Module 3

Introduction to Statistical Analysis and EHR Data

Cynthia Sandor

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

UK Dementia Research Institute
Imperial College London
THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH
Founding funders:
UKRI Medical Research Council
EDMOND J. SAFRA PHILANTHROPIC FOUNDATION

Announcements

Please use the calendar on MT or the link sent by Claire — **not** the automatically generated one. There is **nothing scheduled on 19th November**.

•Please ensure you are using the **correct Python version and libraries** (avoid **Python 3.14**).

•**Windows encoding issue** when writing the .sam file — please use:
with open("../results/extracted_sam.sam", "w", encoding="utf-8") as file_w:

•Students who have not finished **Tutorial 1**, please move on to **Tutorial 2** and ask questions on MT. You can use the **consolidation sessions on 13 / 14 / 19 / 20 November** to finish tutorials and exercises.

What We'll Do Today

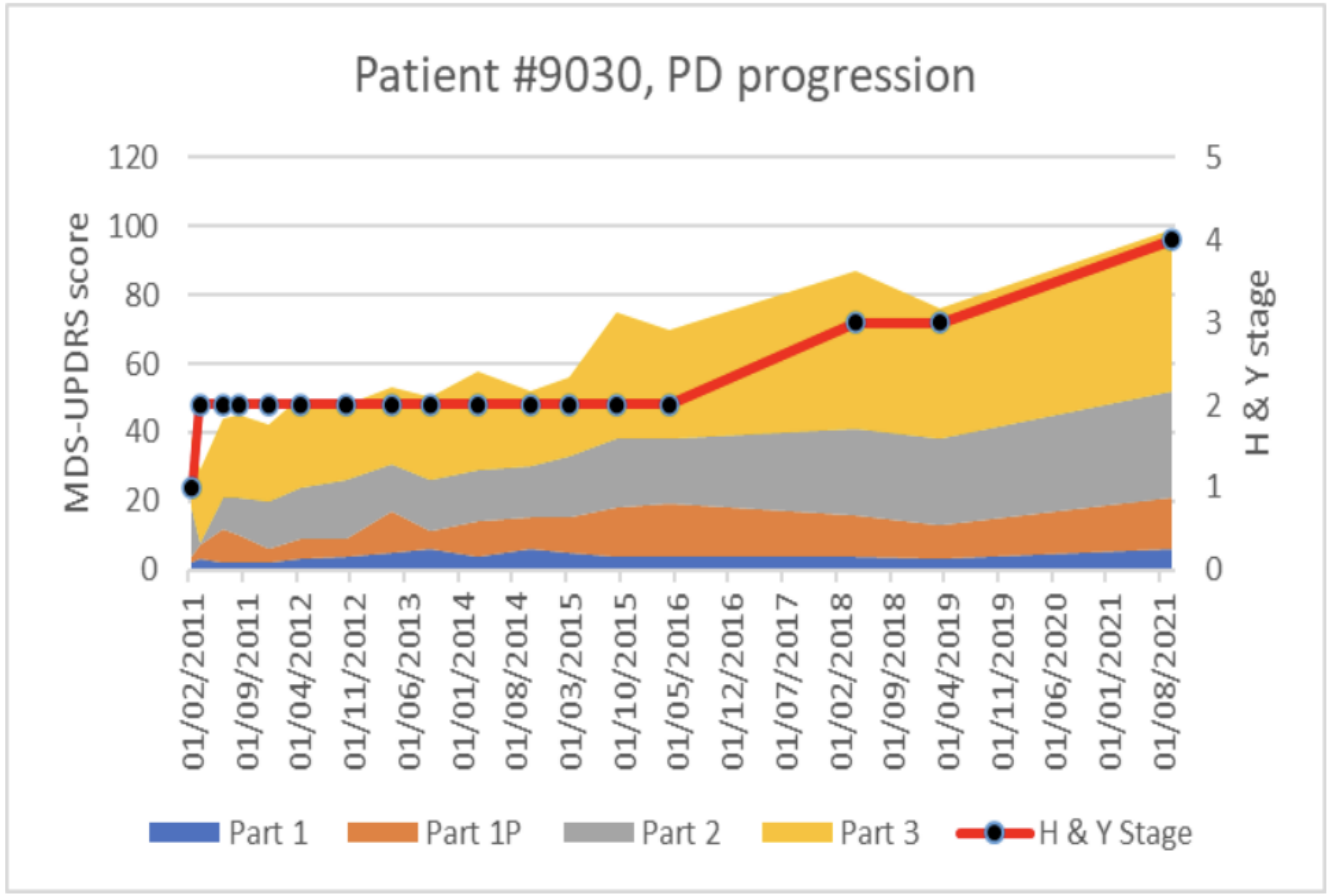
- **Lecture:** Short introduction to EHR and Parkinson's medications
- **Lecture:** Introduction to Statistics
- **3 Tutorials / Challenges** — Overall aim: develop skills for statistical analysis of longitudinal medical data
  1/ Clean and transform EHR tables in pandas; create new features with functions and .apply
  2/Visualise distributions and relationships, identify outliers and duplicates, and handle missing data
  3/Run and interpret statistical tests, check assumptions, measure associations, handle multiple comparisons, and visualise results

