Business Analytics

Basic Data Exploration with Statistics

Proschool
An IMS Initiative

Introduction

- Importance of Analytics.
    - Attrition
    - Car Pooling
    - Credit Approval
    - Marketing
    - Presidential Campaigns
- Backbone of Analytics
    - Mathematics
    - Stats

Data/Attribute Types

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - E.g., customer _ID, name, address
- Types:
  - Nominal = Categorical
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or “names of things”
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes

- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to the important outcome (e.g., HIV positive)

Attribute Types

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- Interval
  - Measured on a scale of equal-sized units
  - Values have order
    - » E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
  - Rare Attribute Type
- Ratio
  - Inherent zero-point
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - – e.g., temperature in Kelvin, length, counts, monetary quantities

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Analysis of Attribute Types

Scale	Definition	Examples
Nominal	Categorizes but does not rank	Industries, Gender, Occupation
Ordinal	Ranked Categories. Differences between ranks not consistent	Organizational Hierarchy. Star Ratings
Interval	Ranks Data. Differences between ranks equal. No True Zero Point.	Celsius or Fahrenheit Scale. Dates
Ratio	Ranks Data. Differences between ranks equal. Also has a True Zero Point.	Rate of Return, Money

Frequency Distribution

- Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval.
- In Frequency distribution, we find the number of counts for a particular observation when the observations are repeated.

For example:

The height of children can be split into several different categories or ranges. When measuring the height of 100 children, some are tall and some are short, but there is a high probability of a higher frequency or concentration in the middle range. The most important factors are that the intervals used must be non-overlapping and must contain all of the possible observations.

Frequency Distribution in Excel

Steps:

- Enter the data in column and sort it in ascending order by **Data → Sort**
- Note the minimum & maximum values and prepare the bins of appropriate sizes e.g , suppose the data ranges from 0 to 100 then the bins will be (1-10,11-21,…,91-100).
- Now create a table with column headers ‘Bins’ and ‘Frequency’ and type in the bin ranges.
- Use the **countif** function to fill in the frequency column for each of the bin.

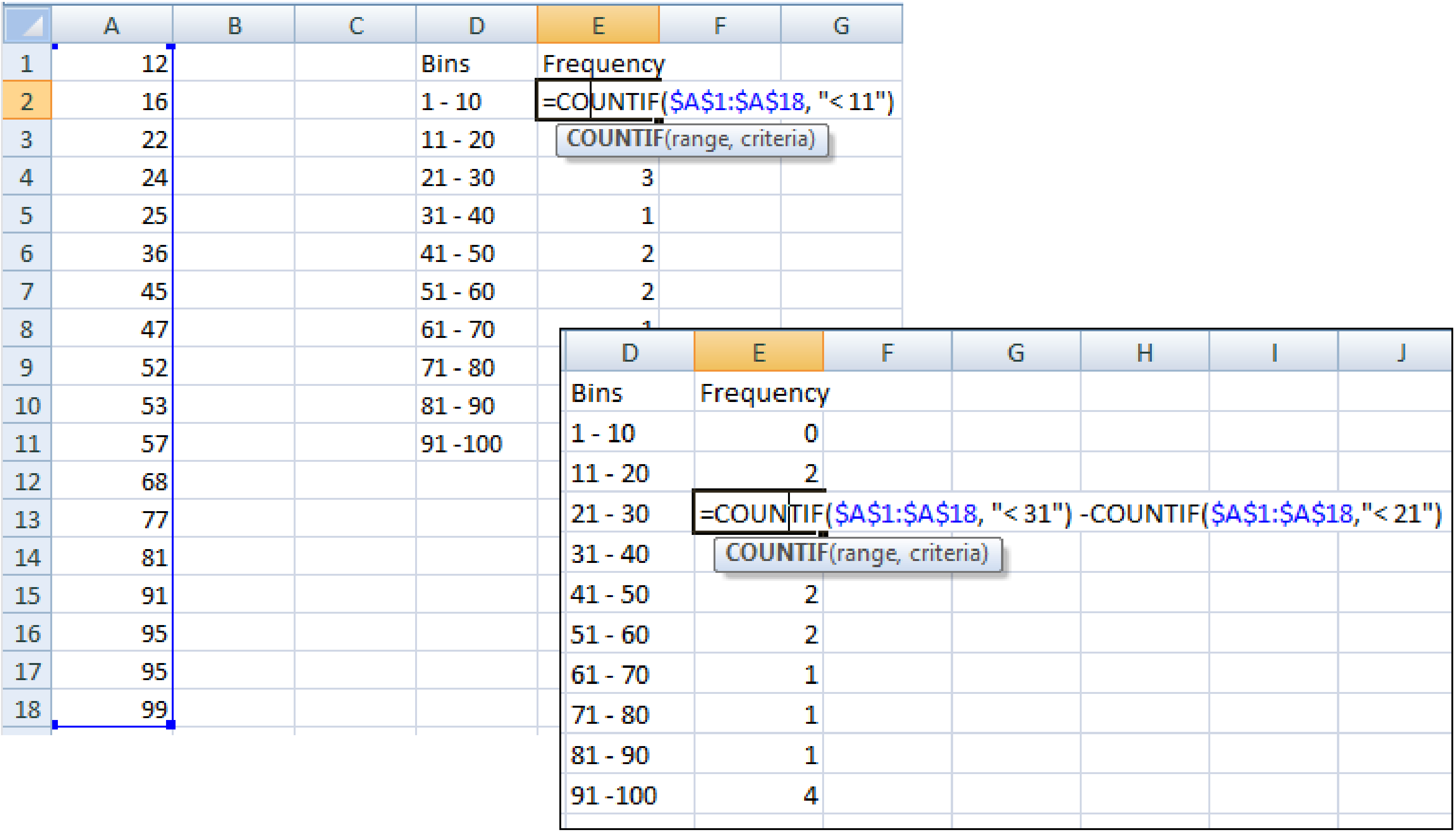
For the first bin:

countif([select data], “[< starting value of second bin]”)

For rest of the bins:

countif([select data], “[< starting value of next highest bin]”) – countif([select data], “[< starting value of selected bin]”)

Frequency Distribution in Excel

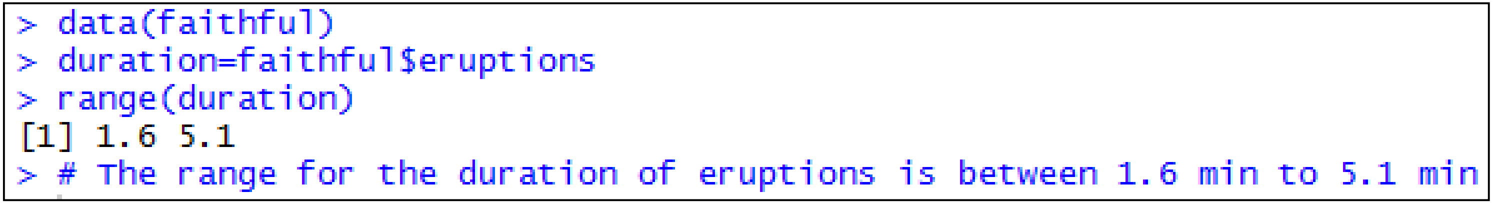


	A	B	C	D	E	F	G
1	12			Bins	Frequency		
2	16			1 - 10	=COUNTIF($A$1:$A$18, "< 11")		
3	22			11 - 20			
4	24			21 - 30	3		
5	25			31 - 40	1		
6	36			41 - 50	2		
7	45			51 - 60	2		
8	47			61 - 70	1		
9	52			71 - 80			
10	53			81 - 90			
11	57			91 -100			
12	68						
13	77						
14	81						
15	91						
16	95						
17	95						
18	99						

D	E	F	G	H	I	J
Bins	Frequency					
1 - 10	0					
11 - 20	2					
21 - 30	=COUNTIF($A$1:$A$18, "< 31") -COUNTIF($A$1:$A$18,"< 21")					
31 - 40						
41 - 50	2					
51 - 60	2					
61 - 70	1					
71 - 80	1					
81 - 90	1					
91 -100	4					

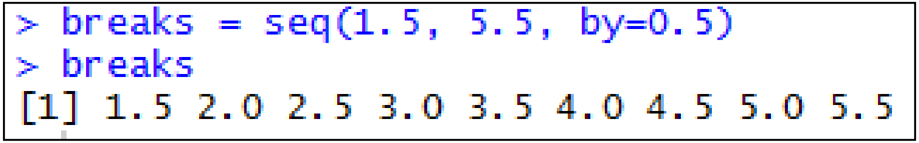
Frequency Distribution in R

Consider the inbuilt R dataset faithful.
We will find the frequency distribution for the variable ‘eruptions’.



```
> data(faithful)
> duration=faithful$eruptions
> range(duration)
[1] 1.6 5.1
> # The range for the duration of eruptions is between 1.6 min to 5.1 min
```

