University of Alberta

STAT151

Introduction to Applied Statistics I

Fall 2017

Term Test 1

Prof: Gregory Wagner

Exam Guide

Chapter 1
Stats Starts Here

What is/are Statistics?

Statistics: A way of reasoning along with a collection of tools and methods designed to help further our understanding of how the world deals with data

What are Data?

Statistical methods are used in areas such as:

- Personal life
- Biology
- Business
- Psychology

The process of making a data driven decision goes as follows

1) State the data
2) How to collect the data
3) Collect data, draw a SRS (simple random sample) from population
4) Summarize data
5) Use statistical tools to make an inference

Parameter: A number that describes certain characteristics of a population
Statistic: A number that describes certain characteristics of a sample

Y - (a Y-bar) = sample mean, can also be symbolized as an X-bar

μ = population mean

We can usually make an inference about the population from the sample statistics since we cannot test every member of a population as it would be too difficult and time consuming, which is why we take random sample and test those instead.

EX. 1) I want to estimate the proportion of male students for this Stat 151 Lecture by randomly selecting 10 students in this class.

Population of Interest: Every Stat 151 student in that class

Sample: The 10 randomly selected Stat students

Parameter: Proportion of male students in the lecture

Statistic: Proportion of male students in this sample

EX. 2) An investigator wants to estimate the average height of Canadian females by measuring the height of 1000 randomly picked Canadian females.

Population of Interest: All Canadian females

Sample: 1000 randomly selected Canadian females

Parameter: Average height of Canadian females

Statistic: Average height of 1000 females

Variables

- Not all data is represented by numbers, these are called *categorical variables*, however these can be coded into numerical/qualitative values (eg. 1 = male, 2 = female)
- Experimental units include animals, plants, and inanimate subjects

Categorical variables

Nominal = Categories have no order (eg. hair colour, gender)

Ordinal = Categories have order

Example:

a) Gender (M or F) - Nominal

b) Hair color (blonde, white, black, red, etc…) - Nominal

c) Nationality (American, Canadian, Chinese, French, German, Japanese, etc…) - Nominal

d) Letter grade (A+, A, A-, B+, B, B-, C+, C, C-, D+, D, F) - Ordinal

e) Car manufacturer (Dodge, Ford, Honda, Others) - Nominal

Quantitative variables

Continuous: Can take on any number within an interval on the number line (e.g. systolic blood pressure or level of Calcium in the blood)

Discrete: Can only be distinct values (e.g. number of siblings bc can't a large number such as 100 siblings)

Chapter 2
Displaying and Describing Categorical Data

One Categorical Variable

We can very rarely make conclusions about a variable just by looking at raw data, this is why we summarize raw data into a more manageable form in order to draw a conclusion.

- For categorical variables we can use bar charts or pie charts
- For numerical variables we can use dot plots, stem plots, time plots, box plots, scatter plots, or histograms

Different types of graphs are needed for different types of data

- After sampling the data we have to summarize it and answers two questions

1) What values have been observed? (Ex: Gender, female or male)
2) How often did every value occur?

To distribute categorical values we can put it into a table which includes the following information

- Each possible category
- Frequency (or number) of individuals in which fall under that category
- Relative frequency/percentage of individuals who fall under that category

Relative frequency: is the % of the frequency that the category appears in the data set

= (Frequency ÷ number of observations)

% = relative frequency x 100%

Steps to create a frequency distribution table

1) Add all the different categories in
2) Calculate the frequency (# of observations in that category) and add it in
3) Summarize the results in a table (frequency distribution table)

A relative frequency table is similar, however instead of counts we use percentages for each category

Example: Favorite candy flavor for this STAT 151 class

Frequency table:

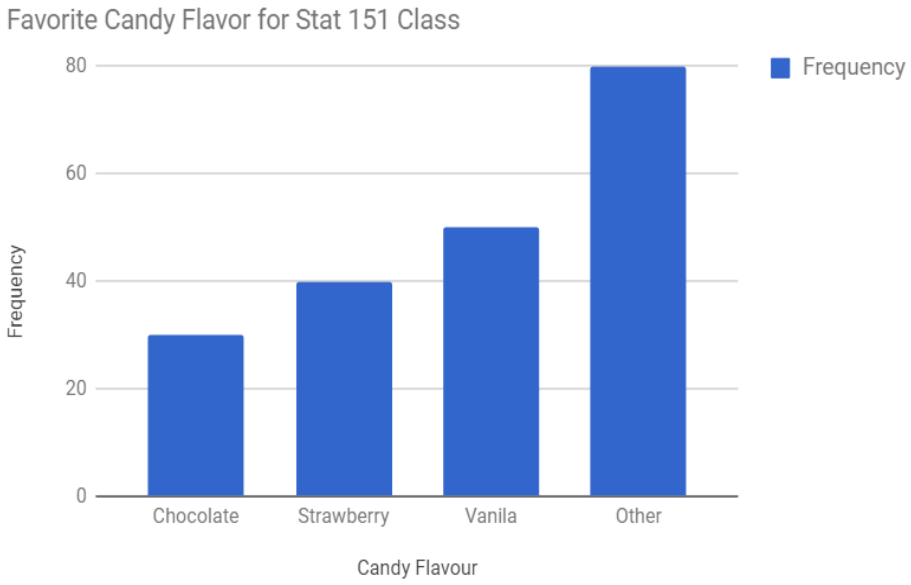
Category	Frequency
Chocolate	30
Strawberry	40
Vanilla	50
Other	80
Total	200

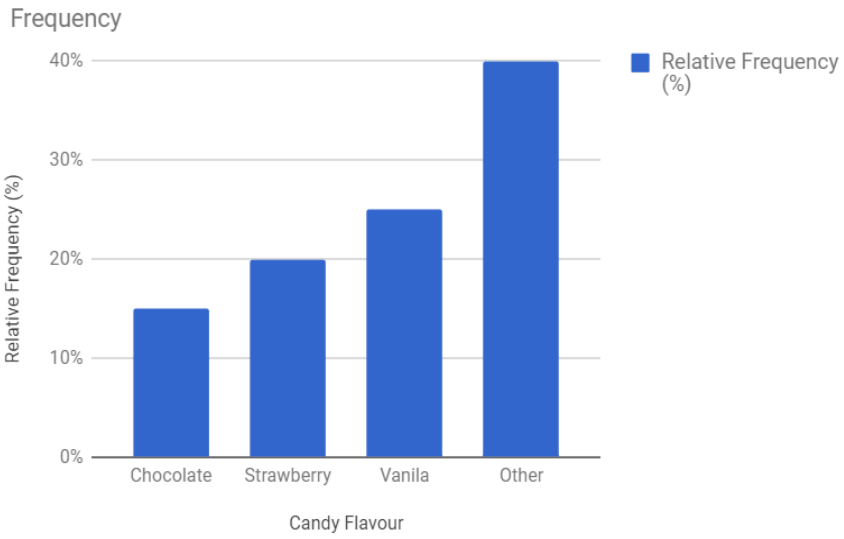
Relative Frequency table:

Category	Frequency
Chocolate	(30÷200) = 15%
Strawberry	(40÷200) = 20%
Vanilla	(50÷200) = 25%
Other	(80÷200) = 40%
Total	1 = 100%

**** Total should always equal 1 or 100%****

After the data is summarized in a frequency distribution table it can displayed on a bar chart or a pie chart

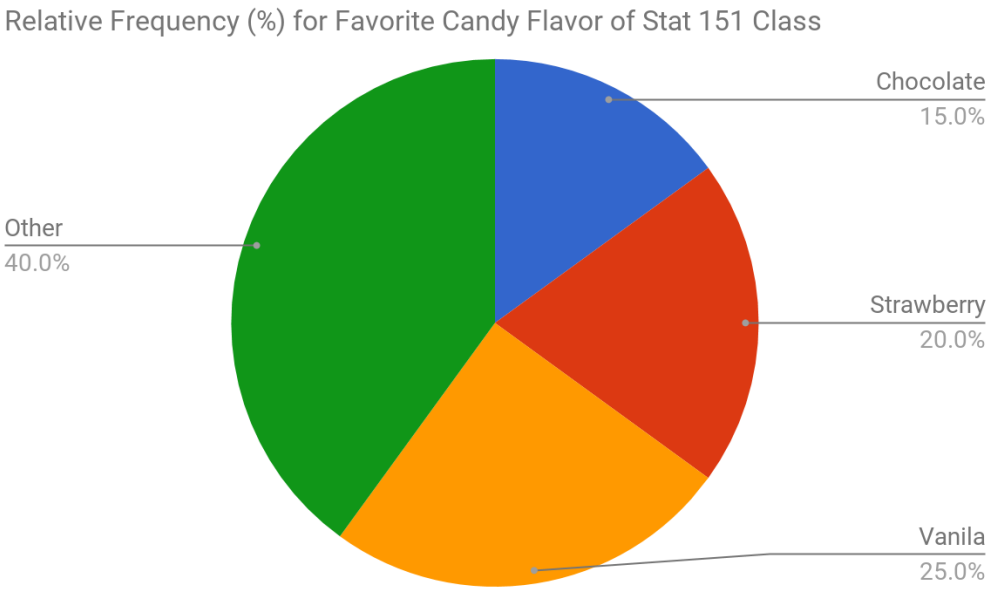




As we can see, the same pattern is shown for the graphs of relative frequency and frequency.

We can also create a pie chart. The steps for creating one goes as follows

1) Draw a circle
2) Calculate slice size (category relative frequency x 360°)



Two categorical variables and the relationship

Contingency table

- Allows us to explore the relationships between two categorical variables
- Also shows how a individuals are distributed in each variable contingent (dependent on) the value of the other variable (Ex. we can examine the gender and see whether or not a person likes salt and vinegar chips)
- The margins for the table (bottom and the right) give frequency distributions for each of the variables
- Every frequency distribution is also called a marginal distribution of its respective variable
- Conditional distribution is when it shows the distribution of one variable when it only meets some condition of the other variable
- The variables are deemed independent when every value in a contingency table is the

Example 1:

A study was conducted to see if smoking had any effect on heart disease

Smoker

Heart

Disease

	Yes	No	Total
Yes	23	15	38
No	69	259	328
Total	92	274	366

Marginal distribution of Heart Disease

Heart

Disease

	Total	% of column
Yes	38	(38/366) 10.4%
No	328	(328/366) 89.6%
Total	366	100%

Marginal distribution of Smoker

Smoker

	Yes	No	Total
Total	92	274	366
% of Row	25.1% (92/366)	74.9% (274/366)	100%

DON'T CONFUSE SIMILAR SOUNDING PERCENTAGES

- % of people with heart disease and smoke = 23/366
- % of smokers with heart disease = 23/92
- % of people with heart disease who smoked 23/38

Independent variables: Independent variables are ones in which the conditional distribution is the same for each category, they are not dependent of each other. However, this is extremely rare

Chapter 3

Displaying and Summarizing Quantitative Data

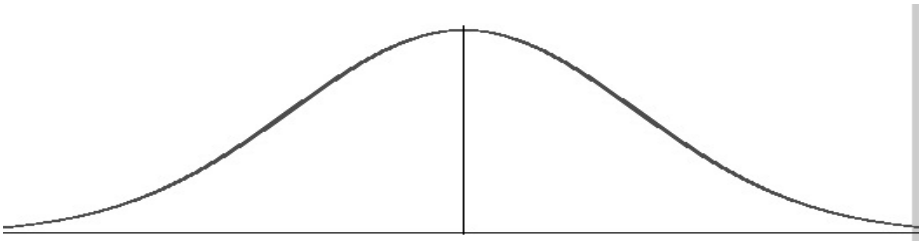
Displaying Quantitative Variables With Graphs

Graphs for numerical data: Since numerical data takes on many forms, different types of graphs need to be introduced in order to represent these quantitative values in such a way that the distribution of data is apparent.

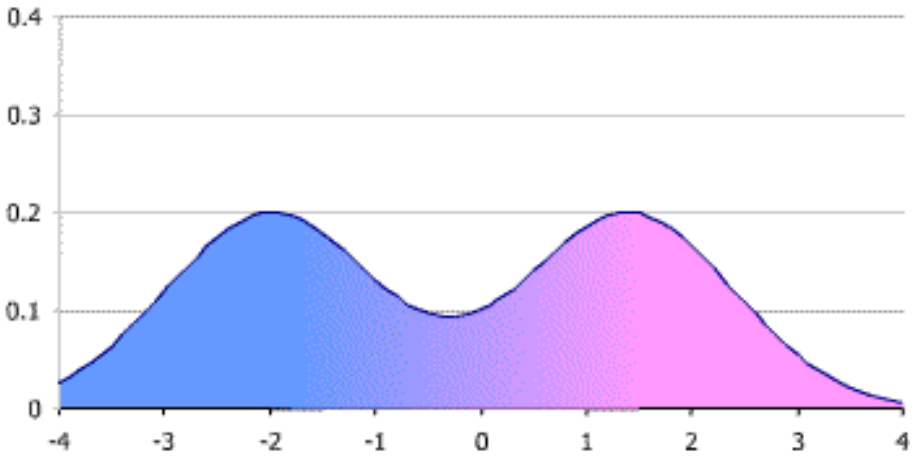
Describing a distribution of a plot (3 factors):

1) Shapes: one characterization of general shapes pertains to the number of humps, or modes

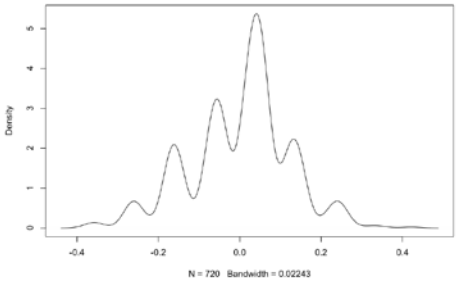
a) Unimodal = a single hump



b) Bimodal = two peaks/humps

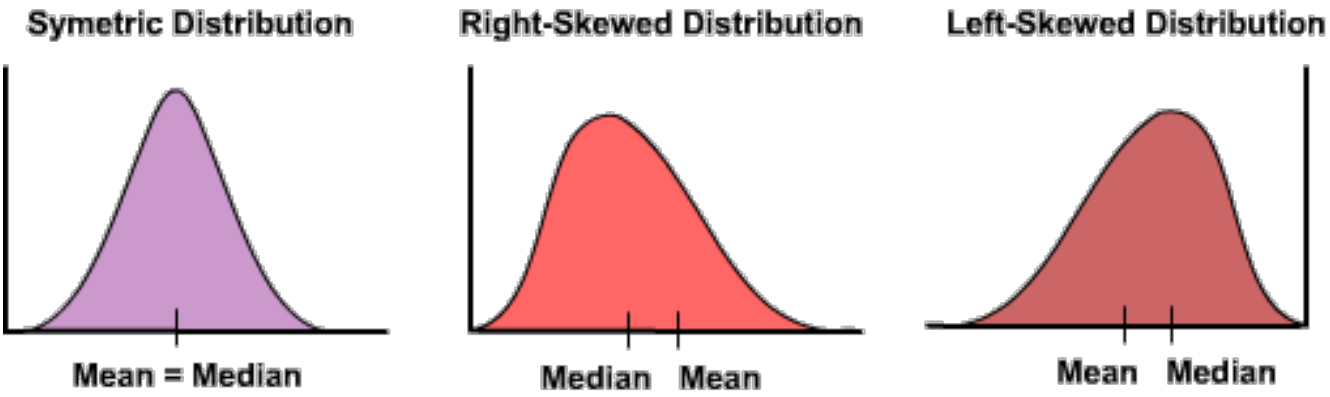


c) A multimodal graph/distribution is very rare since we would usually not place so much data onto one graph

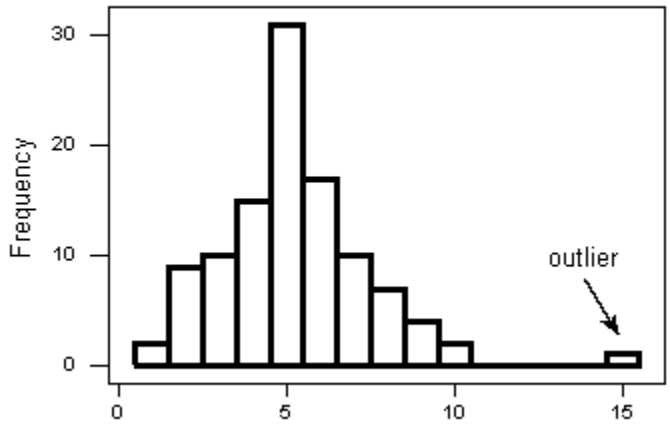


d) Non symmetric graphs (graphs in where if you draw a vertical line in the middle and the left side DOESN'T mirror the right side, or vise versa) are skewed either to the left or the right

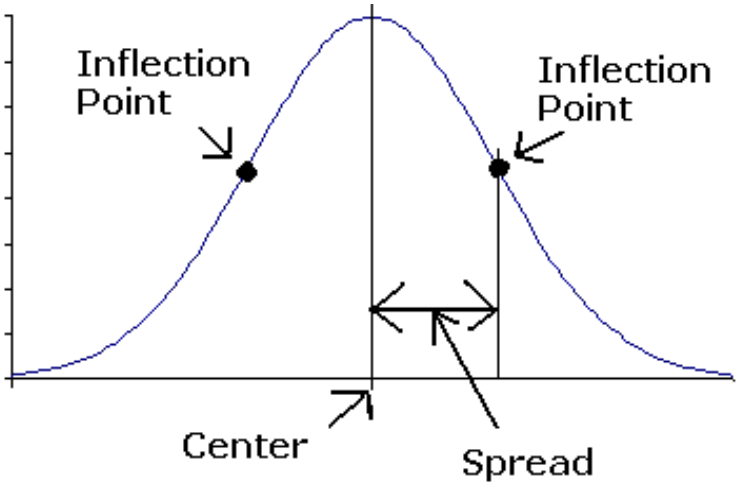
Skewed to the left looks as if there's a tail trailing on the left side of the histogram and skewed to the right looks as if there's a tail trailing on the right side of the histogram



e) Values that fall out of the overall pattern is deemed an outlier



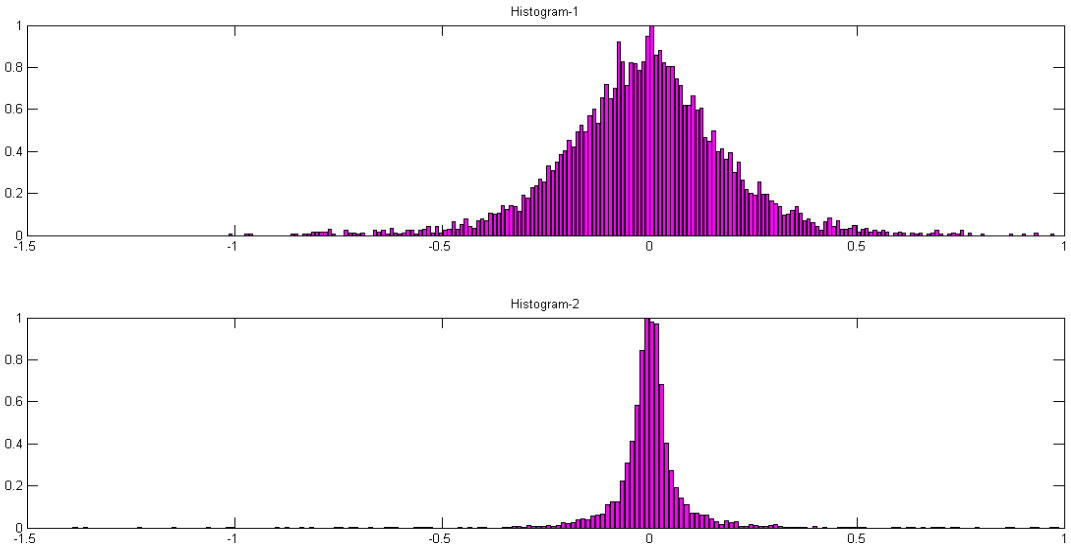
2) Center - this is the value that typically splits the data in half or a typical range of values on the center of the graph (mean, median, mode). Refer to the center on the image below



3) Spread - this is the range of values, are the values close or far from the center (concentration)?

