Statistics deal with variables, and relies on how the numbers are chosen, and how the statistics are interpreted.
Descriptive statistics only use numbers to describe the data. E.g: Average height of teenagers.
Inferential statistics go beyond the data, by having tests and estimations.
E.g: Hypothesis that 8th graders are happier than 9th graders.

Case: Teacher asks 10 students sitting on the front row of their latest test scores. He concludes from their report that the class did well.
Sample: 10 students
Population: The whole class
Potential problem: The students sitting on the front row generally pays more attention, which may not be the same for the rest of the class. Thus, the sample can be higher than the population

Three main types of data in statistics: Quantitative, Ranked, and Qualitative.
Quantitative only consist of numbers
Ranked represents the standings of the data in a population.
Qualitative data can be classified as: categorical data. (Dont contain numbers)

There are also 3 main measurement types: nominal, ordinal, and interval / ratio.
nominal measurement is usually used for classification. (like male coded as 1, female as 2)
ordinal measurement is used for rankings. (1 star, 2, 3, 4, 5)
Interval / ratio measurement a variable / number appears in the data.

Discrete variables is a finite number (usually the count of something) // Num of students in class
Continuous variables can be infinite. (usually measuring something) // Height of a person.

- Descriptive Statistics
- Inferential Statistics
- Types of Data
  - Quantitative
  - Ranked
  - Qualitative
- Types of Measurement
  - Nominal
  - Ordinal
  - Interval
- Types of Variables
  - Discrete variables
  - Continuous variables.

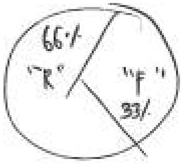
Graphing Data   Qualitative = Count   Quantitative = Act. Values

Graphing Qualitative variables:

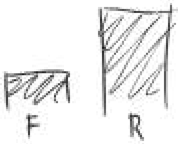
Name	Freq	Relative Freq
"F"	10	0.33
"R"	20	0.66
Total	30	1

10/30
20/30

Frequency Table



Pie Chart.
(Recommended only for small num of categories)



Bar Charts

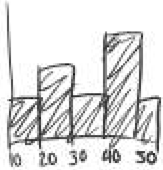
Dont use 3D bar charts for data representation (May cause difficulties and errors)
Bar charts values on the y-axis must start from 0.

Graphing Quantitative variables

3 | 2 3 3 7
2 | 0 0 1 1 1
1 | 2 2 4 4 4
0 | 6 9
(Stem and Leaf)

=>

Two Parts:
1 1 | 4 |
  | 3 | 7
3 2 1 | 0 | 2 3 3

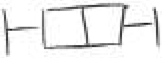


(Histogram)

Lower Limit	Upper Limit	Count	Cumulative Count
10	19	10	10
20	30	20	30

(Frequency Polygons)
Usually graphed using line graphs

Box plots



Data: 14, 15, 16, 16, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 19, 19, 19, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 24, 29

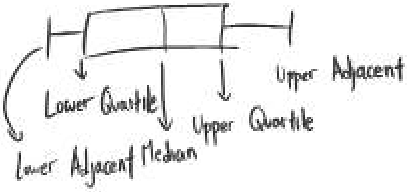
n = 31

H-Spread = Upper Quartile - Lower Quartile

median = $\frac{31+1}{2} = 16 \Rightarrow 19$

(lower hinge)
Lower Quartile = $\frac{31+1}{4} = 8 \Rightarrow 17$

Upper Quartile = $\frac{3(31+1)}{4} = 24 \Rightarrow 20$
(upper Hinge)



Step = 1.5 x H-Spread

Upper Inner Fence = Upper Quartile + 1 step
Lower Inner Fence = Lower Quartile - 1 step

Upper Adjacent : largest value below Upper Inner Fence
Lower Adjacent: Smallest value above Lower Inner Fence

H-Spread = 20 - 17 => 3

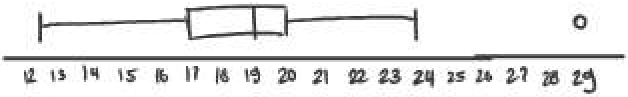
Step = 1.5 · 3 => 4.5

Upper Inner Fence = 20 + 4.5 => 24.5
Lower Inner Fence = 17 - 4.5 => 12.5

Upper Adjacent = 24
Lower Adjacent = 14

Lower Quartile = 17 Median = 19 Upper Quartile = 20 Upper Adjacent = 24 Lower Adj = 12.5
Outer Fences = Quartiles ± 2 steps
Upper Outer Fences = 20 + 9 => 29
Lower Outer Fences = 17 - 9 => 8

Outside value = values beyond inner fence, but not outer fence
Far Out value = values beyond outer fences.



Outside Value = 29

14, 15, 16, 16, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 19, 19, 19, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 24, 29

$n = 31$

median $= \frac{(31+1)}{2} = 16 \Rightarrow 19$ Lower Quartile $= \frac{(31+1)}{4} = 8 \Rightarrow 17$

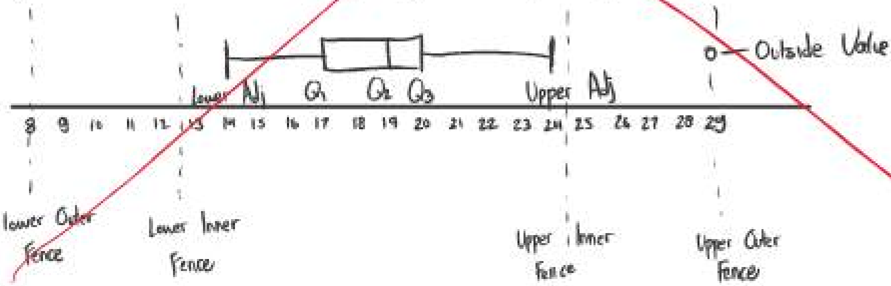
Upper Quartile $= \frac{3(31+1)}{4} = 24 \Rightarrow 20$ H-Spread $= 20 - 17 \Rightarrow 3$

Step $= 3 \cdot 1.5 \Rightarrow 4.5$

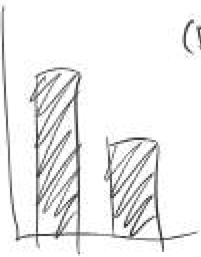
Lower Inner Fence $= 17 - 4.5 \Rightarrow 12.5$ Upper Inner Fence $= 20 + 4.5 = 24.5$

