Textbook: LINK

1.1: An Overview of Statistics

-Definitions
- -Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.
- -Data consists of information coming from observations, counts, measurements, or responses.

-2 Data Sets
- -Population: the collection of all outcomes, responses, measurements, or counts that are of interest.
  - -i.e. All Tacoma households. (Usually “all” & “each”)
- -Sample: A subset, or part, of a population
  - -i.e. 5000 Tacoma households. (Usually a specific number)
  - -There can be bias in sample selection, so watch out.
  - -Samples are necessary as getting stats for a population can be too hard.
- -Ex:
  - -”The number of wireless devices in each U.S. household.”
    - -the population is each U.S. household.
  - -”A survey of 300 people from an auditorium with 13,000 people.” is a sample.
    - -The sample is 300 (out of 13,000) people.
    - -The population would be 13,000 people
  - -”A survey of 186 U.S. adults ages 25 to 29 found that 76% have read a book in the past 12 months”
    - -The sample is 186 U.S. adults (out of the whole US) ages 25 to 29.
    - -The population is “All U.S. adults ages 25 to 29” here.

-Parameters VS Statistics
- -A parameter is a numerical description of a population's characteristics.
- -A statistic is a numerical description of a sample's characteristics.
- -All Parameters & Statistics Symbols
  - -Population Mean (μ) is a paramter.
  - -Sample Mean ($\bar{x}$) is a statistic.
  - -Population Variance (σ^2) is a parameter.
  - -Sample Variance (s^2) is a statistic.
  - -Population Correlation (ρ^2) is a parameter.
  - -Sample Correlation (r^2) is a parameter.
  - -Population Proportion (p) is a parameter.
  - -Sample Proportion ($\hat{p}$) is a parameter.
- -Ex:
  - -”The average salary for 45 of a consulting firm’s 300 engineers is $72,000.” means $\bar{x}$ = $72,000.
  - -”In a recent year, the average math score on the ACT for all graduates was 20.6.” means μ = 20.6 points.

-Branches of Statistics
- -Descriptive Statistics: The organization, summarization, and display of data.
- -Inferential Statistics: Using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.
  - -Probability is basically how confident you are about your inference.
- -Ex: “A survey of 186 U.S. adults ages 25 to 29 found that 76% have read a book in the past 12 months.”
  - -”76% of U.S. adults have read a book in the past 12 months” is an ex of Descriptive Statistics.
  - -Inferential Statistics then infers/generalizes that all U.S. adults from ages 25 to 29 (population) have read a book in the past 12 months.

1.2: Data Classification

-Types of Data
- -Qualitative (aka Categorical): Consist of attributes, labels, or non-numerical entries.
- -Quantitative: Consist of numbers that are measurements or counts.
  - -Sometimes, data with numbers can actually be qualitative data. Unless you can perform math on it, it’s qualitative.
    - -i.e. can’t say 98404 < 98444 if they are zip codes
- -Ex:
  - -Hair color of classmates: Qualitative, attributes.
  - -Zipcodes: Qualitative, labels.
  - -Weights of dogs: Quantitative, measurements
  - -Student ID#: Qualitative, labels.
  - -Wait time at a DMV: Quantitative, measurements

-Level of Measurements
- -Nomial (Qualitative) [weakest]
  - -No mathematical significance
  - -Ex: hair color, zipcodes, cardinal direction, political parties
- -Ordinal (Qualitative & Quantitative)
  - -Can be ordered but the difference between data entries is not meaningful.
  - -Ex: computer case sizes, the doctor asking patient for pain scale, hurricane-level
    - -There is a difference between entries, but it is subjective or not clearly defined.
- -Interval (Quantitative)
  - -Can be ordered & the difference between entries has mathematical meaning. But, the zero value only represents a position on a scale.
  - -Ex: temperature, test scores, 10-meter platform scores
    - -0*C is different than 0*F. A 0 test score doesn’t mean the student doesn’t know anything.
- -Ratio (Quantitative) [Strongest]
  - -Can be ordered, the difference between entries has mathematical meaning, & zero is inherently zero/nothing/none.
  - -Ex: age, $, number of performances of a show, mass
    - -Inherent Zero: $0 means no money. 0 age means not born yet.

2.1: Visual Display for Univariate/One Data

-Histogram (Quantitative Data)
- -A frequency distribution is a table that shows classes or intervals of data entries with a count of the number of entries in each class.
  - -The frequency “f” of a class is the number of data entries in the class.

1. Class width has to be uniform
2. Classes limits do not overlap ********** Mutually exclusive ******
3. Classes cannot be omitted, even if the frequency is zero
4. Number of classes is 5 to 20.

- -Ex: Years of service of 28 Ohio government employees using 5 classes (sub-ranges).

13	8	10	14	19	15	13	16	10	11	7	9	14	13
11	12	16	15	13	10	18	17	10	15	16	9	15	19

  - -Range = max - min = 19 - 7 = **12**

-Width of Class = ceiling(Range / # of Classes) = ceiling(12 / 5) = ceiling(2.4) = **3**

-This tells you the distance between lower (or upper) limits of consecutive limits.

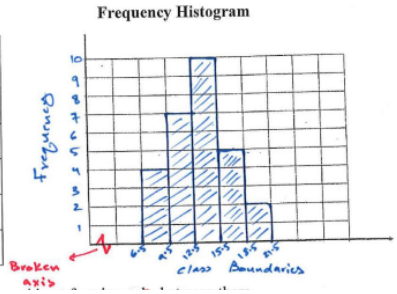
Class	Tally	Frequency
7 – 9	||||	4
10 – 12	|||| ||	7
13 – 15	|||| ||||	10
16 – 18	||||	5
19 – 21	||	2

-Splitting up the x-axis on graph to match classes (Class Boundaries)

-Lower class boundary = lower class limit - 0.5

-Upper class boundary = upper class limit - 0.5

#	Class limit	Frequency *y*-axis	Class boundaries *x*-axis
1	7 – 9	4	6.5 – 9.5
2	10 – 12	7	9.5 – 12.5
3	13 – 15	10	12.5 – 15.5
4	16 – 18	5	15.5 – 18.5
5	19 – 21	2	18.5 – 21.5



-An Interpretation of the data:

-The increase in the number of employees reached a **peak** at about 14 years of service.

-About **1/3** of the employees served between 13 and 15 years.

-R Code

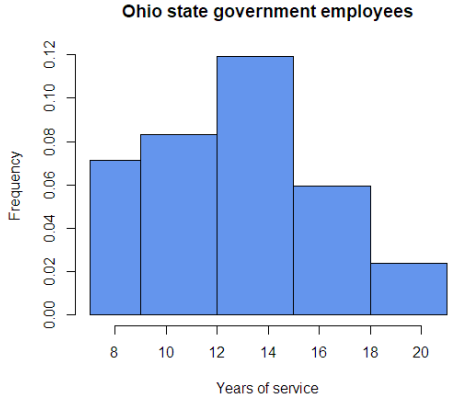
-y=c(13,8,10,14,19,15,13,16,10,11,7,9,14,13,11,12,16,15,13,10,18,17,10,15,16,9,15,19)

-Inputs data in columns

-hist(y,xlab="Years of service",ylab="Frequency",main="Ohio state government employees", breaks=c(7,9,12,15,18,21),col="cornflowerblue",freq=F)

-Make histogram with customized labels

-"breaks" is most important as we are specifying the classes using lower class limits. Else, it'll do automatically



-R defaulted to Relative Frequency here, oops.

-Doing "? hist" gives you the command's help.

-freq=T, xlim=c(0,25), ylim=c(0, 15)

-Class Limits Myth Prac 1: "Min = 7, max = 64, 7 classes"

-Range = 64 - 7 = 57

-Class Width = ceiling(57 / 7) = 9

-So first lower limit is 7, then 16, 25,...

-Then you do upper. 15, 24, 33,...

Class	Frequency
7 - 15	
16 -24	
25 - 33	
34 -42	
43 -51	
52 -60	
61 -69	

-Class Limits Myth Prac 2: "Min = 17, max = 135, 8 classes"

-Range = 135 – 17 = 118

-Class Width = ceiling(118 / 8) = 15

Class	Frequency
17 - 31	
32 - 46	
47 - 61	
62 - 76	
77 - 91	
92 - 106	
107 -121	
122 - 136	

-Find Class Width Myth Prac:

Class	Frequency
25 – 32	86
33 – 40	39
41 – 48	41
: :	

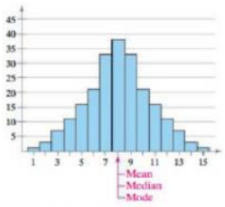
-Class Width = 33 - 25 = 8

-It's not upper - lower!!!

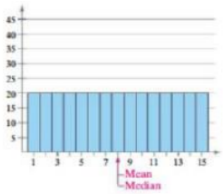
-Histogram Distribution Shapes

-Symmetric, Uniform, Skewed Left/Right

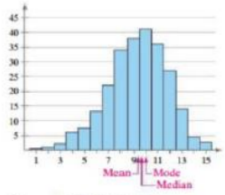
-ie Skewed [Direction] means the mean is to the [Direction] of the median



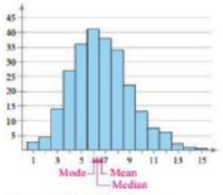
Symmetric Distribution



Uniform Distribution



Skewed Left Distribution

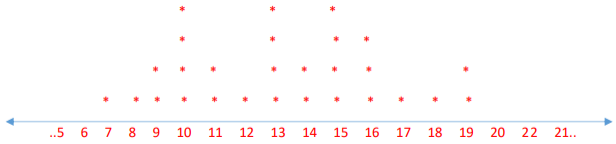


Skewed Right Distribution

-Dot Plot (Quantitative Data)

-Basically a histogram without any classes

-Ex: Ohio government years (same ex)

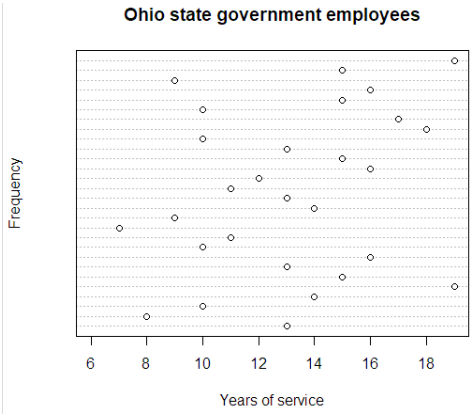


-R Code

-y=c(13,8,10,14,19,15,13,16,10,11,7,9,14,13,11,12,16,15,13,10,18,17,10,15,16,9,15,19)

-Inputs data in columns (Same as histogram)

-dotchart(y,xlab="Years of service",ylab="Frequency",main="Ohio state government employees",xlim=c(6,19), ylim=c(0, 10))



-Stem/Leaf Plot (Quantitative Data)

-Each number is separated into a stem (for instance, the entry's leftmost digits) and a leaf (for instance, the entry's rightmost digits).

1) A stem-and-leaf plot should have as many leaves as there are entries in the original data.
2) Leaves should have single digits.
3) A stem-and-leaf plot is similar to a histogram but has the advantage that the graph still contains the original data.
4) Another advantage of a stem-and-leaf plot is that it provides an easy way to sort data.

-Ex: Data of biology test scores

75	85	90	80	87	67	82	88	95	91	73	80
83	92	94	68	75	91	79	95	87	76	91	85

-Plot by 10 (60-69, 70-79, 80-89, 90-99)

6	7	8							Key: 6|7 = 67
7	3	5	5	6	9				
8	0	0	2	3	5	5	7	7	8
9	0	1	1	1	2	4	5	5	

-Interpretation: Most grades of biology were in the 90's and 80's (aka 80 and 95). (Skewed Left)

-R Code

-y=c(75,85,90,80,87,67,82,88,95,91,73,80,83,92,94,68,75,91,79,95,87,76,91,85)

-stem(y,scale=1,width=80)

```
6 | 78
7 | 35569
8 | 002355778
9 | 01112455
```

-Plot by 5 (60-64, 65-69, 70-74, 75-79,...)

6						Key: 6|7 = 67
6	7	8				
7	3					
7	5	5	6	9		
8	0	0	2	3		
8	5	5	7	7	8	
9	0	1	1	1	2	4
9	5	5				

-R Code

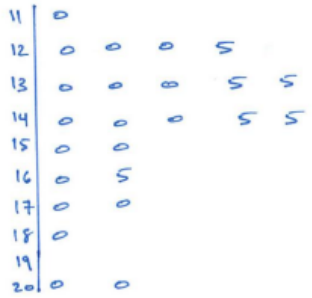
-stem(y,scale=2,width=80)#scale=2

```
6 | 78
7 | 3
7 | 5569
8 | 0023
8 | 55778
9 | 011124
9 | 55
```

-Ex: Systolic blood pressure of 24 patients

120	135	140	145	130	150	120	170	145	125	130	110
160	180	200	150	200	135	140	120	130	170	165	140

-Plot by 10 (110-119, 120-129,...)



-Intepretation: tends to be from 120 to 150 mm of mercury. (Skewed Right?)

-Boxplot (Quantitative Data)

-First, what are quartiles? We need 3

The three **quartiles**, Q_1, Q_2, and Q_3, divide an ordered data set into four equal parts. About one-quarter of the data fall on or below the **first quartile** Q_1. About one-half of the data fall on or below the **second quartile** Q_2 (the second quartile is the same as the median of the data set). About three-quarters of the data fall on or below the **third quartile** Q_3.

-We need to find median first for Q_2. This splits the set in ½. Look for median for both ½ (Q_2 & Q_3)

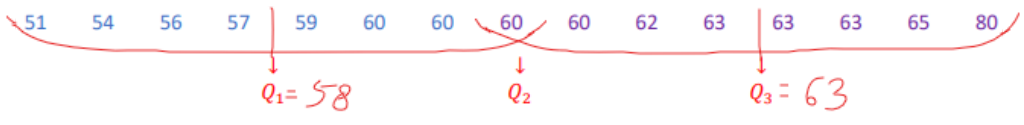
If n is odd then → The median = $\left(\frac{n+1}{2}\right)^{th}$ value (easy)

If n is even then → The median = $\frac{\left(\frac{n}{2}\right)^{th} \text{value} + \left(\frac{n+2}{2}\right)^{th} \text{value}}{2}$ (Avg)

-Ex: Random set

56	63	51	60	57	60	60	54	63	59	80	63	60	62	65

-Arrange them. Find mid, find other 2 mids.



-Our convention: Just include the Q_2 median for the others. (else you have to be consistent)

-R Code find Quartiles

-y=c(56,63,51,60,57,60,60,54,63,59,80,63,60,62,65)

-summary(y)

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 51.00   58.00   60.00   60.87   63.00   80.00
```

-With quartiles, draw it

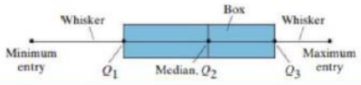
1. The minimum entry
2. The first quartile Q_1
3. The median Q_2
4. The third quartile Q_3
5. The maximum entry

These five numbers are called the **five-number summary** of the data set.

GUIDELINES

Drawing a Box-and-Whisker Plot

1. Find the five-number summary of the data set.
2. Construct a horizontal scale that spans the range of the data.
3. Plot the five numbers above the horizontal scale.
4. Draw a box above the horizontal scale from Q_1 to Q_3 and draw a vertical line in the box at Q_2.
5. Draw whiskers from the box to the minimum and maximum entries.

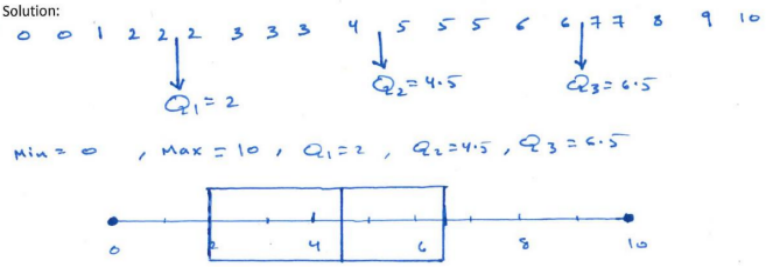


-Ex: Vacation days of 20 employees

3 9 2 1 7 5 3 2 2 6

4 0 10 0 3 5 7 8 6 5

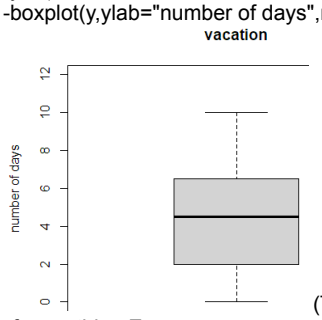
Solution:



-R Code

-y=c(3,9,2,1,7,5,3,2,2,6,4,0,10,0,3,5,7,8,6,5)

-boxplot(y,ylab="number of days",main="vacation",lsa=1, ylim=c(0, 12))



(There's a way to flip it, rip)

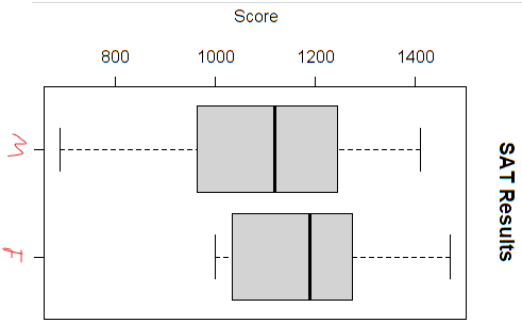
-R Code takes care of everything Ex:

Ex. #7: Find the boxplot for each of the two data sets of SAT scores for eight males and seven females listed. Then, compare the plots:

Males: 1010 1170 1410 920 1320 1100 690 1140

Females: 1190 1010 1000 1300 1470 1250 1060

```
Male=c(1010,1170,1410,920,1320,1100,690,1140)
Female=c(1190,1010,1000,1300,1470,1250,1060)
boxplot(Male,Female,ylab="Score",main="SAT Results")
```



-Interpretation

Females have a slightly higher median SAT scores than males SAT scores. The males SAT score is more spread out thus, it has more variation than the females. Also, the Male SAT scores are skewed left where the female SAT scores are skewed right.

-Outliers

-Interquartile Range (IQR) = Q3 - Q1

The **interquartile range (IQR)** of a data set is a measure of variation that gives the range of the middle portion (about half) of the data. The IQR is the difference between the third and first quartiles.

$$IQR = Q_3 - Q_1$$

-Find Outliers

-If the data entry is less/greater than IQR*1.5, it is an outlier.

Using the Interquartile Range to Identify Outliers

1. Find the first (Q_1) and third (Q_3) quartiles of the data set.
2. Find the interquartile range: IQR = $Q_3 - Q_1$.
3. Multiply IQR by 1.5: 1.5(IQR).
4. Subtract 1.5(IQR) from Q_1. Any data entry less than $Q_1 - 1.5(\text{IQR})$ is an outlier.
5. Add 1.5(IQR) to Q_3. Any data entry greater than $Q_3 + 1.5(\text{IQR})$ is an outlier.

-Ex: Random set from earlier Ex

56	63	51	60	57	60	60	54	63	59	80	63	60	62	65
51	54	56	57	59	60	60	60	60	62	63	63	63	65	80

From $Q_1 = 58$ and $Q_3 = 63$

IQR = $Q_3 - Q_1 = 63 - 58 = 5$

Upper limit = $Q_3 + 1.5 * IQR = 63 + 1.5 * 5 = 63 + 7.5 = 70.5$

Lower limit = $Q_1 - 1.5 * IQR = 58 - 1.5 * 5 = 58 - 7.5 = 50.5$

Any data that is lower than 50.5 or above 72 is an outlier. Thus, 80 is an outlier.

-R Code

```
y=c(56,63,51,60,57,60,60,54,63,59,80,63,60,62,65)
summary(y)
IQR(y)
```

```
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  51.00   58.00   60.00   60.87   63.00
   Max.
  80.00
> IQR(y)
[1] 5
```

(IQR = 5)

-Pie Chart (Qualitative Data)

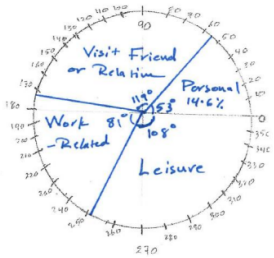
-Basically Circle Histogram for qualitative data.

-Ex: Vacation location survey

-Percentage (frequency / # of entries) to degrees (% * 360deg)

$$\% = \frac{f}{n}$$

Purpose	Number		Degrees
Personal Business	146	$\frac{146}{1000} = .146$	$.146 * 360^\circ \approx 53^\circ$
Visit Friend or Relative	330	$\frac{330}{1000} = .330$	$.330 * 360^\circ \approx 119^\circ$
Work-related	225	$\frac{225}{1000} = .225$	$.225 * 360^\circ \approx 81^\circ$
Leisure	299	$\frac{299}{1000} = .299$	$.299 * 360^\circ \approx 108^\circ$
Total	1,000	100%	360°

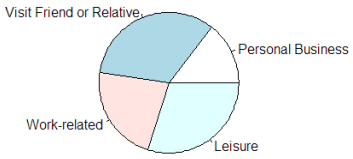


About 1/3 of the travelers visit a friend or a relative, with the fewest traveling for personal business.

-R Code

```
slices = c(146, 330, 225, 299)
lbls = c("Personal Business", "Visit Friend or Relative", "Work-related", "Leisure")
pie(slices,labels = lbls, main="Why Do People Travel in U.S.")
```

-Data set, labels set, then make Pie chart.



-Fancy alternative with percentage

```
slices = c(146, 330, 225, 299)
lbls = c("Personal Business", "Visit Friend or Relative", "Work-related", "Leisure")
pct = round(slices/sum(slices)*100)
bls = paste(lbls, pct) # add percents to labels
lbls = paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)), main="Why Do People Travel in U.S.")
```

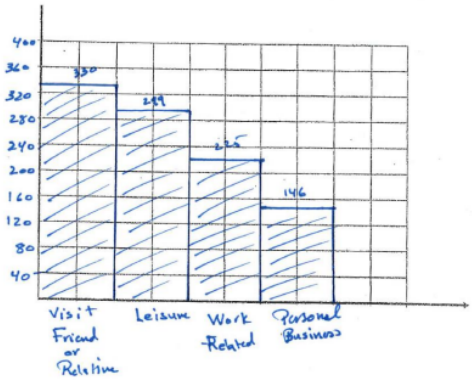
-pct is percentage set, bls concatenate to lbls

-Pareto Chart (Qualitative Data)

-Legit just a histogram but x axis is qualitative labels.

Def: A **Pareto chart** is a vertical bar graph in which the height of each bar represents frequency or relative frequency. The bars are positioned in order of decreasing height, with the tallest bar positioned at the left. Such positioning helps highlight important data and is used in business.

-Ex: Same as pie chart vacation location

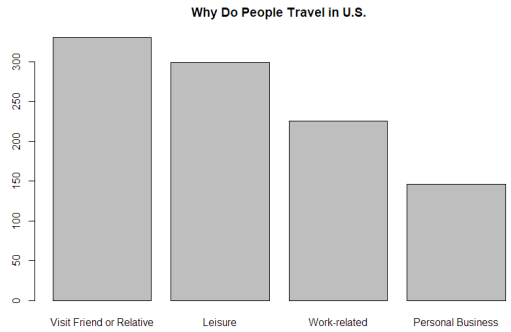


Purpose	Number
Personal Business	146
Visit Friend or Relative	330
Work-related	225
Leisure	299
Total	1,000

-R Code

-slices = c(330,299,225,146)

-barplot(slices, main="Why Do People Travel in U.S.",names.arg=c("Visit Friend or Relative","Leisure", "Work-related", "Personal Business"))



-R won't organize by descending, you have to point entries in order.

2.2: Numerical Summarization of a Quantitative data

-Measures of Central Tendency

-Mean (again)

$$\text{Population Mean: } \mu = \frac{\Sigma x}{N}$$

$$\text{Sample Mean: } \bar{x} = \frac{\Sigma x}{n}$$

(ez, just note symbol)

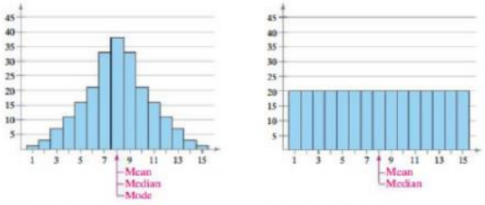
-Median (again)

-Middle entry of the data when in order

If n is odd then → The median = $\left(\frac{n+1}{2}\right)^{th}$ value

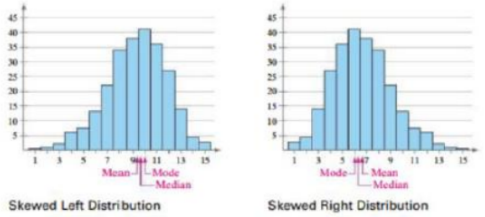
If n is even then → The median = $\frac{\left(\frac{n}{2}\right)^{th} \text{ value} + \left(\frac{n+2}{2}\right)^{th} \text{ value}}{2}$

-Histogram Distribution (again)



Symmetric Distribution

Uniform Distribution



Skewed Left Distribution

Skewed Right Distribution

(note mean and median places relative to each other)

-R Code

-summary([data set]) gives mean and median

-Weighted Mean and Mean of Grouped Data

-"Sum(entries * wieghts) / Sum(weights)"

A **weighted mean** is the mean of a data set whose entries have varying weights. The weighted mean is given by

$$\bar{x} = \frac{\Sigma xw}{\Sigma w}$$ Sum of the products of the entries and the weights / Sum of the weights

where w is the weight of each entry x.

-R Code: "sum(x*w)/sum(w)"

-Ex: Grades (where the points are A = 4, B = 3, C = 2, D = 1, F = 0)

Final Grade	Credit Hours
C	3
C	4
D	1
A	3
C	2
B	3

===>

Final Grade	Credit Hours	grade	# Credit * grade
C	3	2	3*2 = 6
C	4	2	4*2 = 8
D	1	1	1
A	3	4	12
C	2	2	4
B	3	3	9
Σ		16	40

$\bar{X} = \frac{\Sigma X W}{\Sigma W} = \frac{40}{16} = 2.5$

-Ex: Credit card balance at different days

-"For the month of October, a credit card has a balance of $115.63 for 12 days, $637.19 for 6 days, $1225.06 for 7 days, $0 for 2 days, and $34.88 for 4 days. What is the account's mean daily balance for October?"

$$\bar{x} = \frac{(115.63 \cdot 12) + (637.19 \cdot 6) + (1225.06 \cdot 7) + (0 \cdot 2) + (34.88 \cdot 4)[\$ \text{ days}]}{12 + 6 + 7 + 2 + 4\,[\text{days}]}$$

$$= \$449.21$$

-Measures of Variation

-Range

-Ez "Max - Min" entries

-R Code: range([data set])

-Standard Deviation

-Deviation (How far from the mean)

The **deviation** of an entry x in a population data set is the difference between the entry and the mean μ of the data set.

$$\text{Deviation of } x = x - \mu$$

-The sum of all derivation = 0

-Population Variance

The **population variance** of a population data set of N entries is

$$\text{Population variance} = \sigma^2 = \frac{\Sigma(x - \mu)^2}{N}.$$

The symbol σ is the lowercase Greek letter sigma.

-Square it to make everything positive

-Divide by number of entries to avg it, boom.

-Standard Deviation

The **population standard deviation** of a population data set of N entries is the square root of the population variance.

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

-Square root of population variance

-Sample Variance & Standard Deviation

The **sample variance** and **sample standard deviation** of a sample data set of n entries are listed below.

$$\text{Sample variance} = s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

$$\text{Sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

-n-1 because it'll give a larger Variance/Standard Deviation, a way to be less biased as we assume the population var/stddev to be higher.

-Remarks

1) $\sum(x - \bar{x})^2$ **is called the sum of squares for** x **and denoted by** S_{xx}.
2) $\textit{standard deviation} = \sqrt{\textit{variance}}$
3) $\textit{population standard deviation } "\sigma" = \sqrt{\sigma^2}$
4) $\textit{sample standard deviation } "S" = \sqrt{S^2}$

-Ex: Population of alcohol crash faltalities (in thousands) from 2005 through 2011

-14, 13, 13, 12, 9, 11

#	x	$x - \mu$ (Deviation)	$(x - \mu)^2$ (Squared)
1	14	14 − 12 = 2	$(2)^2$ = 4
2	13	13 − 12 = 1	$(1)^2$ = 1
3	13	13 − 12 = 1	$(1)^2$ = 1
4	12	12 − 12 = 0	$(0)^2$ = 0
5	9	9 − 12 = −3	$(-3)^2$ = 9
6	11	11 − 12 = −1	$(-1)^2$ = 1
Σ	72	0	16

Range = 14 − 9 = 5

$\mu = \frac{14+13+13+12+9+11}{6} = \frac{72}{6} = 12$

$\sigma^2 = \frac{16}{6} \approx 2.67$

$\sigma = \sqrt{\frac{16}{6}} \approx 1.63$

-Can't use R for population (unless you trick it into dividing by n instead of n-1)

-Custom Method "sdp <- function(x) sqrt(mean((x-mean(x))^2))" finds Population StdDev

-Ex: the duration (in days) of pregnancies for a random sample of mothers.

277 291 295 280 268 278

#	x	$x-\bar{x}$	$(x-\bar{x})^2$
1	277	$277-281.5=-4.5$	$(-4.5)^2=20.25$
2	291	$291-281.5=9.5$	$(9.5)^2=90.25$
3	295	$295-281.5=13.5$	$(13.5)^2=182.25$
4	280	$280-281.5=-1.5$	$(-1.5)^2=2.25$
5	268	$268-281.5=-13.5$	$(-13.5)^2=182.25$
6	278	$278-281.5=-3.5$	$(-3.5)^2=12.25$
Σ	1689	0	489.5

Range $= 295-268 = 27$

$\bar{x} = \frac{277+291+295+280+268+278}{6} = \frac{1689}{6} = 281.5$

$s^2 = \frac{489.5}{5} = 97.9$

$s = \sqrt{97.9} \approx 9.89$

-R Code

