

STA2023

Summary Notes

Chapter 1 - 10

Dr. Mohammad Shakil

Editor: Jeongmin Correa

Contents

Chapter 1: The Nature of Probability and Statistics

Chapter 2: Frequency distribution and Graphs

Chapter 3: Data Description

Chapter 4: Probability and Counting Rules

Chapter 5: Discrete Probability Distributions

Chapter 6: The Normal Distribution

Chapter 7: Confidence Intervals and Sample Size

Chapter 8: Hypothesis Testing

Chapter 9: Testing the Difference Between Two Means,
Two Variances, and Two Proportions

Chapter 10: Correlation and Regression



Ch1

1 - 1 Descriptive and Inferential Statistics

- **Statistics**

The Methods of classification
and Analysis of numerical & non-numerical data
For Drawing valid conclusion and making reasonable decisions.

< Two Major Areas of Statistics >

Descriptive Statistics	Inferential Statistics
It consists of the collection, organization, summarization, and presentation of data. (It describes the situation as it is).	It consists of making inferences from samples to populations, hypothesis testing, determining relationships among variables, and making predictions. (It is based on probability theory.)

* **Probability**; the chance of an event occurring.
Cards, dice, bingo, & lotteries

In order to gain information about seemingly haphazard events, statisticians study random variables.

1. **Variables**

A variable is a characteristic or an attribute that can assume different values.
Height, weight, temperature, number of phone calls received, etc.

2. **Random Variables**

Variables whose values are determined by chance

< Collection of Data >

The collection of data constitutes the starting point of any statistical investigation. It should be conducted systematically with a definite aim in view and with as much accuracy as is desired in the final results, for detailed analysis would not compensate for the bias and inaccuracies in the original data.

1. **Data;** the measurements or observations (values) for a variable
2. **Data Set;** A collection of data values
3. **Data Value or Datum:** Each value in the data set

Example:

Suppose a researcher selects a specific day and records the number of calls received by a local office of the Internal Revenue Service each hour as follows: {8, 10, 12, 12, 15, 11, 13, 6}, where 8 is the number of calls received during the first hour, 10 the number of calls received during the second hour, and so on.

The collection of these numbers is an example of a data set, and each number in the data set is a data value.

Data may be collected for each and every unit of the whole lot (called population), for it would ensure greater accuracy.

But, however, since in most cases the populations under study are usually very large, and it would be difficult and time-consuming to use all members, therefore statisticians use subgroups called samples to get the necessary data for their studies. The conclusions drawn on the basis of this sample are taken to hold for the population

1. Population

the totality of all subjects possessing certain common characteristics that are being studied.

2. Sample; a subgroup or subset of the population.

3. Random Sample

A sample obtained without bias or showing preferences in selecting items of the population is called a random sample.

1 – 2 Variables and Types of Data

< Classification of Variables (and Data) >

1. Qualitative Variables

– **No mathematical meaning or Non-numerical**

variables that can be placed into distinct categories, according to some characteristic or attribute.

Ex) gender, religious preferences, geographic locations, grades of a student, car's tags, numbers on the uniforms of baseball players, etc.

2. Quantitative Variables

numerical in nature and can be ordered or ranked.

Ex) age, heights, weights, body temperatures, etc.

Discrete Variables	Continuous Variables
assume values that can be counted such as whole numbers	can assume all values between any two specific values by measuring.
Ex) the number of children in a family, the number of students in a class-room, the number of calls received by a switchboard operator each day for one month, batting order numbers of baseball, etc.	Ex) Temperature, height, weight, length, time, speed, etc.

*Since continuous data **must be measured**, rounding answers is necessary because of the limits of the measuring device. Usually, answers are **rounded to the nearest given unit**

→ (there is time between 2 seconds, , it must be rounded up.)

Ex) Heights must be rounded to the nearest inch, weights to the nearest ounce, etc. Hence, a recorded height of 73 inches would mean any measure of 72.5 inches up to but not including 73.5 inches.

Thus, the boundary of this measure is given as 72.5 – 73.5 inches.

(We have taken 72.5 as one of the boundaries since it could be rounded to 73. But, we cannot include 73.5 because it would be 74 when rounded). Sometimes 72.5 – 73.5 is called a class which will contain the recorded height of 73 inches.

The concept of the boundaries of a continuous variable is illustrated in the following Table I:

TABLE I Variable	Recorded Value	Boundaries (Class)
Length	15 cm	14.5 – 15.5 cm
Temperature	86 ⁰ F	85.5 – 86.5 ⁰ F
Time	0.43 sec	0.425 – 0.435 sec
Weight	1.6 gm	1.55 – 1.65 gm

Note: The boundaries of a continuous variable in the above table are given in one additional decimal place and always end with the digit 5.

< MEASUREMENT SCALES OF A DATA: >

1. Nominal-level Data (no order or no comparing values)

– Equality, Categories, No mathematical meaning –Binomial

The nominal-level of measurement classifies data into mutually exclusive (non-overlapping), exhaustive categories in which no ordering or ranking can be imposed on the data.

2. Ordinal-Level Data – Order , Rank (Qualitative data)

The ordinal-level of measurement classifies data into categories that can be ordered or ranked. (only before and after no bigger or less..)
However, precise differences between the ranks do not exist.

Interval-level Data (Quantitative data)

The interval-level of measurement ranks data, and precise differences between units of measure do exist. (equal distances between 2 points)
However, there is no meaningful zero (i.e., starting point)

3. Ratio-level Data (Quantitative data)

possesses all the characteristics of interval measurement (i.e., data can be ranked, and there exists a true zero or starting point).
In addition, true ratios exist between different units of measure.

Nominal	Ordinal	Interval	Ratio
No order or rank Equality, Categories, No mathematical meaning	Order , Rank , No equal distance between 2 ranks	No meaningful zero, Equal distances between 2 points	True zero
Zip code, Gender, Color, Ethnicity, Political affiliation, Religious affiliation, Major field, Nationality, Marital status, Sports player's back numbers, , AM & PM, Date, Credit card numbers	Grade (ABCDF), Judging (1 st , 2 nd , 3 rd), Rating scale (Excellent, good, bad), Ranking of sports players, Week, Months, Mon ~ Fri, left center right, Morning, Afternoon, Evening, Birthdays	Ex) STA score, IQ, Temperature, 12 hours of day, Date of a week, Days of a month, Months of a year	Ex) Height, Weight, Time, Salary, Age, 24 hours of days (0 = 24)

Nominal; Sue is young, and Mary is old.

Ordinal; Sue is younger than Mary.

Interval; Sue is 20 years younger than Mary.

Ratio ; Sue is twice as young as Mary.

1 – 3 Data Collection and Sampling Techniques

When the population is large and diverse, a sampling method must be designed so that the sample is representative, unbiased and random, i.e. every subject (or element) in the population has an equal chance of being selected for the sample.

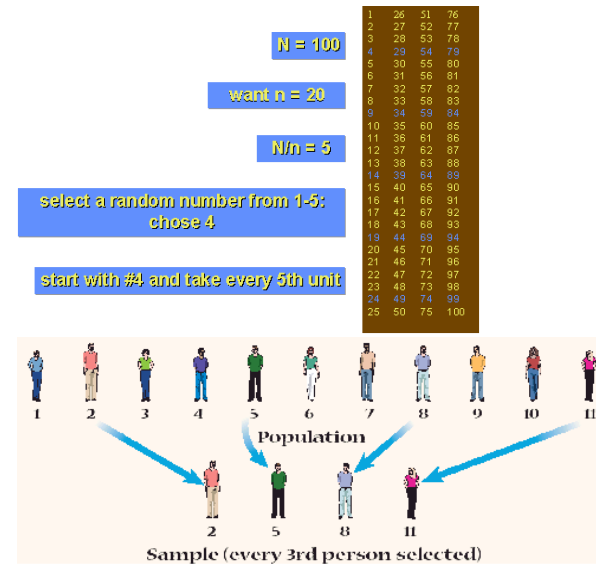
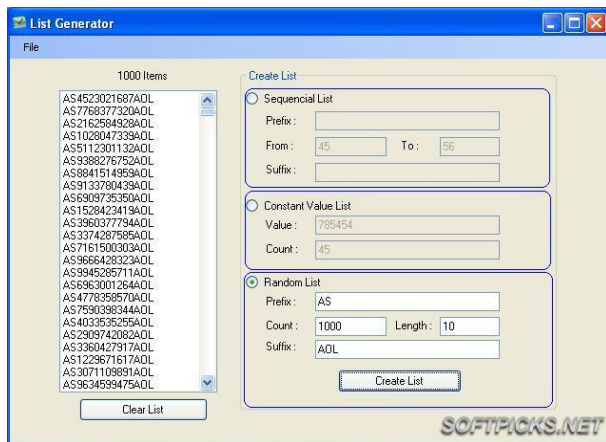
1. Random Sampling

This method requires that each member of the population be identified and assigned a number.

Then a set of numbers drawn randomly from this list forms the required random sample.

Note that each member of the population has an equal chance of being selected.

Ex) For a large population, computers are used to generate random numbers which contain series of numbers arranged in random order.



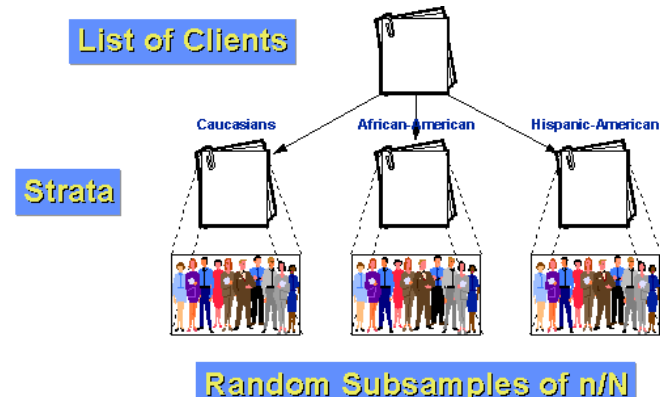
3. Stratified Sampling

This method requires that the population be classified into a number of smaller homogeneous strata or subgroups.

A sample is drawn randomly from each stratum.

= Subdivide the population into at least 2 different subgroups (or strata) so that subject within the same characteristics (such as gender or age bracket) then draw a sample from each subgroup.

Ex) age, sex, marital status, education, religion, occupation, ethnic background or virtually any characteristic.



2. Systemic Sampling – K^{th} – every 5th numbers

This method requires that every k^{th} member (or item) of the population be selected to form the required random sample.

Ex) We might select every 10th house on a city block for the random sample.

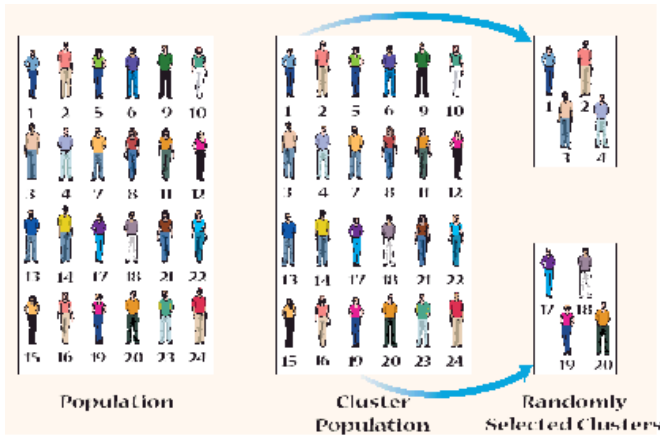
4. Cluster Sampling

The population area is first divided into a number of sections (or subpopulations) called clusters.

A few of those clusters are randomly selected, and sampling is carried out only in those clusters.

(and then choose all members from the selected clusters)

Ex) a community can be divided into city blocks as its clusters. Several blocks are then randomly selected. After this, residents on the selected blocks are randomly chosen, providing a sampling of the entire community.



< Statistical Inference and Measurement of Reliability >

A statistical inference is an estimate or prediction or some other generalization about a population based on information contained in a random sample of the population. That is, the information contained in the random sample is used to learn about the population.

A measure of reliability is a statement (usually quantified) about the degree of uncertainty associated with a statistical inference.

< Elements of Descriptive and Inferential Statistical Problems >

1. Four Elements of Descriptive Statistical Problems

- The population or sample of interest.
- One or more variables (characteristics of the population or sample units) that are to be investigated.
- Tables, graphs, numerical summary tools.
- Identification of patterns in the data.

2. Five Elements of Inferential Statistical Problems

- The population of interest.
- One or more variables that are to be investigated.
- The sample of population units.
- The statistical inference about the population based on information contained in the random sample of the population.
- A measure of reliability for the statistical inference.

5. Convenience Sampling

we use the results that are readily available.

Ex) Someone could say to you, "Do you know...?"

Ch 2

- **Raw (Original) Data:** Data are in original form (Unorganized)

- **Class:** Each raw data value is placed into a quantitative or qualitative category.

- **Frequency Distribution**

The organization of raw data in table form, using classes and frequencies

- Categorical Frequency Distribution - Non numerical data
- Grouped Frequency Distribution - Numerical data
- Ungrouped Frequency Distribution

Rules for Constructing a Frequency Distribution

1. Classes' numbers should be between 5 and 20 classes.

2. The Class Midpoint

$$X_m = \frac{\text{Lower boundary} + \text{Upper boundary}}{2} \text{ or } \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

3. The classes must be mutually exclusive, but the class boundaries are not.

4. The classes must be continuous (No gap)

The only exception is if 1st or the last class starts with 'zero' frequency.

5. The classes must be equal in width.

The only exception that has an open-ended class.

(below, and more, etc.)

- **How to make the Table of Categorical Frequency Distribution**

1. **Class Limit:** Range = Highest value – Lowest value

2. **Class Limit:** The Number of classes desired (5 ~ 20 classes.)

*Ideal of number of classes by Sturges' guideline

$$= 1 + \log n / \log 2 \text{ where } n \text{ is the number of data values}$$

(Round up to the next whole number)

3. **Class Limit:** The Class Width = Range ÷ the number of classes

(Round up to the next whole number)

Class width = low class limit – previous low class limit (Vertical)

= upper class boundary – lower boundary (Horizontal)

(Subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class.)

4. **Class Limit:** Select the starting point for the lowest class limit.

5. **Class Limit:** Subtract one unit from the lower limit of the second class to get the upper limit of the 1st class.

Then add the width to each upper limit to get all the upper limits.

6. **Class boundaries:**

Lower Boundary = Lower Limit – 0.5 (or 0.05) depend on the

Upper boundary = Upper Limit – 0.5 (or 0.05) number of the data

Ex1)

Class Limits	Class boundaries
24 – 30	(24 – 0.5) – (30 + 0.5) 23.5 – 30.5
31 – 37	(31 – 0.5) – (37 + 0.5) 30.5 – 37.5

Ex2)

Class Limits	Class boundaries
2.3 – 2.9	(2.3 – 0.05) – (2.9 + 0.05) 2.25 – 2.95
3.0 – 3.6	(3.0 – 0.05) – (3.6 + 0.05) 2.95 – 3.65

7. **Tally & Frequency:** Count the number of data of each class

8. **Find the sum of all of Frequencies.**

9. **Cumulative Frequency:** adding the frequencies of the classes less than or equal to the upper class boundary of a specific class.

** The number the last class and the frequencies' sum must be same.

10. **Relative Frequency** = frequency \div total number = $\frac{f}{n}$

11. **Percent** = $\frac{f}{n} \times 100$ (%)

12. **Midpoint**

$$X_m = \frac{\text{Lower boundary} + \text{Upper boundary}}{2} \text{ or } \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

P38 Ex) Distribution of Blood types - Categorical F. Distribution

A B B AB O O O B AB B B B O
A O A O O O AB AB A O B A

Blood Type A: 5 people Blood Type B: 7 People

Blood Type O: 9 people Blood Type AB: 4 people Total: 25 people

Class	Tally	Frequency	Relative F.	Percent (%)
A		5	$\frac{5}{25} = \frac{1}{5}$	20%
B		7	$\frac{7}{25}$	28%
O		9	$\frac{9}{25}$	36%
AB		4	$\frac{4}{25}$	16%
Total		$\sum f = 25$	1	100%

Class	Cumulative Frequency
A	5
B	5+7 = 12
O	12+9 = 21
AB	21+4 = 25 (= $\sum f$)

P41 Ex 2-2) Record High Temperatures - Grouped F. Distribution

112 100 127 120 134 118 105 110 109 112 110 118 117 116 118
122 114 114 105 109 107 112 114 115 118 117 118 122 106 110
116 108 110 121 113 120 119 111 104
111 120 113 120 117 105 110 118 112 114 114

Solution)

1. Range = Highest value – Lowest value 134-100 = 34

2. The Number of classes desired that between 5 and 20 classes. 7 classes

3. The Class Width = Range \div the number of classes $34 \div 7 = 4.9 \rightarrow 5$
(Round up to the next whole number)

4. Select the starting point for the lowest class limit. 100

5. Subtract one unit from the lower limit of the second class to get the upper limit of the 1st class.

Then add the width to each upper limit to get all the upper limits.

100-104, 105-109, 110-114, 115-119, 120-124, 125-129, 130-134

6. Find boundaries.

Lower Boundary = Lower Limit – 0.5 (or 0.05) depend on the

Upper boundary = Upper Limit – 0.5 (or 0.05) number of the data

7. **Tally & Frequency:** Count the number of data of each class

8. **Find the sum of all of Frequencies.**

9. **Cumulative Frequency:** adding the frequencies of the classes less than or equal to the upper class boundary of a specific class.

** The number the last class and the frequencies' sum must be same.

10. **Relative Frequency** = frequency \div total number = $\frac{f}{n}$ each class

11. **Percent** = $\frac{f}{n} \times 100$ (%)

12. **Midpoint**

$$X_m = \frac{\text{Lower boundary} + \text{Upper boundary}}{2} \text{ or } \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

$$= \frac{99.5 + 104.5}{2} \text{ or } \frac{100 + 104}{2}$$

<i>C.L.</i>	<i>Class boundaries</i>	<i>Tally</i>	<i>f</i>	<i>m.d.</i>
100 -104	99.5 – 104.5		2	102
105-109	104.5 – 109.5		8	107
110-114	109.5 – 114.5	+	18	112
115-119	114.5 – 119.5	+	13	117
120-124	119.5 – 124.5		7	122
125-129	124.5 – 129.5		1	127
130-134	129.5 – 134.5		1	132

<i>C.L.</i>	<i>f</i>	<i>R. F.</i>	<i>%</i>	<i>C.F.</i>
100 -104	2	$\frac{2}{50} = \frac{1}{25}$	$\frac{1}{25} \cdot 100 = 4\%$	2
105-109	8	$\frac{8}{50} = \frac{4}{25}$	$\frac{4}{25} \cdot 100 = 16\%$	10
110-114	18	$\frac{18}{50} = \frac{9}{25}$	$\frac{9}{25} \cdot 100 = 36\%$	28
115-119	13	$\frac{13}{50}$	$\frac{13}{50} \cdot 100 = 26\%$	41
120-124	7	$\frac{7}{50}$	$\frac{7}{50} \cdot 100 = 14\%$	48
125-129	1	$\frac{1}{50}$	$\frac{1}{50} \cdot 100 = 2\%$	49
130-134	1	$\frac{1}{50}$	$\frac{1}{50} \cdot 100 = 2\%$	50
Total	$\sum f = 50$	1	100%	

C.L. = class limits

f = frequency

c.f. = cumulative frequency

R.F.= relative frequency

<Cumulative Frequency Distribution>

<i>Class</i>	<i>f</i>	<i>C. F.</i>
Less than 105	2	2
Less than 110	8	2+8=10
Less than 115	18	10+18=28
Less than 120	13	28+13=41
Less than 125	7	41+7=48
Less than 130	1	48+1=49
Less than 135	1	49+1=50
	$\sum f = 50$	same as $\sum f$

- **Histogram**

the data by using continuous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes

- **Frequency Polygon**

the data by using lines that connect points plotted for frequencies for the classes.
(starts from zero)

The frequencies are represented by the height of the points.

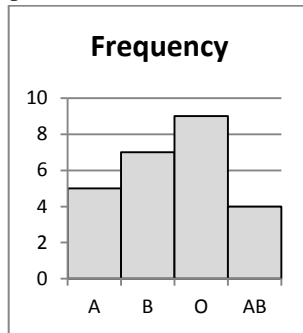
- **Ogive (=Cumulative frequency)**

the cumulative frequencies for the classes in a frequency distribution

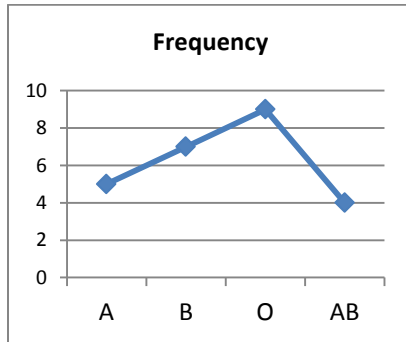
- *****Note:** Those three graphs are used when the data are contained in a grouped frequency distribution

Graphs from the Ungrouped Frequency Distribution of Blood types

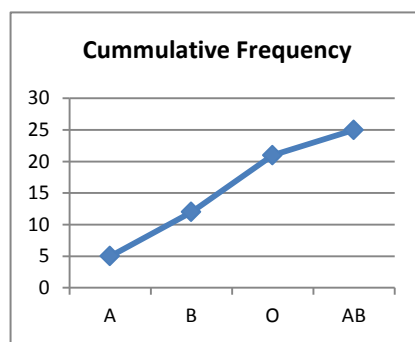
<Vertical Bar Graph >



< Frequency Graph >

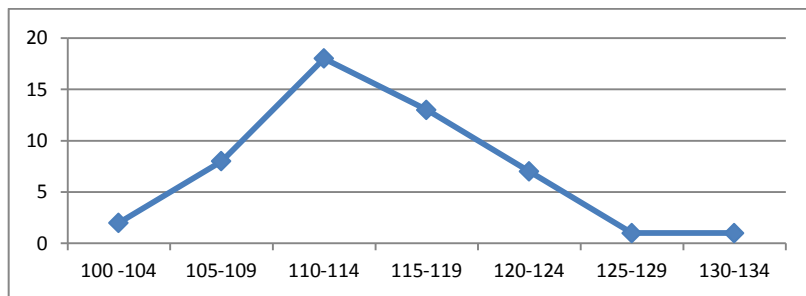


< Cumulative Frequency Graph >



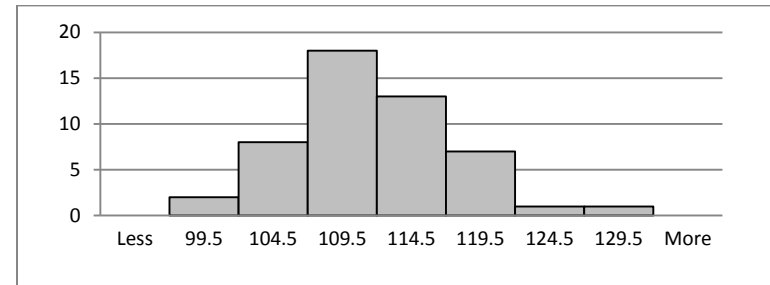
Ex) Drawing Graphs of Grouped Frequency distribution from Ex 2-2)