Lower Adjacent $= 14$ Upper Adjacent $= 24$

Lower Outer Fence $= 17 - 9 = 8$ Upper Outer Fence $= 20 + 9 = 29$

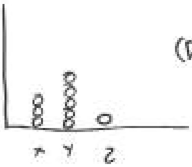


Quantitative Representation



(Bar Graphs)

(Line Graphs)



(Dot plots)

Data = $\boxed{3, 3, 7, 8, 8}$ 10, 11, 12, 15, 18
n = 10

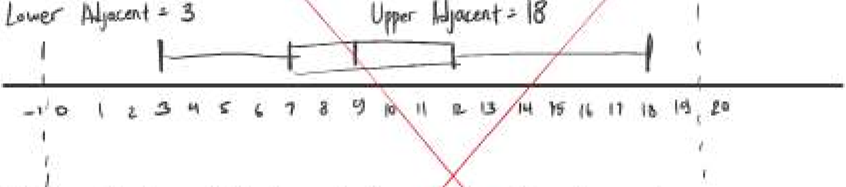
median = $\frac{(10+1)}{2}$ = 5.5 => 9 Lower Quartile = $\frac{(10+1)}{4}$ = 2.75 = 3 => 7

Upper Quartile = $\frac{3(10+1)}{4}$ = 8.25 => 8 => 12 H Spread = 12 - 7 => 5
Step = 5 × 1.5
= 7.5

Lower Inner Fence = 7 - 7.5
= -0.5

Upper Inner Fence = 12 + 7.5
= 19.5

Lower Adjacent = 3 Upper Adjacent = 18



Note: If the calculated position has .25 or .75, we round it up.
e.g. position 1.25 => Take position 1
position 2.75 => Take position 3

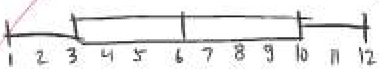
Data = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
n = 12

median = $\frac{(12+1)}{2}$ = 6.5 => 6.5 Lower Quartile = $\frac{(12+1)}{4}$ = 3.25 => 3

Upper Quartile = $\frac{3(12+1)}{4}$ = $\frac{39}{4}$ = 9.75 => 10 => 10

H - Spread = 10 - 3 => 7 Lower Inner Fence = 3 - 10.5 => -7.5
Step = 7 × 1.5 = 10.5 Upper Inner Fence = 10 + 10.5 = 20.5

Lower Adjacent = 1 Upper Adjacent = 12



Summary

Lower Quartile = median of the $\frac{\text{lower}}{\text{upper}}$ half of the values not including the median
Upper Quartile =

$$\text{Data} = \underbrace{\overbrace{3, 3, 6, 8, 10}^{(n=5)}}_{Q_1} \; \underbrace{\overbrace{14, 16, 16, 19, 24}^{(n=5)}}_{Q_3}$$

$n = 10$

$\text{median} = \frac{10+1}{2} = 5.5 \Rightarrow 12$

$Q_1 = \frac{5+1}{2} = 3 \Rightarrow 6$

$\text{H- Spread} = 16 - 6 \Rightarrow 10$

$Q_3 = \frac{5+1}{2} = 3 \Rightarrow 16$

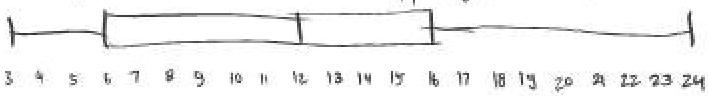
$\text{Step} = 1.5 * 10 = 15$

Lower Inner Fence $= 6 - 15 = -9$

Upper Inner Fence $= 16 + 15 \Rightarrow 31$

Lower Adjacent $= 3$

Upper Adjacent $= 24$



$$\text{Data} = \underbrace{1, 2, 3, 4, 5, 6}_{(n=6)} \; \underbrace{7, 8, 9, 10, 11, 12}_{(n=6)}$$

$\text{median} = \frac{(12+1)}{2} = 6.5 \Rightarrow 6.5$

$Q_1 = \frac{(6+1)}{2} = 3.5 \Rightarrow 3.5$

$Q_3 = \frac{(6+1)}{2} = 3.5 \Rightarrow 9.5$

$\text{H- Spread} = 9.5 - 3.5 = 6$

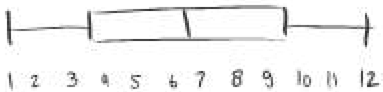
$\text{Step} = 6 * 1.5 = 9$

Lower Inner Fence $= 3.5 - 9 = -5.5$

Lower Adj $= 1$

Upper Inner Fence $= 9.5 + 9 = 18.5$

Upper Adj $= 12$



$$\text{Data} = \overbrace{14, 15, 16, 16, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18}^{n=15} \; 19 \; \underbrace{19, 19, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 24, 29}_{n=15}$$

$n = 31$

$\text{median} = \frac{31+1}{2} = 16 \Rightarrow 19$

$Q_1 = \frac{15+1}{2} = 8 \Rightarrow 17$

$Q_3 = \frac{15+1}{2} = 8 \Rightarrow 20$

$\text{H- Spread} = 20 - 17 \Rightarrow 3$

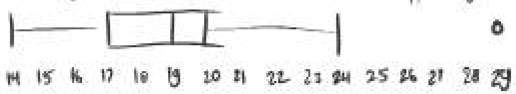
$\text{Step} = 3 \cdot 1.5 = 4.5$

Lower Inner Fence $= 17 - 4.5 = 12.5$

Upper Inner Fence $= 20 + 4.5 = 24.5$

Lower Adj $= 14$

Upper Adj $= 24$



Exercise 1

1. Data = 62, 65, 68, 70, 73, 75, 75, 78, 81, 83, 84, 85, 87, 89, 92, 95, 96, 98, 100

Stem & leaf

6	2 5 8
7	0 3 5 5 8
8	1 3 4 5 7 9
9	2 5 6 8
10	0

2. Dataset = $\underbrace{55, 60, 62, 63, 65, 66, 68}_{n=7}$, 70, $\underbrace{72, 75, 77, 78, 80, 85, 88}_{n=7}$
(n = 15)

a. Five number summary:

minimum = 55
maximum = 88

median = $\frac{15+1}{2} = 8 \Rightarrow 70$

$Q_1 = \frac{7+1}{2} = 4 \Rightarrow 63$ $Q_3 = \frac{7+1}{2} = 4 \Rightarrow 78$

b. Draw boxplot:

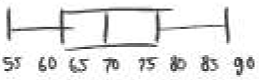
H-Spread = 78 − 63
= 15

Step = 15 × 1.5
= 22.5

Lower Inner Fence = 63 − 22.5
= 40.5

Upper Inner Fence = 78 + 22.5
= 100.5

Lower Adjacent = 55
Upper Adjacent = 88



c. No outliers.