Histogram #1 in the image below would be considered a large spread

Histogram #2 in the image below would be considered a small spread



Dotplots

These plots are ones that show each individual observation and they are usually for small data set.

To construct one we must:

1) Draw a horizontal (or vertical) line
2) Label the line with the name of the variable and mark regular values of the variable on it
3) For each observation, place a dot above (or next to) its value on the number line

Stem-and-Leaf-Displays/ Stemplot

This type of graph offers us another option to display small sets of data for quantitative values, each observed number is broken down into two pieces called the stem and leaf

Steps for Making a Stem Plot:

1) Divide each numerical value into the stem and leaf, the leading number is the stem , and the other number following it is the leaf. For example for the number "60", 6 would be the stem and the number 0 would be the leaf,

NOTE: use stems to label the bins (the equal width intervals), also only use one number for the leaf, the other numbers will be the stem. IF it's a decimal number then round it

2) Order the data, so order the leaves from highest to lowest in each stem
3) List the stems in a column (the smallest will be first) and place a vertical line directly right of this column
4) For each measurement, record the leaf value in the same row with its corresponding stem
5) Provide to the stem and leaf coding

Stemplot of Data Set

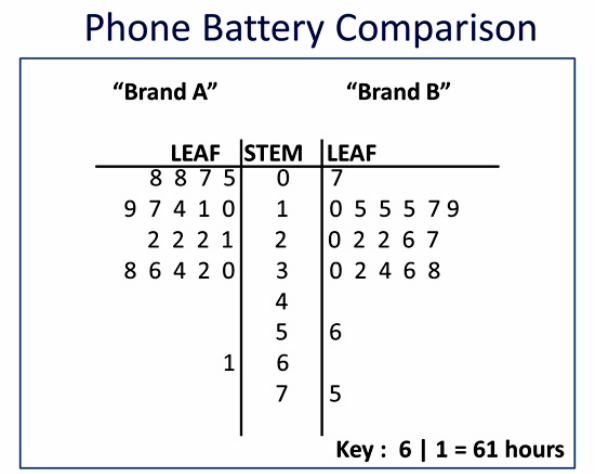
0	4 6
1	2 4 8
2	
3	3 4 4 5 5 7 8
4	2 2 5
5	0 1 8
6	8
7	2

Key: 1|0 = 10

The number values in this set would be 4, 6, 12, 14, 18, 33, 34, 34 ,35, 35, 37, 38, 42, 42, 45, 50, 51, 58, 68, 72.

Although there were no values for the stem 2, we still have to include it in our stem plot since it would disturb the balance of it without the number.

We can also have back-to-back stemplots when we want to compare data



Histograms:

A histogram is the most common graph for describing numerical data because it helps visualize the distribution of the underlying variable very well, it also works for large data sets

A histogram shows how often for measurements falls into a particular equal width interval (bin).

Bin: This is an interval with the equal sized width and the boundaries are if possible, whole numbers of tenth

Informal Rule: 6-10 bins for smaller data sets and 10-25 bins for larger data sets

Steps to plotting a histogram:

1) Decide on the width of the bin
2) Make a frequency table using the method of left inclusion

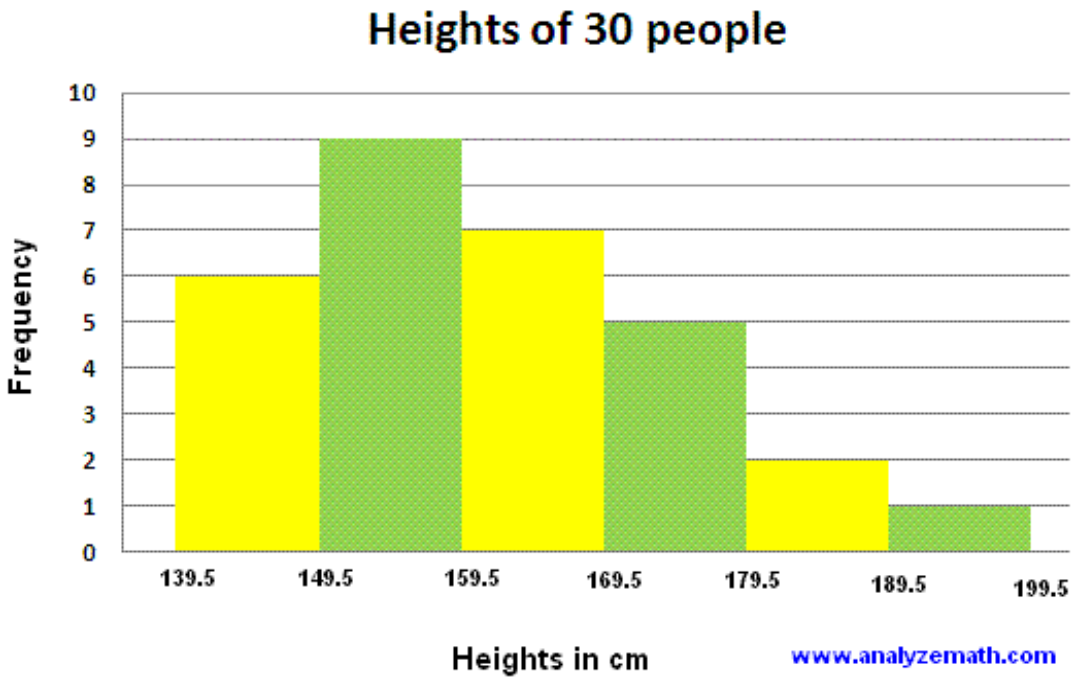
Left inclusion: For the intervals [27.5, 28.0) and [28.0, 28.5), we would put the value "28.0" in the interval [28.0, 28.5), since you have to note the circle and square brackets. The circle brackets (in the first stated interval) denotes that the values go up to 28.0 but are not inclusive of that value

Example: Construct a histogram for prices of walking shoes with bin width of 10.

Prices of walking shoes in $: 40 60 65 65 68 68 70 70 70 70 70 74 75 75 85 90 95

Class intervals	Frequency	Relative Frequency	Percentage
[40;50)	1	1/17	5.88%
[50;60)	0	0/17	0%
[60;70)	5	5/17	29.4%
[70,80)	8	8/17	47.06%
[80;90)	1	1/17	5.88%
[90;100)	2	2/17	11.76%
Total	17	1	100%

EXAMPLE HISTOGRAM BELOW, NOT A HISTOGRAM OF THE VALUES IN THE TABLE ABOVE



Describing Quantitative Variables with Numbers

Numerical summaries: When you have large data sets you should try to summarize # so you get a few numbers that are important from a large quantity of data

*y = the variable for which we have sample data

*n = the number of of observations of the variable y

Example 1A: we have n=4, so four observations on phone battery life

Y1 = 5.9 hours

Y2 = 7.3 hours

Y3 = 6.6 hours

Y4 = 5.7 hours

The Greek letter Σ is used for summation

$$\sum_{i=1}^{4} y_i = y1 + y2 + ...+yn$$

For our example we would add up all the numbers (5.9 + 7.3 + 6.6 + 5.7)

The three most common measures for center of a data set is:

1) Mean (average)
2) Median (middle value)
3) Mode (most frequent value)

To find mean (arithmetic average) we would use the formula (sum of all observations/# of all observations), for our first example it would be “25.5/4” which is equal to 6.375 hours.

Example 1B: If you had a fifth observation for battery lifetime. What is the battery lifetime you will need in your fifth observation in order to have an average battery lifetime of at least 6.5 hours?

Answer: We would need to manipulate this formula
We want y for the fifth observation and the average battery life needs to come up to at least 6.5 hours so the formula would be as follows

$(y1+y2+y3+y4+y5)/5$

$= (25.5+y5)/5 = \geq 6.5$

$= y5 = 6.5x5 - 25.5 = 7$

7 would be the value of y5

Calculating mean from a frequency distribution

If in a class of 20 Chemistry students and their marks were as follows

Grades	Frequency
50%	18
52% (you)	1
100%	1

Mean ($\bar{y}$) = ((18x50%) + (1x52%) + (1x100%)) / 18 + 1 + 1 = 52.6%

52.6% would be the overall average which doesn't exactly reflect the class since it would entail that we are below average even though there are 18 people with marks lower than yours and only one person with a mark higher than yours. In conclusion, if a set of data has an outlier, this can draw the mean towards it this results in not describing the true center of all the values.

NOTE 1: The mean isn't resistant to outliers

NOTE 2: If the data is skewed the mean mostly likely won't represent the center of data accurately

This is why we need an alternative value to measure the center of distribution, this is where **median** comes into play.

The median divides the ordered sample in 2 sets of the same size, one half of the data lies above M and one half lies below M

To find a median we must

1) Order the data set (n observations) from smallest to largest
2) Then if n is odd, it'll be the single middle value, if n is even it's the average of the two middle values

Example 2: We have the data set "2 5 6 7 9", since n is odd the median would be 6

Example 3: We have the data set "2 5 6 7 9 11", since n is odd (six numbers) the median would be (6+7)/2 which would be 6.5

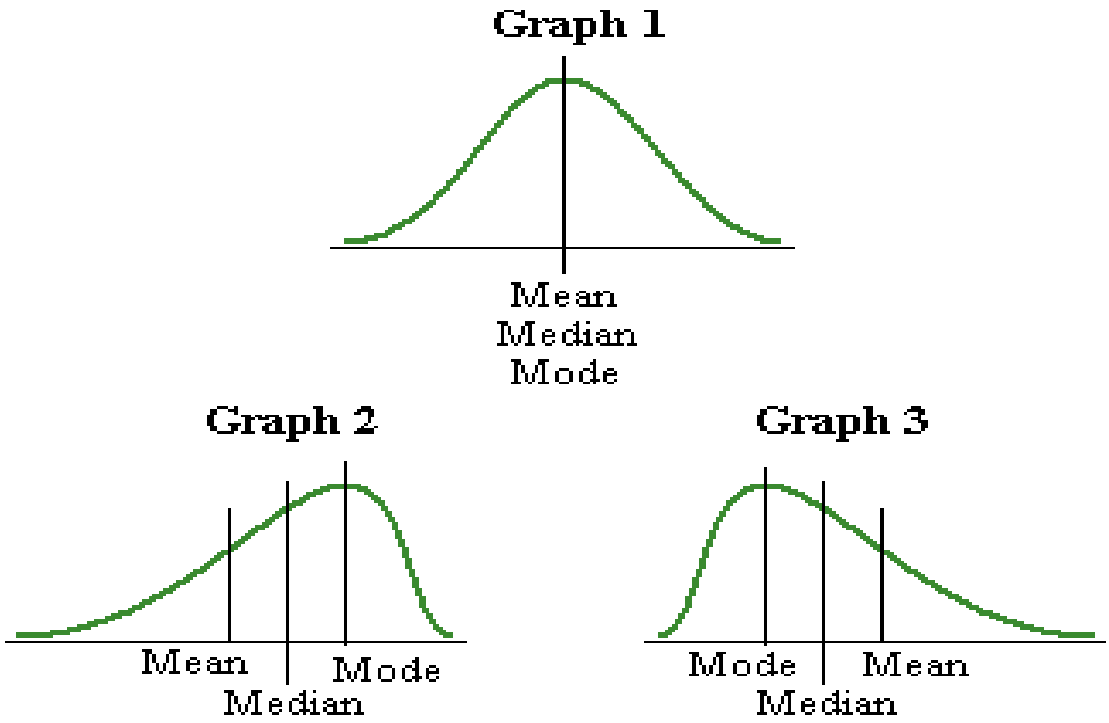
Considering the marks in the class of 20 students again, the median grade in the class is 50% ([50+50]/2)

We also have another method of calculating the center distribution which is called **mode**. Mode is the value that occurs most frequently in a data set.

Example 4: The prices of high heels are as follows "90 70 70 65 84 92 70 45", the mode would be 70 since it is the number that occurs most frequently (three times)

Example 4A: We can also have two modes such as in the data set
"50 67 84 89 69 94 76 67 50 98 64", the two numbers 50 and 67 are the modes

In a symmetrical graph, the mean, median, and mode would be the same, however if a graph is skewed to the left or right they would all be different values, as seen in the image below.



We should use median when there's an outlier/graph is skewed and mean when it's symmetrical since it would be the most accurate. The mean takes into account of the rank of each observation and not its magnitude. Mode is easy to see, however the major setback is that it may have 0 or more than one mode, whereas there will only be one mean or one median.

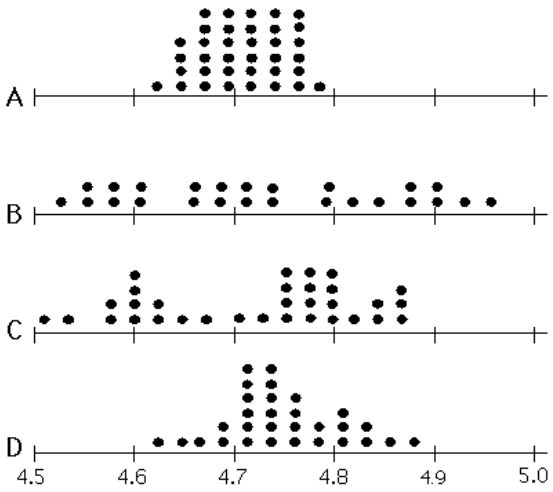
The median is the point in which the distribution is cut into two parts of the same area.

Describing the variability (spread) of a distribution

It's not enough to just provide a number that describes the center of the sample, the spread or variability in a sample is also an important characteristic of a sample. There are three common methods to measure variability

1) Range
2) Variance and standard deviation
3) IQR (interquartile range)

1) Range of the sample is the simplest numerical measure of variability that gives the difference between the largest value and smallest value in the data. Usually the greater the range, the larger the variability. However it depends on much more than the distance between two most extreme values, it also depends on every single observation that contributes to it.



2) Range = max – min

   Range for the numbers "2 5 6 7 9" is (9-2) which is 7

3) The variance and standard deviation: The value $Y_i - \bar{y}$ is the deviation of the observation Y_i from the mean $\bar{y}$. In a sample with n observations, we will get n deviations from the sample mean

Note: A specific deviation is positive is the number is greater than the mean ($\bar{y}$). It will be negative if the number is less than the mean ($\bar{y}$)

For example is the mean is 80 and the number was 90:

90 - 80 = +10% (above average)

If the number was 70

70 - 80 = -10% (below average)

If we added those deviations together (+10 and -10) it would be zero, the sum of deviations is always equal to 0.

$\Sigma(yi - \bar{y}) = 0$

Since the sum of deviations is always 0, introducing calculation techniques to the deviations to characterize the variability in the data set is necessary. This is where squaring comes into play, we need to square each deviation before summing them up.

Variance: It's denoted by s2 (s squared) and is the sum of squared deviations from the mean divided by n-1. Variance is problematic as a measure of spread since it's measured in square units.