Frequency Distribution



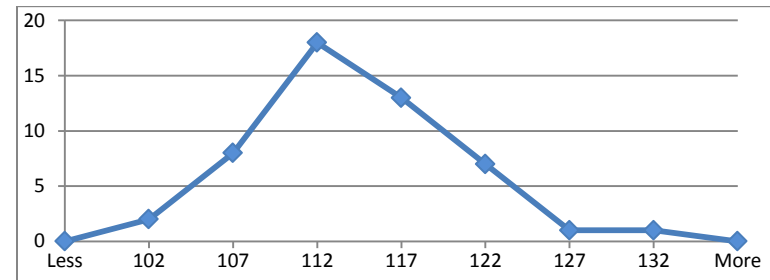
Histogram

*** Using **Class boundaries** for x – axis and Frequencies for y –axis***



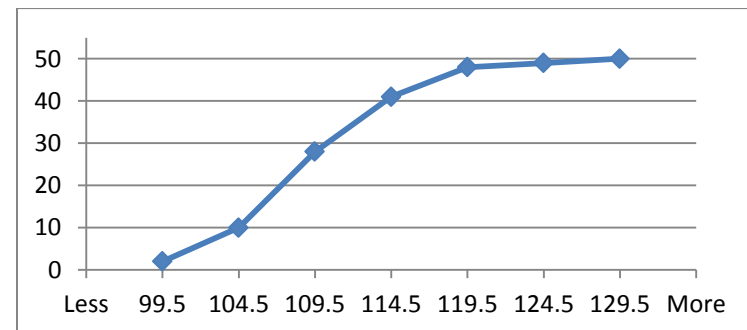
Frequency Polygon

*** Using **Midpoints** for x – axis and Frequencies for y –axis***

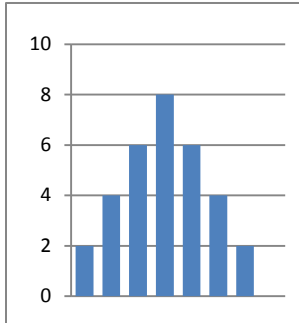
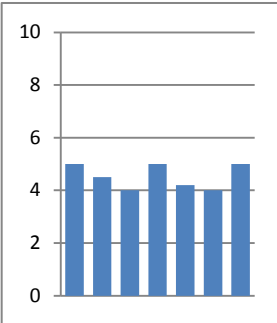
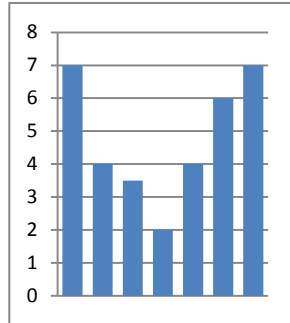
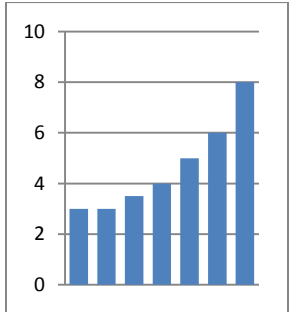
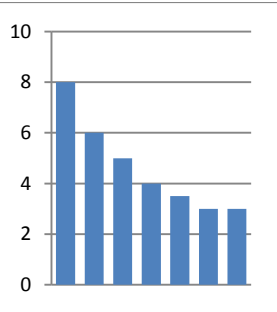
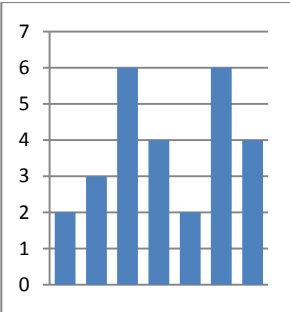
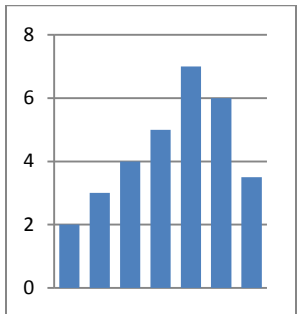
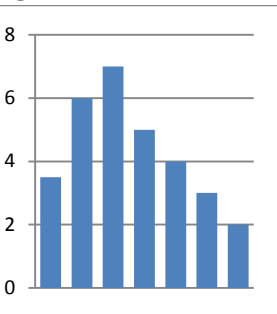


Ogive

*** Using **Class boundaries** for x – axis and
Cumulative Frequencies for y –axis **



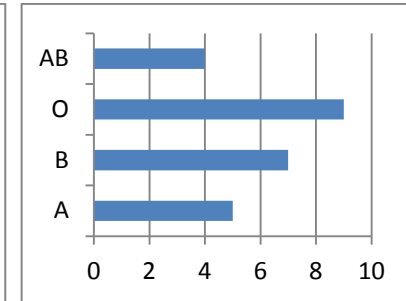
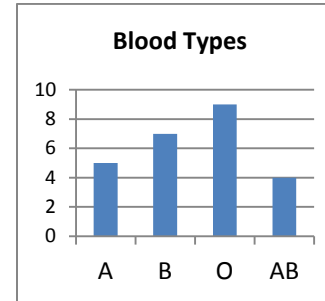
< Distribution Shapes >

Bell shape**Uniform****U shape****J shape****Reverse J shape****Bimodal****Left Skewed****Right Skewed**

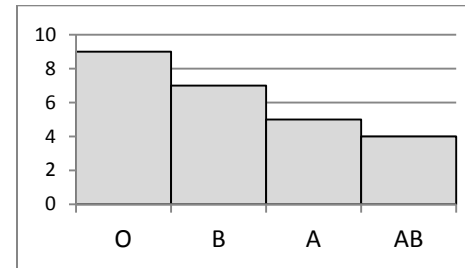
<Other Types of Graphs>

Bar Graph

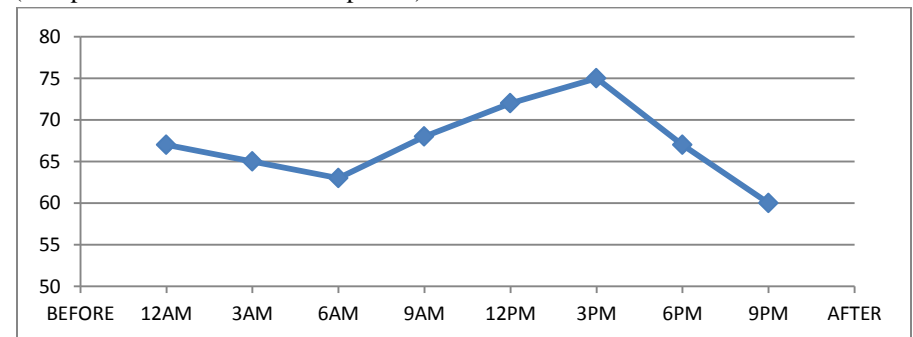
: using vertical or horizontal bars whose heights or lengths represent the frequencies of the data

**Pareto Chart (Horizontal)**

: a Categorical variable and the frequencies are displayed by the heights of vertical bars which are arranged in order from highest to lowest

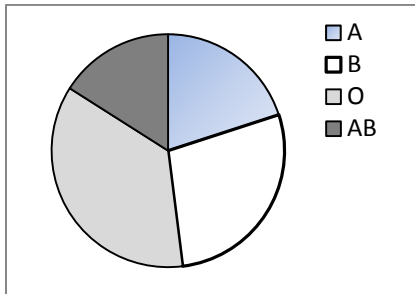
**Time Series Graph** : occur over a specific period of time

(Temperatures over a 24 hours period)



Pie Graph (Percentage or proportions-Nominal or Categorical)

: Divided into sections or wedges according to the percentage of frequencies in each category of the distribution in a circle



Step 1) Degrees = $f/n \cdot 360^\circ$ $\Sigma \text{degrees} = 360^\circ$: to measure

Step 2) % = $f/n \cdot 100\%$ $\Sigma \text{percentages} = 100\%$: to show

- The sum of degrees or percentages does not always sum of 360° or 100% due to rounding

<Scatter Plots>

A graph of order pairs of data values that is used to determine if a relationship exists between the two values \rightarrow a set order pairs (x, y)

x Independent variable

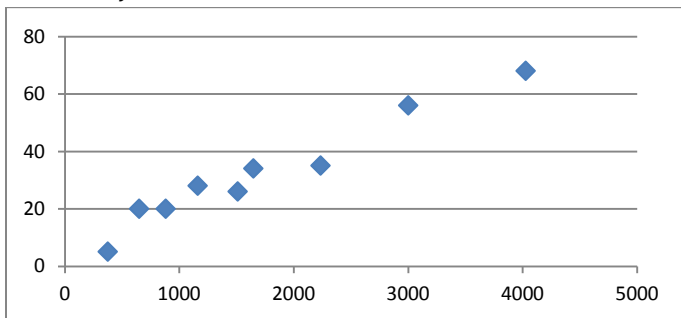
y Dependent variable

Ex)

No. of

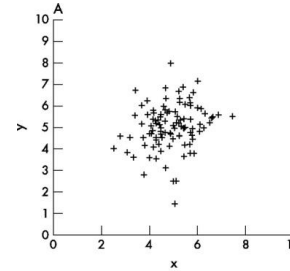
Accidents, x 376 650 884 1162 1513 1650 2236 3002 4028

Fatalities, y 5 20 20 28 26 34 35 56 68

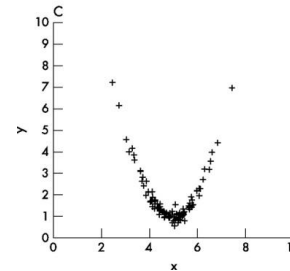
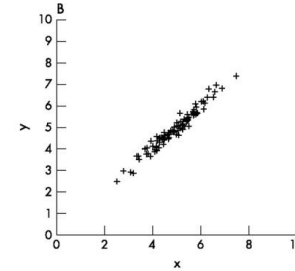


<Analyzing the scatter plot>

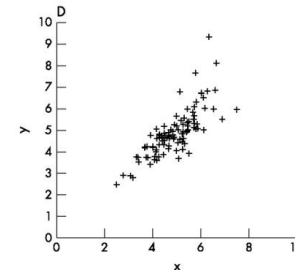
No relationship



Positive Linear relationship



No Linear relationship



Positive Linear relationship

Negative Linear relationship



< Stem and Leaf Plot (Exploratory Data Analysis)>

A data plot that uses part of the data value as the stem as the stem and part of the data value as the leaf to form groups or classes

Step 1) Arrange the data in order

Step 2) Separate the data according to the first digit

Step 3) A display can be made by using the leading digit as the stem and the trailing digit as the leaf.

** If there are no data values in a class, you should write the stem number and leave the leaf row blank. Do not put a zero in the leaf row.

Ex) 24 32 2 56 44 2 13 32 44 31 32 14 105 23 20

Step 1) 2 13 14 20 23 24 31 32 32 32 44 44 56 105

Step 2) 02 31 32 32 32 105
 13 14 44 44
 20 23 24 56

Step3)	Stem (Leading Digit)	Leaf (Trailing Digit)
	0	2
	1	3 4
	2	0 3 4
	3	1 2 2 2
	4	4 4
	5	6
	10	5

Ex) Atlanta: 26 29 30 31 36 36 40 40 50 52 60
 N.Y. : 25 31 31 32 36 39 40 43 51 52 56

Step 1) It's arranged already.

Atlanta: 26 29 30 31 36 36 40 40 50 52 60

N.Y. : 25 31 31 32 36 39 40 43 51 52 56

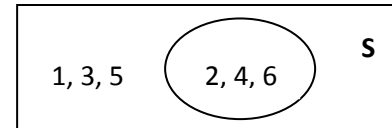
Step 3)

Atlanta	Stem	N.Y.
9 6	2	5
6 6 1 0	3	1 1 2 6 9
0 0	4	0 3
2 0	5	1 2 6
0	6	

Ch 3

3 – 1 Measures of Central Tendency

- **Statistics** : a characteristic or measurer obtains by using the data values from a sample
- **Parameter**: a characteristic or measurer obtains by using all the data values from a specific population



Ex) Statistics → 2, 4, 6 (a sample)

Parameter → 1, 2, 3, 4, 5, 6 (population)

- **Mean** (=Arithmetic Average) Affected by the highest and lowest values

$$\text{Sample Mean} : \bar{X} = \frac{\sum X}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$\text{Population Mean} : \mu (\text{mu}) = \frac{\sum X}{N} = \frac{X_1 + X_2 + \dots + X_N}{N}$$

** n = the numbers of the sample

** N = the numbers of the population

Ex) 2 6 9 10 5 7

$$\mu (\text{mu}) = \frac{\sum X}{N} = \frac{2 + 5 + 6 + 7 + 9 + 10}{6} = \frac{39}{6} = 6.5$$

- **Median (MD)** : the midpoint of the data array

- 1) Arrange the all the data in order
- 2) Select the midpoint
- 3) If there are 2 numbers of MD, adding the 2 numbers
And then divide by 2.

** Data array = the data set is ordered

Ex) 3 5 4 9 2 3 4 6 10

2 3 3 4 4 5 6 9 10 → 4 is MD

Ex) 20 41 66 27 21 24

20 21 24 27 41 66 → 24 & 27 are in the middle

$$\frac{24 + 27}{2} = 25.5 \rightarrow 25.5 \text{ is } \mathbf{MD}$$

• **Mode** the value that occurs most often in a data set

- No mode: all different data ex) 2 3 5 9 7 12 1 4
- Unimodal: one mode ex) 2 3 5 9 7 5 2 4
- Bimodal: two mode ex) 2 3 5 9 3 7 2 11
- Multimodal: more than two mode ex) 2 3 9 2 7 3 4 9

• **Midrange (MR)**: approximate of data values

$$\frac{(\text{Lowest Value} + \text{Highest Value})}{2} = \frac{\text{Range}}{2}$$

• **Weighted Mean**: (ex) GPA

$$\bar{X} = \frac{W_1X_1 + W_2X_2 + \dots + W_nX_n}{W_1 + W_2 + \dots + W_n} = \frac{\sum WX}{\sum W} \quad W_n: \text{Weight}, X_n: \text{Value}$$

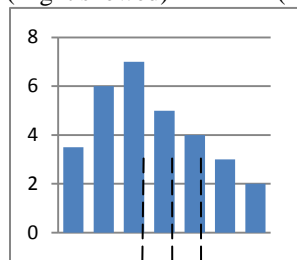
Ex)	Course	Credits (W)	Grade (X)
	Math	3	A (4points)
	English	4	C (2points)
	Biology	2	B (3points)

$$\bar{X} = \frac{WX}{W} = \frac{(3 \cdot 4) + (4 \cdot 2) + (2 \cdot 3)}{3 + 4 + 2} = \frac{26}{9} \approx 2.9$$

• **Distribution Shapes**

Positively skewed

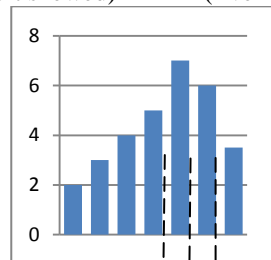
(Right skewed)



Mode | Mean
Median

Negatively skewed

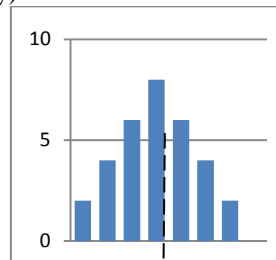
(Left skewed)



Mean | Mode
Median

Bell shape Symmetric

(Evenly)



Mean
Median
Mode

3 – 2 Measures of Variation

• Population Variance and Standard Deviation

: to have a more meaningful statistic to measure the variability, using variance and standard deviation

: When the means of 2 sets of data are equal, the larger the variance or standard deviation is more variable the data are.

– **Range (R)** = Highest value – Lowest value

**Distance between highest and lowest values

$$\text{Variance} = \sigma^2 = \frac{\sum(X - \mu)^2}{N} = \frac{\sum(X - \frac{\sum X}{N})^2}{N}$$

**average of the squares of the distance that each value

$$\text{Standard Deviation} = \sqrt{\sigma^2} = \sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} \quad (\sigma \geq 0)$$

** μ (mu): Population Mean

** X = Individual value

** N = Population size

** σ = sigma Greek lowercase

• Sample Variance and Standard Deviation

Not usually used, but since in most cases the purpose of calculating the statistics is to estimate the corresponding parameter

$n - 1$: because giving a slightly larger value and an unbiased estimate of the population variance ($n \leq 30$)

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

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

** \bar{X} = Sample mean n = Sample size

***Short cut or Computation Formulas (No need \bar{X})

$$s^2 = \frac{(\sum X^2) - \frac{(\sum X)^2}{n}}{n - 1} \quad s = \sqrt{\frac{n(\sum X^2) - (\sum X)^2}{n - 1}} \quad ** (\sum X^2) \neq (\sum X)^2$$

Ex) 131p Find population variance and population standard deviation

Comparison of outdoor paint (how long each will last before fading)

A	10	60	50	30	40	20
B	35	45	30	35	40	25

Step1) $\sum X = A = B = 210$, $N = A = B = 6$

Step2) $\mu = \frac{\sum X}{N} = \frac{210}{6} = 35 \text{ months}$ ($A = B$) *The Means of A&B are equal.*

Step3) Range A: $60 - 10 = 50$ months

B: $45 - 25 = 20$ months

Step4) Variance

A: $10-35=35$, $60-35=25$, $50-35=15$, $30-35=-5$, $40-35=5$, $20-35=-15$

B: $35-35=0$, $45-35=10$, $30-35=-5$, $35-35=0$, $40-35=5$, $25-35=-10$

A: $(-25)^2 = 625$, $(25)^2 = 625$, $(15)^2 = 225$, $(-5)^2 = 25$, $5^2 = 25$, $(-15)^2 = 225$

B: $0^2 = 0$, $(10)^2 = 100$, $(-5)^2 = 25$, $0^2 = 0$, $(5)^2 = 25$, $(-10)^2 = 100$

A: $\sigma^2 = \frac{\sum(X - \mu)^2}{N} = \frac{625 + 625 + 225 + 25 + 25 + 225}{6} = \frac{1750}{6} = 291.7$

B: $\sigma^2 = \frac{\sum(X - \mu)^2}{N} = \frac{0 + 100 + 25 + 0 + 25 + 100}{6} = \frac{250}{6} = 41.7$

Step5) Standard deviation A: $\sigma = \sqrt{291.7} \approx 17.1 \rightarrow \text{more variable}$

B: $\sigma = \sqrt{41.7} \approx 6.5$

• **For Variance and Standard Deviation for Grouped Data**

- Using it uses the midpoints of each class

Ex) Class	Frequency(f)	Midpoint (X_m)	$f \cdot x_m$	$f \cdot X_m^2$
05.5 - 10.5	1	8		
10.5 - 15.5	2	13		
15.5 - 20.5	3	18		
20.5 - 25.5	5	23		
25.5 - 30.5	4	28		
30.5 - 35.5	3	33		
35.5 - 40.5	2	38		

Step 1) Find the mid points of each class.

Step 2) $f \cdot x_m \rightarrow 1 \cdot 8 = 8$, $2 \cdot 13 = 26$, ..., $2 \cdot 38 = 76$

Step 3) $f \cdot X_m^2 \rightarrow 1 \cdot 8^2 = 64$, $2 \cdot 13^2 = 338$... $2 \cdot 38^2 = 2888$

Class	f	X_m	$f \cdot X_m$	$f \cdot X_m^2$
05.5 - 10.5	1	8	8	64
10.5 - 15.5	2	13	26	338
15.5 - 20.5	3	18	54	972
20.5 - 25.5	5	23	115	2645
25.5 - 30.5	4	28	112	3136
30.5 - 35.5	3	33	99	3267
35.5 - 40.5	2	38	76	2888
Sum (\sum)	$n = 20$		$\sum f \cdot X_m = 490$	$\sum f \cdot X_m^2 = 13310$

Step 4) $s^2 = \frac{(\sum X^2) - \frac{(\sum X)^2}{n}}{n - 1} = \frac{13310 - \frac{490^2}{20}}{20 - 1} = 68.7$

Step5) $s = \sqrt{68.7} = 8.3$

<Uses of the Variance and Standard Deviation>

1. Variance and Standard Deviation can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two(or more) data sets to determine which is more(most) variable.
2. The measure of variance and standard deviation are used to determine the consistency of a variable. for example, in the manufacture of fitting, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.
3. The variance or standard deviation are used determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's Theorem shows that, for an distribution, at least 75% of the data values will fall within 2 standard deviations of the mean.
4. finally, the variance or standard deviation are used quite often in inferential statistics. These uses will be shown in later chapters.

• Coefficient of Variation with percentage(%)

For populations: $CVar = \frac{\sigma}{\mu} \cdot 100\%$

For Samples : $CVar = \frac{s}{\bar{X}} \cdot 100\%$

*** To compare standard deviations when the units are different the larger coefficient of variance is more variable than the other.

Ex 3-25) p140 The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is \$5225, and the standard deviation is \$773.

Compare the variations of the two.

-Solution

Sales	$CVar = \frac{s}{\bar{X}} = \frac{5}{87} \cdot 100\% = 5.7\%$
Commissions	$CVar = \frac{s}{\bar{X}} = \frac{773}{5225} \cdot 100\% = 14.8\%$

Since the coefficient of variation is larger for commissions, the commissions are more variable than sales.

• Range rule of Thumb

$$S = \frac{\text{range}}{4} \quad ** \text{ a rough estimate of the standard deviation}$$

$$- \text{Largest Data Value} = \bar{X} + 2S$$

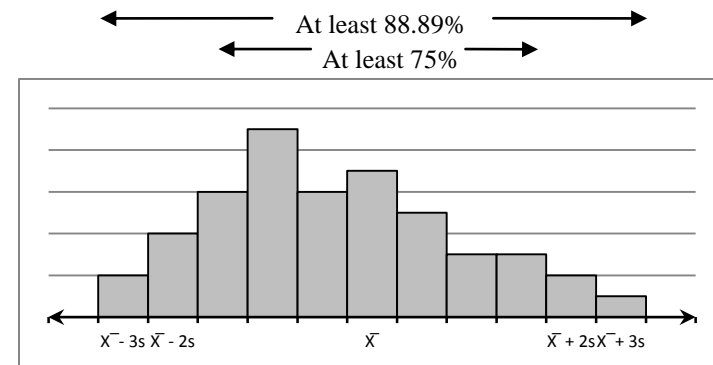
$$- \text{Smallest Data Value} = \bar{X} - 2S$$

• Chebyshev's Theorem

1. The proportion of values from a data set that will fall within k standard deviation of the mean will be at least $1 - \frac{1}{k^2}$, where k is a number greater than 1 (k isn't necessarily an integer).

$$k = \frac{\text{the larger value} - \text{the mean}}{\text{the standard deviation}}$$

2. Find the minimum % of data values that will fall between any two given values.
3. This states at least 75% of the data values will fall within 2 standard deviations of the mean of the data set.



ex) The mean price of houses in a certain neighborhood is \$50,000,

and the standard deviation is \$10,000. Find the price range for which at least 75%, of the houses will sell.

-Solution $450,000 + 2(\$10,000) = \$50,000 + \$20,000 = \$70,000$

$$450,000 - 2(\$10,000) = \$50,000 - \$20,000 = \$30,000$$

Hence, at least 75% of all homes sold in the area will have a price range from \$30,000 to \$70,000.

- **Standard Scores or z score (z)**

- a comparison of a relative standard similar to both can be made the mean and standard deviations
- Number of standard deviations a data value is above or below the mean for a specific distribution of values

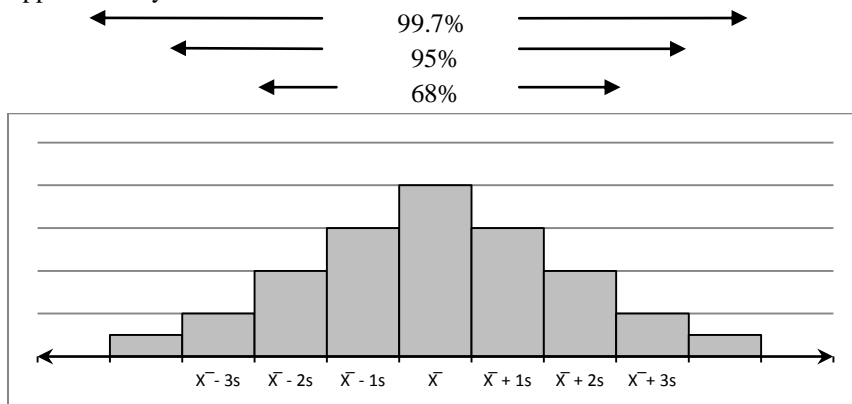
$$z = \frac{\text{Value} - \text{Mean}}{\text{Standard Deviation}} = \frac{X - \bar{X}}{s} \text{ (sample)} = \frac{X - \mu}{\sigma} \text{ (population)}$$

- **The Empirical (normal) rule** in the bell- shaped of graph

Approximately 68% of the data value will fall 1 standard deviation of mean

Approximately 95% of the data value will fall 2 standard deviation of mean

Approximately 99.7% of the data value will fall 3 standard deviation of mean



3-3 Measures of Position

- **Percentiles (P_n)**

- Divide the data set into 100 equal groups
- Position in hundredths that a data value holds in the distribution

$P_1, P_2, \dots, P_{98}, P_{99} = 100 \text{ parts}$ (each part = 1%)

$$\text{Percentile} = \frac{(\text{number of value below } X) + 0.5}{\text{Total number of values}} \cdot 100\%$$

***To find the approximate percentile rank of the data value

Ex 3-32) A teacher gives a 20 point test to 10 students.

