

BSc Neurosciences

Introduction to statistics for neuroscientists

Cynthia Sandor

Edmond J Safra Assistant Professor
UKRI Future Leader Fellow
UK Dementia Research Institute Group Leader



UK Dementia
Research Institute

Imperial College
London



Founding funders:



Medical
Research
Council



EDMOND J. SAFRA
PHILANTHROPIC FOUNDATION

Agenda

Session 1- 45 minutes

Theory 30 minutes

- Introduction & Motivation

- Descriptive statistics & visualization

- Probability basics & expected value

- Distributions & Central Limit Theorem

Practical with R 15 minutes

Break 10 minutes

Session2 - 45 minutes

Theory 30 minutes

- Hypothesis testing, p-values, and errors

- Multiple testing & effect size

- Parametric vs non-parametric tests, t-tests

- Experimental design & power

Practical with R 15 minutes

Why Statistics?

- **Statistics** = science of data (collection, organization, analysis, interpretation)
- **Central in neuroscience:** separating *signal* from *noise*
- **Key questions:**
 - What is the same/different between groups?
 - What is significant vs. insignificant?
 - How to design studies with power?

Motivating example: Is Tasty Beer cheating its customers?

- “Is Tasty Beer cheating its customers?”
- Small deviations in 0.5L bottles → suspicion of underfilling
- How can we test this?

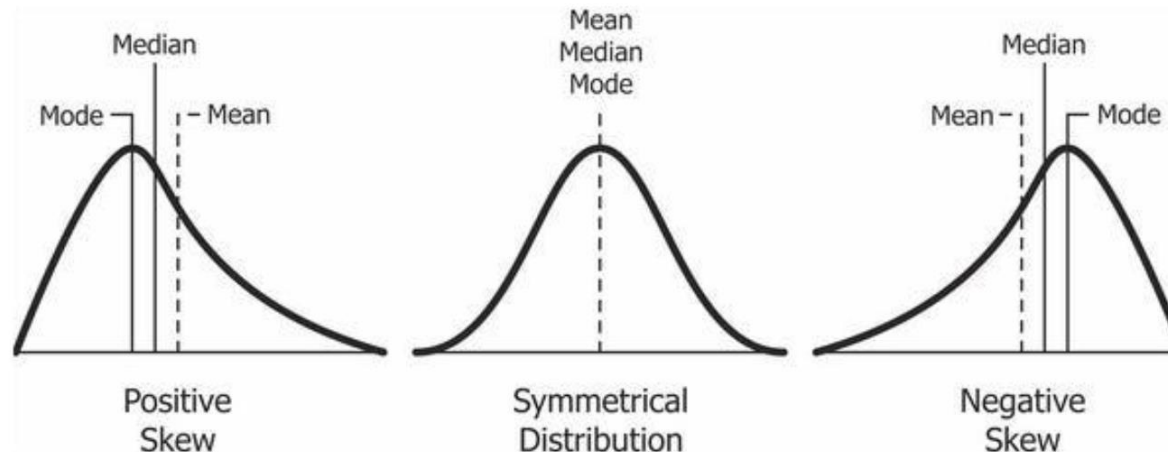


What are Descriptive Statistics?

- Summarize and describe data
- First step before testing hypotheses
- Measures of central tendency & variability

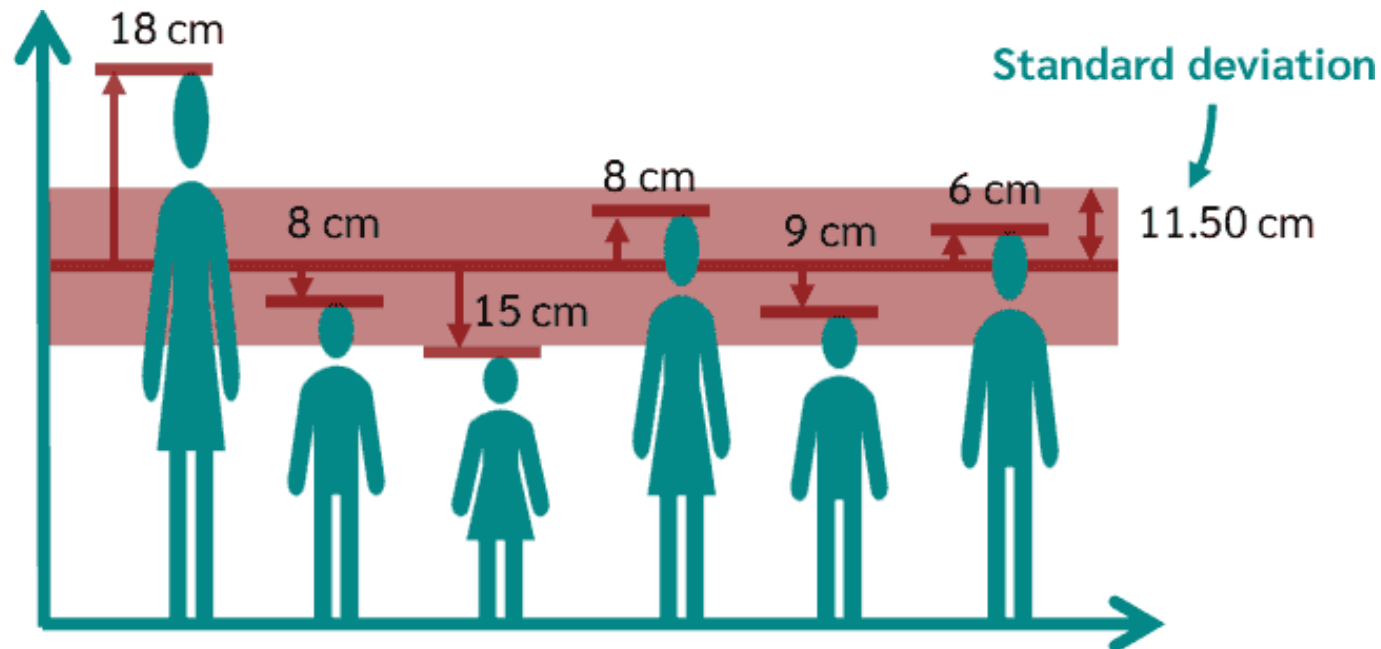
Measures of Central Tendency

- **Mean:** Sum of all values divided by the total number of values
- **Median:** Middle value (when the data is arranged in order)
- **Mode:** Most common value (the value with the most frequency)