The standard deviation the square root of the variance

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

In these formulas the letters x and y mean the same thing, so it could be the letter y instead of x in the formula

Things to know about standard variation

- Standard deviation is the most used measure of variability
- The standard deviation value lets us know how closely values are clustered around the mean and it isn't resistant to outliers
- It's also measured in the same units as the original data

Properties of the standard deviation:

- The standard deviation measures the spread about the mean and should only be used when mean is chosen as the measure for the center
- S = 0 when there is no spread
- Otherwise always S>0, it increases as the observations become more spread out

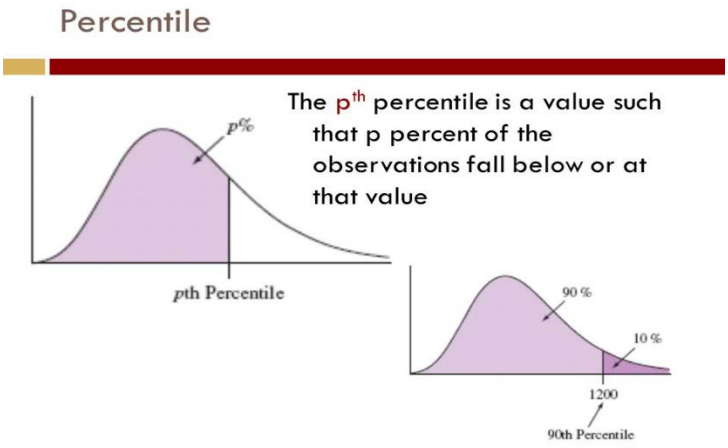
IQR (interquartile Range):

This is another measure of variability and it's the range of the middle half of the data, and similar to the median, it's resistant to outliers. It's based on quartiles in which divide the data into four different sections.

Percentiles

The pth percentile is so that p% of the measurements fall below the pth percentile and (100-p%) fall above it.

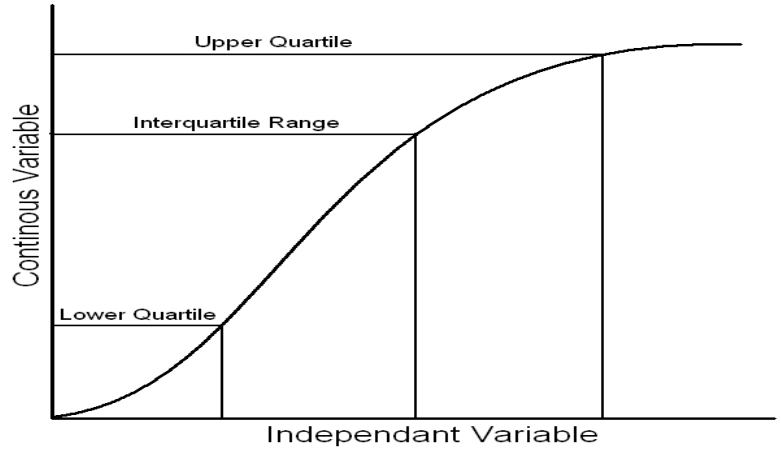
The set of measurements on the variable x must be arranged in order of magnitude. P is a number between 0 and 100 and the median is the 50th percentile.



Quartiles

The lower quartile Q1 is the 25th percentile (separates the bottom 25% of the measurements from the top 75%)

The upper quartile Q3 is the 75th percentile & separates the top 25% of measurements from the bottom 75%



IQR = Q3 - Q1, the distance between Q3 and Q1

When the IQR is small the data is closely clustered around the middle and when the IQR is large the data is further scattered from the center

Ignore the extreme values and concentrate on the middle of the data

Example: for the data set “2 5 6 7 9” exclude the median which is 6

Q1 = median between 2 and 5, which is 3.5

Q3 = median between 7 and 9 which is 8

IQR = Q3 - Q1 = (8 - 3.5 = 4.5)

IQR = 4.5

Example: for the data set “2 5 6 7 9 11” the median would be 6.5

Q1 = 5

Q3 = 9

IQR = Q3 - Q1 = (9 - 5 = 4)

IQR = 4

Chapter 4

The Five - Number Summary and Boxplots

The five - number summary of a group of measurements includes the minimum, first quartile, the median, the third quartile, and the maximum. These numbers offer an accurate summary of the distribution of quantitative variables.

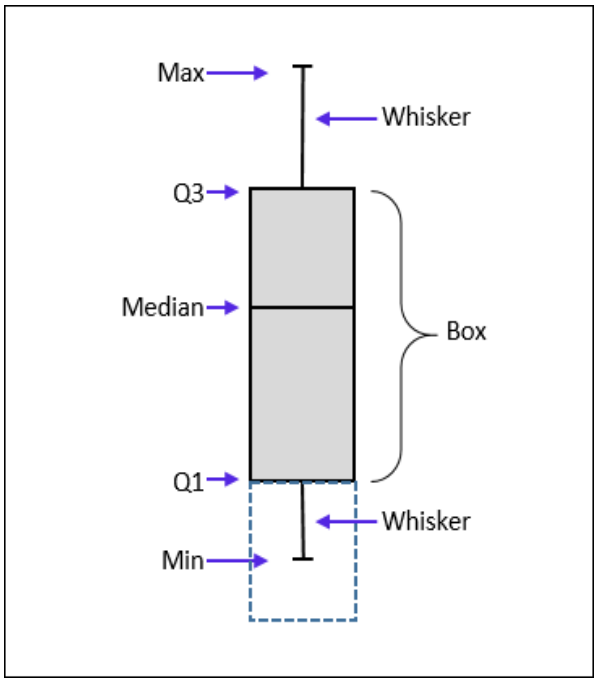
The five symbols are:

Minimum Q1 M(median) Q3 Maximum

Example 1: If we had the data set 2, 5, 6, 7, 9

M = 6 Q1 = 3.5 Q3 = 8 Min = 2 Max = 9

So the five number summary would be 2, 3.5, 6, 8, 9, the boxplot would look somewhat like this



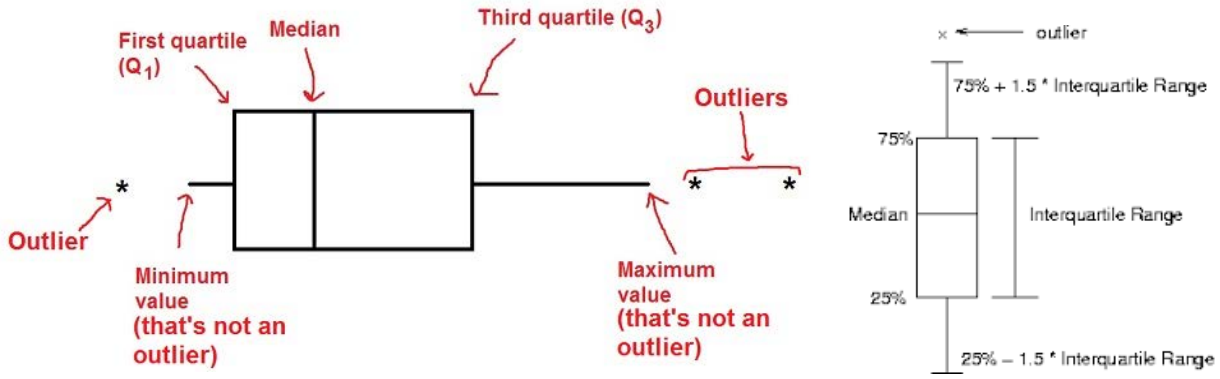
Instructions for constructing a boxplot

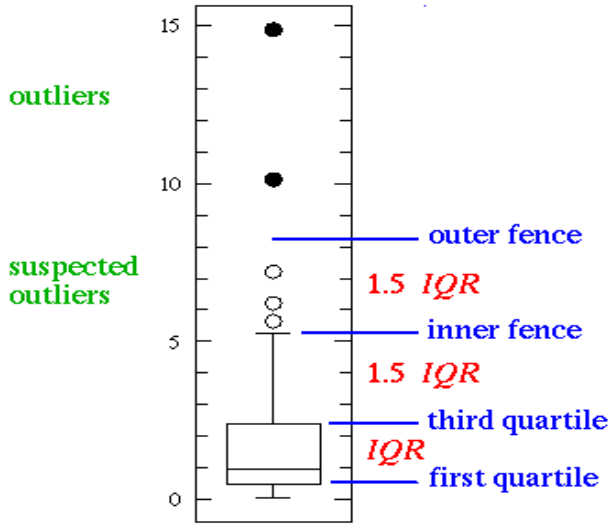
1) Draw a vertical line measurement scale which includes the whole range of the data
2) Draw a rectangular box in which the bottom lines up with the lower quartile and the top of the box lines up with the upper quartile
3) Draw a horizontal line inside the box which lines up with the median
4) Determine the "fences" for outliers. A value is an outlier if it falls more than 1.5 IQR above the third quartile or below the first quartile

- Upper fence = Q3 = (1.5 x IQR)
- Lower fence = Q1 = (1.5 x IQR)

5) Add whiskers (line segments) from each end of the box to the most extreme values, so the highest or lowest value before the outlier zones (past upper and lower fences)
6) Add the outliers by displaying them outside the fence with symbols (such as circles), if an outlier is further than 3 IQR's from the fence we call this a far outlier

**The whisker will extend to the smallest or largest value that isn't an outlier, outliers are included in boxplots however we use dots to represent them, and they aren't connected by the whiskers. A value is considered an outlier if it exceeds Q3 + (1.5 x IQR). In our example the upper outliers would be anything above 14.75 (Q3 + (1.5 x 4.5)). **



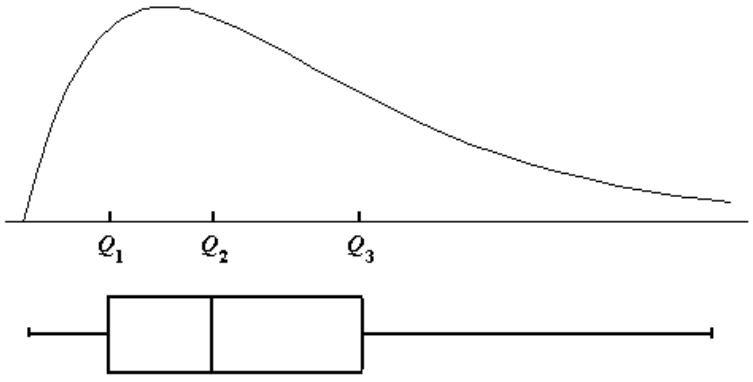


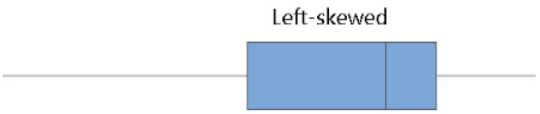
Distribution of a boxplot

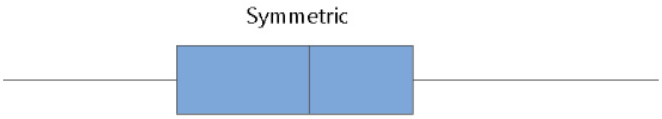
Symmetric distribution: If a box plot has the median line in the exact center of the box and the whiskers are of equal length it is symmetrical meaning 50% of the data is spread within the rectangle/ box.

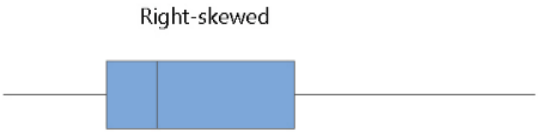
A boxplot can be skewed to the right when the median is closer to the first quartile rather than the third one.

A boxplot is skewed to the left when the median is closer to the third quartile rather than the first one

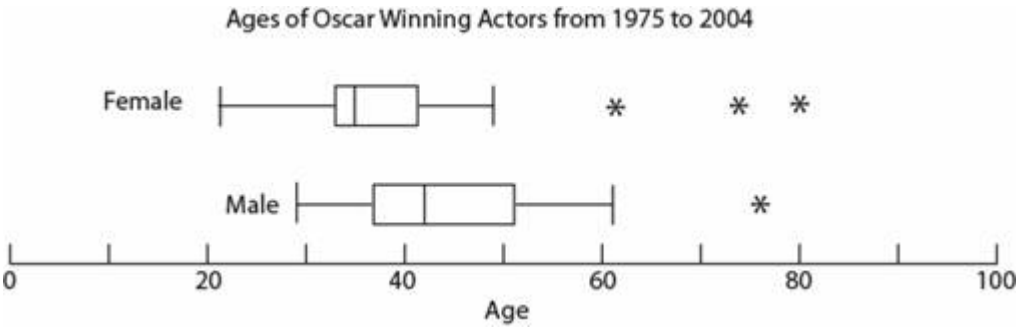








Comparative boxplot: This can also be called side by side boxplot, this is useful for comparing variables in different categories

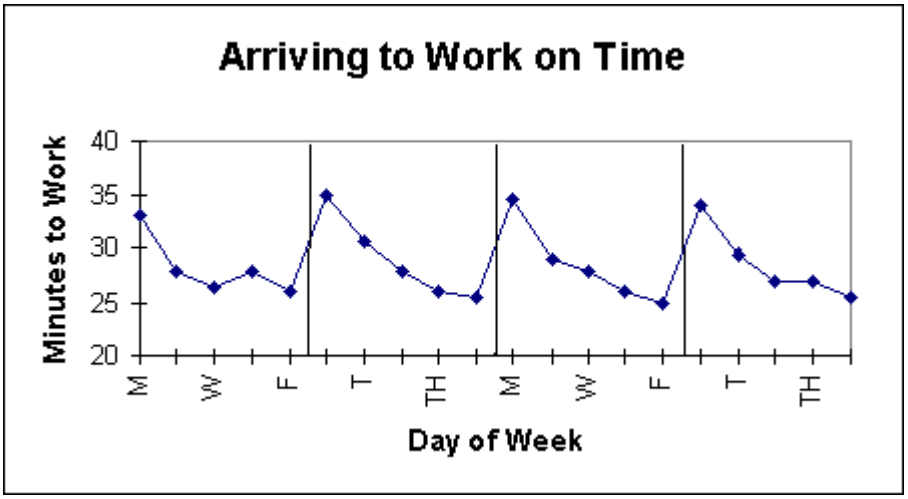


As you can see, at first it looks as the female boxplot is more concentrated, however, as we look closer it's clear that there are more outliers in that one. This makes the male boxplot more concentrated.

Time Plots

Time plots are used in data sets in which we are interested in how the data behaves over time.

Time is always on the horizontal axis



Chapter 5

The Standard Deviation as a ruler and the Normal Model

Shifting and Scaling

Shifting data:

- Adding or subtracting a constant to every data values adds or subtracts the same constant to the measures of position (percentiles, quartiles, z-Scores)
- Adding (or subtracting) a constant to each value will increase (or decrease) measures of position: center, percentiles, max or min by the same constant.
- - Its shape and spread - range, IQR, standard deviation will stay the same

Example 1: If you have a data set: y1 = 1, y2 = 2, y3 = 3, y4 = 4, y5 = 5, and you want to add a constant c = 2 to each value in this data set, how does it affect the mean, median, Q1, Q3, max, min, range, standard deviation, IQR, and its shape?

column	n	mean	variance	Std. dev.	Std. err.	median	range	min	max	Q1	Q3
y	5	3	2.5	1.5811	0.707	3	4	1	5	2	4
y+2	5	5	2.5	1.5811	0.707	5	4	3	7	4	6

The mean, median, range, min, max, and quartiles do change however the spread is unchanged

Summary:

If ynew = yoriginal + c for each observation

For measures of center or position:

- Center new = center original +c
- Position new = position original + c

For measures of Spread and Shape:

- Spread new = Spread original
- Shape new = Shape original

Rescaling Data

When we multiply or divide all the data values by any constant, all measures of position (such as the mean, median, and percentiles) and measures of spread (such as the range, the IQR, and the standard deviation) are multiplied (or divided) by that same constant.

Example 2: You have a data set: y1 = 1, y2 = 2, y3 = 3, y4 = 4, y5 = 5. If you want to multiply each observation with a constant d = 2

column	n	mean	variance	Std. dev.	Std. err.	median	range	min	max	Q1	Q3
y	5	3	2.5	1.58	0.707	3	4	1	5	2	4
yx2	5	6	10	3.16	1.414	6	8	2	10	4	8

Shape is the same but it is more spread out, it'll affect mean, std dev, median, range, min, max, both quartiles, only the shape is unchanged

When checking variance we must square the standard deviation

$S^2new = (2xSold)^2 = 4 \times S^2old$

S = standard deviation

Summary for rescaling data:

If new original y = d × y for each observation

For measures of center or position:

- Centernew = d × Centeroriginal
- Positionnew = d × Positionoriginal

For measures of spread and shape:

- Spreadnew = d × Spreadoriginal
- **Shapenew = Shapeoriginal**

Thus, if you rescale and shift data: (Ynew = d x Yoriginal + c) for each observation

For measures of center or position:

- Centernew = d x Centeroriginal + c
- Positionnew = d x positionoriginal + c

For measures of spread and shape:

- Spreadnew = d × Spreadoriginal
- Shapenew = Shapeoriginal

Example 2: Students taking an intro econ class reported the number of credit hours that they were taking that quarter. Summary statistics are shown below

Mean	Std. dev.	min	max	Q1	Q3	median
16.65	2.96	5	28	15	19	16

If the college charges $73 per credit hour plus a flat fee of $35 per quarter. For example, a student taking 12 credit hours would pay $35 + 12($73) = $911 for that quarter.

What is the mean fee paid?

It would be 35 + (16.65x73) = $1250.48

What is the standard deviation for the fees paid?

Recall that standard deviation of fees is spread so it's only affected by the rescaling factor (not adding)

S.D. = 73 x 2.96 = $216.08

What is the median fee paid?

It would be 35 + (16x73) = $1203

The Standard Deviation as a Ruler

The distance to the mean in a specific observation measured in standard deviations gives information about the location of this value relative to the other values in the data set. (For example +1 standard deviation away from the average math mark of the class)

Standardizing with z-score

Z- score or standardized value is a measure of relative standing

The formula for this is

$$Z = \frac{Y - \bar{Y}}{S_y}$$

ȳ = shifting, s = rescaling

Where y is an observation from a sample with mean (ȳ) and standard deviation s, the formula for z-score is above.

The definition for z-score is "how many standard deviations away from the mean does the measurement lie and in which direction?

We can have a positive z-score, a negative z-score, and a z-score of 0 which the value is not deviated at all from the centre.

Example: In a class of 3 students, the average score on the final exam is 75% and standard deviation of 5%.

