	15	15

Example 2: Hours Working

The distribution is skewed left; $\bar{x} = 22$ hours; $M = 25$ hours. The median is the better measure of central tendency



Outliers = 0,5 affect the MEAN Measure
So MEDIAN is better measurement
To deal with this case : you must Drop the outlier from the table.
To analyze the data between range 15:40

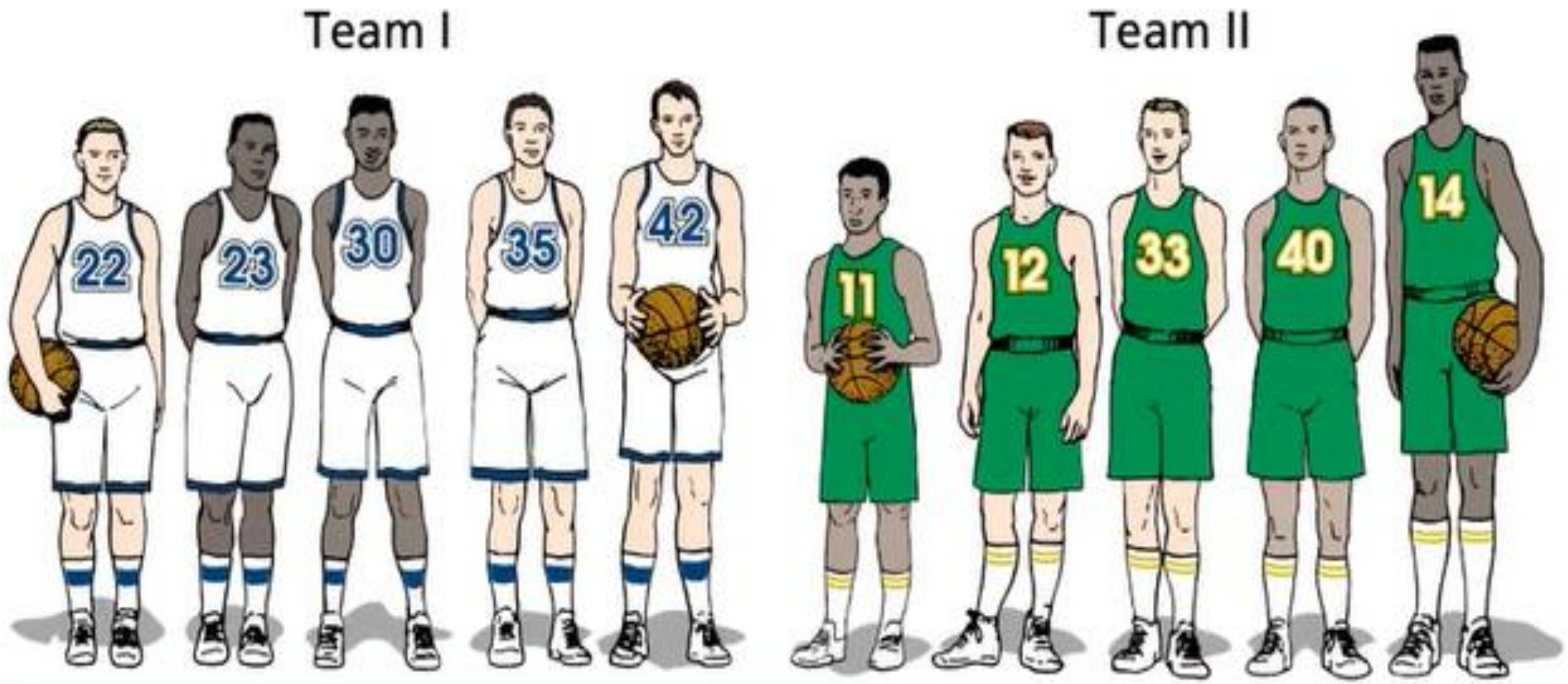
Determine the shape of the distribution of hours worked by drawing a frequency histogram.
Find the mean and median.
Which measure of central tendency better describes hoursworked?

