Descriptive Statistics

***Average/Sample mean $\bar{X}$:** $\frac{1}{n}\sum_{i=1}^{n} X_i$

Maximum/Minimum: The largest/smallest order statistic

- The order statistic are the observations sorted into ascending order; $X_{(1)} \leq X_{(2)}$

Range: Difference between the maximum and the minimum

***Median:** Equal numbers of observations above and below (in an order statistic)

- If n is odd: $X(\frac{n+1}{2})$ – e.g. X(680) is the 680th number which is ANS: 73
- If n is even: $\frac{X(\frac{n+1}{2}) + X(\frac{n+1}{2})}{2}$ e.g. X(3.5) → 3.5th observation → mean of the two middle numbers

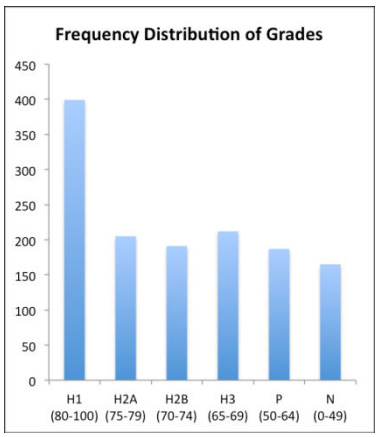
***Mode:** The most common value – it may not be unique

*measures of central location

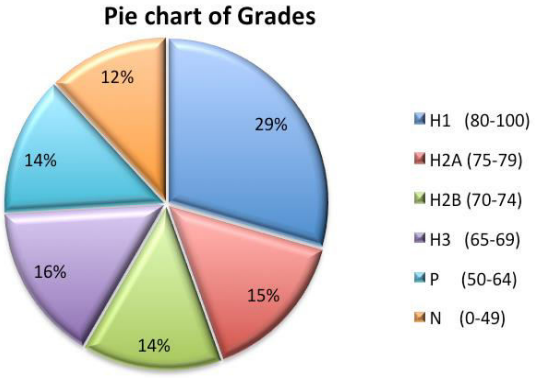
	Distribution	
Final Marks	**Frequency**	**Relative Frequency**
H1 (80-100)	399	29.4%
H2A (75-79)	205	15.1%
H2B (70-74)	191	14.1%
H3 (65-69)	212	15.6%
P (50-64)	187	13.8%
N (0-49)	165	12.1%
		100.0%

Frequency distribution: Number in each category

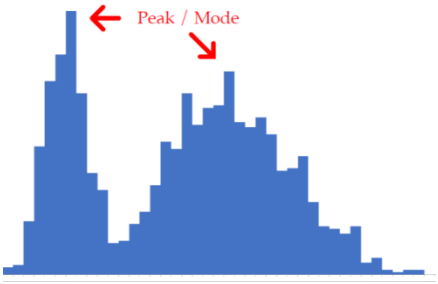
Relative Frequency Distribution: Percentage in each category



Frequency Distribution → **(Clustered) Column Chart (Number)**
Relative Frequency Distribution → **(Clustered) Column Chart (Percentage) OR Pie Chart**



Histogram: Group all observations into (usually) equally sized bins and count. Example: 1st Bin: Exam mark ≤ 5 = 4
2nd Bin: 5 < Exam mark ≤ 10



Bimodal Histogram (Figure 1): Two peaks/modes

Unimodal Histogram: Single peak/modes

Symmetric and Normal/Bell Shaped Histogram: Single peak; Mean ≈ Median

Negatively skewed Histogram: Long left tail; Unimodal; Mean < Median

Positively skewed Histogram: Long right tail; Unimodal; Mean > Median

Uniform Histogram (Figure 2): No distinct peak/mode; Mean ≈ Median

When Mean is very different to Median:

Mean is heavily influenced by outliers; the median is not influenced in this way. <u>In this case median gives a more realistic overall idea.</u>

- Median, quartiles, percentiles are generally influenced much less or not at all by outliers.