Approaches for tracking PD progression: MDS-UPDRS

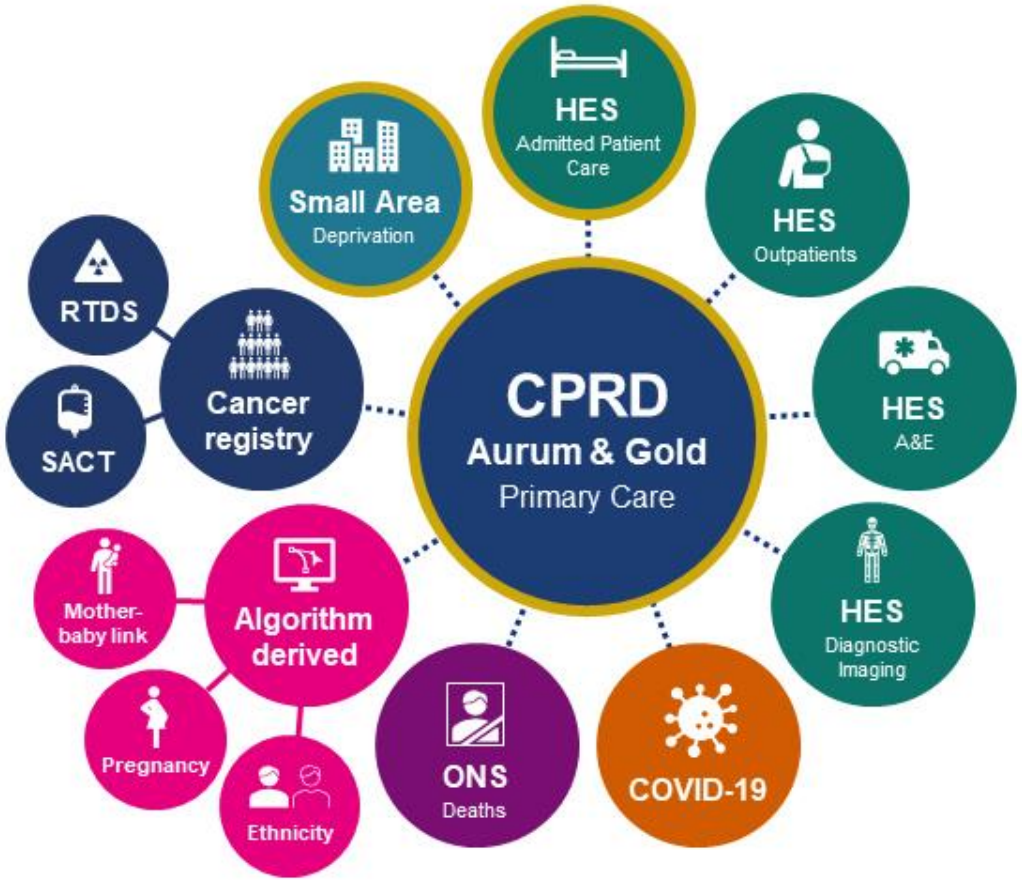
- **Current gold standard:** Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (Holden et al., 2017)
- Clinical questionnaire that assesses **both motor and non-motor** manifestations of PD. Performed by clinicians.