```
y=c(277,291,295,280,268,278)
range(y)        #provide the max and min
mean(y)         #provide the sample mean
sd(y)^2         #provide the sample variance
sd(y)           #provide the sample standard deviation
```

-Properties of the Sample Mean and the Sample Variance

-Summation Review

$\sum_{i=1}^{n}(a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$

$\sum_{i=1}^{n}(a_i - b_i) = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} b_i$

$\sum_{i=1}^{n}(a_i * b_i) \neq \sum_{i=1}^{n} a_i * \sum_{i=1}^{n} b_i$

$\sum_{i=1}^{n}(a_i / b_i) \neq \sum_{i=1}^{n} a_i / \sum_{i=1}^{n} b_i$

$\sum_{i=1}^{n} ca_i = c\sum_{i=1}^{n} a_i$

$\sum_{i=1}^{n} c = c * n$ $\rightarrow \sum_{i=1}^{5} 3 = 3+3+3+3+3 = 5*3$

-Theorem 1: Let a and b be any real numbers and let $x = (x_1, x_2 \dots \dots . , x_i, \dots \dots , x_n)$ be a data set with a sample mean $\bar{x}$ and a sample variance S_x^2.

-1) If $y_i = ax_i$ then $\bar{y} = a\bar{x}$

-Aka, if the entries of the data set is multiplied by some constant, the avg is also of the same multiple.

-Direct Proof:

$x \Rightarrow x_1, x_2, x_3, \dots, x_i, \dots, x_n \Rightarrow y \Rightarrow ax_1, ax_2, ax_3, \dots, ax_i, \dots, ax_n$

(Start with Def of Mean)

Want To show that $\bar{y} = a\bar{x}$

L.H.S $\rightarrow$ $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ but $y_i = ax_i$

$= \frac{\sum_{i=1}^{n} ax_i}{n}$ but a is a constant by properties of Σ it can go out of Σ.

$= \frac{\sum_{i=1}^{n} a x_i}{n}$

$= a\frac{\sum_{i=1}^{n} x_i}{n}$ but $\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

$= a\bar{x}$ R.H.S

$\therefore \bar{y} = a\bar{x}$

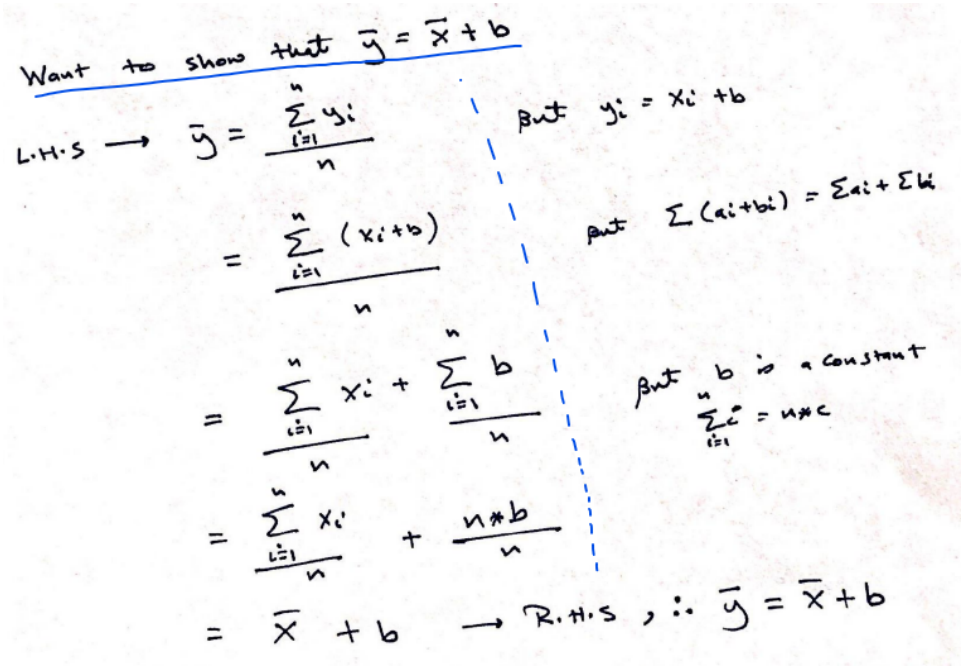
-2) If $y_i = (x_i + b)$, then $\bar{y} = (\bar{x} + b)$

-Aka, if you add the old data set entries by a constant, the avg is changed by adding the constant too.

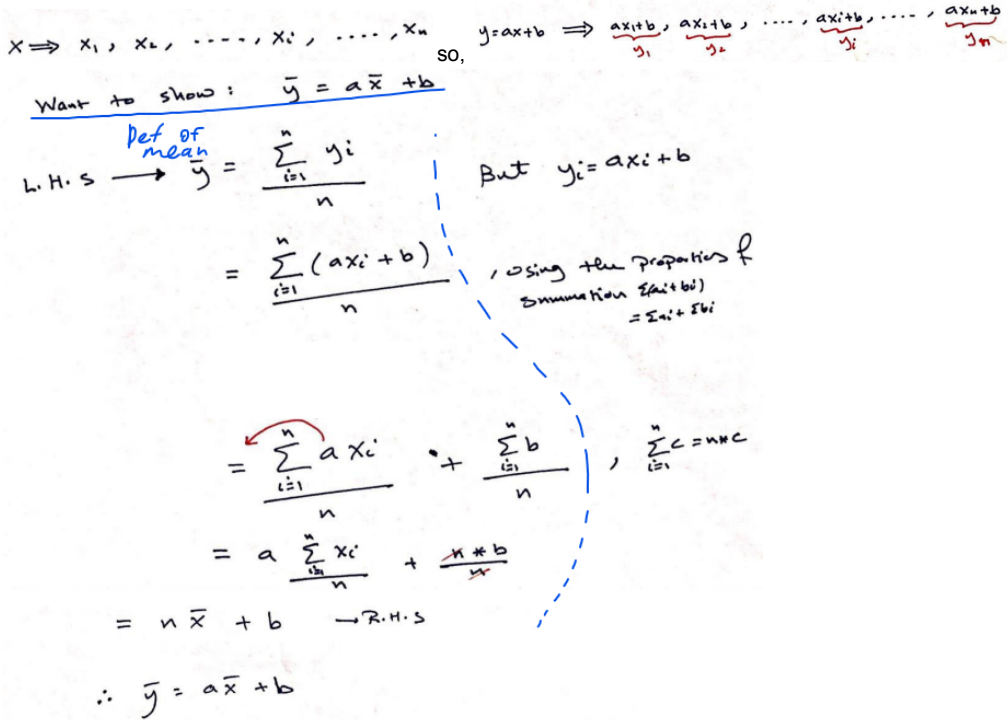
-Direct Proof:

$x \Rightarrow x_1, x_2, x_3, \dots, x_i, \dots, x_n$ so, $y = x+b \Rightarrow \underbrace{x_1+b}_{y_1}, \underbrace{x_2+b}_{y_2}, \underbrace{x_3+b}_{y_3}, \dots, \underbrace{x_i+b}_{y_i}, \dots, \underbrace{x_n+b}_{y_n}$

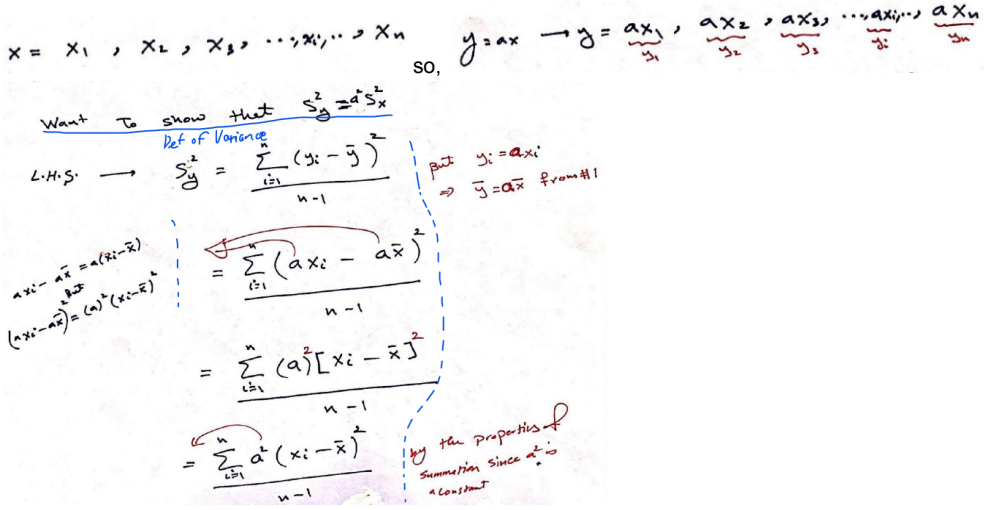
(Start with Def of Mean)



-3) If $y_i = (ax_i + b)$, then $\bar{y} = (a\bar{x} + b)$
-Aka, combine the 2 from above
-Direct Proof



-4) If $y_i = ax_i$ then $Sy^2 = a^2 S_x^2$
-Aka, if you multiply the data entries by a constant, then the variance is multiplied by the constant squared.
-Direct Proof: (requires proof from 1)



$= a^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

$= a^2 S_x^2 \rightarrow$ R.H.S

$\therefore S_y^2 = a^2 S_x^2$

-5) If $y_i = (x_i + b)$, then $S_y^2 = S_x^2$

-Aka, if you add a constant to a data set, the variance do not change

-Direct Proof: (requires proof from 2)

$x = x_1, x_2, \dots, x_i, \dots, x_n$ so, $y = \underbrace{x_1+b}_{y_1}, \underbrace{x_2+b}_{y_2}, \dots, \underbrace{x_i+b}_{y_i}, \dots, \underbrace{x_n+b}_{y_n}$

Want To show $S_y^2 = S_x^2$

L.H.S $\rightarrow$ $S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$ put $y_i = x_i + b$

$\Rightarrow \bar{y} = \bar{x} + b$ from #2

$S_y^2 = \frac{\sum_{i=1}^{n} ([x_i + b] - [\bar{x} + b])^2}{n-1}$

$= \frac{\sum_{i=1}^{n} (x_i + b - \bar{x} - b)^2}{n-1}$

$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ $\rightarrow S_x^2$ by definition

$= S_x^2 \rightarrow$ R.H.S $\Rightarrow S_y^2 = S_x^2$

$$S_y^2 = \frac{\sum_{i=0}^{n} (y_i - \bar{y})^2}{n-1} \text{ (Defn Sample Variance)}$$
$$= \frac{\sum_{i=0}^{n} ((x_i + b) - (\bar{x} + b))^2}{n-1} \text{ (Premise } y_i = x_i + b \text{ \& Thm1.2 } \bar{y} = \bar{x} + b)$$
$$= \frac{\sum_{i=0}^{n} (x_i + b - \bar{x} - b)^2}{n-1}$$
$$= \frac{\sum_{i=0}^{n} (x_i - \bar{x})^2}{n-1}$$
$$= S_x^2 \text{ (Defn Sample Variance)}$$
$$\therefore \quad S_y^2 = S_x^2$$

-6) If $y_i = (ax_i + b)$, then $S_y^2 = a^2 S_x^2$ which is equivalent to $S_y = aS_x$.

-Aka, if you add & multiply by constants, then the variance only changes by the multiplying constant, changed by that constant squared.

-Direct Proof: (requires proof from 3)

$x \Rightarrow x_1, x_2, \dots, x_i, \dots, x_n$ so, $y = ax + b \Rightarrow \underbrace{ax_1+b}_{y_1}, \underbrace{ax_2+b}_{y_2}, \dots, \underbrace{ax_i+b}_{y_i}, \dots, \underbrace{ax_n+b}_{y_n}$

We want to show $S_y^2 = a^2 S_x^2$

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \qquad \text{defn of S.Var.}$$
$$= \frac{\sum_{i=1}^{n} (ax_i + b - \bar{y})^2}{n-1} \qquad y_i = ax_i + b$$
$$= \frac{\sum_{i=1}^{n} (ax_i + b - (a\bar{x} + b))^2}{n-1} \qquad 3)\ \bar{y} = a\bar{x} + b$$
$$= \frac{\sum_{i=1}^{n} (ax_i + b - a\bar{x} - b)^2}{n-1}$$
$$= \frac{\sum_{i=1}^{n} a^2 (x_i - \bar{x})^2}{n-1} \qquad (ax_i - a\bar{x})^2 = a^2 (x_i - a\bar{x})^2$$
$$= \frac{a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$
$$\therefore S_y^2 = a^2 S_x^2 \qquad \text{defn of S.Var.}$$
$$\therefore S_y = aS_x$$

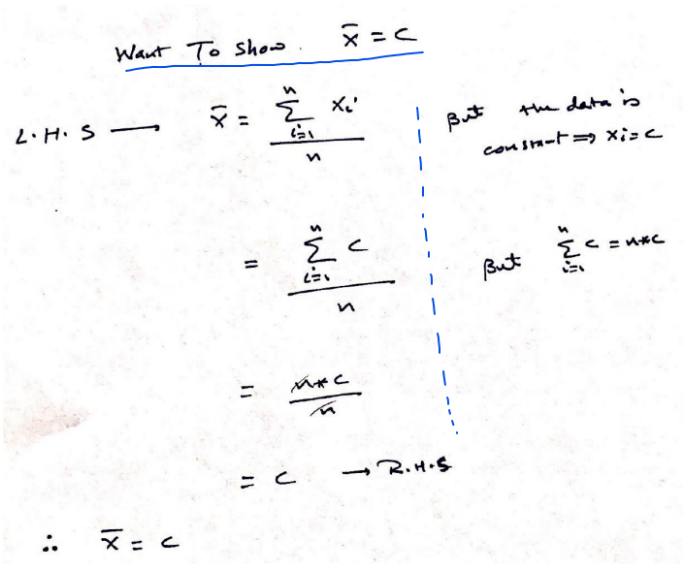
-Theorem 2: : Let c be any real numbers and let $x = (c, c, \dots\dots, c, \dots\dots, c)$ be a data set with a sample mean $\bar{x}$ and a sample variance S_x^2 .

-1) $\bar{x} = c$

-Aka, Mean of a data set of the same constant = the constant.

-Direct Proof

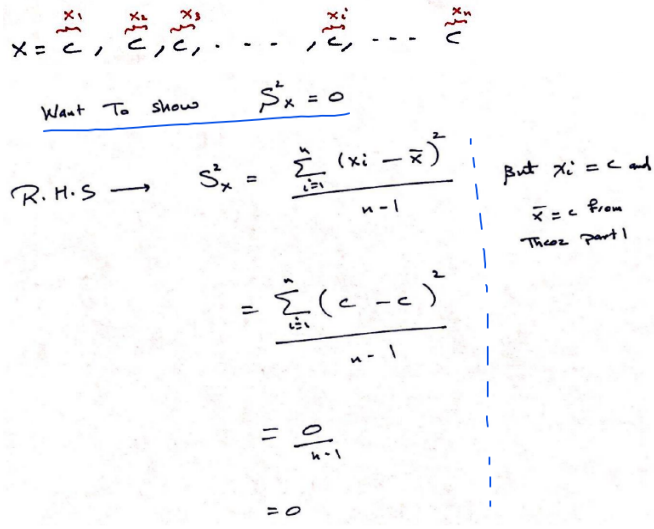
$x = \underbrace{c}_{x_1}, \underbrace{c}_{x_2}, \underbrace{c}_{x_3}, \dots, \underbrace{c}_{x_i}, \dots \underbrace{c}_{x_n}$



-2) $S_x^2 = 0$

-Aka, the Sample Variance of a data set of the same constant is zero

-Direct Proof: (requires proof from 1)



-Theorem 3:

Theorem3: Let $x = (x_1, x_2 \ldots\ldots, x_i, \ldots\ldots, x_n)$be a data set with a sample mean $\bar{x}$ and a sample variance $S_x^2$and Let $y = (y_1, y_2 \ldots\ldots, y_i, \ldots\ldots, y_n)$be a data set with a sample mean $\bar{y}$ and a sample variance S_y^2. Let $z = x - y$ such that $z = (\underbrace{x_1 - y_1}_{z_1}, \underbrace{x_2 - y_2}_{z_2} \ldots\ldots, \underbrace{x_i - y_i}_{z_i}, \ldots\ldots, \underbrace{x_n - y_n}_{z_n})$ then,

-Aka: You got 3 data sets, x, y, & z (same n). The entries for z = x - y from corresponding spots.

-1) $\bar{z} = \bar{x} - \bar{y}$

-Aka, avg of z = avg of x - avg of y

-Direct Proof:

$$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n} \qquad \text{Defn of Mean}$$
$$= \frac{\sum_{i=1}^{n} (x_i - y_i)}{n} \qquad z = x - y$$
$$= \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i}{n}$$
$$= \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} y_i}{n} \qquad \text{Defn of Mean}$$
$$= \bar{x} - \bar{y}$$
$$\therefore \bar{z} = \bar{x} - \bar{y}$$

-2) $S_z^2 \neq S_x^2 - S_y^2$

-Aka, S.Var. of z does not equal S.Var. of x - S.Var. of y

HW 2.2

-30)

Let x be a data set such that $\bar{x} = 3$ and $S_x^2 = 15$.

a) If $y = 5x + 2$, find:

i. $\bar{y}$

ii. S_y^2

b) If $z = x - y$, find $\bar{z}$.

a.i)

$$\bar{y} = 5\bar{x} + 2 \,(\text{thm}1.3)$$
$$= 5(3) + 2$$
$$= 17$$

a.ii)

$$S_y^2 = 5^2 S_x^2 \,(\text{thm}2.3)$$
$$= 25 \cdot 15$$
$$= 375$$

b)

$$\bar{z} = \bar{x} - \bar{y} \,(\text{thm}3.2)$$
$$= 3 - 17$$
$$= -14$$

-31) Prove sum of all data deviation is 0

Let $x = (x_1, x_2 \ldots\ldots, x_i, \ldots\ldots, x_n)$ be a data set with a sample mean $\bar{x}$. Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

-Direct Proof: We want to show $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n * \bar{x} = \sum_{i=1}^{n} x_i - n * \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

2.3: Measures of Position

-Percentiles and Other Fractiles

To find the **percentile that corresponds to a specific data entry** x, use the formula

$$\text{Percentile of } x = \frac{\text{number of data entries less than } x}{\text{total number of data entries}} \cdot 100$$

and then round to the nearest whole number.

Fractiles	Summary	Symbols
Quartiles	Divide a data set into 4 equal parts.	Q_1, Q_2, Q_3
Deciles	Divide a data set into 10 equal parts.	$D_1, D_2, D_3, \ldots, D_9$
Percentiles	Divide a data set into 100 equal parts.	$P_1, P_2, P_3, \ldots, P_{99}$

-Ex: I scored in the 100th percentiles in the math test because im asian

-Always round to the nearest whole if fraction

-Ex: 56, 63, 51, 60, 57, 60, 60, 54, 63, 59, 80, 63, 60, 62, 65

-Percentile of 62:

-(9 entries below / 15 entries in total) * 100 = 60th percentile

-That means P_{60}=62.

-60% scored less than 62 points (aka The person who got 62 scored higher than 60% of others)

-Percentile of 60

-(5/15)*100 = 60

-P_{30}=60, 30% scored less than 62 points

-R Code = quartile([data set], [set of n-tiles, ie]

-Standard Score / Z-Score & T-Score

The **standard score**, or **z-score**, represents the number of standard deviations a value x lies from the mean μ. To find the z-score for a value, use the formula

$$z = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}} = \frac{x - \mu}{\sigma}.$$

()

-Gives you how far the data point is from the mean by units of standard deviation.

-Pos = Above the mean

-Neg = Below the mean

-Zero = Same as the mean

-For a sample data set, it's t = (x - x̄) / s

-Ex: An aptitude test has a mean of 220 and a standard deviation of 10.

-Z-Score for 200

-z(200) = (200 - 220) / 10 = -20 / 10 = -2

-meaning 200 is less than the mean by 2 standard deviation (-2 * 10 points)

-Z-Score for 232

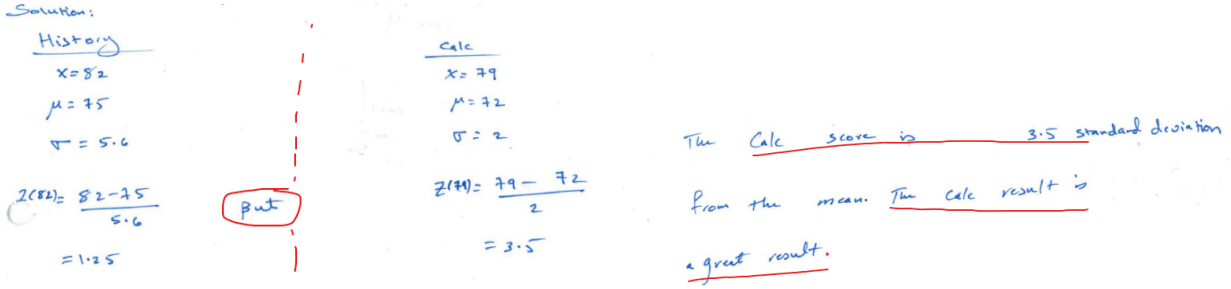
-z(232) = (232 - 220) / 10 = 12 /10 = 1.2

-meaning 232 is above the mean by 1.2 standard deviation (1.2 * 10 points)

-With Z/T-Scores, you can compare data points from different data sets, since you are looking at their position instead of value

-Ex: Calc VS History test scores

Ex. #3: Which score has the higher relative position, a score of 82 on a history test with $\mu = 75$ and $\sigma = 5.6$ or a score of 79 on a Calculus test with $\mu = 72$ and $\sigma = 2$? Please explain your reasoning.



Exam 1

22)

$$\bar{y} = 6\bar{x} - 5$$
$$= 6(17) - 5$$
$$= 97$$
$$S_y^2 = 6^2 S_x^2$$
$$= 36 \cdot 5$$
$$= 180$$

$$\bar{z} = \bar{x} - \bar{y}$$
$$= 17 - 97$$
$$= -80$$

22 agane)

$$\bar{y} = 8\bar{x} - 7$$
$$= 8(12) - 7$$
$$= 89$$
$$S_y^2 = 8^2 S_x^2$$
$$= 64 \cdot 3$$
$$= 192$$
$$\bar{z} = \bar{x} - \bar{y}$$
$$= 12 - 89$$
$$= -77$$

23) sd to sdp factor

$$S^2 c = \sigma^2$$
$$\frac{\sum (x-\bar{x})^2}{n-1} c = \frac{\sum (x-\mu)^2}{N}$$
$$\frac{\sum (x-\mu)^2}{N-1} c = \frac{\sum (x-\mu)^2}{N} \quad (N = n \,\&\, \bar{x} = \mu \text{ because we want to } \sigma^2)$$
$$\sum (x-\mu)^2 c = \frac{(N-1)\sum (x-\mu)^2}{N}$$
$$c = \frac{(N-1)\sum (x-\mu)^2}{N\sum (x-\mu)^2}$$
$$c = \frac{N-1}{N}$$
$$c = 1 - (1/N)$$

```
> t=c(1,2,3,4,5,6,7,8,9,10)
> sd(t)^2
[1] 9.166667
> (sum((t-mean(t))^2))/10
[1] 8.25
> sd(t)^2*(1-(1/10))
[1] 8.25
```

24)mean(y)=a*mean(x)+b*mean(y)+c

LHS

$$z = \frac{\sum_{i=1}^n z_i}{n} \text{ (Defn Sample Mean)}$$
$$= \frac{\sum_{i=1}^n (ax_i + by_i + c)}{n} \text{ (Premise of z)}$$
$$= \frac{\sum_{i=1}^n ax_i + \sum_{i=1}^n by_i + \sum_{i=1}^n c}{n} \text{ (Summation Properties)}$$
$$= \frac{a\sum_{i=1}^n x_i + b\sum_{i=1}^n y_i + c\sum_{i=1}^n 1}{n}$$
$$= a\frac{\sum_{i=1}^n x_i}{n} + b\frac{\sum_{i=1}^n y_i}{n} + c\frac{\sum_{i=1}^n 1}{n}$$
$$= a\frac{\sum_{i=1}^n x_i}{n} + b\frac{\sum_{i=1}^n y_i}{n} + c\frac{n^1}{n}$$
$$= a\bar{x} + b\bar{y} + c \text{ (Defn Sample Mean)}$$
RHS
$$\therefore \quad z = a\bar{x} + b\bar{y} + c$$

3.1: Basic Concepts of Probability and Counting

-Probability Experiment, Outcome, Samples Space, Event

A **probability experiment** is an action, or trial, through which specific results (counts, measurements, or responses) are obtained. The result of a single trial in a probability experiment is an **outcome.** The set of all possible outcomes of a probability experiment is the **sample space.** An **event** is a subset of the sample space. It may consist of one or more outcomes.

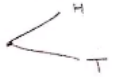
-Simple Event: An event consisting of a single outcome.

-Unusual Event: An event with a slim probability (like 5%)

-Ex: Toss 1 coin

-Sample Space (aka S) = {Heads, Tails}

-1 Event, 2 possible outcomes



-Ex: Roll 6-sided die

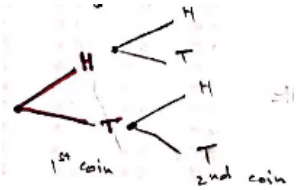
-S = {1,2,3,4,5,6}

-1 Event, 6 possible outcomes

-Ex: Toss 2 coins

-S = {(H,H), (H,T), (T,H) (T,T)}

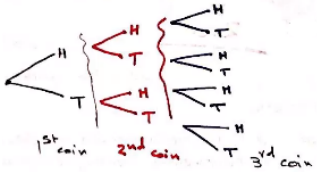
-2 Events, each event has 2 possible outcomes, meaning 2*2 = 4 total possible outcomes



-Ex: Toss 3 coins

$S = \{(H,H,H), (H,H,T), (H,T,H), (H,T,T), (T,H,H), (T,H,T), (T,T,H), (T,T,T)\}$

-3 events, each with 2 possible outcomes, 2*2*2 = 2^3 = 8 total possible outcomes



-Ex: Roll 2 6-sided dice

$S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),$
$(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),$
$(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),$
$(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),$
$(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),$
$(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$

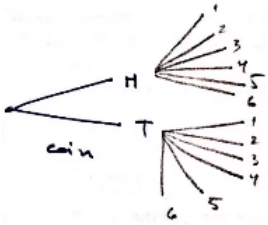
-2 events, each with 6 possible outcomes, 6*6 = 36 total possible outcomes

-Tree diagram is cancerous

-Ex: Toss a coin, then roll a 6-sided die

$S = \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),$
$(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$

-2 Events, 2 and 7 possible outcomes, 2*6 = 12 possible outcomes



-Standard Deck of Card for Gamba

Hearts	Diamonds	Spades	Clubs
A♥	A♦	A♠	A♣
K♥	K♦	K♠	K♣
Q♥	Q♦	Q♠	Q♣
J♥	J♦	J♠	J♣
10♥	10♦	10♠	10♣
9♥	9♦	9♠	9♣
8♥	8♦	8♠	8♣
7♥	7♦	7♠	7♣
6♥	6♦	6♠	6♣
5♥	5♦	5♠	5♣
4♥	4♦	4♠	4♣
3♥	3♦	3♠	3♣
2♥	2♦	2♠	2♣

-Ex: Randomly select a card. Event A is selecting a diamond

$A = \{A\diamondsuit, K\diamondsuit, Q\diamondsuit, 10\diamondsuit, \ldots, 2\diamondsuit\}$

-13 possible outcomes, not a simple event

-Ex: Randomly select a card. Event B is selecting ace of spade

$B = \{A\spadesuit\}$

-1 possible outcome, a simple event

-Fundamental of Counting Principle

If one event can occur in m ways and a second event can occur in n ways, then the number of ways the two events can occur in sequence is $m \cdot n$. This rule can be extended to any number of events occurring in sequence. (aka multiply the possible outcomes)

-Ex: Toss a coin, then roll a 6-sided die (again)

-2 Events, 2*6 = 12 possible outcome

-Ex: 4 digit code, but the 1st digit can't be 0, and the last digit has to be even.

-4 events, the outcomes for each events are 9*10*10*5 = 4500 possible outcomes.

-Ex: Six question true or false quiz

-6 events, each with 2 outcomes, 2^6 = 64 possible outcomes, possible ways student can answer the quiz (if all are answered).

-Types of Properties

-Classical/Theoretical Probability

-Each outcome of a sample space has the same weight/is equally likely to occur..

$$P(E) = \frac{Number\ of\ outcomes\ in\ the\ event\ E}{Total\ number\ of\ outcomes\ in\ sample\ space}$$

-Ex: Flipping coin, drawing 6 marbles in a bowl

-Empirical Probability

-Based on observations obtained from probability experiments. Concerns the frequency of each outcome.

$$P(E) = \frac{frequancy\ of\ event\ E}{Total\ frequency}.$$

-Ex: Chance for a candidate to win election

-Subjective Probability

-Like Empirical where each event cannot be assumed to be equally likely, but you couldn't collect observations via probability experiments to determine relative frequency and resort to an educated guess.

-Ex: Weather forecast by using a model

-Ex: 12-sided die rolling (Classical)

-S = {1,2,3,..., 12}

-Event A: Rolling a 2

-A = {2}

-**P(A)** = 1/12

-Event B: Rolling 15

-B = {} (empty set)

-**P(B)** = 0/12 = 0

-Event C: Rolling # > 4

-C = {5,6,7,8,9,10,11,12}

-**P(C)** = 8/12 = ⅔

-Complementary Events

The **complement of event E** is the set of all outcomes in a sample space that are not included in event E. The complement of event E is denoted by E' and is read as "E prime." The Venn diagram at the left illustrates the relationship between the sample space, event E, and its complement E'. or $\bar{E}$

-Ex: Rolling 6-sided die

-Event A: Rolling 4

-A = {4}

-A' = {1,2,3,5,6}

-P(A') = ⅚

-Axium/Rule: P(E) + P(E') = 1

- $P(E) = 1 - P(E')$
- $P(E') = 1 - P(E)$

-Ex: Finding P(E') from P(E)

E

Ex: Sixty-nine percent of adults favor gun licensing in general. Choose one adult at random. What is the probability that the selected adult doesn't believe in gun licensing?

Let E = an adult favor gun licensing in general

$P(E) = .69$ $P(\text{an adult favor gun licensing}) = .69$

$P(E') = 1 - .69$ $P(\text{an adult } \underline{\text{does}} \text{ not favor gun licensing})$

$= .31$ $= 1 - .69$

$= .31$

-Rules of Probability

- $0 \le P(E) \le 1$ Probability of an event can't be negative and can't be greater than 1.
- $P(E) + P(E') = 1$ Probability of an event + probability of not an event = 1

-The sum of all probabilities of elements in the sample space = 1

-Using Tree Diagram and the Fundamental Counting Principle to find probability.

-Ex: 3 digit code

-Possible outcome = 10*10*10 = 1000

-Event: 1st try

-P(1st try) = 1/1000 = 0.001

-Event: Failing the 1st try

-P(failing 1st try) = 1 - 0.001 = 0.999

3.2: Conditional Probability and the Multiplication Rule

-Conditional Probability

A **conditional probability** is the probability of an event occurring, given that another event has already occurred. The conditional probability of event B occurring, given that event A has occurred, is denoted by $P(B|A)$ and is read as "probability of B, given A."

-P(B|A) is the probability of Event B (2nd Event), given that Event A (1st Event) already occurred.

-Ex: Picking out marbles from a bowl with 5 red & 3 blue marbles. Event B is picking out a blue marble.

-On the second pick, what is the probability of Event B? (aka P(B|B) or P(B|R) = ?)

-With replacements (2nd Event is independent from 1st Event)

-P(2nd B| 1st B) (meaning both picks are blue) = 3/8

-P(2nd B| 1st R) (meaning 1st pick was red and 2nd was blue) = 3/8 still because you put back the marble

-Without replacements (2nd Event is dependent on 1st Event)

-P(2nd B| 1st B) = 2/7 because after the 1st pick, we have 5 red & 2 blue left.

-P(2nd B| 1st R) = 3/7 because after the 1st pick, we have 4 red & 3 blue left.

-Independent/Dependent Events

Two events are **independent** when the occurrence of one of the events does not affect the probability of the occurrence of the other event. Two events A and B are independent when

$P(B|A) = P(B)$ Occurrence of A does not affect probability of B

or when

$P(A|B) = P(A)$. Occurrence of B does not affect probability of A

Events that are not independent are **dependent.**

-Dependent events are harder to deal with.

-Ex: The probability of

-Tossing a coin, then drawing a card are independent events.

-Putting money in a parking meter and getting a parking ticket is dependent.

-A dad having hazel eyes and a daughter having hazel eyes are dependent.

-The amount of rain and the Lakers winning the NBA world championship finals are independent.

-The Multiplication Rule

The probability that two events A and B will occur in sequence is

$P(A \text{ and } B) = P(A) \cdot P(B|A)$. Events A and B are dependent.

If events A and B are independent, then the rule can be simplified to

$P(A \text{ and } B) = P(A) \cdot P(B)$. Events A and B are independent.

This simplified rule can be extended to any number of independent events. (Dependent needs IN SEQUENCE!)

-Ex: Picking 2 cards from a standard deck without replacement

-Probability of both being spades

-P(1st spade and 2nd spade) = P(1st spade) * P(2nd spade | 1st spade)

= **13/52** * **12/51** = 156/2652 = 1/17

-Both are kings

-P(1st king and 2nd king) = P(1st king) * P(2nd king | 1st king)

= **4/52** * **3/51** = 1/221

-Ex: Coin toss and 6 sided die roll

-Probability of tails and rolling > 2

-P(T and {3,4,5,6}) = P(T) * P({3,4,5,6} | T)

= P(T) * P({3,4,5,6}) (because they're **independent**)

= 1/2 * 4/6 = 1/3

-Ex: In a sample of 1000 U.S. adults, 300 said they know a murder victim. 4 people selected without replacement

-Probability all 4 knows a murder victim (KMV)

-P(1st KMV and 2nd KMV and 3rd KMV and 4th KMV)

= 300/1000 * 299/999 * 298/998 * 297/997 = 0.00799 (unusual)

-Probability none of 4 KMV

-P(1st !KMV and 2nd !KMV and 3rd !KMV and 4th !KMV)

= 700/1000 * 699/999 * 698/998 * 697/997 = 0.2395 (unusual)

-Probability at least 1 KMV

-P(at least 1 KMV) = **1 - P(none KMV)** (Use of Complementary Event, not "at least one/there exists" means "none")

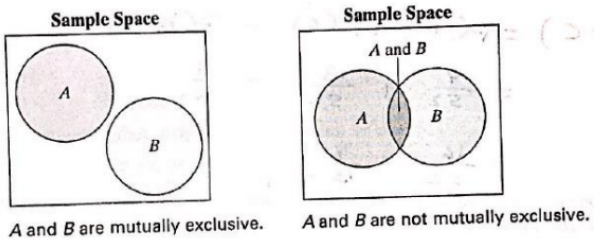
= 1 - P(1st !KMV and 2nd !KMV and 3rd !KMV and 4th !KMV)

= 1 - 0.2395 = 0.7605

3.3: The Addition Rule

-Mutually Exclusive

Two events A and B are **mutually exclusive** when A and B cannot occur at the same time. That is, A and B have no outcomes in common.



A and B are mutually exclusive. A and B are not mutually exclusive.

-Ex:

Event A: Roll a 6-sided die and get a prime number (2, 3, and 5).

Event B: Roll a 6-sided die and get an odd number.

-Not Mutually Exclusive

$A = \{2, 3, 5\}$

$B = \{1, 3, 5\}$

Event A: Roll a 6-sided die and get a prime greater than 3.

Event B: Roll a 6-sided die and get a number less than 3.

-Mutually Exclusive

$A = \{5\}$

$B = \{1, 2\}$

Event A: Select a student that has blond hair.

Event B: Select a student that has blue eyes.

-Not Mutually Exclusive because a person can have both at the same time.

-The Addition Rule (for Probability of A and B)

The probability that events A or B will occur, $P(A \text{ or } B)$, is given by

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. (Minus P(A and B) is because we P(A) + P(B) gives 2*P(A and B), we only want 1)

-But if A & B are mutually exclusive

$P(A \text{ or } B) = P(A) + P(B)$. Events A and B are mutually exclusive.

This simplified rule can be extended to any number of mutually exclusive events. (Also: P(A or B or C) = P(A) + P(B) + P(C)

-Ex: A single card is drawn from a deck. Find the probability that it is a king or a club.

$P(K \text{ or } C) = P(K) + P(C) - P(K \text{ and } C)$ Not M.E.

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$

$= \frac{16}{52} = \frac{4}{13}$

-Ex: In a hospital unit there are 8 nurses and 5 physicians, of which 7 nurses and 3 physicians are females as in the chart below. If a staff person is selected…

Label	Females	Males	Total
Nurses	7	1	8
Physicians	3	2	5
Total	10	3	13

-find the probability that the subject is a nurse (N) and a male (M).
P(N and M) = 1/(7+1+8) = 1/13
-find the probability that the subject is a nurse or a male.
P(N or M) = P(N) + P(M) - P(N and M)
= 8/13 + 3/13 - 1/13
= 10/13

-Ex: Not aids

The table below shows the results of a survey that asked 1056 adults from a certain country if they favored or opposed a tax to fund education. A person is selected at random.

	Support	Oppose	Unsure	Total
Males	154	317	15	486
Females	242	297	31	570
Total	396	614	46	1056

Compute the probability that the randomly selected person is not unsure or is female.