Amy got 80% on the final exam. How many standard deviations better than the mean is that?

It would be +1 standard deviation away since 80% is 5% away from 75%

Mandy got 70%. How many standard deviations is Mandy's score deviate from the mean?

It would be -1 standard deviation away from the mean

Judy got 75%. How many standard deviation is Judy's score deviate from the mean?

It would be 0 standard deviations away from the mean, it isn't deviated away from the centre at all.

Benefits of standardizing:

- Since standardized values have been converted from their original units to the standard deviations away from the mean we can compare values that are measured on different scales with different units
- Standardizing z-scores doesn't change the shape of the distribution

However standardizing into z-scores does change the center by always making the mean 0, it also changes the spread by making the standard deviation always 1.

The Normal Model

Recall that the y-bar is for shifts and the S is for rescaling, these both affect the center but doesn't change the shape. Standardizing the center makes the mean 0 and standardizing spread makes the standard deviation 1

$$Z = \frac{Y - \overline{Y}}{S_y}$$

Example 1:

Two graduate students:

- One is an accounting major and gets a job offer for $ 36000
- One is an advertising major and gets a job offer for $ 33000

Other students of:

- Accounting: $\bar{y}$ = 36500 and s = 1500
- Advertising: $\bar{y}$ = 32500 and s = 1000
-

Q: Which one would more likely be happier about their job?

A: They're both receiving 500 over the mean however we should look at their z-scores

Accounting - (36 500 - 36 000)/1500 = +0.3333

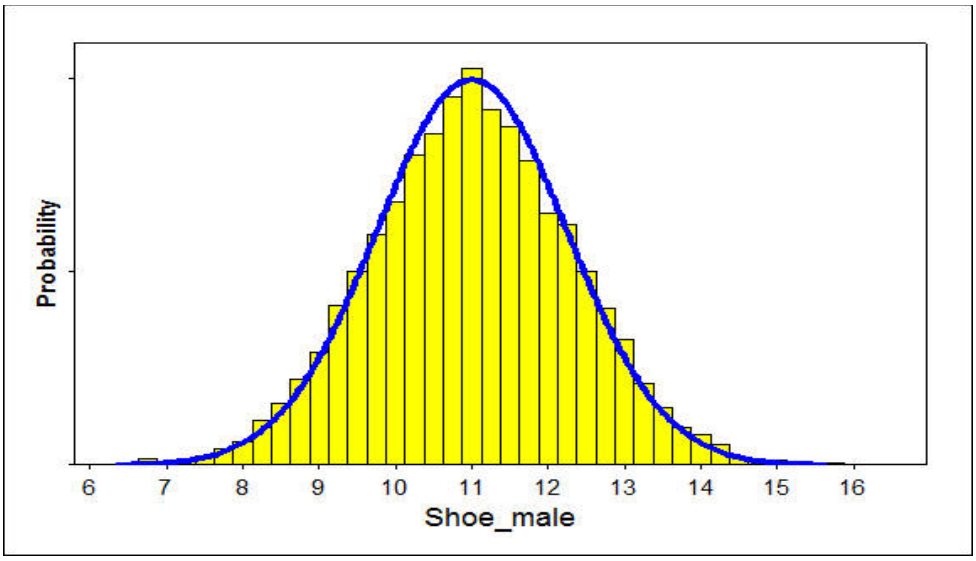
Advertising - (33000-32500)/1000 = +0.5

Looking at their z-score, most luckily the advertising student would be happier with their income

Density Curve and Normal Model

Continuous random variables: These are described by histograms

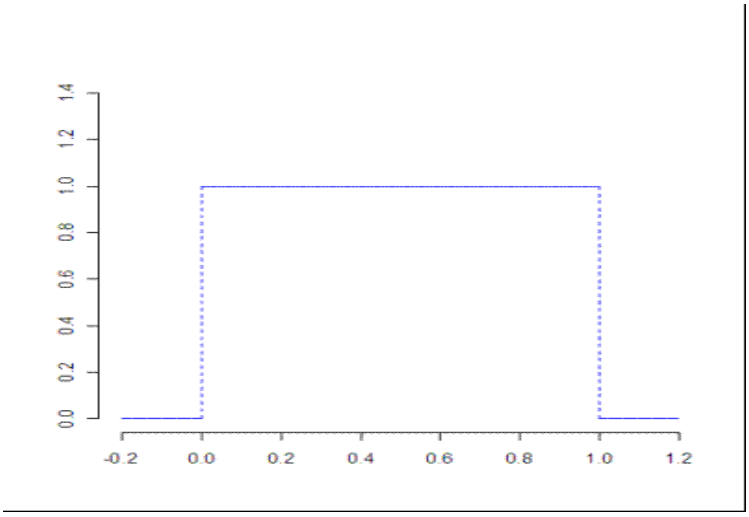
The curve through the histogram is the density curve, also the count is always positive so the curve will always be on or above the horizontal axis. The total area below the curve will always equal to one



The smooth curve on the histogram is called a density curve, they show probability and have to abide by the following rules:

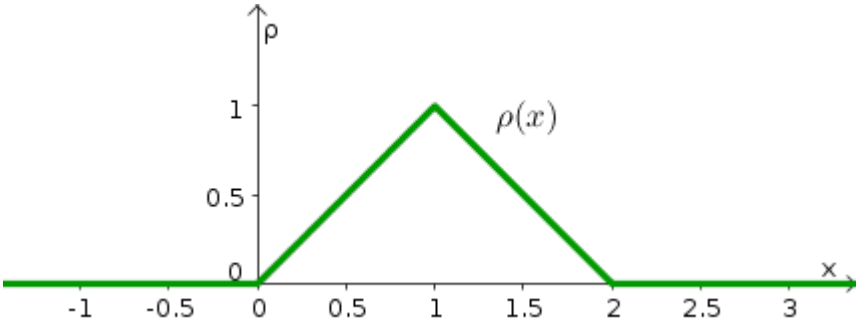
1) A density curve is always on or above the horizontal axis
2) The total area under the density curve is equal to 1
3) The area under the curve above a certain interval is the proportion of all observations that fall in that range. I.e. P(a<x<b) = area under the curve between a and b
4) There is no probability attached to any single value. I.e. P(x = a) = 0, this rings true for continuous values but not discrete since there isn't area attached to the continuous random variables

This graph displays the density of a uniform distribution in an interval [0.0, 1.0]



We can verify that the area underneath is equal to one by using the formula A = L x W, which would be A = 1.0 x 1.0 = 1

We can also calculate the area for density curves similar to this one



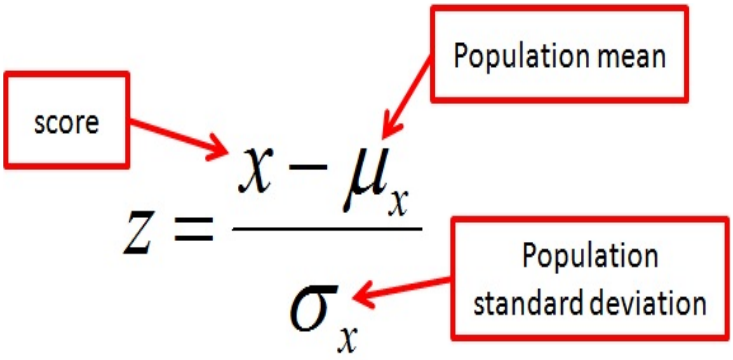
A = 2 x 0.5 = 1

Normal Distributions

They are shaped like bell curves and is the most important and widely used of all probability distributions. They always have the following properties:

1) Symmetric, unimodal (one mode), and bell shaped
2) For every combination of mean and standard deviation there is a different curve
3) μ is the center of the distribution (right at the highest point of the density distribution function) and σ controls the spread of the distribution

* Notation: N (μ,σ) represents a Normal model with a mean of μ and a standard deviation of σ *

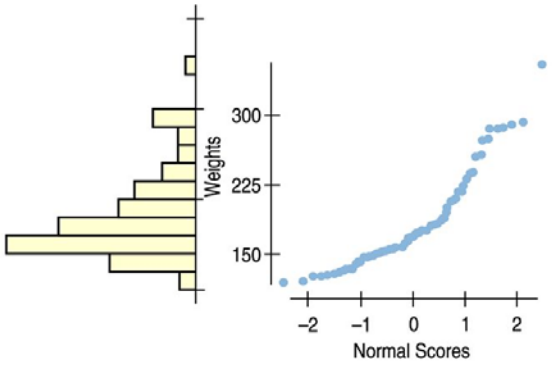


The x variable can also be replace with y

Checking normality assumption: When we use the normal model we assume the distribution is normal however we need to check this by either making a histogram or normal distribution plot

Skewed right probability plot (distribution isn't normal)

- A skewed distribution might have a histogram and Normal probability plot like this:

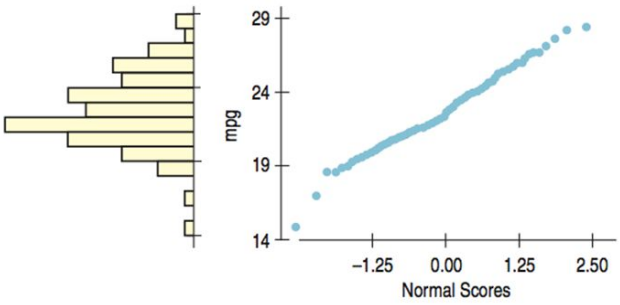


Normal probability plot and histogram

Normal Probability Plots

The *Normal probability plot is a* specialized graph that can help decide whether the Normal model is appropriate.

If the data are approximately normal, the plot is roughly a diagonal straight line. Histogram and Normal probability plot for gas mileage (mpg) for a Nissan Maxima are nearly normal, with 2 trailing low values.

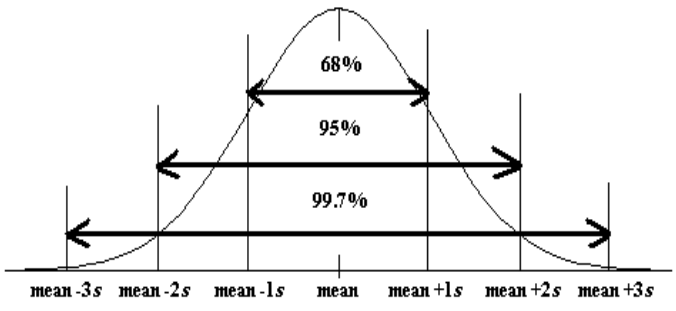


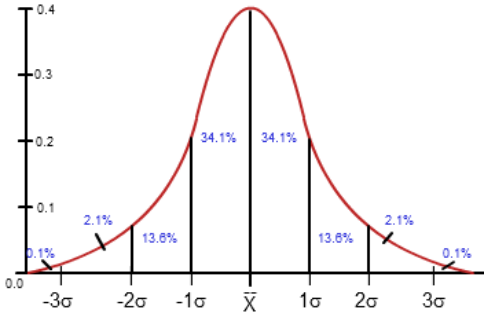
Empirical Rule

This only applies to normal curves and basically states that

- Approx. 68% of all observations fall within 1 standard deviation of the mean
- Approx. 95% of all observations fall within 2 standard deviations of the mean
- Approx. 99.7% of all data falls within 3 standard deviations of the mean

Note that these are all APPROXIMATE percentages, not exact





Example: The height of 112 children follows a normal distribution with $\bar{y}$ =104.5 and standard deviation s = 16.3.

n	$\bar{y}$ + ns	Empirical
1	(88.2, 120.8)	~ 68%
2	(71.9, 137.1)	~ 95%
3	(55.6, 153.4)	~ 99.7%

Example: The time to complete a math exam is approximately normal with a mean of 70 minutes and a standard deviation of 10 minutes. Using the 68-95-99.7 rule

1) What percentage of students will complete the exam in under an hour?

P(y<60) = (100% - 68%)/2 = **~ 16%**

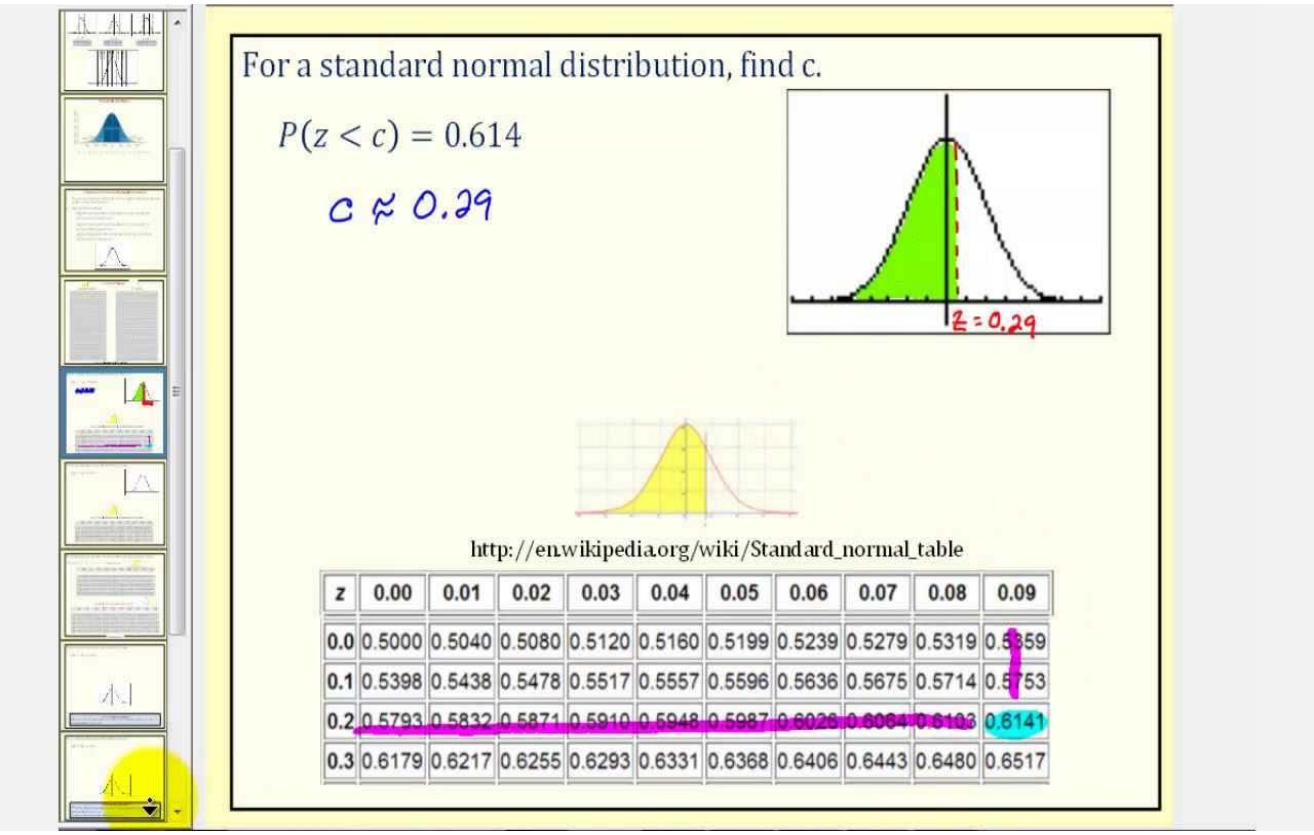
2) What percentage of students will complete the exam between 60 and 70 minutes?

P (60 < y < 70) = 68/2 = **~34%**

Using Z-Tables

If we want to find the probability within 1.5 standard deviations away from the mean, how to find such probability value?

Use Table z for many different values of z*, the area under the curve from -∞ to z, which is called the *cumulative area* or *cumulative proportion*, for standard normal distributed variables.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517

Note this example is showing how to find the area under the curve to the left of the value of 0.29. We would first look in the left column and find 0.2, then we would go horizontally across on the table until we lined up with the value 0.9 on the upper column. Then, we found our value for area which turns out to be 0.614.

Area for Normal Distribution

Examples from past lecture lesson continued below

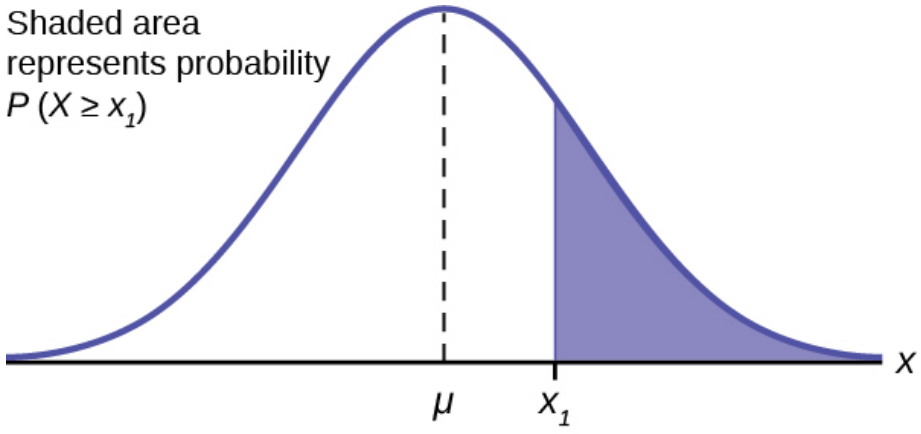
Finding the area to the right of 1.28:

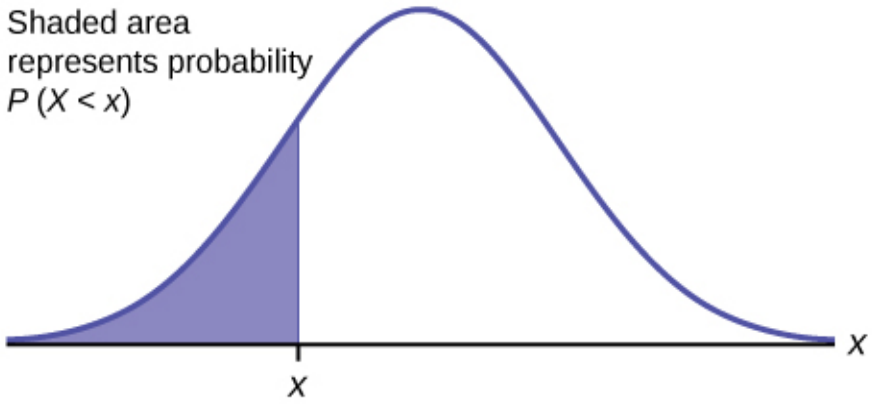
P (Z>1.28), you could find it two ways

1) 1 - P(Z ≤ 1.28) = 1 - 0.8997 = 0.1003

OR

2) Treat it like a mirror image of the flip





1.28 = -1.28 which would equal 0.1003

Finding the area between -1 and 1:

P (-1≤ -1)

Step 1 - P (2≤ 1) - P (2≤ -1)

Step 2 - 0.8413 - 0.517 = 0.6826

If you wanted to find the area left of -6 it would be 0, since if you look at your z-table after 3.9 the area is 0

How to find the area for any normal distribution?