2.2 Measures of Dispersion

1. Determine the range of a variable from raw data
2. Determine the standard deviation of a variable from raw data
3. Determine the variance of a variable from raw data
4. Use the Empirical Rule to describe data that are bell shaped

Standard deviation
Is the difference between it
and Mean

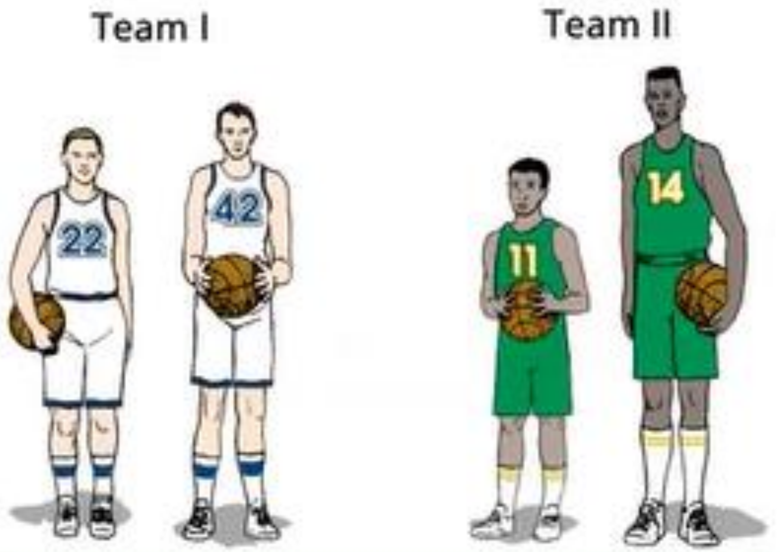
What happen if STD far from Mean : it Means a lot of variation not clustered data and
Can’t relay on it on prediction (ML).
Ask for more data to work on predict.
Save your time and refuse it.
But can only analyze the data



	Team I					Team II				
Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

The two teams have the same mean height, 75 inches (63); the same median height, 76 inches (64); and the same mode, 76 inches (64). Nonetheless, the two data sets clearly differ. In particular, the heights of the players on Team II vary much more than those on Team I. To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set. Such descriptive measures

1. Determine the Range of a Variable from Raw Data



	Team I		Team II	
Feet and inches	6'	6'6"	5'7"	7'
Inches	72	78	67	84

The **range** of a data set is the difference between the maximum (largest) and minimum (smallest) observations.

Team I: Range = 78 - 72 = 6 inches,

Team II: Range = 84 - 67 = 17 inches

1. Determine the Range of a Variable from Raw Data

The **range**, ***R***, of a variable is the difference between the largest and the smallest data value. That is,

$$\text{Range} = R = \text{largest data value} - \text{smallest data value}$$

Computing the Range of a Set of Data

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Problem The data in Table 8 represent the scores on the first exam of 10 students enrolled in Introductory Statistics. Compute the range.

Approach The range is the difference between the largest and smallest data values.

Solution The highest test score is 94 and the lowest test score is 62. The range is $R = 94 - 62 = 32$

All the students in the class scored between 62 and 94 on the exam. The difference between the best score and the worst score is 32 points.

2-Determine the Standard Deviation of a Variable from Raw Data

The **population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. That is, it is the square root of the mean of the squared deviations about the population mean.

The population standard deviation is symbolically represented by σ (lowercase Greek sigma).

sum

Mean

SIGMA
Standard deviation

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

ObservationX1+x2+x3

where $x_1, x_2, \ldots, x_N$ are the N observations in the population and μ is the population mean.

3-Determine the Variance of a Variable from Raw Data

The **variance** of a variable is the square of the standard deviation. The **population variance** is σ^2 and the **sample variance** is s^2.