Displaying Changes over time: Use a line chart/time series plot (Y = number ('000s) OR percentage (%), X=years)

Measures of relative standing

What marks were needed to reach the top 1% in 2017? 99th percentile

Percentiles: p% of the observations are less than or equal to the p'th percentile (using order statistics)

Let i = $\frac{P}{100}(n+1)$

[i] : round down, e.g. [12.1] = 12, [12.9] = 12

E.g. Find the 99th percentile from the data set: 10, 20, 30, 40, and 50 → 99/100 (5+1) = 5.94 = 5 → 50

(99% of the observations are less than or equal to the 99th percentile (50))

p'th percentile = $X_{(i)} + (i - [i])(X_{(i+1)} - X_{(i)})$

First Quartile: 25th percentile

Second Quartile (Median): 50th percentile

Third Quartile: 75th percentile

Measuring Variability/Dispersion (how spread out the observations are)

Sample Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$

- Generally uninterpretable because it is s^2 → the square of the units of observation

Standard Deviation (s): the square root of Variance

Interquartile Range (IQR): a measure of how spread out around the median the values are

Third Quartile – First Quartile in order statistics → 50% lie between the IQR

The lower the answer the lower the variability

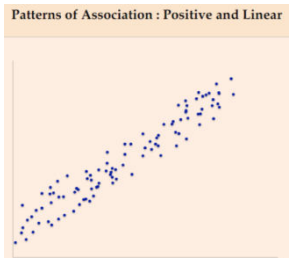
Range: Maximum minus the Minimum observation in order statistics (measures the extremes)

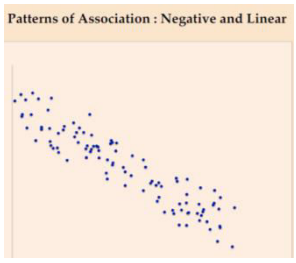
When units/scale of two measurements are different (e.g. Assignment: marks out of 7.5, Exam: marks out of 70), we use the coefficient of variation to compare between the two measurements

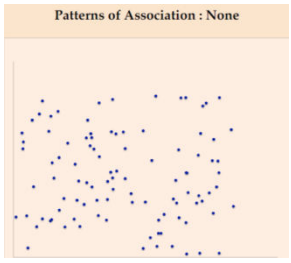
Coefficient of Variation: standard deviation divided by the mean cv = $\frac{s}{\bar{x}}$ x 100%; it is unit free

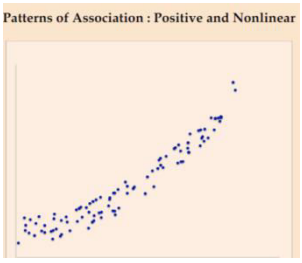
Measuring Relationships/Associations

2 or more Numerical data are best represented in a Scatter plot









All of them have a low variation except the scatterplot with zero patterns of association

Covariance: Measuring Direction of Association (positive or negative association)

Cov(X_i, Y_i) = S_{xy} = $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})$; Unit of measurement is ANS^2

$(x_i - \bar{x})(y_i - \bar{y})$ is **positive** if:

1) $x_i > \bar{x}$ and $y_i > \bar{y}$, or
2) $x_i < \bar{x}$ and $y_i < \bar{y}$

$(xi - \bar{x})(yi - \bar{y})$ is **negative** if:

1) $x_i > \bar{x}$ and $y_i < \bar{y}$, or
2) $x_i < \bar{x}$ and $y_i > \bar{y}$

The covariance is not really useful measure as it is ANS^2 and thus not interpretable, a much better measure would be correlation

Correlation/Correlation coefficient: Measuring **direction** and **strength** of <u>linear association</u>

r = $\frac{cov(xi,yi)}{sd(xi)sd(yi)} = \frac{S_{xy}}{S_x S_y}$

$-1 \leq r \leq 1$:

- **r close to + 1 ⇒** strong positive association
- **r close to − 1 ⇒** strong negative association
- **r close to 0 ⇒** weak or no linear association

Correlation cannot explain **why** there is a relationship – it cannot prove a relationship

Probability is the quantitative technique for dealing with uncertainty and events that (appear to) happen with chance.