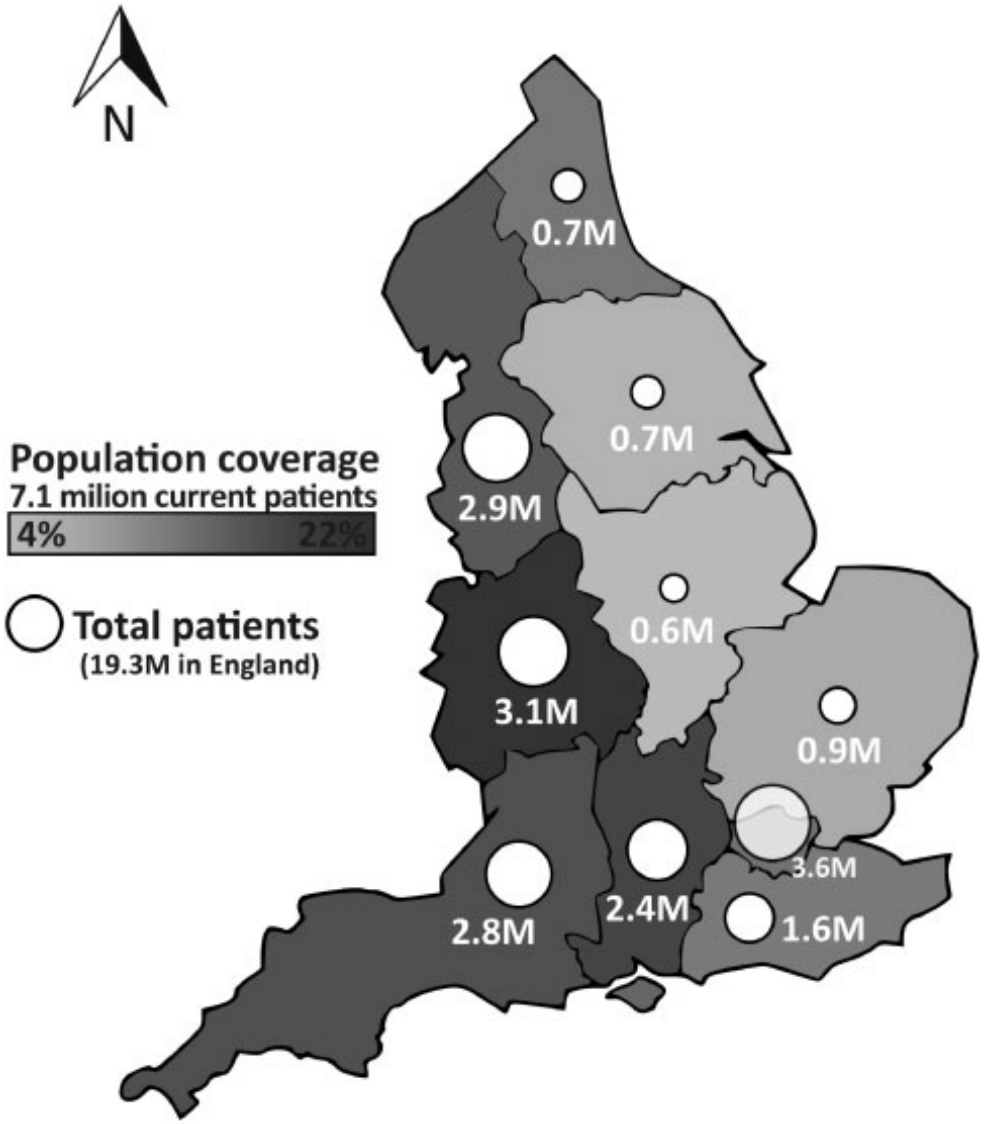
Problems:

- **Sparse** metric (~1-2 times per year, lower in low socio-economic areas) **=> no continuous progression**
- Require visit to a clinic
- Subjective
- Lack of sensitivity



