

COMM 3720

STATISTICS AND EXCEL

PRIMER

McIntire School of Commerce
University of Virginia

Robert Parham w/ Robert Zhang, Joon Lee, John Sun

Table of Contents

1: Introduction and Basics	3
1.1 Statistics/ and Statistics	3
1.2 Descriptive Statistics	4
1.3 Measures of Location	6
1.4 Measures of Dispersion	9
2: Probability	11
2.1 Definitions	11
2.2 Probability Properties	11
2.3 Combinations and Permutations	12
2.3 Conditional Probability and Baye's Theorem	12
3: Random Variables and Probability Distributions	14
3.1: Random Variables	14
3.2 Probability Distributions and Cumulative Distribution Function	14
3.3 Discrete Probability Distributions	14
3.4 Continuous Probability Distributions	17
3.5 Discretizing a Continuous Probability Distribution	20
4: Expectation, Variance, Covariance, Correlation	21
4.1 Expectation and Variance Properties'	21
4.1 Correlation	21
4.2 Covariance	22
4.3 Properties	22
5: Testing and Confidence Intervals	23
5.1 Different Statistical Tests and When to Use Them	23
5.2 Confidence Intervals for a Population Mean	24
5.3 Confidence Intervals for Proportions	25
6: Regression Analysis	25
6.1 Linear Regression Basics	25
6.2 Interpreting a Linear Regression in Excel	25
Excel Tips and Tricks	27
1.1 General Shortcuts	27
1.2 Standard Formulas	27

1.3 Statistical Formulas	27
1.4 More Complex Formulas.....	27
1.5 Running Regression	29
1.6 Monte Carlo	30

1: Introduction and Basics

1.1 Statistics/ and Statistics

- What are statistics and what is statistics?
 - Statistics are: a number used to communicate a piece of information
 - Ex: GPA of 3.7 and other data
 - Statistics is: science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions
- Types of Statistics
 - Descriptive:
 - Methods of organizing, summarizing, and presenting data in an informative way
 - Ex: Grade distributions for classes
 - Inferential
 - Methods used to estimate a property of a population on the basis of a sample
 - Population: The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
 - Any measurable characteristic of a population is called a parameter. The population mean is an example of a parameter.
 - Sample: a portion, or part, of the population of interest
 - Any measurable characteristic of a sample is called a statistic. The sample mean is an example of a statistic.
 - Example: The US Census obtains information about the US' population while polls survey a smaller portion of the population.
- Types of Data
 - Qualitative data: Data that approximates and characterizes, non-numerical in nature.
 - Ex: Brand of laptop, hair color
- Important Definitions
 - Mutually Exclusive: events are mutually exclusive (disjoint) if they cannot both occur at the same time
 - Ex: When tossing a coin, heads and tails are mutually exclusive.
 - Independent: events are independent if the occurrence of one doesn't affect the probability of occurrence of the other
 - Ex: Landing a heads after flipping a coin and rolling a 2 on a 6-sided die are independent events.

1.2 Descriptive Statistics

- Frequency table: a grouping of qualitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class
 - Example: Frequency Table for Vehicles Sold at X Dealership by Location

Location	Number of Cars
A	5
B	7
C	2
D	6

Location was used to develop a frequency table with 4 mutually exclusive classes; in other words if a vehicle is sold at one location, it's not possible for it to have been sold at any of the other locations. Additionally, a frequency table must be collectively exhaustive, which means each vehicle is accounted for in the table.

- Relative Frequency: the relative frequency of a value is the fraction or proportion of times the value occurs
 - Example: You flip a coin 200 times with 70 heads and 130 tails
 - *frequency of the heads value* = 70
 - *relative frequency of the heads value* = $\frac{70}{200} = .35$
 - Example: Relative Frequency Table for Vehicles Sold at X Dealership by Location

Location	Number of Cars
A	.25
B	.35
C	.1
D	.3

- Frequency Distribution: a grouping of quantitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.
 - Example:
 - Table 2-4 presents a table showing the profit on vehicles sold last month by the Applewood Auto Group
 - Step 1: Decide on the Number of Classes
 - A useful guide is the “2 to the k rule.” We would want to select a k such that $2^k > \#$ of observations where k represents the number of classes
 - 180 vehicles were sold. In order to start off let k=7 which would result in $2^7 = 128$, which is less than our target of 180. When k=8, $2^8 = 256$, which is greater than 180 so the recommended number of classes is 8.

- Step 2: Determine the Class Interval
 - Generally, the class interval is the same for all classes. The classes together must cover at least the distance from the minimum value in the data to the maximum
 - $i \geq \frac{\text{Maximum value} - \text{minimum value}}{k}$
 - $i \geq \frac{3292 - 294}{8} = 374.75$ (we'll round to 400 for this example)
- After establishing these parameters, we should have a frequency distribution of:

Profit	Frequency
200 up to 600	8
600 up to 1000	11
1000 up to 1400	23
1400 up to 1800	38
1800 up to 2200	45
2200 up to 2600	32
2600 up to 3000	19
3000 up to 3400	4

- Frequency Distribution Presented as Histogram
 - Histogram: a graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to one another.
 - Differs from bar graph as the data is quantitative, which is measured on a continuous scale instead of a discrete one; thus the horizontal axis represents all possible values

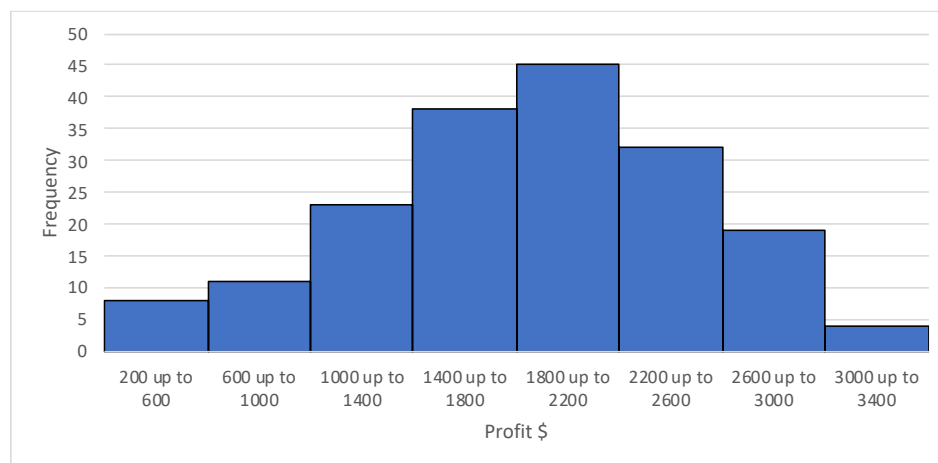


TABLE 2-4 Profit on Vehicles Sold Last Month by the Applewood Auto Group

\$1,387	\$2,148	\$2,201	\$ 963	\$ 820	\$2,230	\$3,043	\$2,584	\$2,370
1,754	2,207	996	1,298	1,266	2,341	1,059	2,666	2,637
1,817	2,252	2,813	1,410	1,741	3,292	1,674	2,991	1,426
1,040	1,428	323	1,553	1,772	1,108	1,807	934	2,944
1,273	1,889	352	1,648	1,932	1,295	2,056	2,063	2,147
1,529	1,166	482	2,071	2,350	1,344	2,236	2,083	1,973
3,082	1,320	1,144	2,116	2,422	1,906	2,928	2,856	2,502
1,951	2,265	1,485	1,500	2,446	1,952	1,269	2,989	783
2,692	1,323	1,509	1,549	369	2,070	1,717	910	1,538
1,206	1,760	1,638	2,348	978	2,454	1,797	1,536	2,339
1,342	1,919	1,961	2,498	1,238	1,606	1,955	1,957	2,700
443	2,357	2,127	294	1,818	1,680	2,199	2,240	2,222
754	2,866	2,430	1,115	1,824	1,827	2,482	2,695	2,597
1,621	732	1,704	1,124	1,907	1,915	2,701	1,325	2,742
870	1,464	1,876	1,532	1,938	2,084	3,210	2,250	1,837
1,174	1,626	2,010	1,688	1,940	2,639	377	2,279	2,842
1,412	1,762	2,165	1,822	2,197	842	1,220	2,626	2,434
1,809	1,915	2,231	1,897	2,646	1,963	1,401	1,501	1,640
2,415	2,119	2,389	2,445	1,461	2,059	2,175	1,752	1,821
1,546	1,766	335	2,886	1,731	2,338	1,118	2,058	2,487