Variability

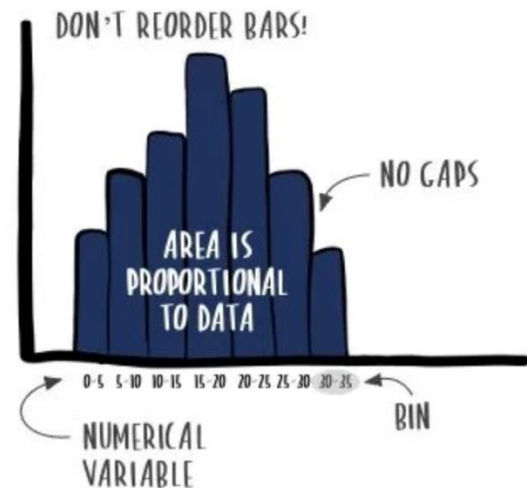
- **Range** = Highest value – Lowest value
→ Quick measure of spread
- **Variance (Var)** = Average squared distance from the mean
- **Standard Deviation (SD or σ)** = Square root of variance
→ “Average distance” of values from the mean



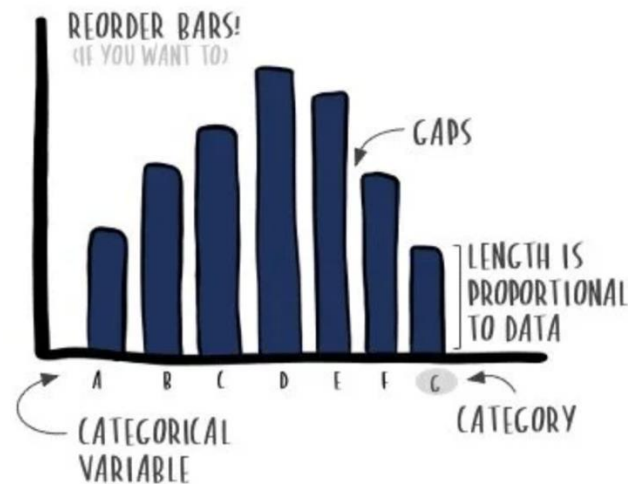
Visualisation Data

- **Histogram** → shows frequency of values in ranges
- **Bar chart** → compares categories
- **Pie chart** → shows proportions of categories

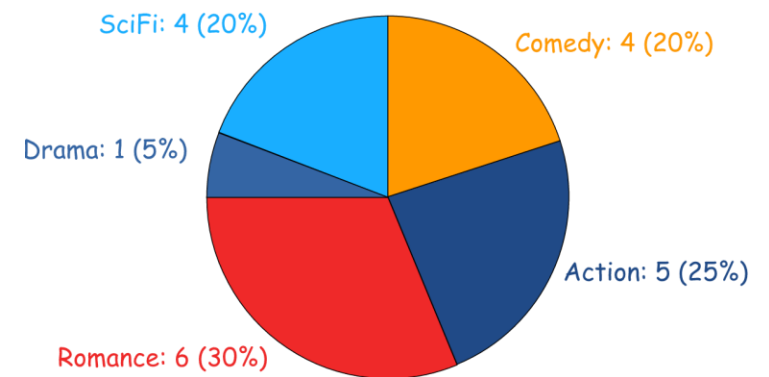
This is a **histogram**...



This is a **bar chart**...



Favorite Type of Movie



Visualisation Data

Boxplot (five-number summary):

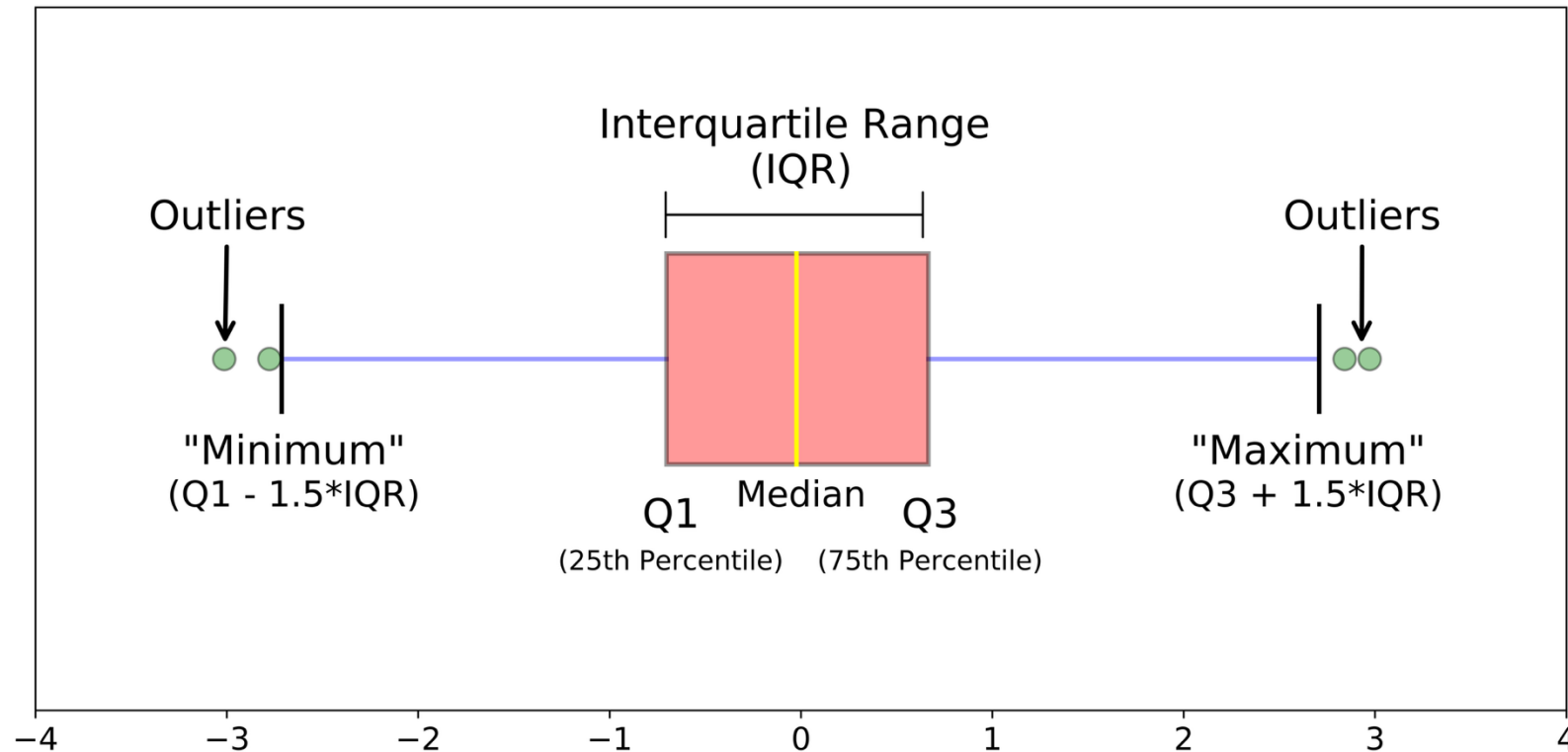


Table of summary statistics: Mean, Median, SD, etc.

Probability Basics

- **Probability** = likelihood of an event (0 = impossible, 1 = certain)

Example The probability of getting heads in a coin toss

A probability is always between 0 (impossible) and 1 (certainty)

Certainty (Probability = 1):

A packed London Tube at rush hour (inevitable chaos!)

