Probability	Theoretical	Empirical/Experimental	Subjective
Description	Based on **mathematical calculations** and known possible outcomes.	Based on **actual trials or experiments** and observed outcomes.	Based on **personal judgment, intuition, or past experience.**
Formula $P(E)$	$\frac{\text{Favourable Outcomes}}{\text{Total Possible Outcomes}}$	$\frac{\text{No. of Times Event Occurred}}{\text{Total Possible Outcomes}}$	N/A

Axioms of Probability

(i) $P(A) \geq 0$ for all $A \subseteq S$. (ii) $P(S) = 1$.

(iii) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \quad \forall i = 1, 2, \ldots$ such that $A_i \cap A_j = \emptyset$ for $i \neq j$.

Rules of Probability

(i) $P(A') = 1 - P(A)$

(ii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(iii) If A and B are **mutually exclusive**,

$P(A \cup B) = P(A) + P(B)$

(iv) $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

(v) If A and B are **independent**,

$P(A \cap B) = P(A)P(B)$

(vi) $P(A|B) = \dfrac{P(A \cap B)}{P(B)}, \quad P(B) > 0$

(vii) If A and B are **independent**,

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

(viii) $P(A'|B) = 1 - P(A|B)$

Random Variables

DISCRETE

PMF:	$P(x) = P(X = x)$
CDF:	$F(x) = P(X \leq x) = \sum_x p(x)$
Expectation:	$E(X) = \sum_x xp(x)$
	$E(g(x)) = \sum_x g(x)p(x)$
Variance:	$\text{Var}(X) = E(X^2) - (E(X))^2$

CONTINUOUS

PDF:	$f(x)$
CDF:	$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$
Expectation:	$E(X) = \int_{-\infty}^{\infty} xf(x)dx$
	$E(g(x)) = \int_{-\infty}^{\infty} g(x)f(x)dx$
Variance:	$\text{Var}(X) = E(X^2) - (E(X))^2$

Permutation and Combination

Permutation (✓ order): ${}_nP_r = \dfrac{n!}{(n-r)!}$

Combination (✗ order): ${}_nC_r = \binom{n}{r} = \dfrac{n!}{r!(n-r)!}$

Partition of Sample Space

If $A_1, A_2, \ldots, A_n$ is a partition of S

(i.e., $\bigcup_{i=1}^{n} A_i = S$ and $A_i \cap A_j = \emptyset$ for $i \neq j$), then:

(i) $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$

(ii) **Law of total probability**

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

(iii) **Bayes' Rule**

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Basic Integration

Constant	$\int a\,dx = ax + C$
x^n	$\int x^n\,dx = \dfrac{x^{n+1}}{n+1} + C, \; n \neq -1$
$1/x$	$\int 1/x\,dx = \ln\lvert x\rvert + C$
e^x	$\int e^x\,dx = e^x + C$
Indefinite	$\int f(x)\,dx = F(x) + C$
Definite	$\int_a^b f(x)\,dx = [F(x)]_a^b = F(b) - F(a)$
	$\int kf(x)\,dx = k\int f(x)dx$
	$\int [f(x) \pm g(x)]\,dx = \int f(x)\,dx \pm \int g(x)\,dx$
	$\int \ln(x)\,dx = x\ln(x) - x + C$
	$\int \dfrac{f'(x)}{f(x)}\,dx = \ln\lvert f(x)\rvert + C$
	$\int e^{f(x)}\,dx = \dfrac{e^{f(x)}}{f'(x)} + C$
By parts	$\int u\,dv = uv - \int v\,du$
By substitution	$\int_a^b f(u(x))\dfrac{du}{dx}\,dx = \int_{u(a)}^{u(b)} f(u)\,du$

Gamma function

For positive integer,

$$\Gamma(n) = (n-1)!$$

For positive real and complex numbers,

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}\,dt, \quad \text{for } x > 0$$

Study Type	Description	Examples
Observational	Researchers observe and collect data *without manipulating any variables*.	Study a population of birds by recording changes in their migration patterns over time.
Experimental	Involve actively *manipulating one or more variables* to determine their effects on an outcome. This is often done with controlled environments and random assignments, like in clinical trials.	Testing the effectiveness of a new drug by giving it to one group (treatment group) while giving a placebo to another (control group).