The scores are shown here. Find the percentile rank of a score of 12.

18 15 12 6 8 2 3 5 20 10

Step 1) Arrange the data 2 3 5 6 8 10 12 15 18 20

Step 2) number of value below $X = 6$

$$P = \frac{(\text{number of value below } X) + 0.5}{\text{Total number of values}} \cdot 100\% = \frac{6 + 0.5}{10} \cdot 100\% = 65\text{th percentile}$$

Step 3) a student whose score was 12 did better than 65% of the class.

- **Finding a Value corresponding to a Given Percentile**

$$c = \frac{n \cdot p}{100} \quad \text{where } n = \text{total number of values} \quad p = \text{percentile}$$

If c^{th} is not a whole number, round it up to the next whole number.

If c^{th} is a whole number $\frac{c+(c+1)}{2}$ $(c+1)^{\text{th}}$ is the next value number of c .

Ex 3-34) from 3- 32 find the value corresponding to the 25th percentile.

Step 1) total number = 10, percentile = 25th

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 25}{100} = 2.5 \text{ (not a whole number)} \rightarrow \text{round up } c = 3$$

Step 2) The 3rd value is 5.

Hence, the value 5 corresponds to 25th percentile.

Ex 3-35) from 3- 32 find the value corresponding to the 60th percentile.

Step 1) total number = $n = 10$, percentile = $p = 60$ th

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 60}{100} = 6 \text{ (a whole number)}$$

Step 2) The 6th value(=c) is 10 and 12 is 7th value(=c+1).

$$\frac{10 + 12}{2} = 11$$

Hence, 11 corresponds to the 60th percentile.

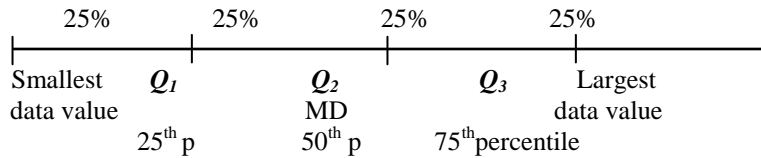
Anyone scoring 11 would have done better than 60% of the class.

• Quartiles (Q_n)

: Position in fourths that a data value holds in the distribution

Step 1) Arrange the data in order from lowest to highest

Step 2) Divide into 4 groups



Ex 3-36) 15 13 6 5 12 50 22 18 Find Q_1 , Q_2 , & Q_3

Step 1) Arrange the data set $\rightarrow 5 \ 6 \ 12 \ 13 \ 15 \ 18 \ 22 \ 50$

Step 2) To find $Q_2 \rightarrow$ divide into 2 $\frac{13 + 15}{2} = 14 = \text{MD} = Q_2$

Step 3) $Q_1 = \frac{6 + 12}{2} = 9$

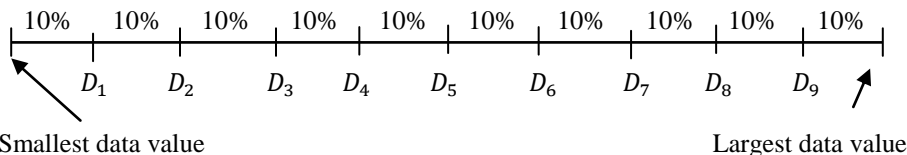
Step 4) $Q_3 = \frac{18 + 22}{2} = 20$

• Deciles (D_n)

Position in tenths that a data value holds in the distribution

Step 1) Arrange the data in order from lowest to highest

Step 2) Divide into 10 groups



• Outliers

- An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.

- Strongly affect with the mean and standard deviation

Step 1) Arrange the data in order and find Q_1 and Q_3 .

Step 2) Find the interquartile range = **IQR** = $Q_3 - Q_1$

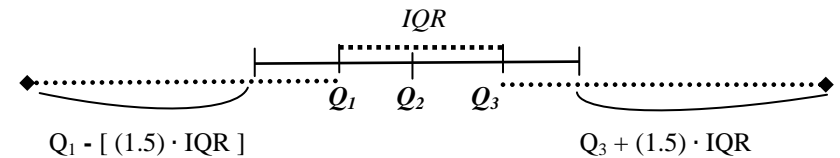
Step 3) $(1.5) \cdot \text{IQR}$

Step 4) $Q_1 - [(1.5) \cdot \text{IQR}]$

$$Q_3 + [(1.5) \cdot \text{IQR}]$$

Step 5) Check the data set for any value that is smaller than

$$Q_1 - [(1.5) \cdot \text{IQR}] \text{ or larger than } Q_3 + [(1.5) \cdot \text{IQR}]$$



Ex 3 - 37) Set for outliers from ex3-36)

Step 1) Arrange the data set $\rightarrow 5 \ 6 \ 12 \ 13 \ 15 \ 18 \ 22 \ 50$

Step 2) To find $Q_2 \rightarrow$ divide into 2 $\frac{13 + 15}{2} = 14 = \text{MD} = Q_2$

$$Q_1 = \frac{6 + 12}{2} = 9 \quad Q_3 = \frac{18 + 22}{2} = 20 \quad Q_3 - Q_1 = 20 - 9 = 11$$

Step 3) $(1.5) \cdot \text{IQR} = 1.5 \cdot 11 = 16.5$

Step 4) $Q_1 - [(1.5) \cdot \text{IQR}] = 9 - 16.5 = -7.5$

$$Q_3 + [(1.5) \cdot \text{IQR}] = 20 + 16.5 = 36.5$$

Step 5) Check the data set for any data values that fall outside the interval from -7.5 to 36.5.

The value 50 is outside this interval.

Hence, it can be considered an outlier.

3 - 5 Exploratory Data Analysis

• The Five - Number summary and Boxplots

1. The 5-Number Summary

- 1) The lowest value of the data set (Minimum)
- 2) Q_1
- 3) Q_2 = The Median
- 4) Q_3
- 5) The highest value of the data set (Maximum)

2. a Boxplot

A graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1 , a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2 .

3. How to make a Boxplot

- Step 1) Arrange the data in order.
- Step 2) Find Q_2 (The Median).
- Step 3) Find Q_1 & Q_3 .
- Step 4) Draw a scale for the data on the x - axis.
- Step 5) Locate the lowest value, Q_1 , the median, Q_3 , and the highest value on the scale.
- Step 6) Draw a box around Q_1 & Q_3 , draw a vertical line through the median, and connect the upper and lower values.

4. Information Obtained from a Boxplot

- 1) If the median is near the center of the box, the distribution is approximately symmetric.
- 2) If the median falls to the left of the center of the box, the distribution is positively skewed.
- 3) If the median falls to the right of the center, the distribution is negatively skewed.
- 4) If the lines are about the same length, the distribution is approximately symmetric.
- 5) If the right line is larger than the left line, the distribution is positively skewed.
- 6) If the right line is larger than the left line, the distribution is negatively skewed.

Ex 3-39) A dietitian is interested in comparing the sodium content of real cheese with the sodium content of a cheese substitute. The data for two random samples are shown. Compare the distributions, using boxplots.

Real Cheese				Cheese Substitute			
310	420	45	40	270	180	250	290
220	240	180	90	260	340	310	

Step 1) Real cheese : 40 45 90 180 220 240 310 420
 Cheese Substitute : 130 180 250 260 270 290 310 340

Step 2) Q_2 (The Median)

$$\text{Real cheese} : \frac{180+220}{2} = 200 = Q_2$$

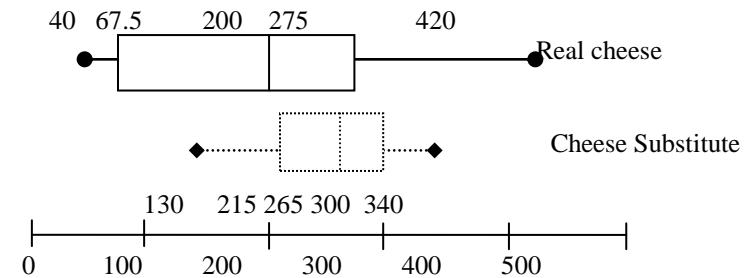
$$\text{Cheese Substitute} : \frac{260+270}{2} = 265 = Q_2$$

Step 3) Q_1 & Q_3

$$\text{Real cheese} : \frac{45+90}{2} = 67.5 = Q_1 \quad \frac{240+310}{2} = 275 = Q_3$$

$$\text{Cheese Substitute} : \frac{180+250}{2} = 215 = Q_1 \quad \frac{240+310}{2} = 275 = Q_3$$

Step 4, 5, & 6)



** Compare the plots. It is quite apparent that the distribution for the cheese substitute data has a higher median than the median for the distribution for the real cheese data. The variation or spread for the distribution of the real cheese data is larger than the variation for the distribution of the cheese substitute data.

Traditional	Exploratory Data Analysis
Frequency distribution	Stem and Leaf Plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile range

The most three commonly used measures of central tendency are mean, median, and mode.

The most three commonly used measurements of variation are range, variance, and standard deviation.

The most common measures of position are percentiles, quartiles, and deciles.

The coefficient of variation is used to describe the standard deviation in relationship to the mean.

These methods are commonly called traditional statistical methods and are primarily used to confirm various conjectures about the nature of the data.

The boxplot and 5-number summaries are part of exploratory data analysis; to examine data to see what they reveal.

CH 4

4-1 Sample Space and counting Rules

- **Probability** - The chance of an Event occurring

1. Probability Experiments

A chance process that lead to well-fined results called outcomes. (not known in advance of an act)

2. Outcome; The result of a single trial of a probability experiment

3. Event (= E)

a subject(a sample from total) of the given sample space denoted by A, B, C, D, etc. (it can consist more than one outcomes.)

Ex 1) A question has multiple choices that 4 possible results (Outcomes) such as Ⓐ Ⓑ Ⓒ and Ⓓ.

Only one of them is the right answer.

What is a chance that a person gets the answer?

$$\frac{\text{the right answer}}{\text{total}(\text{Ⓐ Ⓑ Ⓒ and Ⓓ})} = \frac{1}{4} \quad \frac{1}{4} \times 100 = 25\% \text{ chance}$$

Ex 2) Tossing a fair and balance coin.

(Well- defined, outcomes Head & Tail)

What is the possibility (of chance) of getting "Head" ?

2 possible outcomes (Head & Tail = H & T)

$$\frac{\text{Head}}{\text{total (Head \& Tail)}} = \frac{1}{2} \quad \frac{1}{2} \times 100 = 50\% \text{ chance}$$

* Fair- each side(face) if equally likely

* Balance- it should fall on either side (Head and Tail)

Ex 3) Rolling a die (a six-faced cube from 1 to 6),
what is the probability of getting 4?

$$\frac{\text{the number 4}}{\text{total (from 1 to 6)}} = \frac{1}{6} \quad \frac{1}{6} \times 100 \approx 16.67\% \text{ chance}$$

4. Sample Space (= S)

the set (or collection) of all possible outcomes of a probability experiment

* A die is rolled $S = \{1, 2, 3, 4, 5, 6\}$ (=a set of notation)

* A coin is tossed $S = \{H, T\}$

Ex 4) A die is rolled $S = \{1, 2, 3, 4, 5, 6\}$ Let $E = \{2, 4, 6\}$

Observing an event number

$$\frac{3 \text{ (2,4,6)}}{\text{total (1,2,3,4,5,6)}} = \frac{3}{6} = \frac{1}{2} \times 100 = 50\% \text{ chance}$$

• **Sample Space of Rolling 2 Dice**

		Second throw					
		1	2	3	4	5	6
First throw	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Experiment	Sample Space
Toss a coin	Head, Tail
Toss 2 coins	H-H, H-T, T-T, T-H
Roll a die	1, 2, 3, 4, 5, 6
Roll 2 dice	1-2, 1-2 1-3, 1-4, 1-5, 1-6,, 2-1, 2-2, 2-3, etc 36 outcomes.

• **Playing Cards in a deck**

Diamonds (Red); 13 Cards	A 2 3 4 5 6 7 8 9 10 J Q K ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦
Spades (Black) ; 13 Cards	A 2 3 4 5 6 7 8 9 10 J Q K ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠
Clubs (Black) ; 13 Cards	A 2 3 4 5 6 7 8 9 10 J Q K ♣ ♣ ♣ ♣ ♣ ♣ ♣ ♣ ♣ ♣ ♣ ♣ ♣
Hearts (Red) ; 13 Cards	A 2 3 4 5 6 7 8 9 10 J Q K ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥

Total A deck of 52 Cards = 26 of Red Cards + 26 Black Cards

Face or picture cards = 12 = 4 Jacks(J) + 4 Queens(Q) + 4 Kings(K)

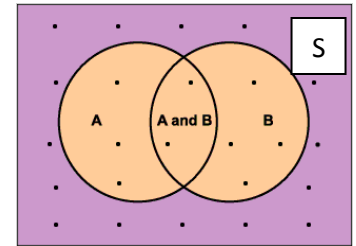
• **Venn Diagram**

A diagram used as a pictorial representative for a probability concept or rule

S = Sample Space

= all the possible outcomes

Event A , Event B



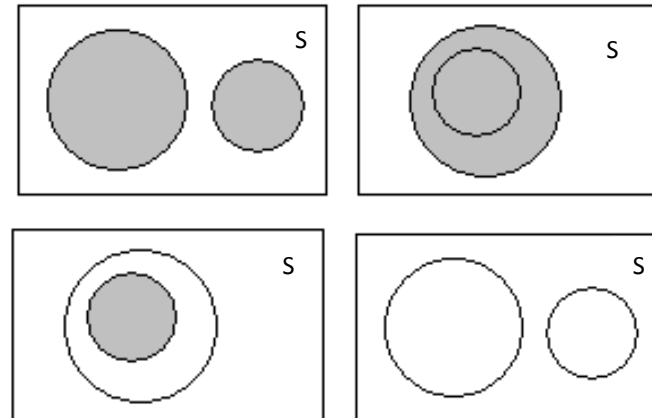
You can represent the Probability of the Events using a Venn diagram from set theory. (can't use this method with all cases)

The rectangle is Sample Space (S).

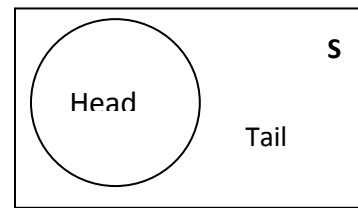
The circle (set) of A or B is the event, and they are dependent of each other.

The intersection area of events A and B is a nice correspondence between "events A and B both occurring" and "being inside both circle A and circle B".

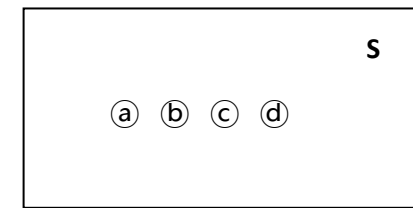
The union area of event A or B is covered the maximum combined area of A and B, when they do not overlap and it's the maximum possible area of A-union-B.



From Ex2) Tossing a coin

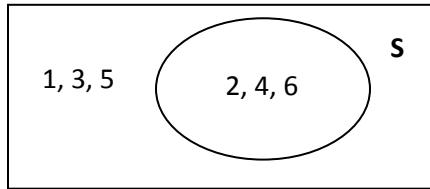


From Ex1) (a)(b)(c) and (d)



From Ex 4) Let $E = \{2, 4, 6\}$ (Observing an event number)

$$\frac{3(2,4,6)}{\text{total } (1,2,3,4,5,6)} = \frac{3}{6} = \frac{1}{2} \times 100 = 50\% \text{ chance}$$

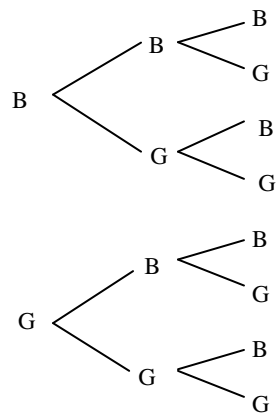


- **Tree Diagram;** the method of constructing a sample space

P193 [Ex 4 - 4] Gender of Children

- a) Find the probability of all possibility outcomes that a married couple has 3 children. (Girls and boys)

1st Child 2nd Child 3rd Child



BBB	BGB	S
BBG	BGG	
GBB	GBG	
GGB	GGG	

$n(S) = 8$ outcomes

$$n(S) = 2 \cdot 2 \cdot 2 = (1\text{st } B \& G) \cdot (2\text{nd } B \& G) \cdot (3\text{rd } B \& G) = 2^3$$

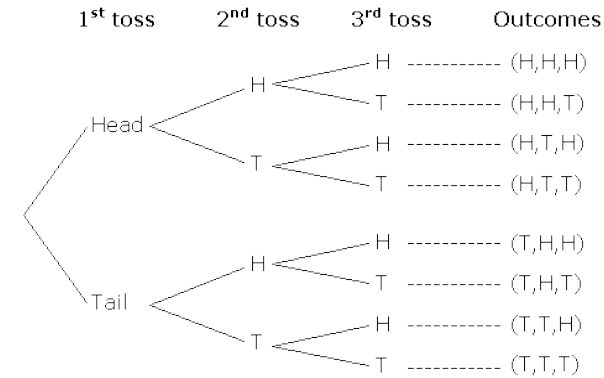
- b) Find the probability of all children are boys

$$\frac{1(BBB)}{8 \text{ outcomes}} = \frac{1(\text{none of } G)}{8} = 12.5\% \text{ chance (unlikely)}$$

Ex 5) A coin is tossed 100 times, find the $n(S)$

$$S = \{H, H, H, \dots, T, T, T, \dots\}$$

$$n(S) = (H \& T)^{\text{times}} = 2^{100}$$



Ex) A coin is tossed only one time = $2^1 = 2$

A coin is tossed 3 times = $2^3 = 8$

$$\begin{array}{cccccc} 1^{\text{st}} \text{ time} & 2^{\text{nd}} \text{ time} & 3^{\text{rd}} \text{ time} & 4^{\text{th}} \text{ time} & 5^{\text{th}} \text{ time} & \text{etc.} \\ 2^1 & & 2^2 = 4 & 2^3 = 8 & 2^4 = 16 & 2^5 = 32 \end{array}$$

Ex 6) A coin is tossed 10 times, find $P(\text{all are Heads})$

$$\frac{1}{2^{10}} = \frac{1}{1024}$$

Ex 7) A die is rolled,

1. Find Odds in favor of getting of less than 4.

$$\frac{\#F}{\#A} = \frac{3}{3} = \frac{1}{1} = 1 : 1 \text{ (it isn't "1".)} \quad 3 + 3 = 6 \text{ numbers}$$

2. Find Odds in favor of getting of less than 5.

$$\frac{\#F}{\#A} = \frac{4}{2} = 2 : 1 \quad 4 + 2 = 6 \text{ numbers}$$

3. Find Odds in against of getting of less than 5

$$\frac{\#A}{\#B} = \frac{2}{4} = 1 : 2$$

• When a coin is tossed N times

Proceeding in the same number if a coin is tossed N times

$$\text{then, } n(S) = 2^N$$

• Odds

The **Actual Odds Against event A** occurring are the ratio

$\frac{P(\bar{A})}{P(A)}$, usually expresses in the form of $a:b$ (or "a to b"),

where a and b are integers having no common factors.

The **Actual odds in favor of event A** are the reciprocal of the actual odds against that event. If the odds against A are $a:b$, then the odds in favor of A are $b:a$.

The **Payoff Odds against event A** represent the ratio of net profit (if you win) to the amount bet.

Payoff Odds Against Event A = (Net profit) : (Amount bet)

Favor	Against	S
#F	#A	

#T= number of Total

☉ **Number of Total = Number of F + Number of A = $n(S)$**

$$\#A = \#T - \#F$$

$$\#F = \#T - \#A$$

$$\text{Odds (favor)} = \frac{\#F}{\#A} = \#F : \#A \quad \text{Odds(Against)} = \frac{\#A}{\#B} = \#A : \#F$$

$$\text{Prob(in favor)} = \frac{\#F}{\#Total} \quad \text{Prob(in against)} = \frac{\#A}{\#Total}$$

Ex 8) A card is drawn from a deck (4+48=52)

$$\text{Odd(in favor of Queen)} = \frac{\#F}{\#A} = \frac{4 \text{ Queens}}{52 - 4} = \frac{4}{48} = \frac{1}{12} = 1 : 12$$

$$\text{Prob(in favor of Queen)} = \frac{4}{52} = \frac{1}{13}$$

Three Basic Interpretations of Probability

1. Classical probability
2. Empirical or Relative Frequency Probability
3. Subjective Probability

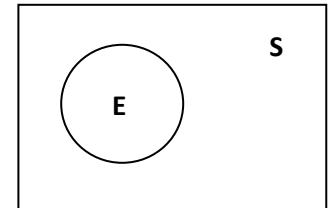
1. Classical Probability

$$\text{Probability of an Event} = P(E) = \frac{n(E)}{n(S)}$$

Probability Rule $0 \leq P(E) \leq 1$

- a) $P(E) = 0$; an event E is uncertain (0%)
 Φ (Phi) = no number in the sample place
- b) $P(E) = 1$; an event E is certain (100%)

The sum of probabilities of all outcomes in the Sample Space



at least (no less than)	at most (no more than)	less than	greater than
$x \geq \#$	$x \leq \#$	$x < \#$	$x > \#$

Ex 9) A die is rolled, let $A = \{1\}$. $P(1)$?

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{6} \approx 16.67\% \text{ chance}$$

Ex 10) A die is rolled Let $B = \{2,4,1,3\}$

* The order isn't important in the set of notation.

$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{6} \approx 66.67\% \text{ chance}$$

Ex 11) A die is rolled Let $P(S)$

$$P(S) = \frac{n(S)}{n(S)} = \frac{6}{6} = 1 = 100\% \text{ chance} \quad \text{certain}$$

Ex 12) An event of observing the 13 when a die is rolled.