Random Experiment: A process for which the outcome cannot be predicted with certainty e.g. "The grade achieved by a randomly chosen student – possible outcomes: H1, H2A, H2B, H3, P, N"

Sample Space: The sample space of a random experiment is the set of all possible outcomes.

- **Outcomes can be denoted O_1, O_2,….**
- **Sample space: S = {O_1, O_2,…}**

Unimelb Grades:

S = {H1, H2A, H2B, H3, P, N}

Event (denoted E): Collection of outcomes

Example: A randomly chosen QM1 students passes the subject:

E = {H1, H2A, H2B, H3, P}

An event may also be a single outcome

Example: Trumps gets impeached:

E = {Yes}

Probability: The probability of an event E is denoted P (E), with $0 \leq P(E) \leq 1$

- **P (E) = 0** implies E is impossible
- **P (E) = 1** implies E is certain to occur
- **P (E) = 0.5** implies E is equally likely to occur or not occur

Probabilities across all outcomes must sum to one: P (O_1) + P (O_2) +…= P(S) = 1

Assigning Probabilities

1. **Using past information**
2. **Objective / Relative Frequency Approach:** P(E) = The relative frequency (proportion) of occurrences of E if the random experiment is repeated indefinitely.

Random Experiment: toss a coin

Repeat n times: $\frac{Number\ of\ heads}{n} \rightarrow \frac{1}{2}$ as n $\rightarrow \infty$

$\rightarrow$ P (Heads) = $\frac{1}{2}$

3. **Subjective Approach:** The relative frequency approach is not always applicable, as assigning probabilities in some cases can be very difficult. Therefore, some judgement or subjectivity is required, as it is important for decision making

Complement: The complement of an event A is an event denoted $\overline{A}$. $\overline{A}$ is the set of outcomes in S that are not included in A.

$$P(\overline{A}) = 1 - P(A)$$

Intersection: The intersection of events A and B is denoted A∩B. A∩B is the set of elements common to both A and B

- **No Intersection:** A ∩ B = ∅ → empty set/event
  A and B are mutually exclusive events → no intersection

Union: The union of events A and B is denoted A ∪ B. A ∪ B is the set of elements in A or B or both.

Probability of an Event: If an event E contains outcomes $\{O_1, O_2, \ldots, O_n\}$ then
$P(E) = P(O_1) + P(O_2) + \ldots + P(O_n)$
If E = ∅ then P (E) = 0

Addition Rule for Probabilities: P (A ∪ B) = P (A) + P (B) – P (A ∩ B)

Joint Probability: P (A ∩ B) is called the joint probability of A and B i.e. the probability that both A and B occur
Marginal Probability: P (A) is called the marginal probability of A.

Independent Events: A and B are independent events if **P (A ∩ B) = P(A)P(B)**
Independence means the probability of A occurring is unaffected by knowledge of whether B occurred, and the reverse.

Conditional Probability: The probability of an event A conditional on an event B is

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

- Condition on an event: treat the event as known to have occurred.
- Probability of A given B has occurred

Multiplication Rule for Probabilities: P(A ∩ B) = P(A|B) x P(B)
If A and B are independent, i.e. P(A ∩ B) = P(A)P(B), then
P(A)P(B) = P(A|B) x P(B)
And hence P(A|B) = P(A) *3 tests for independence
i.e. the probability of A is unaffected by knowledge of whether or not B has occurred

Remember: Two independent variables have a zero covariance *4 tests for independence

Bayes Therom: P(A|B) = $\frac{P(B|A)P(A)}{P(B)}$
It is the combination of the two multiplication rules:
P(A ∩ B) = P(A|B) x P(B)
P(A ∩ B) = P(B|A) x P(A)
To imply
P(A|B)P(B) = P(B|A)P(A)
And then divide both sides by P(B)

Law of Total Probability: P(A) = P(A ∩ B) + P(A ∩ $\bar{B}$)

= P(A|B)P(B) + P(A| $\bar{B}$)P($\bar{B}$)

First line: Some outcomes in A may also be in B - P (A ∩ B)

The rest of the outcomes in A are not in B - P (A ∩ $\bar{B}$)

Second line: Apply the Multiplication rule to each term.