```
> ((396+614)/1056)+(570/1056)-((242+297)/1056)
[1] 0.9857955
```

-Fat Type of Probability and Probability Rules

Type of probability and probability rules	In words	In symbols
Classical Probability	The number of outcomes in the sample space is known and each outcome is equally likely to occur.	$P(E) = \frac{\text{Number of outcomes in event } E}{\text{Number of outcomes in sample space}}$
Empirical Probability	The frequency of each outcome in the sample space is estimated from experimentation.	$P(E) = \frac{\text{Frequency of event } E}{\text{Total frequency}} = \frac{f}{n}$
Range of Probabilities Rule	The probability of an event is between 0 and 1, inclusive.	$0 \le P(E) \le 1$
Complementary Events	The complement of event E is the set of all outcomes in a sample space that are not included in E, and is denoted by E'.	$P(E') = 1 - P(E)$
Multiplication Rule	The Multiplication Rule is used to find the probability of two events occurring in sequence.	$P(A \text{ and } B) = P(A) \cdot P(B\|A)$ Dependent events $P(A \text{ and } B) = P(A) \cdot P(B)$ Independent events
Addition Rule	The Addition Rule is used to find the probability of at least one of two events occurring.	$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ $P(A \text{ or } B) = P(A) + P(B)$ Mutually exclusive events

-Ex: In a sample of 1,000 people (525 men, 475 women), 113 are left-handed (63 men, 50 women). Selecting a person at random, what is the probability

		Gender		
		Male M	Female F	Total
Dominant Hand	Left L	63	50	113
	Right R	462	425	887
	Total	525	475	1000

-The person being left-handed and female
P(L and F) = 50/1000 = 0.05
-Extra: Via dependent events
-P(L and F) = P(L) * P(F|L) (aka P(F) * P(L|F))
= 113/1000 * 50/113 = 50/1000
-The person being left-handed or female (not ME btw)
P(L or F) = P(L) + P(F) - P(L and F) = 113/1000 + 475/1000 - 50/1000 = 0.538
-The person being left-handed or male
P(R or M) = P(R) + P(M) - P(R and M) = 887/1000 + 525/1000 - 462/1000 = 0.950
-The person is not right-handed (aka left-handed) or is a male
P(L or M) = P(L) + P(M) - P(L and M) = 113/1000 + 525/1000 - 63/1000 = 0.575
-The person is right handed and is a female
P(R and F) = 425/1000 = 0.425
-Given that the person is left-handed, what is the probability that it is a female?
P(F|L) = 50/113
-Given that the person is a female, what is the probability that the person is left-handed?
P(L|F) = 50/475
-Are the events "being left-handed" and "being female" independent?

If A and B are independent → $P(A \text{ and } B) = P(A) * P(B)$

$P(L \text{ and } F) \stackrel{?}{=} P(L) * P(F)$ Test it

$\frac{50}{1000} \stackrel{?}{=} \frac{113}{1000} * \frac{475}{1000}$ NO !!

being left-handed and being female are dependent events.

-Are the events "being left-handed" and "being female" mutually exclusive?
-No, a person can be both at the same times. The sample has 50 people.

3.4: (nPr & nCr) Additional Topics in Probability and Counting

-Factorial
-n! = (n)(n-1)(n-2)...(1)
-R Code: factorial(5)
-Trick: 10! / 7! = 10 * 9 * 8

-Choosing *r* objects at a time from *n* objects w/o replacement:

-Permutation (aka when order matters)

A permutation is an ordered arrangement of objects. The number of different permutations of n distinct objects is $n!$.

-Opposite of Permutation is Combination, where order doesn't matter

-Permutations of n Objects Taken r at a time v.s. Combinations of n Objects Taken r at a time. (w/o replacement)

Permutation	Combination
Order does matter	*Order does NOT matter*
Ex1: The Password "2459" is different than 4259	*Ex1: A meal that consists of "burger, fries and a cock " is the same meal as "burger, fries and a cock "*
Ex2: The SSN "436 − 12 − 8778" is different than "463 − 12 − 8778"	*Ex1: In playing cards a full house "5♠, 5♦, 5♣, 7♡ and 7♣" is the same as"5♠, 5♣, 5♦, 7♡ and 7♣"*
$nPr = \frac{n!}{(n-r)!}$	$nCr = \frac{n!}{(n-r)!r!}$
Permutation of *n* objects taking *r* at a time	Combination of *n* objects taking *r* at a time

-Ex: Evaluating

-$7P_2$ = 7! / (7-2)! = 7! / 5! = 7 * 6 = 42

-No R function

-$7C_2$ = 7! / ((7-2)! * 2!) = 7! / (5! * 2!) = 7 * 6 / 2 = 21

-R Code:choose(7, 2)

-Ex: President of the USA

Given *John, Tim and Adam* . How many different committees can you form for

a) {president, vice president } b) { secretary , secretary }

-a. $3P_2$ = 6 (Permutation because 2 different roles, order matters)

-b. $3C_2$ = 3 (Combination because same roles, order doesn't matter)

-Ex: How many different tests can be made from a test bank of 20 questions if the test consists of 5 questions?

-Combination, order does not matter, $20C_5$ = 15504

-Ex: There are 16 finalist in a singing competition. The top five singers receive prizes. How many ways can the singers finish first through fifth?

-Permuation, order does matter by places, $16P_5$ = 524160

-Ex: A restaurant offers a dinner special that lets you choose from 10 entrees, 8 side dishes, and 13 desserts. You can choose one entrée, one side dish, and two desserts. How many different meals are possible

-Many Combinations, order those not matter, $10C_1 * 8C_1 * 13C_2$ = 6240

-Distinguishable Permutations

The number of **distinguishable permutations** of n objects, where n_1 are of one type, n_2 are of another type, and so on, is

$$\frac{n!}{n_1! \cdot n_2! \cdot n_3! \cdots n_k!}$$

where

$$n_1 + n_2 + n_3 + \cdots + n_k = n.$$

(This means you should just count the types with duplicates for denominator)

-Ex: 4 people for 4 positions, {president, vice president, secretary, secretary}.

$$\frac{4!}{1!*1!*2!}$$ (2! is the secretary positions, the president is 1! and VP is 1!)

-Ex: In how many ways can the letters A, B, C, D, E, and F be arranged for a six-letter security code? (no replacements)

$$\frac{6!}{1! \cdot 1! \cdot 1! \cdot 1! \cdot 1! \cdot 1!}$$ (1! for each letter)

-Ex: In how many ways can the letters A, B, C, D, D, and D be arranged for a six-letter security code

$$\frac{6!}{1! \cdot 1! \cdot 1! \cdot 3!}$$ (3! are the Ds)

-Ex: In how many ways can the letters A, B, B, C, C, and C be arranged for a six-letter security code?

$$\frac{6!}{1! \cdot 2! \cdot 3!}$$

-Ex: You are putting 9 pieces of blue beach glass, 3 pieces of red beach glass, and 7 pieces of green beach glass on a necklace. In how many distinguish ways can the beach glass be put on the necklace?

$$\frac{(9+3+7)!}{9! \cdot 3! \cdot 7!}$$

-Application of Counting Principles

-Probability from Permutations w/o replacements

$$Probability = \frac{\text{Permutations of interest}}{\text{Total number of permutations}}$$

-Ex: The university of California Health Services committee has five members are chosen to serve as the committee chair and vice chair. Each committee member is equally likely to serve in either of these positions. What is the probability of randomly selecting the chair and the vice chair?

$$5P_2 = \frac{5!}{(5-2)!} = 20$$

P(selecting (aka guessing who) the chair and vice chair (is)) = 1/20

-Ex:

A warehouse employs 24 workers on first shift and 17 workers on second shift. Eight workers are chosen at random to be interviewed about the work environment. Find the probability of choosing six first-shift workers.

Since we are not interested in the order of the people chosen, order is not important and thus we need to use a **combination**.

We will be selecting 8 out of $24 + 17 = 41$ workers:

$$_{41}C_8 = \frac{41!}{8!(24-8)!} = \frac{41!}{8!33!} = \frac{41 \cdot 40 \cdot \ldots \cdot 1}{(8 \cdot 7 \cdot \ldots \cdot 1) \cdot (33 \cdot 32 \cdot \ldots \cdot 1)} = 95,548,245$$

We are interested in how many ways result in 6 people being selected from the 24 workers on the first shift, while the remaining 2 people are selected from the 17 workers on the second shift.

$$_{24}C_6 \cdot {}_{17}C_2 = \frac{24!}{6!(24-6)!} \cdot \frac{17!}{2!(17-2)!} = \frac{24!}{6!18!} \cdot \frac{17!}{2!15!} = 18,305,056$$

The probability is the number of favorable outcomes divided by the number of possible outcomes.

$$\begin{aligned} &P(\text{Six first-shift workers}) \\ &= \frac{\#\text{ of favorable outcomes}}{\#\text{ of possible outcomes}} \\ &= \frac{_{24}C_6 \cdot {}_{17}C_2}{_{41}C_8} \\ &= \frac{18,305,056}{95,548,245} \\ &= \frac{56,672}{295,815} \\ &\approx 0.1916 \end{aligned}$$

4.1: Probability Distribution

-Random Variables

-The outcome of a probability experiment is often a count or a measure. When this occurs, the outcome is called a random variable.

-Def: A random variable x represents a value associated with each outcome of a probability experiment.

-(Counts) Discrete: a random variable that has a finite or accountable number of possible outcomes that can be listed.

-(Measurements) Continuous: a random variable that has an uncountable number of possible outcomes, represented by an interval on a number line.

-Ex:

-Let x represent the number of cars in a university parking lot.

-Discrete, the amount of cars is countable.

-Let x represent the length of time it takes to complete an exam.

-Continuous, a measurement of time, not countable.

-Let x represent the volume of blood drawn for a blood test.

-Continuous, a measure of volume, not countable

-Probability Distribution

-Continuous Random Variable (probability density function)

Definition: A probability density function $f(x)$ is used to describe (at least approximate) the population or process distribution of a continuous variable. A probability density function (p.d.f) must satisfy the following properties:

1) $f(x) \geq 0$ (non-negative)
2) $\int_{-\infty}^{\infty} f(x)dx = 1$
3) $P(a \leq x \leq b) = \int_a^b f(x)dx$

Since it is a continuous random variable, then there are infinitely many possibilities. Thus, the probability of any single point is zero. That is:

$P(x = a) = P(a \leq x \leq a) = \int_a^a f(x)dx = 0$

Furthermore,

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

$$P(a < x < b) = \int_a^b f(x)dx$$

Since $P(x = a) = 0$ and $P(x = b) = 0$

Thus, $P(a \leq x \leq b) = P(a < x < b)$ if x is a continuous random variable.

(ie You can't ask what is the probability that you will finish the exam right at 12:00? It's actually 12:00.00000… bruh)

-Discrete Random Variable (probability mass function)

Definition: A probability mass function $P(x)$ is used to describe (at least approximate) the population or process distribution of a discrete variable. A probability mass function (p.m.f) must satisfy the following properties.

1) $p(x) \geq 0$ (non-negative)
2) $\sum p(x) = 1$
3) $P(a \leq x \leq b) = \sum_a^b p(x)$

$P(x = a) = p(a)$ (Not necessarily zero)

Furthermore,

$$P(a \leq x \leq b) = p(a) + p(a+1) \ldots + p(b) = \sum_a^b p(x)$$

$$P(a < x < b) = p(a+1) + \cdots + p(b-1) = \sum_{a+1}^{b-1} p(x)$$

Thus, $P(a \leq x \leq b) \neq P(a < x < b)$ if x is a discrete random variable.

(aka bounds matters unlike continuous, because each point has weight/meaning)

-Ex: Determine whether the distribution is a probability distribution

x	1	2	3	4
$P(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{5}{4}$	-1

-Oh hell nah. P(4) = -1 doesn't make sense. P(3) = 5/4 is > 1.

-Ex: Let $f(x) = cx^2(1 - x^2)$ be a density function where c is an unknown positive constant

-1. The Domain of f(x)

-f(x) is non-negative?

-c, x^2, > 0

-$1-x^2 \geq 0$ iff $-1 \leq x \leq 1$

-Yes, therefore domain is [-1, 1] or $-1 \leq x \leq 1$

-2. Finding c (the fudge factor that let f(x)'s area under the curve = 1)

-x is continuous random variable so it's integral time

$$\int_{-1}^{1} cx^2\left(1-x^2\right)dx = 1$$

$$\int_{-1}^{1} c\left(x^2-x^4\right)dx = 1$$

$$c\int_{-1}^{1} \left(x^2-x^4\right)dx = 1$$

$$2c\int_{0}^{1} \left(x^2-x^4\right)dx = 1 \,[\text{even function}]$$

$$2c\left[\frac{x^3}{3}-\frac{x^5}{5}\right]_0^1 dx = 1$$

$$2c\left(\frac{5}{15}-\frac{3}{15}\right) = 1$$

$$2c\frac{2}{15} = 1$$

$$\frac{4c}{15} = 1$$

$$c = \frac{15}{4}$$

$$\therefore f(x) = \frac{15}{4}x^2\left(1-x^2\right),\ -1 \leq x \leq 1$$

$$\text{aka } f(x) = \begin{cases} \frac{15}{4}x^2\left(1-x^2\right),\ -1 \leq x \leq 1 \\ 0 \,\text{elsewhere} \end{cases}$$

-R Code: curve(15/4*x^2*(1-x^2),xlim = c(-1,1)) # we don't know c yet

-R Code (verify answer):

y <- function(x) {15/4*x^2*(1-x^2)} # define the integrated function

integrate(y, lower = -1, upper = 1) # integrate the function

-3. P(x ≤ 1/2)

$$P\left(x \leq \frac{1}{2}\right) = P\left(-1 \leq x \leq \frac{1}{2}\right) = \int_{-1}^{1/2} f(x)dx$$

$$= \int_{-1}^{1/2} \frac{15}{4}x^2\left(1-x^2\right)dx$$

$$= \frac{15}{4}\left[\frac{x^3}{3}-\frac{x^5}{5}\right]_{-1}^{1/2}$$

$$= 81/128 \approx 0.6328$$

-R Code: integrate(y, lower = -1, upper = 1/2) # integrate the function

-4. P(x = 1/2)

-0 ez, because it is a continuous random variable, infinite possibility, can't just pinpoint 1.

-Ex: A contractor is required by a county planning department to submit 1, 2, 3, 4, or 5 forms when applying for a building permit.

Let y denote the number of forms required for an application, and suppose the mass function is given by $p(y) = cy$ for y = 1, 2, 3, 4, or 5.

-FYI: "mass function" is because each form has weight in a sense

-1. What kind of random variable?

-Discrete, count the forms.

-2. Find c (fudge factor that makes f(x)'s summation = 1)

$$\sum_{y=1}^{5} p(y) = 1$$

$$\sum_{y=1}^{5} cy = 1$$

$$c\sum_{y=1}^{5} y = 1$$

$$c15 = 1$$

$$c = \frac{1}{15}$$

$$\therefore f(y) = \frac{1}{15}y$$

-3. The long run proportion (aka probability) of application that requires at most 3 forms.

$$P(y \leq 3) = P(1) + P(2) + P(3)$$

$$= \frac{1}{15}1 + \frac{1}{15}2 + \frac{1}{15}3$$

$$= 2/5 \approx 0.4$$

-4. application that requires between 2 to 4 forms, inclusive

$$P(2 \le y \le 4) = P(2) + P(3) + P(4)$$
$$= \frac{1}{15}2 + \frac{1}{15}3 + \frac{1}{15}4$$
$$= 3/5 \approx 0.6$$

-Ex: HW 4.1 Using Geometric Series to finding c of pmf

Determine c so that the function can serve as the probability distribution of a random variable with the given range. (**Hint:** Recall the geometric series.)

$f(x) = c\left(\frac{1}{4}\right)^x$ for $x = 1, 2, 3, \ldots$.

$$\sum_{x=1}^{\infty} c\left(\frac{1}{4}\right)^x = 1$$
$$c\sum_{x=1}^{\infty}\left(\frac{1}{4}\right)^x = 1$$
$$c\left(\sum_{x=0}^{\infty}\left(\frac{1}{4}\right)^x - \left(\frac{1}{4}\right)^0\right) = 1$$
$$c\left(\frac{1}{1-\frac{1}{4}} - 1\right) = 1$$
$$c\left(\frac{4}{3} - 1\right) = 1$$
$$c\left(\frac{1}{3}\right) = 1$$
$$c = 3$$

$\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ if $|r| < 1$

-Mean (Expected Value), Variance, and Standard Deviation of a Probability Distribution

-Mean

Definition 1: The expected value of a random variable is equal to the mean notation-wise $E(x) = \mu$

-aka the mean is the expected value of a probability distribution (mu because we concern the population).

Definition 2: The expected value of a discrete random variable is $E(x) = \mu = \sum xp(x)$

-For discrete, the mean is the sum of all value * its probability

Definition 3: The expected value of a continuous random variable is $E(x) = \mu = \int_{-\infty}^{\infty} xp(x)dx$

-For continuous, it's the integral instead.

Definition 4: In general, $E(g(x)) = \sum g(x)p(x)$ or $E(g(x)) = \int_{-\infty}^{\infty} g(x)p(x)dx$

-The two combined. the end

-Variance & Standard Deviation

Definition 5: The variance of a discrete random variable is $Var(x) = \sigma^2 = E((x-\mu)^2) = \sum(x-\mu)^2p(x)$ and the standard deviation is $\sigma = \sqrt{\sigma^2}$.

-aka for discrete, variance is sum of all of standard deviation squared

Definition 6: The variance of a continuous random variable is $Var(x) = \sigma^2 = E((x-\mu)^2) = \int_{-\infty}^{\infty}(x-\mu)^2p(x)dx$ and the standard deviation is $\sigma = \sqrt{\sigma^2}$.

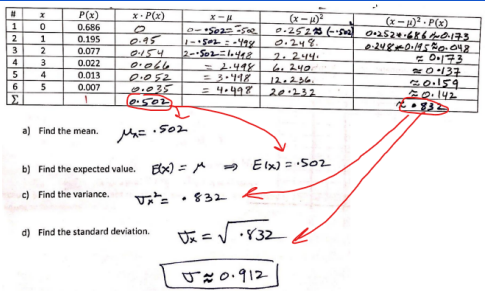
-continuous, it's integral.

-Ex: The number of dogs per household is shown in the table below.

#	x	$P(x)$	$x \cdot P(x)$	$x - \mu$	$(x-\mu)^2$	$(x-\mu)^2 \cdot P(x)$
1	0	0.686				
2	1	0.195				
3	2	0.077				
4	3	0.022				
5	4	0.013				
6	5	0.007				
Σ						

(Fat Probability Distribution chart. Use excel ez)

-Answer https://docs.google.com/spreadsheets/d/1KvEdgwkkzZT9rwVMNmy2V-e8JpYkdQL4OE_PxFZP0yw/edit?usp=sharing



-Properties of the expected value and the variance of a random variable

-Theorem: Let x be a continuous/discrete random variable, and a and c are any real numbers.

-Then $E(ax + c) = aE(x) + c$ (linear transformation on expected value/mean)

-(Direct, Continuous) Let x be a continuous random variable, and let f(x) be the probability distribution function

$$E(ax+c) = \int_{-\infty}^{\infty}(ax+c)f(x)dx \,[\text{defn } E(x)]$$
$$= \int_{-\infty}^{\infty}(axf(x) + cf(x))dx$$
$$= a\underbrace{\int_{-\infty}^{\infty}xf(x)dx}_{E(x)} + c\underbrace{\int_{-\infty}^{\infty}f(x)dx}_{1\,(\text{prob.dist.func.})}$$
$$= aE(x) + c \qquad (\text{aka } \mu_{(ax+c)} = a\mu_x + c)$$

-Special Cases:

$$\text{if } c = 0 \implies E(ax) = aE(x)$$
$$\text{if } a = 0 \implies E(c) = c \text{ (expected value of constant is constant)}$$

-Then $Var(ax + c) = a^2Var(x)$ (linear transformation on variance)

-(Direct)

$$\begin{aligned} Var(ax+c) &= E\Big(((ax+c) - E(ax+c))^2\Big) \text{ [defn } Var(x), \text{ btw } E(x) = \mu] \\ &= E\Big((ax + c - aE(x) - c)^2\Big) \text{ [From above]} \\ &= E\Big((ax - aE(x))^2\Big) \\ &= E\Big(a^2(x - E(x))^2\Big) \\ &= a^2 \underbrace{E\Big((x - E(x))^2\Big)}_{Var(x)} [a \text{ is a constant}] \\ &= a^2 Var(x) \end{aligned}$$

$$\left(\text{aka } \sigma^2_{(ax+c)} = a^2\sigma^2_x\right)$$

-Special Cases:

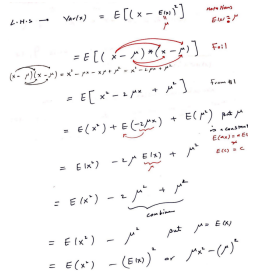
$$\text{if } c = 0 \implies Var(ax) = a^2Var(x)$$
$$\text{if } a = 0 \implies \underbrace{Var(c) = 0}_{Var(0\cdot x+c)=0^2Var(x)}$$

-Then $Var(x) = E(x^2) - [E(x)]^2$ (an alternative to variance)

$$\begin{aligned} Var(x) &= E\Big((x - E(x))^2\Big) \text{ [defn } Var(x)] \\ &= E\Big(x^2 - 2E(x)x + (E(x))^2\Big) \\ &= E(x^2) - E(2E(x)x) + E\Big((E(x))^2\Big) [E(x) = \mu \text{ is a constant}] \\ &= E(x^2) - 2E(x)E(x) + E\Big((E(x))^2\Big) \\ &= E(x^2) - 2(E(x))^2 + (E(x))^2 \\ &= E(x^2) - (E(x))^2 \end{aligned}$$

$$\left(\text{aka } \sigma_{(ax+c)} = \mu_{(x^2)} - \mu^2\right)$$

-Alt mu notation version prof did



-Ex: Redo past Ex with alternative ways using these properties

Example #7: A contractor required by a county planning department to submit 1, 2, 3, 4, or 5 forms (depending on the nature of the project) when applying for a building permit. Let y denote the number of forms required for an application, and suppose the mass function is given by $p(y) = \frac{1}{15}y$ for $y = 1, 2, 3, 4, or\ 5$.

#	x	$P(x)$	$E(x) = \sum x \cdot P(x)$	$E(x^2) = \sum x^2 \cdot P(x)$
1	1	1/15=0.0667	1*0.0667 = 0.0667	1^2*0.0667 = 0.0667
2	2	2/15=0.1333	2*0.1333 = 0.2666	2^2*0.1333 = 0.5332
3	3	3/15=0.2	3*0.2 = 0.6	3^2*0.2 = 1.8
4	4	4/15=0.2667	4*0.2667 = 1.0668	4^2*0.2667 = 4.2672
5	5	5/15=0.3333	5*0.3333 = 1.6665	5^2*0.3333 = 8.3325
∑		1	$E(x) = 3.6666$	$E(x^2) = 14.9996$

a) Find the expected value/mean.

Solution:

$E(x) = 3.6666$

b) Find the variance.

Solution:

$Var(x) = E(x^2) - [E(x)]^2 = 14.9996 - (3.6666)^2 \approx 1.56$

➔ $\sigma^2 = 1.56$

c) Find the standard deviation.

$\sigma = \sqrt{1.56} \approx 1.25$

d) Find $E(2x + 1)$

Shortcut:

$E(2x + 1) = 2E(x) + 1 = 2 * 3.6666 + 1 = 8.3332$

Long Long Long way: eww

$E(g(x)) = \sum g(x)p(x)$

$= \underbrace{(2*1+1)}_{g(1)} * \underbrace{0.0667}_{P(1)} + \underbrace{(2*2+1)}_{g(1)} * \underbrace{0.1333}_{P(2)} + \underbrace{(2*3+1)}_{g(3)} * \underbrace{0.2}_{P(3)} + \underbrace{(2*4+1)}_{g(4)} * \underbrace{0.2667}_{P(4)} + \underbrace{(2*5+1)}_{g(5)} * \underbrace{0.3333}_{P(5)}$

$=3.6666$

e) Find $Var(2x + 1)$

Shortcut:

$Var(2x + 1) = 2^2Var(x) \approx 4 * 1.56 = 6.24$

Long Way: eww

$Var(g(x)) = E([g(x)]^2) - [E(g[x])]^2 = \cdots = 6.24$

Example #8: The Let $f(x) = \frac{15}{4}x^2(1 - x^2)$, $-1 \le x \le 1$.

Example #6: Redo *Example #5* using the alternative formula for variance. The number of dogs per household is shown in the table below.

#	x	$P(x)$	$E(x) = \sum x \cdot P(x)$	$E(x^2) = \sum x^2 \cdot P(x)$
1	0	0.686	0*0.686≈0	0^2*0.686≈0
2	1	0.195	1*0.195≈0.195	1^2*0.195≈0.195
3	2	0.077	2*0.077≈0.154	2^2*0.077≈0.308
4	3	0.022	3*0.022≈0.066	3^2*0.022≈0.198
5	4	0.013	4*0.013≈0.052	4^2*0.013≈0.208
6	5	0.007	5*0.007≈0.035	5^2*0.007≈0.175
∑			$E(x) = 0.502$	$E(x^2) = 1.084$

a) Find the expected value/mean.

$\mu = E(x) = 0.502$

b) Find the variance.

$Var(x) = E(x^2) - [E(x)]^2 = 1.084 - (0.502)^2 \approx 0.83$

➔ $\sigma^2 = 0.832$

c) Find the standard deviation.

$\sigma = \sqrt{0.832} \approx .912$

$$
\begin{aligned}
E(x) &= \int_{-1}^{1} x f(x)dx \\
&= \int_{-1}^{1} x\frac{5}{14}x^2(1-x^2)dx \\
&= \frac{5}{14}\int_{-1}^{1}(x^3-x^5)dx \\
&= \frac{5}{14}\left[\frac{x^4}{4}-\frac{x^6}{6}\right]_{-1}^{1} \\
&= \frac{5}{14}\left[\frac{1}{4}-\frac{1}{6}-\frac{1}{4}+\frac{1}{6}\right] \\
&= \frac{5}{14}[0] = 0
\end{aligned}
$$

$$
\begin{aligned}
Var(x) &= E(x^2) - [E(x)]^2 \\
&= E(x^2) - [0]^2
\end{aligned}
$$

$$
\begin{aligned}
E(x^2) &= \int_{-1}^{1} x^2 f(x)dx \\
&= \int_{-1}^{1} x^2\frac{5}{14}x^2(1-x^2)dx \\
&= \frac{5}{14}\int_{-1}^{1}(x^4-x^6)dx \\
&= \frac{5}{14}\left[\frac{x^5}{5}-\frac{x^6}{6}\right]_{-1}^{1}dx \\
&= \frac{5}{14}\left[\frac{1}{5}-\frac{1}{6}-\left(\frac{-1}{5}-\frac{-1}{6}\right)\right] \\
&= \frac{5}{14}\left[\frac{1}{15}\right] \approx 0.0408
\end{aligned}
$$

$$Var(x) = 0.0408 - [0]^2 = 0.0408$$

-Dependent and Independent Random Variables

-We talked about it earlier, but this is the formal math defn.

Two random variables x and y are independent if the outcomes x do not influence the outcomes of y and vice versa.

Mathematically: $P(x \text{ and } y) = f(x) * g(y)$, or the joint probability density/mass function equal the product of the marginal probability density/mass functions and $x \in \mathcal{D}_1$ and $y \in \mathcal{D}_2$. (Review Chapter 3)

-To truly check if f(x,y) (joint probability distribution function) are independent, take integral wrt x & y (to respective limits).

-If the 2 multiplied together = f(x,y) again, they are independent.

-Continuous VS Discrete (Joint)

Both Random Variables are Continuous	Both Random Variables are Discrete
Definition: A function $f(x,y)$ is called a joint density function if it satisfies the following properties: 1) $f(x,y) \geq 0$ (non-negative) 2) $\iint_D f(x,y)dxdy = 1$	Definition: A function $\mathcal{P}(x,y)$ is called a joint mass function if it satisfies the following properties: 1) $\mathcal{P}(x,y) \geq 0$ (non-negative) 2) $\sum\sum_D \mathcal{P}(x,y) = 1$

-Theorem: If x and y are random variables, then $E(ax \mp by) = aE(x) \mp bE(y)$

-(Direct) Assume that x and y are continuous random variables with a probability joint density function $f(x, y)$

$$
\begin{aligned}
E(ax \pm by) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax \pm by)f(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}axf(x,y)dxdy \pm \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}byf(x,y)dxdy \\
&= a\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dxdy}_{E(x)} \pm b\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dxdy}_{E(y)} \\
&= aE(x) \pm bE(y)
\end{aligned}
$$

-Theorem: If x and y are <u>independent</u> random variables, then $E(xy) = E(x)E(y)$.

-(Direct) Assume that x and y are continuous random variables with a probability joint density function $f(x, y)$

$$
\begin{aligned}
E(xy) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyg(x)h(y)dxdy \ [f(x,y) = g(x)h(y)] \\
&= \underbrace{\int_{-\infty}^{\infty}xg(x)dx}_{E(x)}\underbrace{\int_{-\infty}^{\infty}yh(y)dy}_{E(y)} \\
&= E(x)E(y)
\end{aligned}
$$

-Notes

a) $E\left(\frac{x}{y}\right) \neq \frac{E(x)}{E(y)}$.

b) If x and y are dependent, then $E(xy) \neq E(x)E(y)$.

-Ex: Let $f(x, y) = c(1 - xy)$ [dependent], where $0 \leq x, y \leq 1$ is a joint density function.

-1. Find c

$$
\begin{aligned}
\int\int f(x,y)dxdy &= 1 \\
\int_0^1\int_0^1 c(1-xy)dxdy &= 1 \\
blah\,blah &= \\
c(1-1/4) - c\cdot 0 &= 1 \\
c &= 4/3 \\
\therefore f(x,y) &= \frac{4}{3}(1-xy)
\end{aligned}
$$

where $0 \leq x, y \leq 1$

-R Code:

```
library(cubature)# load the package "cubature"
f <- function(x) { 4/3 * (1-x[1] * x[2] ) } #x[1] is the 1st variable (x) & x[2] is the 2nd (y)
adaptIntegrate(f, lowerLimit = c(0, 0), upperLimit = c(1, 1)) # c(0,0) and c(1,1) correspond to x[1] & x[2]
```

-2. E(x)

$$E(x) = \int\int xf(x,y)dxdy$$
$$= \int_0^1 \int_0^1 x\frac{4}{3}(1-xy)dxdy$$
$$= blah\,blah$$
$$= 4/3 \cdot 1/3$$
$$\approx 0.44$$

-R Code:

```
f <- function(x) {x[1]* 4/3 * (1-x[1] * x[2] ) }
adaptIntegrate(f, lowerLimit = c(0, 0), upperLimit = c(1, 1))
```

-3. Var(x)

$$Var(x) = E\left(x^2\right) - (E(x))^2$$
$$= E\left(x^2\right) - (0.44)^2$$
$$E\left(x^2\right) = \int\int x^2 f(x,y)dxdy$$
$$= blah\,blah$$
$$\approx 0.28$$
$$Var(x) = 0.28 - (0.44)^2 \approx 0.08$$

-4. E(5x)

$$E(5x) = 5E(x) = 5 \cdot 0.44 \approx 2.22$$

-5.

$E(y) = \frac{4}{9}$ (Please verify on your own.)

Also, $Var(y) = \frac{13}{162} \approx 0.08$ (Please verify on your own.)

-6. E(x+y)

$$E(x+y) = E(x) + E(y) = 0.44 + 0.44 \approx 0.89$$

-R Code:

```
f <- function(x) {(x[1]+x[2])* 4/3 * (1-x[1] * x[2] ) }
adaptIntegrate(f, lowerLimit = c(0, 0), upperLimit = c(1, 1))
```

-7. E(xy)

-It's Dependent

$$E(xy) = \int\int xyf(x)dxdy$$
$$= \int_0^1 \int_0^1 xy\frac{4}{3}(1-xy)dxdy$$
$$= blah\,blah$$
$$\approx 0.185$$

-R Code:

```
f <- function(x) {(x[1]*x[2])* 4/3 * (1-x[1] * x[2])}
adaptIntegrate(f, lowerLimit = c(0, 0), upperLimit = c(1, 1))
```

-8. E(x/y)

$$E(xy) = \int\int xyf(x)dxdy$$
$$= \int_0^1 \int_0^1 \frac{x}{y}\frac{4}{3}(1-xy)dxdy$$
$$= blah\,blah$$
$$= \infty$$

-R Code:

```
f <- function(x) {(x[1]/x[2])* 4/3 * (1-x[1] * x[2] ) }
adaptIntegrate(f, lowerLimit = c(0, 0), upperLimit = c(1, 1))
```

-Covariance

-Given two quantitative variables, x and y, we wish to understand the extent to which they vary with one another.

-Def: A measure of how changes in one variable are associated with changes in a second variable. Specifically, it measures the degree to which two variables are linearly associated. However, it is also often used informally as a general measure of how monotonically related two variables are.

-Math Def: $Cov(x, y) = E\{[x - E(x)][y - E(y)]\}$ or $Cov(x, y) = E\{[x - \mu_x][y - \mu_y]\}$

-This is kind of aids to use because E(x) and E(y) like 10 times

-Theorem: $Cov(x, y) = E(xy) - E(x)E(y)$ (aka $Cov(x, y) = \mu_{xy} - \mu_x\mu_y$)

$$Cov(x,y) = E[(x-\mu_x)(y-\mu_y)]$$
$$= E[xy - x\mu_y - \mu_x y + \mu_x\mu_y]$$
$$= E[xy] - E[x\mu_y] - E[\mu_x y] + E[\mu_x\mu_y]$$
$$= \mu_{(xy)} - \mu_x\mu_y \cancelto{0}{-\mu_x\mu_y + \mu_x\mu_y} \text{ [because } \mu \text{ is a constant]}$$
$$\therefore Cov(x,y) = \mu_{(xy)} - \mu_x\mu_y \ (aka\ E(xy) - E(x)E(y))$$

-Ex: Find the $Cov(x, y)$ in the example above. ($f(x, y) = c(1 - xy)$)

$$Cov(x,y) = E(xy) - E(x)E(y) = \frac{5}{27} - \left(\frac{4}{9} \cdot \frac{4}{9}\right) = \frac{1}{81}$$

-Lemma (Property of Cov): If x and y are independent random variables, then $Cov(x, y) = 0$.

$$\begin{aligned}Cov(x,y) &= E(xy) - E(y)E(x) \\ &= E(y)E(x) - E(y)E(x) \; [E(xy) = E(x)E(y) \text{ when independent}] \\ &= 0\end{aligned}$$

-Note: It doesn't go the other way. When Cov(x,y) = 0, it doesn't imply x & y are independent.

-Lemma (Property of Cov): If x and y are two random variables, then (Exam 2)

$$\begin{aligned}Cov(x+a, y+b) &= E([(x+a) - E(x+a)][(y+b) - E(y+b)]) \\ &= E([x+a-(E(x)+a)][y+b-(E(y)+b)]) \; [\text{because } E(x+a) = E(x)+a] \\ &= E([x+a-E(x)-a][y+b-E(y)-b]) \\ &= E([x-E(x)][y-E(y)]) \\ &= E(xy - xE(y) - E(x)y + E(x)E(y)) \\ &= E(xy) - E(xE(y)) - E(E(x)y) + E(E(x)E(y)) \\ &= E(xy) - E(x)E(y) - E(x)E(y) + E(x)E(y) \; [\text{because } E(x) = \mu \text{ is a constant}] \\ &= E(xy) - E(x)E(y) \\ &= Cov(x,y)\end{aligned}$$

-More properties for Cov: [Lesson 30 Properties of Covariance | Introduction to Probability (dlsun.github.io)]

1. *Covariance-Variance Relationship:* $\text{Var}[X] = \text{Cov}[X, X]$
2. *Pulling Out Constants:*

   $\text{Cov}[cX, Y] = c \cdot \text{Cov}[X, Y]$

   $\text{Cov}[X, cY] = c \cdot \text{Cov}[X, Y]$
3. *Distributive Property:*

   $\text{Cov}[X+Y, Z] = \text{Cov}[X, Z] + \text{Cov}[Y, Z]$

   $\text{Cov}[X, Y+Z] = \text{Cov}[X, Y] + \text{Cov}[X, Z]$
4. *Symmetry:* $\text{Cov}[X, Y] = \text{Cov}[Y, X]$
5. *Constants cannot covary:* $\text{Cov}[X, c] = 0$.

//TODO prove these

$$\begin{aligned}Cov(x, yc) &= E((x-E(x))(yc - E(yc))) \; [\text{defn } Cov] \\ &= E((x-E(x))(yc - cE(y))) \;\; [E(xc) = cE(x)] \\ &= E((xyc - xcE(y) - cE(x)y + cE(x)E(y))) \\ &= E(c(xy - xE(y) - E(x)y + E(x)E(y))) \\ &= cE((xy - xE(y) - E(x)y + E(x)E(y))) \;\; [E(xc) = cE(x)] \\ &= cE((x-E(x))(y-E(y))) \\ &= cCov(x,y) \; [\text{defn } Cov]\end{aligned}$$

$$\begin{aligned}Cov(x,x) &= E((x-E(x))(x-E(x))) \\ &= E\left((x-E(x))^2\right) \\ &= Var(x) \; [\text{defn } Var]\end{aligned}$$

-Lemma (Property of Var): If x and y are two random variables, then $Var(ax + by) = a^2Var(x) + b^2Var(y) + 2abCov(x, y)$

$$\begin{aligned}Var(ax+by) &= E([ax+by]^2) - (E[ax+by])^2 \\ &= E(a^2x^2 + 2abxy + b^2y^2) - (aE(x) + bE(y))^2 \\ &= a^2E(x^2) + 2abE(xy) + b^2E(y^2) - a^2[E(x)]^2 - 2abE(x)E(y) - b^2[E(y)]^2 \quad \text{rearrange.} \\ &= a^2E(x^2) - a^2[E(x)]^2 + b^2E(y^2) - b^2[E(y)]^2 + 2abE(xy) - 2abE(x)E(y) \\ &= a^2\underbrace{[E(x^2) - [E(x)]^2]}_{Var(x)} + b^2\underbrace{[E(y^2) - [E(y)]^2]}_{Var(y)} + 2ab\underbrace{[E(xy) - E(x)E(y)]}_{Cov(x,y)} \\ &= a^2Var(x) + b^2Var(y) + 2abCov(x,y)\end{aligned}$$

-Lemma (Property of Var): If x and y are two independent random variables, then $Var(ax + by) = a^2x + b^2y$

$Var(ax+by) = a^2Var(x) + b^2Var(y) + 2abCov(x,y)$ from Lemma 2 above. $Cov(x,y) = 0$ by Lemma 1

$= a^2Var(x) + b^2Var(y) + 2ab * 0$

$= a^2Var(x) + b^2Var(y)$

Note: A special case $a = b = 1$ and x, y are two independent random variables, then $Var(x+y) = Var(x) + Var(y)$

-Examples using Var Properties

-Ex: Find the $Var(x + y)$ in the example above. ($f(x, y) = c(1 - xy)$)

$$Var(x+y) = Var(x) + Var(y) + 2Cov(x,y)$$

$$= \frac{13}{162} + \frac{13}{162} + 2 * \frac{-1}{81} = \frac{11}{81}$$

-Ex: Find the $Var(3x - 5y)$ in the example above. ($f(x, y) = c(1 - xy)$)

$$Var(3x - 5y) = Var\left(\underbrace{3}_{a}x + \underbrace{-5}_{b}y\right)$$

$$= 3^2Var(x) + (-5)^2Var(y) + 2*3*(-5)Cov(x,y)$$

$$= 9 * \frac{13}{162} + (-5)^2 * \frac{13}{162} - 30 * \frac{-1}{81} = \frac{117}{162} + \frac{325}{162} + \frac{30}{81} = 3\frac{8}{81} \approx 3.099$$

-Theorem: Let $x_1, x_2, \dots\dots, x_n$ be <u>independent</u> random variables such that $E(x_i) = \mu$ and $Var(x_i) = \sigma$.

-One of the important applications for statistics is to make an inference **about the population** mean.

-In order to make an inference, we need to **obtain the expected value and the variance of the sample** mean

-Then $E(\bar{x}) = \mu$

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} * n\mu = \mu$$

-Then $Var(\bar{x}) = \sigma^2/n$

Independence ➔ implies that $Cov(x_i, x_j) = 0$

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma = \frac{1}{n^2} * n\sigma = \frac{\sigma}{n}$$

-Fat 4.1 Formulas and Theorems

1) $E(x) = \int_D xf(x)dx$ or $E(x) = \sum_D xP(x)$
2) $E(g(x)) = \int_D g(x)f(x)dx$ or $E(g(x)) = \sum_D g(x)P(x)$
3) $Var(x) = E\{(x - E[x])^2\} = E(x^2) - (E[x])^2$
4) $Cov(x, y) = E\{(x - E[x])(y - E[y])\} = E(xy) - E(x)E(y)$
5) $E(ax \mp by) = aE(x) \mp bE(y)$
1) If x and y are independent, then $P(x \text{ and } y) = p(x) * p(y)$.
6) If x and y are independent, then $E(xy) = E(x)E(y)$.
7) If x and y are independent, then $Cov(x, y) = 0$.
8) Then, $Var(ax + by) = a^2Var(x) + b^2Var(y) + 2abCov(x, y)$.
9) If x and y are independent, then $Var(ax + by) = a^2Var(x) + b^2VarE(y)$

4.2: Binomial Distributions

-Binomial Experiment

-Success or Failure

A **binomial experiment** is a probability experiment that satisfies these conditions.

1. The experiment has a fixed number of trials, where each trial is independent of the other trials.
2. There are only two possible outcomes of interest for each trial. Each outcome can be classified as a success (S) or as a failure (F).
3. The probability of a success is the same for each trial.
4. The random variable x counts the number of successful trials.

-Ex: Which are Binomial Experiments?

-Surveying 100 people to see if they like “Sudsy Soap”

-Yes, either they like or don’t.

-Drawing a card with replacement from a deck and getting a heart card.

-Yes, heart or not.

-Asking 1000 people which brand of cigarettes they smoke.

-No, lots of brands

-Binomial Probability Formula

In a binomial experiment, the probability of exactly x successes in n trials is:

$$P(X = x) = nCxp^xq^{n-x} = \frac{n!}{x!(n-x)!}p^xq^{n-x} \quad , x = 0, 1, 2, \ldots\ldots, n$$

Where n is the number of trails P is the probability of successes in each trial

x is the exact number of successes $q_{=1-P}$ is the probability of failures in each trial

-R Code: dbinom(x, size=n, prob=p) # we don't need q, because q=1-p

-Notation: $x \sim bin(n, p)$ where $x = 0, 1, 2, \ldots, n$

-Ex: A fair coin is tossed 4 times. Let x be the number of heads that will be obtained.

-1. n, p, q?

-n = 4 (trials), p = 1/2 (half heads), q = 1/2 (half tails)

-2. Give the probability mass function and the domain.

$$P(x) = 4Cx\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{4-x} \text{ where } x = 0, 1, 2, 3, 4$$

-3. P(x=0) (the chance to get no head😢)

$$P(0) = 4C0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{4-0} = 4C0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 = 0.0625$$ (😈 it’s low)

-R Code: dbinom(x=0,size=4,prob=1/2)

-4. P(1)

$$P(1) = 4C1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{4-1} = 4C1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = 0.25$$

-5,6,7. P(2), P(3), P(4)

-Ok whatever, just replace

-Ex: A student takes a 10-question, multiple-choice exam with five choices for each question and guesses on each question.

-1. Probability of getting 6 right [aka P(6)]?

-n = 10 (10 questions), p = ⅕ (one correct choice), q = ⅘ (4 wrong choices)

-Plug in for 5

$$P(x) = 10Cx(0.2)^x(0.8)^{10-x} \text{ where } x = 0, 1, 2, \cdots, 10$$

$$P(x = 6) = 10C6(0.2)^6(0.8)^4 \approx 0.0055$$

-R Code: dbinom(x=6,size=10,prob=1/5)

-2. Find the probability of passing the test (Passing the test means getting at least 6 correct out of 10).

$P(x \geq 6) = P(6) + P(7) + P(8) + P(9) + P(10)$ Or independent

$= 10C6(0.2)^6(0.8)^4 + 10C7(0.2)^7(0.8)^3 + 10C8(0.2)^8(0.8)^2 + 10C9(0.2)^9(0.8)^1 + 10C10(0.2)^{10}(0.8)^0$

$= 0.005505024 + 0.000786432 + 0.000073728 + 0.000000687 + 0.0000000102$

$= 0.006366189$

-R Code: sum(dbinom(x=6:10,size=10,prob=1/5)) # sum() and use seq()

-3. Find the probability of getting at least 2 correct out of 10.

$P(x \geq 2) = P(2) + P(3) + \cdots\cdots + P(10)$ use the binomial theorem 9 times. ➔ use the complement.

$= 1 - P(0) - P(1)$

$= 1 - 10C0(0.2)^0(0.8)^{10} - 10C1(0.2)^1(0.8)^9$

$\approx 1 - 0.1074 - 0.2684 = 0.6242$

-Use the complement (kinda like subtracting out summation terms that you don’t want), to cancel out P(0) & P(1) from 100%.

-R Code: sum(dbinom(x=2:10,size=10,prob=1/5)) or 1- sum(dbinom(x=0:1,size=10,prob=1/5))

-The binomial distribution is a probability mass function (pmf)

1) $P(x) = nCxp^xq^{n-x} \geq 0$ 😊
2) Want to show that $\sum_{x=0}^{n} nCxp^xq^{n-x}$=1 ?

-1. is true because

-n is amount of trials, can't be negative
-x is the number of successes, can't be negative
-nCx is positive, a combination
-p & q are probabilities, always ≥ 0
-raised to the x & n-x is always positive. 😊

-2. Direct (using binomial theorem

$$\sum_{x=0}^{n} nCxp^xq^{n-x} = (p+q)^n \text{ [Binomial Theorem]}$$
$$= 1^n \, [p+q=1]$$
$$= 1$$

-If x~$bin(n, p)$, then $E(x)$ = np

$$E(x) = \sum_{x=0}^{n} x \cdot p(x)$$
$$= \sum_{x=0}^{n} xnCxp^xq^{n-x}$$
$$= 0nC0p^0q^{n-0} + \sum_{x=1}^{n} xnCxp^xq^{n-x} \text{ [skip } x=0]$$
$$= 0 + \sum_{x=1}^{n} xnCxp^xq^{n-x}$$
$$= \sum_{x=1}^{n} xnCxp^xq^{n-x}$$
$$= \sum_{x=1}^{n} x\frac{n!}{(n-x)!x!}p^xq^{n-x}$$
$$= \sum_{x=1}^{n} x\frac{n!}{(n-x)!x(x-1)!}p^xq^{n-x}$$
$$= \sum_{x=1}^{n} \frac{n!}{(n-x)!(x-1)!}p^xq^{n-x}$$
$$= \sum_{x=1}^{n} \frac{n!}{(n-x-1+1)!(x-1)!}p^xq^{n-x-1+1}$$
$$= \sum_{x=1}^{n} \frac{n!}{((n-1)-(x-1))!(x-1)!}p^xq^{(n-1)-(x-1)}$$
$$= \sum_{x=1}^{n} \frac{n!}{((n-1)-(x-1))!(x-1)!}p^1p^{x-1}q^{(n-1)-(x-1)}$$
$$= \sum_{x=1}^{n} \frac{n(n-1)!}{((n-1)-(x-1))!(x-1)!}p^1p^{x-1}q^{(n-1)-(x-1)}$$
$$= pn\sum_{x=1}^{n} \frac{(n-1)!}{((n-1)-(x-1))!(x-1)!}p^{x-1}q^{(n-1)-(x-1)}$$
$$\text{Let } (m=n-1),\ (y=x-1)$$
$$= pn\underbrace{\sum_{y=1-1}^{m} \frac{m!}{(m-y)!y!}p^yq^{m-y}}_{\text{binomial thm}}$$
$$= np(p+q)^m$$
$$= np(p+q)^{n-1}$$
$$= np(1)^{n-1}$$
$$= np$$

(QUESTION: Why not m+1 in summation when letting)

-If x~$bin(n, p)$, then $E(x^2)$ = npq+n^2p^2 (Setting up for Var(x))

$$
\begin{aligned}
E(x) &= \sum_{x=0}^{n} x^2 \cdot p(x) \\
&= \sum_{x=0}^{n} x^2 \left(nCxp^x q^{n-x}\right) \\
&= \left(0^2 nC0 p^0 q^{n-0}\right) + \sum_{x=1}^{n} x^2 nCx p^x q^{n-x} [\text{skip } x = 0] \\
&= 0 + \sum_{x=1}^{n} x^2 nCx p^x q^{n-x} \\
&= \sum_{x=1}^{n} x^2 nCx p^x q^{n-x} \\
&= \sum_{x=1}^{n} x^2 \frac{n!}{(n-x)!x!} p^x q^{n-x} \\
&= \sum_{x=1}^{n} x^2 \frac{n!}{(n-x)!x(x-1)!} p^x q^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{(n-x)!(x-1)!} p^x q^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{(n-x+1-1)!(x-1)!} p^x q^{n-x+1-1} \\
&= \sum_{x=1}^{n} x \frac{n!}{((n-1)-(x-1))!(x-1)!} p^x q^{(n-1)-(x-1)} \\
&= \sum_{x=1}^{n} x \frac{n(n-1)!}{((n-1)-(x-1))!(x-1)!} p p^{x-1} q^{(n-1)-(x-1)} \\
&= np \sum_{x=1}^{n} x \frac{(n-1)!}{((n-1)-(x-1))!(x-1)!} p^{x-1} q^{(n-1)-(x-1)} \\
&\quad \text{Let } (m = n-1),\ (y = x-1) \\
&= np \sum_{y=0}^{m} (y+1) \frac{m!}{(m-y)!y!} p^y q^{m-y} \\
&= np \left(\underbrace{\sum_{y=0}^{m} y \frac{m!}{(m-y)!y!} p^y q^{m-y}}_{E(y)} + \underbrace{\sum_{y=0}^{m} \frac{m!}{(m-y)!y!} p^y q^{m-y}}_{\text{Binomial Thm}} \right) \\
&= np(mp + (p+q)^m) \\
&= np((n-1)p + 1) \\
&= np(np - p + 1) \\
&= n^2p^2 - np^2 + np \\
&= n^2p^2 - n(p^2 - p) \\
&= n^2p^2 + np(1-p) \\
&= n^2p^2 + npq
\end{aligned}
$$

-If $x \sim bin(n, p)$, then $Var(x) = npq$

$$
\begin{aligned}
Var(x) &= E(x^2) - (E(x))^2 \\
&= E(x^2) - (np)^2 \ [E(x) = np] \\
&= npq + n^2p^2 - (np)^2 \ [E(x^2) = npq + n^2p^2] \\
&= npq + n^2p^2 - n^2p^2 \\
&= npq
\end{aligned}
$$

-Mean, Variance, and Standard Deviations of a Binomial Distribution

Mean	Variance	Standard Deviation
$\mu = np$	$\sigma^2 = npq$	$\sigma = \sqrt{npq}$

-Ex: A coin is tossed 4 times. Find the mean and variance of the number of heads that will be obtained. (again)

-Slow - calculate each one - way

#	x	$P(x)$	$x \cdot P(x)$	$x^2 \cdot P(x)$
1	0	0.0625	0	0
2	1	0.25	0.25	0.25
3	2	0.375	0.75	1.5
4	3	0.25	0.75	2.25
5	4	0.0625	0.25	1
$\sum$		1	$E(x) = 2$	$E(x^2) = 5$

$\mu = E(x) = 2$

$\sigma^2 = E(x^2) - (E(x))^2 = 5 - 2^2 = 1$

-Cool faster way

$n = 4, p = .5$, q=.5

$\mu = E(x) = n * p = 4 * .5 = 2$

$\sigma^2 = npq = 4 * .5 * .5 = 1$

-Ex: Find the mean, variance, and the standard deviation for each of the values of n and p when the conditions for binomial distributions are met.
-Trash Exs, it's just plug and chug.
-Ex: It has been reported that 83% of federal government employees use email. If a sample of 200 federal government employees is selected.
-1. Binomial Distribution Equation?
n=200, p=0.83, q=1-0.83=0.17
$P(x) = (200Cx)(0.83^x)\left(0.17^{200-x}\right),\ x = 1, 2, 3, \ldots, n$
-2. Find the expected number of employees who use email.
$E(x) = np = 200 \cdot 0.83 = 166$
-3. Find the variance number of employees who use email.
$\sigma^2 = npq = 200 \cdot 0.83 \cdot 0.17 = 28.22$
-4. Find the standard deviation number of employees who use email.
$\sqrt{\sigma^2} = \sqrt{28.22} \approx 5.31$
-5. Find the probability that at least 120 employees use email.
$P(x \geq 120) = P(120) + P(121) + \ldots + P(200)$

4.3a: (Geometric) More Discrete Probability Distributions

-Fat list of Distribution equations (Part 1: Discrete Distributions)

#	Name	distribution	Sample space	Mean /Expected value	Variance
1	Binomial (discrete)	$P(x) = nCxp^x q^{n-x}$	$x = 0, 1, 2, \ldots, n$	$n * p$	$n * p * q$
2	Geometric (discrete)	$P(x) = q^{x-1}p$	$x = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{q}{p^2}$
3	Poisson (discrete)	$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$	$x = 0, 1, 2, \ldots$	λ	λ
4	Exponential (Continuous)	$f(x) = \frac{1}{\lambda} e^{\frac{-x}{\lambda}}$	$x \geq 0$	λ	λ^2
5	Normal (Continuous)	$f(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	$-\infty < x < \infty$	μ	σ^2

-Geometric Distribution
-Many actions in life are repeated until a successful attempt occurs. For example, you might have to send an email several times before it is successfully sent.
-A Probability Mass Function of a discrete random variable x that satisfy these conditions:

1. A trial is repeated until a success occurs.
2. The repeated trials are independent of each other.
3. The probability of success $\boldsymbol{p}$ is the same for each trial.
4. The random variable x represents the number of trials in which the first success occurs.

-The probability that the x trial will be the first success.

$P(x) = q^{x-1}p$, where $x = 1, 2, \ldots\ldots \infty$, and $q = 1 - p$.

-Notation

$x \sim Geom(p) \quad , x = 1, 2, \ldots\ldots \infty$

-R Code:
dgeom(x=x-1, prob = p) #x represents failures before first success

-Ex: The probability that you will make a sale on any given telephone call is 0.19. Find the probability that you…
-Find function: Let x represent the number of calls in which the first sale occurs.
p = 0.19, q = 1 - p = 0.81

$P(x) = (0.81)^{x-1} * 0.19$, $x = 1, 2, 3, 4, 5, 6 \ldots\ldots$

-1. Make your first sale on the fifth call.
-x = 5

$P(5) = (0.81)^{5-1} * 0.19 = 0.08178877$ -R Code: dgeom(x=4, prob = 0.19)

-2. Make your first sale on the first, second, or third call.
-x = 1, 2, or 3

$P(1) + P(2) + P(3) = (0.81)^{1-1} * 0.19 + (0.81)^{2-1} * 0.19 + (0.81)^{3-1} * 0.19 = .19 + .1539 + .1247 \approx 0.4686$

-R Code: sum(dgeom(x=0:2, prob = 0.19))

-3. Do not make a sale on the first three calls.
-x != 1, 2, or 3

$1 - \{P(1) + P(2) + P(3)\} = 1 - 0.4686 = 0.5314$ -R Code: 1-sum(dgeom(x=0:2, prob = 0.19))

-Alt: Fail 3 times

$(0.81)^3 \approx 0.5314$

-Ex: An auto parts seller finds that 1 in every 100 parts sold is defective. Find the probability that…
-Find function: Let x represent the number of sales in which the first defective part is sold.
p = 0.01, q = 0.99

$P(x) = (0.99)^{x-1} * 0.01$, $x = 1, 2, 3, \ldots\ldots$

-1. The first defective part sold is the tenth part sold.
-x = 10

$P(10) = (0.99)^{9} * 0.01 = 0.009$ R Code: dgeom(x=9, prob = .01)

-2. The first defective part is the first, second, or third part sold.
x = 1, 2, or 3

$P(1) + P(2) + P(3) = (0.99)^{1-1} * 0.01 + (0.99)^{2-1} * 0.01 + (0.99)^{3-1} * 0.01$

$= .01 + .0099 + .009801 \approx 0.029701$

-R Code: sum(dgeom(x=0:2, prob = .01))

-3. None of the first 10 parts sold are defective.
-x != 1, 2, …, 10
-aka selling 10 parts non defective back to back

P(first 10 parts sold are not defective) = $(0.99)^{10} \approx 0.9044$

-Alt: 1 - P(first 10 parts are defective)

R Code: 1-sum(dgeom(x=0:9, prob = .01))

-Geometric Series Review

$\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ if $|r| < 1$

If a and r are real numbers and $r \neq 0$, then

$$\sum_{j=0}^{n} ar^j = \begin{cases} \frac{ar^{n+1} - a}{r-1} & \text{if } r \neq 1 \\ (n+1)a & \text{if } r = 1. \end{cases}$$

-Ex:

$\sum_{x=0}^{\infty} 3(\frac{2}{5})^x = 3(\frac{2}{5})^0 + 3(\frac{2}{5})^1 + 3(\frac{2}{5})^2 + 3(\frac{2}{5})^3 + \cdots$ (computer can't do it, except symbolab and wolfram cheese)

-Use geometric series

$\sum_{x=0}^{\infty} 3(\frac{2}{5})^x$ ➔ geometric series where a=3 and $r = \frac{2}{5}$

$\sum_{x=0}^{\infty} 3(\frac{2}{5})^x = \frac{3}{1-\frac{2}{5}} = \frac{3}{\frac{5}{5}-\frac{2}{5}} = \frac{3}{\frac{3}{5}} = 3 * \frac{5}{3} = 5$

-Ex: Lower bound = 1

$\sum_{x=1}^{\infty} 3(\frac{2}{5})^x$

-Skip/Subtract the 0th term from the Sum of 0 to ∞.

$$\sum_{x=1}^{\infty} 3\left(\frac{2}{5}\right)^x = \sum_{x=0}^{\infty} 3\left(\frac{2}{5}\right)^x - 3\left(\frac{2}{5}\right)^0 = \frac{3}{1-\frac{2}{5}} - 3 = 2$$

-Alt: Replace lower bound (x=1) to (y=x-1, y+1=x)

We must restart the boundary by substitution:

Let $y = x - 1$

$x = 1$ ➔ $y = 1 - 1 = 0$ ☺ the lower bound now is starting at 0.

$x = \infty$ ➔ $y = \infty - 1 = \infty$

Also, $y = x - 1$ ➔ $y + 1 = x$

$\sum_{x=1}^{\infty} 3(\frac{2}{5})^x = \sum_{y=0}^{\infty} 3(\frac{2}{5})^{(y+1)} = \sum_{y=0}^{\infty} 3\left(\frac{2}{5}\right)^1 * \left(\frac{2}{5}\right)^y$

$= \left(\frac{2}{5}\right) \sum_{y=0}^{\infty} 3\left(\frac{2}{5}\right)^y = \frac{2}{5} * \left(\frac{3}{1-\frac{2}{5}}\right) = \frac{2}{5} * \left(\frac{3}{\frac{5}{5}-\frac{2}{5}}\right) = \frac{2}{5} * \left(\frac{3}{\frac{3}{5}}\right) = \frac{2}{5} * \left(3 * \frac{5}{3}\right) = 2$

-Change of Index Review

$$\sum_{j=1}^{5} j^2 = \sum_{k=0}^{4} (k+1)^2$$ eww

-Proof for Geometric Distribution is a pmf

-1. $P(x) = q^{x-1}p \geq 0$

-Yessir, q is always + because x-1 is always +. p is always + because it's a probability.

-2. Sum of all probabilities = 1

-Direct by subtracting the 0th term (fail? 😢 idk how to)

$$\sum_{x=1}^{\infty} P(x) = 1$$

$$\sum_{x=1}^{\infty} q^{x-1}p = 1$$

$$\sum_{x=0}^{\infty} q^{x-1}p - q^{0-1}p = 1$$

$$\sum_{x=0}^{\infty} q^{x-1}p - \frac{p}{q} = 1$$

$$\sum_{x=0}^{\infty} (p+1)^{x-1}p^1 - \frac{p}{p+1} = 1$$

$$fat\ hole =$$

-Direct by letting x w/ k=1-x, x=k+1 (success 😎)

$$\sum_{x=1}^{\infty} q^{x-1}p = 1$$

$$\sum_{k+1=1}^{\infty} q^{k+1-1}p = 1 \ [k = 1 - x,\ x = k + 1]$$

$$\sum_{k=0}^{\infty} q^{k}p = 1$$

$$\frac{p}{1-q} = 1 \ [\text{geo. series}]$$

$$\frac{p}{p} = 1$$

$$1 = 1$$

-Prove that if $x \sim Geom(p)$, then $E(x) = 1/p$.

-Issue, this is not a geometric series

$$E(x) = \sum_{x=1}^{\infty} xq^{x-1}p$$

-Let's start at what we proof above instead, sum of the geometric distribution = 1

$$\sum_{x=1}^{\infty} q^{x-1}p = 1$$

$$p\sum_{x=1}^{\infty} q^{x-1} = 1$$

$$\sum_{x=1}^{\infty} q^{x-1} = \frac{1}{p}$$

$$\sum_{x=1}^{\infty} q^{x-1} = \frac{1}{(1-q)}$$

$$\frac{d}{dq}\left(\sum_{x=1}^{\infty} q^{x-1}\right) = \frac{d}{dq}\left(\frac{1}{(1-q)}\right)$$

$$\sum_{x=1}^{\infty}\left(\frac{d}{dq}\left(q^{x-1}\right)\right) = \frac{d}{dq}\left((1-q)^{-1}\right) \text{ [derive sum = derive what's being sum]} \qquad \left(\sum_{x=1}^{\infty} q^{x-1} = 1 + q + q^2 + q^3 + \cdots\right)$$

$$\sum_{x=2}^{\infty}(x-1)q^{x-2} = -1(1-q)^{-2}(-1) \text{ [derive sum of power series, drop 1st term]}$$

$$\sum_{x=2}^{\infty}(x-1)q^{x-2} = (1-q)^{-2}$$

Let $y = x - 1$ ➔ if $x = 2$ ➔ y=2-1 =1

➔ if $x = \infty$ ➔ y=∞ −1=∞

$y = x - 1$ ➔ $y + 1 = x$

$$\sum_{x=2}^{\infty}(x-1)q^{x-2} = (1-q)^{-2}$$

$$\sum_{y+1=2}^{\infty}(y+1-1)q^{y+1-2} = (1-q)^{-2} \text{ [let } y = x-1]$$

$$\sum_{y=1}^{\infty} yq^{y-1} = \frac{1}{p^2}$$

$$\sum_{y=1}^{\infty} yq^{y-1}p = \frac{1}{p}$$

$$E(y) = \frac{1}{p}$$

-Prove that if x~$Geom(p)$, then Var(x) = q/p^2.

$$Var(x) = E(x^2) - (E(x))^2$$

$$= E(x^2) - \left(\frac{1}{p}\right)^2$$

-Like with E(x), E(x^2)'s multiplying x^2 stops it from being a geometric series. Let's start at E(x)

$$\sum_{x=1}^{\infty} xq^{x-1}p = \frac{1}{p}$$
$$\sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{p^2}$$
$$\sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{(1-q)^2}$$
$$\frac{d}{dq}\left(\sum_{x=1}^{\infty} xq^{x-1}\right) = \frac{d}{dq}\left((1-q)^{-2}\right)$$
$$\sum_{x=1}^{\infty} \frac{d}{dq}\left(xq^{x-1}\right) = -2(1-q)^{-3}(-1)$$
$$\sum_{x=2}^{\infty}\left((x-1)xq^{x-2}\right) = 2(1-q)^{-3}$$
$$\sum_{y+1=2}^{\infty}\left((y+1-1)(y+1)q^{y+1-2}\right) = \frac{2}{(1-q)^3}[\text{let } y = x-1]$$
$$\sum_{y=1}^{\infty}\left((y)(y+1)q^{y-1}\right) = \frac{2}{(1-q)^3}$$
$$\sum_{y=1}^{\infty} q^{y-1}\left(y^2+y\right) = \frac{2}{p^3}$$
$$\sum_{y=1}^{\infty}\left(q^{y-1}y^2 + q^{y-1}y\right) = \frac{2}{p^3}$$
$$\sum_{y=1}^{\infty} q^{y-1}y^2 + \sum_{y=1}^{\infty} q^{y-1}y = \frac{2}{p^3}$$
$$\underbrace{\sum_{y=1}^{\infty} y^2q^{y-1}p}_{E(y^2)} + \underbrace{\sum_{y=1}^{\infty} yq^{y-1}p}_{E(y)=\frac{1}{p}} = \frac{2}{p^2}$$
$$E(y^2) + \frac{1}{p} = \frac{2}{p^2}$$
$$E(y^2) = \frac{2}{p^2} - \frac{1}{p}$$
$$E(y^2) = \frac{2}{p^2} - \frac{1-q}{p^2}$$
$$E(y^2) = \frac{2-(1-q)}{p^2}$$
$$E(y^2) = \frac{1}{p^2} + \frac{q}{p^2}$$

-Then E(y^2) + E(y)^2

$$Var(x) = E(x^2) - (E(x))^2$$
$$= E(x^2) - \left(\frac{1}{p}\right)^2$$
$$= \frac{1}{p^2} + \frac{q}{p^2} - \left(\frac{1}{p}\right)^2$$
$$= \frac{1}{p^2}^{0} + \frac{q}{p^2} - \frac{1}{p^2}^{0}$$
$$= \frac{q}{p^2}$$

-Ex: A patient of blood type $O-$ is waiting for a suitable matching kidney donor for a transplant.
If the probability that a randomly selected donor is a suitable match is p = 0.05,
what is the expected number of donors who will be tested until a matching donor is found?

-Let x be number of donors who will be tested until a matching donor is found.

$$x \sim Geo(0.05)$$
$$P(x) = (.95)^{x-1} * 0.5$$
$$E(x) = \frac{1}{0.05} = 20$$

20 is the expected number of donors who will be tested until a matching donor is found.

4.3b: (Poisson) More Discrete Probability Distributions

-Poisson Distribution

In a binomial experiment, you are interested in finding the probability of a specific number of successes in a given number of trials. Suppose that you want to know the probability that a specific number of occurrences take place within a given unit of time, area, or volume. For instance, to determine the probability that an employee will take 15 sick days within a year, you can use the **Poisson distribution.** (aka use mean to find probability)

-Probability Mass Function, a random variable x that satisfies these conditions:

1. The experiment consists of counting the number of times x an event occurs in a given interval. The interval can be an interval of time, area, or volume.
2. The probability of the event occurring is the same each interval.
3. The number of occurrences in one interval is independent of the number of occurrences in other intervals.

The probability of exactly x occurrences in an interval is $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ where $x = 0, 1, 2, \ldots, \infty, e \approx 2.71828$ and λ is the rate/mean number occurrences per interval unit.

-Notation: $x \sim Pois(\lambda)$ where x = 0, 1, 2, … ∞

-R Code: dpois(x=4, lambda=3)

-Ex: The mean number of accidents per month at a certain intersection is 3.
What is the probability that in a given month four accidents will occur at this intersection?

-Let x be the mean number of accidents per month at the intersection

-$x \sim Pois(\lambda = 3)$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{3^x e^{-3}}{x!} \text{ where } x = 0, 1, 2, \ldots\ldots$$

-Four accidents P(4)

$$P(4) = \frac{3^4 e^{-3}}{4!} \approx 0.1680$$

-R Code: dpois(x=4,lambda = 3)

-Ex: A newspaper finds that the mean number of typographical errors per page is four.
Find the probability that the number of typographical errors found on any given page is…
-Find function: Let x be the number of typographical errors per page.
-$x \sim Pois(\lambda = 4)$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{4^x e^{-4}}{x!} \text{ where } x = 0, 1, 2, \dots \dots \infty$$

-1. Exactly three. (x=3)

$$P(3) = \frac{4^3 e^{-4}}{3!} \approx 0.195$$

-R Code: dpois(x=3,lambda = 4)
-2. At most three. (x=0, 1, 2, or 3 aka x≤3)

$$P(x \le 3) = P(0) + P(1) + P(2) + P(3)$$

$$= \frac{4^0 e^{-4}}{0!} + \frac{4^1 e^{-4}}{1!} + \frac{4^2 e^{-4}}{2!} + \frac{4^3 e^{-4}}{3!}$$

$$\approx 0.018 + 0.073 + 0.147 + .195$$

$$\approx 0.433$$

-R Code: sum(dpois(x=0:3,lambda = 4))
-3. More than 3. (x≥3)
-use complement

$$P(x > 3) = P(4) + P(5) + \cdots \dots \dots \dots \dots$$

$$= 1 - P(0) - P(1) - P(2) - P(3)$$

$$= 1 - 0.433$$

$$= 0.567$$

-R Code: 1-sum(dpois(x=0:3,lambda = 4))

-Prove that the Poisson distribution is a pmf
-1. P(x) ≥ 0
-Yessir, everything positive
-2. Sum of P(x) to inf = 1?
-First, Taylor/Maclaurin/Power Series for e^x

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$$

-Proof Direct

Assume that the exponent function can be represented as a series with unknown coefficients:

$$e^x = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots = \sum_{n=0}^{\infty} a_n x^n$$

Recall the fundamental property of exponent $\frac{d}{dx}(e^x) = e^x$. Applying this property to the series for of exponent, we get

$$\begin{aligned}
&\frac{d}{dx} e^x = e^x \implies \frac{d}{dx}\left(\sum_{n=0}^{\infty} a_n x^n\right) = \sum_{n=0}^{\infty} a_n x^n \\
&\implies \frac{d}{dx}(a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots \\
&\implies 0 + a_1 + 2\,a_2\,x + 3\,a_3\,x^2 + \ldots = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots \\
&\implies \sum_{n=1}^{\infty} n\,a_n\,x^{n-1} = \sum_{n=0}^{\infty} a_n x^n \iff \sum_{n=0}^{\infty} (n+1)\,a_{n+1}\,x^n = \sum_{n=0}^{\infty} a_n x^n \\
&\implies (n+1)\,a_{n+1} = a_n \\
&\implies a_{n+1} = \frac{a_n}{n+1}
\end{aligned}$$

The last equation can be rewritten as $a_n = \frac{1}{n} a_{n-1}$, so that

$$a_n = \frac{1}{n} a_{n-1} = \frac{1}{n}\frac{1}{n-1} a_{n-2} = \frac{1}{n}\frac{1}{n-1}\frac{1}{n-2} a_{n-3} = \ldots = \frac{1}{n!} a_0 \qquad (1.1)$$

Observer that $e^0 = 1$, so we can write

$$e^x\Big|_{x=0} = (a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots)\Big|_{x=0} = a_0 = 1$$

This fact combined with equation (1.1) gives us the explicit expression for coefficient $a_n = \frac{1}{n!}$.
Therefore we finally write

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

//TODO (minor) understand this proof
-Direct

$$\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = 1$$

$$e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1$$

$$e^{-\lambda} e^{\lambda} = 1 \; [e^x \text{ Taylor Series}]$$

$$e^0 = 1$$

$$1 = 1$$

-Prove that if $x \sim Pois(\lambda)$, then $E(x) = \lambda$

$$
\begin{aligned}
E(x) &= \sum_{x=0}^{\infty} x\frac{\lambda^x e^{-\lambda}}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x x}{x!} \\
&= e^{-\lambda}\left(\frac{\lambda^0 0}{0!}+\sum_{x=1}^{\infty}\frac{\lambda^x}{x!}\right) \text{ [take out 1st term, else explodes later]} \\
&= e^{-\lambda}\left(\frac{0}{1}+\sum_{x=1}^{\infty}\frac{\lambda^x}{x!}\right) \\
&= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x x}{[(x)(x-1)(x-2)\dots(1)]} \\
&= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!} \\
&= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\
&= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\
&= e^{-\lambda}\lambda\left(\frac{1}{0!}+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\dots\right) \\
&= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\lambda e^{\lambda} \\
&= \lambda
\end{aligned}
$$

-Prove that if $x \sim Pois(\lambda)$, then Var(x) = λ

-We need that $E(x^2)$ cuh

$$
\begin{aligned}
E(x^2) &= \sum_{x=0}^{\infty} x^2\frac{\lambda^x e^{-\lambda}}{x!} \\
&= e^{-\lambda}\left(0^2\frac{\lambda^0}{0!}+\sum_{x=1}^{\infty}x^2\frac{\lambda^x}{x!}\right) \\
&= e^{-\lambda}\left(0+\sum_{x=1}^{\infty}x^2\frac{\lambda^x}{x!}\right) \\
&= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x x^2}{x!} \\
&= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x x^2}{[(x)(x-1)(x-2)\dots(1)]} \\
&= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x x}{(x-1)!} \\
&= e^{-\lambda}\sum_{y+1=1}^{\infty}\frac{\lambda^{y+1}(y+1)}{(y+1-1)!} \text{ [let } y = x-1] \\
&= e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^{y+1}y+\lambda^{y+1}}{y!} \\
&= \lambda\sum_{y=0}^{\infty}\frac{\lambda^y y e^{-\lambda}+\lambda^y e^{-\lambda}}{y!} \\
&= \lambda\underbrace{\sum_{y=0}^{\infty}y\frac{\lambda^y e^{-\lambda}}{y!}}_{E(y)=\lambda}+\lambda\underbrace{\sum_{y=0}^{\infty}\frac{\lambda^y e^{-\lambda}}{y!}}_{\sum_{y=0}^{\infty}P(x)=1} \\
&= \lambda\lambda+\lambda 1 \\
&= \lambda^2+\lambda
\end{aligned}
\qquad
\begin{aligned}
Var(x) &= E(x^2)-(E(x))^2 \\
&= (\lambda^2+\lambda)-\lambda^2 \\
&= \lambda
\end{aligned}
$$

Ez game

5.1: Exponential Distribution Functions

-R Rounding/Significant Digits

-options(digits=22) # ask R to round to 22 significant digits

-Continuous Functions

-Points dont have weights, interval do

Instead of modeling the number of accidents on a junction or the number of typos, we may be interested in how long it will take until an accident occurs on the junction or how long it takes until we make a typo. In order to achieve that objective, we use the exponential probability distribution function (**pdf**).

-Exponential Distribution Function (w/ rate/average λ)

$$f(x) = \frac{1}{\lambda} e^{\frac{-x}{\lambda}} \quad , x \geq 0$$

-Notation $x \sim Exp(\lambda)$ where $x \geq 0$ and λ is the rate/average.
-Useful with things like

i. Waiting time in line at a bank.
ii. Distance between consecutive knots on a thick plank of wood.
iii. Time between two consecutive stops at a bus stop.
iv. Time between consecutive epileptic episodes.

-Ex: Lightbuld lifetime

Example #1: The average lifetime of a certain brand of compact florescent lightbulbs is 8500 hours. The time until the lightbulb fails can be modeled as the exponential random variable $\boldsymbol{Exp(8500)}$. Let x denote the time it takes the lightbulb to fail (in hours).

-1. Give the pdf

λ = 8500 hrs

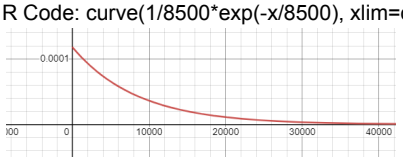
$$x \sim Exp(8500) \rightarrow f(x) = \frac{1}{8500} e^{-\frac{x}{8500}} \quad , x \geq 0$$

-R Code:

```
y <- function(x) {1/8500*exp(-x/8500)} # define the integrated function
integrate(y, lower = 0, upper = Inf) # integrate the function, should = 1
```

-2. Graph the pdf

R Code: curve(1/8500*exp(-x/8500), xlim=c(0,20000))



-3. Find the probability that the lightbulb last between 8,000 to 10,000 hours.

-Integral of e^x review

$$\int e^{ax} dx = \frac{e^{ax}}{a} + c = \frac{1}{a} e^{ax} + c$$

-Lower limit is 8000, Upper is 10000, integrate

$$P(8{,}000 < x < 10{,}000) = \int_{8{,}000}^{10{,}000} \frac{1}{8500} e^{-\frac{x}{8500}} dx = \frac{1}{8500} \int_{8{,}000}^{10{,}000} e^{-\frac{x}{8500}} dx = \frac{1}{8500} * -\frac{8500}{1} e^{-\frac{x}{8500}} \Big|_{8{,}000}^{10{,}000}$$

$$= -e^{-\frac{x}{8500}} \Big|_{8{,}000}^{10{,}000} = -e^{-\frac{10{,}000}{8500}} - \left(-e^{-\frac{8{,}000}{8500}}\right) = -e^{-\frac{100}{85}} + e^{-\frac{80}{85}} \approx 0.0818$$

-R Code:

```
y <- function(x) {1/8500*exp(-x/8500e)} # define the integrated function
integrate(y, lower = 8000, upper = 10000) # integrate the function
```

-4. Find the probability that the lightbulb will last exactly 8,000 hours.

-It's not P(8000) because it's continuous. So its P(x = 8000) = 0

$$P(x = 8{,}000) = P(8{,}000 < x < 8{,}000) = \int_{8{,}000}^{8{,}000} \frac{1}{8500} e^{-\frac{x}{8500}} dx = 0$$

$$P(x = 8{,}000) \neq f(8{,}000)$$

-5. Find the probability that the lightbulb will last at most 8,000 hours.

-P(0 < x < 8000) ez

$$P(0 < x < 8{,}000) = \int_{0}^{8{,}000} \frac{1}{8500} e^{-\frac{x}{8500}} dx = \frac{1}{8500} \int_{0}^{8{,}000} e^{-\frac{x}{8500}} dx$$

$$= \frac{1}{8500} * -\frac{8500}{1} e^{-\frac{x}{8500}} \Big|_{0}^{8{,}000} = -e^{-\frac{x}{8500}} \Big|_{0}^{8{,}000} = -e^{-\frac{8{,}000}{8500}} - \left(-e^{-\frac{0}{8500}}\right)$$

-R Code:

```
y <- function(x) {1/8500*exp(-x/8500)} # define the integrated function
integrate(y, lower = 0, upper = 8000) # integrate the function
```

-6. Find the probability that the lightbulb will last more than 10,000 hours.

-P(x > 10000), so 10000 to ∞

$$P(x > 10{,}000) = \int_{10{,}000}^{\infty} \frac{1}{8500} e^{-\frac{x}{8500}} dx = \frac{1}{8500} \lim_{b \to \infty} \int_{10{,}000}^{b} e^{-\frac{x}{8500}} dx$$

$$= \frac{1}{8500} * -\frac{8500}{1} \lim_{b \to \infty} e^{-\frac{x}{8500}} \Big|_{10{,}000}^{b} = -\lim_{b \to \infty} e^{\frac{-x}{8500}} \Big|_{10{,}000}^{b}$$

$$= \lim_{b \to \infty} -e^{-\frac{b}{8500}} - \left(-e^{-\frac{10{,}000}{8500}}\right) = 0 + e^{-\frac{100}{85}} \approx 0.3084$$

(improper integral, use limit to ∞)

-R Code:

```
y <- function(x) {1/8500*exp(-x/8500)} # define the integrated function
integrate(y, lower = 10000, upper = Inf) # integrate the function
```

-7. Find the median number of hours that the lightbulb last.

-The median is the middle data entry, if it was discrete. Continuous probability means the median is 0.5. So, we just work backwards now.

-$\boldsymbol{P(x < a) = 0.5}$ (Let a be the Number of hours where 50% of the lightbulbs fail before it.) Solve for a and ez game.

$$\int_0^a \frac{1}{8500} e^{-\frac{x}{8500}} dx = 0.5$$

$$\frac{1}{8500}\int_0^a e^{-\frac{x}{8500}} dx = 0.5$$

$$\frac{1}{8500} * -\frac{8500}{1} e^{-\frac{x}{8500}}\Big|_0^a = 0.5$$

$$-e^{-\frac{x}{8500}}\Big|_0^a = 0.5$$

$$-e^{-\frac{a}{8500}} - \left(-e^{-\frac{0}{8500}}\right) = 0.5$$

$$-e^{-\frac{a}{8500}} + 1 = 0.5 \rightarrow -e^{-\frac{a}{8500}} = -0.5 \rightarrow e^{-\frac{a}{8500}} = 0.5 \rightarrow -\frac{a}{8500} = \ln(0.5) \rightarrow a = -8500 * \ln(0.5) \approx 5{,}891.75$$

-Ex: Rainfall duration

Example #2: Data collected at Toronto Pearson International Airport suggests that an exponential distribution with $\boldsymbol{\lambda} = \frac{100}{37}$ is a good model for rainfall duration in hours (*Urban Storm Management Planning with Analytical Probabilistic Models, 2000, p. 69*)

-1. Find pdf

-λ = 100/37, plug in and ez

Let x denote the rainfall duration in hours.

$$\lambda = \frac{100}{37} \rightarrow \ x \sim Exp(\frac{100}{37}) \rightarrow f(x) = \frac{37}{100} e^{-\frac{37x}{100}} \quad , x > 0$$

-R Code to check full interval of P(x) = 1:

```
y  <- function(x) {37/100*exp(-37*x/100)} # define the integrated function
integrate(y, lower = 0, upper = Inf) # integrate the function, should = 1
```

-2. What proportion of rainfall durations at this location are at least 2 hours?

-P(x > 2)

$$= \int_2^\infty \frac{37}{100} e^{-\frac{37x}{100}} dx = \frac{37}{100} \lim_{b\to\infty} \int_2^b e^{-\frac{37x}{100}} dx$$

$$= \frac{37}{100} * -\frac{100}{37} \lim_{b\to\infty} e^{-\frac{37x}{100}}\Big|_2^b = -\lim_{b\to\infty} e^{-\frac{37x}{100}}\Big|_2^b$$

$$= \lim_{b\to\infty} -e^{-\frac{37b}{100}} - \left(-e^{-\frac{37*2}{100}}\right) = 0 + e^{-\frac{37}{50}} \approx 0.4771$$

-R Code: integrate(y, lower = 2, upper = Inf)

-3. What proportion of rainfall durations at this location are at most 3 hours?

-P(x < 3)

$$P(0 < x < 3) = \int_0^3 \frac{37}{100} e^{-\frac{37x}{100}} dx = \frac{37}{100}\int_0^3 e^{-\frac{37x}{100}} dx$$

$$= \frac{37}{100} * -\frac{100}{37} e^{-\frac{37x}{100}}\Big|_0^3 = -e^{-\frac{37x}{100}}\Big|_0^3$$

$$= -e^{-\frac{37*3}{100}} - \left(-e^{-\frac{0}{100}}\right) = 1 - e^{-\frac{111}{100}} \approx 0.6704$$

-R Code: integrate(y, lower = 0, upper = 3)

-4. What proportion of rainfall durations at this location are between 2 and 3 hours?

-P(2 < x < 3)

$$P(2 < x < 3) = \int_2^3 \frac{37}{100} e^{-\frac{37x}{100}} dx = \frac{37}{100}\int_2^3 e^{-\frac{37x}{100}} dx$$

$$= \frac{37}{100} * -\frac{100}{37} e^{-\frac{37x}{100}}\Big|_2^3 = -e^{-\frac{37x}{100}}\Big|_2^3$$

$$= -e^{-\frac{37*3}{100}} - \left(-e^{-\frac{37*2}{100}}\right) = -e^{-\frac{111}{100}} + e^{-\frac{74}{100}} \approx 0.1476$$

-R Code: integrate(y, lower = 2, upper = 3)

-5. What must the duration of a rainfall be to place it among the longest 5% of all times?

-2 ways to set it up

Let a be the rainfall duration in hours where 5% of rainfall events last longer than a.

$$P(x > a) = 0.05 \qquad\qquad P(x < a) = 0.95$$

$$\int_a^\infty \frac{37}{100} e^{-\frac{37x}{100}} dx = 0.05 \qquad \text{or} \qquad \int_0^a \frac{37}{100} e^{-\frac{37x}{100}} dx = 0.95$$

-P(x < a) = 0.95 because screw ∞ limits

$$\frac{37}{100}\int_0^a e^{-\frac{37x}{100}}dx = 0.95$$

$$\frac{37}{100} * -\frac{100}{37}e^{-\frac{37x}{100}}\Big|_0^a = 0.95$$

$$-e^{-\frac{37x}{100}}\Big|_0^a = 0.5$$

$$-e^{-\frac{37a}{100}} - \left(-e^{-\frac{0}{100}}\right) = 0.95$$

$$-e^{-\frac{37a}{100}} + 1 = 0.95 \rightarrow -e^{-\frac{37a}{100}} = -0.05 \rightarrow e^{-\frac{37a}{100}} = 0.05 \rightarrow -\frac{37a}{100} = \ln(0.05) \rightarrow a = -\frac{100}{37} * \ln(0.05) \approx 8.10$$

-Formal Finding Bound from Probability (see above example)

$$f(x) = \int_0^a \lambda e^{\frac{-x}{\lambda}} dx$$
$$a = -\lambda \ln(f(x))$$

-Prove that the Exponential distribution is a pdf

-1. f(x) is non-negative when x is non-negative

-Yessir. λ is positive, e^(-x/λ) is positive, probability is always positive, ez

-2. The full interval probabilities of f(x) = 1

$$\int_0^\infty \frac{1}{\lambda} e^{\frac{-x}{\lambda}} dx = 1$$
$$\left[\frac{-\lambda}{\lambda} e^{\frac{-x}{\lambda}}\right]_0^\infty = 1$$
$$\lim_{a\to\infty} -e^{\frac{-a}{\lambda}} + e^{\frac{0}{\lambda}} = 1$$
$$\lim_{a\to\infty} e^{\frac{-a}{\lambda}} + 1 = 1$$
$$e^{\frac{-\infty}{\lambda}} + 1 = 1$$
$$e^{-\infty} + 1 = 1$$
$$1 = 1$$

-Expected Value and Variance

-First, a review of integration by parts

-Pick one function to be you, another to be dv (dv includes dx, so that integration both sides cancels it out).

$$\int_a^b u\,dv = u\,v\Big|_a^b - \int_a^b v\,du$$

-Ex:

$$\int \underbrace{x}_{u} \underbrace{e^x dx}_{dv}$$

-x~$Exp(\lambda)$, then $E(x) = \lambda$.

$$\int_0^\infty x\frac{1}{\lambda}e^{\frac{-x}{\lambda}}dx = \frac{1}{\lambda}\int_0^\infty xe^{\frac{-x}{\lambda}}dx$$

$$\text{by parts} \quad \begin{bmatrix} u = x & dv = e^{\frac{-x}{\lambda}}dx \\ du = 1 & v = -\lambda e^{\frac{-x}{\lambda}} \end{bmatrix}$$

$$= \frac{1}{\lambda}\int_0^\infty u dv$$
$$= \frac{1}{\lambda}\left(\left[-x\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty - \int_0^\infty 1\left(-\lambda e^{\frac{-x}{\lambda}}\right)dx\right)$$
$$= \frac{1}{\lambda}\left(\left[-x\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty - \left[\lambda^2 e^{\frac{-x}{\lambda}}\right]_0^\infty\right)$$
$$= \frac{1}{\lambda}\left(\left[\lim_{a\to\infty} -a\lambda e^{\frac{-a}{\lambda}} + 0\lambda e^{\frac{-0}{\lambda}}\right] - \left[\lim_{a\to\infty} \lambda^2 e^{\frac{-a}{\lambda}} - \lambda^2 e^{\frac{-0}{\lambda}}\right]\right)$$
$$= \frac{1}{\lambda}\left(\left[-\infty\lambda e^{\frac{-\infty}{\lambda}}\right] - \left[\lambda^2 e^{\frac{-\infty}{\lambda}} - \lambda^2 1\right]\right)$$
$$= \frac{1}{\lambda}\left(0 - \left[0 - \lambda^2\right]\right)$$
$$= \frac{1}{\lambda}\lambda^2$$
$$= \lambda$$

-x~$Exp(\lambda)$, then Var(x) = λ^2.

$$
\begin{aligned}
E(x^2) = \int_0^\infty x^2 \frac{1}{\lambda} e^{\frac{-x}{\lambda}} dx &= \frac{1}{\lambda}\int_0^\infty x^2 e^{\frac{-x}{\lambda}} dx \\
\text{by parts} \quad & \begin{bmatrix} u = x^2 & dv = e^{\frac{-x}{\lambda}} dx \\ du = 2x & v = -\lambda e^{\frac{-x}{\lambda}} \end{bmatrix} \\
&= \frac{1}{\lambda}\int_0^\infty u dv \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty - \int_0^\infty 2x\left(-\lambda e^{\frac{-x}{\lambda}}\right)dx\right) \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty + 2\lambda\int_0^\infty x e^{\frac{-x}{\lambda}} dx\right) \\
\text{by parts (similar to } E(x)) &= \begin{bmatrix} u = x & dv = e^{\frac{-x}{\lambda}} dx \\ du = 1 & v = -\lambda e^{\frac{-x}{\lambda}} \end{bmatrix} \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty + 2\lambda\left(\left[-x\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty - \int_0^\infty 1\left(-\lambda e^{\frac{-x}{\lambda}}\right)dx\right)\right) \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty + 2\lambda\left(\left[-x\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty - \left[\lambda^2 e^{\frac{-x}{\lambda}}\right]_0^\infty\right)\right) \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty + 2\lambda\left(\left[\lim_{a\to\infty} -a\lambda e^{\frac{-a}{\lambda}} + 0\lambda e^{\frac{-0}{\lambda}}\right] - \left[\lim_{a\to\infty} \lambda^2 e^{\frac{-a}{\lambda}} - \lambda^2 e^{\frac{-0}{\lambda}}\right]\right)\right) \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty + 2\lambda\left(\left[-\infty\lambda e^{\frac{-\infty}{\lambda}}\right] - \left[\lambda^2 e^{\frac{-\infty}{\lambda}} - \lambda^2 1\right]\right)\right) \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty + 2\lambda\left(0 - \left[0 - \lambda^2\right]\right)\right) \\
&= \frac{1}{\lambda}\left(\left[-x^2\lambda e^{\frac{-x}{\lambda}}\right]_0^\infty + 2\lambda^3\right) \\
&= \frac{1}{\lambda}\left(\left[\lim_{a\to\infty}\left(-a^2\lambda e^{\frac{-a}{\lambda}}\right) - \left(-0^2\lambda e^{\frac{-0}{\lambda}}\right)\right] + 2\lambda^3\right) \\
&= \frac{1}{\lambda}\left(-\infty^2\lambda e^{\frac{-\infty}{\lambda}} + 0 + 2\lambda^3\right) \\
&= \frac{1}{\lambda}\left(0 + 2\lambda^3\right) \\
&= \frac{1}{\lambda}\left(2\lambda^3\right) \\
&= 2\lambda^2
\end{aligned}
$$

$$
\begin{aligned}
Var(x) &= E(x^2) - (E(x))^2 \\
&= 2\lambda^2 - \lambda^2 \\
&= \lambda^2
\end{aligned}
$$
(cancer, dont even know if steps are right)

-TODO (minor) Tabular Method

From # 4 $E[x] = \lambda$ thus we need $E(x^2)$.

$E(x^2) = \int_0^\infty x^2 * f(x)dx = \lim_{b\to\infty}\int_0^b x^2 * \frac{1}{\lambda}e^{\frac{-x}{\lambda}}dx = \lim_{b\to\infty}\int_0^b \underbrace{x^2}_{u} * \underbrace{\frac{1}{\lambda}e^{\frac{-x}{\lambda}}dx}_{dv}$ → integration by parts twice or use the tabular method to organize and speed the integration.

Tabular method:

Alternating signs	u and its derivatives until 0	dv and its integrals
+	$u = x^2$	$dv = \frac{1}{\lambda}e^{\frac{-x}{\lambda}}$
-	$(x^2)' = 2x$	$\int \frac{1}{\lambda}e^{\frac{-x}{\lambda}}dx = -e^{\frac{-x}{\lambda}}$
+	$(2x)' = 2$	$\int \frac{1}{\lambda}e^{\frac{-x}{\lambda}}dx = \lambda e^{\frac{-x}{\lambda}}$
-	$\left(\frac{2}{1}\right)' = 0$ Stop	$\int \frac{1}{\lambda}e^{\frac{-x}{\lambda}}dx = -\lambda^2 e^{\frac{-x}{\lambda}}$

$$\lim_{b\to\infty}\int_0^b \underbrace{x^2}_{u} * \underbrace{\frac{1}{\lambda}e^{\frac{-x}{\lambda}}dx}_{dv} = \lim_{b\to\infty}\left(-x^2e^{\frac{-x}{\lambda}}\right) + \left(-2\lambda xe^{\frac{-x}{\lambda}}\right) + \left(-2\lambda^2e^{\frac{-x}{\lambda}}\right)\Big|_0^b$$

$$= \lim_{b\to\infty}\left\{\left(-b^2e^{\frac{-b}{\lambda}}\right) + \left(-2\lambda be^{\frac{-b}{\lambda}}\right) + \left(-2\lambda^2e^{\frac{-b}{\lambda}}\right)\right\} - \left\{\left(-0^2 * e^{\frac{0}{\lambda}}\right) + \left(-2\lambda * 0 * e^{\frac{0}{\lambda}}\right) + \left(-2\lambda^2e^{\frac{0}{\lambda}}\right)\right\}$$

But $\lim_{b\to\infty}\left(-b^2e^{\frac{-b}{\lambda}}\right) = -\infty * 0$ which is indeterminate → $\lim_{b\to\infty}\left(\frac{-b^2}{e^{\frac{b}{\lambda}}}\right) = \frac{-\infty}{\infty}$ → use l'Hôpital's Rule.

$= \lim_{b\to\infty}\left(\frac{-2b}{\frac{b}{\lambda}e^{\frac{b}{\lambda}}}\right) = \frac{-\infty}{\infty}$ indeterminate, use l'Hôpital's Rule again.

$= \lim_{b\to\infty}\left(\frac{-2}{\frac{b}{\lambda^2}e^{\frac{b}{\lambda}}}\right) = \frac{-2}{\infty} = 0$

$= \{(0) + (0) + (0)\} - \{(0) + (0) + (-2\lambda^2)\} = 2\lambda^2$ → $E(x^2) = 2\lambda^2$

$Var(x) = E(x^2) - \left(\underbrace{E[x]}_{\text{From \# 4}}\right)^2 = 2\lambda^2 - \lambda^2 = \lambda^2$

5.2: Normal Distribution Functions

-Normal Distributions

-A pdf

-Needs mean μ (where peak density is) and standard deviation σ.

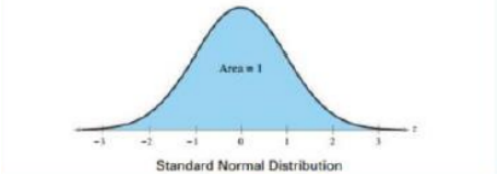
The normal distributions are a very important *class* of probability density distributions. All normal distributions are symmetric and have bell-shaped density curves with a single peak.

To speak specifically of any normal distribution, two quantities have to be specified: the mean μ, where the peak of the density occurs, and the standard deviation σ, which indicates the spread. A special case of the normal distribution is the standard normal distribution.

-Standard Normal Distribution

-μ = 0, σ = 1, & integral of the full interval = 1

The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1. The total area under its normal curve is 1.



Standard Normal Distribution

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \text{ where } -\infty < z < \infty$$

-Notation: $z \sim N(0, 1)$

-R Code: Graphing

```
curve(1/sqrt(2*pi)*exp(-x^2/2),xlim=c(-5,5)) #it should have been from -infinity to infinity
                                              #but the curve command require finite values thus 5 is large enough
```

-Ex:

Example #1: Find the indicated probability $P(z < 1.3)$ assuming that z is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$.

$P(z < 1.3) = P(-\infty < z < 1.3) = \int_{-\infty}^{1.3} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$ = 0.903199515414

-It's actually hard by hand (double integral and polar transformation). Just use R.

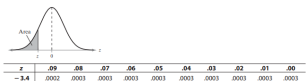
```
y <- function(x) {1/sqrt(2*pi)*exp(-x^2/2)} # define the integrated function
integrate(y, lower = -Inf, upper = 1.3) # integrate the function
```

-R Code: Ez way