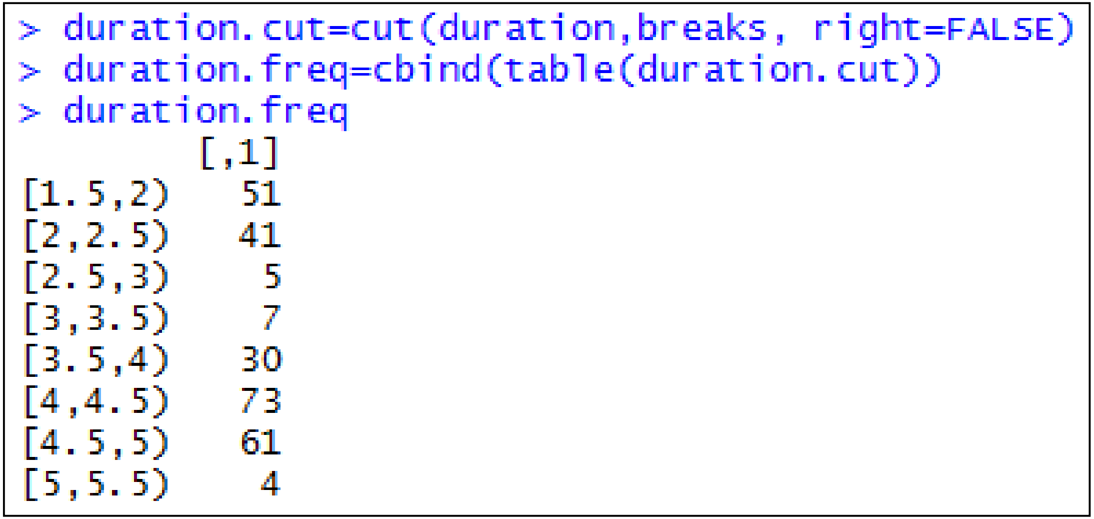
Now we break the range into non-overlapping sub-intervals by defining a sequence of equal distance break points. If we round the endpoints of the interval [1.6, 5.1] to the closest half-integers, we come up with the interval [1.5, 5.5]. Hence we set the break points to be the half-integer sequence { 1.5, 2.0, 2.5, ... }.



```
> breaks = seq(1.5, 5.5, by=0.5)
> breaks
[1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

Frequency Distribution in R

Classify the eruption durations according to the half-unit-length sub-intervals with cut. As the intervals are to be closed on the left, and open on the right, we set the right argument as FALSE.



```
> duration.cut=cut(duration,breaks, right=FALSE)
> duration.freq=cbind(table(duration.cut))
> duration.freq
         [,1]
[1.5,2)    51
[2,2.5)    41
[2.5,3)     5
[3,3.5)     7
[3.5,4)    30
[4,4.5)    73
[4.5,5)    61
[5,5.5)     4
```

Histogram

- Histogram is a graph used for the representation of the frequency distribution.
- It is series of adjacent rectangles erected on X-axis with class interval as base and Frequency on the Y-axis.
- Histograms are useful to find mode and understand the spread of the distribution (which will be discussed later).

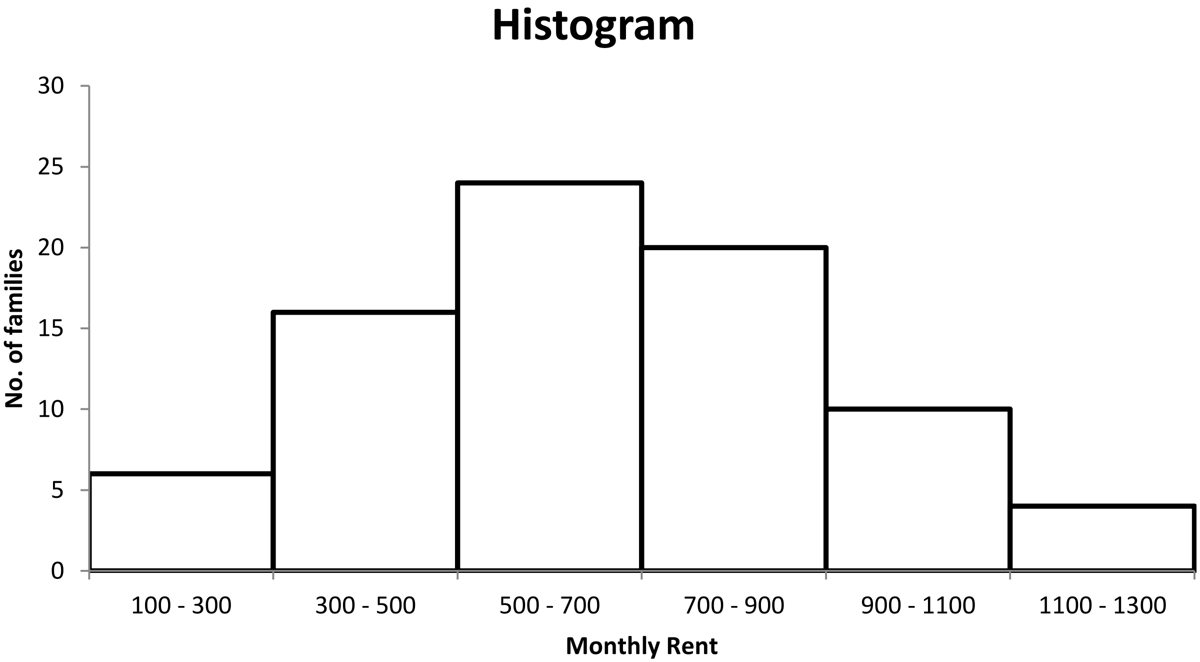
Example:
Consider the following data to plot the histogram

Monthly House Rent	100 - 300	300 - 500	500 - 700	700 - 900	900 - 1100	1100 - 1300
No. of families	6	16	24	20	10	4

Histogram in Excel

- Select the Frequency column and go to
  Insert → Charts → Column → 2D Clustered Column
- To get the correct bin labels on the horizontal axis, right click anywhere on the chart and go to **Select Data**. Click on the small box to the right of the **"Category (X) axis labels."** Highlight the bin labels in the frequency distribution table and hit **Return** and click **Ok**.
- Double click on one of the bars in the chart, which will open the **"Format Data Series"** box. Under **Options**, change the **"Gap width"** to be 0% and click **Ok**.
- Give appropriate title name and axes names.

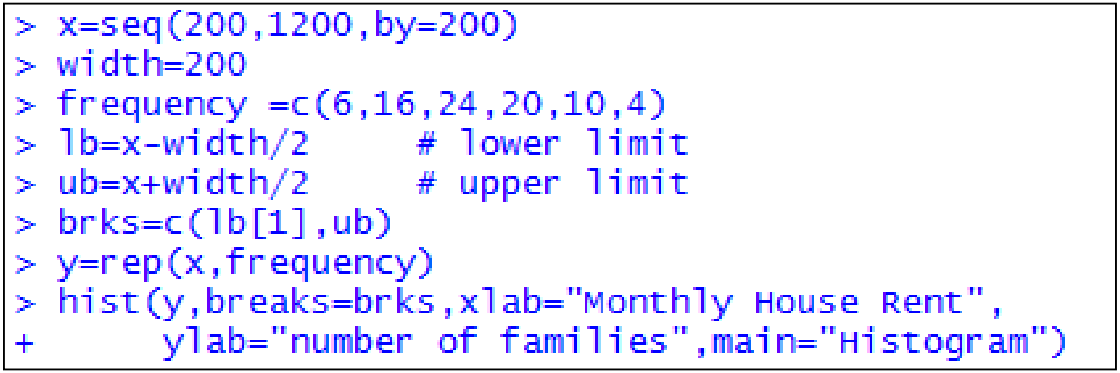
Histogram in Excel



Histogram in R

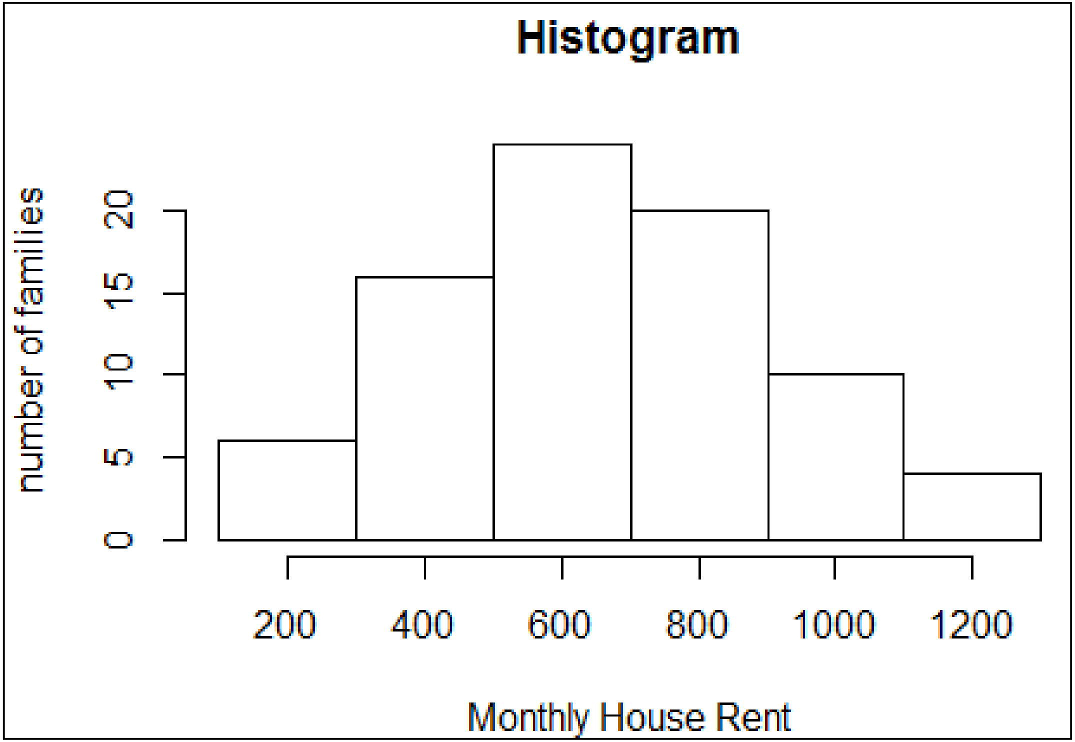
Consider the same data used for histogram in excel.

- Find the midpoints ((lower limit + upper limit)/2) for all the intervals. e.g. if the interval is 100 – 300 then 100 is lower limit and 300 is upper limit.
- These mid points will form x in our example.