Impossibility (Probability = 0):

Snow falling in the Sahara

Somewhere in Between ($0 < p < 1$):

Probability that your favourite team will win



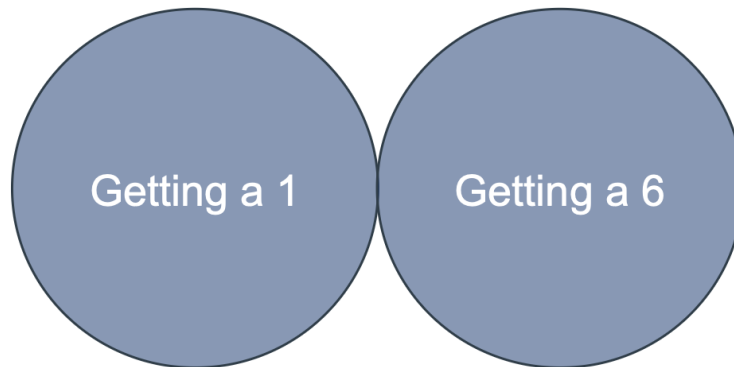
Probability Basics

Addition rule: The joint probability of two mutually exclusive events is the sum of their probabilities: $P(A \text{ or } B) = P(A) + P(B)$

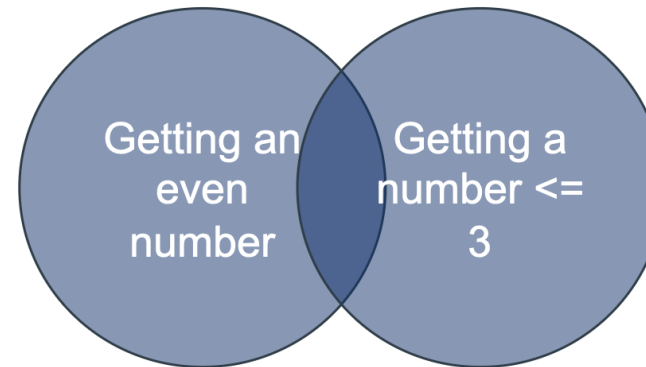
- **Mutually exclusive:** Only one of the events can occur



Exclusive



Non-exclusive



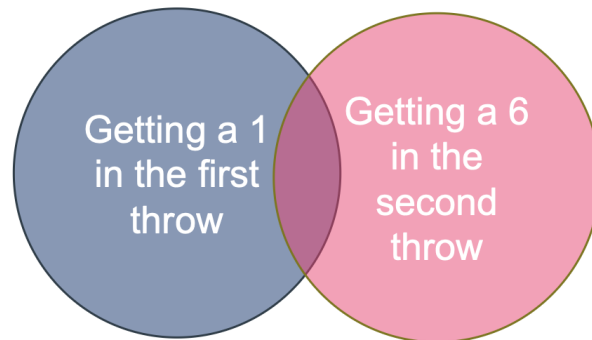
Probability Basics

Multiplication rule: The probability of two independent events both occurring is the product of their probabilities: $P(A \text{ and } B) = P(A) * P(B)$

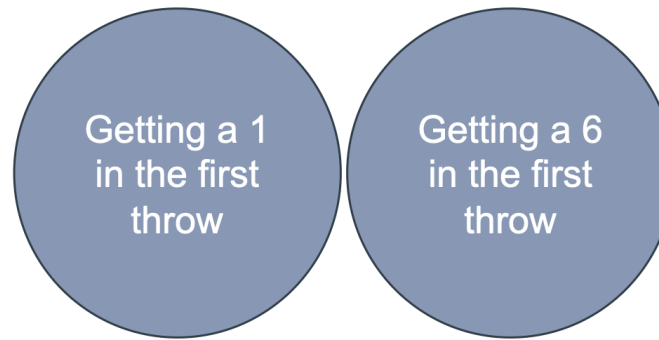
- **Independent** = the occurrence of one event does not affect the other
- Example: Getting heads in the first round and tails in the second round



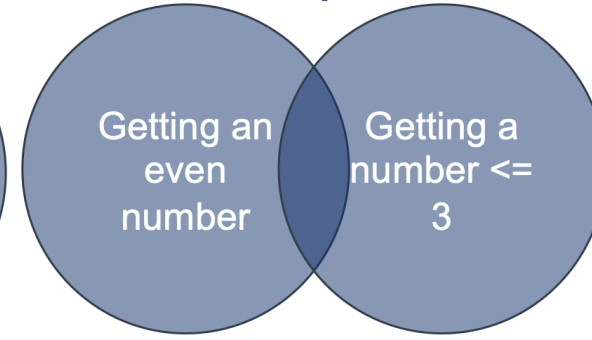
Independent



Not independent



Not independent

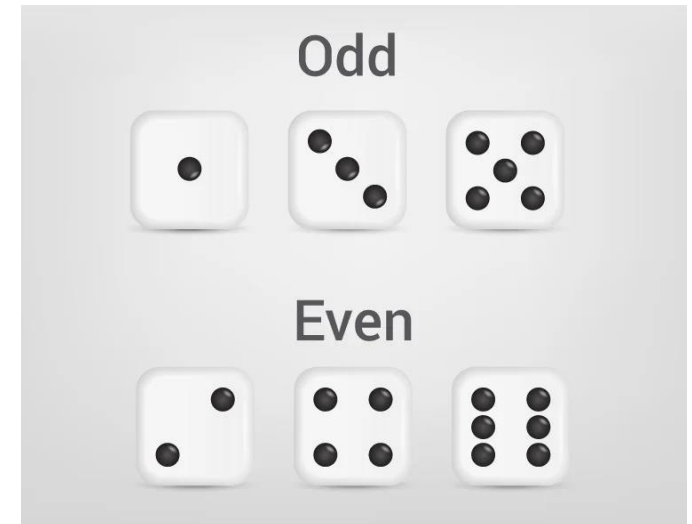


Probability Basics

Calculating probability ! = Summing up events

- $\sum \text{Events that lead to outcome} / \sum \text{All events}$
- This only works for equally probable and mutually exclusive events

What is the probability of throwing an even number?



