Cofounders
-to deal with cofounders into sub groups into subgroups

Observational vs-controlled studies
-controlled uses a treatment and control group and are allocated. Uses placebo(effect occurs when subject respond to the idea of the treatment) and bias. A observational study they two groups are not assigned, can establish association not causation.

Simpsons paradox
Sometimes there is a clear trend in individual groups of data that disappears when groups are pooled together

Histogram
Area of block =percentage of subjects
Height of block = % of block / length of class interval

Boxplot
-box is 50 percent of data (Q3-Q1= IQR)

Centre
Mean - good for symmetrical

Medium- robust therefore good for skewed data
For symmetrical data mean and medium are the same
For left skewed data mean is smaller than medium
For right skewed data mean is larger than medium

Spread
- -the mean gap must be 0 as the mean is the balancing point

Sd
- The standard deviation measures the **spread** of the data.

$$SD_{pop} = RMS \text{ of (gaps from the mean)}$$

Formally, $SD_{pop} = \sqrt{\text{Mean of (gaps from the mean)}^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$

The area 1 SD from mean 0.68%
The area 2SD from mean is 0.95%
The area 3SD out from men is 0.997%

Population RMS of gaps from the
mean $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(gaps)^2}$ $\quad$ sd(data)*sqrt((n-1)/n))

Sample Adjusted RMS of gaps from
the mean
$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(gaps)^2}$ $\quad$ sd(data)

Shortcut
$SD = (\text{big -small})\sqrt{\text{proportion of big} \times \text{proportion of small}}$

Standard zscore
Standard unit of data = how many sds it is below or below the mean

$$\text{Standard units} = \frac{\text{data point - mean}}{\text{standard deviation}}$$

IQR
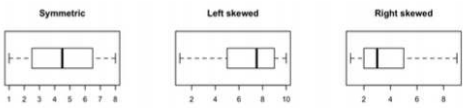Range of middle 50 percent(Q3-Q1)
Q3= 75% Q1= 25%

Coefficient variant

$$CV = \frac{\text{mean}}{\text{standard deviation}}$$

Symmetric | Left skewed | Right skewed

Normal curve
-use normal curve when histogram looked normal
Standard normal curve
-mean 0 and SD of 1

Special properties of curves
-satisfy the 68%,95% and 99.7%
-any general curve can be rescaled to standard curve

Measurement error
No matter how carefully any measurement is made, it could of turned out differently
Individual measurement = exact value + chance error

Chance error
The best way to estimate chance error is to replicate the measurement under the same conditions and calculate SD

Correlation coefficient
Measures clusters around a line
-1 <r< 1 the closer to -1 or 1 the more tightly points cluster
R =0 : points don't fit on a line
In R cor(x,y)
Properties
Symmetrical (changing variable wont effect) and scaling invariant (shifting variables wont effect)

R population = mean of the product of the variables in. S=

$$\frac{\text{data pt - mean(x)}}{SD(x)} \times \frac{\text{data pt - mean(y)}}{SD(y)}$$

Misleading conclusions
1)outliers overly influence
2)non-linear associations
3)rates of av tends to inflate the cor-coef
4)association is not causation
5)small sd makes the correlation look bigger

SD line
-on the point of av and plots line using SDs away from mean

Regression line
- **lm(y~x)**

	SD Line	Regression Line
	$(\bar{x}, \bar{y})$ to $(\bar{x}+SD_x, \bar{y}+SD_y)$	$(\bar{x}, \bar{y})$ to $(\bar{x}+SD_x, \bar{y}+rSD_y)$
	$\frac{SD_y}{SD_x}$	$r\frac{SD_y}{SD_x}$

- **Predicting percentiles:** pnorm[qnorm(0.9)*cor(x,y)]
- **Predicting y value:** mean(y) + qnorm(0.9)*cor(x,y)*sd(y)

Graph of av
-plots the average y for each x
-the regression line is a smooth version

Predictions
1. Baseline p, given a certain value for x a basic prediction of y would be the av of y over all the x values
2. Prediction in a strip, given a certain value x, y would be the av of all the y values in the data corresponding to that x value
3. Regression line
4. Predicting percentile rates

Regression fallacy- test retest situations means changes in results is due to chance.