Largest EHR: Clinical Practice Research Datalink | CPRD





UK Dementia Research Institute

LEDD trajectory calculation methodology

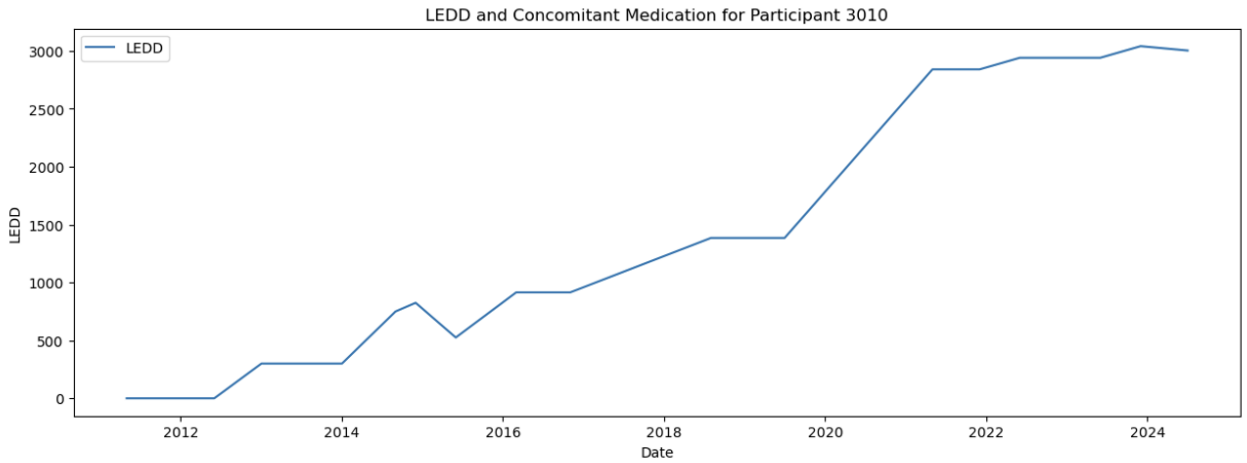
1) Calculate LEDD for each prescription
   1) $LEDD = daily\_dose * substance\_strength * conversion\_factor$
   2) e.g. LEDD (controlled-release levodopa):
      4 tablets a day * 125 mg * 0.75 = 375
   3) Some medication have fixed LEDD despite of the dose
      => LEDD(Safinamide) = 150
2) Establish the end date for each prescription
3) For each time point, sum the LEDDs for active prescriptions at this timepoint
   1) COMT do not contribute to summation
4) For each timepoint, If COMT prescription(s) are active at this time point, multiply by product of COMT conversion factors

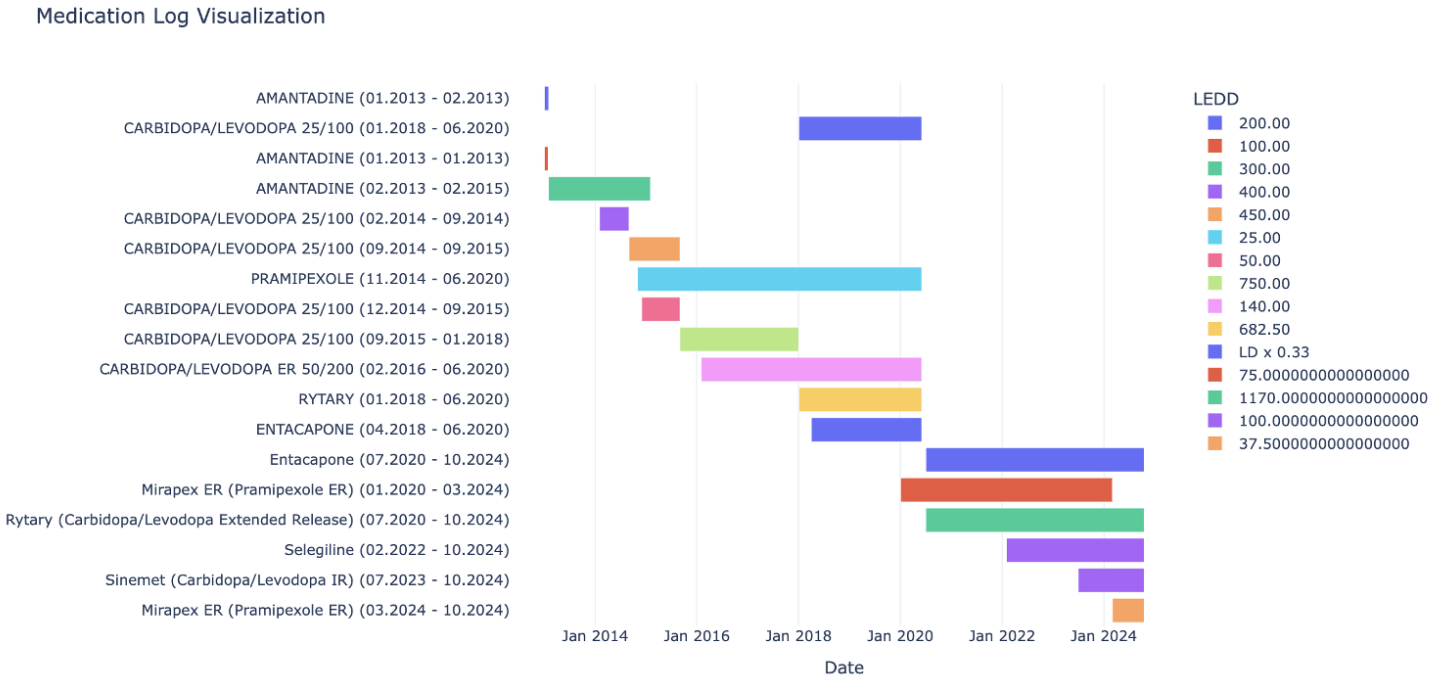
Table 1. Conversion factors used to calculate LEDDs for PD drugs