Maximum

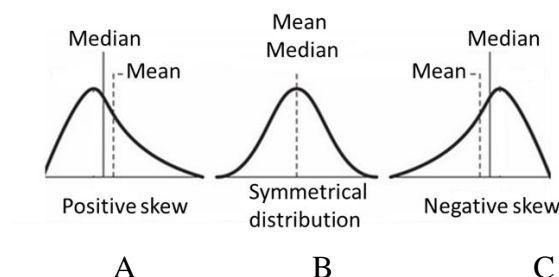
Minimum

1.3 Measures of Location

- Arithmetic Mean:
 - Sample Mean
 - $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$
 - Where:
 - \bar{x} represents the sample mean. Read as “x bar”
 - n is the number of values in the sample (also known as the sample size)
 - x represents any particular value
 - Σ is the Greek capital letter “sigma” and indicates the operation of addition
 - Σx is the sum of the x values in the sample
 - Population Mean
 - $\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$
 - Where:
 - μ represents the population mean. It is the Greek lowercase letter “mu”
 - N is the number of values in the population
 - x represents any particular value
 - Σx is the sum of the x values in the population
 - Properties of the Arithmetic Mean

- All values are included in computing the mean.
 - The mean is unique: there is only one mean in the data set.
 - The sum of the deviation of each value from the mean is zero.
 - $\sum(x - \bar{x}) = 0$
 - Ex: Mean of 0 and 2 is 1
 - $\sum(x - \bar{x}) = (0-1) + (2-1)$
 $= -1 + 1$
 $= 0$
- Weakness of the Mean
- While the mean is usually used to as a measure of center, its value is greatly affected by the presence of even a single outlier (unusually large or small observation).
 - For example, suppose company X has four employees who are paid 30,000, 20,000, 25,000, and 225,000. The mean of 75,000 is not representative of the group as all but one employee earns between 20,000 to 30,000.
- Median
- The median is the middle most value. If n observations were ordered from smallest to largest (with any repeated values included) then:
- If n is odd:

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ordered values}$$
 - If n is even
 - Median = arithmetic average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2} + 1\right)^{th}$ ordered values
- 50% of values should be above the median and 50% of values should be below the median
- Relative positions of mean and median



- A B C
- In Chart A, if a distribution is positively skewed, the mean is greater than the median due to the influence of a few extremely high observations.
 - In Chart B, a symmetrical distribution is shown which is when the distribution has the same shape on either side of the center. In this situation, the mean and the median are the same.

- In Chart C, if a distribution is negatively skewed, the mean is less than the median due to the influence of a few extremely low observations.
- In distributions that are skewed, the median tends to be a better representation of the “center”.
- Weighted Mean
 - The weighted mean is a convenient way to compute the arithmetic mean when there are several observations of the same value.

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n}{w_1 + w_2 + w_3 + \dots + w_n} = \frac{\Sigma(wx)}{\Sigma w}$$

Where:

n – set of numbers

w – weights or frequencies corresponding to the set of numbers

- Ex: Suppose company X pays its employees at a rate of \$16, \$12, or \$9 per hour. There are 6 hourly employees with 3 whom are paid at the \$9 rate, 2 paid at the \$12 rate, and 1 paid at the \$16 rate. What is the mean hourly rate paid to employees?
 - While this could be solved with the formulas under arithmetic mean it would be easier to solve as:

$$\bar{x}_w = \frac{3(\$9) + 2(\$12) + 1(\$16)}{6} = \$11.17$$

- Geometric Mean
 - The geometric mean is useful in finding the average change of percentages, ratios, indexes, or growth rates over time.
 - Defined as the nth root of the product of n values

$$GM = \sqrt[n]{(x_1)(x_2) \dots (x_n)}$$

More info: <https://www.statisticshowto.com/geometric-mean-2/>

1.4 Measures of Dispersion

While measures of location describe the center of the data, it doesn't give information on the spread of the data.

- Range
 - Difference between the maximum and minimum values in a data set.

Range = Maximum value – Minimum value

 - However, range is limited by the fact that it only takes into account these two values.
- Variance (Var)
 - Measures the arithmetic mean of the squared deviations from the mean

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element \bar{x} = sample mean n = sample size

- Standard Deviation (SD)

STANDARD DEVIATION FORMULA

The standard deviation formula can be represented using Sigma Notation:

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

sample standard deviation

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

population standard deviation

The standard deviation formula is the square root of the variance.

2: Probability

2.1 Definitions

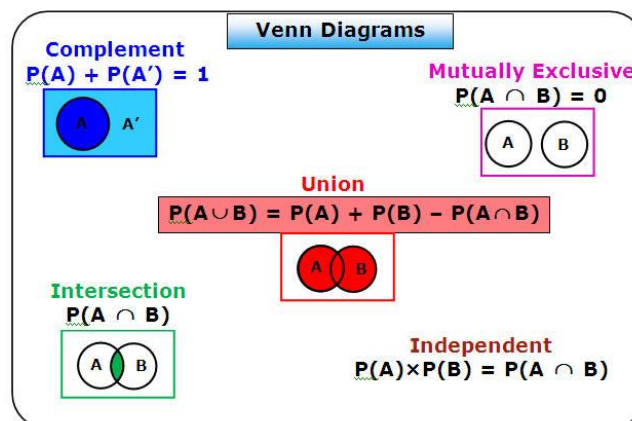
- Probability: a value between 0 and 1, describing the relative possibility (chance or likelihood) an event will occur
- Experiment: a process that leads to the occurrence of one and only one of several possible results
- Outcome: A particular result of an experiment
- Event: a collection of one or more outcomes of an experiment
- Sample Space: set of all possible outcomes of that experiment

Ex: Experiment: Roll a die
Outcome: 1
Sample Space: {1,2,3,4,5,6}
Event: Observe an even number

- Classical probability: based on the assumption that the outcomes of an experiment are equally likely.
 - $P(\text{Event}) = \text{Number of favorable outcomes} / \text{Total number of possible outcomes}$
- Empirical probability: the probability of an event happening is the fraction of the time similar events happened in the past
 - Empirical probability = Number of times the event occurs / Total number of observations
 - Law of Large Numbers: Over a large number of trials, the empirical probability of an event will approach its true probability.
 - Suppose we toss a fair coin. Suppose we flip it once and it lands on heads. Our current empirical probability of heads is 100% while the empirical probability of tails is 0%. However, as we continue to flip the coin, the empirical probability will start converging to 50% as that is the true probability of flipping either side on a fair coin.