Terminology	Description
Random	Random allocation, usually by a computer.
Controlled	Includes a baseline group for comparison.
Blinded	Participants don't know their group/treatment.
Double blinded	Both participants and researchers are unaware.
Triple blinded	Participants, researchers, and analysts are unaware.
Factorial design	Studies multiple variables and their interactions.

Survey	Research method used to gather information from a group of people by *asking them questions* in order to understand opinions, behaviours, attitudes, or experiences on a specific topic.	A company surveys customers about their satisfaction with a product.

	Definition	Characteristics
Population	The *entire group of individuals* or items that researchers are interested in studying.	Large, and studying them directly can be time-consuming, expensive, or impractical.
Samples	A smaller, *selected group taken from the population*. We study the sample and use it to *make inferences* about the population.	Manageable, cost-effective, and allow for quicker data collection. The accuracy of results depends on how representative the sample is.

Sample Strategy	Description	R Example (consider `iris` data)
Simple Random	Randomly select k numbers without replacement	`set.seed(123) # For reproducibility` `s <- iris[sample(1:nrow(iris), 30), ]`
Systematic	Selects individuals at regular intervals from a list.	`start <- sample(1:10, 1) # Random start` `s <- iris[seq(start, nrow(iris), by=10), ]`
Cluster	Randomly selects entire groups or clusters	`c <- split(iris, iris$Species)` `c[[sample(unique(iris$Species), 1)]]`

Stratified	Divides the population into groups and samples each group.	`library(dplyr)`
		`s <- iris %>% group_by(Species) %>%`
		`sample_n(10)`

Purposive	Selects individuals based on specific criteria or purpose.	`s <- subset(iris, Sepal.Length > 7)`
Convenience	Uses easily accessible individuals for the sample.	Note: Very susceptible to bias

Variable Type	Description	Examples
Quantitative	Numerical and can be counted or measured.	
– **Discrete**	Countable numbers	Number of students
– **Continuous**	Measurable values with infinite possibilities	Height, weight
Qualitative (Categorical)	Descriptive and refers to things that can be observed but not measured.	
– **Binary**	Variable that has only two possible values	Yes/No, Success/Failure
– **Ordinal**	Categories with a logical order	Small, medium, large
– **Nominal**	Categories without numerical or logical order	Hair colour: black, brown, blonde

Statistics | **Description**

Descriptive (EDA)

Summarises and describes data using *numerical summaries* like averages (mean, median, and mode), spread (variance, standard deviation, and range), or visualisations like graphs and data tables (*graphical summaries*).

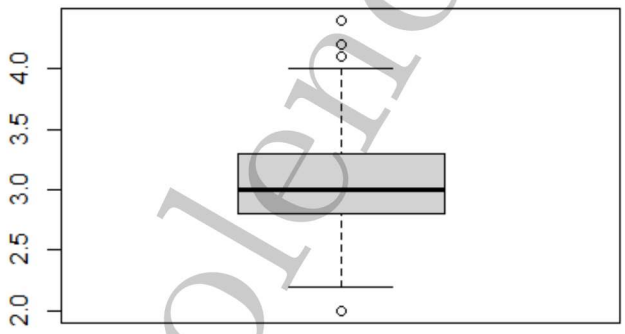
– Numerical summaries

Central Tendency	Population	Samples	Remark
Mean	$\mu = \sum x_i/N$	$\bar{x} = \sum x_i/n$	Best for symmetric dist.
Median	$P(X \leq m) = P(X \geq m) = 0.5$		Best for skewed dist. w/ outliers
Mode	The value that appears most frequently in a dataset		

Spread	Population	Samples	Remark
Variance	$\sigma^2 = \sum(x_i - \bar{x})^2/N$	$s^2 = \sum(x_i - \bar{x})^2/(n-1)$	Best for symmetric dist.
Range	Max – Min		Large, not good w/ outliers
IQR	$Q_3 - Q_1$		50% of the data

R Example (consider `iris` data)	Comment
`summary(iris)`	Five-number summary + Mean
`by(iris$Sepal.Width, INDICES=iris$Species, FUN=summary)`	Five-number summary by group
`mean(iris$Sepal.Length)`	Mean
`quantile(iris$Sepal.Length)`	Quantile
`sd(iris$Sepal.Length)`	Standard deviation