Drug Class	Drug	Conversion factor
Dopamine Replacement	Levodopa	DD x 1
	Dual-release levodopa (Madopar DR)	DD x 0.85
	Controlled-release levodopa	DD x 0.75
	Extended-release levodopa (Rytary)	DD x 0.5
	Intrajejunal levodopa/carbidopa infusion	DD x 1.11
	Intrajejunal levodopa/carbidopa/entacapone infusion	DD x 1.11 (morning) + DD x 1.46 (maintenance and extra doses)
	Subcutaneous foslevodopa/foscarbidopa	DD x 0.75
	Inhaled levodopa (Inbrija)	DD x 0.69 (capsules)
COMT Inhibitors*	Entacapone	LD x 0.33
	Tolcapone	LD x 0.5
	Opicapone	LD x 0.5
Dopamine Agonists	Pramipexole, Lisuride, Pergolide	DD x 100
	Ropinirole	DD x 20
	Rotigotine	DD x 30.3
	Piribedil	DD x 1
	Apomorphine injected	DD x 10
	Apomorphine sublingual (Kynmobi)	DD x 1.5
	Bromocriptine	DD x 10
	Cabergoline	DD x 66.7
	Dihydroergocryptine (DHEC)	DD x 5
MAOB Inhibitors	Selegiline oral	DD x 10
	Selegiline sublingual	DD x 80
	Rasagiline	DD x 100
Other	Amantadine	DD x 1
	Amantadine ER (Gocovri)	DD x 1.25
	Amantadine ER (Osmolex)	DD x 1
	Safinamide (Xadago)	LED = 150
	Zonisamide	LED = 100
	Trihexyphenidyl	LED = 100
	Istradefylline (Nourianz)**	LD x 0.2
	Mucuna pruriens	DD x 0.013

DD = daily dose of drug being converted; LD = levodopa dose; LED = levodopa equivalent dose.

Example of LEDD trajectory and medication log from PPMI





UK Dementia Research Institute

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?

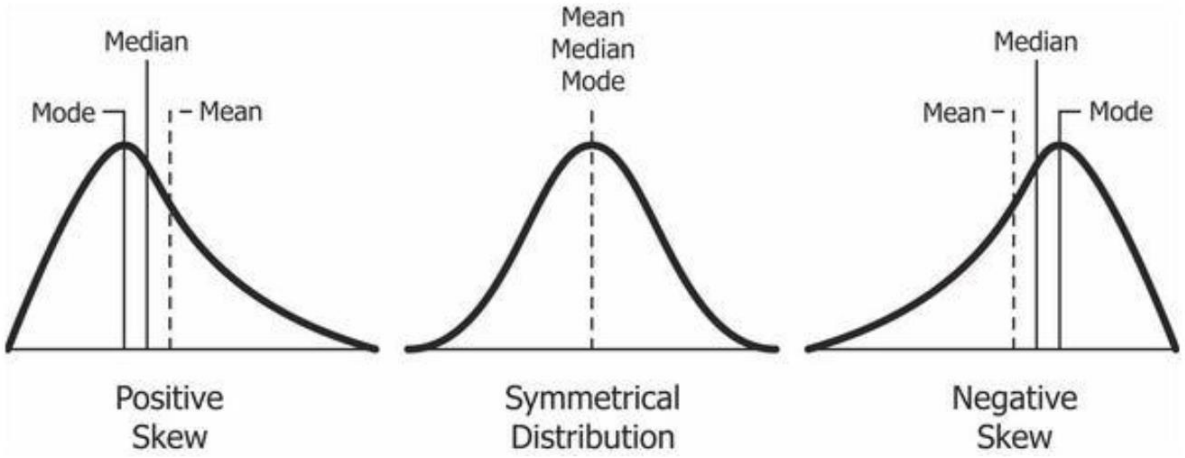
- "Is Tasty Beer cheating its customers?"
- Small deviations in 0.5L bottles → suspicion of underfilling
- How can we test this?

What are Descriptive Statistics?