If y is normally distributed with the mean μ and the standard deviation σ ((ie. y ~ N (μ, σ)), then the standardized variable

$$z = \frac{x - \mu}{\sigma}$$

Is normal distributed with mean = 0 and standard deviation = 1

Example:

Let y be normal distributed with μ = 100 and σ = 5, so y ~ N (100; 5).

Calculate the area under the curve between 98 and 107 for the distribution.

This is not a standard normal distribution because the mean is 100 and not 0, also standard deviation is 5 not 1. WE CANNOT USE Z-TABLES FOR UNSTANDARDIZED Y SCALE VALUES.

To standardize results we need to put it into the formula:

$$z = \frac{x - \mu}{\sigma}$$

P ((98 - 100/5) < (y - 100/5) - (107 - 100/5))

P (-0.4< Z < 1.4)

Using table Z = 0.9192 - 0.3446 = 0.5746

We always need to standardize our variables if they aren't already

Y-scale → Z-scale

Summary to find normal proportions:

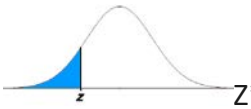
1) State the problem in terms of the observed variable Y
2) Standardize Y to restate the problem in terms of a standardized normal variable Z
3) Find the required area using the Z table

Inverse Normal Calculations:

In some cases, the area we want to find under the graph will be given, and we will have to find the interval in which it lies in.

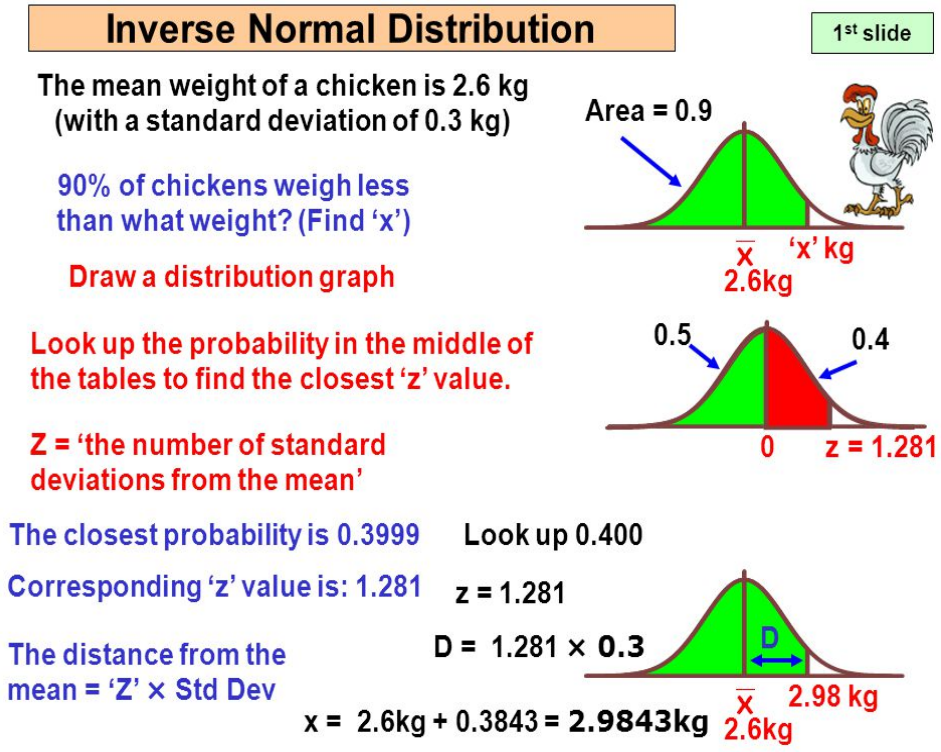
Example: If we wanted to find the z-value that makes up the smallest 2% what would we do?

In other words we would like to find Z* such that P (z<z*) = 0.02



Z* ~ 2.05

We would look at the numbers in between to see which one is closest to 0.02, after we found that we look match it with the corresponding values in the side and top columns.

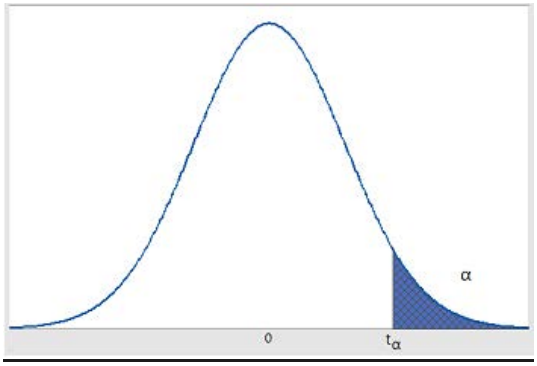


Example:

Now we are interested in the largest 5%, So we are looking for z*, with P(z > z*) = 0.05.

P (Z ≤ Z*) = 0.95

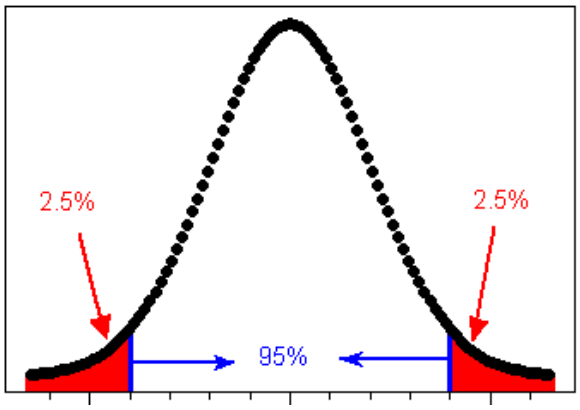
Z -> (1.04 + 1.65/2) = <u>1.645</u>



Example:

Now suppose we are interested in the middle 95%. So we are looking for -z* and z*, with

P (-z* ≤ z ≤ z*) = 0.95.



P(Z ≤ Z*) = 0.025

-Z* = -1.96 can't have negative so flip the sign

Z* = 1.96

Example: **Backward Normal Calculation**

Let y be normal distributed with $\mu = 100$ and $\sigma = 5$. Find the value that makes up the smallest 30% for this distribution. $Y \sim N(100,5)$

Note: This isn't normal distribution so we have to standardize it using

$$Z = \frac{x - \mu}{\sigma}$$

$P(Z < Z^*) = 0.3$

$Z^* = -0.52$

No z-value because if it's on the scale -> $Y^* = \mu + Z^*\sigma$

$= 100 + (-0.52)(5)$

$= 97.4$

<u>Summary for Backward normal calculations:</u>

1) State the problem in terms of the proportion
2) Look in the body of Table z for the number closest to the proportion, and find the corresponding z* value
3) Unstandardized to transform the z* value back to the original y scale using $y^* = \mu + z^*\sigma$

Example:

a) Assume that the length of a human pregnancy follows a normal distribution with mean 266 and standard deviation 16. What is the probability that a human pregnancy lasts longer than 280 days?

P(y > 280) this isn't on z-scale so we must standardize it

= (280 - 266/16)

= P(z > 0.88)

1 - P(z < 0.88)

Or P(Z < -0.88)

= 0.1894

b) How long do the 10% shortest pregnancies last?

P(Z < Z*) = 0.1

Z* = -1.28 (unstandardize to get back to Y*)

$y^* = \mu + z^*\sigma = 266 + (-1.28)(16)$

= 245.5 days

Chapter 12
From Randomness to Probability

Random phenomenon

- Many things in our world depends on randomness, such as what is the chance of rolling a "6" with a fair die? Or what is the probability that the bus will be on time today?
- Although these events occur randomly, there is an underlying pattern in the occurrence of these events, this is the basis for probability theory
- A phenomenon is random is we know the possible outcomes that can occur but we don't know which ones did or will. The individual outcomes are unpredictable and with a large number of observations, predictable outcomes will occur

1) Examples for random phenomenon include:

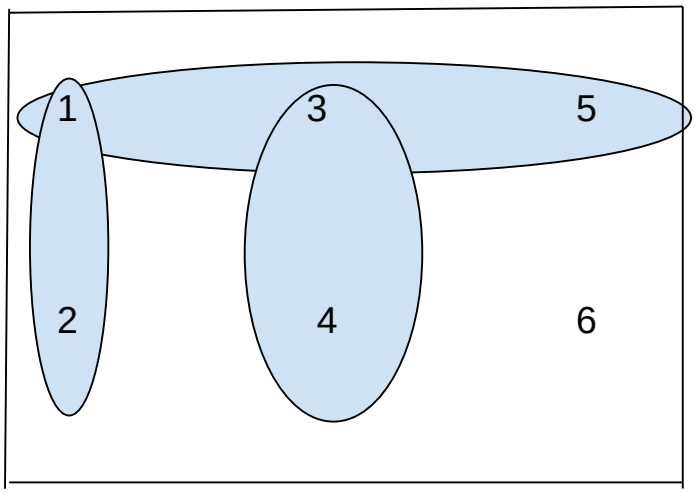
- Count red blood cells in a blood sample {1 - 2M/ml}
- Toss a coin twice {HT, HH, TT, TH}

2) We should put all possible outcomes in brackets that represent a sample space {}, use S to represent.
3) When we combine all outcomes, the combination is called an event which is a collection of one or more outcomes of an experiment
4) Events are usually denoted by the letters beginning the alphabet such as A, B, C
5) The collection of all possible outcomes is called the sample space

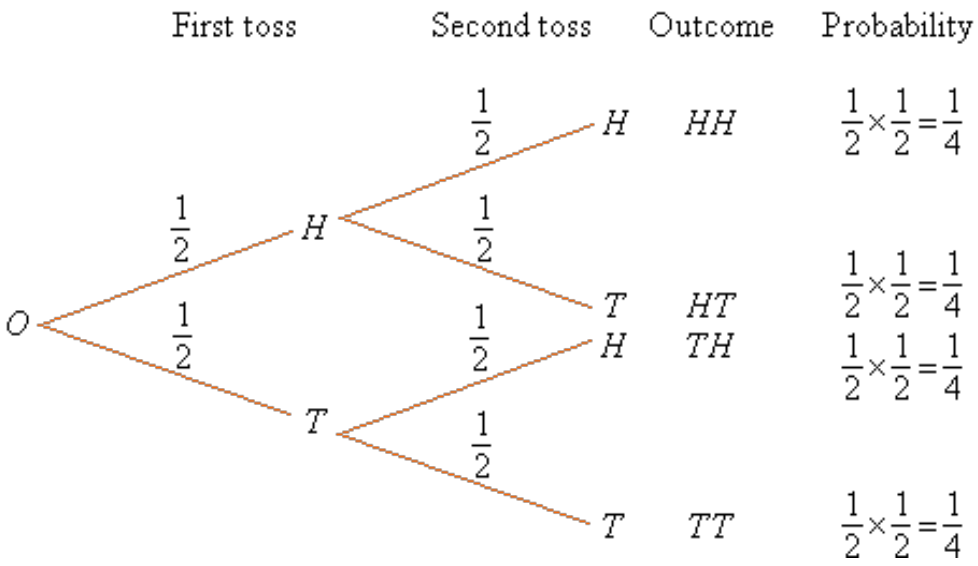
Venn diagrams

A Venn diagram is a picture that shows all the possible outcomes for an experiment. The outer box represents the sample space, which contains all of the outcomes, and the appropriate

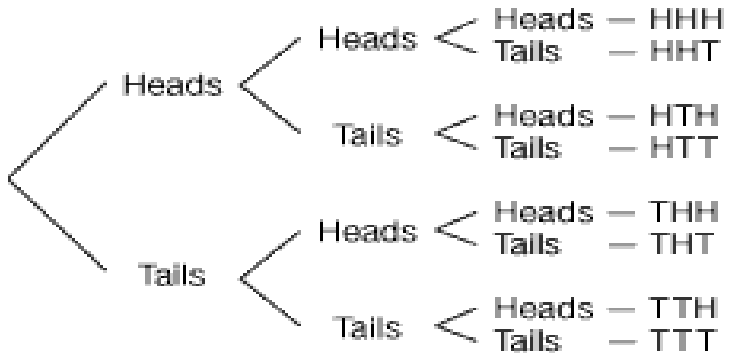
events are circled and labeled. A = {1,2} B = {3,4} C = {1,3,5}



Example: If we toss 2 coins, what is the sample space?



(1st toss outcomes)2 x 2 (2nd toss outcomes) = 4 possible outcomes



2 x 2 x 2 = 8 possible outcomes

A = {HHH} A is the outcome of getting all heads

B = {HHH, HHT, HTH, THH} B is the outcome of getting at least two heads

Probability = Proportion of times the outcome would occur in a series of very long repetitions

This definition is based on the law of large numbers which states that "As the number of experiments increases, the actual ratio of outcomes will converge on the theoretical, or expected, ratio of outcomes."

<u>Modelling probability</u>

Equally likely which means the same thing as fair, unbiased or same chance of occurring, however events aren't always equally likely, a skilled basketball player has more than a 50/50 chance of getting it in

THEORETICAL PROBABILITIES
$P(\boxdot) = \frac{1}{6}$
$P(\boxdot) = \frac{1}{6}$
$P(\boxdot) = \frac{1}{6}$
$P(\boxdot) = \frac{1}{6}$
$P(\boxdot) = \frac{1}{6}$
$P(\boxdot) = \frac{1}{6}$
Total = 1

OR

THEORETICAL PROBABILITIES
$P(1) = \frac{1}{6}$
$P(2) = \frac{1}{6}$
$P(3) = \frac{1}{6}$
$P(4) = \frac{1}{6}$
$P(5) = \frac{1}{6}$
$P(6) = \frac{1}{6}$
Total = 1

Probability of an event with equally likely outcomes is

P (A) = (Total # of outcomes in A/ Total number of possible outcomes)

P (A) = sum of the probabilities of the individual outcomes contained in A.

Find the probability of A.

1) Let A be the event of obtaining a '2' in one roll of an unbiased die.

P(A) = ⅙ A = {2}

2) Let A be the event of obtaining at least one tail in two tosses of an unbiased coin.

A = {HT, TH, TT} **P (A) = ¾**

P (A) = P (HT) + P (TH) + P (TT) = ¼ + ¼ + ¼ = **¾**

Formal Probability Plots

The probabilities must follow the following rules

1) 2 requirements for a probability:

- The probability of each simple event (individual outcome) is between 0 and 1. i.e. For any event A, ($0 \leq P \leq 1$)

- P (A) = 0, if the event A never occurs

- P (A) = 1, if the event always occurs.

2) Total Probability Rule: the probability of the set of all possible outcomes of a trial must be 1. I.e. Then P(S) = 1

Ex: Roll a 6 sided die P ({1, 2, 3, 4, 5, 6}) = 1

Example: Blood type

a) All human blood can be typed as one of O, A, B, or AB, but the distribution varies a bit with race. Here is the model for a randomly chosen black American male.

Blood type	A	B	O	AB
Probability	0.49	0.27	0.2	?

The probability of him having AB blood type would be 1 minus all those values

P(S) = 1 = P (O) + P (A) + P (B) + P (AB)

P (AB) = 1 - (0.49 + 0.27 + 0.2) = 0.04

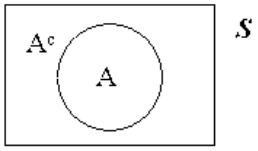
b) Maria has type B blood. She can safely receive blood transfusions from people with blood types B and O. What is the chance that a randomly selected black American can donate blood to Maria?

P(B,O) = P(B) + P(O) = 0.2 + 0.49 = 0.69

Complement Rule

Complement of an event A, denoted by A^c

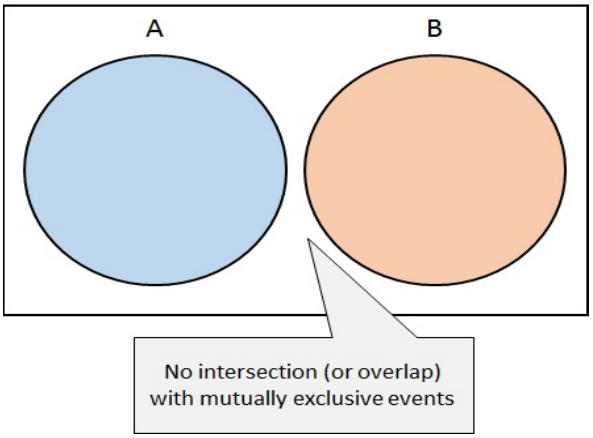
This symbol means all outcomes in the sample size that IS NOT A



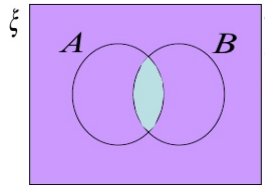
The probabilities of A and Ac add up to 1

If A & B are disjoint events (mutually exclusive), this means the two events observed

have no common outcomes, if one event occurs the other cannot, they do not overlap.