Let $P(\Phi) = \{13\}$ (Φ ; phi Greek)

$$P(\Phi) = \frac{n(\Phi)}{n(S)} = \frac{0}{6} = 0 = 0\% \text{ chance} \quad \text{uncertain}$$

P194 [Ex4-7] Drawing a card from a deck (52 cards)

a) Of getting a Jack

$$P(\text{Jack}) = \frac{4 \text{ Jacks}}{52 \text{ Cards}} = \frac{1}{13} \approx 7.69\% \text{ (Unlikely)}$$

b) Of getting the 6 of clubs

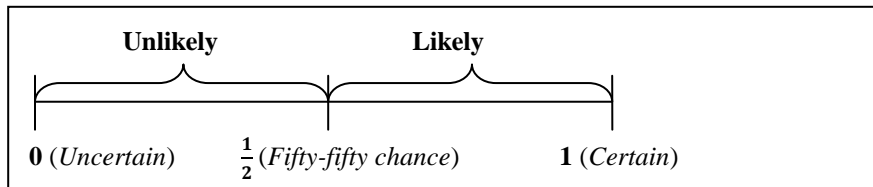
$$P(6 \text{ of Clubs}) = \frac{1 (6 \text{ of clubs})}{52 \text{ Cards}} \approx 1.92\% \text{ (Unlikely)}$$

c) Of getting a 3 or a diamond

$$P(3 \text{ or diamond}) = \frac{4 (3_s) + 13 (\spadesuit \text{ cards}) - 1 (3 \text{ of } \spadesuit)}{52 \text{ Cards}} = \frac{16}{52} = \frac{4}{13} \approx 30.77\% \text{ chance (Unlikely)}$$

d) Of getting a 3 or a 6

$$P(a 3 \text{ or a } 6) = \frac{4(3_s) + 4(6_s)}{52 \text{ Cards}} = \frac{8}{52} \approx 15.39\% \text{ chance}$$



$$* \text{ Real Frequency} = \frac{f}{n} \quad (n = \sum f)$$

2. Empirical Probability

Given a frequency distribution, the probability of an event being in a given class and it is based on observation.

$$P(E) = \frac{f}{n} = \frac{\text{frequency for the class}}{\text{total frequencies in the distribution}}$$

p200 [Ex4-13]

Distribution of Blood Type - Find the following probabilities

Type	A	B	AB	D	Total
Frequency	22	5	2	21	50

a) A person has type O blood

$$P(O) = \frac{f}{n} = \frac{21}{50} = 42\% \text{ chance}$$

b) A person type A or type b blood

$$P(A \text{ or } B) = \frac{22 + 5}{50} = \frac{27}{50} = 0.54 = 54\% \text{ chance (Likely)}$$

c) A person neither type A nor O blood

$$P(nAnO) = \frac{5 + 2}{50} = \frac{7}{50} = 0.14 = 14\% \text{ chance (Unlikely)}$$

d) A person doesn't have type AB blood

$$P(\text{not AB}) = 1 - \frac{2}{50} = \frac{48}{50} = \frac{24}{25} = 96\% \text{ (Likely)}$$

P201 [Ex4-14]

Number of days of maternity patients stayed in the hospital in the distribution

Number of days stayed	3	4	5	6	7
Total= 127	15	32	56	19	5

a) A patient stayed Exactly 5 days

$$P(5) = \frac{56}{127} \approx 44.10\% \text{ chance (Unlikely)}$$

b) Less than 6 days

$$P(\text{less than } 6) = \frac{15 + 32 + 56}{127} = \frac{103}{127} \approx 81.10\% \text{ chance (likely)}$$

c) At most 4 days

$$P(\text{at most } 4) = \frac{15 + 32}{127} = \frac{47}{127} \approx 37.01\% \text{ chance (unlikely)}$$

d) At least 5 days

$$P(\text{at least } 5) = \frac{56 + 19 + 5}{127} = \frac{80}{127} \approx 62.99\% \text{ chance (likely)}$$

3. Subjective Probability

; The type of probability that uses a probability value based on an educated guess or estimate, employing opinions and inexact information
(based on the person's experience and education of a solution)

• Complementary Events

$P(\overline{E})$ = Event of those outcomes which are not in E

1. $P(\overline{E}) = P(S) - P(E)$
2. $P(E) = 1 - P(\overline{E})$
3. $P(S) = P(E) + P(\overline{E}) = 1$
4. "at least one" = complementary of "none"
"none" = "complementary of "at least one"
 $P(\text{at least one}) = 1 - P(\text{none})$
 $P(\text{none}) = 1 - P(\text{at least one})$

P197 Ex 4-10]

Finding Complements

- a) Rolling a die and getting a 4

$$P(\overline{4}) = 1 - \frac{1}{6} = \frac{5}{6}$$

- b) Selecting a month and getting a month that begins with J.

$$P(\overline{\text{with J}}) = 1 - \frac{3(\text{Jan, Jun, Jul})}{12} = \frac{9}{12} = \frac{3}{4}$$

- c) Selecting a day of the week and getting a weekday

$$P(\overline{\text{weekday}}) = 1 - \frac{5}{7} = \frac{2}{7} = \frac{2(\text{Sat, Sun})}{7 \text{ days}}$$

P205 24] Computers in Elementary School

Elementary and secondary schools were classified by number of computers they had.

Choose one of these schools at random.

Computers	1-10	11-20	21-50	51-100	100+
Schools	3170	4590	16,741	23,753	34,803

Find the probability that it has.

- a. 50 or fewer computers 0.295

Find total 83057, no intersection

$$\frac{3170}{83057} + \frac{4591}{83057} + \frac{16741}{83057} = \frac{24501}{83057} \approx 0.2950$$

- b. More than 100 computers 0.419

$$\frac{34803}{83057} \approx 0.4190$$

- c. No more than 20 computers 0.093

$$\frac{3170}{83057} + \frac{4591}{83057} = \frac{7761}{83057} \approx 0.0934$$

*(in class) Choose class "50-100"

$$\frac{23753}{83057} \approx 0.2860 \quad 12.6\% \text{ chance unlikely}$$

P205 19] Prime Numbers

A prime number is a number that is evenly divisible only 1 and itself. Those less than 100 are listed below.

2 3 5 7 11 13 17 19 23 29 31 37 41
43 47 53 59 61 67 71 73 79 83 89 97

Choose one at random and find the probability that

- a. The number is even

$$P(\text{even}) = \frac{n(\text{even})}{n(S)} = \frac{1}{25} = 0.04$$

- b. The sum of the number's digit is even

$$P(\text{sum, ven}) = \frac{n(E)}{n(S)} = \frac{13}{25} = 0.52$$

- c. The number is greater than 50

$$P(> 50) = \frac{n(E)}{n(S)} = \frac{10}{25} = 0.40$$

4-2 The Addition Rules for Probability

• Mutually Exclusive Events

; Probability events that cannot occur at the same time

• Event

1. **Simple**; can't break the event ex) $E=\{1\}$
2. **Compound**; "and" ; "or" ex) $\{1, 2, 3\} = \{1\} \cup \{2\} \cup \{3\}$

$\cup = \text{Union} = \text{or} = \text{total area}$

$\cap = \text{Intersection} = \text{and} = \text{common area}$

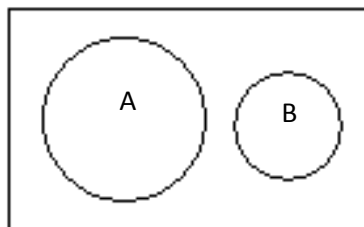
Case 1

(Mutually exclusive events)

$$P(A \text{ or } B) = P(A) + P(B)$$

*In only single trial, event A or B occurs
and no intersection

*A and B are mutually exclusive
(i.e., disjoint $A \cap B = \emptyset$)

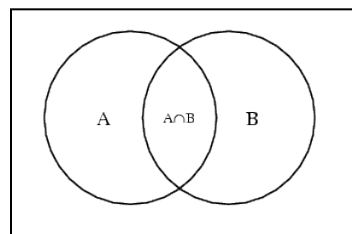


Case 2

(No Mutually exclusive events)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

*A and B aren't mutually exclusive
(i.e., $A \cap B \neq \emptyset$)



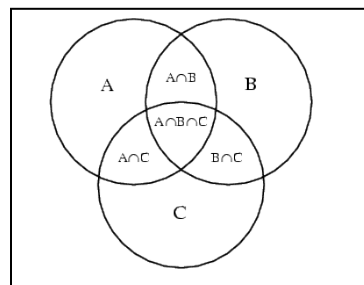
Case 3 (an extra case)

$$P(A \text{ or } B \text{ or } C)$$

$$= P(A) + P(B) + P(C)$$

$$- P(A+B) - P(A+C) - P(B+C)$$

$$+ P(A+B+C)$$



P212 [2] Determine whether these events are mutually exclusive.

- a. Roll a die: Get an even number, and get a number less than 3.
- b. Roll a die: Get a prime number (2,3,5), and get an odd number.
- c. Roll a die: Get a number greater than 3, and get a number less than 3.
- d. Select a student in your class:
The student has blond hair, and the student has blue eyes.
- e. Select a student in your college:
The student is a sophomore, and the business major.
- f. Select any course:
It is a calculus course, and it is an English course.
- g. Select a registered voter: The voter is a Republican, and the voter is a Democrat.

Ans: Yes- c, f, and g.

P212 [5] At a convention there are instructors of 7 mathematics, 5 computer science, 3 statistics, and 4 science.
If an instructor is selected, find the probability of getting a science or math instructor.

$$\text{Total} = P(S) = 7+5+3+4=19$$

$$P(\text{math or sci}) = \frac{n(\text{math})}{n(S)} + \frac{n(\text{sci})}{n(S)} = \frac{7}{19} + \frac{4}{19} = \frac{11}{19}$$

Ex] A die is rolled one time, find $P(E)$ getting 4 or less than 6.

$$P(4) + P(\text{less than } 6) - P(4 \text{ in less than } 6) = \frac{1}{6} + \frac{5}{6} - \frac{1}{6} = \frac{5}{6}$$

Ex] A card is drawn randomly from an ordinary deck of 52 cards

Find $P(\text{the card is diamond or an ace})$

$$P(\text{diamond}) + P(\text{Ace}) - P(\text{Ace of diamond}) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}$$

P209 [Ex4- 20]

A single card is drawn at random from an ordinary deck of cards. Find the probability of either an ace or a black card.

$$P(\text{an ace}) + P(\text{a black card}) - P(\text{ace + vlack card})$$

$$= \frac{4}{52} + \frac{26(\text{all black})}{52} - \frac{2(2 \text{ black aces})}{52} = \frac{28}{52} \approx 53.85\% \text{ chance}$$

(likely)

P209 [Ex 4-24]

In a hospital unit, 8 nurses and 5 physicians; 7 nurses and 3 physicians are females. Find the probability that the subject is a nurse or a male.

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

$$P(\text{nurse}) + P(\text{male}) - P(\text{nurse} + \text{male})$$

$$= \frac{8}{13} + \frac{3}{13} - \frac{1}{13} = \frac{10}{13} \approx 76.92\% \text{ chance (likely)}$$

p213 [13] $P(\text{male}) + P(18 \sim 24) - P(\text{Male in } 18 \sim 24) =$
 $\frac{10456}{17230} + \frac{13701}{17230} - \frac{7922}{17230} = \frac{16235}{17230} \approx 0.9423 = 94.23\% \text{ chance}$

Ex] In a statistics class there are 18 juniors, 10 seniors; 6 of the seniors are females, and 12 of the juniors are males. If a student is selected at random, find the probability of selecting the following:

- a. A junior or a female

18 Juniors = 12 males + 6 females

10 Seniors = 4 males + 6 females

28 students = 16 males + 12 females

$$P(\text{Junior}) + P(\text{Female}) - P(\text{Junior} \cap \text{Female})$$

$$= \frac{18}{28} + \frac{12}{28} - \frac{6}{28} = \frac{24}{28} = \frac{6}{7}$$

- b. A senior or a female

$$P(\text{Senior}) + P(\text{Female}) - P(\text{Senior} \cap \text{Female})$$

$$= \frac{10}{28} + \frac{12}{28} - \frac{6}{28} = \frac{16}{28} = \frac{4}{7}$$

- c. A junior or a senior

$$P(\text{Junior}) + P(\text{Senior}) - P(\text{Junior} \cap \text{Senior})$$

$$= \frac{18}{28} + \frac{10}{28} - \frac{0}{28} = \frac{28}{28} = 1$$

4 – 3**The Multiplication Rules and Conditional Probability**

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

= (An event A occurs in the 1st trial and event B occurs in the 2nd trial)

(* “and” or “both” is in a sentence.)

Case 1 Independent Event $P(A) \cdot P(B)$

When A and B are independent

(i.e., the occurrence of A doesn't affect the probability of the occurrence of B)

Ex] Find the probability of getting a Head on the coin and a 4 on the die

$$\frac{n(\text{Head})}{n(\text{Head} \& \text{Tail})} \cdot \frac{n(4)}{n(1 \sim 6)} = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

P220 Ex 4-25]

There are 3 red balls, 2 blue balls, 5 white balls.

2 items selected and replaced the cards.

(→ replaced the cards = independent, 2 events)

- a. 2 blue balls

$$P(\text{Blue}) \cdot P(\text{Blue}) = \frac{2}{10} \cdot \frac{2}{10} = \frac{1}{25}$$

- b. A blue and a white

$$P(\text{Blue}) \cdot P(\text{white}) = \frac{2}{10} \cdot \frac{5}{10} = \frac{1}{10}$$

- c. A red and blue

$$P(\text{Red}) \cdot P(\text{Blue}) = \frac{3}{10} \cdot \frac{2}{10} = \frac{3}{50}$$

Case 2 Dependent Event $P(A) \cdot P(B|A)$

Where P(B|A) Probability B, given that A is already occurred.

(* The event A – the 1st outcome, a given event, or previous event - using past sentence)

(* The event B – the 2nd outcome or the last event)

When the probability of the occurrence of the event B is affected by the occurrence of the event A.

P222 Ex 4-30) 3 Cards are drawn from a deck and not replaced the cards
(Not replaced = Dependent)

a. Getting 3 Jacks

$$P(\text{Jack } 1^{\text{st}}) \cdot P(\text{Jack } 2^{\text{nd}}) \cdot P(\text{Jack } 3^{\text{rd}}) = \frac{4}{52} \cdot \frac{3(4-1)}{51(52-1)} \cdot \frac{2(4-2)}{50(52-2)} = \frac{1}{5525}$$

b. Getting an Ace, a King, a Queen

$$P(\text{Ace}) \cdot P(\text{King}) \cdot P(\text{Queen}) = \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50} = \frac{8}{16575}$$

c. Getting a club, a spade, a heart

$$P(\text{club}) \cdot P(\text{spade}) \cdot P(\text{heart}) = \frac{13}{52} \cdot \frac{13}{51} \cdot \frac{13}{50} = \frac{169}{10200}$$

d. Getting 3 clubs

$$P(\text{Club } 1^{\text{st}}) \cdot P(\text{Club } 2^{\text{nd}}) \cdot P(\text{Club } 3^{\text{rd}}) = \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} = \frac{11}{850}$$

Ex) 30% chance to get sick. Find of the probability of selecting
2 students and they both are sick in the school.

(→ It's a dependent case and there is already probability)

$$P(1^{\text{st}} \text{ student}) \cdot P(2^{\text{nd}} \text{ student}) = (0.3) \cdot (0.3) = 0.09 = 9\% \text{ chance}$$

Conditional Probability

$P(B|A)$ = Probability of B given that of A

$$= \frac{P(A \cap B)}{P(A)} = \frac{\frac{n(A \cap B)}{n(S)}}{\frac{n(A)}{n(S)}} = \frac{n(A \cap B)}{n(A)}$$

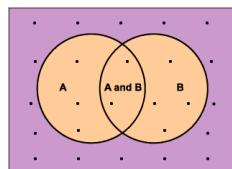
($P(A) \neq 0, n(A) \neq 0$)

When $P(A \cap B) = P(A) \cdot P(B)$, the case is Independent

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

When $P(A \cap B) \neq P(A) \cdot P(B)$, the case is Dependent

$$P(B|A) \neq P(B)$$



Ex] A die is rolled twice. Find the probability of getting 4
after getting an even number.

→ Event A = P(even number) ; 1st outcome

→ Event B = P("4") ; 2nd outcome

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(4 \text{ from } 2,4,6)}{P(2,4,6 = \text{even\#s})} = \frac{\frac{n(A \cap B)}{n(S)}}{\frac{n(A)}{n(S)}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

P225 Ex 4-32]

A box contains black chips and white chips.

A person selects 2 chips without replacement.

If the probability of selecting a black chip *and* a white chip is $\frac{15}{56}$, and
if the probability of selecting a black chip on the first draw is $\frac{3}{8}$,

and it's given that.

Find the probability of selecting a white chip on the second draw.

$$P(\text{white}|\text{black}) = \frac{P(B \text{ and } W)}{P(B)} = \frac{\frac{15}{56}}{\frac{3}{8}} = \frac{5}{7}$$

P225 Ex 4-34]

A recent survey asked 100 people if they thought women in
the armed forces should be permitted to participate in combat

Gender	Yes	No	Total
Male	32	18	50
Female	8	42	50
Total	40	60	100

a. The respondent answered yes,

given that the person was a female.

(→ was a female; 1st event, yes; 2nd event)

$$\frac{P(\text{Female and yes})}{P(\text{Female})} = \frac{\frac{8}{100}}{\frac{50}{100}} = \frac{4}{25} \text{ or } \frac{n(\text{yes in female})}{n(\text{total of female})}$$

b. The resident was a male, **given that** the person answered no

(→ answered no; 1st event, male; 2nd event)

$$\frac{P(\text{Male and no})}{P(\text{no})} = \frac{\frac{18}{100}}{\frac{60}{100}} = \frac{3}{10} \text{ or } \frac{n(\text{male in no})}{n(\text{total of no})} = \frac{18}{60}$$

P230 [33] At an exclusive country club,

68% of the members play bridge and drink champagne,
and 83% play bridge .

If a member is selected at random, find the probability
that the member drinks champagne,
given that he or she plays bridge.

$$P(\text{champagne}|\text{bridge}) = \frac{P(\text{cham. \& bri.})}{P(\text{bridge})} = \frac{0.68}{0.83} = 81.93\%$$

Try P230 [34]

4 - 4 Counting Rules

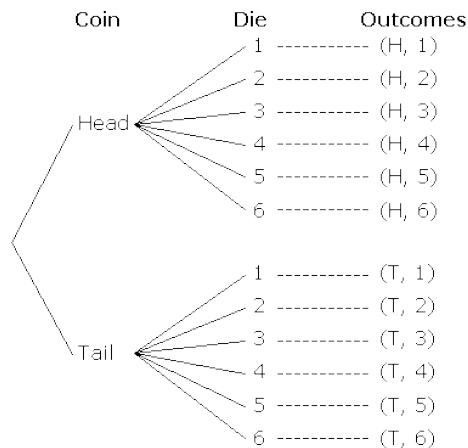
1. The Fundamental Counting Rule

In a sequence of n events in which the 1st one has k_1 , possibilities and so on, the total number of possibilities of the sequence.

$$= k_1 \cdot k_2 \cdot k_3 \cdots k_n$$

- When events are just listed with “and”, it’s counting rule case.
- Event A, event B **and** event C = Event A and event B and event C
(In this case “and” means to multiply)

P233 Ex 4-38] Tossing a coin **and** rolling a die, find the number of outcomes for the sequence of events.



(\rightarrow 2 different event = 1st outcome 2nd outcome)
 $n(\text{coin}) \cdot n(\text{die}) = 2 \cdot 6 = 12 \text{ outcomes}$

P233 Ex 4-38]

A paint manufacturer wishes to manufacture several different paints.

Color	Red, blue, white, black, green, brown, yellow
Type	Latex, oil
Texture	Flat, semi gloss, high gloss
Use	Outdoor, indoor

How many different kinds of paint can be made if a person select one color, one type, one texture, **and** one use?

$$n(\text{color}) \cdot n(\text{type}) \cdot n(\text{texture}) \cdot n(\text{use}) = 7 \cdot 2 \cdot 3 \cdot 2 = 84 \text{ ways}$$

Ex] How many ways can a dinner patron select 2 appetizers, 2 drinks, 3 foods, **and** 2 desserts on the menu?

$$n(\text{appetizer}) \cdot n(\text{drink}) \cdot n(\text{food}) \cdot n(\text{dessert}) \\ = 2 \cdot 2 \cdot 3 \cdot 2 = 24 \text{ ways}$$

Ex] The digit 0, 1, 2, 3, and 4 are to be used in a four-digit ID card. How many different cards are possible

- if it can be repeated.

$$n(1\text{st}) \cdot n(2\text{nd}) \cdot n(3\text{rd}) \cdot n(4\text{th}) = 5 \cdot 5 \cdot 5 \cdot 5 = 625 \text{ cards}$$

$$n(5 \text{ digits}) \cdot n(5 \text{ digits}) \cdot n(5 \text{ digits}) \cdot n(5 \text{ digits})$$

- If it cannot be repeated

$$n(1\text{st}) \cdot n(2\text{nd}) \cdot n(3\text{rd}) \cdot n(4\text{th}) = 5 \cdot 4 \cdot 3 \cdot 2 = 120 \text{ cards}$$

$$n(5 \text{ digits}) \cdot n(5 - 1 \text{ digits}) \cdot n(4 - 1 \text{ digits}) \cdot n(3 - 1 \text{ digits})$$

$${}_5P_4 = \frac{n!}{(n-r)!} = \frac{5!}{(5-4)!} = 5 \cdot 4 \cdot 3 \cdot 2 = 120$$

• Factorial Notation

;the number of ways a square of n events can over
 if the 1st event can occur in k_1 ways, the 2nd event can occur
 in k_2 ways, etc.

$$n! = n(n-1)(n-2) \cdots i$$

$$\text{ex) } 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120 \quad 0! = 1$$

P241 [1] How many ways can a base ball manager arrange
 A batting order of 9 player? (no repeat)

(\rightarrow 9 positions 9 players)

$$n(1\text{st}) \cdot n(2\text{nd}) \cdot n(3\text{rd}) \cdot n(4\text{th}) \cdot n(5\text{th}) \cdot n(6\text{th}) \cdot n(7\text{th}) \cdot n(8\text{th}) \\ \cdot n(9\text{th}) = 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 9! \text{ 362,880 ways}$$

Ex] Florida lottery = {1, 2, 3, ..., 53}

By choosing any six numbers out of 53 numbers and the
 picked numbers are not in order.

$${}_{53}C_6 = 22,957,480 = n(S) = \text{total (using calculator)} \rightarrow \text{Very unlikely}$$

$$P(\text{winning}) = \frac{n(E)}{n(S)} = \frac{1}{22,957,480} \approx 0.000000044\% \text{ chance}$$

$${}_{53}C_6 = 22,957,480 \text{ outcomes}$$



One chance to win

2. Permutation Rule

Ordered arrangement of different things

$$nPr = \frac{n!}{(n-r)!} \quad \text{where } r \leq n$$

$$\text{Note; } nPn = n! = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1}$$

3. Combination Rule

A set of different objects in which ordering isn't important.

$$nCr = \frac{n!}{(n-r)!r!} = \frac{nPr}{r!} \quad \text{where } r \leq n$$

(A set of items in which ordering isn't important)

$$\text{Note; } nCn = 1 = \frac{n!}{(n-n)!n!} = \frac{n!}{0!n!} = \frac{n!}{1 \cdot n!}$$

'n'; items (all different)

'r'; items selected out of 'n'

p238 Ex 4-46]

Given the letters A, B, C, and D list the permutations and Combinations for selecting 2 letters.

Permutation	Combination
AB BA CA DA AC BC CB DB AD BD CD DC	AB BC AC BD AD CD
12 ways	6 ways
	The elements of a combination are usually listed alphabetically.

p238 Ex 4-49]

In a club there are 7 women **and** 5 men. A committee of 3 women **and** 2 men is to be chosen. Hpw many different Possibilities are there?

$$P(\text{women}) \cdot P(\text{men}) = {}_7C_3 \cdot {}_5C_2 = \frac{7!}{(7-3)!3!} \cdot \frac{5!}{(5-2)!2!} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} \cdot \frac{5 \cdot 4}{2 \cdot 1} = 350 \text{ ways}$$

P241 [1] How many 5-digit zip codes are possible

a. if digit can be repeated?

(→ 5 places, 10 digits = 0~9)

$$n(1st) \cdot n(2nd) \cdot n(3rd) \cdot n(4th) \cdot n(5th) \\ = 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 100,000 \text{ ways}$$

b. If there cannot be repetitions?

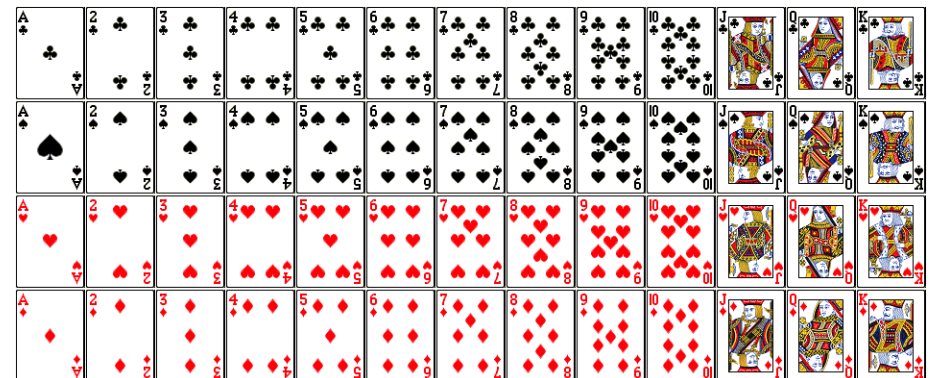
$$n(1st) \cdot n(2nd) \cdot n(3rd) \cdot n(4th) \cdot n(5th) \\ = n(10 \text{ digits}) \cdot n(10-1) \cdot n(9-1) \cdot n(8-1) \cdot n(7-1) \\ = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 = 30,240 \text{ ways}$$

$${}_{10}P_5 = \frac{10!}{(10-5)!} = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 = 30,240$$

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

(→ order & repetition are not important.= Combination)

$${}_{20}C_5 = \frac{20!}{(20-5)!5!} = 15,504 \text{ tests}$$



Ch5

- From Ch1

A Discrete Variable: assume values that can be counted

A Continuous Variable: can assume all values in the interval between any 2 values

Discrete Probability Distribution

Consists of the values a random variable can assume and corresponding probabilities of values. The probabilities are determined theoretically or by observation.

P262 Ex5-1] Construct a probability distribution for rolling a die.