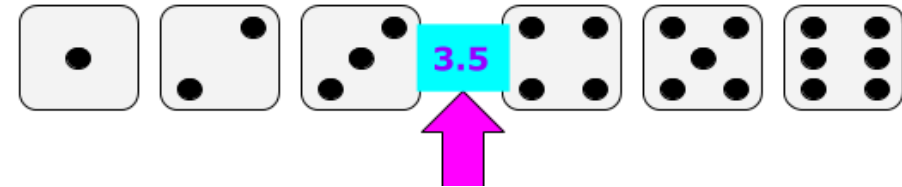
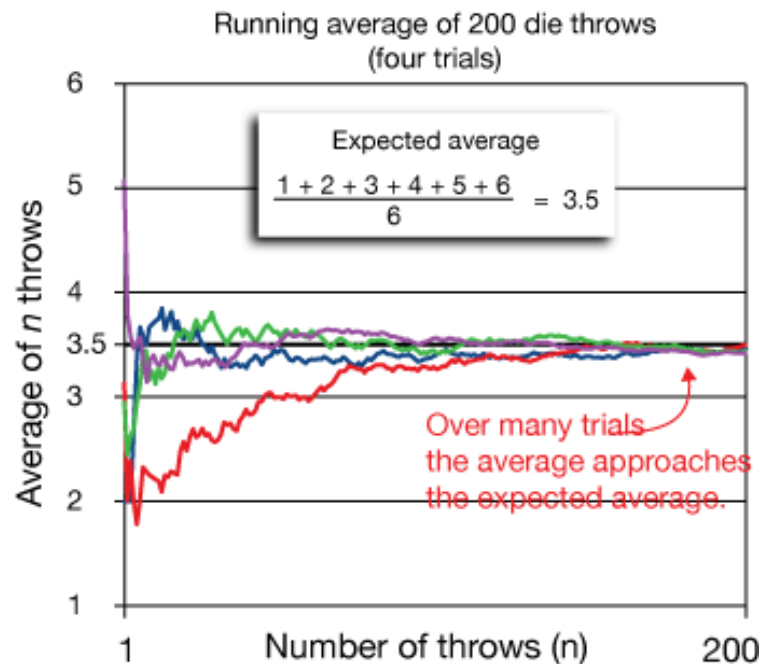
Expected Value & Law of Large Numbers

Expected Value (EV): long-run average outcome

Example (die roll): $EV = (1+2+3+4+5+6)/6 = 3.5$

Law of Large Numbers:

As $n \rightarrow \infty$, the sample average \rightarrow expected value



Probability Distributions

Discrete: only certain values (coin toss, number of goals a player has scored this season)

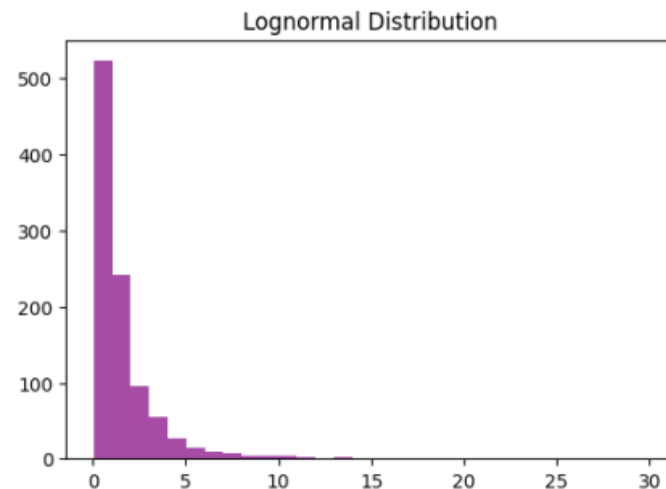
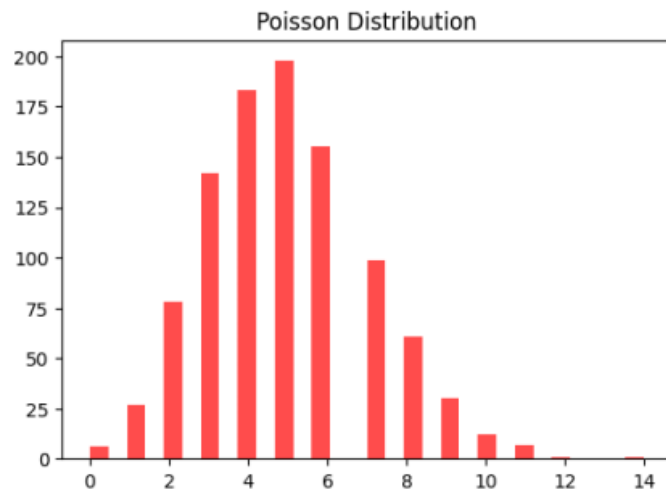
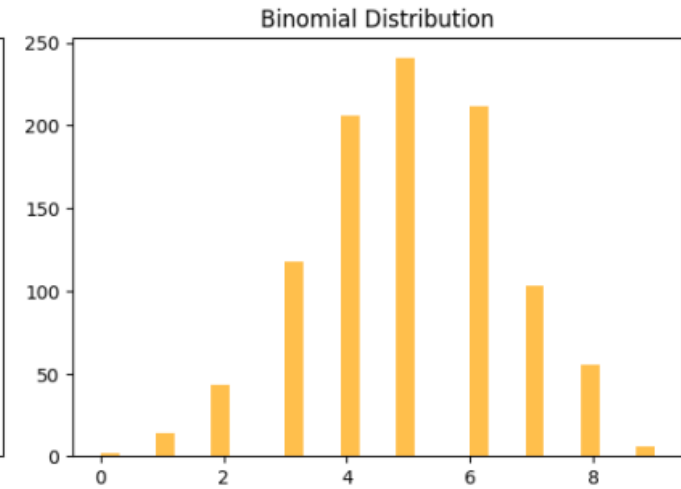
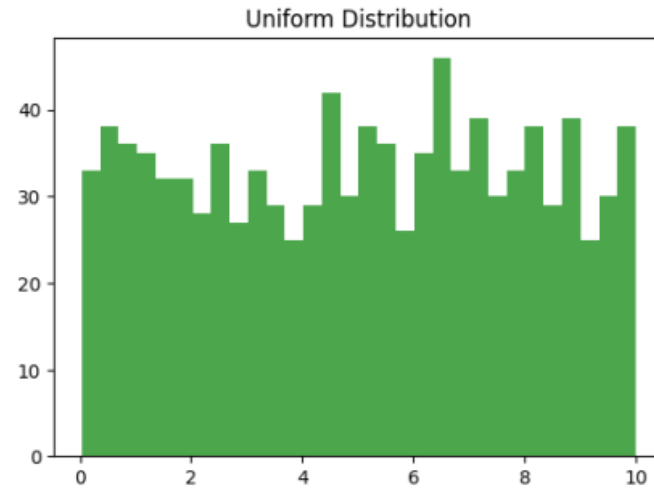
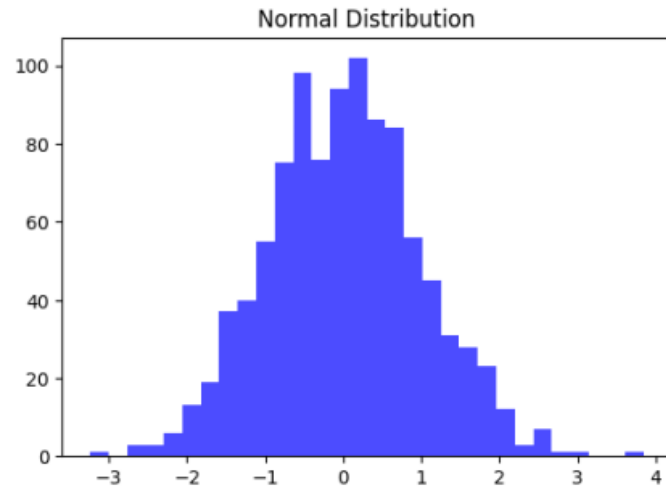