```
pnorm(1.3) # interval from -∞ to 1.3
```

-Alternative: Using the Standard Normal Table

Table 4 — Standard Normal Distribution



z	.09	.08	.07	.06	.05	.04	.03	.02	.01	.00
−3.4	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005
−3.2	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007
−3.1	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010
−3.0	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013
−2.9	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019
−2.8	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026
−2.7	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035
−2.6	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047
−2.5	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062
−2.4	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082
−2.3	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107
−2.2	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139
−2.1	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179
−2.0	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228
−1.9	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287
−1.8	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359
−1.7	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446
−1.6	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548
−1.5	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668
−1.4	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808
−1.3	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968
−1.2	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151
−1.1	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357
−1.0	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587
−0.9	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841
−0.8	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119
−0.7	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420
−0.6	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743
−0.5	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085
−0.4	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446
−0.3	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821
−0.2	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207
−0.1	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602
−0.0	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Critical Values

Level of Confidence c	z_c
0.80	1.28
0.90	1.645
0.95	1.96
0.99	2.575

-So you'd find z = 1.30 (1.3 is y axis, 0.00 is x axis, boom)

-Using the Standard Normal Distribution to Find the Probability

Example #2: Find the indicated probability using the standard normal distribution. If it is convenient, use technology to find the probability.

-1. $P(z < -0.18)$ = P(-∞ < z < -0.18) = 0.428576284099

-R Code: pnorm(-0.18)

-2. $P(z > -0.18)$ = 1 - $P(z < -0.18)$ = 1 - P(-∞ < z < -0.18) = 0.5714

-R Code: 1 - pnorm(-0.18)

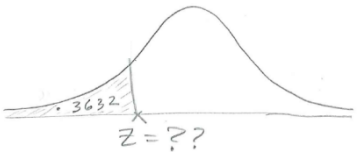
-3. $P(-1.54 < z < 1.54)$ = P(z < 1.54) - P(−1.54 < z)

pnorm(1.54) - pnorm(-1.54)

-Finding z-scores from Given Area

Example #3: Find the z −score that corresponds to a cumulative area of 0.3632.

-This means



(Left of the mean because area < 0.5)

-R Code: qnorm(.3632) # = -0.35 & q is for quantize

-Standardizing Normal Distribution

-What if the mean is not 0, and the standard deviation is not 1?

$$z(vlue) = \frac{the\ value - Mean}{standard\ deviation}$$

$$z = \frac{x - \mu_x}{\sigma_x}$$

-Proofs:

If the x is normally distributed with mean μ and standard deviation σ, let $z = \frac{x-\mu_x}{\sigma_x}$. Then, prove that:

-$\mu_z = 0$

$$u_z = E(z) = E\left(\frac{x-u_x}{\sigma_x}\right)$$
$$= \frac{1}{\sigma_x}E(x-u_x)$$
$$= \frac{1}{\sigma_x}(E(x)-u_x)\,[E(x)\text{ is a constant}]$$
$$= \frac{1}{\sigma_x}(E(x)-E(x))$$
$$= \frac{1}{\sigma_x}(0)$$
$$= 0$$

-$\sigma_z = 1$

$$\sigma_z = Var(z) = Var\left(\frac{x-\mu_x}{\sigma_x}\right)$$
$$= \left(\frac{1}{\sigma_x}\right)^2 Var(x-\mu_x)$$
$$= \frac{1}{\sigma_x^2}(Var(x)-Var(\mu_x))$$
$$= \frac{1}{\sigma_x^2}Var(x)$$
$$= \frac{1}{\sigma_x^2}\sigma_x^2$$
$$= 1$$

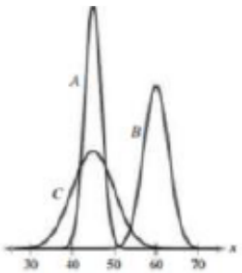
//TODO (minor) how is σ squared?

-Summary Standard Normal Distribution

Standard Normal Distribution (Special Case)	Normal Distribution
$f(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$ where $-\infty < z < \infty$. $\mu_z = 0$ and $\sigma_z = 1$ Notations: $z \sim N(0,1)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ where $-\infty < x < \infty$. $\mu_x = \mu$ and $\sigma_x = \sigma$ $x \sim N(\mu,\sigma)$

-Understanding mean and standard deviation:

i. Which normal curve has a greater mean?
ii. Which normal curve has the greater standard deviation?



-Answer: B has the greater mean (peak is to the right) and C has the greater standard deviation (wider)

-Ex:

Example #5: Find the indicated probability $P(x < 200)$ assuming that x is normally distributed with mean $\mu = 174$ and standard deviation $\sigma = 20$.

-We need to standardize it.

$$P(x < 200) = P\left(\frac{x-\mu}{\sigma} < \frac{200-174}{20}\right)$$
$$= P(z < 1.3)$$
$$\approx 0.9032$$

(x becomes z because notation crap)

-R Code: pnorm(200,mean = 174,sd=20)

-Ex:

Example #6: In a survey of U.S. women, the height in the 20 -to 29-years age group were normally distributed, with a mean of 64.2 inches and standard deviation of 2.9 inches. Find the probability that the a randomly selected study participant has a height that is:

-1. Less than 56.5 inches.

$$P(x < 56.5) = P\left(z < \frac{56.5-64.2}{2.9}\right)$$
$$= P(z < -2.655)$$
$$= 0.00396339411435$$

pnorm(56.5,mean = 64.2,sd=2.9)

-2. Between 61 and 67 inches.

$$P(61 < x < 67) = P(x < 67) - P(x < 61)$$
$$= P\left(z < \frac{67-64.2}{2.9}\right) - P\left(z < \frac{61-64.2}{2.9}\right)$$
$$= 0.697940826807$$

pnorm(67,mean = 64.2,sd=2.9)- pnorm(61,mean = 64.2,sd=2.9)

-3. More than 70.5 inches.

$$P(x < 70.5) = 1 - P(x < 70.5)$$
$$= 1 - P\left(z < \frac{70.5-64.2}{2.9}\right)$$
$$= 0.0149122327801$$

1-pnorm(70.5,mean = 64.2,sd=2.9)

-Transforming a z-score to an x-value

-Algebruh

To transform a standard z-score to an x-value in a given population, use the formula $x = \mu + z\sigma$

-Ex:

Example #7: The undergraduate grade point average (UGPA) of students taking the Law School Admission Test in a recent year can be approximated by normal distribution with mean $\mu = 3.36$ and standard deviation $\sigma = 0.18$.

-1. What is the minimum UGPA that would still place a student in the top 5% of UGPAs?

-Top 5% means P(z) = 0.95. use inverse and z = 1.645

-The corresponding x-value from that z-score is 3.36 + (1.645)0.18 = 3.6561

-R Code: x=3.36+qnorm(0.95)*0.18 aka **qnorm(0.95,mean=3.36,sd=0.18)**

-2. Between what two values does the middle 50% of the UGPAs lie?

-The middle 50%, that means the other 50% are halves from the left and right side

-z = 0.25 & 0.75 (because 25 + 25 = 50) (Extra: if we had 80% mid, it'll be z = 0.1 & 0.9)

-P(z)=0.25 & P(z)=0.75, inverse and get -0.67 & 0.67 (because symmetric)

-find x-values for each z-scores

3.36 + (-0.67)0.18 = 3.2394

3.36 + (0.67)0.18 = 3.4806

-R Code:

LEP=3.36+qnorm(0.25)*0.18

REP=3.36+qnorm(0.75)*0.18

-Prove that the normal distribution is a pdf.

-1. Yessir, it's positive probabilities

-2. Area under curve = 1 (Absolute Cancer)

Let $I = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$ integrate by substitution Let $z = \frac{x-\mu}{\sigma}$ ➔ $dz = \frac{dx}{\sigma}$ ➔ $(\sigma\, dz) = dx$

if $x = \mp\infty$ then $z = \mp\infty$

Thus, $I = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} (\sigma\, dz)$ ➔ , $I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} (dz)$ ➔ z is a dummy variable square both sides

$I * I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} (dx) * \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} (dy)$ ➔ $I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} (dxdy)$ now using polar transformations.

$x = r\cos(\theta)$, $y = r\sin(\theta)$ ➔ $x^2 + y^2 = r^2$ and $dxdy = rdrd\theta$

Also, $-\infty < x, y < \infty$ ➔ $0 < \theta < 2\pi$ and $0 < r < \infty$

$I^2 = \int_0^{2\pi}\int_0^{\infty} \frac{1}{2\pi} e^{-\frac{r^2}{2}} (rdrd\theta)$ by substitution. Let $u = \frac{r^2}{2}$ ➔ $du = rdr$ also, $0 < r < \infty$ ➔ $0 < u < \infty$

$I^2 = \frac{1}{2\pi}\int_0^{2\pi}\left(\int_0^{\infty} e^{-u} du\right) d\theta$

➔ $I^2 = \frac{1}{2\pi}\int_0^{2\pi}\left(\lim_{b\to\infty} \frac{e^{-u}}{-1}\Big|_0^b\right) d\theta = \frac{1}{2\pi}\int_0^{2\pi}\left(\lim_{b\to\infty} \frac{e^{-b}}{-1} - \frac{e^{-0}}{-1}\right) d\theta = \frac{1}{2\pi}\int_0^{2\pi}(1) d\theta = \frac{1}{2\pi}(2\pi - 0) = 1$

➔ $I^2 = 1$ ➔ $I = \mp 1$ but since $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} > 0$

➔ $I = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$

-Prove that E(x) = μ for Standard Normal Distribution

Prove that if $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$,then $E(x) = \mu$.

$E(x) = \int_{-\infty}^{\infty} x f(x) dx$

$E(x) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$ by substitution let $z = \frac{x-\mu}{\sigma}$ ➔ $dz = \frac{dx}{\sigma}$ ➔ $\sigma\, dz = dx$ also $z = \frac{x-\mu}{\sigma} \rightarrow x = \sigma z + \mu$

$$E(x) = \int_{-\infty}^{\infty} \underbrace{(\sigma z + \mu)}_{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \left[\underbrace{\sigma\, dz}_{dx}\right]$$

$$E(x) = \int_{-\infty}^{\infty} (\sigma z + \mu) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} [dz]$$

$$E(x) = \int_{-\infty}^{\infty} \mu \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz + \int_{-\infty}^{\infty} \sigma z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} [dz]$$

$$E(x) = \underbrace{\int_{-\infty}^{\infty} \mu \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz}_{I_1} + \underbrace{\int_{-\infty}^{\infty} \sigma z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} [dz]}_{I_2}$$

$I_1 = \int_{-\infty}^{\infty} \mu \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz}_{Standard\ Normal} = \mu * 1 = \mu$

$I_2 = \int_{-\infty}^{\infty} \sigma z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \sigma \int_{-\infty}^{\infty} \underbrace{z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}}_{an\ odd\ function} dz = \sigma E(z) = \sigma * 0 = 0$

$E(x) = \mu + 0 = \mu$

-Prove that Var(x) = σ^2 for Standard Normal Distribution

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

Let $\left(z = \frac{x-\mu}{\sigma},\ x = \sigma z + \mu\right), \left(dz = \frac{dx}{\sigma}\right), (\sigma dz = dx)$

$$= \int_{-\infty}^{\infty} (\sigma z + \mu)^2 \frac{1}{\cancel{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} \cancel{\sigma} dz$$

$$= \int_{-\infty}^{\infty} \left(\sigma^2 z^2 + 2\sigma z\mu + \mu^2\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz$$

$$= \int_{-\infty}^{\infty} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz + \int_{-\infty}^{\infty} 2\sigma z\mu \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz$$

$$= \sigma^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz + 2\sigma\mu \underbrace{\int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz}_{\text{Odd Function}} + \mu^2 \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz}_{\text{Standard Normal} = 1}$$

$$= \sigma^2 \underbrace{\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} dz}_{\text{TODO}=1} + 0 + \mu^2$$

$$= \sigma^2 + \mu^2$$

//TODO complicated Integral

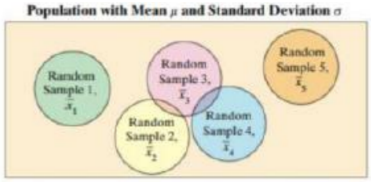
$$Var(x) = E(x^2) - (E(x))^2$$
$$= \sigma^2 + \mu^2 - \mu^2$$
$$= \sigma^2$$

-Why is the normal distribution important?

In this class or in real life we are interested in making inferences over two things.

1. The population proportion(s).
2. The population mean(s).

Under certain conditions, the normal distribution can be used to make inferences about both.



Properties of Sampling Distributions of Sample Means

1. The mean of the sample means $\mu_{\bar{x}}$ is equal to the population mean μ.
   $\mu_{\bar{x}} = \mu$
2. The standard deviation of the sample means $\sigma_{\bar{x}}$ is equal to the population standard deviation σ divided by the square root of the sample size n.
   $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

The standard deviation of the sampling distribution of the sample means is called the **standard error of the mean.**

-mean of sample VS mean of population

$$\mu_{\bar{x}} = \mu$$

-std dev of sample VS std dev of population

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

-Theorem from Chapter 4:

Let $x_1, x_2, \ldots\ldots, x_n$ be independent random variables such that $E(x_i) = \mu$ and $Var(x_i) = \sigma$. Then

1) $E(\bar{x}) = \mu$ and
2) $Var(\bar{x}) = \frac{\sigma^2}{n} \Leftrightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

-The Central Limit Theorem

-1. A sample size ≥ 30 approximates a normal distribution regardless of the original sample's distribution

-2. A sample from a population that is normally distributed is also normally distributed

If $n > 30$ then, $\bar{x}$ is normally distributed regardless of the distribution of the original data.

If x is normally distributed then, $\bar{x}$ is normally distributed regardless n.

The Central Limit Theorem

1. If random samples of size n, where $n \geq 30$, are drawn from any population with a mean μ and a standard deviation σ, then the sampling distribution of sample means approximates a normal distribution. The greater the sample size, the better the approximation. (See figures for "Any Population Distribution" below.)
2. If random samples of size n are drawn from a population that is normally distributed, then the sampling distribution of sample means is normally distributed for *any* sample size n. (See figures for "Normal Population Distribution" below.)

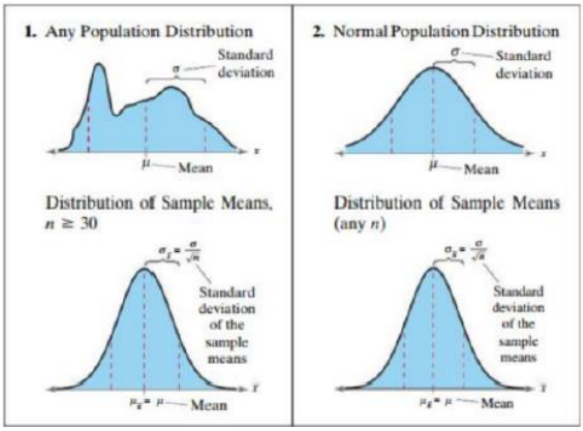
In either case, the sampling distribution of sample means has a mean equal to the population mean.

$\mu_{\bar{x}} = \mu$ Mean of the sample means

The sampling distribution of sample means has a variance equal to $1/n$ times the variance of the population and a standard deviation equal to the population standard deviation divided by the square root of n.

$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ Variance of the sample means

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Standard deviation of the sample means



-Ex:

Example #11: A machine is set to fill milk containers with a mean of 64 ounces and a standard deviation of 0.11 ounce. A random sample of 40 containers has a mean of 64.05 ounces. The machine needs to be reset when the mean of a random sample is unusual. Does the machine need to be reset? Explain your reasoning.

-Let x represents the milk containers.

-Given

$$\mu_x = 64 \text{ oz}$$
$$\sigma_x = 0.11 \text{ oz}$$
$$n = 40$$
$$\mu_{\bar{x}} = 64.05 \text{ oz}$$

-n > 30, the sample approximates a normal distribution, even if we don't know the distribution of the original.

$$\therefore \mu_{\bar{x}} = \mu_x = 64$$
$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \approx 0.0174$$

-P(x > 64.05)

$$P(x > 64.05) = P\left(z > \frac{64.05 - 64}{0.0174}\right)$$
$$= P(z > 2.87)$$
$$= 1 - P(z < 2.87)$$
$$= 1 - 0.9979$$
$$= 0.0021$$

1-pnorm(64.05,mean = 64,sd=0.0174)

-Uh oh, the machine suck dick and needs to be reset. Yes, it is very unlikely that you would have randomly sampled 40 containers with $\bar{x} = 64.05$ ounces because the probability of obtaining a sample mean of 64.05 or more is much smaller than 5%.

-Normal Approximation to Binomial Distribution

-Ex: from Chapter 4

It has been reported that 83% of federal government employees use email. If a sample of 200 federal government employees is selected,

-Find the probability that at least 120 employees use email.

Solution: $P(x \geq 120) = P(x = 120) + P(x = 121) + P(x = 122) + \cdots + P(x = 200)$

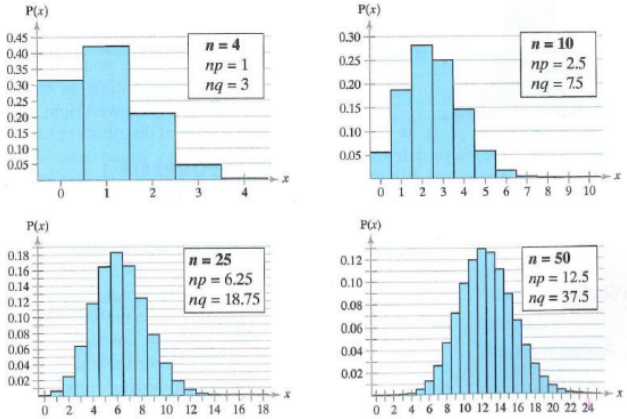
$= 200C120(.83)^{120}(.83.17)^{80} + 200C120(.83)^{121}(.83.17)^{79} + \cdots 200C200(.83)^{200}(.83.17)^{0}$

-Rember? We had to use the binomial theorem 80 times. Very ugly.

-Approximate a Binomial Distribution

To see why a normal approximation is valid, look at the binomial distributions for $p = 0.25$, $q = 1 - 0.25 = 0.75$, and $n = 4$, $n = 10$, $n = 25$, and $n = 50$ shown below. Notice that as n increases, the shape of the binomial distribution becomes more similar to a normal distribution.

(aka the higher n is, the more normal like it becomes)



Normal Approximation to a Binomial Distribution

If $np \geq 5$ and $nq \geq 5$, then the binomial random variable x is approximately normally distributed, with mean

$$\mu = np$$

and standard deviation

$$\sigma = \sqrt{npq}$$

where n is the number of independent trials, p is the probability of success in a single trial, and q is the probability of failure in a single trial.

(if **np & nq ≥ 5**, we good and you can use approximation)

-Ex: Check each binomial distribution to see whether it can be approximated by a normal distribution.

-Given n = 50, p = 0.2 (q = 0.8)

-np = 50 * 0.2 = 10, which is > 5

-nq = 50 * 0.8 = 40, which is > 5

-Ayyy, we good!

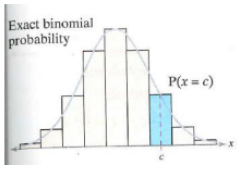
-Continuity Correction

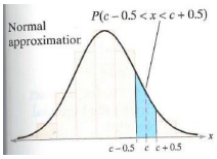
-Uh oh, a little stinky. Binomial Distribution is discrete, we need n interval for continuous

-Just take ±0.5 from the point

A binomial distribution is discrete and can be represented by a probability histogram. To calculate exact binomial probabilities, you can use the binomial formula for each value of x and add the results. Geometrically, this corresponds to adding the areas of bars in the probability histogram. Remember that each bar has a width of one unit and x is the midpoint of the interval.

When you use a continuous normal distribution to approximate a binomial probability, you need to move 0.5 units to the left and right of the midpoint to include all possible x –values in the interval. When you do this, you are making a **continuity correction**.





Binomial	Normal
When finding:	Use:
1) $P(x = a)$	$P(a - 0.5 < x < a + 0.5)$
2) $P(x \geq a)$	$P(x \geq a - 0.5)$
3) $P(x > a)$	$P(x > a + 0.5)$
4) $P(x \leq a)$	$P(x \leq a + 0.5)$
5) $P(x < a)$	$P(x < a - 0.5)$

-Ex: Use a continuity correction to convert the binomial probability to a normal distribution probability

-1. $P(x < 25)$

$= p(0) + p(1) \ldots + p(24) \approx P(x < 24.5)$ (here we use -0.05)

-2. $P(x \leq 25)$

$= p(0) + p(1) \ldots + p(25) \approx P(x < 25.5)$ (here we use +0.05)

-Approximating Binomial Probabilities

GUIDELINES

Using a Normal Distribution to Approximate Binomial Probabilities

In Words	In Symbols
1. Verify that a binomial distribution applies.	Specify n, p, and q.
2. Determine whether you can use a normal distribution to approximate x, the binomial variable.	Is $np \geq 5$? Is $nq \geq 5$?
3. Find the mean μ and standard deviation σ for the distribution.	$\mu = np$ $\sigma = \sqrt{npq}$
4. Apply the appropriate continuity correction. Shade the corresponding area under the normal curve.	Add 0.5 to (or subtract 0.5 from) the binomial probability.
5. Find the corresponding z-score(s).	$z = \frac{x - \mu}{\sigma}$
6. Find the probability.	Use the Standard Normal Table.

(aka combine np&nq > 5 w/ **μ=np** and **σ=sqrt(npq)**)

-Ex: A survey of U.S. adults found that 39% of those who have online account use the same or very similar passwords for many of their accounts. You randomly select 500 U.S. adults who have online accounts.

-Determine whether you can use a normal distribution to approximate the binomial distribution.

-Given: n = 500 , p = 0.39 (q=0.61)

$n * p = 500 * 0.39 = 195 > 5$ 😊

$n * q = 500 * 0.61 = 305 > 5$ 😊😊😊😊😊

-That means we can use normal distribution

$\mu = n * p = 500 * 0.39 = 195$

$\sigma = \sqrt{n * p * q} = \sqrt{500 * 0.39 * 61} \approx 10.91$

b) Find the probability that the number who use the same or very similar passwords for many of their accounts is exactly 175 by using the normal approximation, then use the binomial formula to check the approximation.

Binomial	Normal Approximation
$P(x = 175) = 500C175 * (.39)^{175} * (.61)^{325}$	$P(174.5 < x < 175.5)$ $= P\left(\frac{174.5 - 195}{10.91} < \frac{x - \mu}{\sigma} < \frac{175.5 - 195}{10.91}\right)$ $= P(-1.88 < z < -1.79)$ $= P(z < -1.79) - P(z < -1.88)$ from Z-table $= 0.0367 - 0.0301 = 0.0066$
R code: dbinom(x=175,size = 500,prob = 0.39) # 0.006798961	R code: pnorm(175.5,mean=195,sd=10.91)-pnorm(174.5,mean=195,sd=10.91)# 0.006818825

c) Find the probability that the number who use the same or very similar passwords for many of their accounts is no more than 225.

Binomial	Normal Approximation
$P(x \leq 225) = P(0) + P(1) + \cdots + P(225)$	$P(x \leq 225.5)$
R code: sum(dbinom(x=0:225,size = 500,prob = 0.39))# 0.9972582	R code: pnorm(225.5,mean=195,sd=10.91)# 0.9974098

d) Find the probability that the number who use the same or very similar passwords for many of their accounts is at least 200.

Binomial	Normal Approximation
$P(x \geq 200) = P(200) + P(201) + \cdots + P(500)$	$P(x \geq 200.5)$
R code: sum(dbinom(x=200:500,size = 500,prob = 0.39))# 0.3389487	R code: 1-pnorm(199.5,mean=195,sd=10.91)# 0.3399991

-Disadvantage

-irl, to investigate the population mean, we need std dev. That's hard, so we use S, but that means it's no longer a Normal Distribution matematically.

If x is normally distributed or if $n \geq 30$, then $\bar{x}$ is normally distributed with mean μ and a standard deviation $\frac{\sigma}{\sqrt{n}}$. In order to investigate the population, mean we need to know the population standard deviation which is unknown in real life. In order to make inference about μ, we must know σ or replace it by its estimate. But if we replace it by its estimate, S, then the distribution of the data changes to t/student distribution.

Take home message:

If x is normal or if $n \geq 30$, then $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(\mu = 0, \sigma = 1)$. But σ is unknown thus,

If x is normal or if $n \geq 30$, then $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t(df = n-1)$.

6.1: Confidence Intervals for the Mean (σ Known) and Population Proportions

-Estimating Population Parameters

-Point Estimate

A **point estimate** is a single value estimate for a population parameter. The most unbiased point estimate of the population mean μ is the sample mean $\bar{x}$. $\bar{X}$ for μ aka, use s^2 for σ^2 to estimate for population.

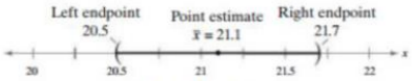
-Interval Estimate

An **interval estimate** is an interval, or range of values, used to estimate a population parameter.

-aka like ±0.5 thing, but you apply it to estimate the population through sample data.

21.1 ± 0.6 or $20.5 < \mu < 21.7$.

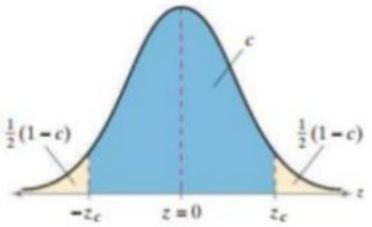
The point estimate and interval estimate are shown in the figure.



Interval Estimate

-Level of Confidence (for Interval Estimate)

The **level of confidence** c is the probability that the interval estimate contains the population parameter, assuming that the estimation process is repeated a large number of times.



If c = 90%:	
$c = 0.90$	Area in blue region
$1 - c = 0.10$	Area in yellow regions
$\frac{1}{2}(1-c) = 0.05$	Area in one tail
$-z_c = -1.645$	Critical value separating left tail
$z_c = 1.645$	Critical value separating right tail

-aka, if your interval estimate contains ≥ 90% of the area under the curve, you get special estimation for population parameters.

-Solving (generic) level of confidence for Standard Normal Distribution

$P(-1.645 < z < 1.645) = 0.90$

-Our target is μ for a specific confidence level c and assuming that σ is known. Then,

$$P(-z_c < z < z_c) = c$$

$$P\left(-z_c < \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} < z_c\right) = c$$

$$P\left(-z_c < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_c\right) = c \,[\text{Central Limit Thm}]$$

$$P\left(-z_c\left[\frac{\sigma}{\sqrt{n}}\right] < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\left[\frac{\sigma}{\sqrt{n}}\right] < z_c\left[\frac{\sigma}{\sqrt{n}}\right]\right) = c$$

$$P\left(-z_c\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_c\frac{\sigma}{\sqrt{n}}\right) = c$$

$$P\left(-z_c\frac{\sigma}{\sqrt{n}} + [-\bar{x}] < \bar{x} - \mu + [-\bar{x}] < z_c\frac{\sigma}{\sqrt{n}} + [-\bar{x}]\right) = c$$

$$P\left(-1\left(-z_c\frac{\sigma}{\sqrt{n}} - \bar{x} < -\mu < z_c\frac{\sigma}{\sqrt{n}} - \bar{x}\right)\right) = c$$

$$P\left(z_c\frac{\sigma}{\sqrt{n}} + \bar{x} > \mu > z_c\frac{\sigma}{\sqrt{n}} + \bar{x}\right) = c$$

$$P\left(-z_c\frac{\sigma}{\sqrt{n}} + \bar{x} < \mu < z_c\frac{\sigma}{\sqrt{n}} + \bar{x}\right) = c$$

$$P\left(\underbrace{\bar{x} - z_c\frac{\sigma}{\sqrt{n}}}_{-} < \mu < \underbrace{z_c\frac{\sigma}{\sqrt{n}} + \bar{x}}_{+}\right) = c$$

$$\therefore \mu = \bar{x} \pm z_c\frac{\sigma}{\sqrt{n}}$$

$$\mu = \bar{x} \pm E$$

$E = z_c * \frac{\sigma}{\sqrt{n}}$ is called the margin of error for μ (σ is known)

Left End Point L.E.P = $\bar{x} - E = \bar{x} - z_c\sigma_{\bar{x}} = \bar{x} - z_c * \left[\frac{\sigma}{\sqrt{n}}\right]$

Right End Point R.E.P = $\bar{x} + E = \bar{x} + z_c\sigma_{\bar{x}} = \bar{x} + z_c * \left[\frac{\sigma}{\sqrt{n}}\right]$

The z-value depends on the level of confidence c (90%, 95%, 99%, or other values). The level of confidence c is the area under the standard normal curve between the ***critical values*** $-z_c$ and z_c. Critical values are values that separate sample statistics that are probable from sample statistics that are improbable, or unusual. You can see from the figure shown below that c is the percent of the area under the normal curve between $-z_c$and z_c. The area remaining is $1 - c$, so the area in one tail is $\frac{1}{2}(1 - c)$.

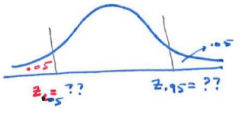
-aka, the tails are the improbable probabilites.

-Ex: Find the critical value z_c necessary to construct a confidence interval.

-1. 90%

-90% percent means the tails are a total of 1 - 0.9 = 0.1. Each tail is then 0.1 * 0.5 = 0.05.

-Therefore the z-scores of the left and right tails are 0.05 and 0.95 respectively.



(qnorm(0.05) = -1.645 & qnorm(0.95) = 1.645)

-2. 95%

-z = 0.025 & 0.975

-qnorm(0.025) = -1.96 & qnorm(0.975) = 1.96

-3. 0.97

-z = 0.015 & 0.985

-qnorm(0.015) = -2.17 & qnorm(0.985) = 2.17

-Table of Common Critical Values

Critical Values

Level of Confidence c	z_c
0.80	1.28
0.90	1.645
0.95	1.96
0.99	2.575

-Confidence Intervals for a Population Mean <u>When σ Is Known</u>

-Checks/Steps to see if you can use Confidence Interval

Constructing a Confidence Interval for a Population Mean (σ Known)

	In Words	In Symbols
1.	Verify that σ is known, the sample is random, and either the population is normally distributed or $n \geq 30$.	
2.	Find the sample statistics n and $\bar{x}$.	$\bar{x} = \frac{\Sigma x}{n}$
3.	Find the critical value z_c that corresponds to the given level of confidence.	Use Table 4 in Appendix B.
4.	Find the margin of error E.	$E = z_c\frac{\sigma}{\sqrt{n}}$
5.	Find the left and right endpoints and form the confidence interval.	Left endpoint: $\bar{x} - E$ Right endpoint: $\bar{x} + E$ Interval: $\bar{x} - E < \mu < \bar{x} + E$

(Again, in irl, σ is rarely known. That makes it a T distribution $E = t_{c*}\frac{s}{\sqrt{n}}$.)

-aka

$$\bar{x} \pm \text{qnorm}\left(1 - \frac{1-c}{2}\right)\frac{\sigma}{\sqrt{n}} = \bar{x} \pm z_c\frac{\sigma}{\sqrt{n}}$$ when (n ≥ 30) & σ is known/assumed

-Ex: Construct the indicated confidence interval for the population mean μ.

-c = 0.95, $\bar{x}$= 31.39, σ = 0.80, n = 82

-1. We know σ=0.8, and n=82 is ≥30.

-2. Sample mean is already given $\bar{x}$= 31.39.

-3. Finding critical value.

-We know c=0.95, meaning z_c=qnorm(.975)=1.96

-4. Margin of Error

-E = 1.96 * (0.8/sqrt(82)) = 0.17

-5. Left and Right Endpoints

-Left: $\bar{x}$-E = 31.39 + 0.17 = 31.22 (31.39-qnorm(.975)*0.80/sqrt(82))

-Right: $\bar{x}$-E = 31.39 - 0.17 = 31.56 (31.39+qnorm(.975)*0.80/sqrt(82))

-With 95% confidence, the population mean is between 31.22 and 31.56

-Ex: No way batchest iphone 21000

Example #3: From a random sample of 36 business days from February 24, 2016, through 24, 2017, the mean closing price of Apple stock was $116.16. Assume the population standard deviation is $10.27. Construct a 95% confidence interval for the population mean closing price of Apple stocks.

-c = 0.95, $\bar{x}$= 116.16, σ = 10.27, n = 36

-Plug and chug

$$\bar{x} \pm \left|\text{qnorm}\left(\frac{1-c}{2}\right)\right|\frac{\sigma}{\sqrt{n}}$$

$$\text{LEP} = 116.16 - \left|\text{qnorm}\left(\frac{1-0.95}{2}\right)\right|\frac{10.27}{\sqrt{36}} = 112.8052$$

$$\text{REP} = 116.16 + \left|\text{qnorm}\left(\frac{1-0.95}{2}\right)\right|\frac{10.27}{\sqrt{36}} = 119.5148$$

With 95% confidence, you can say that the population mean price is between $112.81 and $119.51.

-R Code: 116.16-qnorm(1-((1-0.95)/2))*(10.27/sqrt(36))
116.16+qnorm(1-((1-0.95)/2))*(10.27/sqrt(36))

-Ex:

Example #4: From a random sample of 24 months from January 2006 through December 2016, the mean number of tornadoes per month in the U.S. was about 100. Assume the population standard deviation is 114. Construct a 90% confidence interval for the population mean of number of tornados. Interpret the results.

-c = 0.90, $\bar{x}$= 100, σ = 114, n = 24

Solution:

$E = z_c * \frac{\sigma}{\sqrt{n}} \longrightarrow z_{.90} = 1.645$

$E = 1.645 * \frac{114}{\sqrt{24}} = 38.28$

$L.E.P = \bar{x} - E = 100 - 38.28 = 61.72$

$R.E.P = \bar{x} + E = 100 + 38.28 = 138.28$

90% C.I. for μ = (61.72, 138.28)

Interpretation:

With 90% confidence, you can say that the population mean tornadoes ~~per week~~ is between 61.72 and 138.28 tornadoes.

100-qnorm(.95)*114/sqrt(24) & 100+qnorm(.95)*114/sqrt(24)

-Ex: Have to find sample mean first, then do LEP and REP.

Example #5: A group of researchers estimates the mean length of time (in minutes) the average U.S. adult spends watching television using digital video recorders (DVRs) and other forms of time-shifted television each day. To do so, the group takes a random sample of 30 U.S. adults and obtains times (in minutes) below.

29	12	23	24	33	24	28	31	18	27	27	32	17	13	17
12	21	32	26	16	28	28	21	24	29	13	20	13	21	27

From past studies, the researcher can assume that σ is 6.5 minutes. Construct a 98% confidence interval for the population mean of the time that was spent watching DVRs and other forms of time-shifted television.

Solution:

We need the sample mean → Use R to obtain the stats summary.

```
y=c(29,12,23,24,33,24,28,31,18,27,27,32,17,13,17,12,21,32,26,16,28,28,21,24,29,13,20,13,2
1,27)
summary(y)
```

Given: $n = 30$, $\sigma = 6.5$, $\bar{x} = 22.87$, $c = .98$

Solution: $E = z_c * \frac{\sigma}{\sqrt{n}} \Longrightarrow z_{.98} = 2.33$

$= 2.33 * \frac{6.5}{\sqrt{30}} = 2.77$

$L.E.P = \bar{x} - E = 22.87 - 2.77 = 20.1$

$R.E.P = \bar{x} + E = 22.87 + 2.77 = 25.64$

The 98% C.I. for μ = (20.1, 25.64)

Interpretation:

With 98% confidence, you can say that the Population mean length of time is between 20.1 and 25.64 minutes.

R code:

```
L.E.P = 22.87-qnorm(.99)*6.5/sqrt(30)
L.E.P
R.E.P = 22.87+qnorm(.99)*6.5/sqrt(30)
R.E.P
```

-Finding a Minimum Sample Size of Confidence Intervals for a Population Mean When σ Is Known

-The objective is to find the minimum sample size (n) needed to estimate the population mean μ.

$E = z_c \frac{\sigma}{\sqrt{n}}$ Multiply both sides by $\sqrt{n}$.

$\Rightarrow \sqrt{n} * E = z_c \frac{\sigma}{\sqrt{n}} * \sqrt{n}$

$\Rightarrow \sqrt{n} * E = z_c \sigma$ Divide both sides by E.

$\Rightarrow \frac{\sqrt{n} * E}{E} = \frac{z_c \sigma}{E}$

$\Rightarrow \sqrt{n} = \frac{z_c \sigma}{E}$ Square both sides.

$\Rightarrow n = \left(\frac{z_c \sigma}{E}\right)^2$

Finding a Minimum Sample Size to Estimate μ

Given a c-confidence level and a margin of error E, the minimum sample size n needed to estimate the population mean μ is

$$n = \left(\frac{z_c \sigma}{E}\right)^2.$$

If n is not a whole number, then round n up to the next whole number (see Example 6). Also, when σ is unknown, you can estimate it using s, provided you have a preliminary random sample with at least 30 members.

-Ex: Just plug and chug 4head

Example #6: Determine the minimum sample size required when you want to be 99% confident that the sample mean is within two units of the population mean and $\sigma = 4.8$. Assume the population is normally distributed.

Given: $c = .99$, $E = 2$, $\sigma = 4.8$, Find n ??

Solution:

$n = \left(\frac{z_c * \sigma}{E}\right)^2 \longrightarrow c = .99 \Rightarrow z_{.99} = 2.575$

$n = \left(\frac{2.575 * 4.8}{2}\right)^2 = 38.1924 \rightarrow$ Round up

The Min Sample size = 39

Interpretation:

39 is the smallest sample size needed to construct a 99% C.I. with a margin error of 2.

n=(qnorm(0.995)*4.8/2)^2

-Ex:

Example #7: A cheese processing company wants to estimate the mean cholesterol content of all one-ounce servings of a type of cheese. The estimate must be within 0.75 milligram of the population mean. Determine the minimum sample size required to construct a 95% confidence interval for the population mean. Assume the population standard deviation is 3.10 milligrams.

Given: $E = .75$, $c = .95$, $\sigma = 3.10$, $n = ??$

Solution: $n = \left(\frac{z_c * \sigma}{E}\right)^2 \rightarrow z_{.95} = 1.96$

$= \left(\frac{1.96 * 3.10}{.75}\right)^2$

≈ 65.631 (rounded up)

$= 66$

Interpretation:
66 is the smallest sample size needed to construct a 95% C.I. with a margin error of 0.75 milligram.

n=(qnorm(0.975)*3.10/0.75)^2

-Confidence Intervals for Population Proportions (Relating it Binomial Distribution)

-The c confidence interval for μ can be modified for the population proportion as long as $np > 5$ and $nq > 5$.

$P(\bar{x} - z_c * \sigma_{\bar{x}} < \mu < \bar{x} + z_c * \sigma_{\bar{x}}) = c$

➔ Let $\hat{p} = \frac{x}{n}$ be the sample proportion where x is the number of successes in a sample of size n.

-The Expected Value is then

$$E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n} * np = p \quad \rightarrow \mu_{\hat{p}} = p$$ unbiased same for $\bar{x}$ since $\mu_{\bar{x}} = \mu$

-We need σ of p̂

$$\sigma_{\hat{p}} = \sqrt{Var(\hat{P})} = \sqrt{Var\left(\frac{x}{n}\right)} = \sqrt{\frac{1}{n^2}Var(x)} = \sqrt{\frac{1}{n^2} * npq} = \sqrt{\frac{pq}{n}}$$

-And these are End Points

$$P\left(\underbrace{\bar{x}}_{\hat{p}} - z_c * \underbrace{\sigma_{\bar{x}}}_{\sigma_{\hat{p}}} < \underbrace{\mu}_{p} < \underbrace{\bar{x}}_{\hat{p}} + z_c * \underbrace{\sigma_{\bar{x}}}_{\sigma_{\hat{p}}}\right) = c$$

$$P\left(\hat{p} - z_c * \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_c * \sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = c$$

Left End Point L.E.P = $\hat{p} - E = \hat{p} - z_c\sigma_{\hat{p}} = \hat{p} - z_c * \sqrt{\frac{\hat{p}\hat{q}}{n}}$ and

Right End Point R.E.P = $\hat{p} + E = \hat{p} + z_c\sigma_{\hat{p}} = \hat{p} + z_c * \sqrt{\frac{\hat{p}\hat{q}}{n}}$ $\quad \hat{p} \mp z_c * \sqrt{\frac{\hat{p}\hat{q}}{n}}$

-Confidence Interval for a Population Proportion

-Steps

Constructing a Confidence Interval for a Population Proportion

In Words	In Symbols
1. Identify the sample statistics n and x.	
2. Find the point estimate $\hat{p}$.	$\hat{p} = \frac{x}{n}$
3. Verify that the sampling distribution of $\hat{p}$ can be approximated by a normal distribution.	$n\hat{p} \geq 5, n\hat{q} \geq 5$
4. Find the critical value z_c that corresponds to the given level of confidence c.	Use Table 4 in Appendix B.
5. Find the margin of error E.	$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$
6. Find the left and right endpoints and form the confidence interval.	Left endpoint: $\hat{p} - E$ Right endpoint: $\hat{p} + E$ Interval: $\hat{p} - E < p < \hat{p} + E$

-aka

$$\frac{x}{n} \pm \text{qnorm}(1 - \frac{1-c}{2})\sqrt{\frac{\left(\frac{x}{n}\right)\left(1 - \frac{x}{n}\right)}{n}} = \hat{p} \pm z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$$ when $n\hat{p}\ \&\ n\hat{q}\ \geq 5$

-Ex:

Example #8: In a survey of 2241 U.S. adults in a recent year, 650 made a New Year's resolution. Construct a 90% confidence interval for the population proportion for the people that made a New Year's resolution. Interpret the results.

Solution:

$n = 2241, \quad x = 650$

$\hat{p} = \frac{650}{2241} \approx 0.29 \rightarrow \hat{q} = 0.71$

$n * \hat{p} = 650 > 5$ ☺ and $n * \hat{q} = 1591 > 5$ ☺ We can use the normal approximation.

The c confidence interval for P is $\hat{p} \mp z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$ → $c = 0.90$ → $z_c = 1.645$

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.645 * \sqrt{\frac{0.29 * 0.71}{2241}} \approx 0.016$$

$$L.E.P = \hat{p} - E = 0.29 - 0.016 = 0.274$$

$$R.E.P = \hat{p} + E = 0.29 + 0.016 = 0.306$$

R code:

```
L.E.P=.29-qnorm(.95)*sqrt(.29*.71/2241)
L.E.P
R.E.P=.29+qnorm(.95)*sqrt(.29*.71/2241)
R.E.P
```

Interpretation:

With 90% confidence, you can say that the population proportion of U.S. adults who say they made a New Year's resolution is between 27.4% and 30.6%.

-Ex:

Example #9: In a survey of 1000 U.S. adults, 700 think police officers should be required to wear body cameras while on duty. Construct a 95% confidence interval for the population proportion for the people that think police officers should be required to wear body cameras while on duty. Interpret the results.

Solution:

$n = 1{,}000, \quad x = 700$

$\hat{p} = \frac{700}{1{,}000} = 0.7 \rightarrow \hat{q} = 0.3$

$n * \hat{p} = 700 > 5$ ☺ and $n * \hat{q} = 300 > 5$ ☺ We can use the normal approximation.

The c confidence interval for P is $\hat{p} \mp z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$ → $c = 0.95$ → $z_c = 1.96$

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 * \sqrt{\frac{0.7 * 0.3}{1{,}000}} \approx 0.028$$

$$L.E.P = \hat{p} - E = 0.70 - 0.028 = 0.672$$

$$R.E.P = \hat{p} + E = 0.70 + 0.028 = 0.728$$

R code:

```
L.E.P=.7-qnorm(.975)*sqrt(.7*.3/1000)
L.E.P
R.E.P=.7+qnorm(.975)*sqrt(.7*.3/1000)
R.E.P
```

Interpretation:

With 90% confidence, you can say that the population proportion of U.S. adults think police officers should be required to wear body cameras while on duty is between 67.2% and 72.8%.

-Finding a Minimum Sample Size of Confidence Intervals for a Population Proportion

-For a specific given E, what is the smallest sample size that is needed to produce a confidence interval that is within that margin of error.

$E = z_c * \sqrt{\frac{\hat{p}\hat{q}}{n}}$ Solve for n → Divide both sides by z_c.

$\frac{E}{z_c} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ Square both sides.

$\left(\frac{E}{z_c}\right)^2 = \frac{\hat{p}\hat{q}}{n}$ Flip both sides of the equations.

$\left(\frac{z_c}{E}\right)^2 = \frac{n}{\hat{p}\hat{q}}$ Multiply both sides of the equations by $\hat{p}\hat{q}$.

$\hat{p}\hat{q} * \left(\frac{z_c}{E}\right)^2 = \hat{p}\hat{q} * \frac{n}{\hat{p}\hat{q}}$ Simplify and rearrange.

$n = \hat{p}\hat{q} * \left(\frac{z_c}{E}\right)^2$

Finding a Minimum Sample Size to Estimate p

Given a c-confidence level and a margin of error E, the minimum sample size n needed to estimate the population proportion p is

$$n = \hat{p}\hat{q}\left(\frac{z_c}{E}\right)^2.$$

If n is not a whole number, then round n up to the next whole number . Also, note that this formula assumes that you have preliminary estimates of $\hat{p}$ and $\hat{q}$. If not, use $\hat{p} = 0.5$ and $\hat{q} = 0.5$.

-Ex: (If p and q not given, assume 0.5), ("within %" means E) (Always round up n

Example #10: You wish to estimate, with 95% confidence, the population proportion of U.S. adults who think Congress is doing a good or excellent job. Your estimate must be accurate within 4% of the population proportion.

a) No preliminary estimate is available. Find the minimum sample size needed.

Solution:

$E = 0.04, c = 0.95 \rightarrow z_c = 1.96$ Since no preliminary estimate is available → Assume $\hat{p} = 0.5$ and $\hat{q} = 0.5$

$n = \hat{p}\hat{q} * \left(\frac{z_c}{E}\right)^2 \rightarrow n = 0.5 * 0.5 * \left(\frac{1.96}{0.04}\right)^2 = 600.25\ round\ up$

⇨ $n = 601$

b) Finding the minimum sample size needed using a prior survey that found that 25% of U.S. adults think that Congress is doing a good or excellent job.

Solution:

$E = 0.04, c = 0.95 \rightarrow z_c = 1.96$ From the prior survey we assume that $\hat{p} = 0.25 \rightarrow \hat{q} = 0.75$

$n = \hat{p}\hat{q} * \left(\frac{z_c}{E}\right)^2 \rightarrow n = 0.25 * 0.75 * \left(\frac{1.96}{0.04}\right)^2 = 450.1875\ round\ up$

⇨ $n = 451$

c) Compare the results from parts **a** and **b**.

Solution:

The sample size for b is smaller than for a. Having an estimate of the population proportion reduces the minimum the sample size.

6.2: Confidence Intervals for the Mean (σ Unkown)

-The t-distribution

The Standard Normal z-distribution	t-distribution
$z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
$E = z_c \frac{\sigma}{\sqrt{n}}$	$E = t_c \frac{s}{\sqrt{n}}$

In many real-life situations, σ is unknown. So how can you construct a confidence interval for a population mean when σ is not known? For a simple random sample that is drawn from a population that is normally distributed or has a sample size of 30 or more, you can use the sample standard deviation (s) to estimate the population standard deviation (σ). However, when using (s), the sampling distribution of ($z_{\bar{x}}$) does not follow a normal distribution. In this case, the sampling of ($z_{\bar{x}}$) follows a t-distribution.

-Degress of Freedom

To find t_c, use Table 5. To use Table 5, you need the degrees of freedom ($d.f.$) where $d.f. = n - 1$.

-Ex: Find the critical value t_c for the level of confidence c and sample size n.

-1. c = 0.95, n = 12

-df = 12 - 1 = 11

-t_c =

$$\text{inversecdf}\left(\text{tdist}(11), 1 - \frac{1 - 0.95}{2}\right)$$

$t_c = 2.20098516009$ R Code: qt(0.975,df=11)

-2. c = 0.98, n = 40

df = 39

-t_c = qt(1-((1-0.98)/2), 39)

-Confidence Intervals and t-distributions

-Margin of Error for t-distribution (when σ is not known)

If σ is not known then, the margin of error $E = t_c \frac{s}{\sqrt{n}}$.

-Steps

Constructing a Confidence Interval for a Population Mean (σ Unknown)

In Words	In Symbols
1. Verify that σ is not known, the sample is random, and either the population is normally distributed or $n \geq 30$.	
2. Find the sample statistics n, $\bar{x}$, and s.	$\bar{x} = \frac{\Sigma x}{n}, s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$
3. Identify the degrees of freedom, the level of confidence c, and the critical value t_c.	d.f. $= n - 1$ Use Table 5 in Appendix B.
4. Find the margin of error E.	$E = t_c \frac{s}{\sqrt{n}}$
5. Find the left and right endpoints and form the confidence interval.	Left endpoint: $\bar{x} - E$ Right endpoint: $\bar{x} + E$ Interval: $\bar{x} - E < \mu < \bar{x} + E$

-aka

$$= \bar{x} \pm \text{qt}(n - 1)\frac{S}{\sqrt{n}}$$

$$= \bar{x} \pm \text{qt}(df)\frac{S}{\sqrt{n}}$$

$$= \bar{x} \pm t_c \frac{S}{\sqrt{n}}$$

-Ex:

Example #2: Find the margin of error for the values of c, s, and n.

$$c = 0.90, s = 2.4, n = 35$$

Solution: $E = t_c * \frac{s}{\sqrt{n}} \Rightarrow$ d.f = 34 & c = .90 $\Rightarrow t_c$

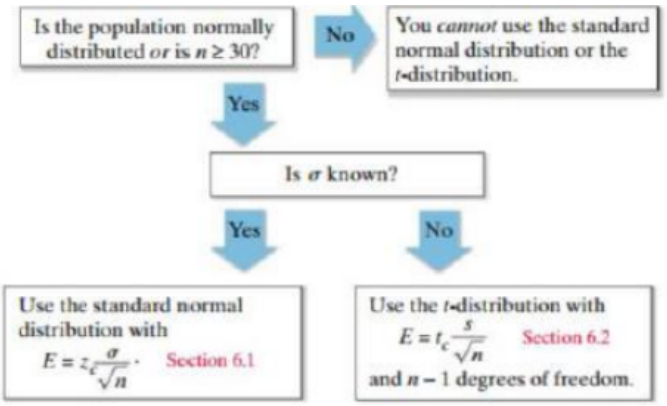
$E = 1.691 * \frac{2.4}{\sqrt{35}}$

d.f = 34 → 1.691

≈ 0.69

R code:
```
qt(0.95,df=34)*2.4/sqrt(35)
```

-When to use t-distribution vs z-distributions



Notice in the flowchart that when both $n < 30$ and the population is *not* normally distributed, you *cannot* use the standard normal distribution or the *t*-distribution.

(aka, use t-dist when n≥30 & when σ is not known)

-Ex: All coming together

Example #3: The growing seasons for a random sample of 35 U.S. cities were recorded, yielding a sample mean of 190.7 days and a sample standard deviation of 54.2 days. Estimate the true mean population of the growing season with 98% confidence. Interpret the results.

Given: $n = 35$, $\bar{x} = 190.7$, $s = 54.2$ → σ is unknown ⇒ use t not z

$c = .98$

Solution:

$E = t_c * \frac{s}{\sqrt{n}} \Rightarrow$ d.f = 34 and $c = .98 \Rightarrow t_c = 2.441$

$E = 2.441 * \frac{54.2}{\sqrt{35}} \approx 22.36$

$L.E.P = \bar{x} - E = 190.7 - 22.36 = 168.34$

$R.E.P = \bar{x} + E = 190.7 + 22.36 = 213.06$

The 98% C.I. for μ is (168.34, 213.06)

Interpretation:

With 90% confidence, you can say that the Population mean of the growing season is between 168.34 and 213.06 days.

R code:
```
L.E.P =190.7- qt(0.99,df=34)*54.2/sqrt(35)
L.E.P
R.E.P =190.7+ qt(0.99,df=34)*54.2/sqrt(35)
R.E.P
```

Example #4: You randomly selected 36 cars of a model that were sold at a car dealership and determine the number of days each car sat on the dealership's lot before it was sold. The sample mean is 9.75 days, with a sample standard deviation of 2.39 days. Construct a 99% confidence interval for the population mean number of days the car model sits on the dealership's lot. Interpret the results.

Given: $n = 36$, $\bar{x} = 9.75$, $s = 2.39$, $c = .99$ "σ is unknown"

Solution: $E = t_c * \frac{s}{\sqrt{n}}$ → d.f = 35, $c = .99 \rightarrow t_c = 2.724$

$= 2.724 * \frac{2.39}{\sqrt{36}} \approx 1.09$

$L.E.P = \bar{x} - E = 9.75 - 1.09 = 8.66$

$R.E.P = \bar{x} + E = 9.75 + 1.09 = 10.84$

99% C.I. for μ is (8.66, 10.84)

Interpretation:

With 99% Confidence, you can say that the Population mean for number of days the car model sits on the dealership's lot is between 8.66 and 10.84 days.

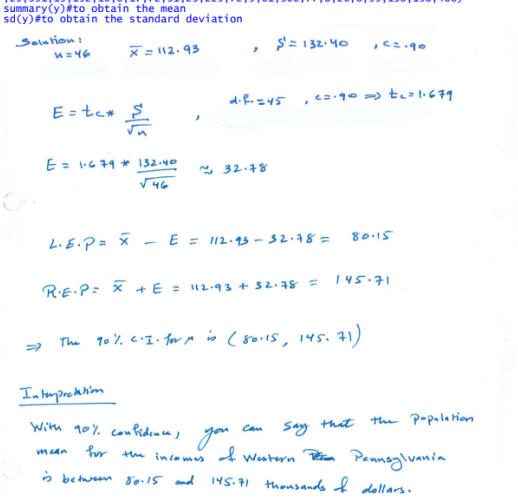
R code:
```
L.E.P =9.75- qt(0.995,df=35)*2.39/sqrt(36)
L.E.P
R.E.P =9.75+ qt(0.995,df=35)*2.39/sqrt(36)
R.E.P
```

Example #5: Find the 90% confidence interval of the population mean for the incomes of Western Pennsylvania credit unions. A random sample of 46 credit unions is shown. The data average is in thousands of dollars.

84	49	3	133	85	340	461	60	28	97
14	252	18	16	24	346	254	29	254	6
31	104	72	29	391	19	152	10	6	17
72	31	23	225	72	5	61	366	77	8
26	8	55	138	158	486				

Solution:

```
R code:
y=c(84, 49, 3, 133, 85, 340, 461, 60, 28, 97,14, 252,18,16,24, 346,254,29,254,6,31,104,72
,29,391,19,152,10,6,17,72,31,23,225,72,5,61,366,77,8,26,8,55,138,158,486)
summary(y)#to obtain the mean
sd(y)#to obtain the standard deviation
```



-R Test function

But this is not convenient. You can produce the results much faster and more accurately using the t.test command in R. This command works only if you have the raw data, which is more realistic.

```
R code:
t.test(y,conf.level =.90)
```

7.1: Hypothesis Testing for the Population Mean (σ is Known)

-Hypothesis Tests

Through the remainder of the quarter, you will study an important technique in inferential statistics called hypothesis testing. A **hypothesis test** is a process that uses sample statistic to test a claim about the value of a population parameter. Researchers in fields such as medicine, psychology, and business rely on hypothesis testing to make informed decisions about new medicine, treatments, and marketing strategies.

-Stating a Hypothesis

Definition: A statement about a population is called a **statistical hypothesis**. To test a population parameter, you should carefully state a pair of hypotheses, one that represents the claim and the other, its complement. When one of these hypotheses is false, the other must be true. Either hypothesis – the **null hypothesis** or the **alternative hypothesis** – may represent the original claim.

-Null (H_0) vs Alternative (H_a) Hypothesis

1. A **null hypothesis** H_0 is a statistical hypothesis that contains a statement of equality, such as ≤, =, or ≥.
2. The **alternative hypothesis** H_a is the complement of the null hypothesis. It is a statement that must be true if H_0 is false and it contains a statement of strict inequality, such as >, ≠, or <.

The symbol H_0 is read as "H sub-zero" or "H naught" and H_a is read as "H sub-a."

(Alternative doesn't have equality)

-Ex: The population mean is [in/equality] k

$$\begin{cases} H_0: \mu \le k \\ H_a: \mu > k \end{cases} \quad \begin{cases} H_0: \mu \ge k \\ H_a: \mu < k \end{cases} \quad \text{and} \quad \begin{cases} H_0: \mu = k \\ H_a: \mu \ne k \end{cases}$$

(with the 1st

-Ex: The statement represents a claim. Write its complement and state which is H_0 and which is H_a.

-1. μ < 128

-"<" is for Alternative

-This means (H_a: μ < 128) & (H_0: μ ≥ 128)

-2. p = 0.21

-"=" is for Null

-This means (H_0: p = 0.21) & (H_a: p ≠ 0.21)

-Ex: Write the claim as a mathematical statement.

-1. A tablet manufacturer claims that the mean life of the battery for a certain model of tablet is more than 8 hours.

-Claim: μ > 8

-(H_a: μ > 8) & (H_0: μ ≤ 8)

-2. An amusement park claims that the mean daily attendance at the park is at least 20,000 people.

-Claim: μ ≥ 20,000

-(H_0: μ ≥ 20,000) & (H_a: μ < 20,000)

-Types of Errors and Level of Significance

No matter which hypothesis represents the claim, you always begin a hypothesis test by **assuming** the equality condition in the **null** hypothesis is true. So, when you perform a hypothesis test, you make one of two decisions:

-Type 1: Reject the Null Hypothesis

-Type 2: Fail to reject the Null Hypothesis

Because your decision is based on a sample rather than the entire population, there is always the possibility you will make the wrong the decision.

-Ex: court trial

	Truth about defendant	
Verdict	Innocent	Guilty
Not guilty	Justice	Type II error
Guilty	Type I error	Justice

-Type of Errors Formal Definition

	Truth of H_0	
Decision	H_0 is true.	H_0 is false.
Do not reject H_0.	Correct decision	Type II error
Reject H_0.	Type I error	Correct decision

A **type I error** occurs if the null hypothesis is rejected when it is true.

A **type II error** occurs if the null hypothesis is not rejected when it is false.

-Ex: Describe type I and type II errors for the hypothesis test of the indicated claim.

-A used textbook selling website claims that at least 60% of its new customers will return to buy their next textbook.

-Claim: p ≥ 0.60

$H_0: p \geq 0.60$

$H_a: p < 0.60$

Type I error will occur when the actual proportion of new customers who return to by their next text book is at least 0.60, but you reject $H_0: p \geq 0.60$.

Type II error will occur when the actual proportion of new customers who return to by their next text book is less than 0.60, but you fail reject $H_0: p \geq 0.60$.

-Level of Significant

In a hypothesis test, the **level of significance** is your maximum allowable probability of making a type I error. It is denoted by α, the lowercase Greek letter alpha.

The probability of a type II error is denoted by β, the lowercase Greek letter beta.

-Why is select a small α?

By setting the level of significance at a small value, you are saying that you want the probability of rejecting a true null hypothesis to be small. Three commonly used levels of significance are:

α = 0.1, 0.05, 0.01

Alpha sets the standard for how extreme the data must be before we can reject the null hypothesis. The P-value indicates how extreme the data are. We compare the P -value with the alpha to determine whether the observed data are statistically significantly different from the null hypothesis:

-Level of Confidencce and Level of Significant

-(c = 1 - α) & (α = 1 - c)

-Enough or Not Enough Evidence Matrix FAT

	Claim	
Decision	Claim is H_0.	Claim is H_a.
Reject H_0.	There is enough evidence to reject the claim.	There is enough evidence to support the claim.
Fail to reject H_0.	There is not enough evidence to reject the claim.	There is not enough evidence to support the claim.

-Statistical Tests and P-value

If the null hypothesis is true, then a ***P*-value** (or **probability value**) of a hypothesis test is the probability of obtaining a sample statistic with a value as extreme or more extreme than the one determined from the sample data.

-When to rejec the Null Hypothesis?

-When the P-value < α (the smaller the P-value, the stronger the evidence against the Null Hyposthesis is)

One way to decide whether to reject the null hypothesis is to determine whether the probability of obtaining the standardized test statistic (or one that more extreme) is less than the level of significance.

-Ex: When to reject Null Hypotheisis with give P-value and α

#	Label	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	$P = 0.0691$	Fail to reject H_0	Fail to reject H_0	Reject H_0
2	$P = 0.0107$	Fail to reject H_0	Reject H_0	Reject H_0
3	$P = 0.0062$	Reject H_0	Reject H_0	Reject H_0

-Finding the P-value for a Hypothesis Test

-Depending on claim if it's > & < vs =, you can have 1 or 2 tails.

After determining the hypothesis test's standardized test statistic and the standardized test statistic's corresponding area, do one of the following to find the P-value.

a. For a left-tailed test, P = (Area in left tail).

b. For a right-tailed test, P = (Area in right tail).

c. For a two-tailed test, P = 2(Area in tail of standardized test statistic).

-"<" means Left Tail (qnorm(z-score), -∞ to z)

-">" means Right Tail (1-qnorm(z-score), z to ∞)

-"≠" means 2-Tail

-If z-score < 0, 2*qnorm(z-score)

-If z-score > 0, 2*(1-qnorm(z-score))

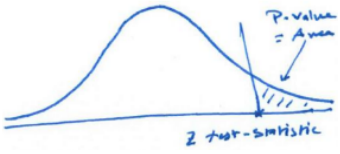
-Ex:

-1. A manufacturer of grandfather clocks claims that the mean time its clocks lose is no more than 0.02 second per day.

-Claim μ ≤ 0.02

-H_0: μ ≤ 0.02, H_a: μ > 0.02

-Since we want to see if we should reject the null hypothesis, we want the P-value of the alternative hypothesis.



(μ > 0.02 means Right Tail Test) (if it was < 0.02, then Left)

-2. A report claims that lung cancer accounts for 25% of all cancer diagnoses.

-Claim: μ = 0.25

-H_0: μ = 0.25, H_a: μ ≠ 0.25

-Since we want to see if we should reject the null hypothesis, we want the P-value of the alternative hypothesis.

(μ ≠ 0.25 means 2-Tail Test)

-Hypothesis Testing for the Population Mean (σ is known)

The test statistic for μ when **(σ is known) Sec. 7.1**	The test statistic for μ when **(σ is known) Sec. 7.2**
$z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
The test statistic for *P* when n*p>5 and n*q>5 Sec. 7.1	
$z_{\hat{p}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$	

-Ex: Test the claim about the population mean μ at the level of significance α. Assume the population is normally distributed.

-1. Claim: μ = 40; α = 0.05; σ = 1.97 & Sample Statistics: $\bar{x}$= 39.2, n = 25 (σ is known, we can use z)

-a. Hypothesis

H_0: μ = 40, H_a ≠ 40. (≠ means 2-tail test)

-b. Standardize / find z-score

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = -2.03$$

-c. P-value

$$\text{P-value} = 2P(z < -2.03) = 2(0.0212) = 0.0424$$ which is less than α = 0.05

-d. Reject the Null Hypothesis?

-Yes, reject H_0 because P-value < α.

-e. Interpretation

-There is enough evidence (P-value = 0.0424) at 5% level of significance to reject the claim.

-2. Claim: $\mu \leq$ 22,500; α = 0.01; σ = 1200 & $\bar{x}$= 23,500, n = 45

-a. Hypothesis

H_0: μ ≤ 22500, H_a > 22500. (> 22500 means right tail test)

-b. Standardize / find z-score

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = 5.59$$

-c. P-value

$$\text{P-value} = P(z < 5.59) = 1.135348061E10$$ which is less than α = 0.01

-d. Reject the Null Hypothesis?

-Yes, P-value < α, by a lot

-e. Interpretation

-There is enough evidence (P-value very small) at 1% level of significance to reject the claim.

-3.

Example #8: A random sample of 100 medical school applicants at a university has a mean score of 502 on the MCAT. According to a report, the mean total score for the school's applicants is more than 499. Assume the population standard deviation is 10.6. At $\alpha = 0.01$, is there enough evidence to support the report's claim?

Claim: $\mu > 499$

Sample Statistics: $n = 100$, $\bar{x} = 502$ assume $\sigma = 10.6$ and $\alpha = .01$

-a. Hypothesis

H_0: $\mu \leq 499$

H_A: $\mu > 499$ → claim one-tailed test ⇒ Right-tailed test

-b. Standardize / find z-score

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{502 - 499}{\frac{10.6}{\sqrt{100}}}$$

$z \approx 2.83$

-c. P-value

$$\text{P-value} = P(z > 2.83) = 1 - .9977 = .0023 < \alpha = .01$$

-d. Reject the Null Hypothesis?

Yes, because P-value < α

-e. Interpretation

-There is evidence to support the claim @ 1% level of confidence

There is enough evidence at the 1% level of significance to support the report's claim that the mean total score for the school's applicants is more than 499. "P = .0023"

-Hypothesis Testing for Proportions Using the Rejection Region(s) Method

z-Test for a Proportion p

The ***z*-test for a proportion *p*** is a statistical test for a population proportion. The *z*-test can be used when a binomial distribution is given such that $np \geq 5$ and $nq \geq 5$. The **test statistic** is the sample proportion $\hat{p}$ and the **standardized test statistic** is

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{pq/n}}.$$ Standardized test statistic for p

(z-test for proportion, review)

-Ex: Determine whether a normal sampling distribution can be used. If it can be used, test the claim.

-Claim: $p < 0.12$, $\alpha = 0.01$. Sample statistics: $\hat{p} = 0.10$, $n = 4$

-We actually cannot use the normal distribution to infer the binomial distribution accurately.

$H_0: p \geq 0.12$

$H_a: p < 0.12$ → Claim

$n * p = 40 * 0.12 = 4.8 < 5$ ☹

Cannot use the normal distribution.

-Ex: Determine whether a normal sampling distribution can be used. If it can be used, test the claim.

Example #10: A researcher center claims that more than 29% of U.S. employees have changed jobs in the past three years. In a random sample of 180 U.S. employees, 63 have changed jobs in the past three years. At $\alpha = 0.10$, is there enough evidence to support the center's claim?

-Claim: P > 0.29, Sample Stats: n = 180, x = 63, α = 0.10

-a. Can we use a normal sampling distribution?

$n * p = 180 * 0.29 = 52.2 > 5$ ☺ p=.29 → q=1-p=1-.29=.71

$n * q = 180 * 0.71 = 127.8 > 5$ ☺ (Yes ☺)

-b. State H_0 and H_a and identify the claim.

$H_0: P \leq .29$

$H_A: P > .29$ "claim" → Right-tailed test

-c. Find the standardized test statistic.

-Find p̂

$P = .29 \rightarrow q = .71$

$\hat{P} = \frac{x}{n} = \frac{63}{180} = .35$

-z-score

$$z = \frac{\hat{P} - P}{\sqrt{\frac{P * q}{n}}}$$

$$= \frac{.35 - .29}{\sqrt{\frac{.29 * .71}{180}}} \approx 1.77$$

-d. Compute the P-value.

$P - value = P(z > 1.77)$

$= 1 - 0.9616 = 0.0384 < \alpha = 0.10$ → reject H_0 & support claim (R Code: 1-pnorm(1.77))

-e. Interpret the decision in the context of the original claim.

There is enough evidence at the 10% level of significance to support the researcher's claim that more than 29% U.S. employees have changed jobs in the past three years. "P = .0384"

-R Code: prop.test(63,180,p=.29,alternative = "greater",correct = FALSE)

-Ex: Test the claim

Example #11: A humane society claims that less than 67% of U.S. household own a pet. In a random sample of 600 U.S. households, 390 say they own a pet. At $\alpha = 0.10$, is there enough evidence to support the society's claim?

Solution:

claim : $p < .67$

statistics: $n = 600$, $x = 390$, $\alpha = .10$

a) Determine whether a normal sampling distribution can be used.

$n * p = 300 * 0.67 = 201 > 5$ 😊

$n * q = 300 * 0.33 = 99 > 5$ 😊

Normal approximation can be used.

b) State H_0 and H_a and identify the claim.

$H_0: p \geq .67$

$H_A: p < .67$ "claim" left-tailed test

a) Find the standardized test statistic.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p * q}{n}}}$$

$p = .67 \Rightarrow q = .33$

$\hat{p} = \frac{390}{600} = 0.65$

$$= \frac{.65 - .67}{\sqrt{\frac{.67 * .33}{600}}} \approx -1.04$$

a) Compute the P-value.

Solution:

$P - value = P(z < -1.04)$

$= 0.1492 > \alpha = 0.10$ → fail to reject H_0

R code.

pnorm(-1.04) → 0.14917

b) Decide whether to reject or fail to reject H_0.

Fail to reject H_0 ⟶ Do not support the claim.

c) Interpret the decision in the context of the original claim.

There is not enough evidence at the 10% level of significance to support the human society's claim that less than 67% of U.S. households own a pet. "P= 0.1492"

-R Code: prop.test(390,600,p=.67,alternative = "less",correct = FALSE)

7.2: Hypothesis Testing for the Mean (σ Unknown)

-Literally the same thing except

-Use sample standard deviation because population standard deviation is unknown

-Use t-distribution instead of normal because population standard deviation is unknown

8.1: Testing the Difference between Two Means (independent Samples with $\sigma 1$ and $\sigma 2$ Known)

-Independent and Dependent Samples

Two samples are **independent** when the sample selected from one population is not related to the sample selected from the second population. Two samples are **dependent** when each member of one sample corresponds to a member of the other sample. Dependent samples are also called **paired samples** or **matched samples.**

-Ex:

Ex1: Classify the two samples as independent and dependent.

1) Sample 1: The IQ scores of 60 random females.
   Sample 2: The IQ scores of 60 random males.

Independent.

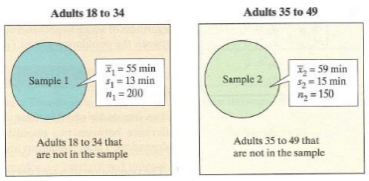
2) Sample 1: The commute times of 10 workers when they use their own vehicles.
   Sample 2: The commute times of same 10 workers when they use public transportation.

Dependent.

-An Overview of Two-Sample Hypothesis Testing

In this section and next, you will be learning how to test a claim comparing the means of two different populations using independent / dependent samples.

For instance, an advertiser is developing a marketing plan and wants to determine whether there is a difference in the amounts of time adults ages 18 to 34 and adults ages 35 to 49 spend on social media each day.



-2 Independent Samples

For a two-sample hypothesis test with independent samples,

1. the **null hypothesis H_0** is a statistical hypothesis that usually states there is no difference between the parameters of two populations. The null hypothesis always contains the symbol $\leq$, $=$, or $\geq$.
2. the **alternative hypothesis H_a** is a statistical hypothesis that is true when H_0 is false. The alternative hypothesis contains the symbol $>$, $\neq$, or $<$.

$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 \neq \mu_2 \end{cases}$ $\begin{cases} H_0: \mu_1 \leq \mu_2 \\ H_a: \mu_1 > \mu_2 \end{cases}$ $\begin{cases} H_0: \mu_1 \geq \mu_2 \\ H_a: \mu_1 < \mu_2 \end{cases}$

$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 \neq 0 \end{cases}$ $\begin{cases} H_0: \mu_1 - \mu_2 \leq 0 \\ H_a: \mu_1 - \mu_2 > 0 \end{cases}$ $\begin{cases} H_0: \mu_1 - \mu_2 \geq 0 \\ H_a: \mu_1 - \mu_2 < 0 \end{cases}$

-2-sample z-Test for the Difference between Means with $\sigma 1$ and $\sigma 2$ Known

We are interested in the difference of the two means $\mu_1 - \mu_2$. First, we start with an unbiased point estimate of $\mu_1 - \mu_2$ then, we need to obtain the variance of the estimate. And lastly, we need to standardizes it.

If x_1 and x_2 are two independent sample sets with sizes n_1 and n_2 and let $x_i \sim N(\mu_i, \sigma_i)$ for $i = 1, 2$ then,

-Proof: $\bar{x_1} - \bar{x_2}$ is an unbiased point estimator for $\mu_1 - \mu_2$.

$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2)$ but $E(\bar{x}_i) = \mu_i$

$= \mu_1 - \mu_2$

→ $\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_1 - \mu_2$

-Proof: Standard deviation of the unbiased point estimator

$$\text{Claim: } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Var(\bar{x}_1 - \bar{x}_2) = Var\left(\underbrace{1}_{a} \cdot \bar{x}_1 + \underbrace{[-1]}_{b} \cdot \bar{x}_2\right) = 1^2 Var(\bar{x}_1) + [-1]^2 Var(\bar{x}_2) + 2 * 1 * [-1]\, Cov(\bar{x}_1, \bar{x}_2)$$

From lemma 2 in chapter 4 $Var(ax + by) = a^2 Var(x) + b^2 Var(y) + 2ab\, Cov(x, y)$

$$Var(1 \cdot \bar{x}_1) + Var([-1] \cdot \bar{x}_2) + Cov(1 \cdot \bar{x}_1, [-1] \cdot \bar{x}_2) = 1^2 Var(\bar{x}_1) + [-1]^2 Var(\bar{x}_2) + 2 * 1 * [-1] \underbrace{Cov(\bar{x}_1, \bar{x}_2)}_{=0}$$

$$= Var(\bar{x}_1) + Var(\bar{x}_2) + 2 * 1 * [-1] * 0$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

-Remark: If samples were normal distribution, the difference between both means is too.

if $x_i \sim N(\mu_i, \sigma_i)$ for $i = 1, 2$ then $(\bar{x}_1 - \bar{x}_2) \sim N\left(\underbrace{\mu_1 - \mu_2}_{\mu_{(\bar{x}_1 - \bar{x}_2)}}, \underbrace{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}_{\sigma_{\bar{x}_1 - \bar{x}_2}}\right)$ Thus, based on the CLT. If x_i is normal or if $n_i > 30$, then $(\bar{x}_1 - \bar{x}_2)$ is normally distributed.

-Remark: z-score

$$Z_{(\bar{x}_1 - \bar{x}_2)} = \frac{(\bar{x}_1 - \bar{x}_2) - \left(\mu_{(\bar{x}_1 - \bar{x}_2)}\right)}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

-Steps, tricks, and chips

A **two-sample z-test** can be used to test the difference between two population means μ_1 and μ_2 when these conditions are met.

1. Both σ_1 and σ_2 are known.
2. The samples are random.
3. The samples are independent.
4. The populations are normally distributed *or* both $n_1 \geq 30$ and $n_2 \geq 30$.

The **test statistic** is $\bar{x}_1 - \bar{x}_2$. The **standardized test statistic** is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad \text{where} \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

-Ex: test the claim about the difference between two population means $\mu 1$ and $\mu 2$ at the level of significance α. Assume the samples are random and independent, and the populations are normally distributed.

Claim: $\mu_1 = \mu_2$; $\alpha = 0.10$

Population Parameters: $\sigma_1 = 3.4$ and $\sigma_2 = 1.5$.
Sample statistics: $\bar{x}_1 = 16, n_1 = 29$,and $\bar{x}_2 = 14, n_2 = 28$.

a) State $H_0$and H_a and identify the claim.

H_0: $\mu_1 = \mu_2$ → claim → H_0: $\mu_1 - \mu_2 = 0$

H_A: $\mu_1 \neq \mu_2$ "TTT" → H_A: $\mu_1 - \mu_2 \neq 0$ "T.T.T"

b) Find the standardized test statistics.

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(16 - 14) - 0}{\sqrt{\frac{(3.4)^2}{29} + \frac{(1.5)^2}{28}}} \approx 2.89$$

c) Compute the $P - value$.

$P - value = 2 * p(z > 2.89)$

$= 2 * (1 - 0.9981)$ → from z-tab

$= 0.0038 < \alpha = 0.10$

R code.

```
2*(1-pnorm(2.89))      -->   0.003852418
```

d) Decide whether to reject or fail to reject H_0.

→ Reject H_0 ⇒ Reject the claim

e) Interpret the decision in the context of the original claim.

There is enough evidence at the 10% level of significance to reject the claim. P-value = .0039

Claim: $\mu_1 > \mu_2$; $\alpha = 0.10$

Population Parameters: $\sigma_1 = 40$ and $\sigma_2 = 15$.
Sample statistics: $\bar{x}_1 = 500, n_1 = 100$,and $\bar{x}_2 = 495, n_2 = 75$.

a) State $H_0$and H_a and identify the claim.

$H_0: \mu_1 \leq \mu_2$ ⟶ $H_0: \mu_1 - \mu_2 \leq 0$

$H_A: \mu_1 > \mu_2$ → claim ⟶ $H_A: \mu_1 - \mu_2 > 0$ → claim" R.T.T

b) Find the standardized test statistics.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(500 - 495) - 0}{\sqrt{\frac{(40)^2}{100} + \frac{(15)^2}{75}}} \approx 1.15$$

c) Compute the $P-value$.

$P-value = p(z > 1.15)$

$= (1 - 0.8749)$ → from z-tab

$= 0.1251 > \alpha = 0.10$

R code.