2.2 Probability Properties

- Visuals to help provide context:




- For any event A, $1 \geq P(A) \geq 0$
- $P(\text{Sample Space}) = 1$
- $P(\text{Null event}) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ – General Rule of Addition
 - If A and B are mutually exclusive then:
 - $P(A \cup B) = P(A) + P(B)$
- $P(A) = 1 - P(A')$ where A' is “not A” (“not A” is also known as the “complement of A”)
- $P(A \cap B) = P(A)P(B)$ when A and B are independent events

For unknown symbols: <https://www.mathsisfun.com/sets/symbols.html>

2.3 Combinations and Permutations

- Permutation: Any ordered sequence of r objects taken from a set of n distinct objects is called a permutation of size r of the objects.
- Combination: given a set of n distinct objects, any unordered subset of size r of the objects is called a combination

$$C_{(n,r)} = \frac{n!}{r! (n-r)!}$$

$$P_{(n,r)} = \frac{n!}{(n-r)!}$$


n = set size:
the total number of
items in the sample

r = subset size:
the number of items to be
selected from the sample

2.3 Conditional Probability and Bayes' Theorem

- Conditional Probability: the probability of a particular event occurring given that another event has occurred:
 - Use the notation $P(A|B)$ to represent the conditional probability of A given the event B has occurred
- $P(A|B) = P(A \cap B)/P(B)$
- $P(A \cap B) = P(A|B) * P(B)$ – General Rule of Multiplication

- Bayes' Theorem

Assume that events A and A' are mutually exclusive and collectively exhaustive

$$P(A|B) = \frac{P(A) * P(B|A)}{P(A) * P(B|A) + P(A') * P(B|A')}$$

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

Example: Suppose a particular tests for whether someone has been using cannabis has a 90% true positive result and 80% true negative result. (90% of those with a positive result actually use cannabis while 80% of those with a negative result actually don't use cannabis) Assume that 5% of people actually use cannabis, what is the probability that a random person who tests positive is really a cannabis user?

$$P(\text{User}|\text{Positive}) = \frac{P(\text{Positive}|\text{User}) * P(\text{User})}{P(\text{Positive})}$$

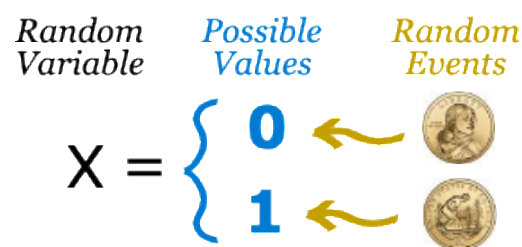
$$= \frac{P(\text{Positive}|\text{User}) * P(\text{User})}{P(\text{Positive}|\text{User}) * P(\text{User}) + P(\text{Positive}|\text{Non} - \text{User}) * P(\text{Non} - \text{User})}$$

$$= \frac{.9 * .05}{.9 * .05 + .2 * .95} \approx 19\%$$

3: Random Variables and Probability Distributions

3.1: Random Variables

- Random variable: a quantity resulting from an experiment that, by chance, can assume different values.
 - For a given sample space S of some experiment, a random variable is any rule that associates a number with each outcome in S . In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.
- Bernoulli random variable:
 - Any random variable whose only possible values are 0 and 1



- Discrete Random Variable
 - A random variable whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on.
- Continuous Random Variable
 - A random variable is continuous if its set of possible values consists either of all numbers in a single interval on the number line, or all numbers in a disjoint union of such intervals AND no possible value of the variable has positive probability, that is $P(X=c) = 0$ for any possible value c

3.2 Probability Distributions and Cumulative Distribution Function

- Probability distribution: a listing of all the outcomes of an experiment and the probability associated with each outcome
- Characteristics:
 - The probability of a particular outcome is between 0-1 inclusive.
 - The outcomes are mutually exclusive.
 - The list of outcomes is exhaustive. So, the sum of the probabilities of the outcomes is equal to one.

3.3 Discrete Probability Distributions

- Mean and Expectation
 - With probability distributions the mean is commonly referred to as the expectation of X where X is a random variable and is denoted $E(X)$

$$E(X) = \sum x_i P(x_i)$$

Where:

X: Random Variable x: a particular value of the RV P(x): probability of that value

- Variance

$$Var(X) = \sum (x_i - E(x))^2 P(x_i)$$

- Binomial Probability Function

- Binomial Probability Experiment

- The outcome for each trial is classified into one of two mutually exclusive categories – a success or a failure.
- The random variable is the number of successes in a fixed number of trials.
- The probability of success is the same for each trial.
- The trials are independent (the success of one doesn't affect the other).

- Binomial Probability Formula

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where:

$\binom{n}{x}$ denotes the combination n choose x

n is the number of trials

x is the random variable defined as the number of successes

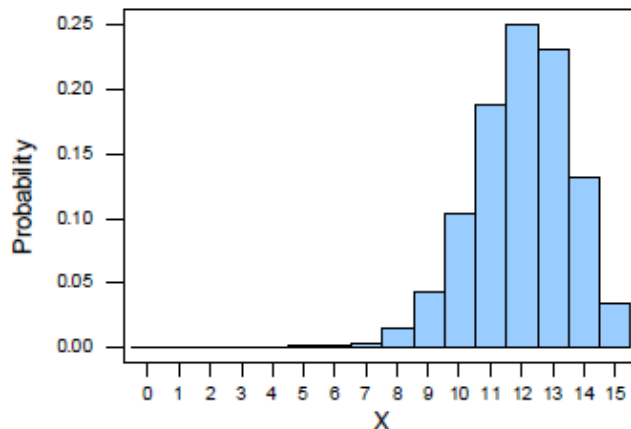
p is the probability of success on each trial