Are you on drugs?

Probabilities provided:

"On any given Saturday night, perhaps 1% of Australians may take the drug ecstasy." **P(E) = 0.01**

Test can "correctly detect the presence of ecstasy with probability 0.99" **P(T|E) = 0.99**

Test can "correctly detect the absence of ecstasy with probability 0.99" **P($\bar{T}$)| $\bar{E}$) = 0.99**

"One Saturday night a randomly selected motorist tests positive for ecstasy - what is the probability they have actually taken it?" **P (E|T) = ??? = 0.5**

Translation into Probability Notation

Events:

- E : individual has actually taken ecstasy that night
- $\bar{E}$: individual has not taken ecstasy that night
- T : test is positive
- $\bar{T}$: test is negative

Four possibilities:

1. Individual took ecstasy and tested positive **(E ∩ T) (Test was correct)**
2. Individual took ecstasy and tested negative **(E ∩ $\bar{T}$) ("False negative")**
3. Individual did not take ecstasy and tested positive **($\bar{E}$∩ T) ("False positive")**
4. Individual did not take ecstasy and tested negative **($\bar{E}$ ∩ $\bar{T}$) (Test was correct)**

Using Bayes Theorem as we need to reverse the order of the conditioning:

$$P(E|T) = \frac{P(T|E)P(E)}{p(T)} = 0.99 \text{ x } 0.01 \text{ / } ??$$

Substitute the Law of Total Probability into equation

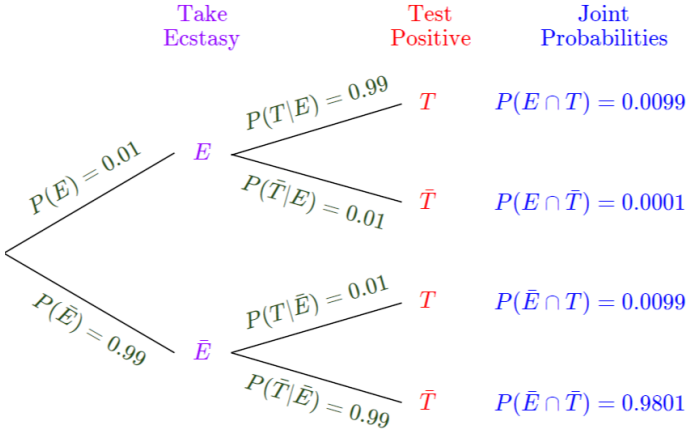
$$P(E|T) = \frac{P(T|E)P(E)}{P(T \cap E) + P(T \cap \bar{E})} = \frac{P(T|E)P(E)}{P(T|E)\,P(E) + P(T|\bar{E})P(\bar{E})}$$

$$= \frac{0.99\ x\ 0.01}{(0.99\ x\ 0.01)+(0.01\ x\ 0.99)} = 0.5$$

REMEMBER: ($\bar{T}$| $\bar{E}$) = 0.99, Thus (T |$\bar{E}$)= 0.01

AND: (T | E) = 0.99, Thus ($\bar{T}$|E)= 0.01

Another way of handling this question is to do a Tree Representation



Using the Law of Total Probability:

P(T) = P(T ∩ E) + P(T ∩ $\bar{E}$)

P(T) = 0.0099 + 0.0099 = 0.0198

Therefore = 0.99 x 0.01/0.0198 = 0.5

Random Variables: is a rule/function that assigns a numerical value to each outcome of a random experiment (e.g. X = heads when flipping a coin twice; X=2, X=1, X=0)

Discrete random variable: has a finite or countably infinite number of outcomes.

- e.g. dice rolling has possible values 1, 2, 3, 4, 5, 6
- e.g. number of shares traded on a randomly selected day.