Continuous: any value in a range (height, car speed,)



Probability Distributions

- A **probability distribution** tells us how likely different outcomes are



Probability Distributions: common

- **Binomial distribution:** “yes/no” outcomes in repeated trials e.g. number of heads in 10 coin tosses
- **Poisson distribution:** counts of events in a fixed time or space e.g. number of spikes in 1 second
- **Normal distribution:** continuous “bell curve” e.g. height, reaction time, blood pressure

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

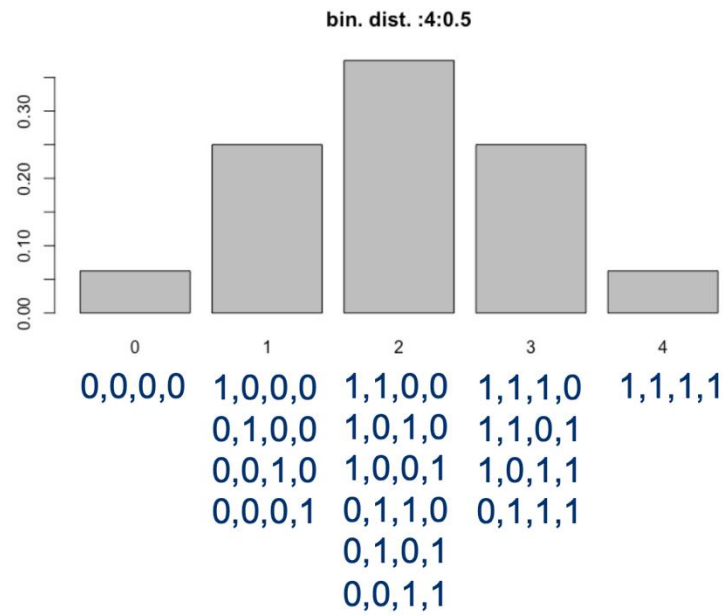
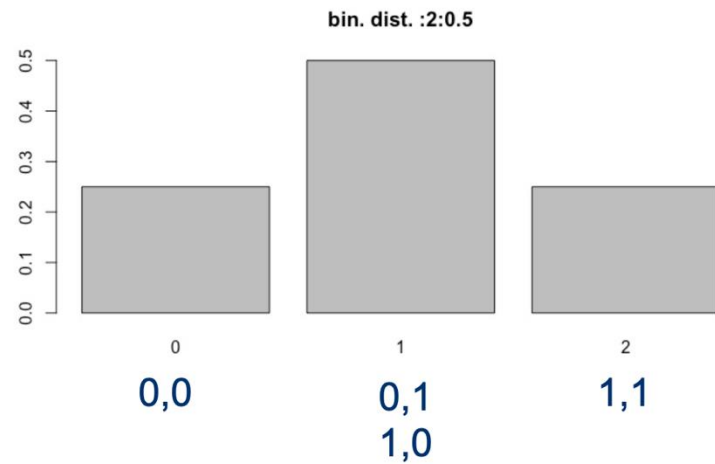
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Probability Distributions: common

- Binomial distribution:** “yes/no” outcomes in repeated trials e.g. number of heads in 10 coin tosses

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$



Probability Distributions: common

- **Poisson distribution: counts of events in a fixed time or space**

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



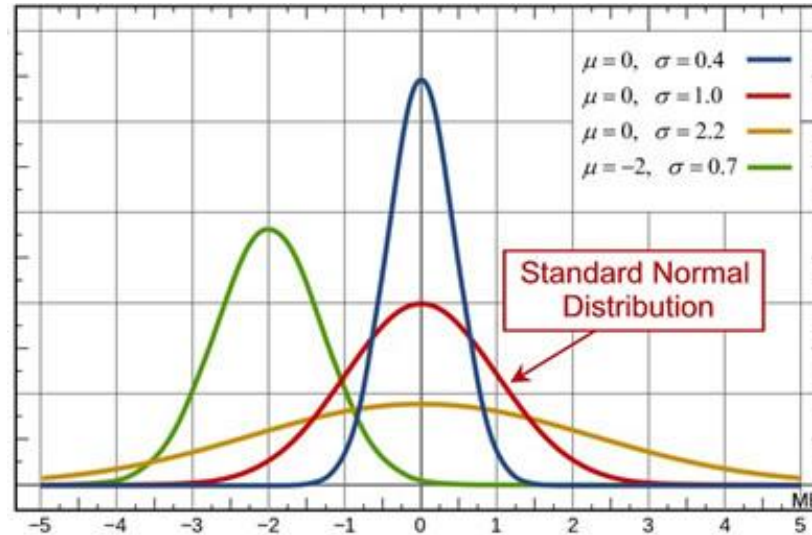
On average, 3.6 people arrive at a booking counter every 10 minutes on weekends. What is the probability that exactly 7 people arrive in 10 minutes?

Probability = 0.0424 ($\approx 4.2\%$)

$$P(7; 3.6) = \frac{e^{-3.6} \cdot 3.6^7}{7!}$$

Probability Distributions: common

- **Normal distribution:** continuous “bell curve” e.g. height, reaction time, blood pressure

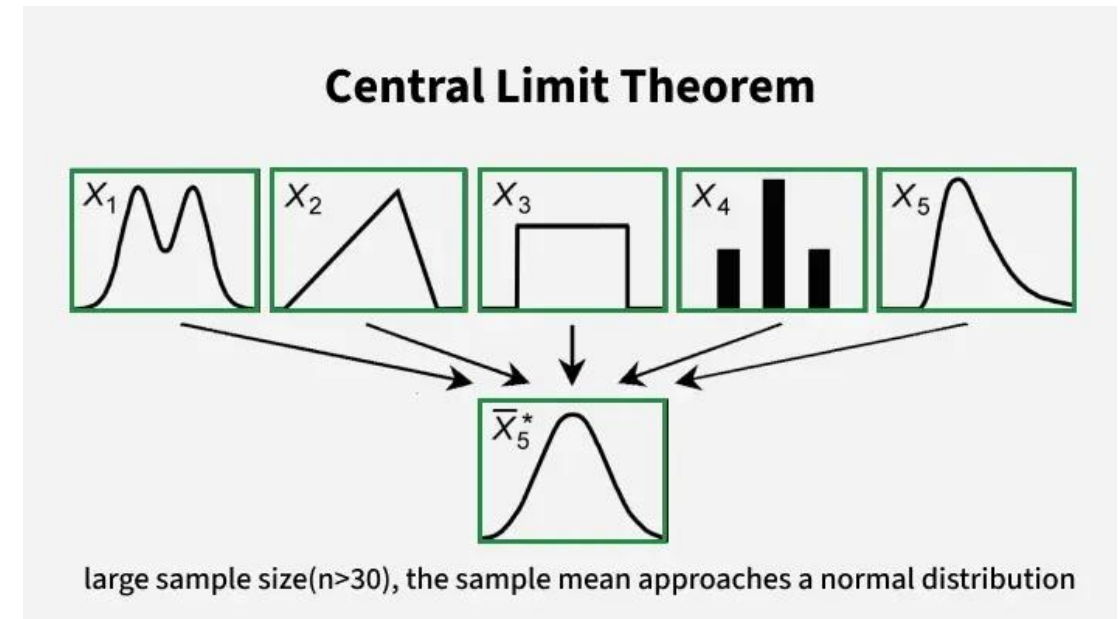


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Probability defined by only two parameters (mean and standard deviation)
- Symmetry
- Mean, median and mode are identical
- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean

Central Limit Theorem (CLT)

- Regardless of population distribution, the sample mean tends to be normal as sample size \uparrow
- The distribution of the sample mean: $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- Why it matters:** Justifies t-tests, ANOVA, regression, Explains why averaging trials reduces noise



Session1: Practical with R 15 minutes

<https://github.com/csandorfr/brains-and-stats>

The screenshot shows the GitHub repository page for **csandorfr / brains-and-stats**. The repository is public and has 0 stars, 0 forks, and 0 watches. The main branch is **main**. The file list includes:

- `.github/workflows` (Create jekyll-gh-pa)
- `course` (Add files via upload)
- `exercises` (Update ex2_hypot)
- `scripts` (Update 07_multiple)
- `README.md` (Update README.m)
- `setup.md` (Rename syllabus.m)
- `syllabus.md` (Create syllabus.md yesterday)

The **README** section is titled **Introduction to Statistics for Neuroscientists (BSc Neuroscience)**. A **Clone** dropdown menu is open, showing options for **Local** and **Codespaces**. The **Clone** options are:

- HTTPS** (selected): `https://github.com/csandorfr/brains-and-st`
- SSH**
- GitHub CLI**

The **About** section on the right includes:

- Readme**
- Activity**
- 0 stars**
- 0 watching**
- 0 forks**
- Releases**: No releases published. [Create a new release](#)
- Packages**: No packages published. [Publish your first package](#)
- Deployments**: 11. [github-pages](#) 38 minutes ago

Session1: Practical with R 15 minutes

Running Scripts 01–03 in Rstudio

- **Open RStudio** → check R and RStudio are installed
- **Open script:** brains-and-stats/exercices/scripts/01–03_*.R
- **Run code:** Ctrl/Cmd + Enter (line) or **Source** (entire script)
- **Explore:** View plots → modify sample size, mean, SD

Session1: Practical with R 15 minutes

01_descriptive_stats_base.R

```
1 # =====
2 # Introduction to Descriptive Statistics (base R)
3 # =====
4
5 # In this exercise, we explore how to describe data using:
6 # - Mean: average value
7 # - Median: middle value
8 # - Mode: most frequent value
9 # - Range: max - min (spread of data)
10 # - Variance and SD: average squared deviation from the mean (and its square root)
11 # - How skewed data affect these measures
12
13 # -----
14 # 1. Define a simple function to compute the MODE
15 # -----
16 # R does not have a built-in mode() function for numeric data,
17 # so we create one that finds the most frequent rounded value.
18 get_mode <- function(x) {
19   tab <- table(round(x))           # Count how often each rounded value appears
20   as.numeric(names(tab)[which.max(tab)]) # Return the value with the highest count
21 }
22
23 # -----
24 # 2. Simulate three different distributions
25 # -----
26 set.seed(1) # ensures reproducibility (same random numbers each time)
```

Console

R - R 4.4.2 · ~/ /

```
> set.seed(1) # ensures reproducibility (same random numbers each time)
>
> # (a) Symmetrical distribution (e.g. IQ scores)
> # Normally distributed data: mean = 100, SD = 15
> iq <- rnorm(1000, mean = 100, sd = 15)
>
> # (b) Positively skewed distribution (e.g. reaction times)
> # Exponential: most values are small, few are large → long right tail
> reaction <- rexp(1000, rate = 1/300)
>
> # (c) Negatively skewed distribution (e.g. age of onset)
> # "Flipped" exponential: most values are large, few are small → long left tail
> age <- 100 - rexp(1000, rate = 1/10)
>
```

Environment

History

Connections

Project: (None)

Import Dataset

29 MiB

List

R

Global Environment

reaction

num [1:1000]

180.3 225.1 50.1 661.6 16.6 ...

rt

num [1:5]

220 230 400 240 250

rtC

num [1:10000]

228 315 427 210 294 ...

rtD

num [1:10000]

406 210 420 170 378 ...

scores

num [1:200]

65 61 69 58 97 86 77 71 65 88 ...

sd

0.01

sd_pop

2.17944947177034

sd_rt

80

Files

Plots

Packages

Help

Viewer

Presentation

Zoom

Export

Publish

Symmetrical (IQ)

Positive Skew (Reaction Tim

Negative Skew (Age of Onse

Frequency

Value

ms

Years



Quick Check Questions – Session 1

Q1: Why do we need statistics in neuroscience?

Q2: You collect reaction times (in ms) from 5 students: **220, 230, 400, 240, 250**.

- What is the **mean**?
- What is the **median**?
- Which measure (mean or median) is more robust to the outlier (400)?

Q3: A neuron fires on each trial with probability **$p = 0.2$** . What is the expected number of spikes in 10 trials?

Q5: Suppose you flip a coin 5 times and calculate the proportion of heads. If you repeat this experiment many times, what will the distribution of these proportions look like as the number of repeats grows?

Hypothesis Testing & P-values

- **Hypothesis testing** asks: is the observed effect real or just chance?
- Two competing hypotheses:
 - **Null hypothesis (H_0)**: no effect / no difference
 - **Alternative hypothesis (H_1)**: there is an effect / difference
- **Test statistic** → measures signal vs. noise
- **P-value** = probability of observing this data (or more extreme) **if H_0 is true**
- If $p < \alpha$ (usually 0.05) → reject H_0

Hypothesis Testing & P-values

Tasty Beer produce 0.5l beer bottles:

- There is a little bit of variation in the amount of beer in each bottle
- The consumer production agency thinks they are cheating their customers, so they take a random sample of bottles to look at:

Sample: 0.499 0.488 0.478 0.508 0.490 0.504 0.488 0.502 0.508 0.506



- How can they test whether *Tasty Beer* is really filling too little beer into their bottles?