```
> x=seq(200,1200,by=200)
> width=200
> frequency =c(6,16,24,20,10,4)
> lb=x-width/2       # lower limit
> ub=x+width/2       # upper limit
> brks=c(lb[1],ub)
> y=rep(x,frequency)
> hist(y,breaks=brks,xlab="Monthly House Rent",
+       ylab="number of families",main="Histogram")
```

Histogram in R



Frequency Polygon

- Frequency polygon is an another way of representing the frequency distribution graphically.
- It enables us to understand the pattern in the data more clearly.
- Mid-values are taken on X-axis and frequencies are taken on Y-axis and the successive points are joined by the line segments.
- To complete the polygon we obtain closed figure by taking two more classes, one preceding to first class and the other succeeding to last class. Frequency of these classes is taken to be zero.

Frequency Polygon in Excel

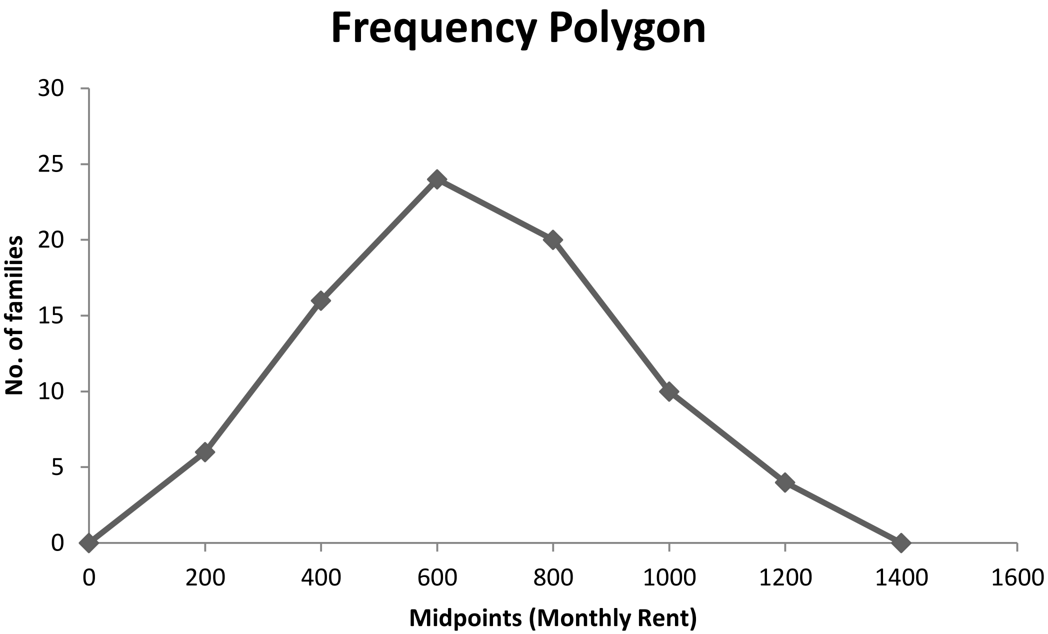
Consider the same data used for histogram.

- Find the mid-points by taking two more classes and their frequencies as zero.
- Select mid-points and frequency columns and go to

Insert → Chart → Scatter with straight lines and markers

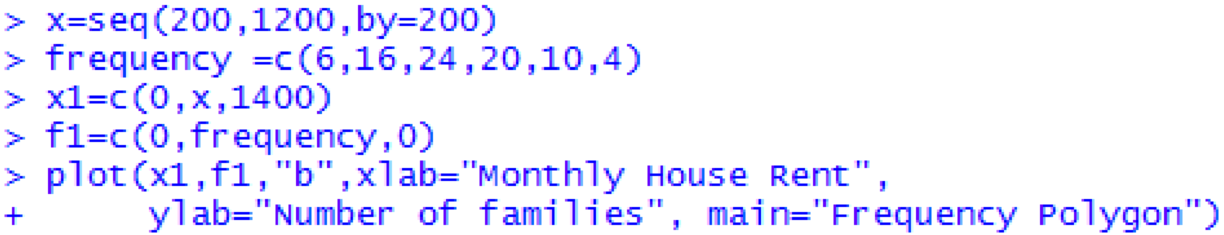
- Give appropriate title name and axes names.

Frequency Polygon in Excel



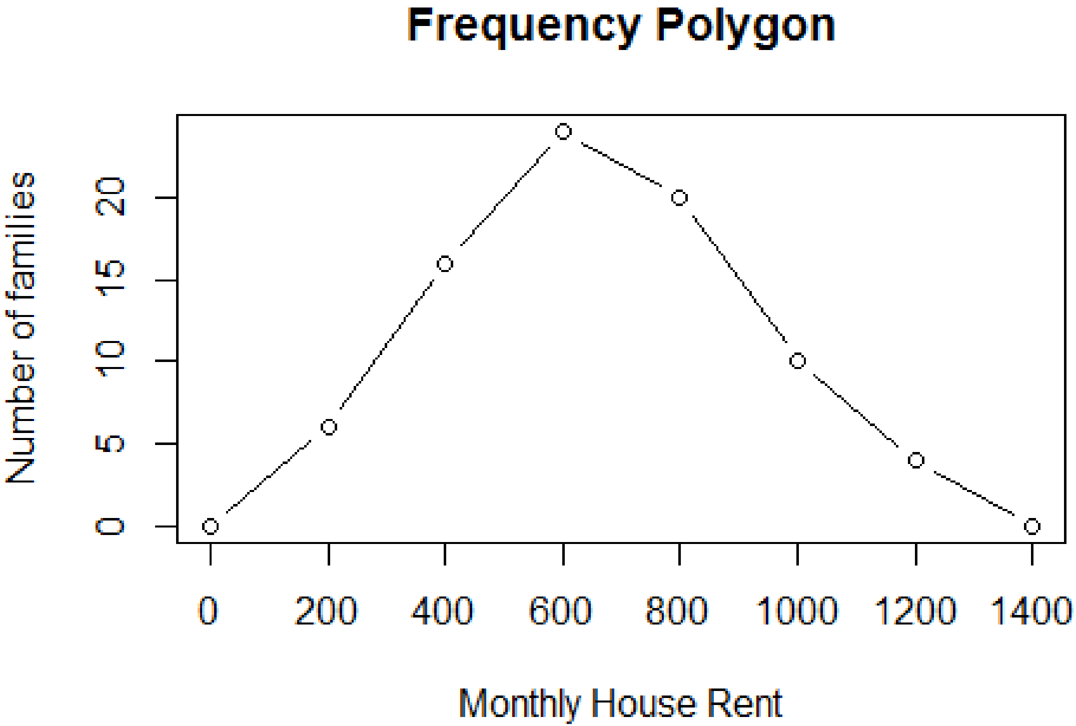
Frequency Polygon in R

Consider the same data used for histogram.