Use the Empirical Rule to Describe Data That Are Bell Shaped

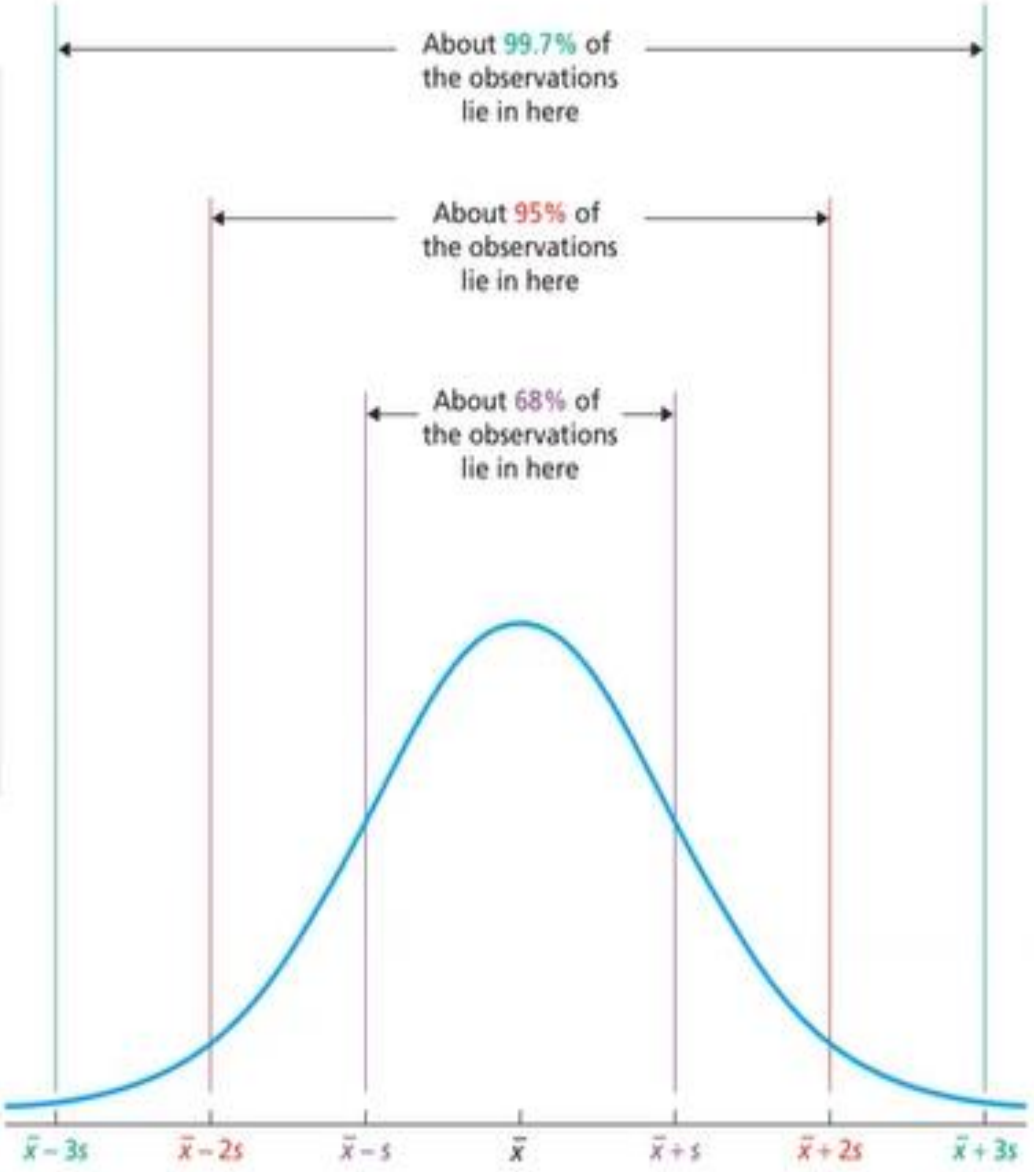
If data have a distribution that is bell shaped, the *Empirical Rule* can be used to determine the percentage of data that will lie within k standard deviations of the mean.