Outcome	X	1	2	3	4	5	6	$\sum P(x)$
Probability	$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

X is Discrete Probability Distribution.

2 Requirements for a Probability Distribution (P.D.)

- $\sum P(x) = 1$
- $0 \leq P(X) \leq 1$

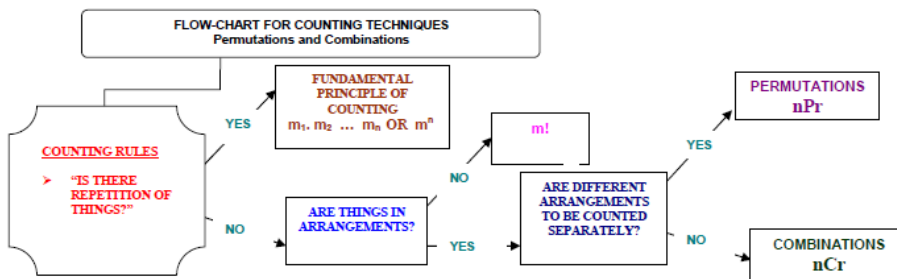
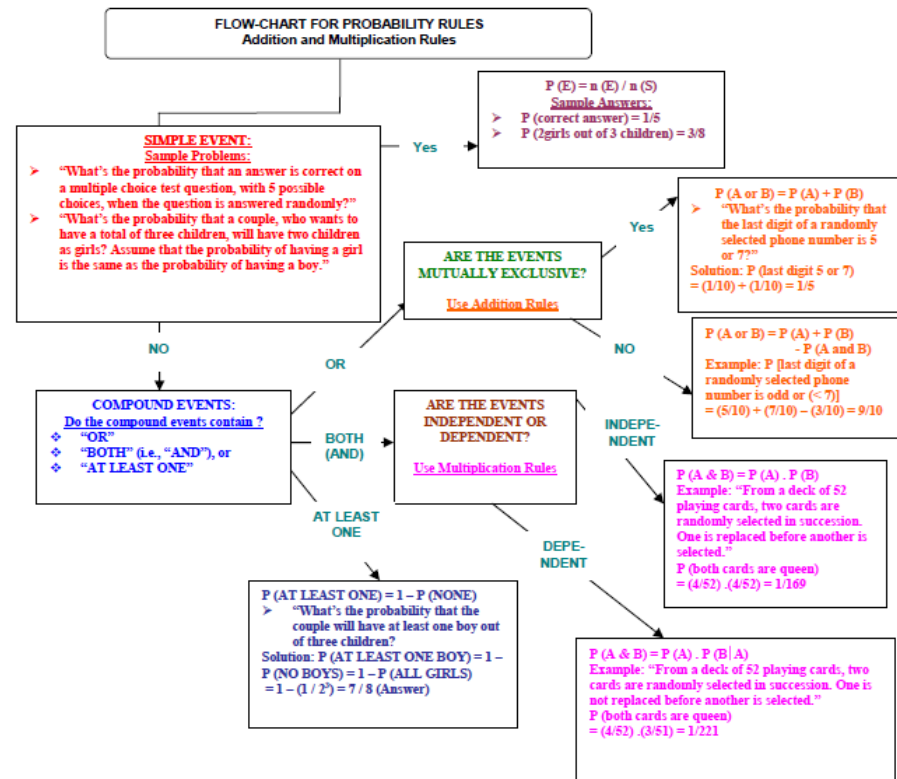
P265 Ex 5-4] Determine whether each distribution is a Probability Distribution (P.D.)

X	0	5	10	15	20	P.D.
$P(X)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\sum P(x) = 1 \quad 0 \leq P(X) \leq 1$

X	0	2	4	6	No P.D.
$P(X)$	-1.0	1.5	0.3	0.2	$\sum P(x) = 1, \text{ but } -1.0 \text{ \& } 1.5 \text{ aren't } 0 \leq P(X) \leq 1$

X	1	2	3	4	P.D.
$P(X)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\sum P(x) = 1 \quad 0 \leq P(X) \leq 1$

X	2	3	7	No P.D.
$P(X)$	0.5	0.3	0.4	$\sum P(x) \neq 1, \quad 0 \leq P(X) \leq 1$



• **Mean of a Probability Distribution (P.D.)**

$$\mu = \sum X \cdot P(X)$$

The mean of a random variable with a discrete probability distribution

X ; outcomes

$P(X)$; corresponding probability

P262 Ex5-5] Construct a probability distribution for rolling a die.

Outcome	X	1	2	3	4	5	6	$\sum P(x)$
Probability	$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1
Mean $\mu = \sum X \cdot P(X)$		$1 \cdot \frac{1}{6}$	$2 \cdot \frac{1}{6}$	$3 \cdot \frac{1}{6}$	$4 \cdot \frac{1}{6}$	$5 \cdot \frac{1}{6}$	$6 \cdot \frac{1}{6}$	3.5

(A die doesn't have 3.5, but the theoretical average is 3.5.)

P262 Ex5-6] In a family with 2 kids, find the mean of the number of the kids who will be girls.

# of girls	0 girl	1 girl	2 girls
$P(X)$	$\frac{n(2 \text{ boys})}{n(S)} = \frac{1}{4}$	$\frac{n(1G1B \text{ or } 1B1G)}{n(S)} = \frac{2}{4} = \frac{1}{2}$	$\frac{n(2 G)}{n(S)} = \frac{1}{4}$

$$n(S) = 1st \text{ kid} \cdot 2nd \text{ kid} = 2 \cdot 2 = 4$$

$$\mu = \sum X \cdot P(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

p275 #3]

X	0	1	2	3	4
$P(X)$	0.18	0.44	0.27	0.08	0.03

a. Is this a probability distribution?

a) all $P(x)$ are $0 \leq P(X) \leq 1$

b) $\sum P(x) = 0.18 + 0.44 + 0.27 + 0.08 + 0.03 = 1$

c) Thus it's probability distribution.

b. Find it's mean. $\mu = 0 \cdot 0.18 + 1 \cdot 0.44 + 2 \cdot 0.27 + 3 \cdot 0.08 + 4 \cdot 0.03 = 1.34$

p275 #8]

X	1	2	3	4
$P(X)$	0.32	0.51	0.12	0.05

Is this a probability distribution?

a) all $P(x)$ are $0 \leq P(X) \leq 1$

b) $\sum P(x) = 0.32 + 0.51 + 0.12 + 0.05 = 1.9$

c) Thus it's not probability distribution.

• **Binomial Probability distribution Requirements**

1. There must be a fixed number of trials $[1, \infty)$
2. The probability of a success must remain the same for each trial.
3. Each trial can have only two outcomes or outcomes that can be reduced to outcomes. $(\frac{1}{2}, \frac{1}{2})$
4. The outcomes of each trial must be independent for each other.

P285 #1] Are they binomial experiments or not?

Yes/No (fixed number of trials, only two outcomes)

1. Surveying 100 people to determine if they like Sudsy Soap.
Yes (100, like or dislike)
2. Tossing a coin 100 times to see how many heads occur.
Yes (100, head or tail)
3. Drawing a card from a deck and getting a heart
Yes (1, heart or no heart)
4. Asking 1000 people which brand of cigarettes they smoke
No (1000, more than 2 brands)
5. Testing 4 different brands of aspirin to see which brands are effective
No (no, 4 brands)
6. Testing 1 brand of aspirin by using 10 people to determine whether it is effective
Yes (10, effective or not)
7. Asking 100 people if they smoke
Yes (100, smoke or no smoke)
8. Checking 1000 applicants to see whether they were admitted to White Oak College
Yes (1000, admitted or not)
9. Surveying 300 prisoners to see how many different crimes they were convicted
No (300, more than 2 crimes)
10. Surveying 300 prisoners to see whether this is their 1st offence
Yes (300, 1st offence or not)

• **Binomial Probability**

$$P(X) = {}_n C_X \cdot p^X \cdot q^{n-X} \quad 0 \leq X \leq n$$

p : the numerical probability of a success

$$p = P(S)$$

q : the numerical probability of a failure

$$q = P(F) = 1 - p$$

n : the number of trials

X : the number of successes in n trials (where 0,1,2 ... n)

$n - X$: the number of failure in n trials

P285 #3] Compute the probability of X success,

Using Table B in Appendix C.(p636)

1. $n = 2, X = 1, p = 0.30 \quad P(x) = 0.420$
2. $n = 4, X = 3, p = 0.60 \quad P(x) = 0.590$
3. $n = 5, X = 0, p = 0.10 \quad P(x) = 0.000$
4. $n = 10, X = 4, p = 0.40 \quad P(x) = 0.418$
5. $n = 12, X = 2, p = 0.90 \quad P(x) = 0.246$
6. $n = 15, X = 12, p = 0.80 \quad P(x) = 0.346$
7. $n = 17, X = 0, p = 0.05 \quad P(x) = 0.251$
8. $n = 20, X = 10, p = 0.50 \quad P(x) = 0.250$
9. $n = 16, X = 3, p = 0.20 \quad P(x) = 0.176$

P285 #3] Compute the binomial probability of X success,

1. $n = 6, X = 3, p = 0.03$
 $q = 1 - 0.03 = 0.97 \quad P(X) = {}_n C_X \cdot p^X \cdot q^{n-X} = {}_6 C_3 \cdot (0.03)^3 \cdot (0.97)^{6-3} = 0.0005$
2. $n = 4, X = 2, p = 0.18$
 $q = 1 - 0.18 = 0.82 \quad P(X) = {}_n C_X \cdot p^X \cdot q^{n-X} = {}_4 C_2 \cdot (0.18)^3 \cdot (0.82)^{4-2} = 0.131$
3. $n = 5, X = 3, p = 0.63 \quad P(x) = 0.420$
 $q = 1 - 0.63 = 0.37 \quad P(X) = {}_n C_X \cdot p^X \cdot q^{n-X} = {}_5 C_3 \cdot (0.63)^5 \cdot (0.37)^{5-3} = 0.342$
4. $n = 9, X = 0, p = 0.42 \quad P(x) = 0.420$
 $q = 1 - 0.42 = 0.58 \quad P(X) = {}_n C_X \cdot p^X \cdot q^{n-X} = {}_9 C_0 \cdot (0.42)^9 \cdot (0.58)^{9-0} = 0.007$
5. $n = 10, X = 5, p = 0.37 \quad P(x) = 0.420$
 $q = 1 - 0.37 = 0.63 \quad P(X) = {}_n C_X \cdot p^X \cdot q^{n-X} = {}_{10} C_5 \cdot (0.37)^{10} \cdot (0.63)^{10-5} = 0.173$

• **Using Table:** If p is 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 or 0.95

• **Binomial Probability**

Mean $\mu = n \cdot p$

Variance $\sigma^2 = n \cdot p \cdot q$

Standard deviation $\sigma = \sqrt{n \cdot p \cdot q}$

P286 # 14] Find mean, variance, and standard deviation

1. $n = 100 \quad p = 0.75$
 $q = 0.25 \quad \mu = 100 \cdot 0.75 = 75$
 $\sigma^2 = 100 \cdot 0.75 \cdot 0.25 = 18.75 \quad \sigma = \sqrt{100 \cdot 0.75 \cdot 0.25} = 4.33$
2. $n = 300 \quad p = 0.3$
 $q = 0.7 \quad \mu = 300 \cdot 0.3 = 90$
 $\sigma^2 = 300 \cdot 0.7 \cdot 0.3 = 63 \quad \sigma = \sqrt{300 \cdot 0.7 \cdot 0.3} = 7.9$
3. $n = 20 \quad p = 0.5$
 $q = 0.5 \quad \mu = 20 \cdot 0.5 = 10$
 $\sigma^2 = 20 \cdot 0.5 \cdot 0.5 = 5 \quad \sigma = \sqrt{20 \cdot 0.5 \cdot 0.5} = 2.2$
4. $n = 10 \quad p = 0.8$
 $q = 0.2 \quad \mu = 10 \cdot 0.8 = 8$
 $\sigma^2 = 10 \cdot 0.8 \cdot 0.2 = 1.6 \quad \sigma = \sqrt{10 \cdot 0.8 \cdot 0.2} = 1.3$
5. $n = 1000 \quad p = 0.1$
 $q = 0.9 \quad \mu = 1000 \cdot 0.1 = 100$
 $\sigma^2 = 1000 \cdot 0.1 \cdot 0.9 = 90 \quad \sigma = \sqrt{1000 \cdot 0.1 \cdot 0.9} = 9.5$
6. $n = 500 \quad p = 0.25$
 $q = 0.75 \quad \mu = 500 \cdot 0.25 = 125$
 $\sigma^2 = 500 \cdot 0.25 \cdot 0.75 = 93.8 \quad \sigma = \sqrt{500 \cdot 0.25 \cdot 0.75} = 9.7$
7. $n = 50 \quad p = \frac{2}{5} = 0.4$
 $q = 0.6 \quad \mu = 50 \cdot 0.4 = 20$
 $\sigma^2 = 50 \cdot 0.4 \cdot 0.6 = 12 \quad \sigma = \sqrt{50 \cdot 0.4 \cdot 0.6} = 3.5$
8. $n = 36 \quad p = \frac{1}{6} \approx 0.17$
 $q = 0.83 \quad \mu = 36 \cdot 0.17 = 6.12$
 $\sigma^2 = 36 \cdot 0.17 \cdot 0.83 = 5.08 \quad \sigma = \sqrt{36 \cdot 0.17 \cdot 0.83} = 2.25$

Ch6

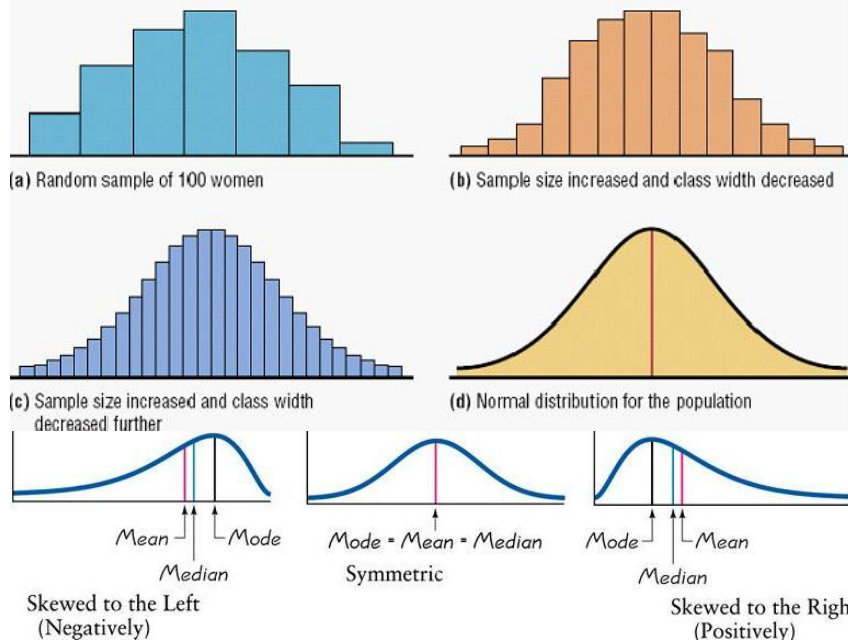
Discrete Random Variable; Binomial Distribution

Continuous Random Variable; Normal distribution interval (a, b)

ex) height, weight, temperature, blood pressure, & time

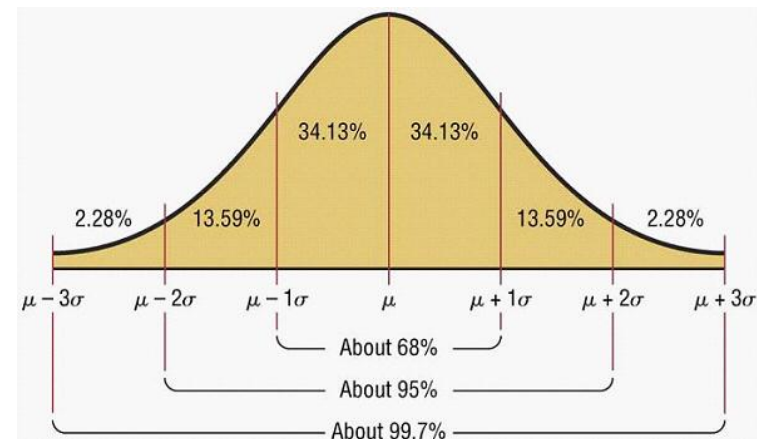
In theory, a normal distribution curve is the theoretical counterpart to a relative frequency histogram for a large number of data values with a very small class width.

Histograms for the distribution of heights



• Properties of a normal distribution

1. A normal distribution curve is **bell-shaped**
2. **The mean, median, and mode are equal** and are located at **the center** of the distribution
3. A normal distribution curve is **unimodal** (i.e., it has only one mode)
4. The curve is **symmetric about the mean**, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center
5. The curve is **continuous**, that is, there are no gaps or holes.
For each value of X, there is a corresponding value of Y
6. **The curve never touches the x axis.**
Theoretically, no matter how far in either direction the curve extends, it never meets the x axis – but it gets increasingly closer
7. **The total area under a normal distribution curve is equal to 1.00**, or 100% ($0 \leq P(x) \leq 1$)
8. The area under the part of a normal curve that lies within 1 standard deviation of the mean is approximately 0.68, or 68%; within 2 standard deviations, about 0.95, or 95%; and within 3 standard deviations, about 0.997, or 99.7%. The Empirical rule applies.



In statistics, a standard score is derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

- Standard scores are also called z-scores.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

- Standard Normal Distribution** is a normal distribution with a mean of 0 and a standard deviation of 1. ($\mu = 0$, $\sigma = 1$)



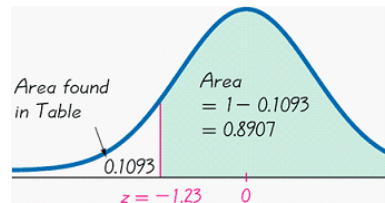
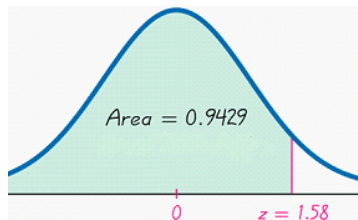
$$y = \frac{e^{-(z^2/2)}}{\sqrt{2\pi}}$$

- Finding Area Under the Standard Normal Distribution Curve**

- Draw a picture
- Put the Z on the graph and shade the area
- Find the value of probability(=area) in the table (Cumulative Standard Normal Distribution Table)

Case 1 For the area to the left of a specified z value, use the table entry directly.

$P(z < a) = P(z > -a)$ the area to the left of $z = a$

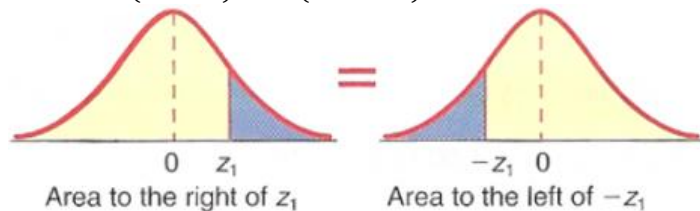


$P(z > -a) = P(z < a)$ the area to the right of $z = a$

Another way $P(z > -a) = 1 - P(z < -a)$

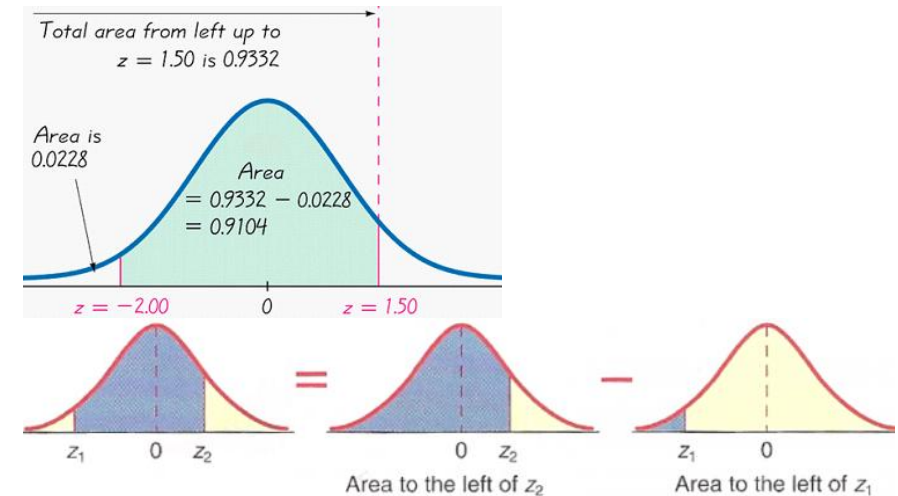
Case 2 $P(z > a) = P(z < -a)$

Since it's symmetric



Case 3 $P(a < z < b)$ the area between $z = a$ and $z = b$

$$= P(z < a \text{ bigger \#}) - P(z < a \text{ smaller \#}) = P(z < b) - P(z < a)$$

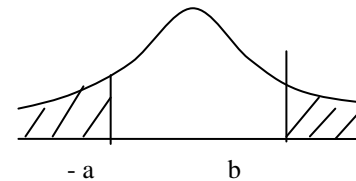


Case 4 $P(z < -a \text{ or } z > b)$

$$= P(z < -a) + P(z < -b)$$

Since it's symmetric

$$= P(z < -a) + [1 - P(z < b)] \quad (A \cap B = \emptyset \text{ mutually exclusive})$$

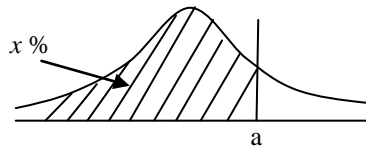


• **Finding z Value that corresponds to the given area**

Find z in the table

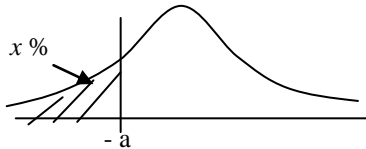
Case 1) $x\%$ of the area or $P(z < a) = x\%$ → Find a

(ex) $89.07\% = 0.8907 \rightarrow a = 1.2 + 0.03 = 1.23$



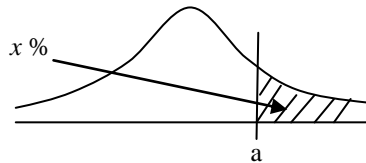
Case 2) $x\%$ of the area or $P(z < -a) = x\%$ → Find $a = P(z > a) = x\%$

(ex) $10.93\% = 0.1093 \rightarrow a = -1.2 + 0.03 = -1.23$



Case 3) $x\%$ of the area or $P(z > a) = x\%$ → Find a

(ex) $10.93\% = 0.1093 \rightarrow a = -1.2 + 0.03 = -1.23 \rightarrow$ answer is 1.23



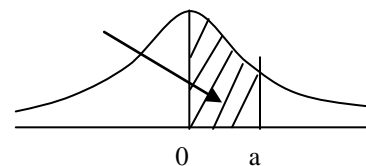
Case 4) $x\%$ of the area or $P(-a < z < a) = x\%$ → Find a

Q; $P(0 < z < a) = 39.07\% = 0.3907$

A; $= P(z < a) - P(z < 0) = P(z < a) - 0.5$

$P(z < a) - 0.5 = 0.3907 \quad P(z < a) = 0.3907 + 0.5 = 0.8907$

$0.8907 \rightarrow a = 1.2 + 0.03 = 1.23$

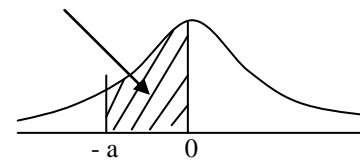


Q; $P(-a < z < 0) = 39.07\% = 0.3907$

A; $= P(z < 0) - P(z < -a) = 0.5 - P(z < -a)$

$0.5 - P(z < a) = 0.3907 \quad P(z < a) = 0.5 - 0.3907 = 0.1093$

$0.1093 \rightarrow a = -1.2 + 0.03 = -1.23$



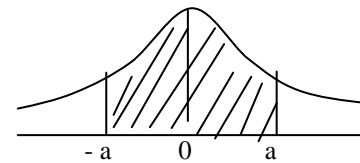
$$P(z < -a) = \frac{(1 - P(-a < z < a))}{2}$$

Q; $P(-a < z < a) = 78.14\% = 0.7814$

A; $0.7814 = 1 - [P(z < -a) \times 2] \quad \text{Since it's symmetric}$

$[P(z < -a) \times 2] = 1 - 0.7814$

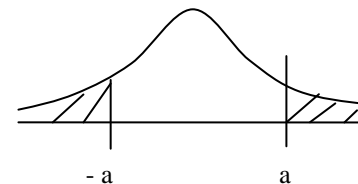
$P(z < -a) = \frac{1 - 0.7814}{2} = 0.1093 \rightarrow \pm a = \pm 1.23$



Q; $P(z < -a \text{ or } z > a) = 21.86\% = 0.2186$

A; $P(z < -a) \times 2 = 0.2186 \quad \text{Since it's symmetric}$

$P(z < -a) = 0.2186 \div 2 = 0.1093 \rightarrow \pm a = \pm 1.23$



- Normal Distribution

Non- Standard Normal distribution; ($\mu \neq 0, \sigma \neq 1$) $\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Standard Normal distribution; ($\mu = 0, \sigma = 1$) $\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$

- Relationship between x and z

Population	Sample
$z = \frac{X - \mu}{\sigma}$	$z = \frac{X - \bar{X}}{s}$
$X = z\sigma + \mu$	$X = zs + \bar{X}$

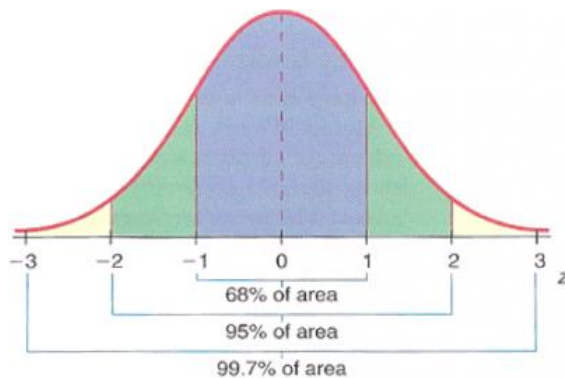
- Continuous Random Value (z has been calculated.)