MUTUALLY EXCLUSIVE EVENTS

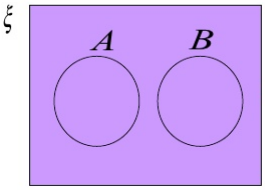


The probability of event A or event B occurring / not mutually exclusive

P(A or B)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) \neq 0$$

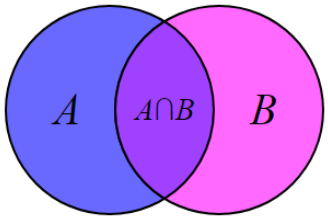


The probability of event A and event B mutually exclusive.

P(A or B)

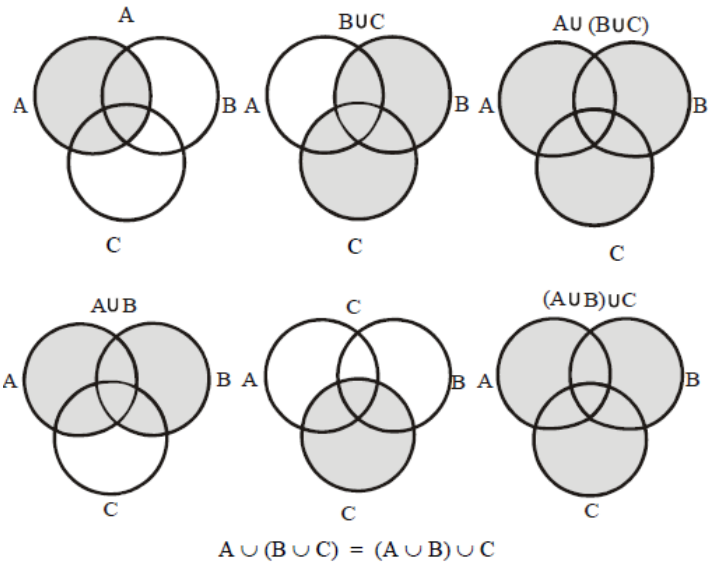
$$P(A \cup B) = P(A) + P(B)$$

because $P(A \cap B) = 0$



The symbol of the upside down U denotes intersection, which means A and B

The symbol union, (U) means and/or



General addition rule and disjoint (mutually exclusive, no overlap) addition rule

General Addition Rule

General Addition Rule:

P(A or B) = P(A) + P(B) - P(A and B)

If A and B are mutually exclusive, then

P(A and B) = 0, so the rule can be simplified:

P(A or B) = P(A) + P(B)

For mutually exclusive events A and B

The reason we need to minus the overlap in the general addition rule is because the first two variables P (A) and P(B) count for two overlaps, we need to minus one of them since we only have one overlap.

Rules of Probability: General Addition Rule

- If A and B are any two events,
  P(A or B) = P(A) + P(B) – P(A & B)
- Example: Toss Two Coins
  A: Getting Head on Coin No. 1
  B: Getting Head on Coin No. 2
  P(A)= ½, P(B)= ½, P(A & B)= ¼
  So, P(A or B)= ½ + ½ - ¼ = ¾ =0.75=75%

Varsha Varde 30

Example: Addition Rule for Disjoint Events

Example: If you flip two coins, what is the probability of getting only heads or ONLY tails?

S = {HH, HT, TH, TT}.
We first need to determine the probabilities. Let's create a probabilty model:

Event	HH	HT	TH	TT
Prob	0.25	0.25	0.25	0.25

Because this is an 'or' question, it is a good time to think about using the addition rule. Before we do so, we must confirm that th events we are interested in are disjoint with each other. In this case, the events {HH} and {TT} are indeed disjoint. Therefore, we can use our addition rule for disjoint events.

Answer: So, the probability that of getting only heads (HH) or only tails (HH):

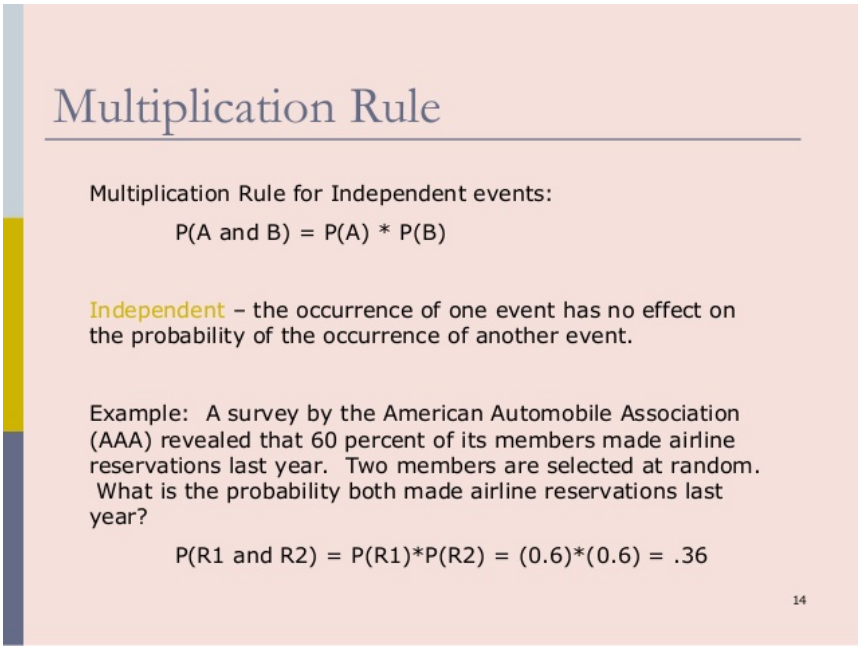
$P(\text{HH or TT}) = P(\text{HH}) + P(\text{TT})$
$= 0.25 + 0.25$
$= 0.50$

Chapter 13
Independence

Multiplication rule: probability of the intersection of independent events

P(A and B) = P(A) X P(B)

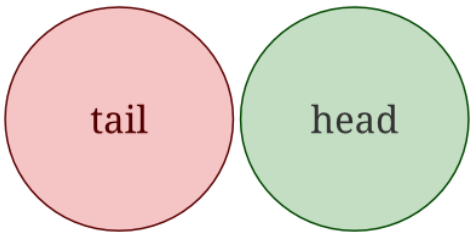
Independence of two events: two events are considered independent when one occurrence of one event has NO impact on the probability for the second event to occur



*These INDEPENDENT events do not and cannot be DISJOINT

Disjoint = P (A∩B) = 0

Independent = P (A∩B) = P (A) X P (B)



For example: If we flipped a one coin toss, A and B are disjoint, if the coin landed on heads there is no chance that the outcome is tail, since it is head. Before we flipped the coin the outcomes were 50/50, however after it landed on heads it was a 0% chance outcome that it was tails, this means it is NOT independent because the probability of the second (tails) occurring changed after the first occurred (heads).

A common error is to treat disjoint events as if they were independent, and apply the Multiplication Rule for independent events—don't do that

Example 1: Toss a fair coin twice. What is the probability to toss two heads?

Recall: P ({HH}) = ¼

Find P (A and B) = P (A)* P (B) = ½ x ½ = ¼

A = head on first toss (independent trials)

B = head on 2nd toss (independent trials)

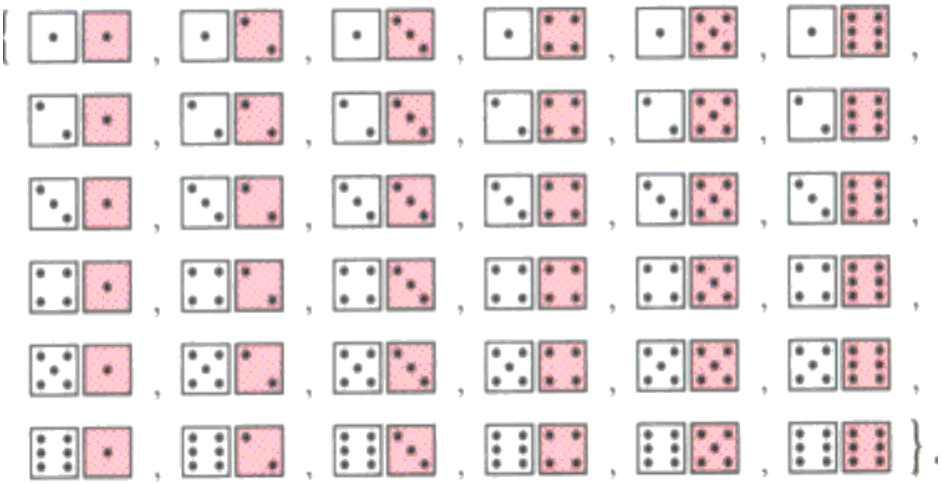
Example 2: What is the probability to get the outcome (HTTH) with a biased coin which has a 0.1 chance to toss a head? P ({HTTH}) = P (H) P (T) P (T) P (H)

P (H) = 0.1

P (T) = 1 - 0.1 = 0.9

P ({HTTH}) (independent trials) = 0.1 x 0.9 x 0.9 x 0.1 = 0.0081

Examples: Rolling two dice



Method 1: counting

a) When you roll two dice, what is the probability to roll two 6's? There are 36 possible outcomes. You could count the outcomes from the picture which would give you 1/36

OR you could use

Method 2: probability multiplication (probability rule 5)

P (6, 1 ∩ 6, 2) = P (6, 1) x P (6, 2) = ⅙ x ⅙ = **1/36**

6, 1 = probability of rolling a 6 with the FIRST die

6, 2 = probability of rolling a 6 with the SECOND die

b) What is the chance that none of the dice are divisible by 3 with 2 fair dice?

Use P (A and B) = P (A) P (B)

Let A = no numbers are divisible by 3 with the **first** dice (1, 2, 4, 5)

Let B = no numbers are divisible by 3 with the **second** dice (1, 2, 4, 5)

P (A∩B) = P(A) x P(B) = 4/6 x 4/6 = 4/9

c) What is the chance of rolling **at least** one 6 with 2 fair dice?

Method 2: Using Probability Rule 5, we have: P (AB) = P (A) + P (B) – P (AB)

Let A to be the event that the first die is 6, and let B to be the event that the second die is 6

Method 2: P (AB) = P (A) + P (B) – P (AB)

⅙ + ⅙ - 1/36 = 11/36

However, if the question has the phrase “at least” in it, we can use the complement rule

Method 3: P (A) = 1 - P(Ac)

$$P(A^C) = 1 - P(A)$$

Ac = A complement, and in this question it would mean the trials that roll no 6’s (NOT AT LEAST one 6)

P(A) = 1 - P(Ac) = 1 - P(no 6,1 ∩ no 6,2)

= 1 - (⅚ x ⅚) = **11/36**

d) What is the chance of rolling at least one 6 with 3 fair dice?

We shouldn’t use method 1 for this one as there would be 216 outcomes and that is way too many, instead we should use the complement rule as we can see the question involves the phrase “at least”

P(at least one 6) = 1 - P(no ‘6’ in any other rolls)

1 - P((no 6,1) x (no 6,2) x (no 6,3))

1 - (⅚ x ⅚ x ⅚) = 91/260 or 0.4213

e) What is the chance of getting not all 6 with 3 fair dice?

For this we must first find the probability of rolling all 6’s with three dice

P(not all '6') = 1 - P(all 6)

1 - ((⅙) x (⅙) x (⅙)) = 215/216

Example - pop quiz:

1) What are the chances that they answer all questions correctly?

   P ({CCC}) = P(C) probability of getting it correct is 0.20

   P (I) probability of getting it incorrect is 0.80

   P (C1 ∩ C2 ∩ C3) = P (C1) x P (C2) x P (C3) =0.2 x 0.2 x 0.2 = **0.08**

2) What is the probability that they answer one question correctly?

   P ({IIC, ICI, CII})

   P (IIC) + P (ICI) + P (CII)

   0.128 + 0.128 + 0.128 = 0.384

   The 0.128 comes from multiplying the chance of getting one correct (0.2) by two incorrect (0.8)

3) What is the probability of the student answering at least 2 questions correctly?

   P ({CCC, CCI, CIC, ICC})

   P (CCC) + P (CCI) + P (CIC) + P (ICC)

   0.008 + 0.032 + 0.032 + 0.032 = 0.104

4) What is the probability of the student answering at least 1 questions correctly?

P ({CCC, CCI, CIC, ICC, IIC, ICI, CII})

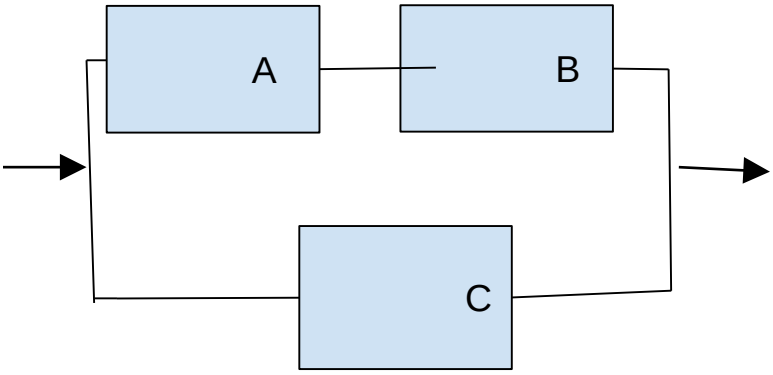
Complement rule

P (at least one correct)

1 - P(none correct) = 1 - P(III)

1 - 0.512 = **0.488**

System Reliability:



The signal will pass through if "A and B" or "C" is working. If these 3 independent components are 95% individually reliable, what is the reliability of the system?

P (A) = P (B) = P(C) = 0.95, they're all the same

P (signal passes) = P (A ∩ B) ∩ (or) C

(Event 1) P (A∩B) + (event 2) P(C) - (P ∩ B)∩C)

(P(A) x P(B) + P(C))- (P(A)xP(B)xP(C)) =

(0.95 X 0.95 + 0.95) - (0.95 X 0.95 X 0.95)= 0.995125

System as a whole is 0.995125% reliable

Example 1: A satellite has two power systems, a main and an independent backup system. If the probability of failure in the first ten years for the main system is 0.05 and for the backup system 0.08.

a) What is the probability that both systems fail in the first 10 years and the satellite will be lost? Let M be the event that the main system fails, and B the event that the backup systems fails in the first 10 years.

MB = both fail (0.05 x 0.08) = 0.04

(Note: McBc = both functional (0.95x0.92) = 0.874, M c is the event that the main system is functional, and B c the event that the backup system is still functional)

b) What is the probability that at least one of the systems is still functional after the first 10 years?

Let O be the event that at least one of the systems is still operational after 10 years

P(O) = 1 - P(Oc)

1 - P(both systems fails) = 1 - 0.04

0.996 is the probability

c) What is the probability that both systems are still functional after the first 10 year?

P(McBc) = P(Mc) x P(Bc) =

1 - P(M) x 1 - P(B)

(1 - 0.05)(1-0.08) = 0.874

d) What is the probability that one functions and the other one fails?

P(McB) or P(MBc) = P(McB) + P(MBc) =

P(Mc)P(B) + P(M)P(Bc)

= 0.95x0.08 +(0.05 x 0.92) = 0.122

Union symbol = add $\cup$

Intersection symbol = multiply $\cap$

Example 2:

Insurance company records indicate that 12% of all teenage drivers have been ticketed for speeding and 9% for going through a red light. If 4% have been ticketed for both, what is the probability that a teenage driver:

Let R = running a red light

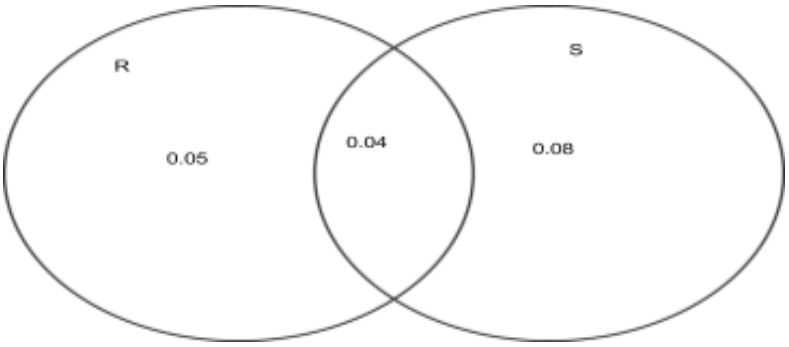
S = get ticket for speeding

P(R and S) = 0.04

P(S) = 0.12

P(R) = 0.09

S & R are NOT INDEPENDENT because the probability of one event happening does change the probability of the other event happening, therefore we cannot use the multiplication rule we have been using for independent events. The P(S and Rc) = P(S) x P(Rc) is NOT APPLICABLE here.



a) Has been issued a ticket for speeding but not for running a red light?

P(S $\cap$ Rc) = P(S) - P(S $\cap$ R), have to minus the overlap between the two

This would give us 0.08

b) Has been issued a ticket for speeding or for running a red light but not both?

P(S or R not both) = 0.05 + 0.08 = 0.13

OR

P (SUR) - P(S $\cap$ R) = 0.013

c) Has not been issued a ticket for speeding nor for running a red light?

(This would be the area outside the circles on the diagram)

P (Sc $\cap$ Rc) = 1 - P (SUR)

1 - 0.05 - 0.08 = 0.83

Often events are NOT independent

Assume the instructor only gives 2 questions in the pop quiz and he finds the proportions for the actual responses of her students:

Outcome	II	IC	CI	CC
Probability	0.26	0.11	0.05	0.58

Let A: {first question correct}

B: {2nd question correct}

Find P(A), P(B) and P(A and B).P(A) =

P({CI,CC}) = 0.05 + 0.58 = **0.63**

P(B) = P({IC, CC}) = 0.11 + 0.58 = **0.69**

P(A $\cap$ B) = P({CC}) = **0.58**

Are A and B independent events? No, they are dependent because if we used the independent multiplication rule we would get a different answer, as shown below

P(A $\cap$ B) = P(A)P(B) = 0.63 x 0.69 = 0.43 which DOESN'T EQUAL 0.58.