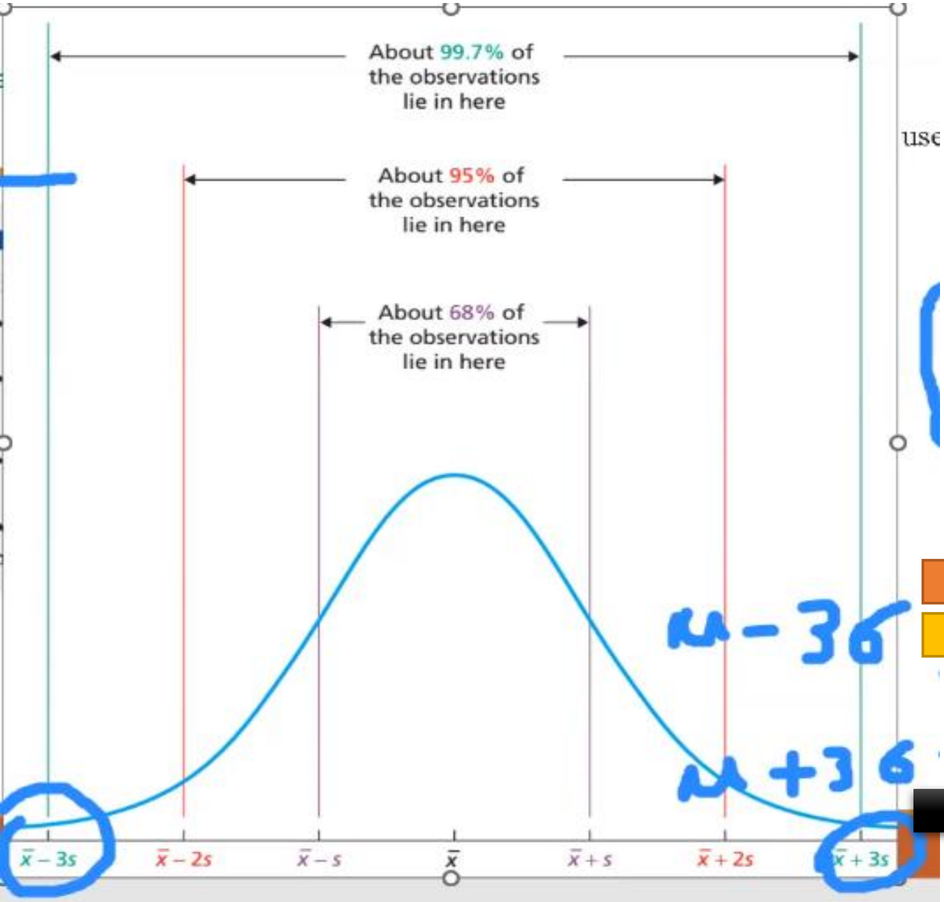
The Empirical Rule

If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: We can also use the Empirical Rule based on sample data with $\bar{x}$ used in place of μ and s used in place of σ.





Benefits from Empirical role:
1- Define the outlier

```
In [4]: 1 df.describe()
Out[4]:
```

	wind_speed	tmin	tmax	t	dayLenght	rh	GDD	PTU	HYTU	Prod
count	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000	9867.000000
mean (average)	4.505868	15.492754	28.065555	21.330225	11.997602	55.031161	6.830225	207.498762	916.131627	2.820439
std (Sigma)	1.203268	5.113966	6.379263	5.446211	1.400406	12.685606	5.446211	83.476255	359.506825	0.098942
min	1.382345	2.152200	10.856100	7.639600	9.985629	3.345892	3.139600	31.415192	59.481575	2.617500
25%	3.718388	10.985500	22.597450	16.319900	10.616327	50.642187	11.819900	127.289356	629.516456	2.778900
50% (Median)	4.406298	15.532700	28.797400	21.642400	12.000000	57.460712	17.142400	208.285750	880.389701	2.781000
75%	5.128551	20.198750	33.707900	26.558200	13.371042	62.870843	22.058200	289.578827	1223.286253	2.895400
max	12.898120	25.362900	44.825700	34.063200	14.014371	93.192306	29.563200	404.932133	1729.098153	3.052500