Descriptive statistics usually use few numbers to describe a distribution (Like Average)

There are 3 different ways in defining the center of distribution:

- Balance scale
- smallest absolute deviation
- Smallest squared deviation

} Measures of central tendency

Measures of central tendency = Mean, Median, Mode.
(Most frequent element)

$$\text{Trimean} = \frac{(Q_1 + 2Q_2 + Q_3)}{4}$$

(n = 15)

e.g Data: 6, 9, 12, 12, 14, 14, 14, 15, 16, 18, 18, 18, 18, 19, 19, 20, 20, 21, 21, 21, 22, 22, 22, 23, 28, 28, 29, 32, 33, 33, 37

(n = 31)

$$\text{median} = \frac{31+1}{2} = 16 \Rightarrow 20$$

$$Q_1 = \frac{(15+1)}{2} = 8 \Rightarrow 15 \qquad Q_3 = \frac{(15+1)}{2} = 8 \Rightarrow 23$$

$$\text{Trimean} = \frac{(15 + 2(20) + 23)}{4} = \frac{78}{4} = 19.5$$

Trimean is a better version of the median. Median is used to avoid outliers.

Geometric mean = $(\text{product } x)^{1/x}$

e.g. data = 1, 10, 100

$$\text{geometric mean} = (1 \cdot 10 \cdot 100)^{1/3} = (1000)^{1/3} = 10$$

Geometric mean is appropriate in measuring the average of rates.

Example Question:

Year	Return
1	13%
2	22%
3	12%
4	-5%
5	-13%

Average annual rate of return = ?

=> 1.13, 1.22, 1.12, 0.95, 0.87

$$= (1.13 \cdot 1.22 \cdot 1.12 \cdot 0.95 \cdot 0.87)^{1/5}$$
$$= 1.05$$

Average annual rate of return = 5%

Aside from using to face outliers, trimmed mean can be used. A trimmed 10% means that the data is trimmed by 5% from top and bottom.

e.g: 6, 8.1, 8.3, 9.1, 9.9 Find the trimmed mean of 40%

$= \cancel{6}, 8.1, 8.3, 9.1, \cancel{9.9}$

$\Rightarrow \frac{8.1 + 8.3 + 9.1}{3} = 8.5$

Bivariate data involves 2 variables, describing the relationship of those variables.

Univariate analysis: Analysis of one variable
Bivariate analysis: Analysis of exactly two variables
Multivariate analysis: Analysis of more than two variables

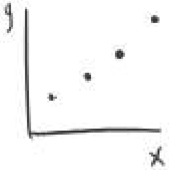
} Variables are directly related

e.g of bivariate analysis:

Calorie Intake X	Weight Y (Lbs)
3500	250
2000	225
1500	110
2250	145
4500	380

// Describes the relationship between calorie intake and weight.

Bivariate Analysis representation:
(Scatter plot)



Regression Analysis: Based on the data, it can give an equation for the line or curve.

Correlation coefficient tells if variables are related to one another or not.

The pearson correlation represents the strength of the linear relationship between 2 variables.
unit: ρ if population
r if sample
It can range from -1 to 1. -1 → Perfect negative linear relationship
0 → No linear relationship
1 → Perfect positive linear relationship

$$r = \frac{\Sigma((x-\bar{x})(y-\bar{y}))}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}}$$

Σ = sum $\bar{x}$ = Average of x
$\bar{y}$ = Average of y

e.g: n=10

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
17	94	1.4	14.3	20.02	1.96	204.44
13	73	-2.6	-6.7	17.42	6.76	44.89
12	59	-3.6	-20.7	74.52	12.96	428.49
15	80	-0.6	0.3	-0.18	0.36	0.09
16	93	0.4	13.3	5.32	0.16	176.89
14	85	-1.6	5.3	-8.48	2.56	28.09
16	66	0.4	-13.7	-5.48	0.16	187.69
16	79	0.4	-0.7	-0.28	0.16	0.49
18	77	2.4	-2.7	-6.43	5.76	7.29
19	91	3.4	11.3	38.42	11.56	127.69
$\bar{x}$ = 15.6	$\bar{y}$ = 79.7			Σ = 134.8	Σ = 42.4	Σ = 1206.1

pearson coefficient = $\frac{134.8}{\sqrt{(42.4)(1206.1)}} = \frac{134.8}{\sqrt{51138.64}} = 0.596$ => Shows moderate correlation between x and y.

$$\frac{\Sigma((x-\bar{x})(y-\bar{y}))}{\sqrt{(x-\bar{x})^2(y-\bar{y})^2}}$$

Exercise 2

1. Trimean of data: 10, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48, 50
$n = 15$

median $= \frac{15+1}{2} = 8 \Rightarrow 30$ $Q_1 = \frac{7+1}{2} = 4 \Rightarrow 18$

$Q_3 = \frac{7+1}{2} = 4 \Rightarrow 42$

Trimean $= \frac{(18 + 30(2) + 42)}{4} \Rightarrow 30$

2. Geometric mean: Year 1: +5% Year 2: +10% Year 3: −3% Year 4: +6%

Values = 1.05, 1.1, 0.97, 1.06
geometric mean $= (1.05 \cdot 1.1 \cdot 0.97 \cdot 1.06)^{1/4}$
$= 1.044$
= 4.4% every year on average

3. ~~65~~, 70, 72, 75, 80, 85, 90, 92, 95, ~~100~~ $n = 10$, Find the trimmed mean of 20%.

$\frac{70 + 72 + 75 + 80 + 85 + 90 + 92 + 95}{8} = 82.375$

Probability

Probability of a single event = $\frac{\text{Possible outcomes}}{\text{Total outcomes}}$

e.g. Probability of getting an ace upon choosing a random card: $\frac{4}{52} = \frac{1}{13}$

Probability of A and B = $P(A) * P(B)$ (Independent events)

e.g. If you flip 2 coins, what is the probability of getting 2 heads?

$$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

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

Permutations = $\frac{n!}{(n-r)!}$ n = population   r = sample taken   combinations = $\frac{n!}{(n-r)!\,r!}$

e.g. There are 4 candy colours, how many different ways can 2 candies be selected?

$$\frac{4!}{(4-2)!} = \frac{24}{2} = 12 \text{ ways (If arrangement matters)}$$

$$\frac{4!}{(4-2)!\,2!} = \frac{24}{4} = 6 \text{ ways (If arrangement doesn't matter)}$$

Permutations and combination are used when only one type of object is used.