1-p is the probability of failure in a trial

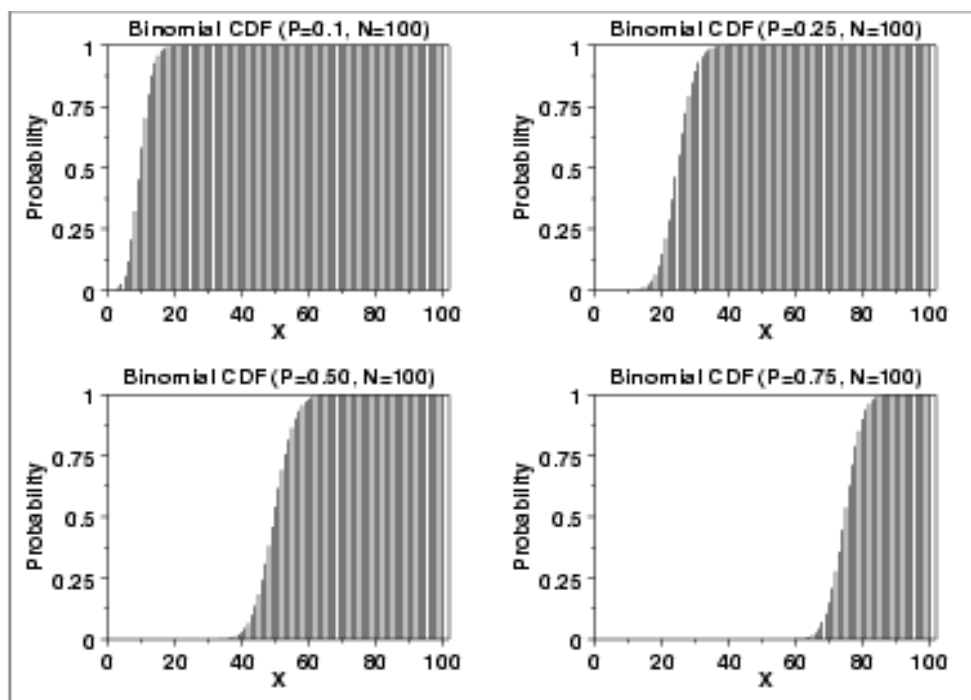
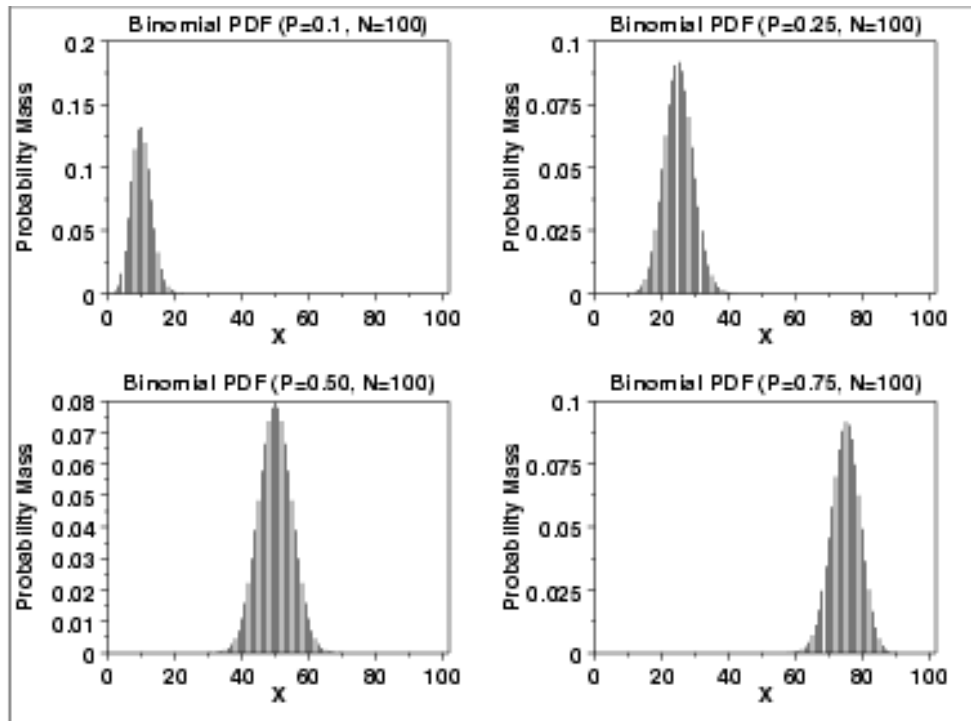
- Mean and Variance

- $E(X) = np$
- $Var(X) = np(1-p)$

Binomial distribution with n = 15 and p = 0.8



- Cumulative Distribution Function
 - Distribution function of the probability that the random variable will take a value less than or equal to.
 - Probability Distribution Function PDF vs. Cumulative Distribution Function CDF



3.4 Continuous Probability Distributions

- Expectation

$$E(X) = \int x_i P(x_i)$$

Where:

X: Random Variable x: a particular value of the RV P(x): probability of that value

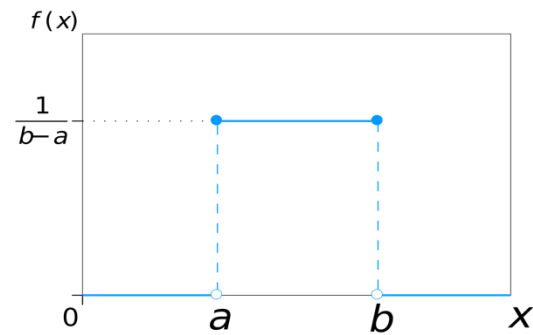
- Variance

$$Var(X) = \int (x_i - E(x))^2 P(x_i)$$

Note: Notice how in a discrete distribution we take the summation while in a continuous distribution we take the integral

- Uniform Distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$



- Probability Distribution $P(x) = 1/(b-a)$
- Notation $X \sim U(a,b)$
- Expectation $E(X) = (a+b)/2$
- Variance $Var(X) = \frac{(b-a)^2}{12}$

- Normal Distribution

- Function

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Notation $X \sim N(\mu, \sigma^2)$

- Expectation

$$E(X) = \mu$$

- Variance

$$\text{Var}(X) = \sigma^2$$

- Characteristics of a Normal Distribution

- Bell-shaped with a single peak at the center of the distribution: Mean = Median
 - Symmetrical about the mean
 - Distribution is asymptotic: Range = $(-\infty, \infty)$
 - Empirical Rule:

- ~68% of observations lie within 1 standard deviation of the mean
 - ~95% of the observations lie within 2 standard deviations of the mean
 - ~99.7% of observations lie within 3 standard deviations of the mean

- The Standard Normal Distribution

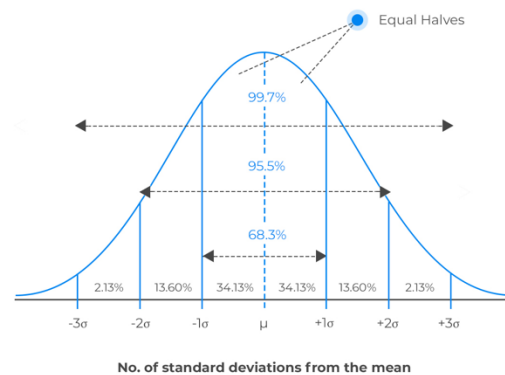
- $X \sim N(0,1)$
 - Any normal probability distribution can be converted into a standard normal probability distribution by obtaining the z scores

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

z-value expresses the distance between a particular value of x and the arithmetic mean in units of the standard deviation

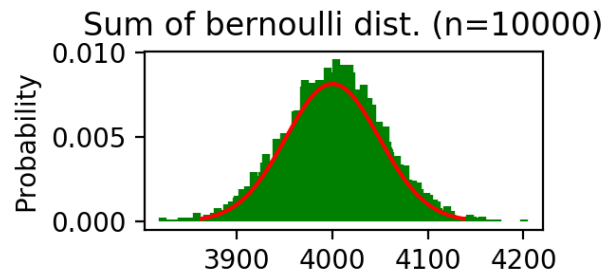
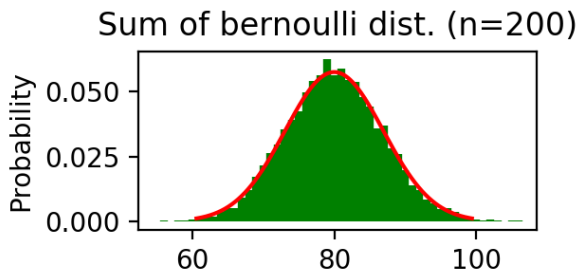
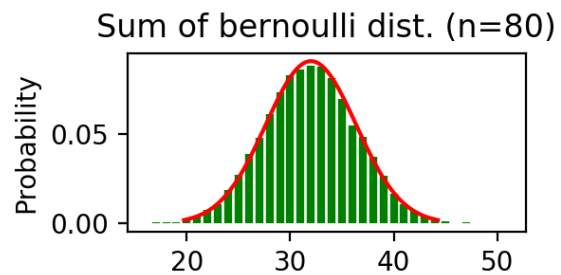
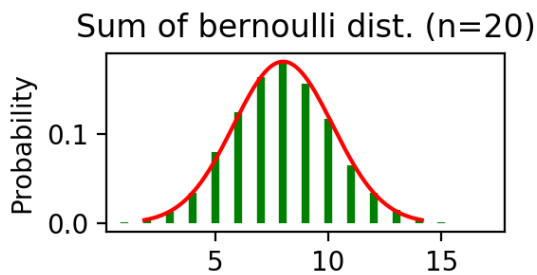
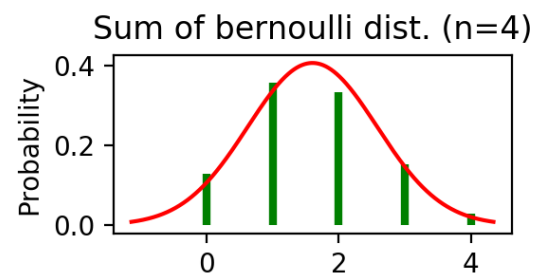
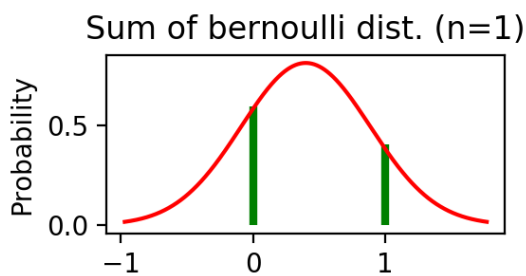