Continuous random variable: has an uncountably infinite number of possible outcomes. e.g. draw a real number at random between 0 and 1

Example. Let X denote the value from a single dice roll. Probability distribution:

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

Discrete Probability Distributions: A probability distribution for a discrete random variable gives the probability of each of the possible outcomes.
If the random variable is denoted X then the probability distribution is written **P(X = x) = p(x), for any number x.**

Expected Value/weighted average of the random variable: A discrete random variable X with possible outcomes $x_1, x_2, \ldots, x_n$ and probability distribution p(x) has expected value

$E(X) = x_1p(x_1) + x_2p(x_2) + \ldots + x_np(x_n) = \sum_{i=1}^{n} x_i p(x_i)$
= units of measurement of the random variable

<u>Addition rule for Expected Values</u>
$X = X_1 + X_2 + \ldots + X_n$ then
$E(X) = E(X_1) + E(X_2) + \ldots + E(X_n)$

<u>Dispersion of the possible values of X from its expected value – (on how uncertain they are – the risk)</u>

- **The variance of a random variable X is**

$var(X) = E[(X-E(X))^2] \rightarrow [P(x_1) * (x_1 - E(x))^2] + [P(x_2) * (x_2 - E(x))^2]$ etc
OR $E(X^2) - E(X)$
Units is ANS^2 i.e. 20 $dollars^2$

- **The standard deviation is**

$sd(X) = \sqrt{var(X)}$.
Units is just ANS

Remember: Integers are ones without a fractional component

Joint Probability Distributions

Consider random variables X and Y with possible outcomes $x_1,\ldots, x_m$ and $y_1,\ldots, y_n$.

- The joint probability distribution of X and Y is **p(x, y) = P(X = x and Y = y) for any numbers x and y.**

The *marginal* probability distributions of X and Y are

$$P(X = x) = \sum_{i=1}^{n} p(x, y_i)$$

$$P(Y = y) = \sum_{i=1}^{m} p(x_i, y)$$

(Law of Total Probability)

Example:

Z	**90**	**120**
P(Z=z)	0.4	0.6

T	**86**	**130**
P(T=t)	0.5	0.5

		Z		
T		**90**	**120**	
	86	0.1	0.4	**0.5**
	130	0.3	0.2	**0.5**
		0.4	**0.6**	**1**

→ Joint probability is within this table

E.g. p(86, 90) = P(T = 86 and Z = 90) = 0.1

These numbers would be given to you (not worked out)

All of the probabilities add up to 1 = 0.1+0.4+0.3+0.2

→ Marginal Probability

e.g. P(Z = 90) = P(Z = 90, T = 86) + P(Z = 90, T = 130)

= 0.1 + 0.3

= 0.4

When given the joint probabilities, marginal probability can be worked out, however this does not work the other way around. **However, if you have independent random variables → then joint probabilities are equal to the <u>product of the marginal probabilities</u>**

<u>What if you split your money equally between the two?</u>

The random payoff from $100 of Facebook shares is Z

→ The random payoff from **$50 of Facebook shares is 0.5Z**

The random payoff from $100 in Thelonious's hedge fund is T

→ The random payoff from **$50 in Thelonious's hedge fund is 0.5T**

Portfolio: $50 of Facebook shares and $50 in Thelonious's hedge fund

Portfolio random payoff: Q = 0.5Z + 0.5T

		Z	
T		**90**	**120**
	86	88	103
	130	110	125

→ Possible outcomes for Q

Probabilities ←

		Z	
T		**90**	**120**
	86	0.1	0.4
	130	0.3	0.2

Q	**88**	**103**	**110**	**125**
P(Q=q)	0.1	0.4	0.3	0.2

→ Probability distribution of Q

If

$$X = w_1X_1 + w_2X_2 + \ldots + w_nX_n,$$

(w_i's are fixed weights) then

$$E(X) = w_1E(X_1) + w_2E(X_2) + \ldots + w_nE(X_n).$$

<u>The Addition Rule for Expected Values with weights</u>

Q = 0.5Z + 0.5T

e.g E(Q) = 0.5E(Z) + 0.5E(T)