Hypothesis Testing & P-values

- H_0 = *Tasty Beer* fills at least 0.5l into each bottle on average
- H_1 = *Tasty Beer* fills less than 0.5l into each bottle on average

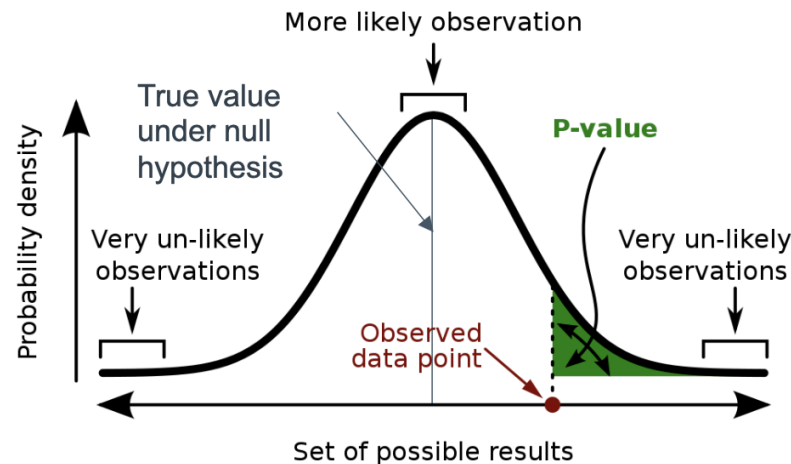
How likely are we to observe the sampled values if we assume that H_0 is true (*Tasty Beer* fills at least 0.5l on average into each bottle)?



Hypothesis Testing & P-values

What is the probability (P) of observing the sampled data under the null hypothesis? • If P is less than a selected threshold α (often 0.05), the null hypothesis is rejected – Statistical significant

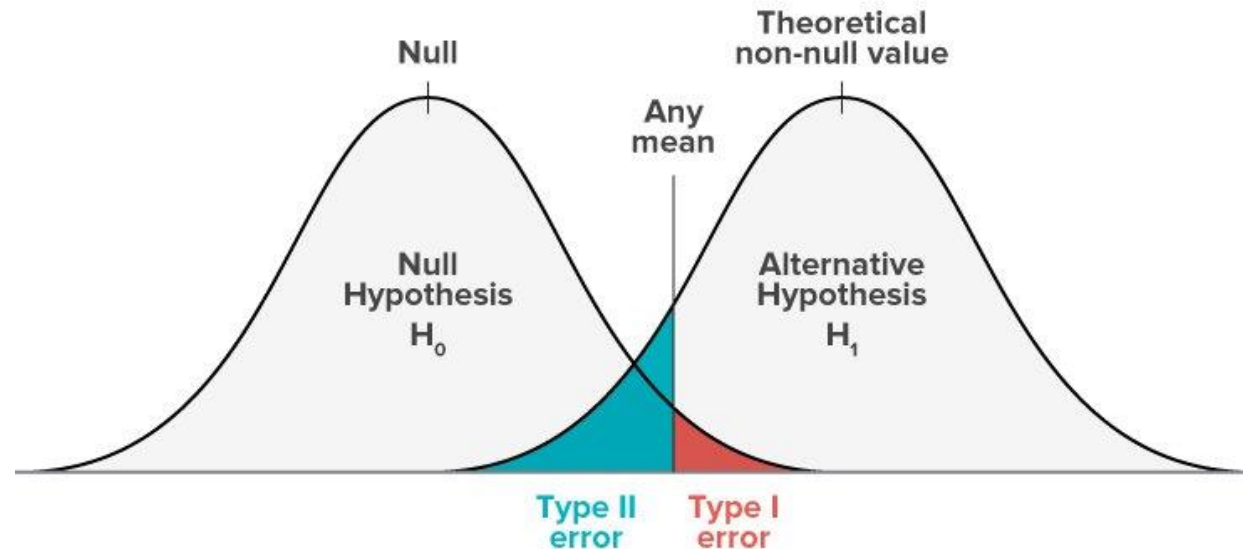
The P value is defined as the probability, under the null hypothesis H_0 , of obtaining a result equal to or more extreme than what was actually observed – $P(\text{result} \mid H_0)$.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Errors in Hypothesis Testing

- **Type I error (false positive):** reject H_0 when it's true
Probability = α (significance level, usually 0.05)
 - **Type II error (false negative):** fail to reject H_0 when H_1 is true
Probability = β
 - **Power = $1 - \beta$** = chance of correctly detecting a true effect
- Balancing errors is key to good study design



Multiple Testing Problem

- Each test has **false positive rate** = α (usually 0.05)
- Doing many tests increases chance of false positives
- Probability of at least one false positive after n tests:

$$P = 1 - (1 - \alpha)^n$$

- Example:

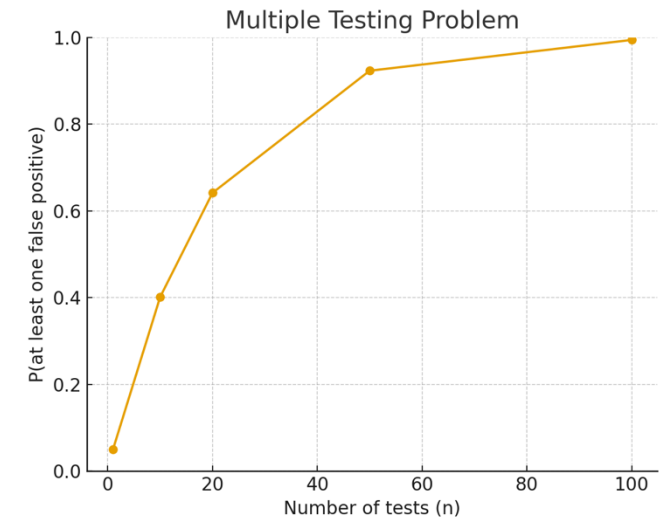
10 tests → 40% chance of ≥ 1 false positive

20 tests → 64%

50 tests → 92%

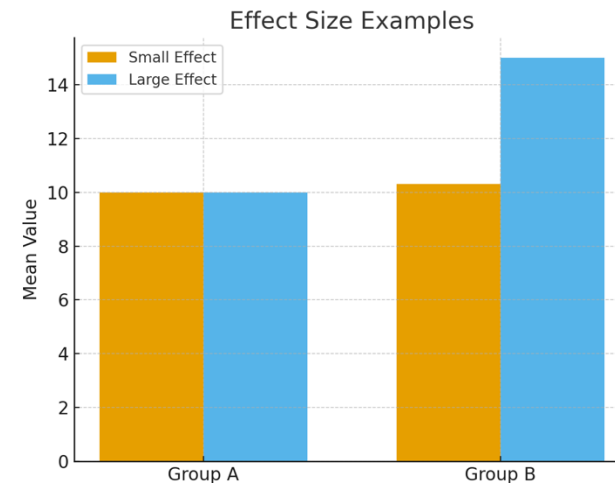
- **Solutions: Bonferroni correction**
- Other methods: FDR, Holm–Bonferroni

$$\alpha_{\text{adj}} = \frac{\alpha}{n}$$



Effect Size

- **P-values** only tell us if an effect is likely not due to chance
- **Effect size** tells us how *big* the difference really is
- Examples:
 - Clinical trial: drug reduces symptoms by 1% vs 30%
 - Neuroscience: brain volume difference of 1 mm³ vs 100 mm³
- Common measures:
 - Cohen's d** (mean difference / SD)
 - Correlation coefficient (r)**