Multiplication rule is used when generating all possible combinations, but only choosing one from each category

e.g. 3 soups, 6 entrees, and 4 deserts.

Possible combination: $3 \times 6 \times 4 = 72$ ways

(1 from each category)

Binomial probabilites = $\frac{N!}{x!(N-x)!} \pi^x q^{n-x}$

n = num of trials

x = num of successes

π = probability of getting a success in a single trial

q = probability of a failure

n − x = num of failures

Toss coin 12 times. Probability of 0-3 heads?

$$P(0) = \frac{12!}{0!(12-0)!} \cdot \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^{12-0}$$

$$= \frac{\cancel{12!}}{\cancel{12!}} \cdot 1 \cdot \left(\frac{1}{2}\right)^{12} = \frac{1}{4096}$$

Binomial Probability = $\frac{n!}{x!(n-x)!} \cdot \pi^x \cdot (1-\pi)^{n-x}$

Independent trials. Outcome of one trial doesn't affect the results of other outcomes.

Toss coin 12 times. Probability of 0 - 3 heads.

$P(0) = \frac{\cancel{12!}}{0!\cancel{(12-0)!}} \cdot \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^{12} = .0002.4^{-4} \Rightarrow 0.0002$

$P(1) = \frac{12!}{\cancel{1!}(12-1)!} \cdot \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^{12-1} = 6 \cdot \left(\frac{1}{2}\right)^{11} = .0029^{-3}$

$\doteq 0.0029$

$P(2) = \frac{12!}{2!(12-2)!} \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^{12-2} = 66 \cdot \frac{1}{4} \cdot \frac{1}{1024} = 0.016$

$P(3) = \frac{12!}{3!(12-3)!} \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^{12-3} = 220 \cdot \frac{1}{8} \cdot \frac{1}{512} = 0.054$

Total $P = 0.0002 + 0.0029 + 0.016 + 0.054 \Rightarrow 0.073$

Binomial probabilities used when finding the probability of a number of successes on a fixed number of trials.

mean of binomial distribution = $(\pi)(N)$ π = Probability of success

variance = $N\pi(1-\pi)$

mean of binomial distribution = $(1/2)(12) = 6$

variance = $(12)(1/2)(1-1/2) = 3$

Poisson distribution is used when an interval of an action is given. And the probability is asked for a different interval value.

$\frac{e^{-\lambda}\lambda^x}{x!}$ x = number of successes

λ = rate of success / Interval provided

mean of daily phone calls = 8. Probability of 11 phone calls? $\frac{e^{-8}8^{11}}{11!} = 0.072$

Binomial Distribution = $\frac{n!}{x!(n-x)!} \cdot \pi^x \cdot (1-\pi)^{n-x}$

Poisson Distribution = $\frac{e^{-\lambda}\lambda^x}{x!}$ λ = given interval

x = num of success

Multinomial Distribution

$$P = \frac{n!}{(n_1!)(n_2!)\ldots} P_1^{n_1} \cdot P_2^{n_2}$$

n = number of events
n_1 = number of time it occurs
P_1 = probability of that event.

Player A win = 0.4 Player B win = 0.35 Draw = 0.25

If the chess players played 12 games, what is the probability that player A win 7 games, B win 2 games, 3 games draw.

$$\frac{12!}{7!2!3!} \cdot 0.4^7 \cdot 0.35^2 \cdot 0.25^3 = 0.0248$$

Formulas:

Binomial Distribution = $\frac{n!}{x!(n-x)!} \cdot \pi^x \cdot (1-\pi)^{n-x}$

n = num of occurences (Total)
x = Success π = P(success)

Poisson Distribution = $\frac{e^{-\lambda}\lambda^x}{x!}$ λ = Given Frequency x = num target

Multinomial Distribution = $\frac{n!}{n_1!n_2!\ldots n_k!} \cdot P_1^{n_1} P_2^{n_2} \ldots P_k^{n_k}$

n = occurences (Total) n_1 = Occurence Target
P_1 = Probability of Target

Poisson Distribution : $\frac{e^{-i} i^x}{x!}$ i = given value
x = target value

Hypergeometric Distribution is used to calculate the probability of an action without replacement.

e.g. getting 3 cards from a deck what is the probability of getting 2 aces.
(without replacement) → Dependent Trials

Data = 600, 470, 170, 430, 300

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

$$\text{Variance} = \frac{(600-394)^2 + (470-394)^2 + (170-394)^2 + (430-394)^2 + (300-394)^2}{5}$$
$= 21704$

$$\text{Standard deviation} = \sqrt{21704}$$
$= 147.32277$
$= 147$

Difference in calculations:
- When calculating for variance, divide by N (Population)
- When calculating for variance, divide by N-1 (Sample)

$$\text{Hypergeometric Distribution} = \frac{(\text{num of ways to chose the objects that result in success})(\text{same but failures})}{(\text{num of ways to choose the total n from the population})}$$

A company has 30 employees, 12 are software developers while the rest aren't. A manager randomly chooses 8 employees. What is the probability of exactly 5 of the employees are developers.

$$\text{Developers} = \frac{12!}{(12-5)!\,5!}$$
$= 792$

$$\text{Non-Developers} = \frac{18!}{(18-3)!\,3!}$$
$= 816$

$$\text{All} = \frac{30!}{(30-8)!\,8!}$$
$= 5852925$

$$\text{Probability} = \frac{792 \times 816}{5852925}$$
$= 0.11$

Exercise 3

1. 8 people and arrange 4 in order. How many possible arrangements?

$$\frac{8!}{(8-4)!} = \frac{8!}{4!} = 5 \times 6 \times 7 \times 8$$

$$= 1680 \text{ ways}$$

2. You have 7 books, you want to choose 4. How many different ways can you select the books?

$$\text{Combinations} = \frac{7!}{(7-4)!4!} = 35 \text{ ways}$$

3. A bag contains 10 red balls and 15 blue balls. If you randomly select 5 balls without replacement, what is the probability that exactly 3 of the balls are red.

$$\text{red balls}: \frac{10!}{(10-3)!3!} = 120 \qquad \text{Blue balls}: \frac{15!}{(15-2)!2!} = 105 \qquad \text{All balls} = \frac{25!}{(25-5)!5!} = 53130$$

$$\text{Probability} = \frac{120 \times 105}{53130} = 0.237$$

$$\frac{\overbrace{\frac{10!}{(10-3)!3!}}^{a} \cdot \overbrace{\frac{15!}{(15-2)!2!}}^{b}}{\underbrace{\frac{25!}{(25-5)!5!}}_{c}}$$

a = number of ways to choose 3 red balls from 10

b = number of ways to choose 2 blue balls from 15

c = number of ways to choose 5 balls from 25

Hypothesis Testing steps: Specify the null hypothesis
Specify the a value (significance level), which is usually around 0.05 and 0.1
Compute the p value, then compare it with the a level.