```
> x=seq(200,1200,by=200)
> frequency =c(6,16,24,20,10,4)
> x1=c(0,x,1400)
> f1=c(0,frequency,0)
> plot(x1,f1,"b",xlab="Monthly House Rent",
+       ylab="Number of families", main="Frequency Polygon")
```

Frequency Polygon in R



Business Analytics

Measures of Central Tendency & Dispersion

Proschool

An IMS Initiative

Introduction to Central Tendency and Dispersion

Lets say you are the captain of the Indian Cricket Team. We have already batted and the Australians are batting now. They require 10 runs from the last over to win. You have two good bowlers that still have an over to bowl. Who should you send to bowl the last over?

Without any extra data each is as good as the other.

Suppose you now know that Bowler A concedes 8 runs on an average, Bowler B concedes 6 runs on an average. Now who should we send in to bowl?

Introduction to Central Tendency and Dispersion

Will a lower average always be better in such cases. Lets add some more data to what we already know. Bowler A concedes 8 runs on an average with a standard deviation of 0 runs which means he always concedes exactly 8 runs, neither more nor less (a little unrealistic). Bowler B concedes 6 runs on an average with a standard deviation of 4 runs that means he is quite erratic and may concede anything between zero to a large number. Now who will you choose?

With a consistent eight runs you know that you have a 100% chance of winning.

Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set. (We have already seen this)
- For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data but it can be used on any data type.

Mean and Median

- The mean is the most common measure of the location of a set of points

  However, the mean is very sensitive to outliers
- Thus, the median or a trimmed mean is also commonly used. Trimmed here means outliers are removed.
- Mean can only be used with numeric whereas median works with both ordinal and numeric.

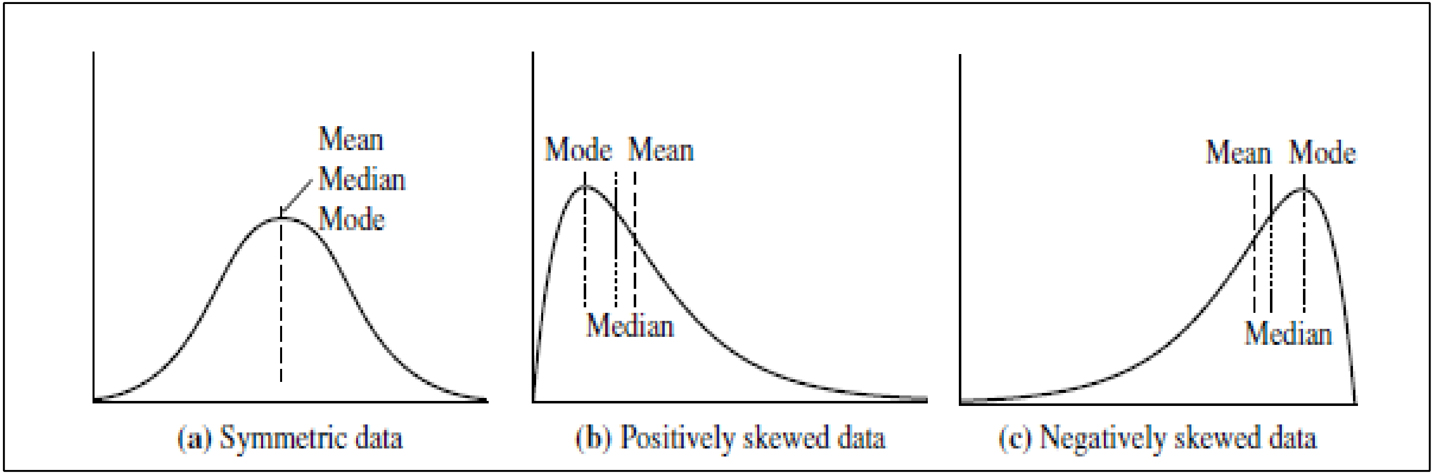
$$\text{Mean(X)} = \bar{X} = \frac{1}{m}\sum_{i=0}^{m} X_i$$

$$Median(X) = \begin{cases} X_{(r+1)} \\ \frac{1}{2}(X_{(r)} + X_{(r+1)}) \end{cases}$$

If *m* is odd, i.e., m = 2r + 1

If m is even, i.e., m = 2r

Skewness and Central Tendency

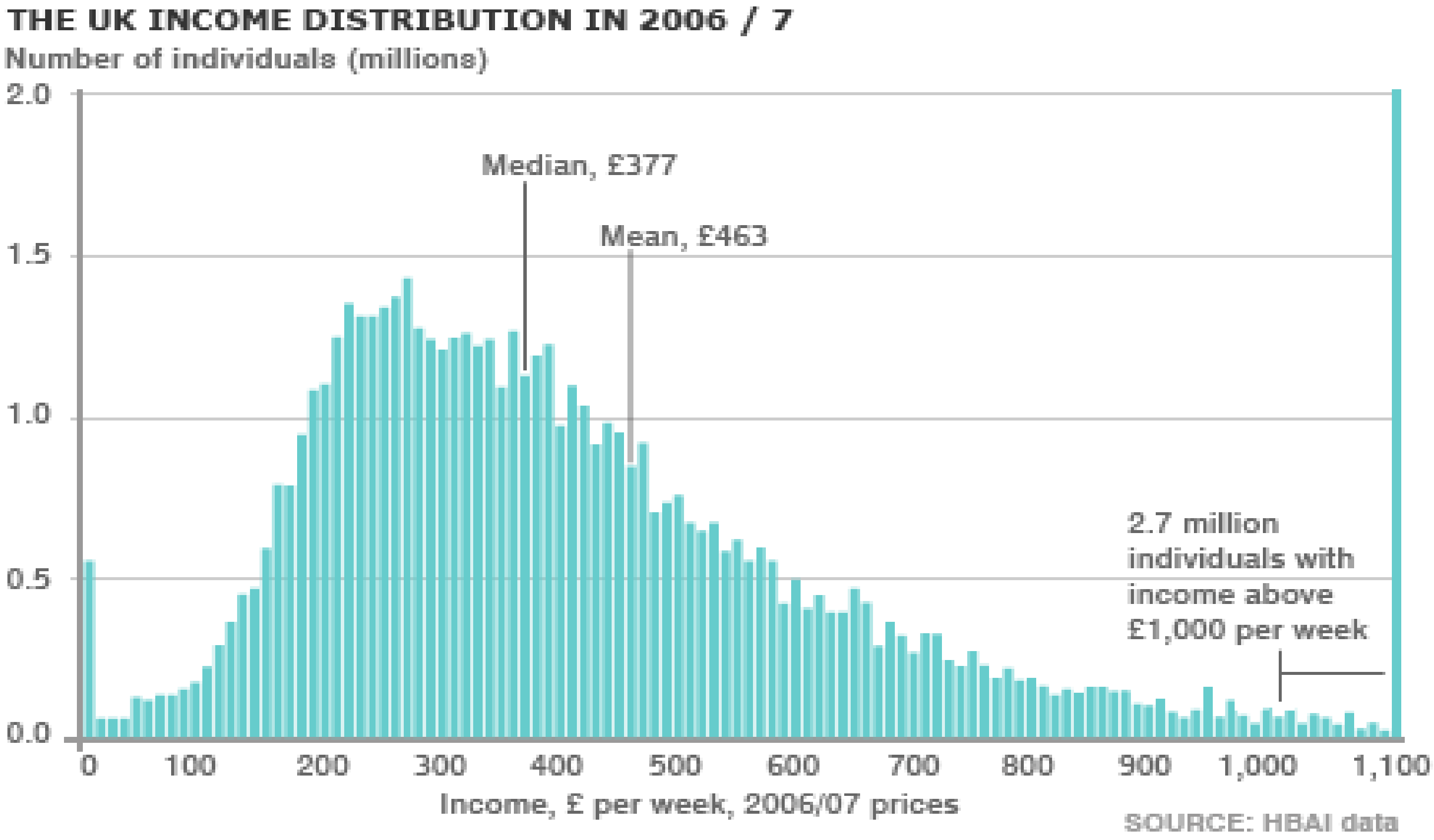


(a) Symmetric data

(b) Positively skewed data

(c) Negatively skewed data

Mean and Median – Resistant Measure?



Measuring Dispersion

- Start with explaining the meaning of dispersion, you can refer again to the bowling example. Continue with percentiles and the following:
- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: $IQR = Q_3 – Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR

Measuring Dispersion

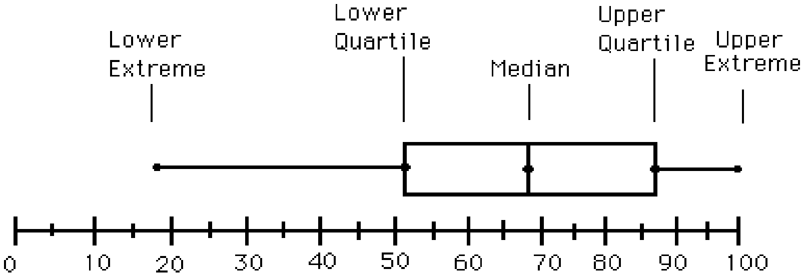
Variance and standard deviation (sample: s, population: σ)

- **Variance**: (algebraic, scalable computation)
- **Standard deviation** s (or σ) is the square root of variance s^2 $^($or $σ^{2)}$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

Boxplots

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum



Boxplot

Data is represented with a box

The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR

The median is marked by a line within the box

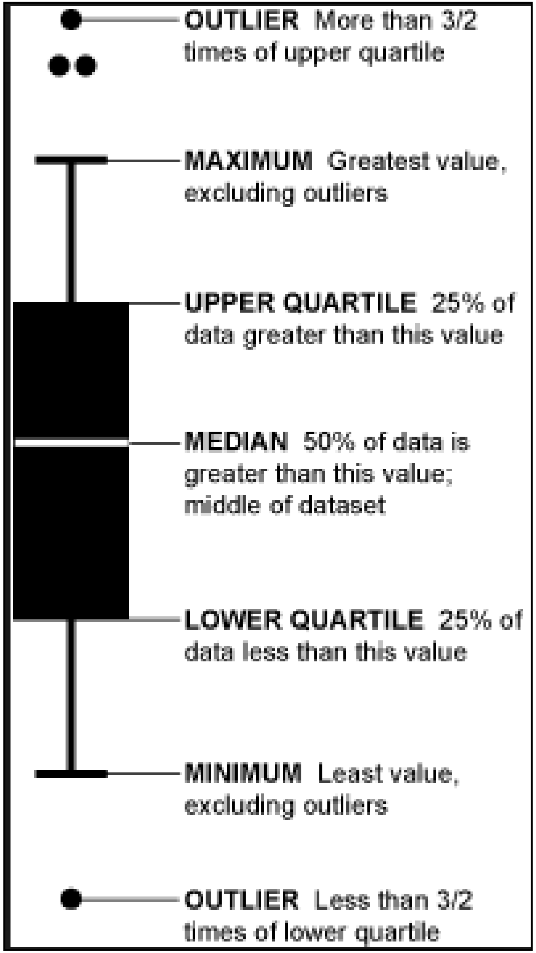
Boxplots

- **Whiskers:**

  Two lines outside the box extended to Minimum and Maximum

- **Outliers:**

  Points beyond a specified outlier threshold, plotted individually


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

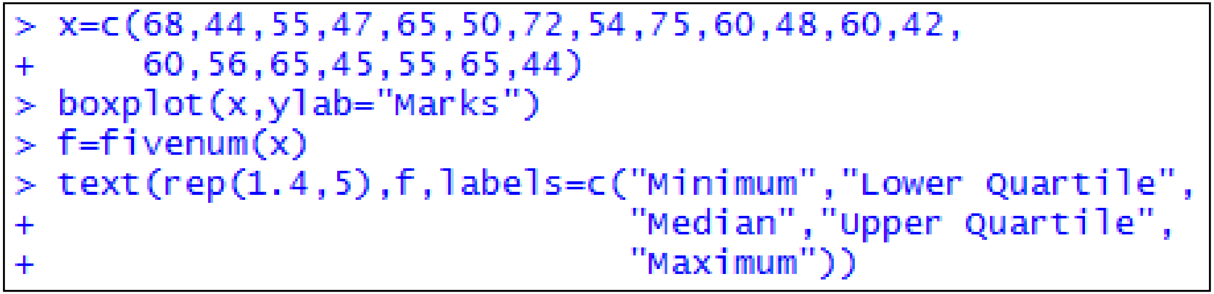
Boxplot in R

Example:

Marks obtained (out of 100) by 20 students are in a class test are as given below:

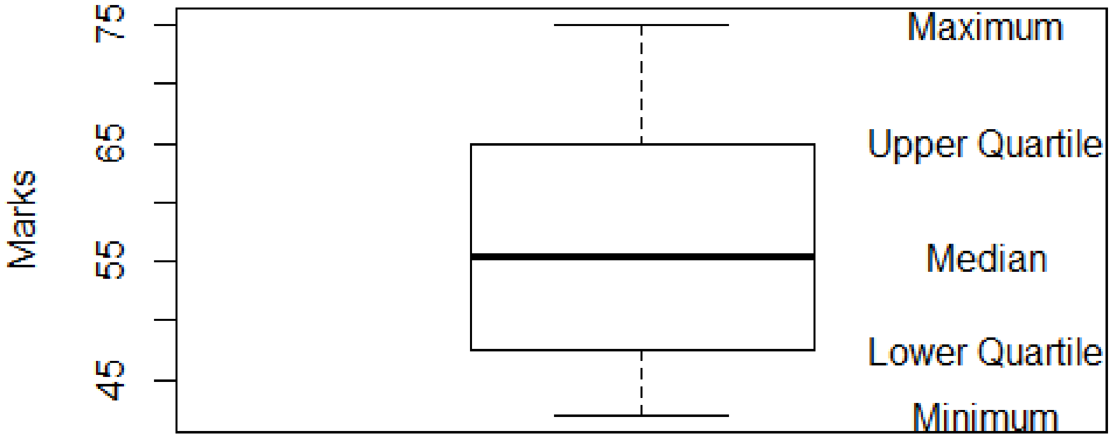
68, 44, 55, 47, 65, 50, 72, 54, 75, 60, 48, 60, 42, 60, 56, 65, 45, 55, 65, 44.

Draw a boxplot.