– Graphical summaries*

R Base Plot Example (consider `iris` data)	Comment
`boxplot(iris$Sepal.Width)` `boxplot(Sepal.Length~Species, data=iris)`	Box-and-whisker plot
	∘ – Outliers Upper whisker = $Q_3 + 1.5IQR$ Q_3 position = $3(n+1)/4$ Q_2 position = $(n+1)/2$ Q_1 position = $(n+1)/4$ Lower whisker = $Q_3 - 1.5IQR$
`hist(iris$Sepal.Length, freq=TRUE)`	Frequency histogram
`hist(iris$Sepal.Length, freq=FALSE)` `lines(density(iris$Sepal.Length))`	Density histogram
`barplot(summary(iris[sample(1:nrow(iris),30),]$Species))` `pie(summary(iris[sample(1:nrow(iris),30),]$Species))`	Bar chart Pie chart
`plot(Sepal.Length~Sepal.Width, data=iris)` `pairs(iris, panel=panel.smooth, col=2+(iris$Petal.Length>2.5))`	Scatter/pair plot

*For more visualisation options, check out `ggplot2` cheat sheet at https://posit.co/wp-content/uploads/2022/10/data-visualization-1.pdf

Inferential

Uses *sample* data to make *predictions* or *generalisations* about a larger population, often involving *hypothesis testing* and *confidence intervals*.

Statistical Tests (Hypothesis testing)	Statistical Models (Prediction)
One-sample z-test	General linear models
t-test (one-sample, two-sample, paired)	Generalised linear models
ANOVA	Time-series models
Chi-squared test	Repeated measures models
F-test	Multivariate models

Population distribution

The distribution of x is called the population distribution of *one measurement*.

$$E(X) = \mu \qquad Var(X) = \sigma^2$$

$$z = \frac{x - \mu}{\sigma} \qquad sd(X) = \sigma$$

Sampling distribution

The distribution of $\bar{x}$ is called the sampling distribution of *average measurement* based on multiple samples drawn from the population.

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \sigma^2/n$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad sd(X) = \sigma/\sqrt{n}$$

Law of Large Numbers (LLN) By LLN, $\uparrow n : \bar{x} \to \mu$

As the sample size n becomes large, the sample mean will become close to the population mean no matter what the distribution of each x_i is.
LLN Example: A simulation with a fair die, which has a population mean of 3.5.

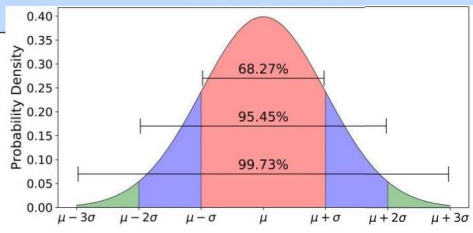
```
# Simulate rolling a fair die 10,000
rolls <- sample(1:6, size = 10000, replace = TRUE) times
# Compute cumulative mean after each roll
cmean <- cumsum(rolls) / seq_along(rolls)
plot(cmean, type="l", col="blue", lwd=2, xlab="Number of Rolls", ylab="Sample Mean")
abline(h=3.5, col="red", lty=2)
legend("topright",legend=c("Sample Mean","True Mean (3.5)"),col=c("blue","red"),lty=c(1,2))
```

Central Limit Theorem (CLT) By CLT, $\uparrow n : \bar{x} \to N(\mu, \sigma^2/n)$

When you take large samples ($n \geq 30$) of independent RVs x_i ($i = 1, \ldots, n$) from a population (with a mean μ and standard deviation σ_X), the sampling distribution $\bar{x}$ is approximately normal with a mean μ and standard deviation $\sigma_X/\sqrt{n}$ regardless of the original population's distribution. This allows us to use properties of the normal distribution (like z-scores) to make inferences about the population mean using the sample mean.
CLT Example: Simulate 10,000 coin tossing experiments (binomial experiments), each tossing a coin 10 times.

```
install.packages("mosaic")
library(mosaic)
# simulate 10,000 experiments
BinomExp10000 <- do(10000) * rflip(10)
# sampling distribution proportions
tally(~heads, data=BinomExp10000, format="proportion")
histogram(~heads, data=BinomExp10000, width=1, labels=TRUE)
dbinom(0:10, 10, p=0.5)
```

The Empirical Rule 68-95-99.7 Rule



If the data is normally distributed, and has a mean μ and standard deviation σ

- 68% of the observations are between $\mu \pm \sigma$
- 95% of the observations are between $\mu \pm 2\sigma$
- 99.7% of the observations are between $\mu \pm 3\sigma$

100(1–α)% Confidence Interval (CI) = Estimate ± (Critical Value × Standard Error)

A range of values, derived from sample data, that is likely to contain the true population parameter (mean μ or proportion p).