Residual plot
a residual plot is the vertical distance gap of points above and below regression line
- error = actual value – predicted value

CLT
The **Central Limit Theorem** (CLT) that a large sample size from a population with a finite level of ... from the same ... ual to the mean

---

RMS error
The RMS error represents the gap between points and the regression line

- **RMS Error (population)**
RMS of (gaps from the line) = $\sqrt{\text{mean of (gaps)}^2}$

- **Baseline Prediction**
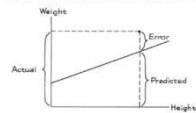$\text{RMS error}_{Pop} = SD_y$

- **Normal Prediction**
$\text{RMS error}_{Pop} = \sqrt{1-r^2}SD_y$

Residual plot
Plots residuals vs x
If the linear fit is appropriate it should show no pattern

Vertical strips
equal spread in y direction data is homoscedastic and RMS can be used.
unequal spread the data is heteroscedastic and RMS cannot be used

The prosecutor's fallacy
Assuming the prob of a random match is equal to the prob that the defendant is innocent
Basic properties of chance
1. Chances are between 0 – 100 %
2. The chance of something equals 100%- it's opposite

Conditional probability
Chance something occurs, given something else has occurred

$$P(\text{Event1}|\text{Event2})$$

Multiplication principal
The prob that two things occur is the chance of the 1st event multiples by the chance of the second

Addition rule
If the 2 things are mutually exclusive then the chance of at least 1 occurring is the sum of the individual chances

Mutually exclusive
2 things are mutually exclusive when the occurrence of the 1st event prevents the other

Column 1

What	When	Formula	Condition
Addition Rule	P(At least 1 of 2 events occurs)	P(Event1) + P(Event2)	if mutually exclusive
Multiplication Rule	P(Both events occur)	P(Event1) x P(Event2)	if independent
		P(Event1) x P(Event2, given Event 1)	if dependent

Independence
P(2nd event|1st event) = p(2nd event), independent of the outcome of the 1st event
Drawing randomly with replacement is independent
2 things are independent if the product of their unconditional probabilities

Dependence
2 things are dependent if the chance of the second given the 1st is not the same as the 2nd, dependent of the outcome of the 1st event
Drawing without replacement ensures dependence.

Binomial model

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

binary trial is where 2 things can occur p = 1-p
binomial theorem is p at every trail is fixed and use equation

In R dbinom(x,n,p) where,
- X = number of heads
- N = total number of tosses
- P = probability of events

Chance variability/error
Observed = expected +chance error
As no. of tosses increase then the size of chance error increases and the absolute percentage size of chance decreases

Law of average
-States that the proportion of heads becomes more stable as the length of the simulation increases and approaches a -fixed number called the relative frequency
The chance error in the number of heads is likely to be large in absolute size, but small relative to no. of tosses

Box model
-mark1 on tickets you are counting
Replicate(N,sum(sample(box,n,rep=T)) n=no.of draws N=replications
Parameter= fact about pop Estimate = cal of sample values which predicts parameter

Column 2

Sum of the Sample
Estimate:
$$EV_{sum} = \text{sample size} \times \text{mean}_{box}$$
Chance Error:
$$SE_{sum} = \sqrt{\text{sample size}} \times SD_{box}$$

Proportion in the Sample
Estimate:
$$EV_{proportion} = \text{population proportion} = \text{mean}_{box}$$
Chance Error:
$$SE_{proportion} = \frac{SD_{box}}{\sqrt{\text{sample size}}}$$

Correction Factor
$$SE_{withoutreplacement} = \text{correction factor} \times SE_{withreplacement}$$
where
$$\text{correction factor} = \sqrt{\frac{\text{number of tickets-number of draws}}{\text{number of tickets -1}}}$$
$$\text{correction factor} = \sqrt{\frac{\text{population size-sample size}}{\text{population size -1}}}$$

bootstrapping
is estimating properties of pop by using the properties of the sample
pop percentage ~ sample percentage chance error

$$SE_{proportion} \approx \frac{SD_{box \ (with \ sample \ proportions)}}{\sqrt{\text{sample size}}}$$