```
> x=c(68,44,55,47,65,50,72,54,75,60,48,60,42,
+     60,56,65,45,55,65,44)
> boxplot(x,ylab="Marks")
> f=fivenum(x)
> text(rep(1.4,5),f,labels=c("Minimum","Lower Quartile",
+                            "Median","Upper Quartile",
+                            "Maximum"))
```

Boxplot in R



Marks
75
65
55
45
Maximum
Upper Quartile
Median
Lower Quartile
Minimum

Standard Deviation

Standard Deviation

- The Standard Deviation is a measure of how spread out numbers are.
- Symbol : σ (the greek letter sigma)
- The formula is easy: it is the square root of the Variance. So now you ask, "What is the Variance?"

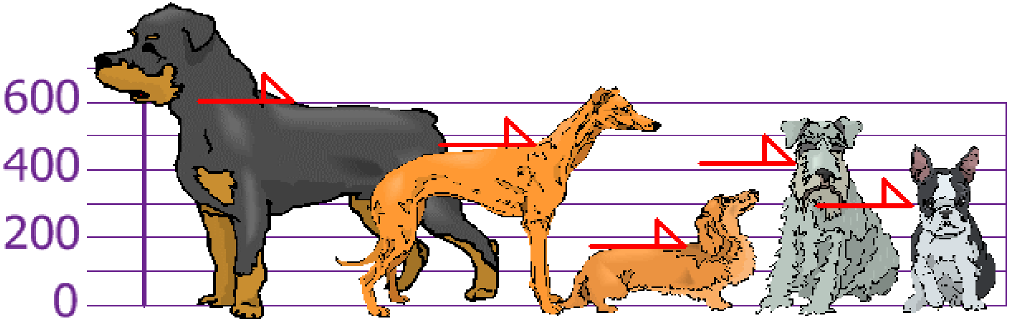
SD & Variance

- **Variance:**
- The Variance is defined as:
- The average of the squared differences from the Mean.

- To calculate the variance follow these steps:
  - Work out the Mean (the simple average of the numbers)
  - Then for each number: subtract the Mean and square the result (the squared difference).
  - Then work out the average of those squared differences.

Example

- You and your friends have just measured the heights of your dogs (in millimeters):



- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.
- Find out the Mean, the Variance, and the Standard Deviation.

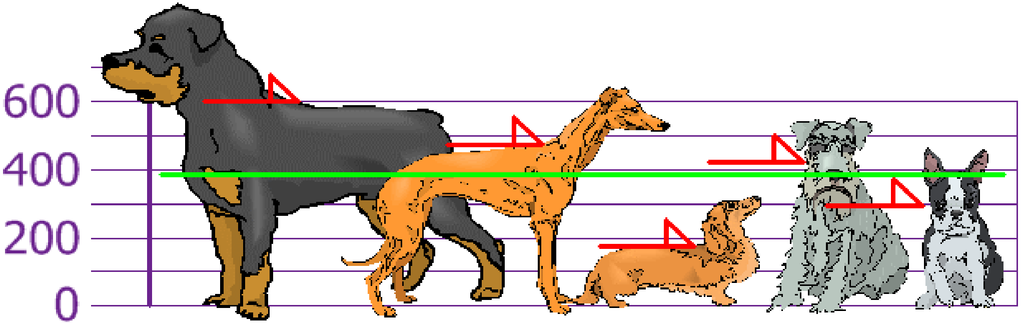
SD & Variance

Your first step is to find the Mean:

Answer:

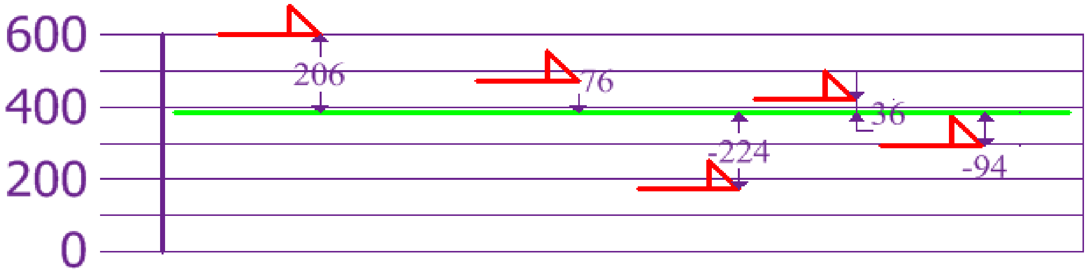
$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



SD & Variance

- Now we calculate each dog's difference from the Mean:



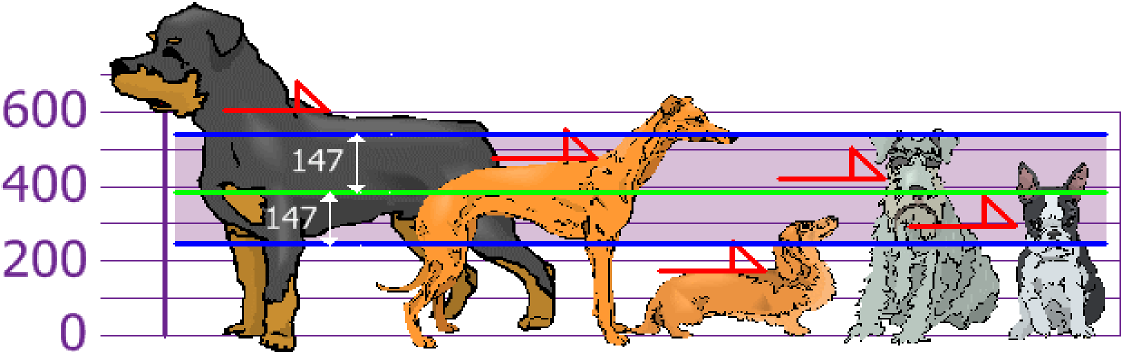
- To calculate the Variance, take each difference, square it, and then average the result:

$$\text{Variance: } \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$
$$= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5}$$
$$= \frac{108,520}{5} = 21,704$$

SD & Variance

- So, the Variance is 21,704.
- And the Standard Deviation is just the square root of Variance, so:
- Standard Deviation: σ = √21,704 = 147.32... = 147 (to the nearest mm)
- And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:

SD & Variance



- So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.
- Rottweilers are tall dogs. And Dachshunds are a bit short ... but don't tell them!

SD & Variance

- Our example was for a Population (the 5 dogs were the only dogs we were interested in).

- But if the data is a Sample (a selection taken from a bigger Population), then the calculation changes!

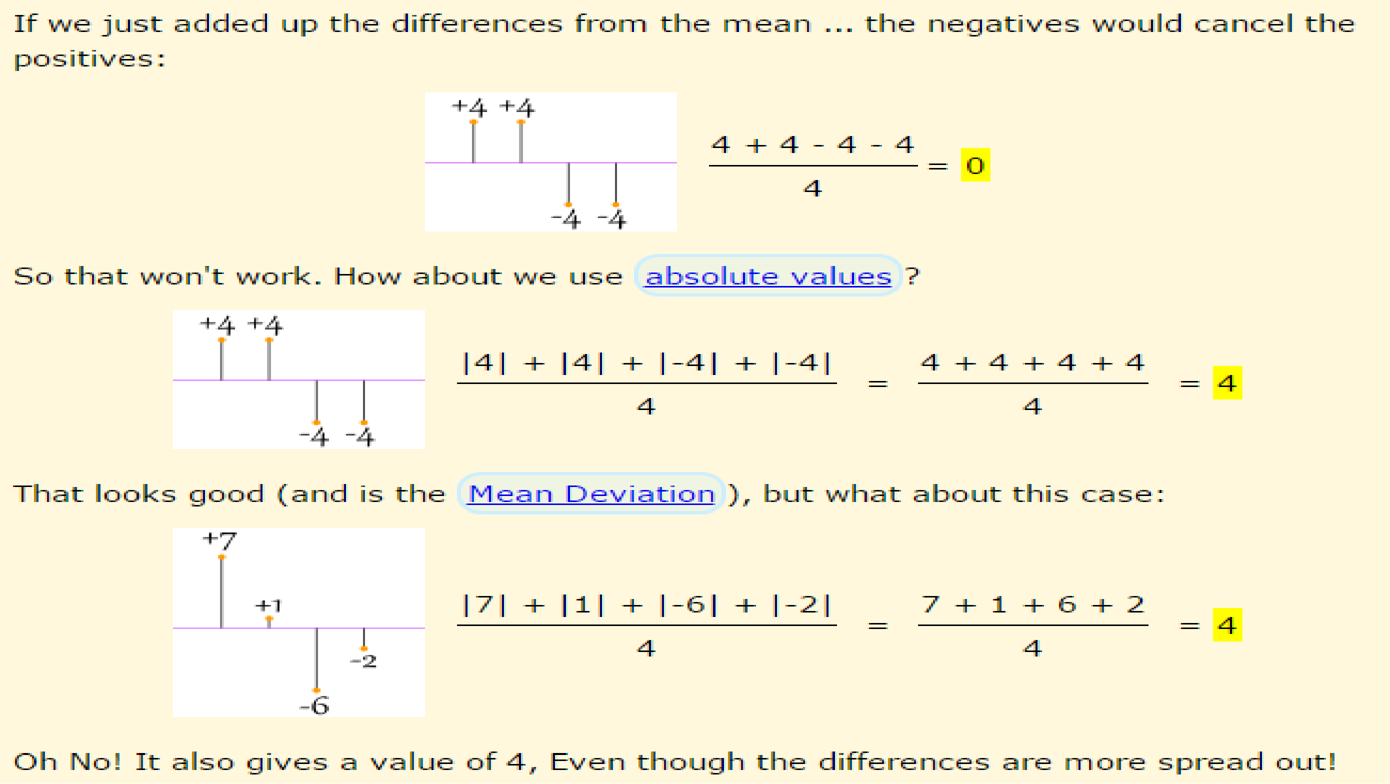
- When you have "N" data values that are:
  - The Population: divide by N when calculating Variance (like we did)
  - A Sample: divide by N-1 when calculating Variance

SD & Variance

- All other calculations stay the same, including how we calculated the mean.
- Example: if our 5 dogs were just a sample of a bigger population of dogs, we would divide by 4 instead of 5 like this:
- Sample Variance = 108,520 / 4 = 27,130
- Sample Standard Deviation = √27,130 = 164 (to the nearest mm)
- Think of it as a "correction" when your data is only a sample.

SD & Variance

- Why Square the difference?



If we just added up the differences from the mean ... the negatives would cancel the positives:

$$\frac{4 + 4 - 4 - 4}{4} = \boxed{0}$$