Type I error = false positive conclusion
Type II error = false negative conclusion
Probability of making Type I error = a (alpha)
Probability of making Type II error = b (beta)

eg Coronavirus.
False positive = The test states that you have the virus, but you don't
False negative = The test states that you do not have the virus, but you actually do.

One and two tailed tests are used to identify the relationship between statistical variables.
- A one tailed test is used to identify the relationship between the variables in one direction
- A two tailed test checks whether the relation is ~~in~~ any direction or not.

eg of one tailed test: Checking if class A students scored higher than class B
two tailed test: Checking if class A scored higher or lower than class B

A one sample T-test compares the mean of the sample data with a value.

(< 25)

mean = 30   new = 25   n = 100

$SD = \sqrt{100 \times 0.7 \times 0.3}$
$= 4.583$

$z\text{-score} = \frac{24.5 - 30}{4.583}$
$= -1.2$

$\Rightarrow 0.1151$

Exercise 4

1. Year 1 = 10% → 1.1
Year 2 = 15% → 1.15
Year 3 = -5% → 0.95
Year 4 = 8% → 1.08
Year 5 = 12% → 1.12

$(1.1 \cdot 1.15 \cdot 0.95 \cdot 1.08 \cdot 1.12)^{1/5} = 1.077$
$\Rightarrow 7.7\%$

2. Group A = 7, 9, 12, $\boxed{13}$, 14, 15, 16
Group B = 5, 7, 8, $\boxed{10}$, 12, 15, 18

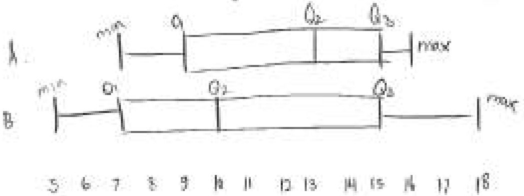
a. minimum: A = 7
B = 5

median = $\frac{7+1}{2} = 4 \Rightarrow$ A = 13
B = 10

$Q_1 = \frac{3+1}{2} = 2 \Rightarrow$ A = 9
B = 7

$Q_3 = \frac{3+1}{2} = 2 \Rightarrow$ A = 15
B = 15

maximum: A = 16
B = 18



b. Group A has a higher median.

c. H-Spread = A: 6
B: 8

Step = A: 9
B: 12

$\left.\begin{array}{l} 9-9=0 \\ 7-12=-5 \end{array}\right\}$ Lower Inner Fence

Lower Adj: 7
5

No values seem to exceed the inner fences, thus there are no outliers for both groups

$\left.\begin{array}{l} 15+9=24 \\ 15+12=27 \end{array}\right\}$ Upper Inner Fence

Upper Adj: 16
18

3. A card is drawn from a deck of 52 cards, and a coin is flipped. What is the probability of drawing king and flipping a tail.

$P = \frac{4}{52} \times \frac{1}{2} = \frac{4}{104} = \frac{1}{26} \Rightarrow 0.038 \Rightarrow 3.8\%$

4. Department x: 12, 14, 17, 19, 21, 24, 26, 28, 30, 32
Department y: 13, 16, 18, 20, 23, 25, 27, 29, 31, 33

Stem & Leaf =

2 6	3	1 3
8 6 4 1	2	0 3 5 7 9
9 7 4 2	1	3 6 8

5. Calculate the probability of getting 3 heads when a coin is flipped 5 times.

$$\frac{n!}{(n-x)!\,x!} \cdot \pi^{x} \cdot (1-\pi)^{n-x} \Rightarrow \frac{5!}{(5-3)!\,3!} \cdot \left(\frac{1}{2}\right)^{3} \cdot \left(1-\frac{1}{2}\right)^{3-3}$$

$$= 10 \cdot \frac{1}{8} \cdot \frac{1}{4}$$

$$= \frac{10}{32} \Rightarrow \frac{5}{16}$$

6. A player has a free throw success rate of 80%. If the player takes 15 free throws, what is the probability of making at least 12 shots.

$$p(12) = \frac{15!}{(15-12)!\,12!} \cdot \left(\frac{4}{5}\right)^{12} \cdot \left(\frac{1}{5}\right)^{3} \Rightarrow 0.25$$

$$p(13) = \frac{15!}{(15-13)!\,13!} \cdot \left(\frac{4}{5}\right)^{13} \cdot \left(\frac{1}{5}\right)^{2} \Rightarrow 0.23$$

$$p(14) = \frac{15!}{(15-14)!\,14!} \cdot \left(\frac{4}{5}\right)^{14} \cdot \left(\frac{1}{5}\right)^{1} \Rightarrow 0.13$$

$$p(15) = \cancel{\frac{15!}{(15-15)!\,15!}} \cdot \left(\frac{4}{5}\right)^{15} \cdot \left(\frac{1}{5}\right)^{0} \Rightarrow 0.035$$

Total Probability $= 0.25 + 0.23 + 0.13 + 0.035$
$= 0.645$

7.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
2	10	-4	-10	40	16	100
4	15	-2	-5	10	4	25
6	20	0	0	0	0	0
8	25	2	5	10	4	25
10	30	4	10	40	16	100
$\bar{x}=6$	$\bar{y}=20$			$\Sigma=100$	$\Sigma=40$	$\Sigma=250$

$$p = \frac{100}{\sqrt{(40)(250)}} \Rightarrow 1.$$ The pearson coefficient is 1 which indicates perfect linear relationship.