Shape of the normal distribution

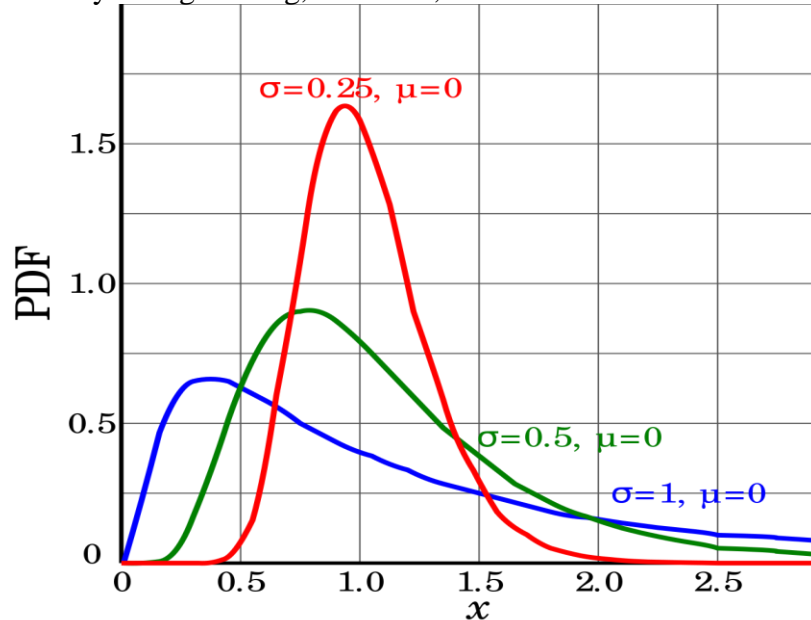


- Central Limit Theorem

- When random variables are independent and identically distributed random variables with the same distribution, zero mean, and variance, then the sum of many random variables approximates the normal distribution
- If you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately Normally distributed
- CLT with a Bernoulli Random Variable



- Log-Normal Distribution
 - Used extensively in engineering, medicine, and finance



$$P(x) = \begin{cases} \frac{1}{\sigma x \sqrt{2\pi}} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} & \text{when } x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

- Expectation $E(X) = e^{\mu + (\frac{\sigma^2}{2})}$
- Variance $\text{Var}(X) = e^{2\mu + \sigma^2} \cdot (e^{\sigma^2} - 1)$

3.5 Discretizing a Continuous Probability Distribution

- You are given a continuous probability distribution for a random variable X , e.g.:
 - X is uniform between values A and B , denoted $X \sim U[A, B]$
 - X is Normal, with given mean and variance, denoted $X \sim N(\mu, \sigma^2)$
- You are then asked questions about some function of the random variable, $Y(X)$, e.g.:
 - $Y = X^3$
 - $Y = \sin(X)$
 - $Y = \max(0, c \cdot (X - d))$, for some values c, d
- To calculate using Excel, you need to make the continuous probability distribution discrete (as computers are discrete, not continuous). This is known as “discretizing the distribution.”
- The basic recipe is:
 - Elect a (large but not infinite) group of possible outcomes for X
 - Assign each one of them an appropriate probability
 - Assume that these are the only possible outcomes
- There is of course much care in choosing the “right” points, and assigning them the “right” probability.

- For the uniform, it is common to assume that only whole numbers between A and B are possible outcomes, assuming A and B are “far enough” from each other, and there are enough whole numbers between them. This gives us the group of possible outcomes.
- For the Normal, it is common to assume the whole numbers from $\mu - 4 \cdot \sigma$ to $\mu + 4 \cdot \sigma$ are all the possible outcomes (again, assuming there’re enough of them).
- For each possible outcome, the probability is the CDF around that outcome. By around we mean from half-way between this outcome and the previous one, to half-way between this outcome and the next one.
- So, e.g., to discretize $X \sim N(500, 100)$:
 - We assume possible outcomes are all whole numbers from 100 to 900
 - Each number n has probability of happening given by $\text{CDF}(n+0.5, 500, 100) - \text{CDF}(n-0.5, 500, 100)$
 - In Excel, the above can be implemented like so:

$$P(X = n) = \text{NORM.DIST}(n+0.5, 500, 100, \text{TRUE}) - \text{NORM.DIST}(n-0.5, 500, 100, \text{TRUE})$$

for all integers in $100 \leq n \leq 900$

4: Expectation, Variance, Covariance, Correlation

4.1 Expectation and Variance Properties'

- Expectation is just a weighted average, with the weights being the probabilities.
- Expectation Properties
 - $E(a) = a$ – (where a is a constant)
 - $E(X) = E(X)$
 - $E(E(X)) = E(X)$
 - $E(X+Y) = E(X) + E(Y)$
 - $E(X-Y) = E(X) - E(Y)$
 - $E(XY) = E(X) \cdot E(Y)$ only if X and Y are independent
 - $E(aX+bY) = aE(X) + bE(Y)$
- Variance Properties
 - $\text{Var}(X) = E[(X-E(X))^2] = E(X^2) - (E[X])^2$
 - $\text{Var}(a) = 0$
 - $\text{Var}(aX) = a^2 \text{Var}(X)$
 - When X and Y are independent $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$
 - $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X, Y)$
 - $\text{Var}(aX+bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab\text{Cov}(X, Y)$

4.1 Correlation

- A quantitative measurement of the relationship between two variables

- Correlation is between -1 and 1 inclusive
 - The closer the correlation is to -1 or 1 the stronger it is

$$\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

- If two variables are independent then their correlation will be 0. However, a correlation of 0 does not imply independence

4.2 Covariance

- Quantitative measure of the extent to which the deviation of one variable from its mean matches the deviation of the other from its mean
- $\text{Cov}(X, Y) = E[XY] - E[Y]E[X]$

4.3 Properties

- Covariance Properties
 - $\text{Cov}(X, X) = \text{Var}(X)$
 - $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
 - $\text{Cov}(X, a) = 0$
 - $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$
 - $\text{Cov}(X+a, Y+b) = \text{Cov}(X, Y)$
 - $\text{Cov}(X+Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$
 - $\text{Cov}(aX+bY, cW+dV) = ac \text{Cov}(X, W) + ad \text{Cov}(X, V) + bc \text{Cov}(Y, W) + bd \text{Cov}(Y, V)$
- These are very important in finance, and you should probably be very comfortable with them and memorize them.