So that won't work. How about we use absolute values?

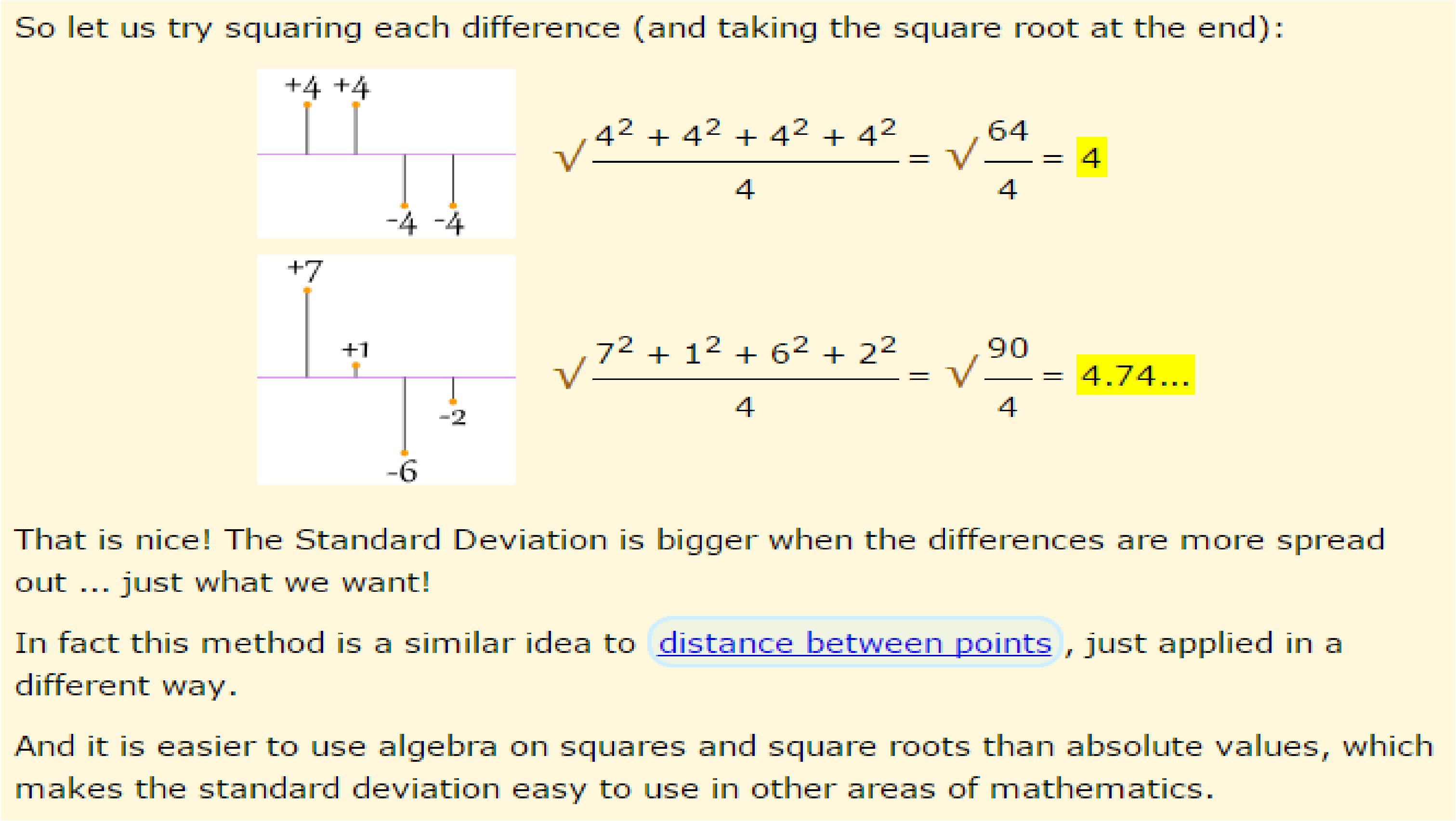
$$\frac{|4| + |4| + |-4| + |-4|}{4} = \frac{4 + 4 + 4 + 4}{4} = \boxed{4}$$

That looks good (and is the Mean Deviation), but what about this case:

$$\frac{|7| + |1| + |-6| + |-2|}{4} = \frac{7 + 1 + 6 + 2}{4} = \boxed{4}$$

Oh No! It also gives a value of 4, Even though the differences are more spread out!

SD & Variance



So let us try squaring each difference (and taking the square root at the end):

$$\sqrt{\frac{4^2 + 4^2 + 4^2 + 4^2}{4}} = \sqrt{\frac{64}{4}} = 4$$

$$\sqrt{\frac{7^2 + 1^2 + 6^2 + 2^2}{4}} = \sqrt{\frac{90}{4}} = 4.74...$$

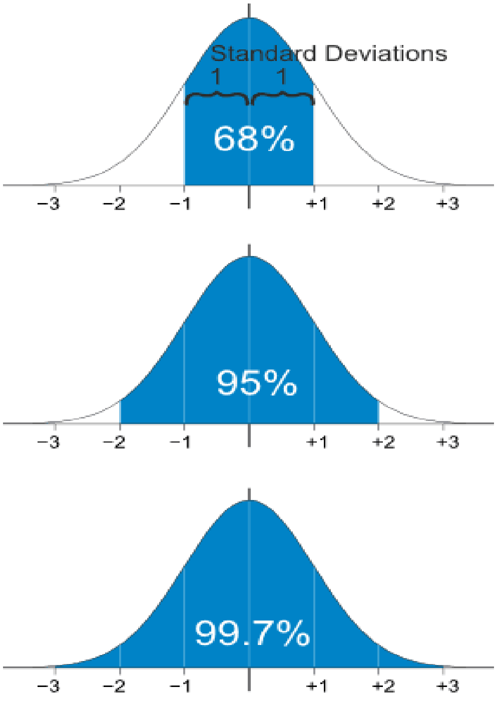
That is nice! The Standard Deviation is bigger when the differences are more spread out ... just what we want!

In fact this method is a similar idea to distance between points, just applied in a different way.

And it is easier to use algebra on squares and square roots than absolute values, which makes the standard deviation easy to use in other areas of mathematics.

Standard Deviation

- The Standard Deviation is a measure of how spread out numbers are
- When we calculate the standard deviation we find that (generally):



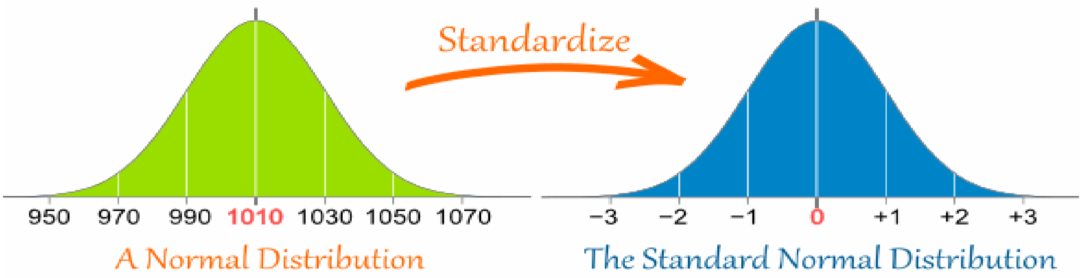
68% of values are within
1 standard deviation of the mean

95% of values are within
2 standard deviations of the mean

99.7% of values are within
3 standard deviations of the mean

Standard Scores

- The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score". Get used to those words!
- So to convert a value to a Standard Score ("z-score"):
- first subtract the mean,
- then divide by the Standard Deviation
- And doing that is called "Standardizing":
- We can take any Normal Distribution and convert it to The Standard Normal Distribution.



Example Travel Time

- A survey of daily travel time had these results (in minutes):

26, 33, 65, 28, 34, 55, 25, 44, 50, 36, 26, 37, 43, 62, 35, 38, 45, 32, 28, 34

- The Mean is 38.8 minutes, and the Standard Deviation is 11.4 minutes (you can copy and paste the values into the Standard Deviation Calculator if you want).

Example Travel Time

- Convert the values to z-scores ("standard scores").

To convert 26:

- first subtract the mean: 26 - 38.8 = -12.8,
- then divide by the Standard Deviation: -12.8/11.4 = -1.12
- So 26 is -1.12 Standard Deviations from the Mean

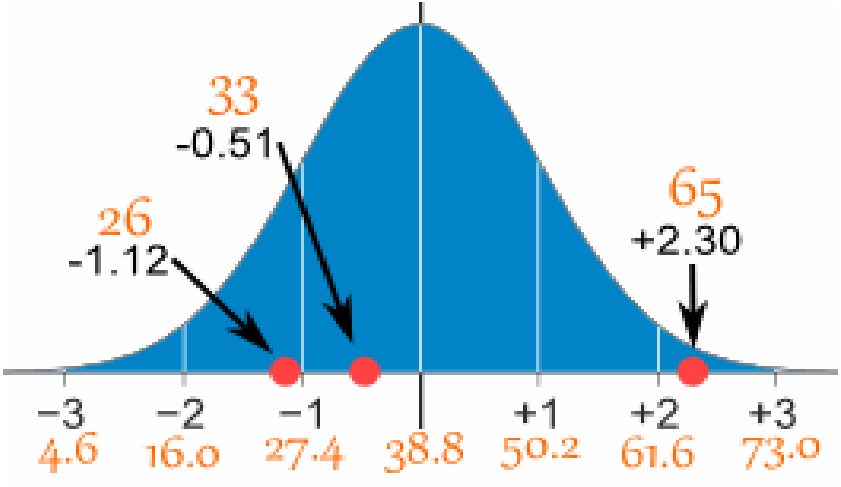
Standard Scores

- Here are the first three conversions

Original Value	Calculation	Standard Score (z-score)
26	(26-38.8) / 11.4 =	**-1.12**
33	(33-38.8) / 11.4 =	**-0.51**
65	(65-38.8) / 11.4 =	**+2.30**
...	...	...

Standard Scores

- And here they are graphically:



You can calculate the rest of the z-scores yourself!

Why Standardize?

- It can help us make decisions about our data.
-
- **Example:-**
- Here are the students results (out of 60 points):
- 20, 15, 26, 32, 18, 28, 35, 14, 26, 22, 17
- Most students didn't even get 30 out of 60, and most will fail.
- The test must have been really hard, so the Prof decides to Standardize all the scores and only fail people 1 standard deviation below the mean.

Why Standardize