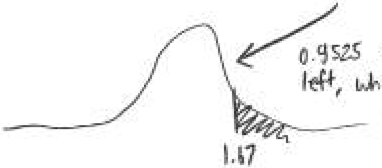
mean = 65 standard deviation = 9 (scores)

Probability of scores < 54

$\frac{53.99 - 65}{9} = -1.223 \Rightarrow -1.22$ for z-score = $-1.22 \Rightarrow 0.1112$

at least 80: $\frac{80-65}{9} = 1.67$ for z-score = $1.67 \Rightarrow 0.9525$

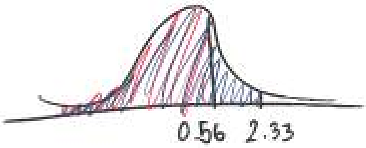
$P = 1 - 0.9525$
$= 0.0475.$



0.9525 is the area on the left, while we're trying to find the area on the right.

between 70 and 86: $\frac{70-65}{9} = \frac{5}{9} = 0.56 \Rightarrow 0.7123$ $\frac{86-65}{9} = \frac{21}{9} = \frac{7}{3} = 2.33 \Rightarrow 0.9901$

$P = 0.9901 - 0.7123$
$= 0.2778$



A survey indicates that 30% of people prefer coffee over tea. If you randomly select 10 people, what is the probability that fewer than 25 people prefer coffee. Use z-table.

mean = $0.3 \times 100 \Rightarrow 30$

SD formula: $\sqrt{n \cdot p \cdot (1-p)} \longrightarrow \sqrt{100 \cdot 0.3 \cdot 0.7} \Rightarrow 4.58$

For $(P < 25) \longrightarrow \frac{24.5 - 30}{4.58} = -1.2 \Rightarrow 0.1151$

Continuity Correction:
$x \leq 45 \longrightarrow x < 45.5$
$x < 45 \longrightarrow x < 44.5$
$x \geq 45 \longrightarrow x > 44.5$
$x > 45 \longrightarrow x > 45.5$

Exercise 5.

1. 70, 85, 78, 90, 88. Standard deviation = ?

$$\text{mean} = \frac{70 + 85 + 78 + 90 + 88}{5} \Rightarrow 82.2$$

$70 - 82.2 = -12.2$
$85 - 82.2 = 2.8$
$78 - 82.2 = -4.2$
$90 - 82.2 = 7.8$
$88 - 82.2 = 5.8$

$$\text{Variance} = \frac{(-12.2)^2 + (2.8)^2 + (-4.2)^2 + (7.8)^2 + 5.8^2}{5}$$

$$= 53.76$$

$$\text{Standard deviation} = \sqrt{53.76} \Rightarrow 7.33$$

2. 30% of people prefer coffee over tea. If you randomly select 100 people, what is the probability that fewer than 25 people prefer coffee? $(x < 25) \rightarrow P < 24.5$

$$\text{mean} = 0.3 \times 100 = 30$$

$$SD = \sqrt{100 \times 0.3 \times 0.7} = 4.583$$

$$z\text{-score} = \frac{24.5 - 30}{4.583} = -1.2$$

$$P(-1.2) = 0.1151$$

3. You are conducting an experiment with 100 trials, and the probability of success is 0.4. Find the probability of at least 45 successes will occur. $(x \geq 45) \rightarrow P > 44.5$

$$\text{mean} = 0.4 \times 100 = 40$$

$$SD = \sqrt{100 \times 0.4 \times 0.6} = 4.899$$

$$z\text{-score} = \frac{44.5 - 40}{4.899} = 0.92$$

$$P(z\text{-value}) = 0.8212$$

Since greater than, $1 - 0.8212 = 0.1788$

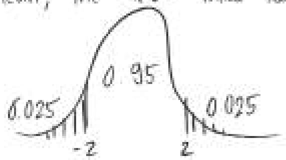
$$\text{Standard deviation} = \sqrt{\text{population} \cdot (P\ \text{success}) \cdot (P\ \text{failure})}$$

One tailed test and two tailed tests.

When to use one tailed and two tailed tests?

Average mass of potato chips = 100g ⟶ Null Hypothesis
An employee believes that the mean is not 100g ⟶ Alternative Hypothesis.

Whenever the alternative hypothesis does not state that a value is greater or less than the mean, the two-tailed test is used.



Shaded = reject region ⟶ 0.025, 0.025
Unshaded = Fail to reject region. ⟶ 0.95

z values = critical values

The employee conducts a test at 95% confidence level. $C = 0.95$

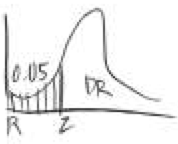
$\alpha = 1 - C$
$= 1 - 0.95$
$= 0.05$

If the calculated z-value is in the reject region, the null hypothesis will be rejected.

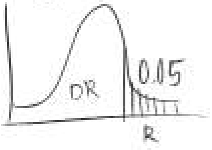
In t-tests, the provided mean or the value calculated from the sample data provided represents the null hypothesis. While the alternative statement represents the alternative hypothesis.

One tailed test: If the alternative hypothesis that the value must be less than a value, the left tailed test is used.
Otherwise, if it states that the value is greater than a value, the right tailed test is used.

Left tailed test:



Right tailed test:



(The Null Hypothesis)
R = Reject
DR = Don't reject

Confidence Intervals

95% confidence → One tailed:

$$1.6 \rightarrow \overset{0.05}{\downarrow} 0.95$$

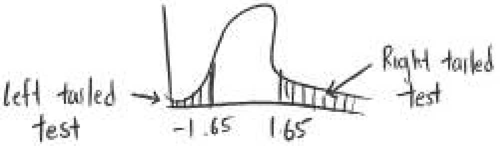
critical value = 1.65

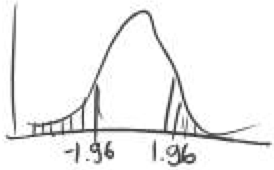
Two tailed:

$$\frac{1 + 0.95}{2} = 0.975$$

$$1.9 \rightarrow \overset{0.06}{\downarrow} 0.975$$

critical value = 1.96





A factory has a machine that dispenses 80 ml. A worker believes that the average amount is not equal to 80. He tested himself and obtained an average of 78 ml with an SD of 2.5 using 40 samples. Confidence level = 95%.

Null Hypothesis = The average is = 80 → Two tailed test used.
Alternative Hypothesis : The average is ≠ 80.

Given mean = 78
Given SD = 2.5

$\alpha = 1 - 0.95$
$= 0.05$

Critical value $= \dfrac{1 + 0.95}{2}$
$= 0.975$ (z-score)
$\Rightarrow 1.96$

Intervals = -1.96, 1.96

$$z \text{ score} = \frac{78 - 80}{2.5/\sqrt{40}} = -5.06$$

The resulting z-score is in the reject area. Thus, we reject the null hypothesis.