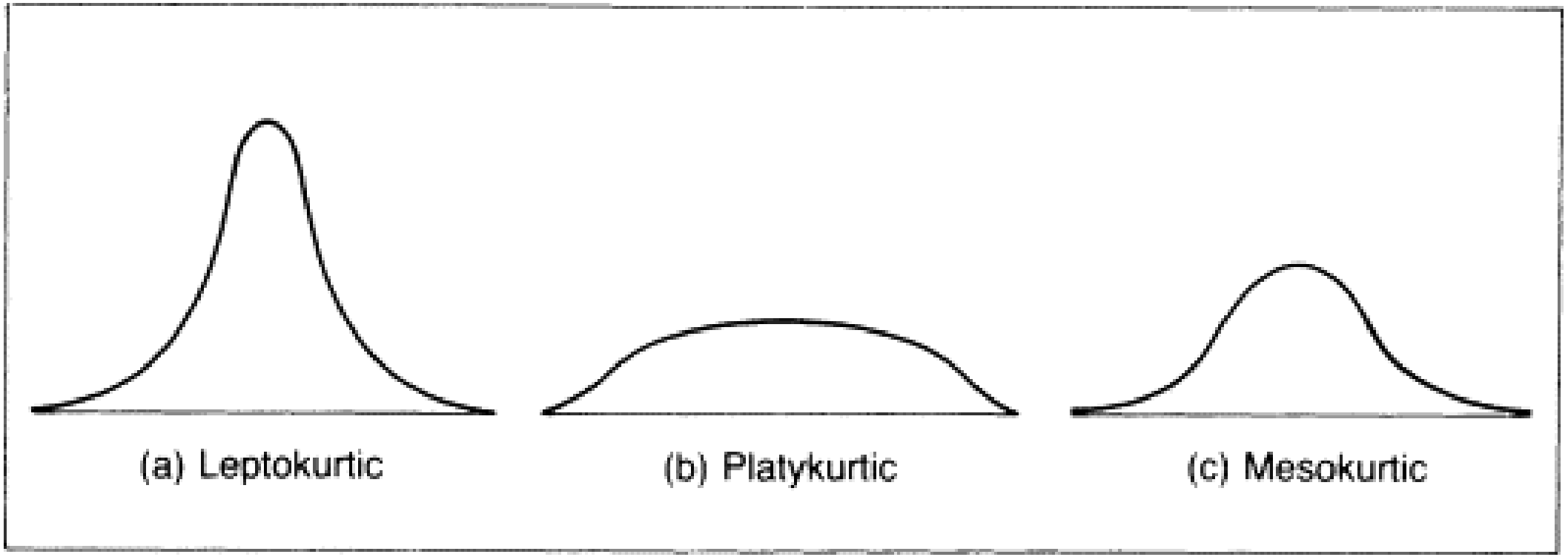
- Summarize and describe data
- First step before testing hypotheses
- Measures of central tendency & variability

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)