- The Mean is 23, and the Standard Deviation is 6.6, and the Standard Scores are:

  -0.45, -1.21, 0.45, 1.36, -0.76, 0.76, 1.82, -1.36, 0.45, -0.15, -0.91

- Only 2 students will fail (the ones who scored 15 and 14 on the test)

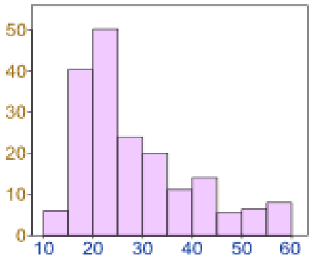
Business Analytics

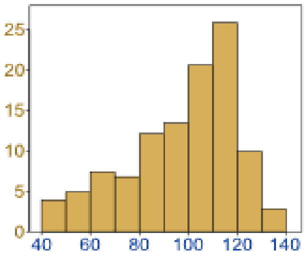
Normal Distribution

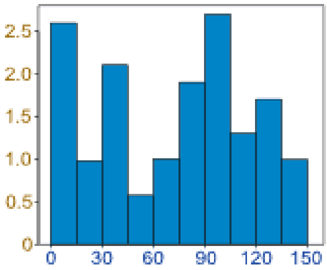
Proschool

An IMS Initiative

- Data can be "distributed" (spread out) in different ways.
- It can be spread out more on the left or to the right or it can be all jumbled up as shown below.

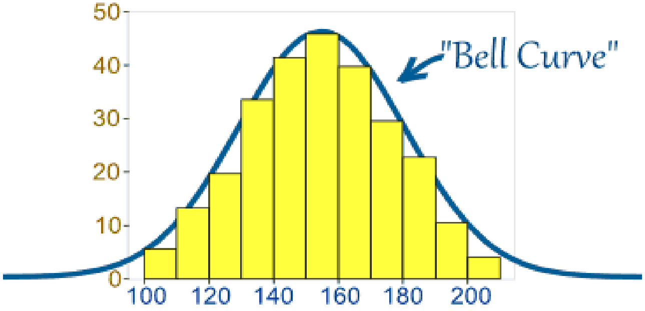






Normal Distribution

- But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



A Normal Distribution

- The "Bell Curve" is a Normal Distribution. And the yellow histogram shows some data that follows it closely, but not perfectly (which is usual).

Normal Distribution

Many things closely follow a Normal Distribution:

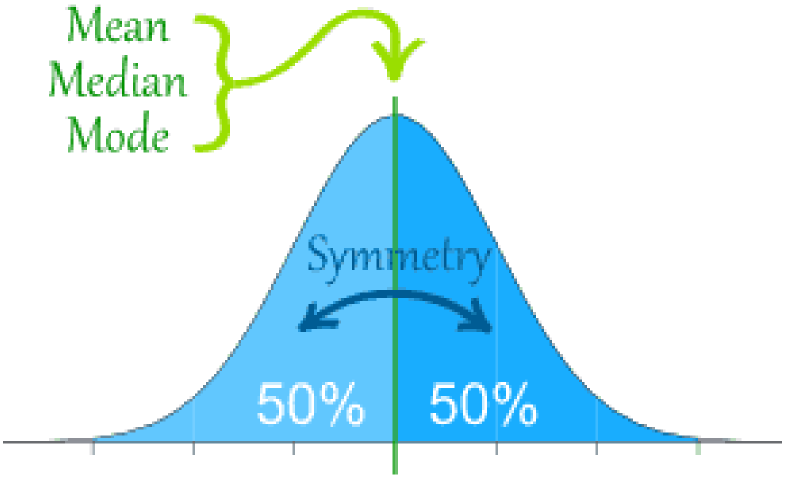
- Heights of people
- Size of things produced by machines
- Errors in measurements
- Blood pressure
- Marks on a test

Normal Distribution

- A normal distribution is a very important statistical data distribution pattern occurring in many natural phenomena, such as height, blood pressure, lengths of objects produced by machines, etc.
- This is important to understand because if a distribution is normal, there are certain qualities that are consistent and help in quickly understanding the scores within the distribution

Normal Distribution

- The Normal Distribution has:
- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean



Business Analytics

Central Limit Theorem

Proschool

An IMS Initiative

Central Limit Theorem

- In a world full of data that seldom follows nice theoretical distributions, the Central Limit Theorem is a beacon of light.

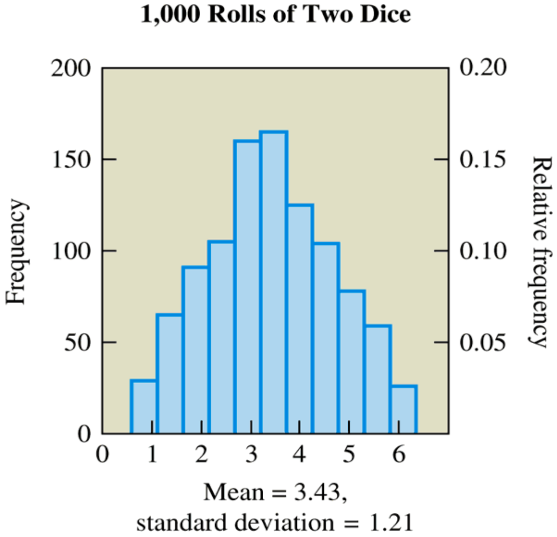
- Often referred to as the cornerstone of statistics, it is an important concept to understand when performing any type of data analysis.

- The Central Limit Theorem states that given a sufficiently large sample size from a population with finite mean and variance, the sampling distribution of mean approaches the normal distribution.

- All this is saying is that as you take more samples, especially large ones, your graph of the sample means will look more like a normal distribution.

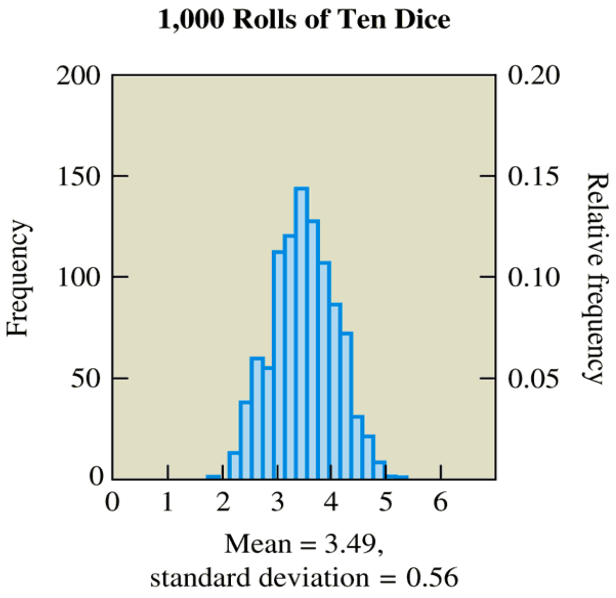
Visualizing the Central Limit Theorem Using Dice

- Now suppose we roll two dice 1,000 times and record the mean of the two numbers that appear on each roll. To find the mean for a single roll, we add the two numbers and divide by 2.



Visualizing the Central Limit Theorem Using Dice

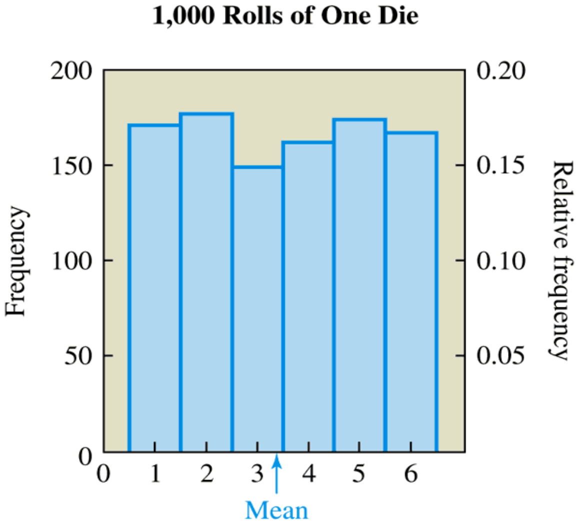
- Now we will further increase the number of dice to ten on each of 1,000 rolls.



(c)

Visualizing the Central Limit Theorem Using Dice

- Suppose we roll one die 1,000 times and record the outcome of each roll, which can be the number 1, 2, 3, 4, 5, or 6.



Mean = 3.41, standard deviation = 1.73

Visualizing the Central Limit Theorem Using Dice

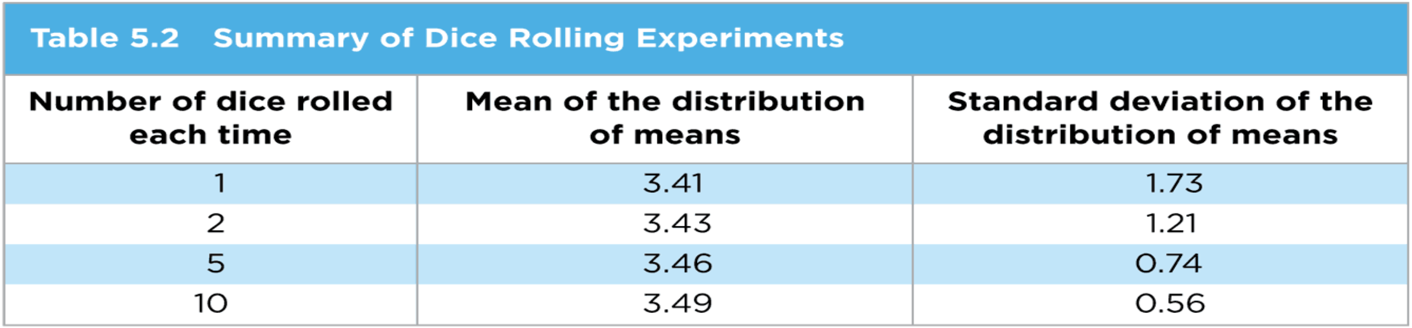


Table 5.2 Summary of Dice Rolling Experiments