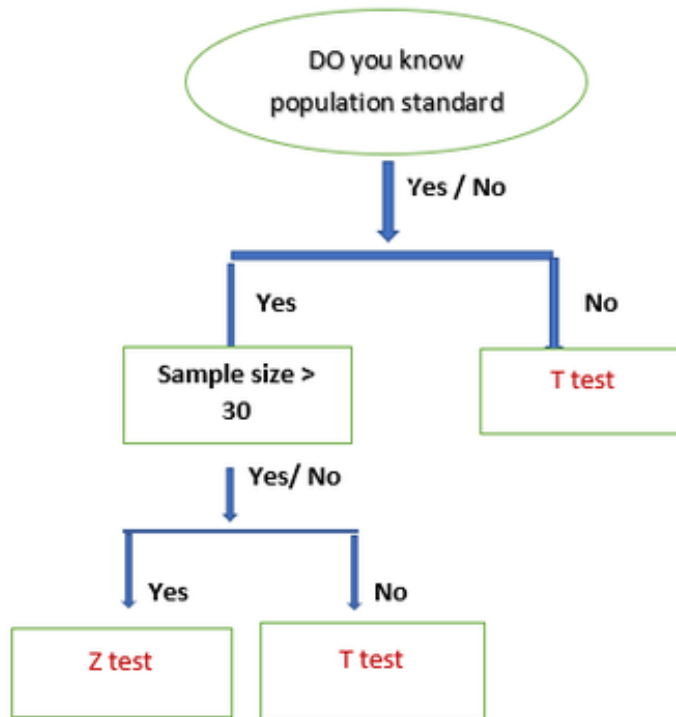
5: Testing and Confidence Intervals

5.1 Different Statistical Tests and When to Use Them

Different Hypothesis Tests

	Hypothesis Test	Underlying Distribution	Purpose
Parametric (Assumes the data subscribes to a distribution)	1 Sample t-Test	Normal	Compares one sample average to a historical average or target
	2 Sample t-Test	Normal	Compares two independent sample averages
	Paired t-Test	Normal	Compares two dependent sample averages
	Test for Equal Variances	Chi-square	Compares two or more independent sample variances or standard deviations
	1 Proportion Test	Binomial	Compares one sample proportion (percentage) to a historical average or target
	2 Proportion Test	Binomial	Compares two independent proportions
	Chi-square Goodness of Fit	Chi-square	Determines whether a data set fits a known distribution
	Chi-square Test for Independence	Chi-square	Determines whether probabilities classified for one variable are associated with the classification of a second
Non-Parametric (Makes no assumption about the underlying distribution of the data)	1 Sample Sign Test	None	Compares one sample median to a historical median or target
	Mann-Whitney Test	None	Compares two independent sample medians

- Z test v. T test



5.2 Confidence Intervals for a Population Mean

- Confidence Interval: a range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. The specified probability is called the level of confidence.
- Understanding the confidence level
 - When we have a 95% confidence interval what it means is if we continued to draw samples and construct confidence intervals, 95% of the intervals constructed would contain the true population parameter.
- Population Standard Deviation Known

$$CI = x \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Looking back at the standard normal distribution, z represents the z -score of the confidence level. Ex: When level of confidence = 95%, $z = 1.96$

- Population Standard Deviation Unknown

$$CI = x \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

When the population standard deviation is unknown we need to use the t-distribution. The t stands for the t statistic for the confidence level.

5.3 Confidence Intervals for Proportions

Sample proportion $p = x/n$

$$CI = p \pm z \sqrt{\frac{p(1-p)}{n}}$$

6: Regression Analysis

6.1 Linear Regression Basics

Simple Linear Regression Formula: $\hat{y} = a + \beta_1 x_1$

General Formula: $\hat{y} = a + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$

6.2 Interpreting a Linear Regression in Excel

Regression Statistics								
Multiple R	0.9830972							
R Square	0.9664801							
Adjusted R Square	0.95307214							
Standard Error	0.36591637							
Observations	8							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	19.3029608	9.6514804	72.0825689	0.00020571			
Residual	5	0.66947395	0.13389479					
Total	7	19.9724348						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.330639	0.63031834	0.52455875	0.62230647	-1.2896459	1.95092389	-1.2896459	1.95092389
X1	5.9039E-06	6.4124E-07	9.20702712	0.0002537	4.2555E-06	7.5523E-06	4.2555E-06	7.5523E-06
X2	0.51639636	0.31543629	1.63708608	0.16253898	-0.2944584	1.32725115	-0.2944584	1.32725115

- The above table gives an example of a regression output from excel
- Multiple R:
 - Also known as the correlation coefficient, it is a measure of linear association between the variables
 - Want the absolute value of your correlation coefficient to be close to 1
- R Square:
 - Measures the proportion of the variance for a dependent variable that is explained by the independent variable(s)
 - Want the absolute value of your r-squared to be close to 1
 - Is the square of the correlation coefficient
 - Ex: In the above model 96.6% of the variation in our dependent variable can be explained by the variation in variables X1 and X2.
- Adjusted R Square
 - Modified version of r-squared that has been adjusted for the number of predictors in the model. (R-squared can increase with the number of variables added to your model but that doesn't necessarily mean that your model is better. It might just be better at predicting values with the data set given. Adjusted R-square serves as a better measure of proportion of variance when there are many independent variables)
- Standard Error
 - A measure of the dispersion of the observed values around the line of regression for a given value of x
 - Generally favor models with lower standard errors
- ANOVA
 - Tells more about the levels of variability within the model
 - Significance F: generally want one less than .05
- Bottom Table
 - Coefficients
 - Looking at the formula in 6.1, these coefficients would plug into the betas in order to get your regression equation
 - Intercept would plug into a
 - X1 would plug into the beta associated with that variable
 - X2 would plug into the beta associated with that variable
 - P-Value
 - In order to make sure your model is valid, make sure all of the p-values for the variables are below the threshold (.05 generally).
 - In this model since X2 has a p-value of .16 associated with it, it would be best to take that variable out and create a new model with just X1.
- Order of relevance
 - Make sure all p-values are less than .05
 - Try to maximize your r-square or adjusted r-square
 - Minimize your standard error

Excel Tips and Tricks

1.1 General Shortcuts

Shortcut	Windows	Mac
Select Large Groups of Data	Ctrl + shift + (up, down, left, right)	Cmd + shift + (up, down, left, right)
Add column/row	Ctrl + shift + +	Cmd + shift + +
Go to extremes of your data	Ctrl + (up, down, left, right)	Cmd + (up, down, left, right)

Extend a formula down: Double click the bottom left of the box which contains the formula

1.2 Standard Formulas

- =average() Gives average of a selection
- =sum() Gives sum of a selection
- =product() Gives product of a selection
- =min() Gives minimum of a selection
- =max() Gives maximum of a selection
- =if(condition,if true, if false) Returns different values based on if condition
- =count() Counts the number of values in a selection
- =percentile(g1,k) Gives the value of the kth percentile in g1

1.3 Statistical Formulas

- =var.s() Gives variance of a selection that is a sample
- =var.p() Gives variance of a selection that is a population
- =stdev.s() Gives standard deviation of a selection that is a sample
- =stdev.p() Gives standard deviation of a selection that is a population
- =covariance.s(g1,g2) Gives covariance between g1 and g2 (sample)
- =covariance.p(g1,g2,) Gives covariance between g1 and 2 (population)
- =correl(g1,g2) Gives correlation between g1 and g2