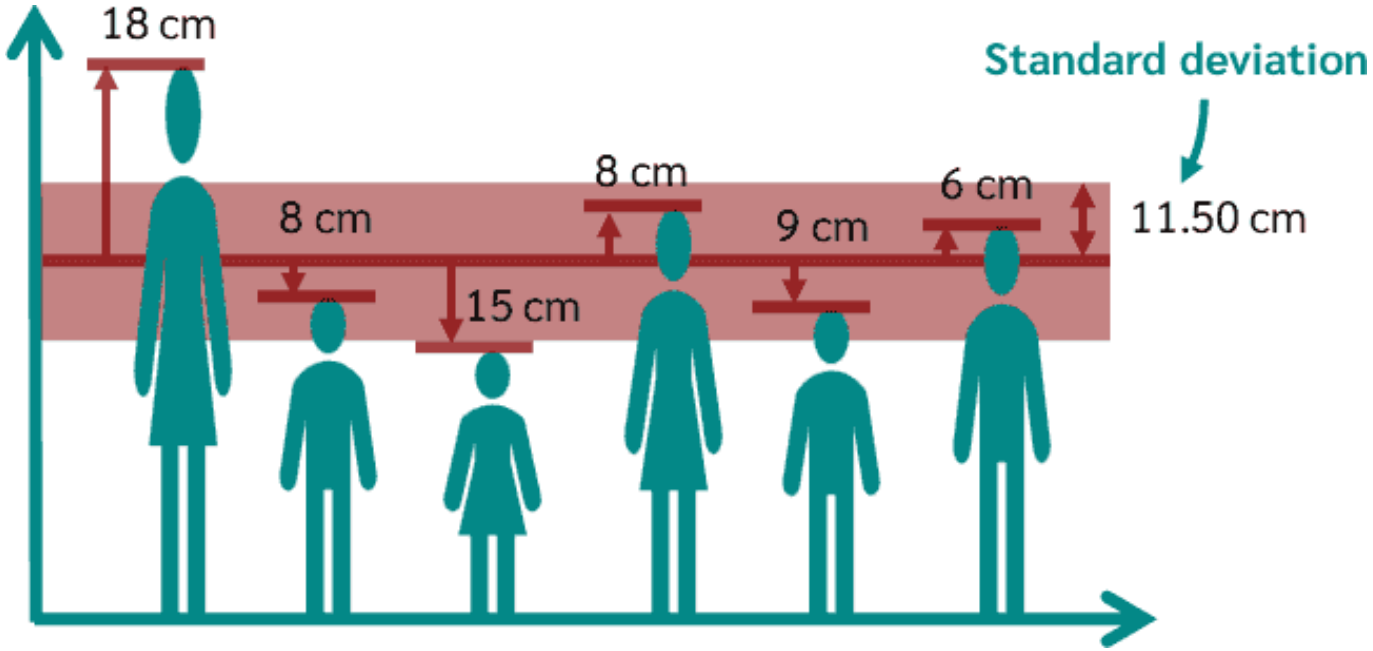
Kurtosis



Variability

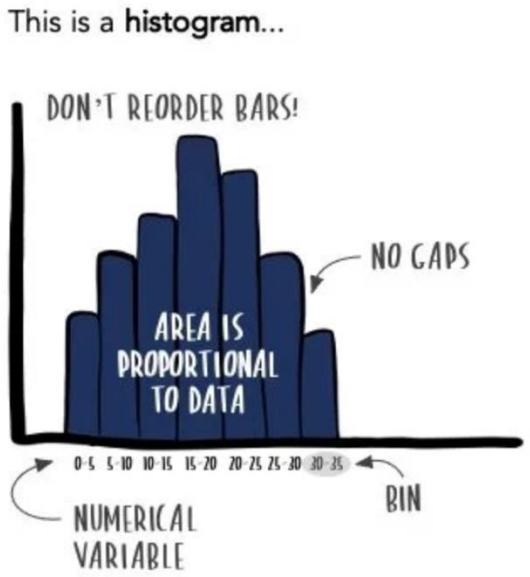
- **Range** = Highest value − Lowest value
→ Quick measure of spread
- **Variance (Var)** = Average squared distance from the mean

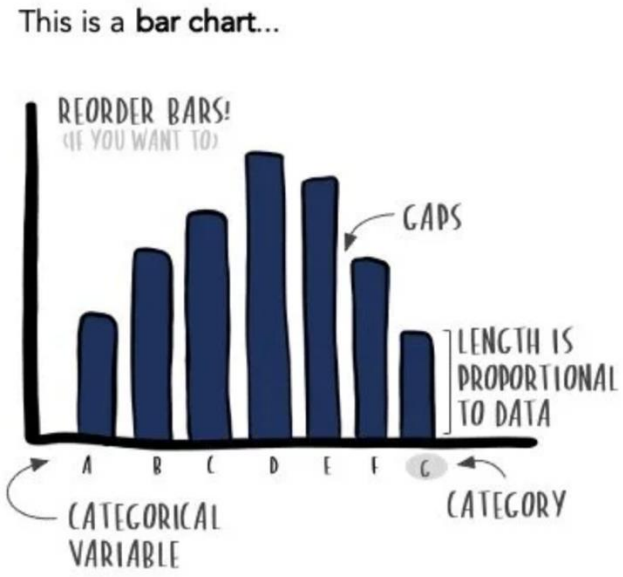
- **Standard Deviation (SD or σ)** = Square root of variance
→ "Average distance" of values from the mean

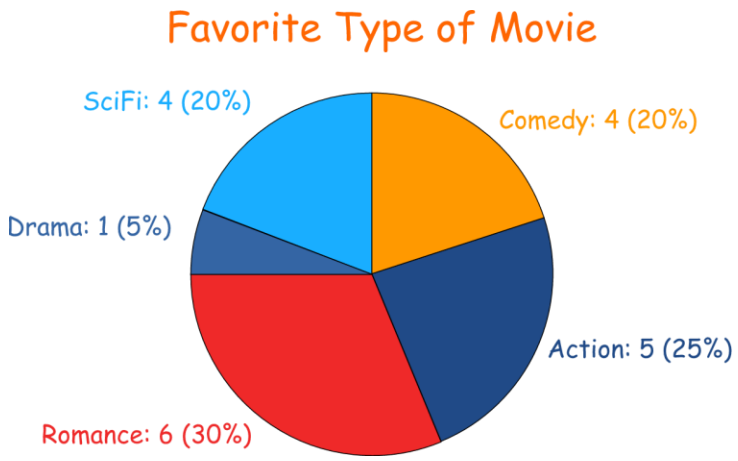


Visualisation Data

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







Visualisation Data

Boxplot (five-number summary):