NOTES!

When solving tailed tests, if the sample ≥ 30, the z-test is used. Otherwise, the t-test is used for smaller samples.

(Area under the curve up to the critical value)

z-test: two tailed $= \dfrac{(1 + \text{confidence})}{2}$ One tailed = confidence

t-test: two tailed $= \dfrac{(1 - \text{confidence})}{2}$ One tailed = 1 − confidence

Degree of freedom = n − 1

A company manufactures batteries with an average lifespan of 2 years. An employee thi that this number is less. By using 10 samples, he obtained an average lifespan of 1.8 years with an SD of 0.15. Confidence level = 99%

Null Hypothesis = Batteries have an lifespan of 2 years
Alternative Hypothesis = Batteries have an lifespan of less than 2 years → left tailed test

As n < 30, t test is used.
Degree of freedom = 10 - 1
= 9

Critical value = 1 - 0.99
= 0.01
For DOF = 9 & Significance level = 0.01,
Critical value = 2.821

$$t\text{-score} = \frac{1.8 - 2}{0.15/\sqrt{10}}$$

$$= -4.2$$

As t score < Critical value, we can reject the null hypothesis.

Exercise 6

1. A company claims their light bulbs last 1000 hours on average. A sample of 10 bulbs yields the following timespans:

950, 960, 970, 980, 1020, 1030, 990, 1010, 1000, 995

Test whether the mean lifespan differ significantly from 1000, using a = 0.05.

Perform two tailed t- test.

$$\text{mean} = \frac{950 + 960 + 970 + 980 + 1020 + 1030 + 990 + 1010 + 1000 + 995}{10}$$

$= 990.5$

$$SD = \sqrt{\frac{(950-990.5)^2 + (960-990.5)^2 + (970-990.5)^2 + (980-990.5)^2 + (1020-990.5)^2 + (1030-990.5)^2 + (990-990.5)^2 + (1010-990.5)^2 + (1000-990.5)^2 + (995-990.5)^2}{9}}$$

$= 25.87$

$t = \frac{0.05}{2}$

$= 0.025$

For $t = 0.025$ and Degree of Freedom = 9, Critical value = 2.262

Critical values = 2.262, −2.262

$$t\text{-value} = \frac{990.5 - 1000}{25.87/\sqrt{10}}$$

$= -1.16$

As the t- value is within the range of the critical values, we fail to reject the null hypothesis.

Null Hypothesis = Bulbs last 1000 hours on average.

Alternative Hypothesis = Bulbs does not last 1000 hours on average.

For population less than 30, the results can't be determined for the whole population. Thus, it is considered as a sample, and on the variance stage, it is divided by n-1.

For One tailed tests, the critical value obtained, the sign is not changed. (Unlike two tailed tests)

7. Conduct a paired t-test to determine if the training program significantly reduced weight. $a = 0.05$.

Use Left-tailed test.

$$\text{mean} = \frac{-3 -3 -5 -2 -3 -3 -3 -3}{8} = -3.125$$

$$SD = \sqrt{\frac{(-3+3.125)^2 + (-3+3.125)^2 + (-5+3.125)^2 + (-2+3.125)^2 + (-3+3.125)^2 + (-3+3.125)^2 + (-3+3.125)^2 + (-3+3.125)^2}{8-1}} = 0.835$$

Degree of freedom $= 8-1 = 7$

For degree of freedom $= 7$, and $a = 0.05$:
critical value $= 1.895$

$$t\text{-value} = \frac{-3.125}{0.835/\sqrt{8}} = -10.6$$

The t-value is beyond the critical value. Thus, we reject the null hypothesis.

3. t-value for independent t-test: $\dfrac{\bar{x}_A - \bar{x}_B}{\sqrt{\dfrac{S^2_A}{n_a} + \dfrac{S^2_B}{n_b}}}$

$\bar{x}_A$ = Mean A
$\bar{x}_B$ = Mean B

S^2_A = $(\text{Standard Dev})^2$ (a)
S^2_B = $(\text{Standard Dev})^2$ (B)

n_a = Population A
n_b = Population B

$$\frac{8-6}{\sqrt{\frac{2^2}{25} + \frac{2.5^2}{25}}} = 3.12$$

Degree of Freedom $= 25 + 25 - 2 = 48$

df $= 48$, $a = 0.05$. Critical value $= 1.679$.

The tailed test is a right tailed test, because we're checking if it leads to a greater weight loss.
Thus, as the t-statistic > Critical value, we reject the null hypothesis.

T-statistic formulas:

Regular: $\dfrac{\text{New mean} - \text{Initial mean}}{\text{SD of new mean}/\sqrt{n}}$

Independent: $\dfrac{\text{mean A} - \text{mean B}}{\sqrt{\dfrac{(SD_A)^2}{n_A} + \dfrac{(SD_B)^2}{n_B}}}$

Paired: $\dfrac{\text{Difference in mean}}{\text{SD of mean}/\sqrt{n}}$

One Way ANOVA

Data :

①	②	③
1	2	2
2	4	3
5	2	4.

①

H_0 : There is no difference in the mean of 3 groups

H_a : There is a difference in the mean of 3 groups

$a = 0.05$

(If F-value > Critical value, we reject)

② Degrees of freedom between groups = 3 - 1 = 2 (No of groups - 1) → Numerator

Degrees of freedom within groups = 9 - 3 = 6 (No of elements - no of groups)
↳ Denominator

To get the critical value, the F Distribution table is used.

For $DF_{between} = 2$, and $DF_{within} = 6$, Critical value = 5.14

③ $\bar{x}_1 = \frac{1+2+5}{3} = 2.67$ $\quad \bar{x}_2 = \frac{2+4+2}{3} = 2.67$ $\quad \bar{x}_3 = \frac{2+3+4}{3} = 3$

Grand mean $= \frac{2.67 + 2.67 + 3}{3} = 2.78$

Sum of squares Total = $\sum(x - \bar{\bar{x}})^2$ ↳ Grand mean

$= (1-2.78)^2 + (2-2.78)^2 + (5-2.78)^2 + (2-2.78)^2 + (4-2.78)^2 + (2-2.78)^2 + (2-2.78)^2 + (3-2.78)^2 + (4-2.78)^2 = 13.6$

Sum of squares within = $(1-2.67)^2 + (2-2.67)^2 + (5-2.67)^2 + (2-2.67)^2 + (4-2.67)^2 + (2-2.67)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 = 13.34$

Sum of squares between = 13.6 - 13.3 => 0.23

④ Mean of squares between = $\frac{0.23}{2}$ → 0.12

Mean of squares within = $\frac{13.34}{6}$ → 2.22

⑤ $F = \frac{MS_{Between}}{MS_{within}} = \frac{0.12}{2.22}$ → 0.05 Critical value = 5.14

As F-value < Critical value, we fail to reject the null hypothesis.