Parametric vs. Non-parametric Tests

- **Parametric tests**
Assume data follow a known distribution (often normal)
Examples: t-test, ANOVA, Pearson correlation
- **Non-parametric tests**
“Distribution-free,” no assumption of normality
Examples: Wilcoxon, Mann-Whitney U, Spearman correlation
- **Why not always non-parametric?**
Parametric tests usually have **more power**
Easier to model population & confounders
More flexible (e.g., regression models)

The T-test

- **Goal:** compare means
- **Types:**
 - One-sample t-test:** compare mean to known value
 - Independent two-sample t-test:** compare means of 2 groups
 - Paired t-test:** compare before/after in the same group
- **Formula**

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

- **Signal** = difference from null mean
- **Noise** = standard error (spread / \sqrt{n})

Experimental Design

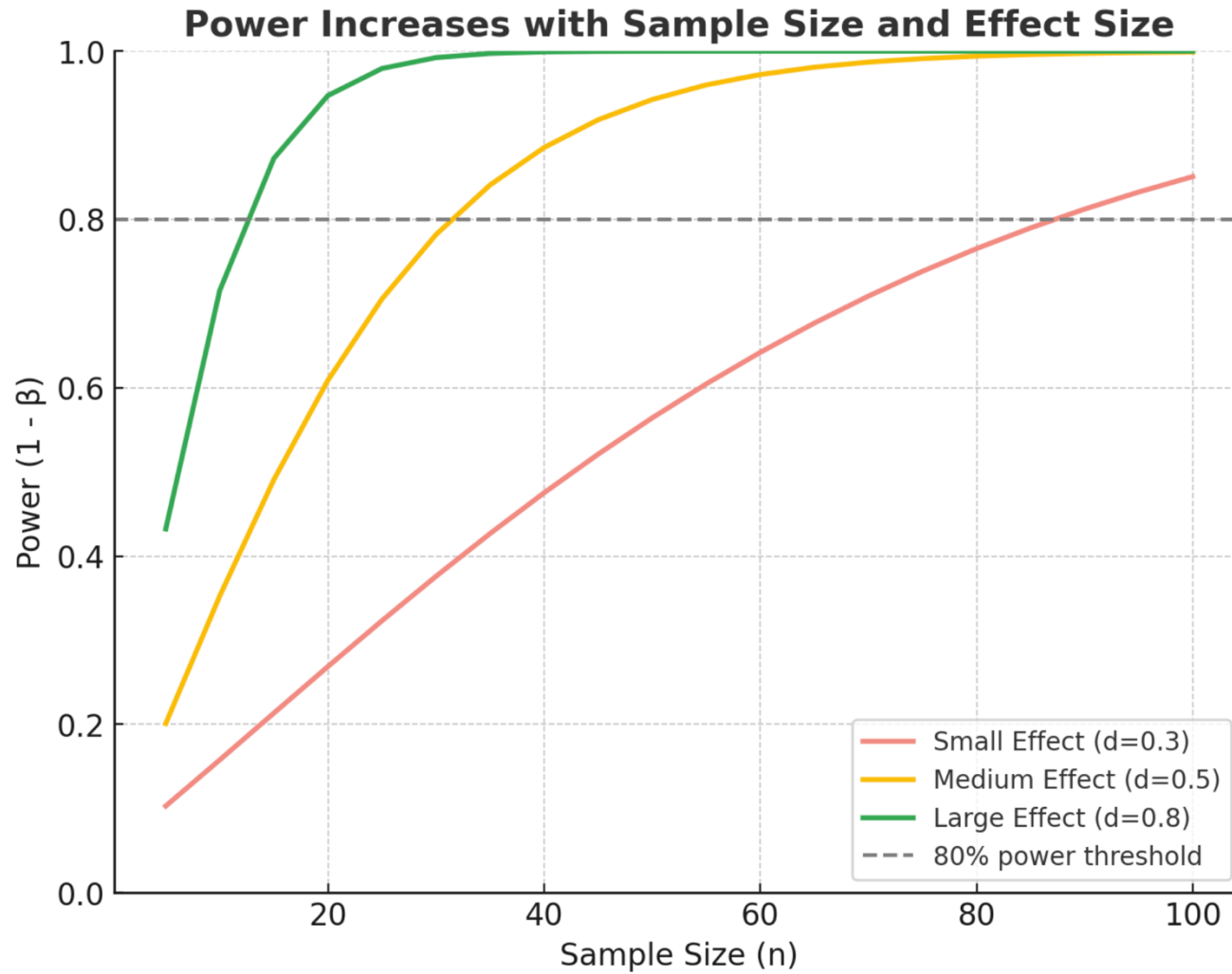
- **Randomized Controlled Trials (RCTs):**
Participants randomly assigned to control vs. experimental group
Goal: reduce bias, isolate treatment effect
- **Advantages:**
Washes out population bias
Reduces observer bias
- **Challenges:**
Needs large enough sample size
Balancing cost vs. power

Power & Sample Size

- **Power = $1 - \beta$**
 - Probability of detecting a true effect
- Depends on:
 - **Effect size** (bigger effects easier to detect)
 - **Sample size (n)** (more data → less noise)
 - **Significance level (α)**
- **Formula (sample size estimate):**
 - Z = z-score for confidence (e.g., 1.96 for 95%)
 - σ = standard deviation
 - E = margin of error

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

Power & Sample Size



Session 2: Practical with R 15 minutes

Running Scripts 04–07 in Rstudio

- **Open RStudio** → navigate to brains-and-stats/exercices/scripts/
- **Open scripts:** 04_pvalue_simulation.R → 07_multiple_testing_effect_size.R
- **Run code:** Ctrl/Cmd + Enter (line) or **Source** (entire script)
- **Explore:** Change α , n , or distributions → see effects on p-value & power

Quick Check Questions – Session 2

Q1: You compare patient and control brain volumes. What is your null hypothesis (H_0)? What does a p-value = 0.01 mean in this context?

Q2: You run 20 independent statistical tests, each with a false positive rate of $\alpha = 0.05$. Compute the probability of getting at least one false positive. Why might effect size be more useful than just a p-value?

Q3: In a class of 30 students, the number of siblings follows a Poisson distribution with mean 1.6. Simulate 30 values and plot a histogram. Calculate the median and interquartile range (IQR). Why might a non-parametric test be more appropriate here?

- Reaction times (ms) are measured before and after an intervention in 20 participants. Simulate pre values from a normal distribution (mean = 320, SD = 40).
- Generate post values that are on average 25 ms faster (with variability). Test whether the difference is significant using a paired t-test.

Q4: Why is an underpowered study problematic in neuroscience?

Take-home Messages

- **Visualize your data first** – plots tell you more than tables
- **P-values \neq truth** – always think about effect size
- **Beware of errors & multiple testing** – false positives are easy to create
- **Good design & power matter** – plan before you collect data