$\mu - 3\sigma =$

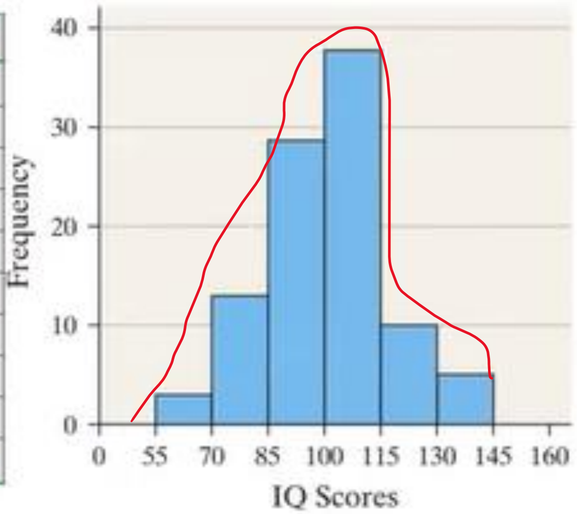
$\mu + 3\sigma =$

```
In [5]: 1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9867 entries, 0 to 9866
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   date       9867 non-null   object
```

4-Use the Empirical Rule to Describe Data That Are Bell Shaped

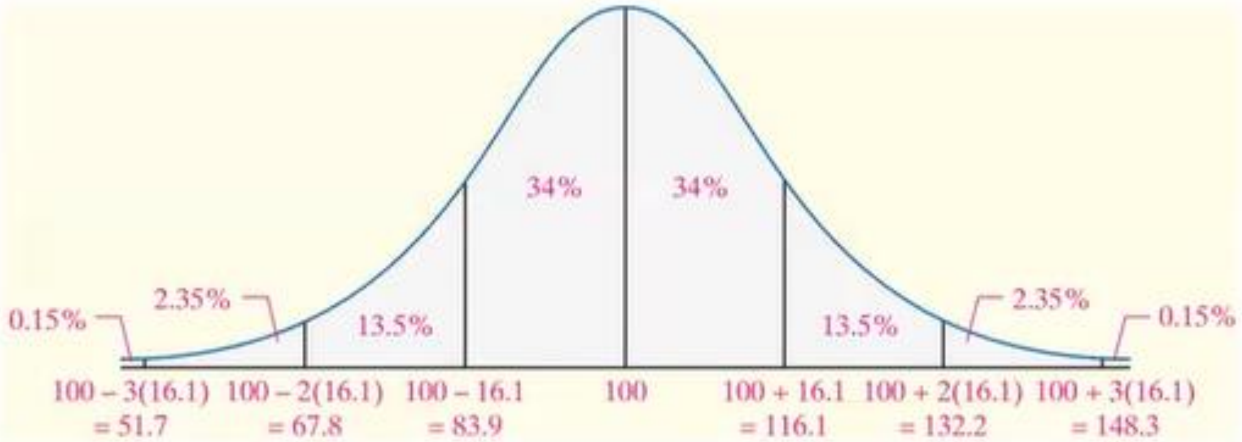
University A									
73	103	91	93	136	108	92	104	90	78
108	93	91	78	81	130	82	86	111	93
102	111	125	107	80	90	122	101	82	115
103	110	84	115	85	83	131	90	103	106
71	69	97	130	91	62	85	94	110	85
102	109	105	97	104	94	92	83	94	114
107	94	112	113	115	106	97	106	85	99
102	109	76	94	103	112	107	101	91	107
107	110	106	103	93	110	125	101	91	119
118	85	127	141	129	60	115	80	111	79



(a) Determine the percentage of students who have IQ scores within 3 standard deviations of the mean according to the Empirical Rule.
(b) Determine the percentage of students who have IQ scores between 67.8 and 132.2 according to the Empirical Rule.
(c) Determine the actual percentage of students who have IQ scores between 67.8 and 132.2.
(d) According to the Empirical Rule, what percentage of students have IQ scores above 132.2?