① Hypothesis ② DOF ③ Sum of Squares ④ mean of squares ⑤ F

1.

A	B	C
15	20	25
16	22	27
14	19	26
15	21	28
17	20	24

① H_0: There is no difference in plant growth for the different fertilizers used

H_A: There is a difference.

② $DF_{between} = 3 - 1 \rightarrow 2$

$DF_{within} = 15 - 3 \longrightarrow 12$

③ $\bar{x}_A = \dfrac{15+16+14+15+17}{5} = 15.4$

$\bar{x}_B = \dfrac{20+22+19+21+20}{5} = 20.4$

$\bar{x}_C = \dfrac{25+27+26+28+24}{5} = 26$

$$\text{Grand Mean} = \frac{15.4+20.4+26}{3} = 20.6$$

$$S_{total} = (15-20.6)^2 + (16-20.6)^2 + (14-20.6)^2 + (15-20.6)^2 + (17-20.6)^2 + (20-20.6)^2 + (22-20.6)^2 + (19-20.6)^2 + (21-20.6)^2 + (20-20.6)^2 + (25-20.6)^2 + (27-20.6)^2 + (26-20.6)^2 + (28-20.6)^2 + (24-20.6)^2$$

$$= 31.36 + 21.16 + 43.56 + 31.36 + 12.96 + 0.36 + 1.96 + 2.56 + 0.16 + 0.36 + 19.36 + 40.96 + 29.16 + 54.76 + 11.56$$

$$= 301.6$$

$$S_{within} = (15-15.4)^2 + (16-15.4)^2 + (14-15.4)^2 + (15-15.4)^2 + (17-15.4)^2 + (20-20.4)^2 + (22-20.4)^2 + (19-20.4)^2 + (21-20.4)^2 + (20-20.4)^2 + (25-26)^2 + (27-26)^2 + (26-26)^2 + (28-26)^2 + (24-26)^2$$

$$= 0.16 + 0.36 + 1.96 + 0.16 + 2.56 + 0.16 + 2.56 + 1.96 + 0.36 + 0.16 + 1 + 1 + 0 + 4 + 4$$

$$= 20.4$$

$$S_{between} = 301.6 - 20.4 = 281.2$$

④ Average of sum of squares: Between: $\dfrac{281.2}{2} = 140.6$ Within: $\dfrac{20.4}{12} = 1.7$

⑤ $F = \dfrac{140.6}{1.7} = 82.7$

Critical value = 2.8068

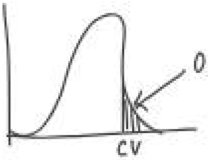
82.7 > 2.8068. Thus, we reject the null hypothesis.

	1	2	3	4	5
1	23	16	14	19	28
2	20	20	20	20	20

Ho: Equal frequencies
HA: Unequal frequencies

$a = 0.05$

Chi Square



0.05 (Reject Null Hypothesis)

$df = n-1$
$= 5-1$
$= 4$

For $df = 4$ and $a = 0.05$, Critical value $= 9.49$

$$\text{Chi Square} = \frac{3^2}{20} + \frac{(-4)^2}{20} + \frac{(-6)^2}{20} + \frac{(-1)^2}{20} + \frac{8^2}{20} = \frac{126}{20} \Rightarrow 6.3$$

(substract 1 from 2)

As $6.3 < 9.49$, we cannot reject the null hypothesis.

Martial status	Middle School	High School	Bachelor's	Master's	PHD	T
Never Married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

Ho = There is no relation between the martial status and educational qualification
HA = There is a relation between the martial status and educational qualification

$$\text{Expected values} = \frac{\text{Row total} \times \text{column total}}{\text{All total (300)}}$$

	MS	HS	Bach	Master's	PHD
NM	11.7	27	25.2	16.2	9.9
M	19.5	45	42	27	16.5
D	3.9	9	8.4	5.4	3.3
W	3.9	9	8.4	5.4	3.3

$$\text{Chi-Square} = \frac{(\text{Actual} - \text{Expected})^2}{\text{Expected}} \text{ (sum)}$$

23.57

$DF = (\text{columns} - 1)(\text{rows} - 1)$
$= 12$

Critical Value = 21.03

As Chi square > CV, we reject the null hypothesis.

Two way ANOVA

	Low Noise	Medium Noise	High Noise	Row total
Male Students	10, 12, 11, 9 $\}$ 42	7, 9, 8, 12 $\}$ 36	4, 5, 6, 5 $\}$ 20	98
Female Students	12, 13, 10, 13 $\}$ 48	13, 15, 12, 12 $\}$ 52	6, 6, 4, 4 $\}$ 20	120
Column Total	90	88	40	

Correction term $= \dfrac{(\Sigma x)^2}{n} \rightarrow \dfrac{(10+12+11+9+7+\cdots+4)^2}{24} = 1980$

$Sum_{total} = (10^2+12^2+11^2+\cdots+4^2) - 1980 = 274$

$Sum_{column} = \dfrac{(90^2+88^2+40^2)}{8} - 1980 = 200$

8 $\rightarrow$ num of values in each column

$Sum_{row} = \dfrac{(98^2+120^2)}{12} - 1980 = 20$

12 $\rightarrow$ num of values in each row

$Sum_{within} = \dfrac{(42^2+48^2+\cdots+20^2)}{4} - 1980 - 200 - 20 = 16.33$

$Sum_{error} = 274 - 200 - 20 - 16.33 = 37$

Source	DF	Sum of squares	MSS	F $\frac{Mss}{Mse}$ (2.06)
Noise	$(C-1)=2$	200	100	$48.73 > F_{(2,18)} = 3.55$
Gender	$(R-1)=1$	20	20	$9.81 > F_{(1,18)} = 4.41$
Interaction	$(C-1)(R-1)=2$	16.33	8.167	$3.97 > F_{(2,18)} = 3.55$
Residual	$C \cdot R \cdot (n-1) = 18$	37	2.06	Reject all
Total	23	59.25		

C = num of rows
R = num of columns
n = num of elements in a group
N = Total population

Critical value = $(DF_x, DF_{Residual})$