Number of dice rolled each time	Mean of the distribution of means	Standard deviation of the distribution of means
1	3.41	1.73
2	3.43	1.21
5	3.46	0.74
10	3.49	0.56

What do you notice about the shape of the distribution as the sample size increases?

It approximates a normal distribution

What do you notice about the mean of the distribution of sample means as the sample size increases in comparison to the true mean of the population (3.5)?

It approaches the population mean

Visualizing the Central Limit Theorem Using Dice

What do you notice about the standard deviation of the distribution of means as the sample size increases?

It gets smaller representing a lower variation

The Central Limit Theorem

- The distribution of means will be approximately a normal distribution for larger sample sizes
- The mean of the distribution of means approaches the population mean, μ, for large sample sizes
- The standard deviation of the distribution of means approaches for large sample sizes, where σ is the standard deviation of the population and n is the sample size

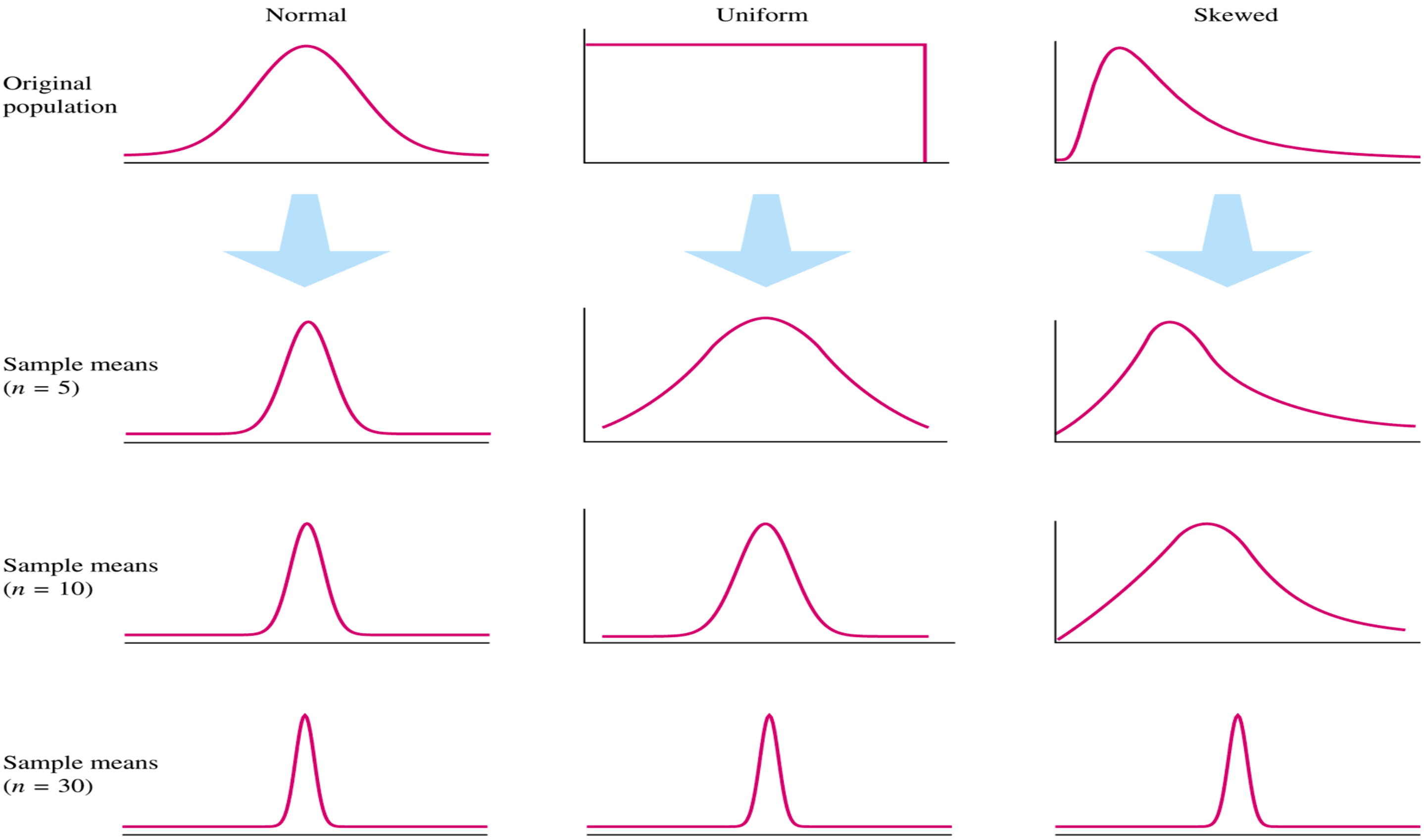
The Central Limit Theorem Notes

- For practical purposes, the distribution of means will be nearly normal if the sample size is larger than 30

- If the original population is normally distributed, then the sample means will remain normally distributed for any sample size n, and it will become narrower

- The original variable can have any distribution, it does not have to be a normal distribution

Shapes of Distributions as Sample Size Increases



Example - Predicting Test Scores

- You are a middle school principal and your 100 eighth-graders are about to take a national standardized test. The test is designed so that the mean score is μ = 400 with a standard deviation of σ = 70. Assume the scores are normally distributed.
- What is the likelihood that one of your eighth-graders, selected at random, will score below 375 on the exam?
- Since the distribution is normal, we can just use z-scores to determine the percentage for one student

$$z = \frac{375 - 400}{70} = -0.36$$

Example - Predicting Test Scores

- According to the table, a z-score of -0.36 corresponds to about 36% which means that about 36% of all students can be expected to score below 375, thus there is a 36% chance that a randomly selected student will score below 375

Example - Predicting Test Scores

- Your performance as a principal depends on how well your entire group of eighth-graders scores on the exam. What is the likelihood that your group of 100 eighth-graders will have Mean score below 375?

- According to the C.L.T. if we take random groups of say 100 students and study their means, then the means distribution will approach normal. Hence, the μ = 400 and its standard deviation is σ/√n = 70/√100 = 70/10 = 7 according to the C.L.T . Therefore, the z-score for a mean of 375 with a standard deviation of 7 is:

$$z = \frac{375 - 400}{7} = -3.57$$

Example - Predicting Test Scores

- The percent that corresponds to a z-score of -3.57 is less than .01%, which means that fewer than .01% of all samples of 100 students will have a mean score of 375.
- In other words, 1 in 5000 samples of 100 students will have a mean score of 375.

Thank You…