– If σ **is known**, CI for the population mean is given by $\boxed{\bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}$,

with confidence level of 90% ($z_{\alpha/2} = 1.645$), 95% ($z_{\alpha/2} = 1.96$), 99% ($z_{\alpha/2} = 2.58$).

– If σ **is unknown**, CI for the population mean is given by $\boxed{\bar{X} \pm t_{\alpha/2,df}\left(\frac{s}{\sqrt{n}}\right)}$,

where the critical $t_{\alpha/2,df}$ is derived from the t-distribution tables with $df = n - 1$.

– If a sample proportion $\hat{p}$ follows a binomial distribution and the sample size n is sufficiently large (typically $n\hat{p}$ and $n(1-\hat{p})$ are both ≥ 5 or ≥ 10), CI for the proportion is given by

$$\boxed{\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}.$$

Interpretation: "I am 100(1-α)% confident that the population mean or the proportion is between the CI."

CORRELATION

a statistical measure that describes the relationship between two continuous variables. It indicates how changes in one variable correspond to changes in another.

- Positive correlation – As one variable increases, the other also increases.
- Negative correlation – As one variable increases, the other also decreases.
- No correlation – The variables do not show a relationship.

$$-1 \leq \text{Correlation Coefficient} \leq 1$$

Correlation coefficient	Description
0	No correlation
∼0.2	Weak correlation
∼0.35	Moderate correlation
∼0.5	Moderate to strong correlation
∼0.7	Strong correlation
0.8	Very or extremely strong correlation
1	Perfect correlation

"Correlation does not imply causation", meaning that just because two variables are related, one does not necessarily cause changes in the other. Here's why:

- **Spurious Correlations:** Sometimes, two things are correlated purely by coincidence. For example, the number of people who drown in swimming pools and the number of films Nicolas Cage appears in each year are correlated, but clearly, Cage is not causing drownings!
- **Third Variables (Confounding Factors):** A hidden variable might be influencing both variables. Imagine that ice cream sales and drowning incidents are positively correlated. This does not mean eating ice cream leads to drowning. Instead, the hidden factor is hot weather, which increases both ice cream consumption and swimming (leading to more drownings).
- **Reverse Causation:** Sometimes, the assumed cause-and-effect relationship is actually the reverse. Suppose data shows that people who own more books tend to be smarter. It might not be that owning books makes someone smarter, it's more likely that smarter people buy more books!

I. Pearson's Correlation (r) – Linear Relationship

For a **sample**, Pearson's correlation is defined as

$$r = \frac{\text{cov}(x,y)}{s_x s_y}, \quad \text{cov}(x,y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}, \quad s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}, \quad s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$$

$$\boxed{r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}}$$

```
cor(iris$Petal.Length, iris$Petal.Width, method="pearson")
```

II. Spearman's Rank Correlation (r_s) – Monotonic Non-linear Relationships

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables. For a sample of size n, the n pairs of raw scores (X_i, Y_i) are converted to ranks $R[X_i], R[Y_i]$, and r_s is computed as

$$\boxed{r_s = \rho\,[\,R[X], R[Y]\,] = \frac{\text{cov}\,[\,R[X], R[Y]\,]}{s_{R[X_i]} s_{R[Y_i]}}}$$ similar to Pearson's but in the form of rank variables.

```
cor(iris$Petal.Length, iris$Petal.Width, method="spearman")
```

III. Kendall's Rank Correlation (τ) – Small Datasets or Many Tied Values

$$\boxed{\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}},$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and n is the sample size.

```
cor(iris$Petal.Length, iris$Petal.Width, method="kendall")
```