The weights do not have to add up to 1

$$\text{cov}(X, Y) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i - E(X))(y_j - E(Y)) \cdot p(x_i, y_j)$$

<u>Covariance:</u> between two random variables X and Y is: **cov(X, Y) = E [(X – E(X))(Y – E(Y))]**

- multiply each possible $(x_i - E(X))$ by each possible $(y_j - E(Y))$
- weight by the joint probabilities $p(x_i, y_j)$
- add up all m x n of these terms

Example

		Z	
T		**90**	**120**
	86	0.1	0.4
	130	0.3	0.2

$$\begin{aligned}\operatorname{cov}(T,Z) &= \sum_{i=1}^{2}\sum_{j=1}^{2}(t_i - E(T))(z_j - E(Z)) \cdot p(t_i, z_j) \\ &= (86-108)(90-108)\cdot p(86,90) \\ &\quad + (86-108)(120-108)\cdot p(86,120) \\ &\quad + (130-108)(90-108)\cdot p(130,90) \\ &\quad + (130-108)(120-108)\cdot p(130,120)\end{aligned}$$

p(86,90) = 0.1
ANS: -132 = units is X^2 e.g. dollars^2
A negative relationship between T and Z
When the payoff for one variable goes up, the other goes down

The *correlation* between two random variables X and Y is

$$\operatorname{cor}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{sd}(X)\cdot\operatorname{sd}(Y)} \qquad (-1 \le \operatorname{cor}(X,Y) \le 1)$$

Correlation
e.g. ANS: -0.41 = a moderately strong negative relationship between T and Z

If w_1 and w_2 are fixed weights then

$$\operatorname{var}(w_1X + w_2Y) = w_1^2 \operatorname{var}(X) + 2w_1w_2 \operatorname{cov}(X,Y) + w_2^2 \operatorname{var}(Y)$$

Example.

$$\begin{aligned}\operatorname{var}(Q) &= \operatorname{var}(\overset{w_1}{\downarrow}0.5T + \overset{w_2}{\downarrow}0.5Z) \\ &= 0.5^2 \times \operatorname{var}(T) + 2 \times 0.5 \times 0.5 \times \operatorname{cov}(T,Z) + 0.5^2 \times \operatorname{var}(Z) \\ &= 0.5^2 \times 484 + 2 \times 0.5 \times 0.5 \times (-132) + 0.5^2 \times 216 \\ &= 109\end{aligned}$$

→ Addition rule for variances

What's the best portfolio? (Highest return with the lowest risk)
Suppose we invest fraction w in Z, and hence 1 − w in T
Define the portfolio:
Q = wZ + (1 − w)T
For any w, an investment of $100 in Q has expected value
E(Q) = wE(Z) + (1 − w)E(T)

= w × 108 + (1 − w) × 108 = *$108*
Can we choose w to minimise risk?

var(Q) = var(wZ + (1 − w)T)
= w^2var(Z) + 2w(1 − w)cov(Z, T) + $(1 − w)^2$var(T) → addition rule for variances
= $w^2 \times 216 + 2w(1 − w) \times (−132) + (1 − w)^2 \times 484$

Algebra:
$216w^2 - 132(2w - 2w^2) + 484 – 968w + 484w^2$
$216w^2 - 264w + 264w^2 + 484 – 968w + 484w^2$
1. var(Q) = $964w^2 – 1232w + 484$
Risk is a quadratic (i.e. parabola) in w, **with minimum solving**
$\frac{dvar\,(Q)}{dw} = 2 \times 964w − 1232$
$= 1928w − 1232 = 0.$
Thus $w = \frac{1232}{1928} =$ **0.639 is the risk-minimising allocation to Z.**
1. var(Q) = $964 \times 0.639^2 – 1232 \times 0.639 + 484 = 90.373$

Remember:
- More than → does not include the number, at least → includes the number
- If random variables are independent, covariance is zero and correlation is 0