4-Use the Empirical Rule to Describe Data That Are Bell Shaped

The histogram is roughly bell shaped. From the mean IQ score of the students enrolled in University A is 100 and the standard deviation is 16.1. To make the analysis easier, we draw a bell-shaped curve like the one in Figure 13, with x = 100 and s = 16.1.



(a) According to the Empirical Rule, approximately 99.7% of the IQ scores are within 3 standard deviations of the mean [that is, greater than or equal to 100 - 3116.12 = 51.7 and less than or equal to 100 + 3116.12 = 148.3].

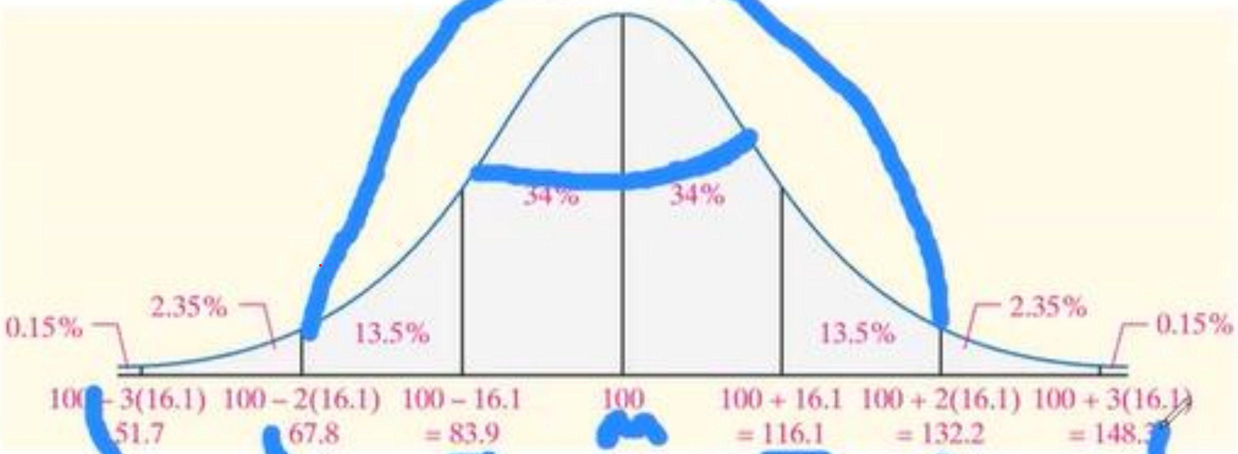
(b) Since 67.8 is exactly 2 standard deviations below the mean [100 - 2116.12 = 67.8] and 132.2 is exactly 2 standard deviations above the mean [100 + 2116.12 = 132.2], the Empirical Rule tells us that approximately 95% of the IQ scores li between 67.8 and 132.2.

(c) Of the 100 IQ scores listed in Table 7, 96, or 96%, are between 67.8 and 132.2. This is very close to the Empirical Rule approximation.

(d) Based on Figure, approximately 2.35% + 0.15% = 2.5% of students at University A will have IQ scores above 132.2.