Suppose X Normal distribution (μ, σ^2)

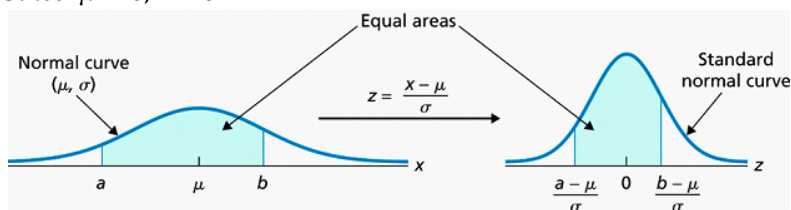
Suppose a typical score = $X = \mu$

Suppose Standard Deviation = σ

$$\text{mean } z = \frac{X - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = \frac{0}{\sigma} = 0$$



Since $\mu = 0, \sigma = 1$



- Finding probability for a normally distributed variable by transformong it onto a standard nomal variable.

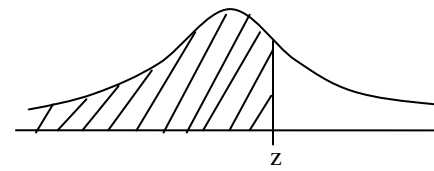
$$z = \frac{X - \mu}{\sigma} \rightarrow X = z \cdot \sigma + \mu$$

Step 1) Find the z value corresponding to a given number X or X_1 and X_2 .

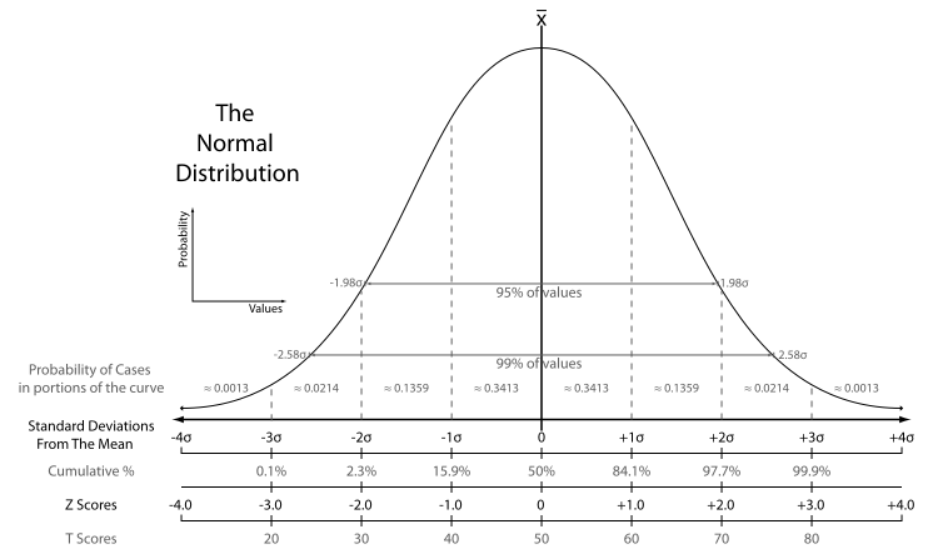
$$z = \frac{X - \mu}{\sigma}$$

Step 2) Drawing the figure and represent the area

(to the left, right, between or union area of the z)



Step 3) Find the probability or the area in the table.



<Finding probabilities (area) for a normally distributed variable by transforming it into a standard normal variable by using the formula>

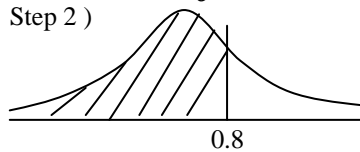
Ex 1] The average or the mean = 3.1 hours The standard deviation = 0.5

Find the percentage of less than 3.5 hours.

Step 1)

$$z = \frac{X - \mu}{\sigma} = z = \frac{3.5 - 3.1}{0.5} = 0.80$$

Step 2)



Step 3) $0.8 + .00 \rightarrow 0.7881$

Ex 2] The mean is 28 lb, and the standard deviation is 2 lb.

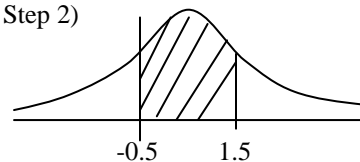
1. Between 27 and 31 lb

Step 1)

$$z_1 = \frac{X - \mu}{\sigma} = z = \frac{27 - 28}{2} = -0.5$$

$$z_2 = \frac{X - \mu}{\sigma} = z = \frac{31 - 28}{2} = 1.5$$

Step 2)



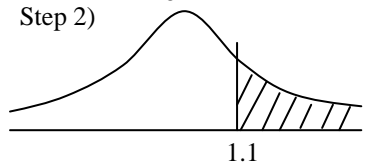
Step 3) $P(z < 1.5) - P(z < -0.5) = 0.9332 - 0.3085 = 0.6247$

2. More than 30.2 lb

Step 1)

$$z = \frac{X - \mu}{\sigma} = z = \frac{30.2 - 28}{2} = 1.1$$

Step 2)



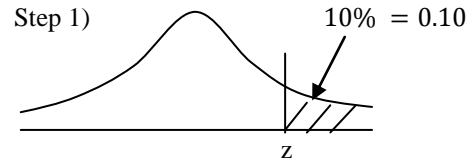
Step 3) $P(z > 1.1) = P(z < -1.1) = 0.1357 = 13.57\%$

<Finding specific data values for given percentage, using the standard normal distribution $\rightarrow X = z \cdot \sigma + \mu$ >

Ex 3) In the top 10%, the mean is 200 and the standard deviation is 20.

Find the lowest possible score to quality.

Step 1)



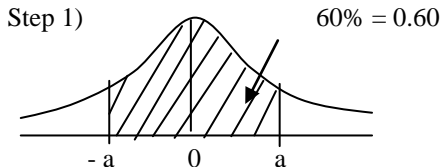
Step 2) Find the z in the table $0.10 \approx 0.003 \rightarrow -1.28 \rightarrow z = 1.28$

Step 3) $X_1 = z \cdot \sigma + \mu = 1.28 \cdot 20 + 200 = 226$

Step 4) $X > 226$

Ex4) To select in the middle 60% of the population, the mean is 120, and the standard deviation is 8. Find the upper and lower values

Step 1)



Step 2) $P(-a < z < a) = 60\% = 0.60$

$= 1 - [P(z < -a) \times 2]$ Since it's symmetric

$P(z < -a) = (1 - P(-a < z < a)) \div 2 =$

$= (1 - 0.60) \div 2 = 0.2 \rightarrow 0.205 \text{ or } 0.1977$

0.205 is the closest z value $\rightarrow -0.84 \rightarrow \pm a = \pm 0.84$

Step 3) $X_1 = z \cdot \sigma + \mu = 0.84 \cdot 8 + 120 = 126.72$

$X_2 = z \cdot \sigma + \mu = -0.84 \cdot 8 + 120 = 113.28$

Step 4) $113.28 < X < 126.72$

Ch 7

<Use the Central Limit Theorem to Solve Problems Involving Sample Means for Large Samples>

- **A Sample Distribution of Sample Means**

A distribution obtained by using the means computed from random samples of a specific size taken from a population

- **Sampling Error**

The difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population.

Ch5. Two Requirements for a Probability Distribution

1. $0 \leq P(X) \leq 1$
2. $\sum P(x) = 1$

Two Requirements for **Distribution of Sample Means**

1. $\sum P(\bar{X}) = 1$
2. $0 \leq P(\bar{X}) \leq 1$

- **Probabilities of the Distribution of Sample Means**

1. The mean of sample means will be the same as the population mean

2. *Standard Error of the Mean* $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

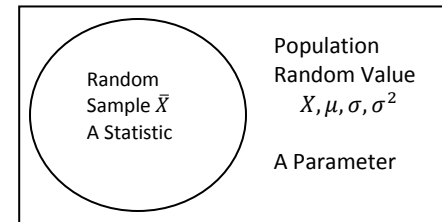
3. *the Central Limit Theorem* $z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$

σ = is the population standard deviation.

n = is the size of sample (or the number of observations in the sample).

Z = depends on the level of confidence.

- **Confidence Interval Estimate of a Parameter, say population mean**



A **Statistic** ; a characteristic or measure obtained by using the data values from a sample.

A **Parameter**; a characteristic or measure obtained by using all the data values for a specific population

< Confidence Intervals for the Mean (σ Known or $n \geq 30$) and Sample Size >

- **A Point Estimate ($\mu \approx \bar{X}$)**

A specific numerical value estimate of a parameter

The best point estimate of the population mean $\mu \approx$ the sample mean \bar{X}
(Consider μ is true value unknown best estimated value)

- **3 Properties of a Good estimator**

1. The estimator should be **an unbiased estimator**.

The expected value or the mean of the estimates obtained from samples of a given size is equal to the parameter being estimated.

$$\mu_{\bar{X}} (\text{Mean of sample means}) = \mu (\text{population mean})$$

2. The estimator should be consistent.

For a **Consistent estimator**, as sample size increase, the value of the estimator approaches the value of the parameter estimated.

$$\text{if } n (\text{sample size}) \rightarrow \infty, \bar{X} (\text{sample mean}) \rightarrow \mu (\text{population mean})$$

3. The estimator should be **a relatively efficient estimator**.

That is, of all the statistics that can be used to estimate a parameter, The relatively efficient estimator has the smallest variance.

An Interval Estimate of a Parameter is an interval or a range of values used to estimate the parameter.

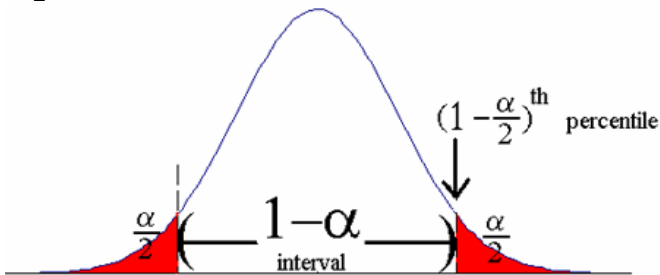
The estimate may or may not contain the value of the parameter being estimated.

The Confidence Level of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter.

A **confidence Interval** is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.

- Range of Values**

- Which may contain μ ; $a < \mu < b$ (a, b)
- It's called interval estimate of to find **confidence level**; $1 - \alpha$
- Probability of success
 - * α is the total area in both tails of the standard normal distribution curve.
 - * $\frac{\alpha}{2}$ is the area in each one of the tails.



Confidence Interval (C.I.)	$1 - \alpha$	α	$\alpha/2$
90%	100% - 90% = 10%	10% = 0.10	0.05
95%	100% - 95% = 5%	5% = 0.05	0.025
99%	100% - 99% = 1%	1% = 0.01	0.005

- 99% Confidence Interval is better than 90% or 95% because the Confidence Level is larger.**

- $|\mu - \bar{X}|$

- Confidence Interval estimate of μ
- Suppose **Confidential Level** = $1 - \alpha$
 $\alpha = 1 - \text{Confidential Level}$
- Find answer with "Z - table"
- The Maximum Error of Estimate ($E = \text{Margin of error}$)**
the maximum likely difference between the point estimate of a parameter and the actual value of the parameter.

$$\bar{X} (\text{actual parameter}) \rightarrow \mu (\text{estimate parameter})$$

When $\sigma = \text{population standard deviation (given, known)}$

$n = \text{sample size}$

$Z_{\alpha/2} = \text{only positive two-tailed critical z value}$

$$E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \frac{Z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$$

- The Confidence Interval of the Mean for a Specific α**

$$\bar{X} - Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \quad \text{or } \mu = \bar{X} \pm Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$\bar{X} - E < \mu < \bar{X} + E \quad \text{or } (\bar{X} - E, \bar{X} + E)$$

* $\bar{X} = \text{Sample mean}$

- The Minimum Sample of Size**

Required for $(1 - \alpha) \cdot 100\%$ confidence

an Interval Estimate of the μ (population mean)

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

**** **Round up to next whole number** ****

Ex) $0.23 \rightarrow 1$; $2.43 \rightarrow 3$; $4.91 \rightarrow 5$

There is a constant multiplier, usually a constant around 2 or a little higher, that comes from the distribution being used and the degree of confidence required.

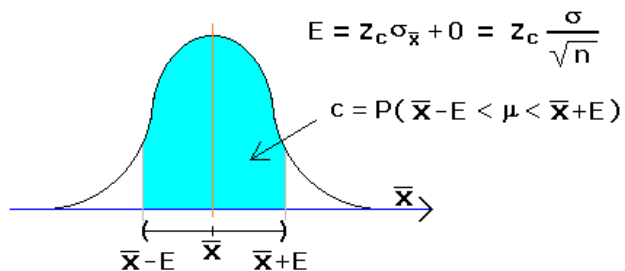
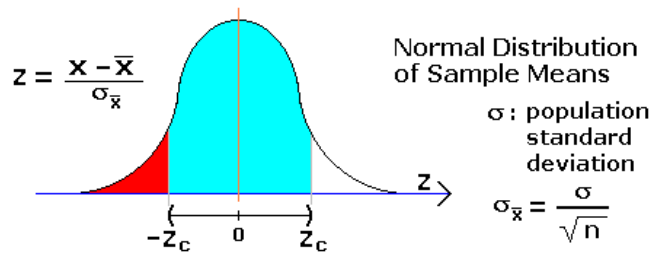
The "margin of error" is some multiplier times the standard error, and it is added to and subtracted from the mean to get the endpoints of the interval.

$$\bar{X} \pm 2 \frac{\sigma}{\sqrt{n}}$$

The sample mean is the best point estimate and so it is the center of the confidence interval.

The standard error of the mean, which is the standard deviation of the sampling distribution, is σ/\sqrt{n} .

Confidence Intervals for the Point Estimate of the Mean



P364 #11

A sample reading score of 35 students

μ **mean** = 82, σ **standard deviation** = 15

A, Find the best point estimate of the mean

$$\mu_{\bar{x}} (\text{Mean of sample means}) = \mu (\text{population mean}) \rightarrow \bar{X} = 82$$

B. Find 95% confidence interval of the mean reading scores of all students

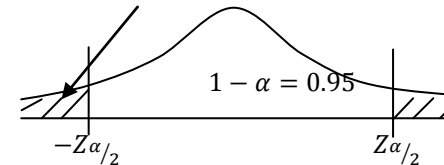
→ to find $\bar{X} - E < \mu < \bar{X} + E$, we need to find E

to find $-E$, at the first, need to find $Z_{\alpha/2}$

[Step 1]

$$\alpha = 1 - \text{Confidential Level} = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$



[Step 2]

In the table, it's less than 0.50 (50%), so look at the \ominus table
the area 0.025 $\rightarrow -1.96 \rightarrow \pm 1.96$

[Step 3] use only +1.96

$$E = \frac{Z_{\alpha/2} \cdot \sigma}{\sqrt{n}} = \frac{1.96 \cdot 15}{\sqrt{35}} \approx 4.96 \approx 5$$

* Round up to next whole number

[Step 4]

$$\bar{X} - E < \mu < \bar{X} + E$$

$$82 - 5 < \mu < 82 + 5 \quad 77 < \mu < 87$$

<How large a sample is necessary to make an accurate estimate?>

a. Depend on 3 things

E (the maximum error of estimate)

σ (the population standard deviation)

The degree of confidence (30%, 95%, 99%, etc.)

b. Use **The Minimum Sample of Size**

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

P365 #21

A university dean of students wishes to the average number of estimate hours students spend doing homework per week.

Standard deviation is 6.2 hours.

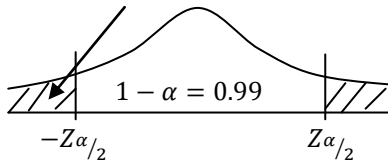
How large a sample must be selected, if he wants to be 99% confidence of finding whether the true mean differs from the sample mean by 1.5 hours?

→ To find **The Minimum Sample of Size**

[Step 1]

$$\alpha = 1 - \text{Confidential Level} = 1 - 0.99 = 0.01$$

$$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$$



[Step 2]

In the table, it's less than 0.50 (50%), so look at the \ominus table

the area 0.005 → 0.0051 → -2.57

→ 0.0049 → -2.58

$$Z \text{ given by average, thus, } \frac{(-2.57) + (-2.58)}{2} = -2.575 \rightarrow \pm 2.575$$

[Step 3] use only +2.575

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E} \right)^2 = \left(\frac{2.575 \cdot 6.2}{1.5} \right)^2 \approx 113.2805 \approx \mathbf{114 \text{ hours}}$$

* Round up to next whole number

<Confidence Intervals for the Mean (σ Unknown or $n < 30$) and Sample Size>

1. -- Characteristics of the t Distribution I

1. A normal distribution curve is **bell-shaped**

2. **The mean, median, and mode are equal** and are located at the **center** of the distribution ($\mu = 0$)

3. A normal distribution curve is **unimodal** (i.e., it has only one mode)

4. The curve is **symmetric about the mean**

5. The curve is **continuous**,

6. **The curve never touches the x axis.**

2. -- Characteristics of the t Distribution II

→ The **variance** $\sigma > 1$

→ A family of curves based on the concept of "**Degrees of Freedom**:",

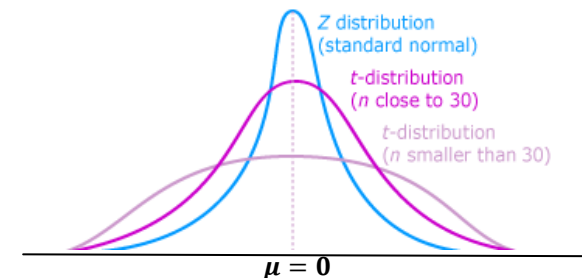
→ which is related to sample size.

$$\text{Degrees of Freedom} = d.f. \quad D.F. = n - 1$$

→ As the sample size increase, the t distribution approaches the standard normal distribution.

*At the bottom of the table where $d.f. = \infty$, the $t_{\alpha/2}$ values can be found for specific confidence intervals

- Normal distribution graph has only one curve, but t - **distribution** graph is changed & it depends on n .



- **A Specific Confidence Interval of the Mean, When σ is Unknown and Sample Size $n < 30$**

$$\bar{X} - t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) < \mu < \bar{X} + t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) \quad \text{or} \quad \mu = \bar{X} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

$$\bar{X} - E < \mu < \bar{X} + E \quad \text{or} \quad (\bar{X} - E, \bar{X} + E)$$

* \bar{X} = Sample mean

s = Sample Standard Deviation

• **How to find the Confidence Interval Estimate of μ**
When σ is Unknown and Sample Size $n < 30$

[Step 1] d.f. = $n - 1$

[Step 2] Critical Values $t_{\frac{\alpha}{2}}$ = Use the t -distribution table;

left column of d.f.'s the number & top row (Confidence Interval)
 which is given 90%, 95%, 99%, etc. $\rightarrow t_{\frac{\alpha}{2}}$

[Step 3] The Maximum Error of Estimate

$$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = \frac{t_{\alpha/2} \cdot s}{\sqrt{n}} \quad s = \sqrt{\frac{\sum(x - \bar{X})^2}{n - 1}}$$

finding sample standard deviation (p136)

* Rounding 2 decimal places (dp = 2 CF significant figure)

[Step 4] $(1 - \alpha) \cdot 100\%$

[Step 5] the Confidence Interval Estimate of μ

$$\bar{X} - E < \mu < \bar{X} + E$$

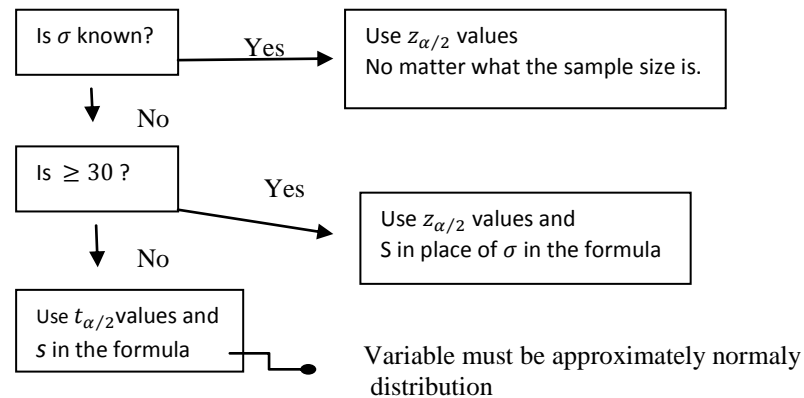
Ex) Find the $t_{\alpha/2}$ value for a 95% confidence interval, sample size is 22.

[Step 1] the d.f. = $22 - 1 = 21$ & C.I. = 95%

[Step 2] Use the t -distribution table;

left column 21 & top row (Confidence Interval) 95% $\rightarrow 2.080$

• **When to Use the z or t Distribution (p371)**



p373 #20

A sample of data; 61, 12, 6, 40, 27, 38, 93, 5, 13, 40

Construct a 98% confidence interval based on the data.