1.4 More Complex Formulas

- =index(array, row_num, col_num, area_num)
 - Returns the value at a given location
 - Select the array, and then select row position (and column position)
- =match(lookup_value, Lookup_array, match_type)
 - Returns a number representing a position in lookup_array
 - Select value you want to look up, select array you want to look it up in, and then select whether you want an exact match or not
- =vlookup(lookup_vale, table_array, col_index_num, [Range lookup])
 - Select the value you want to look up
 - Select the array in which you want to look up said value
 - From that array, select which column's information you would like returned
 - Approximate or exact match

- =sumproduct(g1,g2)
 - Returns the sum of the products between g1 and g2. A.k.a. the dot product operator.
 - Ex: g1 = 1,2 g2 = 3,4
 - Sumproduct(g1,g2) = 3 + 8 =11
- =slope(y_values,x_values)
 - Returns the slope of a regression line based off the data points
- =mmult()
 - Multiplies two matrices together
 - Need to click ctrl + shift + enter in order to use

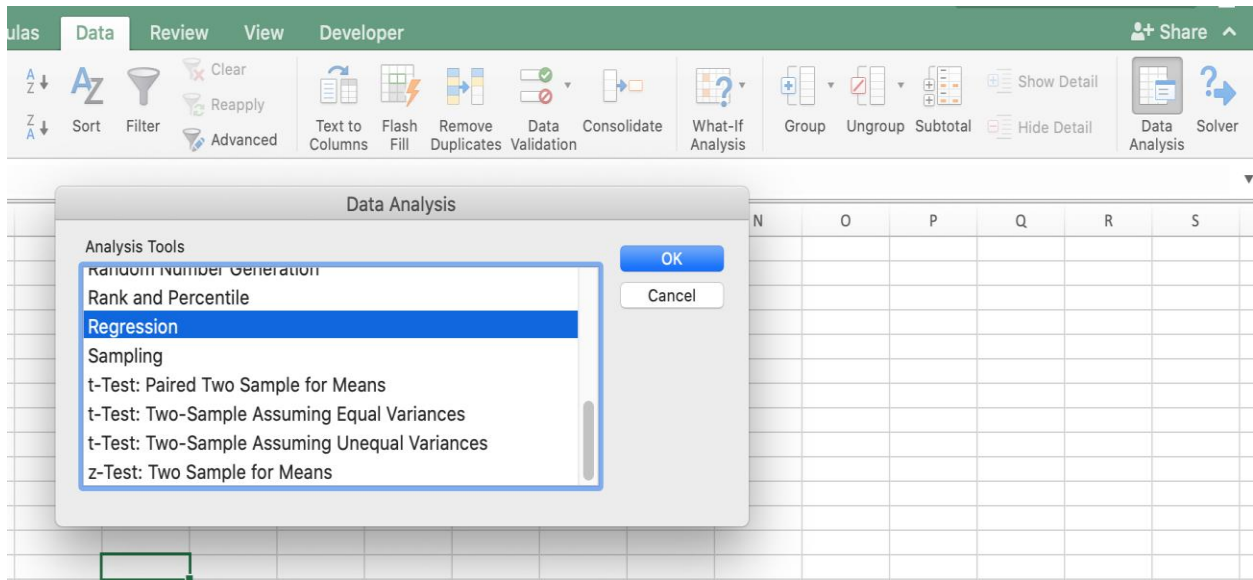
MMULT(array1, array2)

array1 (2 x 3)	array2 (3 x 2)	array result (2 x 2)
0 3 5	3 4	29 -16
5 5 2	3 -2	38 6
	4 -2	

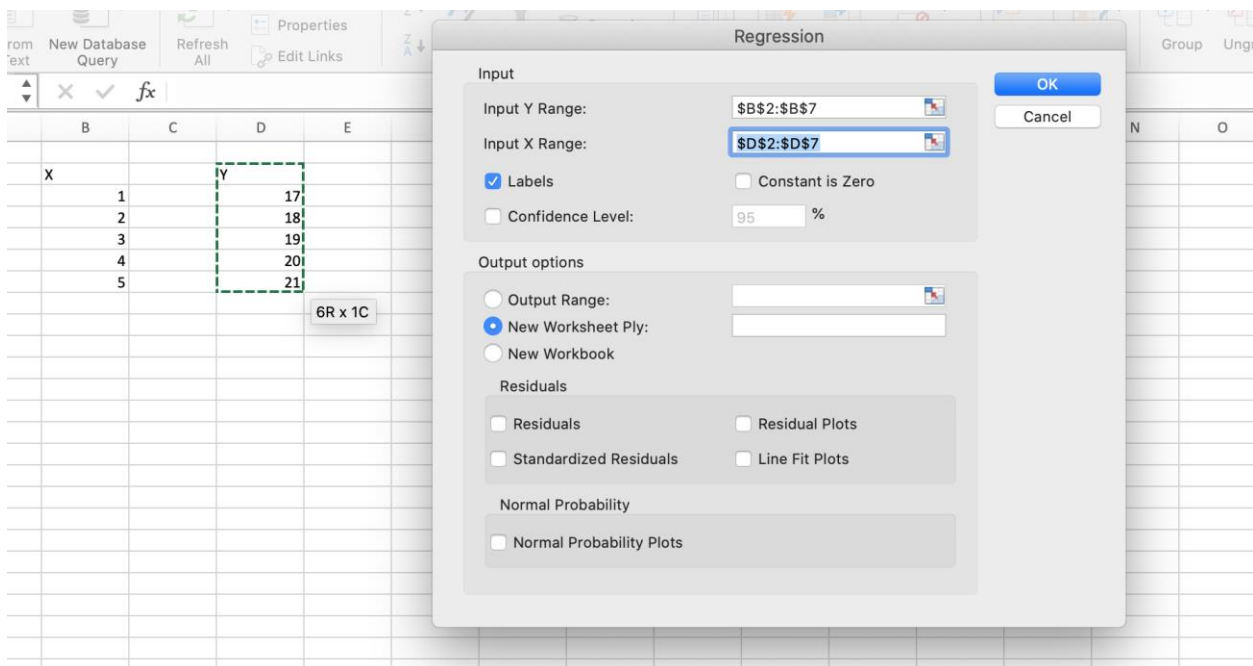
Matrix Multiplication rules: <https://www.mathsisfun.com/algebra/matrix-multiplying.html>

- =norminv(probability, mean, standard_dev)
 - Returns the inverse of the normal cumulative distribution with the specific mean and standard deviation
 - Ex: norminv(.5,10,2) will return 10
- =norm.dist(x, mean, standard_dev, cumulative)
 - Returns the normal distribution for the specified mean and standard deviation (returns the probability of that value)
 - Specify True in cumulative to return the CDF and False in order to return the PDF

1.5 Running Regression



- Go to options -> add-ins in order to enable Data Analysis
- Click on the Data tab in the top and then click Data Analysis
- Find the function you want in this case regression.