Model 4

Hypothesis testing
The null hypothesis h0 assumes that the difference between OV and EV is due to chance alone
1 sided, specifies the change in expected by the treatment h1 p>0.5
2 sided does not specify the change expected by the treatment H1 cannot equal 0.5
Test statistic OV – EV / SE

Column 3

P- value
Small <0.05 statistically different
The p value is a way of weighing up whether the sample is consistent with H0

Proportion test
HO any participants in the treatment group reporting an improvement is due to chance alone(Ho = 0,5)H1 improvement is due to treatment

Assumptions
Participants are independent of each other and chance of becoming desensitised is the same for all participants P test - >4.3

Accuracy of means
Ev mean = ev sum divided by number of draws
SE mean = Se sum divided by number of draws or sdbox divided by square root number of box
-CLT popbox is normally distributed
-The more skewed the bigger the sample size needs to be

Gauss model
a measurement of error, different, tomake it unbiased we set the meanerrorbox = 0

Z test
Hypothesis
- 2 sided → 2*pnorm(x)
H: H_0 : population mean = c vs H_1 : population mean ≠ c
- One sided:
H: H_0 : population mean = c vs H_1 : population mean < c (or >)

Assumptions: Sample is random. Know the population SD or the sample SD can estimate the population SD.
ONLY APPROPRIATE FOR LARGE SAMPLE SIZES FOR the CLT.

Test Statistic: uses normal curve
$$T: \text{Test statistic} = \frac{\text{observed mean - population mean}}{\frac{\text{population SD}}{\sqrt{n}}}$$

T-test
Hypothesis
H: H_0 : population mean = c vs H_1 : population mean ≠ c

Test statistic
P: Use t_{n-1} curve to find tail area for observed test statistic.
$$T: \text{Test statistic} = \frac{\text{observed mean - population mean}}{\frac{\text{sample SD}}{\sqrt{n}}}$$

Column 4

2 Sample T-test
CODE: var.equal = T
Hypothesis:
Let μ_1 = mean heart rate of our control (no Red Bull)
Let μ_2 = mean heart rate of our treatment (Red Bull)
H_0 : There is no difference: $\mu_1 = \mu_2$
H_1 : There is a difference: $\mu_1 \neq \mu_2$

Assumptions:
- Two samples are **independent** (eg. treatments not on same person)
- **Normality:**
  - **Boxplot** → symmetry (if not, then diff spread)
  - **QQ Plot for Normality** → line formed by points should be straight
  - **Shapiro-Wilk Test for Normality** → p > 0.05 is normality
- **Same variation/spread**
  - **Levene's Test (F-test)** for equal spread → p > 0.05 is equal variance
    - H_0 : There is no difference: $\sigma_1^2 = \sigma_2^2$
    - H_1 : There is a difference: $\sigma_1^2 \neq \sigma_2^2$

T-statistic
$$\text{test statistic} = \frac{\mu_1 - \mu_2 - 0}{\text{standard error of the difference}}$$

Paired T-test
CODE: paired = T
- Non-Independence between two groups
- **Assumptions** for paired t-test: data of difference is normal (check using QQ plot)

Hypothesis
H_0 : The difference in means is zero : $\mu_d = 0$
H_1 : The difference in means is not zero : $\mu_d \neq 0$

Welch Sample T-test
CODE: var.equal = FALSE
- **Assumptions:** Unequal variance → Check variance by conducting Levene's Test (F-test)
  - p < 0.05 is unequal variance

Chi-squared Tests
Hypothesis
H_0: Model fits data vs H_1: Model doesn't fit data.

Test Statistic
$$\chi^2 = \text{Sum of}\left[\frac{(\text{Observed frequency - Expected frequency})^2}{\text{Expected frequency}}\right]$$

Types
1) **Goodness of Fit:** distribution of qualitative variable in a population
   CODE: pchisq(x2, degrees of freedom, lower.tail = F)
   Deg of freedom = no. of catergories – 1
2) **Homogeneity:** distribution of qualitative variable in several populations
3) **Independence:** test a hypothesis about the relationship between 2 qualitative variables in a population
   CODE: chisq.test(data)
   Deg of freedom = (m-1)*(n-1) (m = row, n = columns)