NOTE: Responses to different questions on a quiz are typically not independent. Most students do not guess randomly. Students who get the first question correct may have studied more than students who do not get the 1st question correct, and thus they may also be more likely to get the 2nd question correct. NOTE: Don't assume that events are independent unless you have given this assumption careful thought and it seems plausible.

To understand the concept of independent and dependent events even further, we will introduce the idea of **conditional probability**.

Conditional Probability

If A and B are events with P (B) > 0, the conditional probability of A given B is defined by

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

The interpretation of the conditional probability P (A|B) is as follows: Given that you know already event B occurred, what is the probability that A occurs?

Example: Toss a fair coin twice.

A: head on second toss

B: head on first toss

P (A|B) = 1/2 and P(A|Bc) = 1/2

A and B are independent of each other since each toss is independent of each other

Suppose I roll a fair die and ask, “What is the chance to get a 6 ?” The answer is P(6) = 1/6

Now suppose I give you a hint: the number on the die is even. Now what is the chance that it is a 6? P(6|even number) = 1/3

These events are NOT independent (dependent) of each other since the first event happening changed the probability of the second event happening.

General Multiplication Rule

The General Multiplication Rule: Rearranging the equation in the definition for conditional probability, we get the General Multiplication Rule:

1) $P(A \cap B) = P(A)\, P(B|A))$, or

2) $P(A \cap B) = P(B)\, P(A|B))$

Note: $P(A|B) \neq P(A)$; $P(B|A) \neq P(B)$

Summary for independent events

Events A and B are independent if **any one** of the following conditions is satisfied

1) $P(A|B) = P(A)$

2) $P(B|A) = P(B)$

OR 3) $P(A \cap B) = P(A)\, P(B))$

Example:

A bowl contains five M&Ms®, two red and three blue. Randomly select two candies **without replacement**, and find

A: second candy is red

B: first candy is blue

Find P(A|B) = 2/4 and P(A|Bc)= ¼

These two events are dependent since the probability of A changed because of B.

Example Cont:

A bowl contains five M&Ms®, two red and three blue. Randomly select two candies **with replacement**, and find

A: second candy is red.

B: first candy is blue.

Find P (A|B) = ⅖ and P (A|Bc) = ⅖

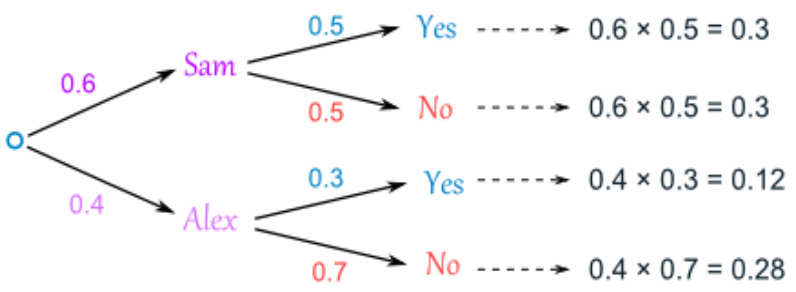
These two events are independent since the probability doesn't change

Example:

Five juniors and four seniors have applied for two open student council positions. School administrators have decided to pick the two new members randomly. What is the probability that one junior and one senior are chosen for the two positions? Let J = the person chosen is junior; S = the person chosen is senior.

For this example you could draw a tree diagram to find the probability or just use a formula

For a tree diagram you would draw it something like this,



Except you would replace the names “Sam” and “Alex” with Junior and Senior, the first value for Junior would be 5/9, then the first value for senior would be 4/9 (make sure they add up to 1)

The second value for junior would be 4/8 and the second value for senior would be 4/8.

P(J1S2 or S1J2) = P(J1S2) + P(S1J2) =

P(J1) x P(S2|J1) + P(S1)P(J2|S1) =

5/9 x 4/8 + (4/9 x ⅝) = 5/9

Multiplication Rule for Finding P (A and B)

For dependent events A and B, the probability that A and B both occur:

• P (A and B) = P (A|B) x P (B); and

• P (A and B) = P (B|A) x P (A)

For independent events A and B, the probability that A and B both occur is

- P(A and B) = P(A) x P(B)

Disjoint events

P (A and B) = 0 b/c no overlap NOT the same as dependent events

Example:

Six soldiers have to decide between themselves which one goes on a suicide mission. They decide to draw straws: there are 5 long straws and 1 short one, and they take turns picking one. The guy with the short straw loses. Is it better to pick first or second?

L = long straw

S = short straw

P(S1) = ⅙

P(L and S2) = P(L1) x P(S2|L2)

= ⅚ x ⅕ = **⅙**

No matter what the probability of picking a short straw will be ⅙ so it's not better to pick first or second since you have to take into account the first guy who pulled the first long straw.

Example: In a criminal trial, a person is being suspected as a murder. In general, the probability that the jury finds the person guilty given the person is innocent is 0.04, and the probability that the jury finds the person innocent given the person is guilty is 0.10. If a city has a criminal rate of 0.25, what is the probability that the jury:

a) Finds the person innocent and he is innocent

It = innocent and true

Id = innocent by decision

Gt = guilty and true

Gd = guilty by decision

P(ItId) = P(It) x P(Id|It) = (1 - 0.25) x (1 - 0.04) = **0.72**

B) finds the person guilty and he is the murderer?

P (Gt) x P(Gd|Gt) = 0.25 x (1 - 0.10) = **0.225**

c) Makes the right decision?

P (right) = P(ItId) + P(GtGd) = 0.72 + 0.225 = **0.945**

d) Finds the person innocent and in fact he is innocent given the jury makes the right decision?

P(ItId|right) = P(ItId and right)/P(right)

(right = ItId, GtGd) =

P(ItId)/P(right) = (0.72/0.945) = **0.76**

Chapter 14
Random Variables

Random variable

Can be represented as r.v. or X, a rv is a variable/value based on a random event

Examples of random variables:

X = number of observed "Tail" while tossing a coin 10 times

X = survival time after specific treatment of a randomly selected patient

X = SAT score for a randomly selected college applicant

Just as variables in sample data, rvs can be categorical or quantitative, and if they are quantitative, they can be either discrete or continuous.

Discrete variables

A finite number of distinct outcomes

Examples:

- The number of stores in a shopping mall
- The number of cars owned by a family
- The number of luggage each traveler carries in the airport

Continuous variables

Any numeric value within an interval (decimal numbers such as 0.587393 whereas discrete values you most likely wouldn't have that number)

Examples:

- Cost of books this term

- Height of football players

Properties of discrete probability distributions:

$0 \leq P(x_i) \leq 1$ (can't have negative variable)

$\sum P(x_i) = 1$ (sum must add up to 1)

Probability model for random variables:

Expected Value: Center (cont.)

- A probability model for a random variable consists of:
  - The collection of all possible values of a random variable, and
  - the probabilities that the values occur.

Roll of a die

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

Random Variable and Probability Distribution

A **probability model** describes the possible outcomes of a chance process and the likelihood that those outcomes will occur.

A numerical variable that describes the outcomes of a chance process is called a **random variable**. The probability model for a random variable is its probability distribution

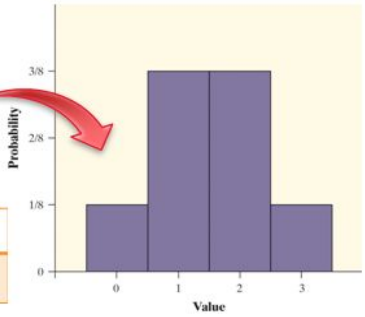
Definition:

A **random variable** takes numerical values that describe the outcomes of some chance process. The **probability distribution** of a random variable gives its possible values and their probabilities.

Example: Consider tossing a fair coin 3 times.
Define X = the number of heads obtained

X = 0: TTT
X = 1: HTT THT TTH
X = 2: HHT HTH THH
X = 3: HHH

Value	0	1	2	3
Probability	1/8	3/8	3/8	1/8



Expected value: Centre

The symbol for expected value is E(X), and this means the same thing as the population mean μ (mu)

∴ E(X) = μ (mu)

The expected value is the value you would expect to see all over again if the experiment is constantly being repeated, it is the center of distribution

If X was a discrete rv with probability distribution P(X). The population mean or expected value of x is given as

$$\mu_x = \sum[x * P(x)]$$

In other words, the expected value of a (discrete) random variable can be found by summing the products of each possible value and the probability that it occurs.

Population Mean (Expected Value) and Population Variance

- Given a discrete random variable **X** with values x_i, that occur with probabilities $p(x_i)$, the population mean of **X** is.
  - 加權平均概念(權數是機率)

$$E(X) = \mu = \sum_{\text{all } x_i} x_i \cdot p(x_i)$$

- Let **X** be a discrete random variable with possible values x_i that occur with probabilities $p(x_i)$, and let $E(x_i) = \mu$. The variance of **X** is defined by

$$V(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \sum_{\text{all } x_i} (x_i - \mu)^2 p(x_i)$$

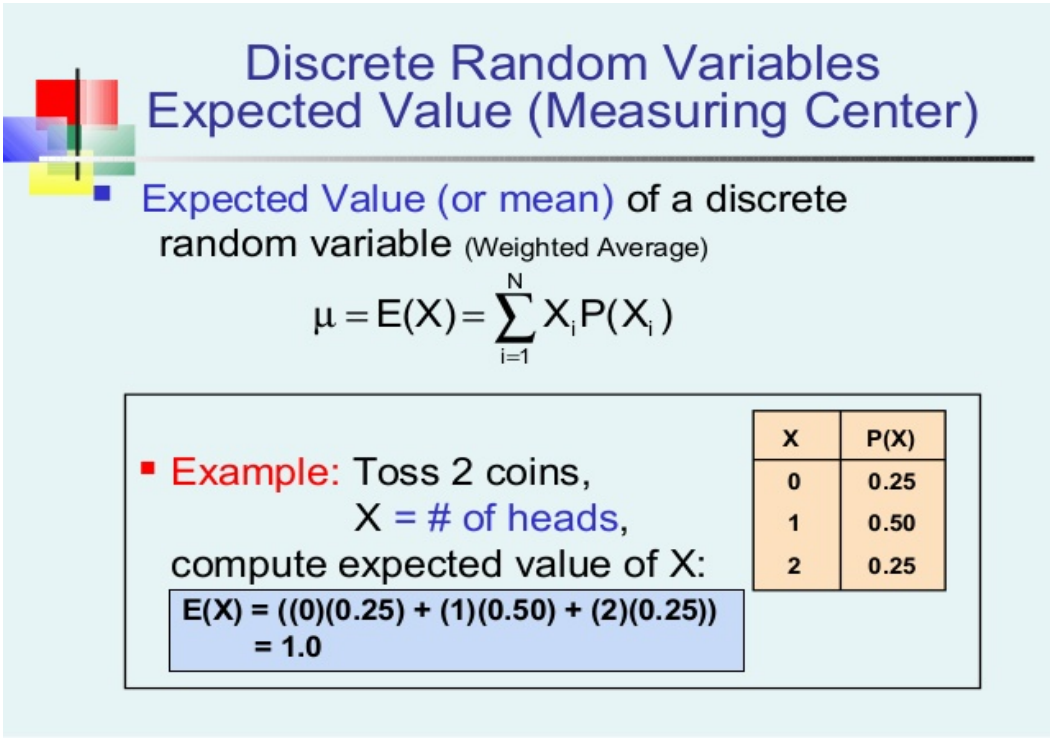
The standard deviation is

$$\sigma = \sqrt{\sigma^2}$$

Make sure:

- Every possible outcome is included in the sum
- You have a valid probability model to start off with

Example: Tossing coins



Example: A wheel comes up green 50% of the time and red 50% of the time. If it comes up green, you win $100, if it comes up red you win nothing. Intuitively, how much do you expect to win on one spin, on average?

E(X) = μ

= 100 x 0.5 + (0 x 0.5) = 50

The average would be winning $50 per spin

Outcomes	X = win	Probability P(X)
Green	100	0.5
Red	0	0.5

Example: Duracel, a company that sells batteries, claims that 99.5% of their batteries are in working order. How many batteries would you expect to buy, on average, to find one that does not work?

E (not working) = 1 - nxP(not working)

= n x 0.05 = 1

(Rearrange formula to get n) n = 1/0.05

n = 200

You would have to purchase 200 batteries on average to find one that isn't working.

Example: Your friend plans to toss a fair coin 200 times. You watch the first 40 tosses, noticing that she got only 16 heads. But then you get bored and leave. If the coin is fair, how many heads do you expect her to have when she has finished the 200 tosses?

200 - 40 = 160 tosses remaining

16 + (160 x 0.5) = 16 + 80 = 96 heads

Standard Deviation: Spread

Let X be a discrete rv with probability distribution P(X). The population variance of X is shown below and the population standard deviation of a random variable x is equal to the square root of its variance

Population Variance

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$$

Variance

- Population Variance: $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
- Where σ^2 means population variance,
- μ means population mean, and the other terms have their usual meaning.
- The variance is equal to the average squared deviation from the mean.
- To compute, take each score and subtract the mean. Square the result. Find the average over scores. Ta da! The variance.
- Think of this as the average squared distance from the mean. The farther scores are from the mean, the bigger the variance.

Find the population variance and standard deviation of X = number of heads observed tossing two coins.

2 = represents squaring, not multiplying by two in my answers

σ^2 = Var(X)

= (0-1)2 ¼ + (1-1)2 ½ + (2-1)2 ¼ = 1/2

σ^2 = SD(X)

= $\sqrt{\sigma^2}$ = 1/√2

Consider tossing an unbiased dice and recording the number on upper face X. Find the expected value, variance and standard deviation of the distribution of X

μ = E(X) = ∑ xi P(xi)

= 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 3.5

σ 2 = (1 – 3.5)2 1/6 + (2 – 3.5) 2 1/6 + (3 – 3.5) 2 1/6 + (4 – 3.5) 2 1/6 + (5 – 3.5) 2 1/6 + (6 – 3.5) 2 1/6

= 2.91666

σ^2 =1.7078

More about Means and Variances

Adding or subtracting a constant from data shifts the mean but doesn't change the variance or standard deviation: E(X ± c) = E(X) ± c Var(X ± c) = Var(X)

Example: The average midterm mark for this Statistics Class is 70% with a standard deviation of 10%. Consider everyone in this class receives an extra 5%. What will be the mean and standard deviation of the average midterm mark after the increase in mark?

X = original mark

S = new mark x+5%

E(S) = E(x+5%) = E(X) + 5% = 70% + 5% = 75%

Var (S) = Var (X + 5%) = Var (X) = (10%)2 = 100%2 (one hundred percent squared)

SD(s) = √ Var(S) = 10%

***sq = squared**

E(S) = E(1.05X) = 1.05 E(X) = 1.05 x 70% =

73.5%

Var(S) = Var(1.05X) = 1.05sqVar(X) = 1.05sq(10%)sq

S(D)S = √ Var(S) = 10.5%

Multiplying each value of a random variable by a constant multiplies the mean by that constant and the variance by the square of the constant:

E(aX) = aE(X)

Var(aX) = a2 Var(X)

Two Random Variables

The mean of the sum (difference) of two random variables is the sum (difference) of the means.

$E(X \pm Y) = E(X) \pm E(Y)$

Thus, $E(aX \pm bY) = aE(X) \pm bE(Y)$

If the random variables are **independent**, the variance of their sum or difference is always the sum of the variances.

$Var(X \pm Y) = Var(X) + Var(Y)$

$= a\ 2\ Var(X) + b2\ Var(Y)$

Remember:

- Variances of independent random variables add. Standard deviations don't.
- The mean of the sum or difference of two random variables, discrete or continuous, is just the sum or difference of their means.
- For independent random variables, the variance of their sum or difference is always the sum of their variances.

Example:

Given independent rv with means and standard deviations as shown, find the mean and standard deviation of each of the following:

	Mean	SD
X	5	3
Y	10	4

a) 2X

$E(X) = 2E(X) = 2(S) = 10$

$Var(2X) = 2sqVar(X) = 4Var(X) = 4(3sq) = 36$

SD (2x) = square root of VAR(2X) = 6

b) Y + 6

E(Y+6) = E(Y)+6 = 10 + 6 = 16

Var (Y+6) = Var(Y) = 4sq = 16

SD(y+6) = 4

c) X - Y

E(X - Y) = E(X) - E(Y) = 5 - 10 = -15

Var (X - Y) = Var(X) + Var(Y) = 3sq + 4cubed = 25

SD(X - Y) = 5

d) X1 + X2

E(x1 + x2) = E(x1) + E(x2) = 5 + 5 = 10

Var (x1 + x2) = Var(x1) + Var(x2) = 3sq + 3sq = 18

SD(x1 + x2) = $\sqrt{18}$

Chapter 15
Sampling Distribution Model

1) Population parameter is a quantitative measure such as the mean, median, mode, range, variance, or standard deviation that is calculated for a population data. It's also written with Greek letters. Eg. μ and σ.
    - Usually unknown and constant
2) A sample statistic is a summary measure calculated for a sample data set; it is written with Latin letters. Eg. y, s
    - Regarded as random before sample is selected
    - Observed after sample is selected
3) The value of the statistic changes from sample to sample, this is called sampling variability.
4) The distribution of all the values of a statistic is called its sampling distribution

∑ = sum

The table for distribution of x follows this outline

X	P(X)
X1	P(X1)
X2	P(X2)
X3	P(X3)
∑	1

Distribution of $\overline{X}$

$\overline{X}$	$P(\overline{X})$
$\overline{X}_1$	$P(\overline{X}_1)$
$\overline{X}_2$	$P(\overline{X}_2)$
…	….
∑	1

Population and Sample Proportions

If we are just interested in looking for one of the characteristics in which occurred in the population of interest, we will call the outcome we are looking for a "Success". The population proportion (P) is obtained by taking the ratio of the number of success divided by the number of the total elements in a population.