```
(1-pnorm(1.15))      -->   0.1250719
```

d) Decide whether to reject or fail to reject H_0.

Fail to reject H_0 ⟶ Fail to support the claim (H_A)

e) Interpret the decision in the context of the original claim.

There is not enough evidence at the 10% level of significance to support the claim. P-value = 0.1251

Ex4: The mean ACT English score for 120 high school students is 20.1. Assume the population standard deviation is 6.8. The mean ACT reading score for 150 high school students is 21.3. Assume the population standard deviation is 6.5. At $\alpha = 0.10$, can you support the claim that the ACT reading score are higher than ACT English score? Assume the samples are random and dependent, and the populations are normally distributed.

Solution: claim: $\mu_1 < \mu_2$ (English, Reading) stat: $n_1 = 120$, $\bar{x}_1 = 20.1$, $\sigma_1 = 6.8$; $n_2 = 150$, $\bar{x}_2 = 21.3$, $\sigma_2 = 6.5$

a) State $H_0$and H_a and identify the claim.

$H_0: \mu_1 \geq \mu_2$

$H_A: \mu_1 < \mu_2$ claim

$H_0: \mu_1 - \mu_2 \geq 0$

$H_A: \mu_1 - \mu_2 < 0$ L.T.T "claim"

b) Find the standardized test statistics.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(20.1 - 21.3) - 0}{\sqrt{\frac{(6.8)^2}{120} + \frac{(6.5)^2}{150}}}$$

≈ -1.47

c) Compute the $P-value$.

$P-value = p(z < -1.47)$

$= .0708$ → from z-tab

$= 0.0708 < \alpha = 0.10$

R code.

```
pnorm(-1.47)      -->   0.07078088
```

d) Decide whether to reject or fail to reject H_0.

Reject H_0 ⟶ support H_A which is the claim.

e) Interpret the decision in the context of the original claim.

There is enough evidence at the 10% level of significance to support the claim that ACT reading scores are higher than ACT English scores. P-value = 0.07078

Ex5: A real estate agency says that the mean home sales price in Casper, Wyoming, is the same as in Cheyenne, Wyoming. The men home sales price for 25 homes in Casper is $294,220. Assume the population standard deviation is $135,387. The mean home sales price for 25 homes in Cheyenne is $287,984. Assume the population standard deviation is $151,996. At $\alpha = 0.01$, is there enough evidence to reject the agency's claim? Assume the samples are random and dependent, and the populations are normally distributed.

claim: $\mu_1 = \mu_2$ stat: $n_1 = 25$, $\bar{x}_1 = 294,220$, $\sigma_1 = 135,387$

Solution: Casper, Cheyenne $n_2 = 25$, $\bar{x}_2 = 287,984$, $\sigma_2 = 151,996$

$\alpha = 1\%$

a) State $H_0$and H_a and identify the claim.

$H_0: \mu_1 = \mu_2$ claim

$H_A: \mu_1 \neq \mu_2$

$H_0: \mu_1 - \mu_2 = 0$ → claim

$H_A: \mu_1 - \mu_2 \neq 0$ "T.T.T"

b) Find the standardized test statistics.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{294,220 - 287,984}{\sqrt{\frac{(135,387)^2}{25} + \frac{(151,996)^2}{25}}}$$

≈ 0.15

c) Compute the $P-value$.

$P-value = 2 * p(z > 0.15)$

$= 2 * (1 - 0.5596)$ → from z-tab

$= .8808 > \alpha = 0.01$

R code.

```
2*(1-pnorm(0.15))      -->   0.8807646
```

d) Decide whether to reject or fail to reject H_0.

Fail to reject H_0 ⟶ Fail to reject the claim.

e) Interpret the decision in the context of the original claim.

There is not enough evidence at the 1% level of significance to reject the real estate agency's claim that the mean home sales prices in Casper, Wyoming, is the same as in Cheyenne, Wyoming. P-value = .8808

-Demonstrate the use of a two–sample z-test for the difference between proportions

Motivation:

1) Claim: The proportion of women voting for Candidate A = The proportion of men voting for Candidate A.
   ➔ Claim: $P_1 = P_2$

2) Claim: The proportion of patients cured using the new treatment > The proportion of patients cured using the old treatment.
   ➔ Claim: $P_1 > P_2$

3) Claim: The proportion of people who own a house ≤ The proportion of people who rent a house.
   ➔ Claim: $P_1 \leq P_2$

Based on the results above the z-test statistic for the difference of two proportions$(P_1 - P_2)$.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$\begin{cases} H_0: p_1 = p_2 \\ H_a: p_1 \neq p_2 \end{cases}$, $\begin{cases} H_0: p_1 \leq p_2 \\ H_a: p_1 > p_2 \end{cases}$, and $\begin{cases} H_0: p_1 \geq p_2 \\ H_a: p_1 < p_2 \end{cases}$. which is equivalent to

$\begin{cases} H_0: p_1 - p_2 = 0 \\ H_a: p_1 - p_2 \neq 0 \end{cases}$, $\begin{cases} H_0: p_1 - p_2 \leq 0 \\ H_a: p_1 - p_2 > 0 \end{cases}$, and $\begin{cases} H_0: p_1 - p_2 \geq 0 \\ H_a: p_1 - p_2 < 0 \end{cases}$.

Simplify the denominator of the previous test: under the assumption that $P_1 = P_2 = \bar{p}$

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{\bar{p}\,\bar{q}}{n_1} + \frac{\bar{p}\,\bar{q}}{n_2}} = \sqrt{\bar{p}\,\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ is called the weighted estimate of P_1 and P_2:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\bar{p}\,\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(weighted estimate $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ and $\bar{q} = 1 - \bar{p}$)

-Ex:

Example #1: In a survey of 480 drivers from the South, 408 wear a seat belt. In a survey of 360 drivers from the Northeast, 288 wear a seatbelt. At $\alpha = 0.05$, can you support the claim that the proportion of drivers who wear seat belts is greater in the South than in the Northeast? Assume that the samples are random and independent.

Solution:

1) Find the weighted estimate of P_1 and P_2. Verify that $n_1\bar{p}$, $n_1\bar{q}$, $n_2\bar{p}$, and $n_2\bar{q}$ are at least 5.

2) Identify the claim and State H_0 and H_a.

4) Find the standardized test statistics.

4) Compute the P-value.

$P - value = p(z > 1.9)$

$= (1 - 0.9713)$ → from z-tab

$= 0.0287 < \alpha = 0.05$

R code.

```
(1-pnorm(1.9))      -->   0.02871656
```

5) Decide whether to reject or fail to reject H_0.

⇒ Reject Ho

⇒ support the claim (Ha)

6) Interpret the decision in the context of the original claim.

There is enough evidence at the 5% level of significance to support the claim that the proportion of drivers who wear seat belts is greater in the South than in the Northeast. P-value = .0287

R code:

```
prop.test(x=c(408,288),n=c(480,360),alternative = "greater", correct = FALSE)
#(408,288) =(n1,n2)→test p1-p2>0.
```

Example #2: In a survey of 200 males ages 18 to 24, 39% were enrolled in college. In a survey of 220 females ages 18 to 24, 45% were enrolled in college. At $\alpha = 0.05$, can you support the claim that the proportion of males ages 18 to 24 who enrolled in college is less than the proportion of females ages 18 to 24 who enrolled in college? Assume that the samples are random and independent.

Solution:

1) Find the weighted estimate of P_1 and P_2. Verify that $n_1\bar{p}$, $n_1\bar{q}$, $n_2\bar{p}$, and $n_2\bar{q}$ are at least 5.

2) Identify the claim and State H_0 and H_a.

claim: $P_1 < P_2$ ⇒ Ha left-tailed test

Ho: $P_1 \geq P_2$

Ha: $P_1 < P_2$ → claim

4) Find the standardized test statistics.

4) Compute the P-value.

P-value = $P(z < -1.24)$ ⇒ using z-table

$= .1075 > \alpha = .05$ → Fail to reject Ho

5) Decide whether to reject or fail to reject H_0.

⇒ Fail to reject Ho

⇒ Do not support the claim (Ha)

6) Interpret the decision in the context of the original claim.

There is not enough evidence at the 5% level of significance to support the claim that the proportion of males ages 18 to 24 who enrolled in college is less than the proportion of females ages 18 to 24 who enrolled in college. P-value = .1075

R code.

```
pnorm(-1.24)      -->   0.1074877
```

R code:

```
prop.test(x=c(78,99),n=c(200,220),alternative = "less",correct = FALSE)
#(78,99) =(n1,n2)→test p1-p2<0.
```

8.2 Testing the Difference Between Two Population Means (with $\sigma 1$ and $\sigma 2$ Unknown.)

-The Two-Sample t − ***Test*** for the difference Between Means for Independent data

Two-Sample $t-test$ with variances **are** equal.	Two-Sample $t-test$ with variances **are not** equal.
A two-sample $t-test$ is used to test the difference between two population means μ_1 and μ_2 when 1) The samples are random. 2) The samples are independent. 3) $\bar{x}_1$and $\bar{x}_2$are normally distributed ➔ both populations are normally distributed or both $n_1 \geq 30$ and $n_2 \geq 30$ 4) σ_1 and σ_2 are unknown. 5) $\sigma_1 = \sigma_2$.	A two-sample $t-test$ is used to test the difference between two population means μ_1 and μ_2 when 1) The samples are random. 2) The samples are independent. 3) $\bar{x}_1$and $\bar{x}_2$are normally distributed ➔ both populations are normally distributed or both $n_1 \geq 30$ and $n_2 \geq 30$ 4) σ_1 and σ_2 are unknown. 5) $\sigma_1 \neq \sigma_2$.
In this class the question will specify whether the population variances are equal or not but in real life you need to justify that by using `boxplot(y1,y2)` or other techniques that were not discussed in this class.	In this class the question will specify whether the population variances are equal or not but in real life you need to justify that by using `boxplot(y1,y2)` or other techniques that were not discussed in this class.
The standardized test statistic of $(\bar{x}_1 - \bar{x}_2)$ is $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{x_1} - \mu_{x_2})}{s_{(\bar{x}_1 - \bar{x}_2)}}$	The standardized test statistic of $(\bar{x}_1 - \bar{x}_2)$ is $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{x_1} - \mu_{x_2})}{s_{(\bar{x}_1 - \bar{x}_2)}}$
The estimate of the variance is called the pooled estimate of the standard deviation: $s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ With $d.f = n_1 + n_2 - 2$	The estimate of the variance is: $s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ With $d.f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{1}{(n_1-1)}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{(n_2-1)}\left(\frac{s_2^2}{n_2}\right)^2\right]}$

//TODO Examples

-The t − ***Test*** for the difference Between Means for a population of Dependent (paired) for before and after data.

-Convert 2 dependent into 1 independent data set

-1. Pick y1 & y2.

-Always pick the data set that is claimed to be “increased”/”improved” for y2, and the other for y1. (or just abs(d_bar))

-The claim is then μ2 - μ1 > 0 (μd > 0).

-Otherwise, the claim is that they are equal, so it doesn't matter.

-2. d (difference) & dbar (avg of all differences)

$$d = y_2 - y_1 \qquad \bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

-3. s_d (Sample StDev of d)

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$$

-4. t & degree of freedom

$$t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}} = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \quad (\mu d \text{ is } 0) \quad d.f = n - 1$$

-5. Plug and chug into t-dist

//TODO Examples

9.1: Correlation

-Motivation

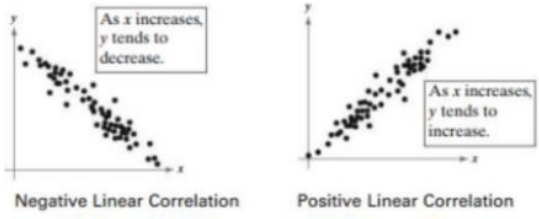
Suppose a safety inspector wants to determine whether a relationship exist between the number of hours of training for an employee and the number of accidents involving the employee.

Supposed a psychologist want to know whether a relationship exists between the number of hours a person sleeps each night and the person's reaction time.

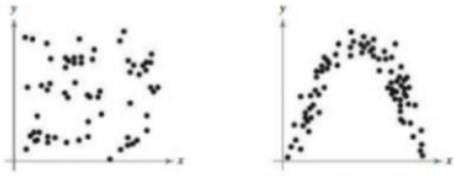
-Correlation

A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y), where x is the **independent** (or **explanatory**) **variable** and y is the **dependent** (or **response**) **variable.**

-Types of Correlation



Negative Linear Correlation | Positive Linear Correlation



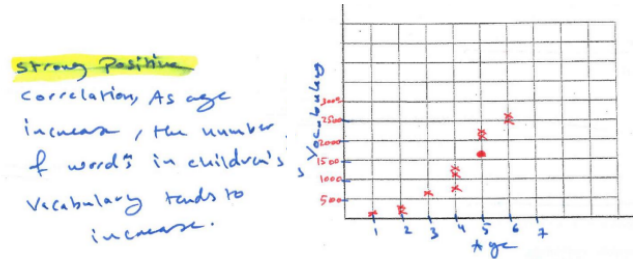
No Correlation | Nonlinear Correlation

-Scatter Plot and Correlation

-A subjective way to see if there is correlation (subjective because no numbers)

-Ex:

Age, *x*	1	2	3	4	5	6	3	5	2	4	6
Vocabulary size, *y*	3	220	540	1100	2100	2600	730	2200	260	1200	2500



-R Code:

```
x=c(1,2,3,4,5,6,3,5,2,4,6)
y=c(3,220,540,1100,2100,2600,730,2200,260,1200,2500)
plot(x, y, main=" speech therapy",xlab=" Age", ylab=" Vocabulary size")
```

-Correlation Coefficient

-Motivation: quantify linear relationship between 2 quantitative variables. Since each have their own Dtdev, We divide Cov by Stdev to (kinda) normalize output.

We wish to quantify the strength of the **linear** relationship between two quantitative variables starting with the covariance, which is a measurement for the strength of the linear relationship between the two quantitative variables. However, the covariance magnitude changes with the nature of the variables. For example, if we compute the covariance between dish soap price and milk price and compare it with the covariance between housing prices and the number of people who drink OJ, the covariance between housing prices and the number of people who drink OJ will be inflated by the nature of the numbers because the numbers are large in comparison to the covariance between dish soap and milk price. Thus, we need to adjust the covariance in order to be able to compare which relation is a stronger relationship.

-Population Correlation

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{E(xy) - E(x)E(y)}{\sqrt{E(x^2) - (E[x])^2}\sqrt{E(y^2) - (E[y])^2}} = \frac{\mu_{xy} - \mu_x \mu_y}{\sqrt{\mu_{x^2} - (\mu_x)^2}\sqrt{\mu_{y^2} - (\mu_y)^2}}$$

-Sample Correlation

$$r_{xy} = \frac{\widehat{Cov(x,y)}}{s_x s_y} = \frac{\frac{\sum_{i=1}^{n}(xy)}{n} - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n}}{\sqrt{\frac{\sum_{i=1}^{n}(x^2)}{n} - \left(\frac{\sum_{i=1}^{n} x}{n}\right)^2}\sqrt{\frac{\sum_{i=1}^{n}(y^2)}{n} - \left(\frac{\sum_{i=1}^{n} y}{n}\right)^2}}$$ multiply by $\frac{n^2}{n^2}$ and simplify,

$$= \frac{n\sum_{i=1}^{n}(xy) - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{\sqrt{n\sum_{i=1}^{n}(x^2) - \left(\sum_{i=1}^{n} x\right)^2}\sqrt{n\sum_{i=1}^{n}(y^2) - \left(\sum_{i=1}^{n} y\right)^2}}$$

-R Code: cor(x, y)

-Theorem: -1 < ϱ_{xy} < 1

$$|Cov(x,y)|^2 = \left|E([x-\mu_x][y-\mu_y])\right|^2 = \left|\langle x-\mu_x, y-\mu_y\rangle\right|^2$$

$$\leq \langle x-\mu_x, x-\mu_x\rangle \cdot \langle y-\mu_y, y-\mu_y\rangle \quad \text{by Cauchy – Schwarz inequality}$$

The Cauchy–Schwarz inequality states that for all vectors u and v of an inner product space, it is true that $|\langle u, v\rangle|^2 \leq \langle u, u\rangle \cdot \langle v, v\rangle$.

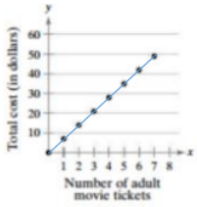
$$= E([x-\mu_x]^2)E\left([y-\mu_y]^2\right)$$

$$= Var(x)Var(y)$$

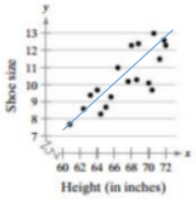
$$|Cov(x,y)|^2 \leq Var(x)Var(y) \rightarrow \frac{|Cov(x,y)|^2}{Var(x)Var(y)} \leq 1 \rightarrow -1 \leq \frac{Cov(x,y)}{\sigma_x\sigma_y} \leq 1$$

-Perfect, Strong, Weak, or No Correlation

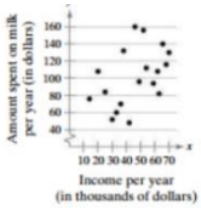
1) The range of the correlation coefficient is between -1 and 1, inclusive. $(-1 \leq r \leq 1)$
2) When x and y have a strong **positive** linear correlation, r is close to 1.
3) When x and y have a strong **negative** linear correlation, r is close to -1.
4) When there is no linear correlation, r is close to 0.
5) If r is close to 0, it does not mean that there is no relationship between x and y, just that there is no linear relation.



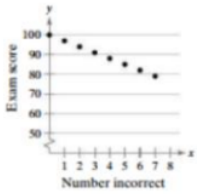
Perfect positive correlation
$r = 1$



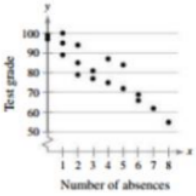
Strong positive correlation
$r = 0.81$



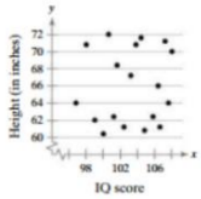
Weak positive correlation
$r = 0.45$



Perfect negative correlation
$r = -1$



Strong negative correlation
$r = -0.92$



No correlation
$r = 0.04$

-Steps

Calculating a Correlation Coefficient

In Words	In Symbols
1. Find the sum of the x-values.	Σx
2. Find the sum of the y-values.	Σy
3. Multiply each x-value by its corresponding y-value and find the sum.	Σxy
4. Square each x-value and find the sum.	Σx^2
5. Square each y-value and find the sum.	Σy^2
6. Use these five sums to calculate the correlation coefficient.	$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

-Ex: The age (in years) of 11 children and the number of words in their vocabulary.

#	Age, x	Vocabulary size, y	$x \cdot y$	x^2	y^2
1	1	3	1*3 = 3	$(1)^2$ = 1	$(3)^2$ = 9
2	2	220	2*220 = 440	$(2)^2$ = 4	$(220)^2$ = 48,400
3	3	540	3*540 = 1620	$(3)^2$ = 9	$(540)^2$ = 291,600
4	4	1100	4*1100 = 4400	$(4)^2$ = 16	$(1100)^2$ = 1,210,000
5	5	2100	5*2100 = 10500	$(5)^2$ = 25	$(2100)^2$ = 4,410,000
6	6	2600	6*2600 = 15600	$(6)^2$ = 36	$(2600)^2$ = 6,760,000
7	3	730	3*730 = 2190	$(3)^2$ = 9	$(730)^2$ = 532,900
8	5	2200	5*2200 = 11000	$(5)^2$ = 25	$(2200)^2$ = 4,840,000
9	2	260	2*260 = 520	$(2)^2$ = 4	$(260)^2$ = 67,600
10	4	1200	4*1200 = 4800	$(4)^2$ = 16	$(1200)^2$ = 1,440,000
11	6	2500	6*2500 = 15000	$(6)^2$ = 36	$(2500)^2$ = 6,250,000
Σ	41	13,453	66,073	181	25,850,504

Σx Σy Σxy Σx^2 Σy^2

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{11 * 66073 - 41 * 13453}{\sqrt{11 * 181 - (41)^2} * \sqrt{11 * 25850504 - (13453)^2}}$$

(11*66073-41*13453)/(sqrt(11*181-41^2)*sqrt(11*25850504-13453^2)) → 0.978871

Interpretation: Strong Positive correlation; As age increases the number of words in children's vocabulary tends to increase.

("tends to" is different than "causes", which correlation doesn't help with)

-R Code:

```
x=c(1,2,3,4,5,6,3,5,2,4,6)
y=c(3,220,540,1100,2100,2600,730,2200,260,1200,2500)
cor(x, y) #0.9788707
```

-Hypothesis Test a Population Correlation Coefficient ϱ

-Possible Combo of Claims

$\begin{cases} H_0: \rho \geq 0 \text{ (no significant negative correlation)} \\ H_a: \rho < 0 \text{ (significant negative correlation)} \end{cases}$ Left-tailed test

$\begin{cases} H_0: \rho \leq 0 \text{ (no significant positive correlation)} \\ H_a: \rho > 0 \text{ (significant positive correlation)} \end{cases}$ Right-tailed test

$\begin{cases} H_0: \rho = 0 \text{ (no significant correlation)} \\ H_a: \rho \neq 0 \text{ (significant correlation)} \end{cases}$ Two-tailed test

-Standardized t-test

A ***t*-test** can be used to test whether the correlation between two variables is significant. The **test statistic** is r and the **standardized test statistic**

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

follows a t-distribution with $n - 2$ degrees of freedom, where n is the number of pairs of data. (Note that there are $n - 2$ degrees of freedom because one degree of freedom is lost for each variable.)

(df = n-2, 2 for each variable)

-Proof: If $y = mx + b$ where m is a positive real number and b is any real number, then $\varrho\_x,y = 1$.

$$\begin{aligned}
\rho_{xy} &= \frac{Cov(x,\, y)}{\sigma_x \sigma_y} \\
&= \frac{Cov(x,\, xm+b)}{\sigma_x \sigma_{xm+b}} \\
&= \frac{Cov(x,\, xm)}{\sigma_x \sigma_{xm+b}} \quad [Cov(x+a, y+b) = Cov(x,y)] \\
&= \frac{mCov(x,\, x)}{\sigma_x \sigma_{xm+b}} \quad [Cov(x, yc) = cCov(x,y)] \\
&= \frac{mVar(x)}{\sigma_x \sigma_{xm+b}} \quad [Cov(x,x) = Var(x,x)] \\
&= \frac{mVar(x)}{\sqrt{Var(x)}\sqrt{Var(xm+b)}} \\
&= \frac{mVar(x)}{\sqrt{Var(x)}\sqrt{m^2 Var(x)}} \quad [Var(ax+c) = a^2 Var(x)] \\
&= \frac{mVar(x)}{\sqrt{Var(x)}m\sqrt{Var(x)}} \\
&= \frac{Var(x)}{\sqrt{Var(x)}\sqrt{Var(x)}} \\
&= \frac{Var(x)}{Var(x)} \\
&= 1
\end{aligned}$$

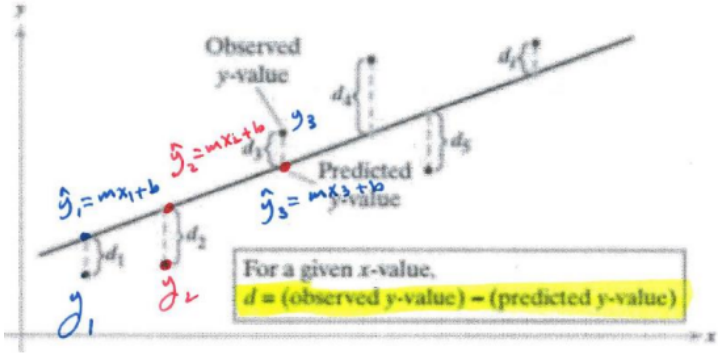
-Proof: Let a and b be any real number, then prove that $\varrho_{(ax)(by)} = \varrho_{xy}$.

$$\begin{aligned}
\rho_{(ax)(by)} &= \frac{Cov(ax, by)}{\sigma_{ax}\sigma_{by}} \\
&= \frac{abCov(x, y)}{\sigma_{ax}\sigma_{by}} [Cov(x, cy) = cCov(x, y) \,\&\, Cov(cx, y) = cCov(x, y)] \\
&= \frac{abCov(x, y)}{\sqrt{Var(ax)}\sqrt{Var(by)}} \\
&= \frac{abCov(x, y)}{\sqrt{a^2Var(x)}\sqrt{b^2Var(y)}} \left[Var(cx) = c^2Var(x)\right] \\
&= \frac{abCov(x, y)}{a\sqrt{Var(x)}b\sqrt{Var(y)}} \\
&= \frac{Cov(x, y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{Cov(x, y)}{\sigma_x\sigma_y} = \rho_{xy}
\end{aligned}$$

9.2: Linear Regression

-Motivation

After verifying that the linear correlation between two variables is significant, the next step is to determine the equation of the line that best models the data. This line is called a regression line and its equation can be used to predict the value of y for a given value of x.

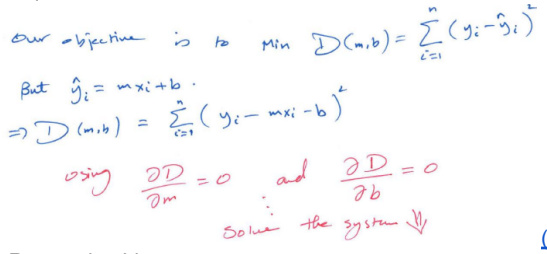


d is aka residual value of a point

-Linear Regression Line (Aka Best Fit Line)

A **regression line,** also called a **line of best fit,** is the line for which the sum of the squares of the residuals is a minimum.

-Sum of the squares of the resudals is a minimum.



[(4) Linear Regression: Derivation - YouTube](#)

-Equation of a Linear Regression Line

The equation of a regression line for an independent variable x and a dependent variable y is

$$\hat{y} = mx + b$$

where $\hat{y}$ is the predicted y-value for a given x-value. The slope m and y-intercept b are given by

$$m = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \quad \text{and} \quad b = \bar{y} - m\bar{x} = \frac{\Sigma y}{n} - m\frac{\Sigma x}{n}$$

where $\bar{y}$ is the mean of the y-values in the data set, $\bar{x}$ is the mean of the x-values, and n is the number of pairs of data. The regression line always passes through the point $(\bar{x}, \bar{y})$.

-Ex: The number of hours 9 students spent studying for a test and their scores on that test.

#	Hours spent studying, x	Test scores, y	$x \cdot y$	x^2
1	0	40	0*40 = 0	(0)² = 0
2	2	51	2*51 = 102	(2)² = 4
3	4	64	4*64 = 256	(4)² = 16
4	5	69	5*69 = 345	(5)² = 25
5	5	73	5*73 = 365	(5)² = 25
6	5	75	5*75 = 375	(5)² = 25
7	6	93	6*93 = 558	(6)² = 36
8	7	90	7*90 = 630	(7)² = 49
9	8	95	8*95 = 760	(8)² = 64
Σ	42	650	3391	244

Σx Σy Σxy Σx²

a) Find the equation of the regression line for the data.

$$m = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{9*3391 - 42*650}{9*244 - (42)^2} = \frac{3219}{432} \approx 7.451$$

$$b = \frac{\Sigma y}{n} - m\frac{\Sigma x}{n} = \frac{650}{9} - 7.45 * \frac{42}{9} \approx 37.450$$

$\hat{y} = 7.451x + 37.450 \Rightarrow$ for x between 0 and 8

-R Code:

```
x=c(0,2,4,5,5,5,6,7,8)
y=c(40,51,64,69,73,75,93,90,95)
lm(y~x)
summary(lm(y~x)) # alternative show results
```

-Predict using regression (Don't do it if it's outside range of data)

b) Use the regression equation to predict the value of y for the x-values, if meaningful. If the x-values are not meaningful to predict the value of y, explain why not.

I. $x = 3$ hours

$\hat{y} = 7.451*3 + 37.450 \approx 60$

II. $x = 6.5$ hours

$\hat{y} = 7.451*6.5 + 37.450 \approx 86$

III. $x = 13$ hours

It is not meaningful to predict the value of y for $x=13$ because $x=13$ is outside the range of the original data.

IV. $x = 4.5$ hours

$\hat{y} = 7.451*4.5 + 37.450 \approx 71$

-Multiple Regression Equation & Model

A multiple regression equation for independent variables $x_1, x_2, x_3, \ldots, x_k$ and a dependent variable y has the form

$$\hat{y} = b + m_1x_1 + m_2x_2 + m_3x_3 + \cdots + m_kx_k$$

where $\hat{y}$ is the predicted y-value for given x_i values and b is the y-intercept. The y-intercept b is the value of $\hat{y}$ when all x_i are 0. Each coefficient m_i is the amount of change in $\hat{y}$ when the independent variable x_i is changed by one unit and all other independent variables are held constant.

(Finding it requires Matrix Algebra, so we use R to do the heavy lifting)

-Ex:

Example #3: A researcher wants to determine how employee salaries at a company are related to the length of employment, previous experience, and education. The researcher selects eight employees randomly from the company and obtains the data shown in the table.

-a. Model selection and goodness-of-fit parameters.

In a multi regression the first step is model selection. Should I include all variables? Or do I need more variables? How well does the model fit the data?

Because the mathematics associated with multiple regression involves matrix algebra, this section focuses on how to use R/technology to find a multiple regression equation and how to interpret the results.

Step #1: Start with the full model → The model that includes all variables (Try with all variables)

-R Code:

```
y=c(57310,57380,54135,56985,58715,60620,59200,60320)
x1=c(10,5,3,6,8,20,8,14)
x2=c(2,6,1,5,8,0,4,6)
x3=c(16,16,12,14,16,12,18,17)
summary(lm(y ~ x1 + x2 + x3)) # show results
```

```
Residuals:
     1      2      3      4      5      6      7      8
-824.76  156.82 -153.52  158.90  -56.65  364.09  804.95 -449.82

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 49764.45    1981.35  25.116 1.49e-05 ***
x1            364.41      48.32   7.542  0.00166 **
x2            227.62     123.84   1.838  0.13991
x3            266.94     147.36   1.812  0.14430
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 659.5 on 4 degrees of freedom
Multiple R-squared:  0.9438,   Adjusted R-squared:  0.9017
F-statistic:  22.4 on 3 and 4 DF,  p-value: 0.005804
```

m = 364.41, 227.62, 266.94 (p-value = 0.00166, 0.13991, 0.14430)

aka lev. of sig. (x1 is not significant here)

-R^2 is aka Coefficient of Determination

-Residual Standard Error is the avg error of a predicted value,

Based on the output, the multiple regression model $\hat{y} = 4976.45 + 364.41x_1 + 227.62x_2 + 266.94x_3$ has a Multiple R-squared (coefficient of determination) of 0.9438 with p-value = 0.005804.

The model is significant and the R-square of 0.9438 tells us that 94.38% of the variation in y can be explained by the multiple regression model. The remaining 5.62% is unexplained and is due to other factors such as sampling error, coincidence, or lurking variables.

Do we need all three variables?

Based on the results above, we can see that we fail to reject $H_0: m_2 = 0$ ($P - value = 0.13991$)and $H_0: m_3 = 0$ ($P - value = 0.14430$), thus we do not need all three variables.

We start by eliminating one of these variables and see the effect of that on the model. (Try removing a variable)

Eliminate Education (in years), x_3	Eliminate Experience (in years), x_2
`mod <- lm(y ~ x1 + x2)` `summary(mod) # show results`	`mod <- lm(y ~ x1 + x3)` `summary(mod) # show results`

`Call:`
`lm(formula = y ~ x1 + x2)`
 	
`Residuals:`
`1 2 3 4 5 6 7 8`
`-309.141 147.124 -482.872 -250.568 -389.966 -8.488 1586.244 -292.332`
 	
`Coefficients:`
`Estimate Std. Error t value Pr(>|t|)`
`(Intercept) 53118.58 851.09 62.412 2e-08 ***`
`x1 375.50 57.84 6.492 0.00129 **`
`x2 372.80 113.92 3.272 0.02214 *`
`---`
`Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`
 	
`Residual standard error: 795.9 on 5 degrees of freedom`
`Multiple R-squared: 0.8977, Adjusted R-squared: 0.8568`
`F-statistic: 21.95 on 2 and 5 DF, p-value: 0.003343`
	`Call:`
	`lm(formula = y ~ x1 + x3)`
	
	`Residuals:`
	`1 2 3 4 5 6 7 8`
	`-1412.3 339.2 -464.3 492.3 665.3 303.6 265.8 -189.7`
	
	`Coefficients:`
	`Estimate Std. Error t value Pr(>|t|)`
	`(Intercept) 48283.71 2198.97 21.957 3.64e-06 ***`
	`x1 336.30 55.68 6.040 0.00179 **`
	`x3 442.23 136.46 3.241 0.02293 *`
	`---`
	`Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`
	
	`Residual standard error: 801.1 on 5 degrees of freedom`
	`Multiple R-squared: 0.8964, Adjusted R-squared: 0.8549`
	`F-statistic: 21.63 on 2 and 5 DF, p-value: 0.003456`

The multiple regression model $\hat{y} = 53118.58 + 375.50x_1 + 372.80x_2$	The multiple regression model $\hat{y} = 48283.71 + 336.30x_1 + 442.23x_3$

The models $\hat{y} = 53118.58 + 375.50x_1 + 372.80x_2$ and $\hat{y} = 48283.71 + 336.30x_1 + 442.23x_3$ are considered to be better models than $\hat{y} = 49764.45 + 364.41x_1 + 227.62x_2 + 266.94x_3$ since they all have close R-square.

But in the first and second models, all parameters are significant. That is not the case in the third model.

-b. Use the regression equations to predict an employee's salary for each set of conditions.

I. 12 years of current employment
II. 5 years of previous experience
III. 16 years of education

Model	$\hat{y} = 49{,}764.45 + 364.41x_1 + 227.62x_2 + 266.94x_3$	$\hat{y} = 53{,}118.58 + 375.50x_1 + 372.80x_2$	$\hat{y} = 48{,}283.71 + 336.30x_1 + 442.23x_3$
Estimate	$\hat{y} = 49{,}764.45 + 364.41 * 12 + 227.62 * 5 + 266.94 * 16$ ≈59,546.51 Based on this model, the employee's predicted salary is $59,546.51	$\hat{y} = 53{,}118.58 + 375.50 * 12 + 372.80 * 5$ ≈59,488.58 Based on this model, the employee's predicted salary is $59,488.58	$\hat{y} = 48{,}283.71 + 336.30 * 12 + 442.23 * 16$ ≈59,394.99 Based on this model, the employee's predicted salary is $59,394.99