[Step1] d.f. = $n - 1 = 10 - 1 = 9$

[Step 2] Critical Values $t_{\frac{\alpha}{2}}$ = Use the t -distribution table;

left column of d.f. 9 & top row (Confidence Interval) 98% $\rightarrow 2.821 = t_{\frac{\alpha}{2}}$

[Step 3] $s = ?$

$$s = \sqrt{\frac{\sum(x - \bar{X})^2}{n - 1}} = 27.678$$

$$E = \frac{t_{\alpha/2} \cdot s}{\sqrt{n}} = \frac{(2.821) \cdot (27.678)}{\sqrt{10}} \approx 24.691 \approx 24.7$$

[Step 4] $(1 - \alpha) \cdot 100\% = 98\%$ (given)

[Step 5] the Confidence Interval Estimate of μ

$$\bar{X} = \frac{\sum x}{n} \approx 32.78 \approx 33.5$$

$$\bar{X} - E < \mu < \bar{X} + E \quad 33.5 - 24.7 < \mu < 33.5 + 24.7$$

$$8.8 < \mu < 58.2$$

p372 #11

$n = 28$, for 95% confidence interval, sample standard deviation=2

[Step1] d.f. = $n - 1 = 28 - 1 = 27$

[Step 2] Critical Values $t_{\frac{\alpha}{2}}$ = Use the t -distribution table;

left column of d.f. 27 & top row (Confidence Interval) 95% $\rightarrow 2.052 = t_{\frac{\alpha}{2}}$

[Step 3]

$$E = \frac{t_{\alpha/2} \cdot s}{\sqrt{n}} = \frac{(2.052) \cdot (2)}{\sqrt{28}} \approx 0.776 \approx 0.78$$

[Step 5] the Confidence Interval Estimate of μ

$$\bar{X} = \frac{\sum x}{n} \approx 14.3$$

$$\bar{X} - E < \mu < \bar{X} + E \quad 14.3 - 0.78 < \mu < 14.3 + 0.78$$

$$13.52 < \mu < 15.08$$

p372 #5 (H.W. #20)

99% Confidence Interval Estimate of μ , $n = 20$, $s = 2$, $\bar{X} = 16$

[Step 1] d.f. = $n - 1 = 20 - 1 = 19$

[Step 2] Critical Values $t_{\frac{\alpha}{2}} = t$ -distribution table;

left column of d.f. 19 & top row (Confidence Interval) 99% $\rightarrow 2.861 = t_{\frac{\alpha}{2}}$

[Step 3] $E = \frac{t_{\alpha/2} \cdot s}{\sqrt{n}} = \frac{(2.861) \cdot (2)}{\sqrt{20}} \approx 1.2795 \approx 1.28$

[Step 4] the Confidence Interval Estimate of μ

$$\bar{X} - E < \mu < \bar{X} + E \quad 16 - 1.28 < \mu < 16 + 1.28 \quad 15 < \mu < 17$$

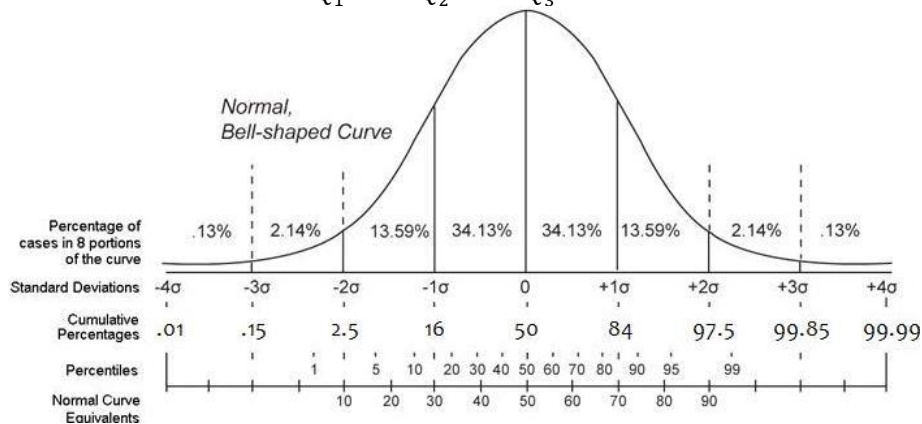
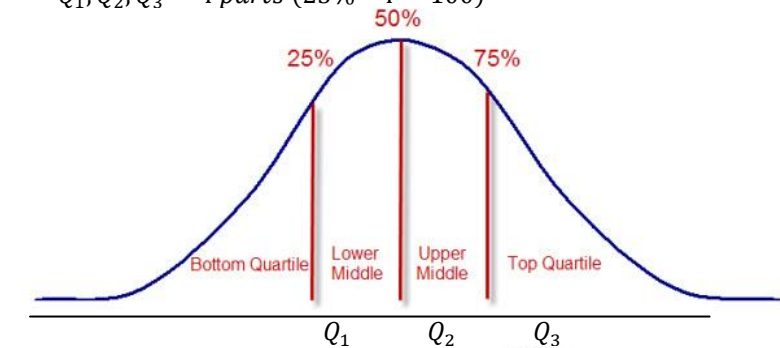
- Percentiles;** Divide the data set into 100 equal groups. (p151)

$$P_1, P_2, \dots, P_{98}, P_{99} = 100 \text{ parts}$$

- Quartiles;** Divide the distribution into 4 equal groups.

$$Q_1, Q_2, Q_3 = 4 \text{ parts } (25\% \cdot 4 = 100)$$

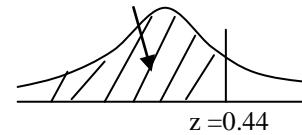
=

**p325 #30**

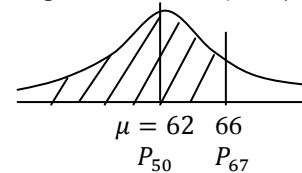
The scores are normally distributed, with a mean of 62 and a standard deviation of 8.

- a. Find 67th percentile.** (similar to H.W. #10)

[Step 1] $P_{67} = 67\% = 0.67 \rightarrow$ in the table $z = 0.44$



[Step 2] $X = z \cdot \sigma + \mu = (0.44)(8) + (62) = 65.52 \approx 66$

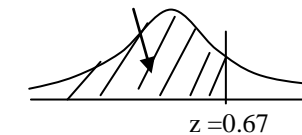


- b. Find Q_3 (3rd quartile)**

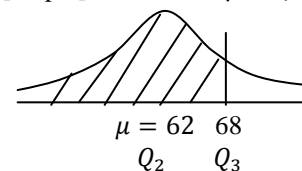
[Step 1] $Q_3 = 75\% = 0.75 \rightarrow$ in the table

0.7486 (0.0014 difference) \rightarrow closer $\rightarrow Z = 0.67$

0.7517 (0.0017 difference)

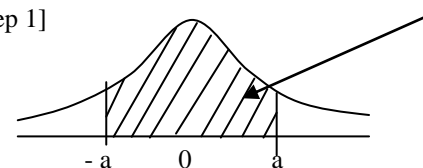


[Step 2] $X = z \cdot \sigma + \mu = (0.67)(8) + (62) = 67.36 \approx 68$



H.W. #7) $\mu = 0$, $s = 1$ Suppose $P(-c < z < c) = 0.9728$ Find $z = c = ?$

Step 1]



Step 2] $0.9728 = 1 - [P(z < -c) \times 2]$ Since it's symmetric

$$P(z < -c) = (1 - P(-c < z < c)) \div 2 =$$

$$= (1 - 0.9728) \div 2 = 0.0136 \rightarrow \text{in the table} \quad -2.21$$

P325 #12

90
150 110
190

N = 4 Sample size = 3 = n $\mu_{\bar{X}} = \mu$

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} \quad \text{Population Standard Deviation}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Ex) N=4, n = 2

2 6
4 8

$$\mu_{\bar{X}} = \mu = \frac{20}{4} = 5$$

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} = \frac{\sqrt{20}}{2} = \sqrt{5}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5}}{\sqrt{2}} \approx 1.581$$

 $\sigma_{\bar{X}}$ = Standard Error of the Mean σ = the population standard deviation. n = the size of sample (or the number of observations in the sample). Z = depends on the level of confidence. $\mu_{\bar{X}}$ = (Mean of sample means) μ = population mean \bar{X} = Sample Mean $1 - \alpha$ = confidence level E = The Maximum Error of Estimate or Margin of error $Z_{\alpha/2}$ = only positive two – tailed critical z value

The degree of confidence (30%, 95%, 99%, etc.)

d.f. = Degrees of Freedom

Ch 8

- Statistical Hypothesis**; a conjecture about a population parameter. This conjecture may or may not be true.

- Hypothesis-Testing Common Phrases (P400)**

$x \geq \#$	$x \leq \#$	$x > \#$	$x < \#$
at least no less than Is greater than or equal to	at most no more than Is less than or equal to	Is greater than Is above Is increased Is longer than Is bigger than Is higher than	Is less than Is below Is decreased or reduced from Is shorter than Is smaller than Is lower than

$x = \#$	Is equal to, is the same as, has not changed from, is exactly the same as
$x \neq \#$	Is not equal to, is not the same as, has changed from, is different from

- Two Types of Statistical Hypotheses for each situation**

	Hypothesis (= claim)	Between a parameter and a specific value or Between 2 parameters	Suggestions
H_0	<i>Null</i>	No difference	$=, \leq, \text{ or } \geq$
H_1 or H_a, H_A	<i>Alternative</i> (= research)	A difference	$\neq, < \text{ or } >$

< Traditional Method >

• Six Steps of Hypothesis- Testing

1.1. The Null Hypothesis (H_0) $\mu = k$ $\mu \leq k$ $\mu \geq k$ 2.2. The Alternative Hypothesis (H_1) $\mu \neq k$ $\mu > k$ $\mu < k$ 3.3. Test Statistics (T.S.) (or Test Value=T.V.) for z or t

$$n \geq 30 \rightarrow z \text{ table} \quad z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$n < 30 \rightarrow t \text{ table} \quad t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

4.4. Critical Values (C.V.s) using z table or t table with $d.f. = n - 1$ $\mu \neq k \rightarrow$ Two - tailed test $\alpha \rightarrow \alpha/2$ (always $\pm a$) $\mu > k \rightarrow$ Right - tailed test (always $+ a$) $\mu < k \rightarrow$ Left - tailed test (always $- a$)5.5. Decision; Always about H_0 (Null) 2 options

Reject \rightarrow (if the T.S. is inside reject region)
 Do not Reject \rightarrow (if the T.S. is inside nonreject region)

6.6. Conclusion: about the Claim that could be H_0 or H_1

Claim	H_0	Reject H_0	B
		Do Not Reject H_0	A
	H_1	Reject H_0	A
		Do Not Reject H_0	B

A: There is a enough evidence to support the claim that ~
 There is not a enough evidence to reject the claim that ~

B: There is not a enough evidence to support the claim that ~
 There is a enough evidence to reject the claim that ~

One-tail α	0.25	0.20	0.10	0.05	0.025	0.02	0.01	0.005
Two-tail α	0.50	0.40	0.20	0.10	0.05	0.04	0.02	0.01
z	0.67	0.84	1.28	1.645	1.96	2.05	2.33	2.58

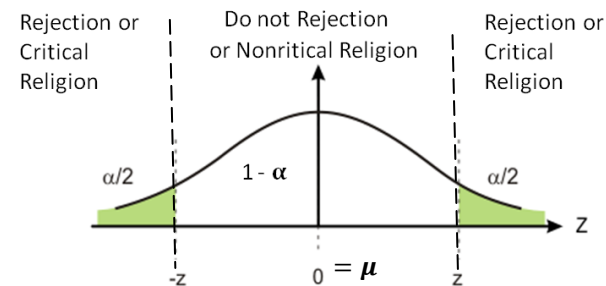
1. Statistical Test

Uses the data obtained from a sample to make a decision about whether the H_0 (null hypothesis) should be rejected.

2. Test Value

The numerical value obtained from a statistical test

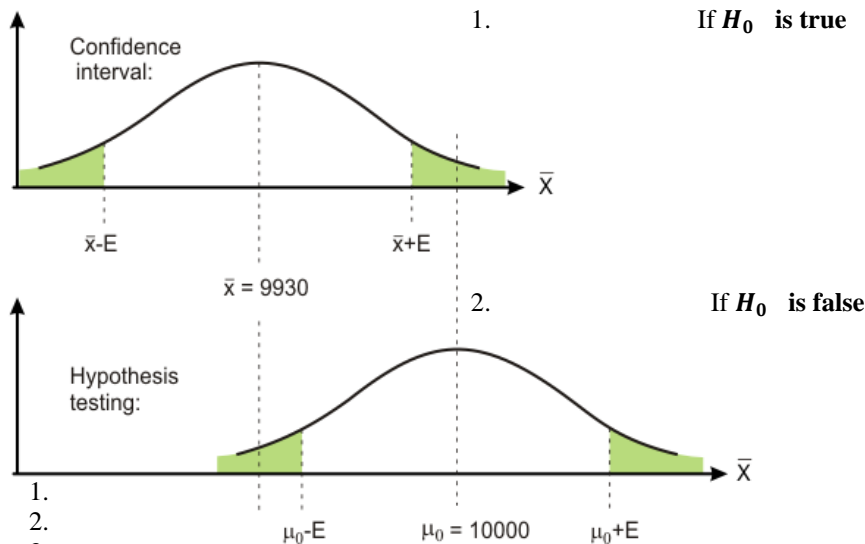
Null (H_0)	Alternative (H_1)	Tails	Graph
$\mu = k$	$\mu \neq k$	Two - Tailed Test (always $\pm a$)	
$\mu \leq k$	$\mu > k$	Right - Tailed Test (always $+ a$)	
$\mu \geq k$	$\mu < k$	Left - Tailed Test (always $- a$)	



Ex) State the null and Alternative hypotheses for each conjecture.

1. A researcher thinks that if expectant mothers use vitamin pills, the weight of the babies will increase. The average birth weight of the population is 8.6 lb.
 Ans.; Increase $\rightarrow \mu > \#$ (H_1),
 Average ~ of the population is 8.6 lb. $\rightarrow \mu \leq 8.6$ (H_0)
 $H_0: \mu \leq 8.6$ or $\mu = 8.6$ and $H_1: \mu > 8.6$
2. An engineer hypothesizes that the mean number of defects can be decreased in a manufacturing process of compact disks by using robots instead of humans for certain tasks. The mean number of defective disks per 1000 is 18.
 Ans.; Decrease $\rightarrow \mu < \#$ (H_1),
 Average ~ is 18 $\rightarrow \mu \geq 8.6$ (H_0)
 $H_0: \mu \geq 18$ or $\mu = 18$ and $H_1: \mu < 8.6$
3. A psychologist feels that playing soft music during a test will change the results of the test. The psychologist is not sure whether the grades will be higher or lower. In the past, the mean of the scores was 73.
 Ans.; Change $\rightarrow \mu \neq \#$ (H_1), Whether ~ higher or lower,
 Mean ~ was 73 $\rightarrow \mu = 73$ (H_0)
 $H_0: \mu = 73$ or $\mu = 73$ and $H_1: \mu \neq 73$

• Situation in Hypothesis Testing



- 1.
- 2.
- 3.
4. We are testing if u_0 is outside the confidence interval.
5. We are testing whether \bar{X} is in one of the rejection regions.

• Hypothesis Testing & a Jury Trial

	$H_0 \rightarrow \text{True}$ (Innocent)	$H_0 \rightarrow \text{False}$ (Not Innocent)
Reject H_0 Find guilty	Type I Error $P(\text{Type I Error}) = \alpha$	Correct Decision
Do not Reject H_0 Find not guilty	Correct Decision	Type II Error $P(\text{Type II Error}) = \beta$

- **Type I Error:** If when $\mu =, \leq,$ or $\geq,$ Reject
- **Type II Error:** If when $\mu \neq, >, \text{ or } < ,$ Do Not Reject
- **Type I Error:** There is not sufficient evidence to support the claim H_0
 (There is enough evidence to reject the claim.)
 Probability of type I error = $P(\text{type I error}) = \alpha$
- **Type II Error:** There is sufficient evidence to support the claim H_0
 (There is not enough evidence to reject the claim.)
 Probability of type II error = $P(\text{type II error}) = \beta$

* **T.S.** = Test Statistics (or Test Value=T.V.)

* **C.V.** = Critical Values (A significance level of α)

\bar{X} = Sample Mean

μ = Hypothesized population mean

σ = the population standard deviation.

n = the size of sample

< Z Test for a Mean >

Ex) Two-tailed Case

Test the claim that $\mu \neq \$24,672$, given that $\alpha = 0.01$ and the sample statistics are $n = 35$, $\bar{X} = \$25,226$, and $\sigma = \$3,251$

[Idea; $\mu \neq \$24,672 \rightarrow H_1$, 2 tailed test, $n=35 \geq 30 \rightarrow z$ table]

Step 1) $H_0: \mu = \$24,672$

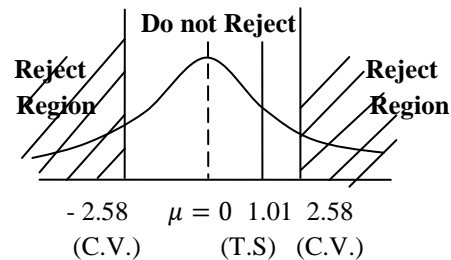
Step 2) $H_1: \mu \neq \$24,672 \leftarrow$ Claim

Step 3) $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{25226 - 24672}{3251/\sqrt{35}} = 1.01$

Step 4) **Critical Values** (C.V.s) using z table

$\alpha = 0.01 \rightarrow \alpha/2 = 0.005 \rightarrow z\text{-table} \rightarrow -2.57 \text{ \& } -2.58$

$$\frac{2.57 + 2.58}{2} = 2.575 \approx 2.58 \rightarrow \pm 2.58$$



H_0 does not reject.

Step 5) Decision always **about H_0 (Null)** 2 options

H_0 (Null) does not reject because the test value is in the nonrejection (noncritical) region

H_0 (Null) : false

H_1 (Alternative) : true

Step 6) Conclusion (about Claim) $H_1: \mu \neq \$24,672$

There is not sufficient evidence to support the claim that the average cost is different from \$24,267. ($\mu \neq \$24,672$)

*** Fail to Reject = Do not Reject

Ex) Left-tailed Case

Test the claim that the average cost is less than \$80, given that $\alpha = 0.10$ and the sample statistics are $n = 36$, $\bar{X} = \$75$, and $\sigma = \$19.2$

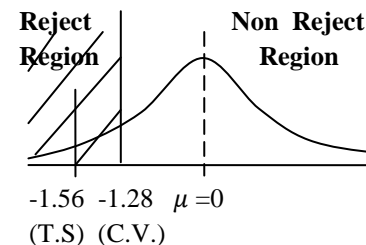
[Idea; $\mu < \$80 \rightarrow H_1$, left tailed test, $n=36 \geq 30 \rightarrow z$ table]

Step 1) $H_0: \mu \geq \$80$

Step 2) $H_1: \mu < \$80 \leftarrow$ Claim

Step 3) $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{75 - 80}{19.2/\sqrt{36}} = -1.56$

Step 4) **Critical Values** (C.V.s) using z table $\alpha = 0.10 \rightarrow z\text{-table} \rightarrow -1.28$



H_0 does reject.

Step 5) Decision always **about H_0 (Null)** 2 options

The decision is to reject H_0 (Null)

Step 6) Conclusion (about **Claim** that $\mu < \$80$)

There is sufficient evidence to support the claim that average cost is less than \$80.

Ex) Right-tailed Case

Test the claim that the average cost is more than \$42,000, given that $\alpha = 0.05$ and the sample statistics are $n = 30, \bar{X} = \$43,260$, and $\sigma = \$5,230$

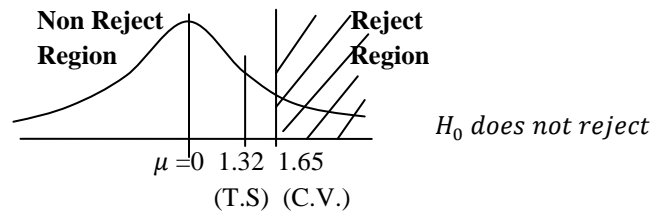
[Idea; $\mu > \$42,000 \rightarrow H_1$, right tailed test, $n=30 \geq 30 \rightarrow z$ table]

Step 1) $H_0: \mu \leq \$42,000$

Step 2) $H_1: \mu > \$42,000 \leftarrow$ Claim

Step 3) $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{43260 - 42000}{5230/\sqrt{30}} = 1.32$

Step 4) **Critical Values (C.V.s)** using z table $\alpha = 0.05 \rightarrow z\text{-table } 1.65$



Step 5) Decision always **about H_0 (Null)** 2 options

The decision is not to reject **H_0 (Null)**

Step 6) Conclusion (about **Claim**) $\mu > \$42,000$

There is not sufficient evidence to support the claim

That average cost is more than \$42,000.

- Hypothesis Test: wording of Final Conclusion**

1. Does the Claim have equality ($=$)?

Yes $\rightarrow H_0$ ($=, \leq, \text{ or } \geq$)	Reject H_0	= Cancel H_0	B
	Do not reject H_0	= Keep H_0	A
No $\rightarrow H_1$ ($\neq, <, \text{ or } >$)	Reject H_0	= Keep H_1	A
	Do not reject H_0	= Cancel H_1	B

A: There is a enough evidence to support the claim that ~
There is not a enough evidence to reject the claim that ~

B: There is not a enough evidence to support the claim that ~
There is a enough evidence to reject the claim that ~

- Hypothesis Test: wording of Final Conclusion (1)**

A: There is a sufficient evidence to support the claim that ~

B: There is not a sufficient evidence to support the claim that ~

if $\alpha = 0.05$	$H_0: \mu = 41$ $H_1: \mu \neq 41 \rightarrow \text{Claim}$	$H_0: \mu = 41 \rightarrow \text{Claim}$ $H_1: \mu \neq 41$
Test Value $z = +2.05$ -1.96 1.96 2.05 (C.V.) (C.V.) (T.S.)	A (Because the claim $\mu \neq 41$ is true)	B (Because the claim $\mu = 41$ could be false)
Test Value $z = +1.65$ -1.96 1.65 1.96 (C.V.) (T.S.) (C.V.)	B (Because the claim $\mu \neq 41$ is false)	A (Because the claim $\mu = 41$ could be true)

if $\alpha = 0.05$	$H_0: \mu = 41$ $H_1: \mu < 41 \rightarrow \text{Claim}$	$H_0: \mu = 41 \rightarrow \text{Claim}$ $H_1: \mu < 41$
Test Value $z = -1.75$ -1.75 -1.645 (T.S.) (C.V.)	A (Because the claim $\mu < 41$ is true)	B (Because the claim $\mu = 41$ could be false)
Test Value $z = -0.3$ -1.645 -0.3 (C.V.) (T.S.)	B (Because the claim $\mu < 41$ is false)	A (Because the claim $\mu = 41$ could be true)

< P-Value Method >

(P - value, p- value, or Probability Value)

The probability of getting a sample statistic (such as the mean) or a more extream sample statistic in the direction of H_1 (the alternative hypothesis) when the H_0 (the null hypothesis) is true.

The actual area under the standard normal distribution curve representing the probability of a particular sample statistic or a more extream sample statistic occurring if H_0 (the null hypothesis) is true

• Guidelines for P-values

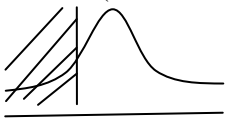
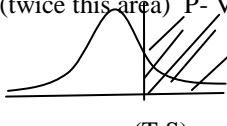
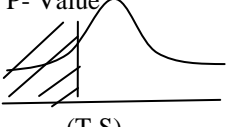
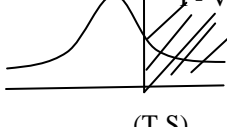
P-value & α	H_0 (Null)	The Difference is
$P - \text{value} \leq 0.01$	Reject	Highly significant
$0.01 < P - \text{value} \leq 0.05$	Reject	Significant
$0.05 < P - \text{value} \leq 0.10$		Type I error
$0.10 < P - \text{value}$	Do not reject	Not significant

*** P- Value is all about Area (Probability)

• Six Steps of Hypothesis- Testing

1. The Null Hypothesis (H_0)
2. The Alternative Hypothesis (H_1)
3. **Test Statistics** (T.S.) (or Test Value=T.V.) for z or t
by respective formulas

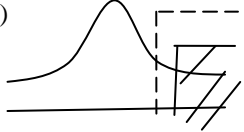
$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$
4. **Find P - value** using z or t table (about the probability) with T.S.

Type of the Test	P – V alue	Graphes
Two tailed	Twice the area to the left of the test statistic	P- Value (twice this area)  (T.S)
	Twice the area to the right of the test statistic	(twice this area) P- Value  (T.S)
Left – Tailed	To the left of the test statistic	P- Value  (T.S)
Right – Tailed	To the right of the test statistic	P- Value  (T.S)

5. Decision; **about H_0 (Null)** $\left[\begin{array}{ll} P \text{ value} \leq \alpha & \text{Reject} \\ P \text{ value} > \alpha & \text{Do not Reject} \end{array} \right]$
6. Conclusion; **about the Claim** (It is H_0 or H_1)
 A: There is a sufficient evidence to support the claim that ~
 B: There is not a sufficient evidence to support the claim that ~

Ex) A significance level of $\alpha = 0.05$ is used in testing the claim that $p > 0.25$ and the sample data result in a test statistic of $z = 1.18$.
[Idea; $p > 0.25 \rightarrow$ right tailed test]

Step1) $z = 1.18$ with z table $\rightarrow 0.1190$

Step 2)  P-value = $P(z=1.18) = 0.119$
 $\alpha = 0.05$

1.18 (T.S)

P- Value = $0.1190 > \alpha = 0.05$

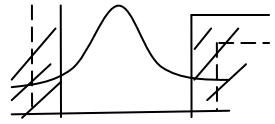
Step 3) Fail to reject H_0 (Null)

Step 4) The P-value of 0.119 is relatively large, indicating that the sample results could easily occur by chance.