SIMPLE LINEAR REGRESSION

From a scatter plot of two variables with the dependent/response variable (Y) on the y-axis and the independent/predictor variable (X) on the x-axis, we find the **line of best fit**

$$\underbrace{\hat{y}}_{\text{Predicted response}} = \underbrace{b_0}_{y-\text{intercept}} + \underbrace{b_1}_{\text{Slope}} \underbrace{x}_{\text{Predictor}}, \quad \text{where} \quad b_1 = \underbrace{r}_{\text{Pearson's correlation}} \cdot \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \underbrace{\bar{y}}_{\text{Mean of } Y} - b_1 \underbrace{\bar{x}}_{\text{Mean of } X}$$

by minimising the **sum of squared residuals (SSR)** denoted by

$$SSR = \sum_{i=1}^{n} (\underbrace{y_i}_{\text{Observed}} - \underbrace{(b_0 - b_1 x)}_{\text{Predicted}})^2, \quad \text{where} \quad \underbrace{e_i}_{\text{Residuals}} = y_i - \hat{y}, \quad \text{and} \quad R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

	F-test for overall model	T-test for coefficients	T-test for correlation
Null hypothesis	$H_0 : \hat{y} = \bar{y}$ has more explanatory power	$H_0 : b_i = 0$	$H_0 : r = 0$
Alternative hypothesis	$H_1 : \hat{y} = b_0 + b_1 x$ has more explanatory power	$H_1 : b_i \neq 0$	$H_1 : r \neq 0$
Test statistics	$F = \frac{SSR_1 - SSR_2}{SSR_2} \cdot \frac{n-p}{p-1}$	$t = \hat{b_i}/SE(\hat{b_i})$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
Degrees of Freedom	$(p, n-p-1)$	$n-p-1$	$n-2$
Critical value	one-tailed F_{α,df_1,df_2}	two-tailed $t_{\alpha/2,df}$	two-tailed $t_{\alpha/2,df}$
Reject H_0 if	$F > F_{\alpha,df_1,df_2}$	$\lvert t\rvert > \mathrm{t}_{\alpha/2,df}$	$\lvert t\rvert > \mathrm{t}_{\alpha/2,df}$

Interpretation of Model Results

- **y-intercept (b_0)**: This is where the line crosses (intercepts) the y-axis or where $x = 0$.
- **Slope (b_1)**: This is the gradient of the line, which tells us how much y increases for every unit increase in x. E.g. for every 1 unit increase in x, y increases/decreases by b_1 units.
- **Coefficient of determination (R^2)** indicates how well a regression model explains the variability of the dependent variable. It ranges from 0 to 1, where 0 suggests that the model does not explain any of the variability in the dependent variable, and 1 suggests that the model perfectly explains this variability. E.g. if $R^2 = 0.8$, this means that 80% of the variability in the dependent variable (Y) can be explained by the independent variable (X) in the regression model.

```
l1 <- lm(Petal.Length~Petal.Width, data=iris)
summary(l1) # show a summary of the model
plot(Petal.Length~Petal.Width, data=iris)
abline(lm(Petal.Length~Petal.Width, data=iris))
```

Assumptions of Linear Regression

- **Lack of (influential) outliers:** All responses were generated from the same process, so that the same regression model is appropriate for all the observations.
- **Linearity:** The linear predictor captures the 'true' relationship between expected value of the response variable and the explanatory variables (and all important explanatory variables are included).
- **Independence:** The observations are statistically independent of each other.
- **Constant variance (homoscedasticity):** The residuals have constant variance.
- **Normality of Residuals:** The residuals are normally distributed.

```
plot(l1) # Diagnostic plots for checking assumptions
hist(l1$residuals) # Histogram of residuals to check the normality
```

- **Residuals vs. Fitted Plot** – Checks **Linearity & Homoscedasticity**
  If residuals are randomly scattered around the horizontal line, the linearity assumption holds. If the spread of residuals increases or decreases systematically, it suggests heteroscedasticity (non-constant variance).
- **Normal Q-Q Plot** – Checks **Normality of Residuals**
  If residuals follow a straight diagonal line, they are normally distributed. Deviations from the line indicate non-normality, which may affect inference.
  Scale-Location Plot (Spread-Location Plot) – Checks **Homoscedasticity (Equal Variance of Residuals)**
  If residuals are evenly spread across fitted values, the assumption holds. A systematic pattern suggests heteroscedasticity.
- **Residuals vs. Leverage Plot** – Checks **Influential Observations (Outliers & High Leverage Points)**
  Points outside Cook's distance lines indicate influential observations that may disproportionately affect the model.