Examples for population proportions:

- Check for N students, how many are "nonresidents"
- Check for how many patients survived for at least five years after a cancer treatment

N = population size

Looking at a SRS of the size n from a large population, the probability p can be estimated by calculating the sample proportion (relative frequency) of Successes, $\hat{p}$. That is,

$\hat{P}$ = (number of successes in the sample / sample size)

Example for sample proportion:

- Flip n coins and observe if "Tail" was tossed.
- Look at n random persons and survey how many have an IQ above 120.
- Look at n random students and survey how many have more than two siblings.

Example: Suppose a total of 10,000 patients in a hospital and 7,000 of them like to play basketball. A sample of 200 patients is selected from this hospital, and 128 of them like to play basketball. Find the proportion of patients who like to play basketball in the population and in the sample.

10 000 = N

200 = n

7000 = count

128 = count

P = (7000/10000) = 0.7

$\hat{P}$ = (128/200) = 0.64

Find the sampling error for this case while assuming that the sample is random and no sampling error has been made.

Sampling error = $\hat{P}$ - P = 0.64 - 0.7 = -0.06

The Sampling Distribution of a Sample Proportion ($\hat{P}$)

Consider two different samples from a population, which you want to use for estimating the proportion of people with more than 2 brothers in the population. Use the statistic "proportion" for both samples. Are the outcomes the same? - Most likely not! This is known as sampling variability.

- Imagine we draw many samples and look at the sample proportions for these samples. 4 of 29 - The histogram we'd get if we could see all the proportions from all possible samples is called the sampling distribution of the proportions.
- What would the histogram of all the sample proportions look like?
- We would expect the histogram of the sample proportions to center at the true proportion, p, in the population.

There are three rules for sampling variability

1) $\mu\hat{P}$ is the mean of the sampling distribution of $\hat{P}$ equals p

$$\mu\hat{P} = P$$

2) The standard deviation of the sampling distribution of the sampling distribution is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

It can also be denoted as SD($\hat{P}$)

The standard error of proportion: $\hat{p}$ (p-hat)

- The standard error is an estimate of the standard deviation of a statistic.
- This is the formula of the Standard Error of an estimated proportion (the *hat* always represents an estimate)

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- $\hat{p}$ = estimated proportion
- n = sample (number of observations)

What about the shape of the sampling distribution of the proportions?

- It turns out that the histogram is unimodal, symmetric, and centered at p.

- More specifically, the sampling proportion $\hat{P}$ is approximately normal distributed for large n.

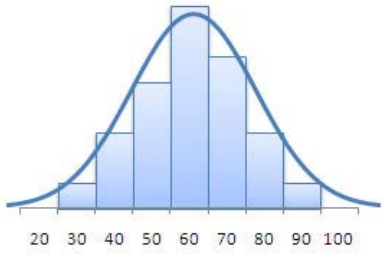
(Central Limit Theorem)

3)

Rules of Thumb and the Normal Approximation

- We can use the normal approximation for p-hat ONLY when $np \geq 10$ AND $n(1-p) \geq 10$. *Notice no hat on p!!
- We can use the formula for the standard deviation of p-hat only when the population is at least 10 times the sample size. In symbols, $population \geq 10n$.

This would be what a normal shape would look like is the sample is large enough



Example: (this is for proving the rules

Which Brand of Pizza Do You Prefer?

• Two Choices: A or D.

• Assume that half of the population prefers Brand A and half prefers Brand D.

• Take a random sample of n = 3 tasters.

Find the sampling distribution for the sample proportion. Find the mean and standard deviation

Sample No.	Prefer Pizza A	Proportion
(A,A,A)	3	1
(A,A,D)	2	2/3
(A,D,A)	2	2/3
(D,A,A)	2	2/3
(A,D,D)	1	1/3
(D,A,D)	1	1/3
(D,D,A)	1	1/3
(D,D,D)	0	0

Sample Proportion	Probability
0	1/8
1/3	3/8
1/3	3/8
1	1/8

$E(X) = \sum x\ P(X) = \mu$

$\mu\hat{P} = (0x1/8) + (1/3x3/8) + (⅔ \times ⅜) + (1x1/8) = 0.5$

$\mu\hat{P} = \hat{P}$

$SD(\hat{P}) = \sqrt{}\ (0 \times 0.05)sq1/8 + (⅓ - 0.5)sq\ ⅜ + (⅔ - 0.5)sq\ ⅜ + (1-0.5)sq1/8$

$\sqrt{}\ 0.083 = 0.288675$

Note: SD($\hat{P}$) = √ (0.5x(1-0.5))/3 = 0.288675, this is the same as the previous answer and our distribution is normal although it isn't larger than 10

n = 3

p = 0.5

n x p ≠ 10

Central Limit Theorem

The Central Limit Theorem (CLT)

Suppose we have a random variable X with expected value $E(X) = \mu$ and $v(X) = \sigma^2$

We extract n observation from X (say $\{x = x_1, x_2, ..., x_n\}$).

Lets define $\hat{X}_n = \frac{\sum_i x_i}{n} = \frac{x_1+x_2+...+x_n}{n}$.

$\hat{X}_n$ distributes with expected value μ and variance $\frac{\sigma^2}{n}$.

In case $n \longrightarrow \infty$ (in pratice $n > 30$)

$\hat{X}_n \sim N(\mu, \frac{\sigma^2}{n})$, whatever the distribution of x be.

N.B. If X is normal distributed, $\hat{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ even if $n < 30$

Basically, it states that under rather general conditions, means of random samples drawn from one population tend to have an approximately normal distribution. We find that it does not matter which kind of distribution we find in the population, it even can be discrete or extremely skewed, but if n is large enough the distribution of the mean is approximately normal distributed. That is, under all the possible distributions, we find one family of distributions that describes approximately the distribution of a sample mean, if only n is large enough.

Note:

1) If the original population is normal, then y bar is exactly normal distributed for any value of n, so that n does not have to be large
2) When the sampled population has a symmetric distribution, the sampling distribution of y becomes quickly normal
3) If the distribution is skewed, usually for n = 30 the sampling distribution is already close to a normal distribution

Assumptions and Conditions:

- If the original population is normal, then y is exactly normal distributed for any value of n, so that n does not have to be large
- When the sampled population has a symmetric distribution, the sampling distribution of y becomes normal
- If the distribution is skewed, usually for n = 30 the sampling distribution is already close to a normal distribution

Example:

Consider tossing n unbiased dice and recording the average number of the upper faces.

Means – The "Average" of One Die

The histogram shape of die starting with a simulation of 10,000 tosses of a die

1 die = uniform

2 dice = triangular

3 dice = mount-shaped

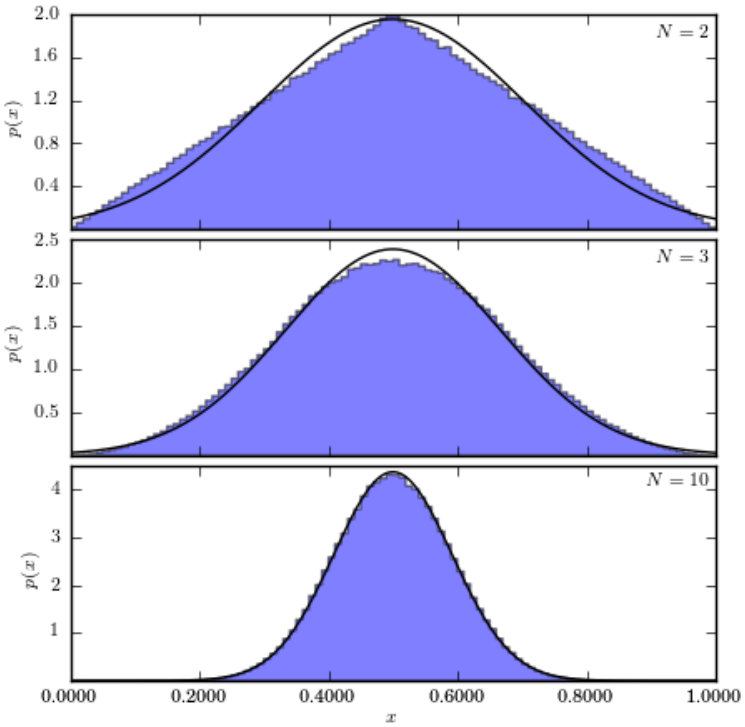
5 dice = fairly N

20 dice = N, very concentrated data, small $\sigma_{\overline{Y}}$

- As the sample size (number of dice) gets larger, each sample average is more likely to be closer to the population mean.
- So, we see the shape continuing to tighten around 3.5 And, it probably does not shock you that the sampling distribution of a mean becomes Normal
- The sampling distribution of any mean becomes more nearly Normal as the sample size grows.
- All we need is for the observations to be independent and collected with randomization. We don't even care about the shape of the population distribution!

The Fundamental Theorem of Statistics is called the Central Limit Theorem (CLT).

- The CLT is surprising and a bit weird:
- Not only does the histogram of the sample means get closer and closer to the Normal model as the sample size grows, but this is true regardless of the shape of the population distribution.
- The CLT works better (and faster) the closer the population model is to a Normal itself. It also works better for larger samples. With these sampling distributions, we can apply standardization in order to find the probability for the mean.



The Z value for y bar is

Probability of Y-bar, given μ

- To determine the probability associated with a particular value, convert to Z-scores
  - p(-1<Z<1) is.68, p(-2<Z<2) is.95, etc
- We use a slightly different Z-score formula than we learned before
  - But it is analogous

$$Z_i = \frac{(Y_i - \overline{Y})}{s_Y} \rightarrow \frac{(\overline{Y} - \mu)}{\sigma_{\overline{Y}}}$$

Example: The scores of students on the ACT college entrance exam has a normal distribution with a population mean of 18.6 and variation of 5.9

1) Take a mean of an SRS of 50 students who took the test. What are the mean and standard deviation of y and describe the shape of its sampling distribution?

a) $\mu_{\bar{y}} = \mu = 18.6$

b) $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 5.9/\sqrt{50} = 0.8344$

c) $Y \sim N = \bar{y} \sim N$

$\bar{y} \sim N (18.6, 5.9/\sqrt{50})$

2) What is the probability that the mean score $\bar{y}$ is 21 or higher?

$P(\bar{y} \geq 21) = P((\bar{y} - \mu_{\bar{y}})/ \sigma_{\bar{y}}) \geq ((21 - 18.6)/ 0.8344)$

$P(z \geq 2.88)$

$= P(Z \leq -2.88)$

Z - table = 0.0020

Note: $P(y \geq 21) \neq P(\bar{y} \geq 21)$

Example:

The duration of a disease from the onset of symptoms until death ranges from 3 to 20 years. The mean is 8 years and the standard deviation is 4 years. Looking at the average duration for 30 randomly selected patients, calculate the mean and standard deviation of $\bar{y}$ and describe the shape of its sampling distribution. What is the probability that the average duration of those 30 patients is less than 7 years?

Y ~ ? (8.4)

a) $\mu_{\bar{y}} = \mu = 8$

b) $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 4/\sqrt{30}$

c) Shape y ~ ? (n large so n $\geq$ 30) $\bar{y}$ ~ N

$P(\bar{y} < 7) = P((\bar{y} - \mu_{\bar{y}})/ \sigma_{\bar{y}}) < ((7-8)/4/\sqrt{30})) =$

$P(7 < -1.37)$

Z - table = 0.0853

*** A refresher:**

Probability of Y-bar, given μ

- Back to the problem: What is the Z-score associated with getting a sample mean of 27 or greater from this population?
  - Sampling distribution mean = 23
  - Standard error = 1.5

$$Z = \frac{(\overline{Y} - \mu)}{\sigma_{\overline{Y}}} = \frac{27 - 23}{1.5} = 2.66$$

Midterm review

Sample midterm questions:

1. A football league tests players for performance enhancing drugs. Officials put all players together and test randomly chosen players. What kind of sample is this?

A) Simple random sample

B) Stratified random sample

C) Voluntary response sample

D) Convenience sample

E) Systematic sample

2. A recent study examined the medical records of hospital patients and found that the incidence of Hepatitis C was twice as high among people who had tattoos or piercings than among people who do not have tattoos or piercings.

A) This is an observational study and causal inferences can be obtained from this study.

B) This is an observational study and causal inferences cannot be obtained from this study.

C) This is an experiment and causal inferences can be obtained from this study.

D) This is an experiment and causal inferences cannot be obtained from this study.

E) This is neither an observational study nor an experiment.

3. A sample was taken of the verbal GRE scores of 20 applicants to graduate school at a large midwestern university. Below are the scores. For convenience, the data are ordered.

280 310 340 350 370 410 420 420 420 470

490 510 520 520 600 610 670 720 750 770

Q1 = median of lower group = (370 + 410)/2 = 390

The first quartile for the applicant scores is

A) 340

B) 390

C) 410

D) 480

E) 605

4) The time to complete a standardized exam is approximately normal with a mean of 70 minutes and a standard deviation of 10 minutes. How much time should be given to complete the exam so that 80% of the students will complete the exam in the time given?

A) 84.0 minutes

B) 78.4 minutes

C) 92.8 minutes

D) 79.8 minutes

E) 87.5 minutes

$Y \sim (70,10)$

$P(Z<Z^*) = 0.8$

$Z^* = 0.84$ on z table

$Y^* = \mu + Z^*\sigma$

$= 70 + (0.84)(10)$

$= 78.4$

5) A random variable, Y, has a mean of 50 and a variance of 25. What is the variance of 2Y+3?

A) 10

B) 53

C) **100**

D) 25

E) 103

$E(aX+b) = aE(x) + b$

$Var(ax + b) = a^2var(x)$

$SD(ax+b) = aSD(x)$

$Var(2x+3) = Var(2Y)$

$2^2Var(Y)$

$4(25) = 100$

Note: $Var(2y+3) \neq 2Var(y) + 3$

6) In a certain lottery, you have a 70% chance of winning $5, a 25% chance of winning $10, and a 5% chance of winning $100. What is the expected value for your winnings? A) 38

B) 33

C) 11

D) 50

E) 75

X = winning	P(Winning)
5	70%
10	25%
100	5%
Sum =	1

E(x) = sum x P(X)

= (5 x 0.7) + (10 x 0.25) + (100 x 0.05) = 11

7) Suppose that, for any survivor of Oceanic flight 815, the probability of getting off the island they land on is approximately 42%. If each person's chances of getting off the island is independent of the others and we are only concerned with the fate of SIX survivors, what is the probability of more than 5 getting off the island?

A) 0.0510

B) 0.9945

C) 0.9619

D) 0.0055

E) 0.0131

P(more than 5) = P(all 6 gets off)

= P(O1∩O2∩)3…..∩)6)

P(O1) x P(O2) x P(O3)x……..xP(O6)

= 0.426

P(at least 1) = 1-P(none gets off)

1 - P(O1C ∩ O2C….. O6C)

1 - (1 - 042) = 0.9619

8) An urn contains 10 red chips, 6 black chips, and 2 blue chips. You are to randomly select 2 chips without replacement. You win the game if both chips are black. What is the probability that you win?

A) 0.111

B) 0.667

C) 0.098

D) 0.333

E) 0.227

$P(win) = P(B1 \cap B2) = P(B1) \times P(B2|B1) = 6/18 \times 5/17$

$= 0.0980$

9) The year is 1936 and Indiana Jones is trapped in a villain's death trap. His life randomly flashes before his eyes and he realizes he's been in death traps before. In fact, his amazingly good memory suggests the meantime he spends in death traps is 19.84 minutes with a standard deviation of 1.3 minutes. From 40 random death traps, what is the probability the sample mean is greater than 20?

A) 0.2216

B) 0.2177

C) 0.4522

D) 0.7823

E) 0.5478

$Y \sim ?\ (19.84, 1.3)$

a) $\mu_{\bar{y}} = \mu = 19.84$

b) $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 1.3/\sqrt{40} = 0.2055$

c) $Y \sim N$ doesn't equal $\bar{y} \sim N$ (CLT)

$P((\bar{y} - \mu_{\bar{y}})/ \sigma_{\bar{y}}) \geq ((20 - 19.84)/0.2055) = P(Z > 0.78)$

$P(Z < - 0.78) = 0.277$

10) Refer to the previous question, suppose “time spent in death traps” is normally distributed. If Doctor Jones considers a new random sample of 10 death traps, then the sampling distribution for the corresponding sample mean

A) is normal with a standard deviation of 0.411.

B) may not be normal because the Central Limit Theorem requires n to exceed 30.

C) is normal with a standard deviation of 1.3.

D) may not be normal with a standard deviation of 0.411.

E) may not be normal with a standard deviation of 1.3.

$Y \sim N(19.84, 1.3) =$

$\bar{Y} \sim N(19.84, 1.3/\sqrt{10})$

$N = 10$

11) Which of the following statements best describes the Central Limit Theorem?

A) In a sample from a population having a mean μ and a standard deviation σ, the respective mean and standard deviation of the sampling distribution of the sample mean are approximately μ and $\sigma/\sqrt{n}$.

B) When sampling from a normal population, the sample mean is always normally distributed.

C) If the sample size n is large enough, then the population being sampled can be assumed to have a normal distribution.

D) If the sample size n is large enough, then the sample is approximately normally distributed.

E) If the sample size n is large enough, then the distribution of the sample mean is approximately normally distributed.

12) Two sections of a class took the same quiz. Section A had 15 students who had a mean score of 80, and Section B had 20 students who had a mean score of 90. Overall, what was the mean score for all of the students on the quiz?

A) 84.3

B) 85.0

C) 85.7

D) None of these.

E) It cannot be determined

	n	y
A	15	80
B	20	90

$\bar{Y}$ Overall = $\sum y/n$ = (15x80) + (20x90)/ (15+20) = 85.71428571

GOOD LUCK!