Ex) A significance level of $\alpha = 0.05$ is used in testing the claim that $p \neq 0.25$ and the sample data result in a test statistic of $z = 2.34$.
[Idea; $p \neq 0.25 \rightarrow$ two- tailed test]

Step1) $z = 2.34$ with z table $\rightarrow 0.0096$

P-value = $0.0096 \cdot 2 = 0.0192$

Step 2)  $\alpha/2 = 0.05/2 = 0.025$
P-value $/2 = 0.019/2 = 0.0096$

-2.34(T.S) 2.34 (T.S)

P- Value $0.0192 \leq 0.05 = \alpha$

Step 3) Reject H_0 (Null)

Step 4) The P-value of 0.0192 is small, indicating that the sample results are not likely to occur by chance.

Ex) The claim that the average age of lifeguards in a city is greater than 24 years. A sample of 36 guards, the mean of the sample to be 24.7 years, and a standard deviation of 2 years.

Is there evidence to support the claim at $\alpha = 0.05$?

[Idea; $\mu > 24 \rightarrow H_1$, right tailed test, $n=36 \geq 30 \rightarrow z$ table]

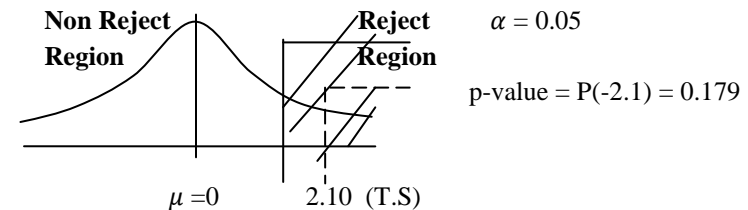
Step 1) $H_0: \mu \leq 24$

Step 2) $H_1: \mu > 24 \leftarrow$ Claim

Step 3) $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{24.7 - 24}{2/\sqrt{36}} = 2.10$

Step 4) **Find P - value** using z table

P (Right tailed $z = 2.10$) = $P(-2.10) = 0.179 \rightarrow$ symmetric



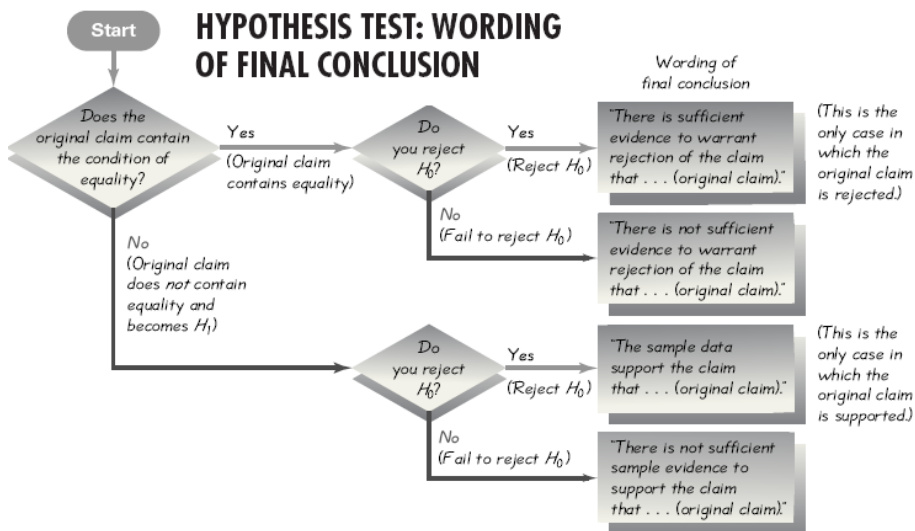
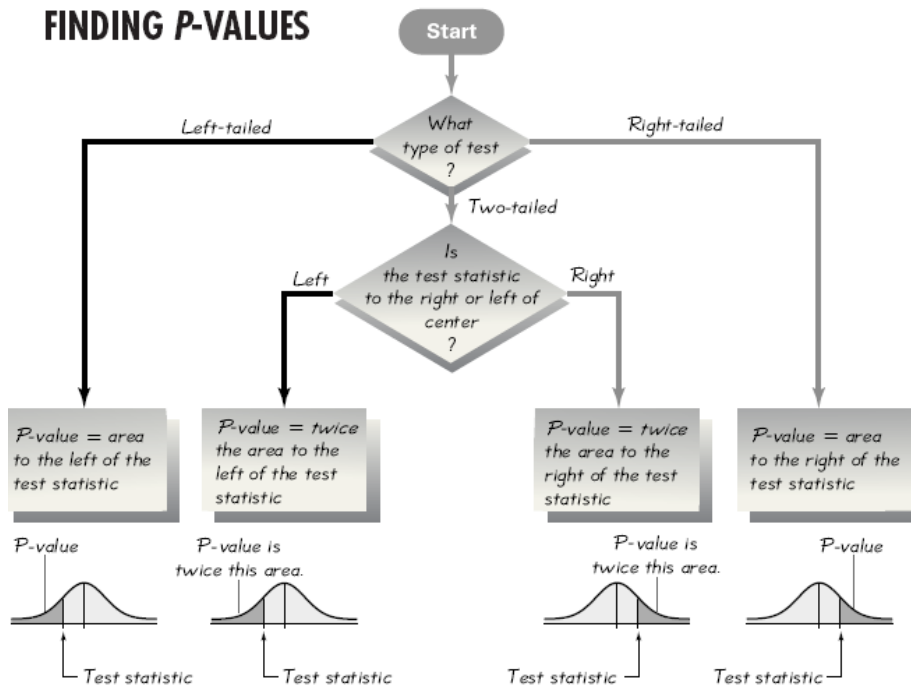
Step 5) Decision always **about H_0 (Null)**

P-value = $0.179 < \alpha = 0.05$; Reject

Step 6) Conclusion (Also about **Claim**) $\mu > 24$

There is sufficient evidence to support the claim
That average age is greater than 24 years.

FINDING P-VALUES



< t Test for a Mean >

Ex) A job director claims that the average starting salary for nurses is \$24,000. A sample of 10 nurses has a mean of \$23,450 and a standard deviation of \$400. Is there enough evidence to reject the director's claim at $\alpha = 0.05$? Assume the variable is normally distributed.
[Idea; $\mu = \$24,000 \rightarrow H_0$, 2 tailed test, $n=10 < 30 \rightarrow t$ table]

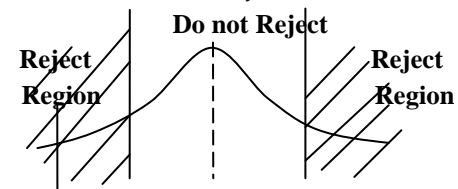
Step 1) $H_0: \mu = \$24,000 \leftarrow$ Claim

Step 2) $H_1: \mu \neq \$24,000$

Step 3) $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{23450 - 24000}{400/\sqrt{10}} = -4.35$

Step 4) **Critical Values (C.V.s)** using t table

$\alpha = 0.05$ & d.f. = $n - 1 = 10 - 1 = 9 \rightarrow t$ table ± 2.262



H_0 rejects.

-4.35	-2.262	$\mu = 0$	2.262
(T.S)	(C.V.)		(C.V.)

Step 5) Decision always about H_0 (Null) 2 options

H_0 (Null) rejects because the test value is in the rejection (critical) region

$$-4.35 < -2.262$$

Step 6) Conclusion (about Claim) $H_0: \mu = \$24,000$

There is not enough evidence to support the claims the starting salary of nurses is \$24,000.

Ex) Find the P - value when the t test value is 2.983, the sample size is 6, and the test is two-tailed.

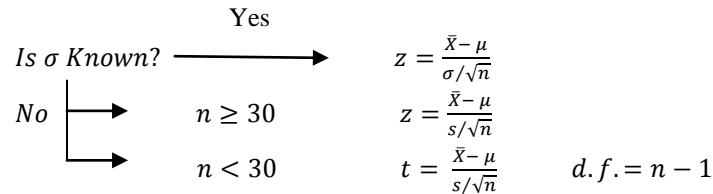
Step 1) d.f. = $6 - 1 = 5$ 2-tailed values = 2.983 with t - table

Step 2) They are 2.571 & 3.365 $\rightarrow 0.05$ & 0.02

Step 3) $0.02 < P\text{-Value} < 0.05$

Tip !

- Using the z or t test



- The claim is right \rightarrow

There is a enough evidence to support the claim ~
(or There is not a enough evidence to reject the claim that ~)

- The claim is wrong \rightarrow

There is not a enough evidence to support the claim that ~
There is a enough evidence to reject the claim that ~

- Six Steps of Hypothesis- Testing

1. The Null Hypothesis (H_0) - Prove Innocent
2. The Alternative Hypothesis (H_1) - Prove not Innocent
3. **Test Statistics** (T.S.) (or Test Value=T.V.) for z or t
4. by respective formulas

$$z^* = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{or} \quad z^* \approx \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

5. **Critical Values** (C.V.s) using z or t table
7. Decision; Always **about H_0 (Null)** 2 options

$$\left[\begin{array}{ll} \text{Reject} & \rightarrow (\text{inside critical region}) \\ \text{Do not Reject} & \rightarrow (\text{inside noncritical region}) \end{array} \right]$$
8. Conclusion; Always **about the Claim** that is H_0 or H_1

Ch 9

<Testing the Difference between 2 Means>

- Assumptions for the Test to determine the Difference between 2 Means

1. The samples must be independent of each other. (They are not related.)
2. Normally distributed, (symmetric)
3. Three Types of Formulas

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_d - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \text{if } \mu_1 - \mu_2 = 0$$

		$\mu_d = \mu_1 - \mu_2$
Two Tailed	$H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$ $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$	$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$
Left Tailed	$H_0: \mu_1 \geq \mu_2 \rightarrow \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 < \mu_2 \rightarrow \mu_1 - \mu_2 < 0$	$H_0: \mu_d \geq 0$ $H_1: \mu_d < 0$
Right Tailed	$H_0: \mu_1 \leq \mu_2 \rightarrow \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 > \mu_2 \rightarrow \mu_1 - \mu_2 > 0$	$H_0: \mu_d \leq 0$ $H_1: \mu_d > 0$

- 2) the standard deviations of variable are unknown, and $n \geq 30$ (both)

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{X}_d - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{X}_d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{if } \mu_1 - \mu_2 = 0$$

- 3) the standard deviations of variable are unknown, and $n < 30$ (one or both) \rightarrow **t - Table**

- The Confidence Interval for Difference between 2 Means
($n \geq 30$ both)

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < (\mu_1 - \mu_2) < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ex) a average hotel room rate in N.Y. is \$88.42 and in Miami is \$80.61. Two samples of 50 hotels each, the standard deviation were \$5.62 & \$4.83. At $\alpha = 0.05$, can it be concluded that there is a significant difference in the rates?

[Idea; $\$88.42 \neq \80.61 , $\mu_1 \neq \mu_2$, $\mu_1 - \mu_2 = \mu_d \neq 0$
 $\rightarrow H_1$ & Two - tailed]

Step 1) $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$

Step 2) $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0 \leftarrow$ Claim

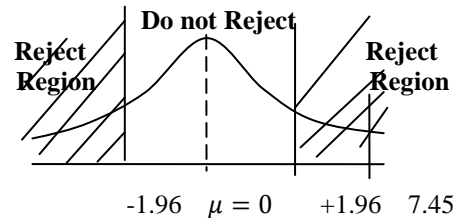
Step 3) **Test Statistics** (T.S.) (or Test Value=T.V.)

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(88.42 - 80.61) - 0}{\sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}}} = 7.45 \text{ if } \mu_1 - \mu_2 = 0$$

Step 4) **Critical Values** (C.V.s) using *z table*

$$\alpha = 0.05 \quad \alpha/2 = 0.025 \rightarrow \text{using } z \text{ table} \rightarrow \pm 1.96$$

Step 5) **Decision** always about $H_0(\text{Null}) \rightarrow H_0$ rejects.



Step 6) **Conclusion; about the Claim** $\rightarrow H_1: \mu_1 \neq \mu_2$

There is enough evidence to support the claim that

$$\$88.42 \neq \$80.61$$

Ex) There are 2 groups who left their profession within a few months after graduation (leavers) and who remained in their profession after they graduated (stayers).

Test the claim that those who stayed had a higher science grade point average than those who left. Use $\alpha = 0.01$.

Leavers	Stayers
$\bar{X}_1 = 3.16$	$\bar{X}_2 = 3.28$
$s_1 = 0.52$	$s_2 = 0.46$
$n_1 = 103$	$n_1 = 225$

[Idea; $3.16 < 3.28$, $\mu_1 < \mu_2$, $\mu_1 - \mu_2 = \mu_d < 0 \rightarrow H_1$ & left - tailed]

Step 1) $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$

Step 2) $H_1: \mu_1 < \mu_2 \rightarrow \mu_1 - \mu_2 < 0 \leftarrow$ Claim

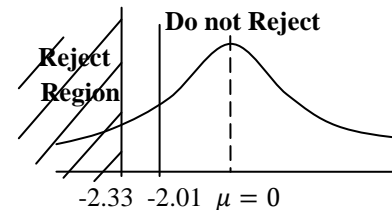
Step 3) **Test Statistics** (T.S.) (or Test Value=T.V.)

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(3.16 - 3.28) - 0}{\sqrt{\frac{0.52^2}{103} + \frac{0.46^2}{225}}} \approx -2.01 \text{ if } \mu_1 - \mu_2 = 0$$

Step 4) **Critical Values** (C.V.s) using *z table*

$$\alpha = 0.01 \rightarrow \text{using } z \text{ table} \rightarrow -2.33$$

Step 5) **Decision** always about $H_0(\text{Null}) \rightarrow H_0$ do not rejects.



Step 6) **Conclusion; about the Claim** $H_1: \mu_1 < \mu_2$

There is not enough evidence to support the claim that $3.16 < 3.28$.

<Testing the Difference between 2 Means : Small Dependent Samples>

	<i>Two Tailed</i>	<i>Left Tailed</i>	<i>Right Tailed</i>
$\mu_d = \mu_1 - \mu_2$	$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$	$H_0: \mu_d \geq 0$ $H_1: \mu_d < 0$	$H_0: \mu_d \leq 0$ $H_1: \mu_d > 0$

- Six Steps of Hypothesis- Testing**

Step 1) The Null Hypothesis (H_0)Step 2) The Alternative Hypothesis (H_1)Step 3) **Test Statistics** (T.S.) (or Test Value=T.V.) for z or t
by respective formulas* Compute the **before and after** datas = x & y (or x_1 & x_2)

x	y	$x - y$	$(x - y)^2$
\vdots	\vdots	\vdots	\vdots
$\sum x$	$\sum y$	$\sum x - y$	$\sum (x - y)^2$

$$\text{Mean} = \bar{\mu}_d = \frac{\sum \mu_d}{n}$$

$$\text{Standard Deviation} = s_d = \sqrt{\frac{\sum \mu_d^2 - \frac{(\sum \mu_d)^2}{n}}{n - 1}}$$

$$\text{Test Value} = t = \frac{\bar{\mu}_d - \mu_0}{s_d / \sqrt{n}}$$

Step 4) **Critical Values** (C.V.s) using t table, degree of freedom = $n - 1$ Step 5) Decision; Always **about H_0 (Null)** 2 options

$\left[\begin{array}{ll} \text{Reject} & \rightarrow (\text{inside critical region}) \\ \text{Do not Reject} & \rightarrow (\text{inside noncritical region}) \end{array} \right]$

Step 6) Conclusion; Always **about the Claim** that is H_0 or H_1 **A:** There is a sufficient evidence to support the claim that ~**B:** There is not a sufficient evidence to support the claim that ~

Ex) A dietitian wishes to see if a person's cholesterol level will change if the diet is supplemented by a certain mineral.

Six subjects were pretested, and then they took the mineral supplement for a 6-week period. The results are shown in the table.

Can it be concluded that the cholesterol level has been changed at $\alpha = 0.10$? (*It's nomally distributed*)

Subject	1	2	3	4	5	6
Before (x_1)	210	235	208	190	172	244
After (x_2)	190	170	210	188	173	228

[Idea; $\$88.42 \neq \80.61 , $\mu_1 \neq \mu_2$, $\mu_1 - \mu_2 = \mu_d \neq 0$
 $\rightarrow H_1$ & *Two - tailed*]

Step 1) $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$ Step 2) $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0 \leftarrow \text{Claim}$ Step 3) **Test Statistics** (T.S.) (or Test Value=T.V.)

x	y	$x - y$	$(x - y)^2$
210	190	20	400
235	170	65	4,225
208	210	-2	4
190	188	2	4
172	173	-1	1
244	228	16	256
$\sum x$	$\sum y$	$\sum x - y$	$\sum (x - y)^2$
1,259	1,159	100	4,890

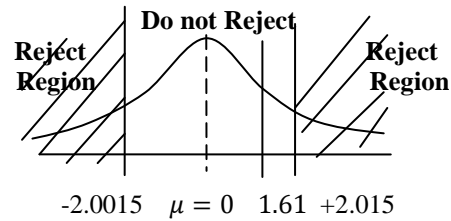
$$\text{Mean} = \bar{\mu}_d = \frac{\sum \mu_d}{n} = \frac{100}{6} \approx 16.7$$

$$s_d = \sqrt{\frac{\sum \mu_d^2 - \frac{(\sum \mu_d)^2}{n}}{n - 1}} = \sqrt{\frac{4890 - \frac{(100)^2}{6}}{6 - 1}} \approx 25.4$$

$$\text{Test Value} = t = \frac{\bar{\mu}_d - \mu_0}{s_d / \sqrt{n}} = \frac{16.7 - 0}{25.4 / \sqrt{6}} \approx 1.61 \quad \text{if } \mu_1 - \mu_2 = 0$$

Step 4) **Critical Values** (C.V.s) using t table
 $\alpha = 0.10 \rightarrow$ using t table $d.f. = 6 - 1 = 5 \rightarrow \pm 2.015$

Step 5) **Decision** always **about** $H_0(\text{Null}) \rightarrow H_0$ *do not rejects*.



Step 6) **Conclusion; about the Claim** $\rightarrow H_1: \mu_1 \neq \mu_2$

There is not enough evidence to support the claim that
\$88.42 \neq \$80.61, .

Ch 10

<Linear Correlation Coefficient r >

• **A scatter plot** : a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x , and the dependent variable y .

• the Correlation Coefficient r

Computed from the sample data measures the strength and direction of a linear relationship between two variables.

The symbol for the sample correlation coefficient is r .

The symbol for the population correlation coefficient is ρ (Greek letter rho)

• Range of Values for the Correlation Coefficient $(-1 \leq r \leq 1)$

-1	-0.5	0	0.5	+1
Perfect	Strong	No Linear	Strong	Perfect
Negative	Negative	Relationship	Positive	Positive
L.C.C.	L.C.C.	L.C.C.	L.C.C.	L.C.C.

* L.C.C.=Linear Correlation Coefficient

< Line of Best Fit >

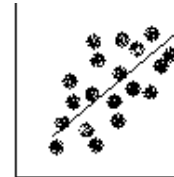
$$\hat{y} = y' = a + bx,$$

$a = y$ - **intercept** & b is the slope of the line

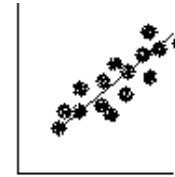
*In Algebra, $y = mx + b$, slope - intercept form,

In statistic, $y' = a + bx$, *Becareful with "b"*

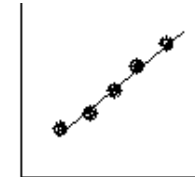
• Relationship Between the Correlation Coefficient and the Line of Best Fit



(a) $r = 0.50$



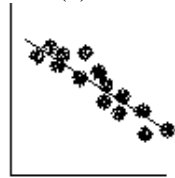
(b) $r = 0.90$ (strong)



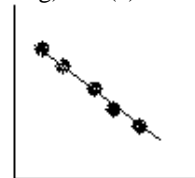
(c) $r = 1.00$ (perfect)



(d) $r = -0.50$

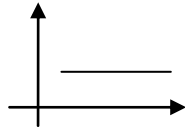


(e) $r = -0.90$ (strong)



(f) $r = -1.00$ (perfect)

*** When $r = 0$, the line is horizontal



• **Four Steps of Hypothesis- Testing**

Step1)

x	y	$x \cdot y$	x^2	y^2
\vdots	\vdots	\vdots	\vdots	\vdots
$\sum x$	$\sum y$	$\sum x \cdot y$	$\sum x^2$	$\sum y^2$

Step 2) **the Correlation Coefficient r**

$$r = \frac{n(\sum x \cdot y) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

n is the number of data pairs

Step3) the values of a and b

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum x \cdot y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum x \cdot y) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Step 4) When r is significant, **The Regression Line Equation**

$$\hat{y} = y' = a + bx$$

Ex) In a study of age and systolic blood pressure of 6 randomly selected subjects.

Subject	A	B	C	D	E	F
Age x	43	48	56	61	67	70
Pressure y	128	120	135	143	141	152

Step 1)

x	y	$x \cdot y$	x^2	y^2
43	128	5504	1849	16384
48	120	5760	2304	14400
56	135	7560	3136	18225
61	143	8723	3721	20449
67	141	9447	4489	19881
70	152	10640	4900	23104
$\sum x$	$\sum y$	$\sum x \cdot y$	$\sum x^2$	$\sum y^2$
345	819	47,634	20,399	112,443

Step 2) **the Correlation Coefficient r**

$$r = \frac{n(\sum x \cdot y) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{6(\sum 47,634) - (\sum 345)(\sum 819)}{\sqrt{[6 \cdot 20,399 - 345^2][6 \cdot 112,443 - 819^2]}} = 0.897$$

* The correlation coefficient suggests a Strong Positive Relationship between age and blood pressure.

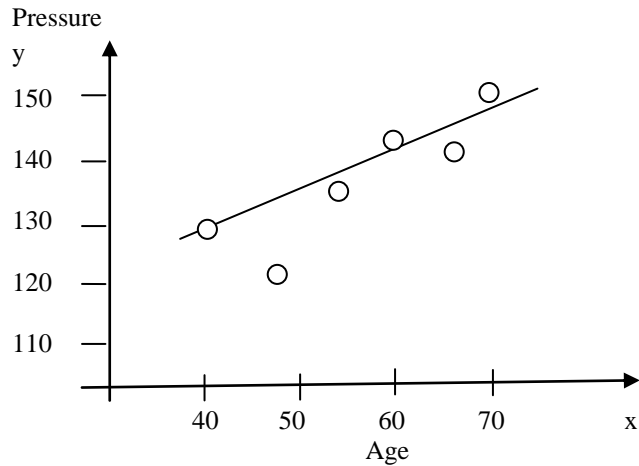
Step 3) the values of a and b

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum x \cdot y)}{n(\sum x^2) - (\sum x)^2} = \frac{819 \cdot 20,399 - 345 \cdot 47,634}{6 \cdot 20,399 - 345^2} = 81.048$$

$$b = \frac{n(\sum x \cdot y) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{6 \cdot 47,634 - 345 \cdot 819}{6 \cdot 20,399 - 345^2} = 0.964$$

Step 4) When r is significant, **The Regression Line Equation**

$$\hat{y} = y' = a + bx = 81.048 + 0.964x$$



Plot (43, 128), (48, 120), (56, 135), (61, 143), (67, 141), & (70, 152)

Ex) Using the equation, predict the blood pressure for a person who is 50 years old.

[Idea: 50 years old \rightarrow a x value]

$$\hat{y} = y' = a + bx = 81.048 + 0.964x = 81.048 + 0.964 \cdot (50) = 129.248$$

In other words, the predicted systolic blood pressure for a 50-years-old person is 129.

- **The t Test for the Correlation coefficient**

$$t = r \sqrt{\frac{n-2}{1-r^2}} \text{ with degrees of freedom equal to } n-2$$

Source

Elementary Statistics - A Brief Version

<http://rchsbowman.wordpress.com/category/statistics/statistics-notes/page/7/>

<http://statistics.laerd.com/statistical-guides/normal-distribution-calculations.php>

<http://rchsbowman.wordpress.com/2009/12/01/statistics-notes-%E2%80%93-the-standard-normal-distribution-2/>

<http://www.anlyzemath.com/statistics/mutually-exclusive.html>