4-Use the Empirical Rule to Describe Data That Are Bell Shaped

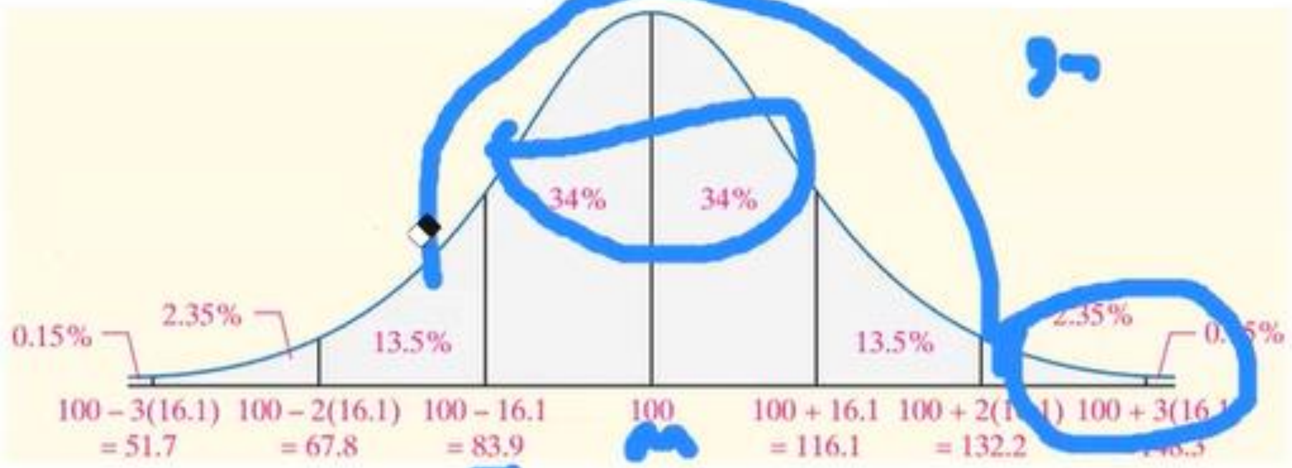
The histogram is roughly bell shaped. From the mean IQ score of the students enrolled in University A is 100 and the standard deviation is 16.1. To make the analysis easier, we draw a bell-shaped curve like the one in Figure 13, with $\bar{x}$ = 100 and s = 16.1.



(a) According to the Empirical Rule, approximately 99.7% of the IQ scores are within 3 standard deviations of the mean [that is, greater than or equal to 100 - 3116.12 = 51.7 and less than or equal to 100 + 3116.12 = 148.3].
(b) Since 67.8 is exactly 2 standard deviations below the mean [100 - 2116.12 = 67.8] and 132.2 is exactly 2 standard deviations above the mean [100 + 2116.12 = 132.2], the Empirical Rule tells us that approximately 95% of the IQ scores lie between 67.8 and 132.2.
(c) Of the 100 IQ scores listed in Table 7, 96, or 96%, are between 67.8 and 132.2. This is very close to the Empirical Rule's approximation.
(d) Based on Figure, approximately 2.35% + 0.15% = 2.5% of students at University A will have IQ scores above 132.2.

se the Empirical Rule to Describe Data That Are Bell Shaped

e histogram is roughly bell shaped. From the mean IQ score of the students enrolled in University A is 100 and the ndard deviation is 16.1. To make the analysis easier, we draw a bell-shaped curve like the one in Figure 13, with $\bar{x}$ =) and s = 16.1.



According to the Empirical Rule, approximately 99.7% of the IQ scores are within 3 standard deviations of the mean it is, greater than or equal to 100 - 3116.12 = 51.7 and less than or equal to 100 + 3116.12 = 148.3].

Since 67.8 is exactly 2 standard deviations below the mean [100 - 2116.12 = 67.8] and 132.2 is exactly 2 standard iations above the mean [100 + 2116.12 = 132.2], the Empirical Rule tells us that approximately 95% of the IQ scores lie ween 67.8 and 132.2.

Of the 100 IQ scores listed in Table 7, 96, or 96%, are between 67.8 and 132.2. This is very close to the Empirical Rule's roximation.

Based on Figure, approximately 2.35% + 0.15% = 2.5% of students at University A will have IQ scores above 132.2.

2-Determine the Standard Deviation of a Variable from Raw Data

The **population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. That is, it is the square root of the mean of the squared deviations about the population mean.

The population standard deviation is symbolically represented by σ (lowercase Greek sigma).

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N}}$$

where $x_1, x_2, \ldots, x_N$ are the N observations in the population and μ is the population mean.

x	μ	$x-\mu$	$(x-\mu)^2$
10	12	-2	4
12		0	0
14		2	4
			8

$$\sigma = \sqrt{\frac{8}{3}} =$$