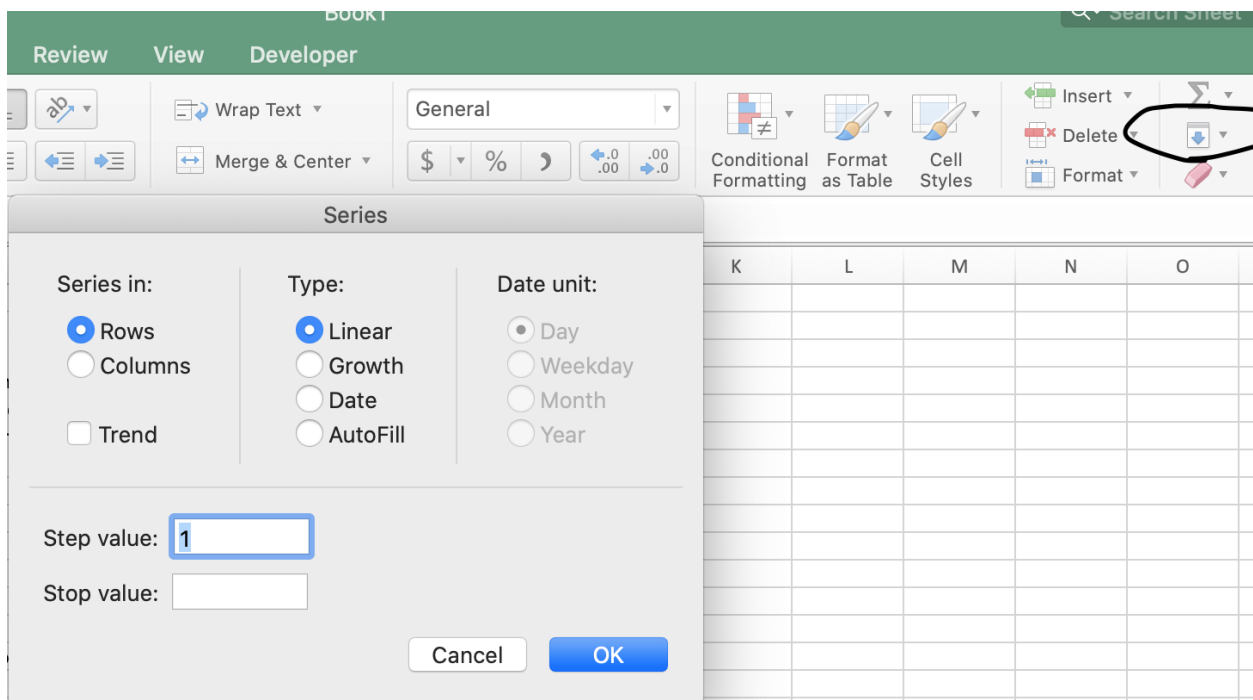


- Put the X and Y Values in their respective ranges.
 - X range can have more than one column for multiple regression
- Can also set a confidence level and/or add residuals

1.6 Monte Carlo

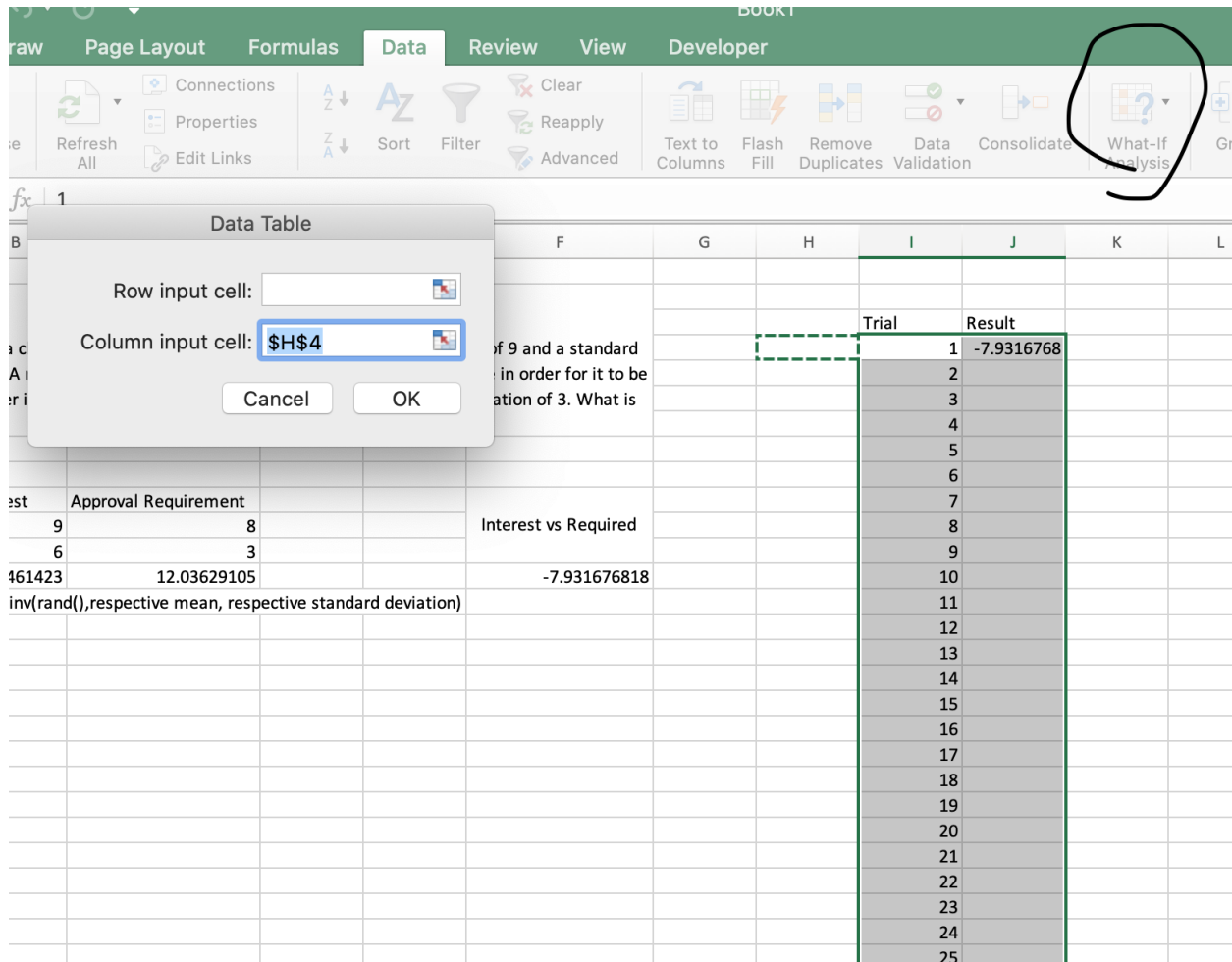
Problem: Will My Class Fill?					
Suppose I (the TA) teach a class where the sign up is normally distributed with a mean of 9 and a standard deviation of 6. Suppose UVA requires a certain number of students to sign up for a course in order for it to be approved and that number is normally distributed with a mean of 8 and a standard deviation of 3. What is the likelihood that my class will be approved?					
	Interest	Approval Requirement			
Mean	9	8			
Standard Deviation	6	3			
First Simulation	-5.7799502	7.216782918			
Formula	norminv(rand()),respective mean, respective standard deviation)				

- To get our values for Monte Carlo we'll be using the formula norminv(rand()),respective mean, respective standard deviation)
 - This essentially gets a random probability from 0-1 and converts it into the respective value based off of the normal distribution given
- We're interested in the difference between the interest and the approval requirement in our first simulation



- Create a column named Trial and another one next to it named result

- To populate the Trial column, select the button circled and click series.
 - This should bring you to this tab and put the number of trials you're running as your stop value, additionally be sure to select the column option in "Series in:". In this example we'll be using the number 1000.



- As you can see in the picture we created a cell from our first simulation that gave the difference between the interest and the amount of interest required.
- Make the first trial equal to that cell
- Then highlight from trial 1 and the first result all the way to the end of your trials (in our case 1000)
- Go into data -> what if analysis -> Data Table
- Select an empty cell for the column input cell and leave the row input cell blank. Click OK.
- Then in another cell use the formula $\text{=countif}(\text{Range of cells relevant}, ">0")/1000$
 - This will give you the likelihood that the class will have enough interest to be approved

- Further analysis can be conducted on things such as min, max, and other measures of dispersion.