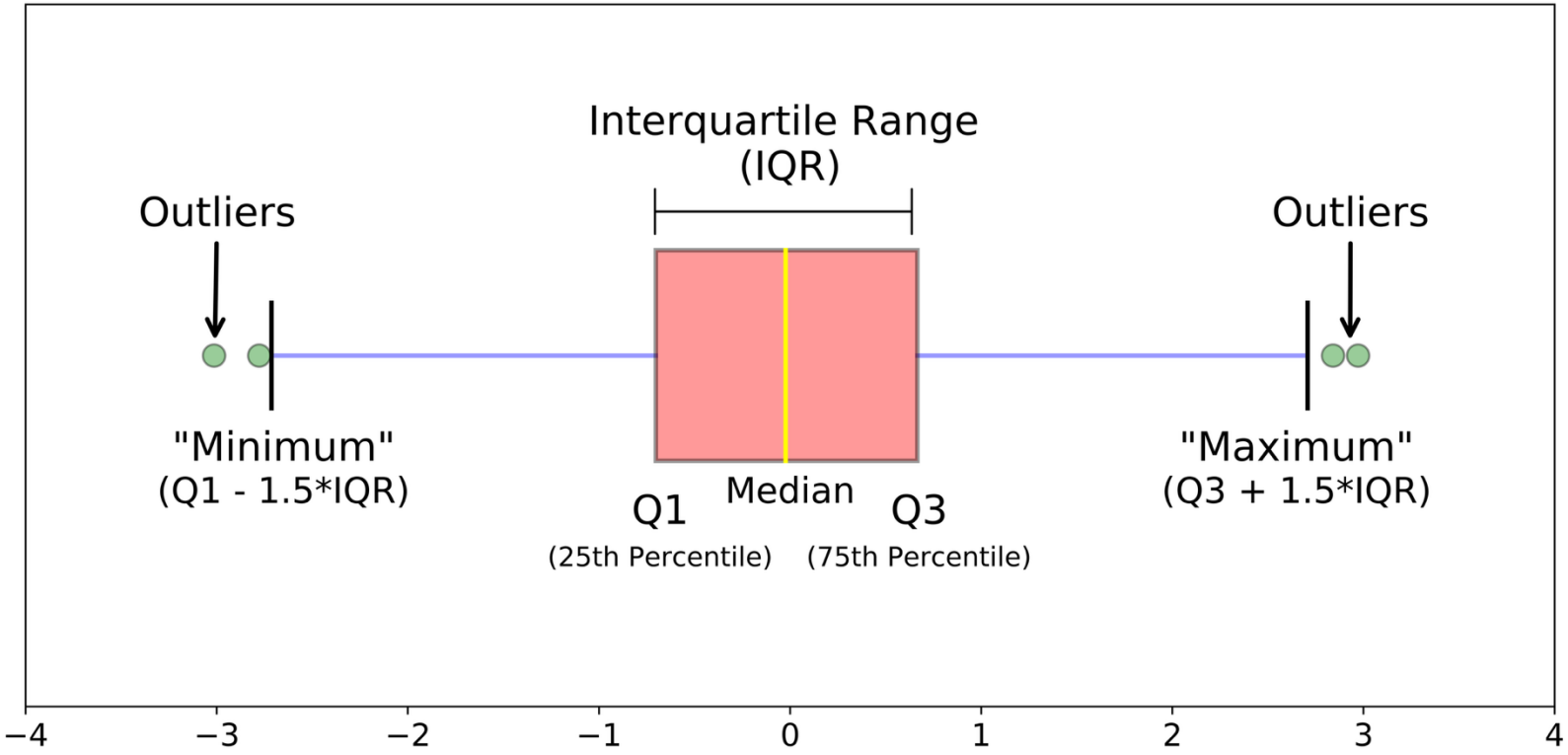


Table of summary statistics: Mean, Median, SD, etc.

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

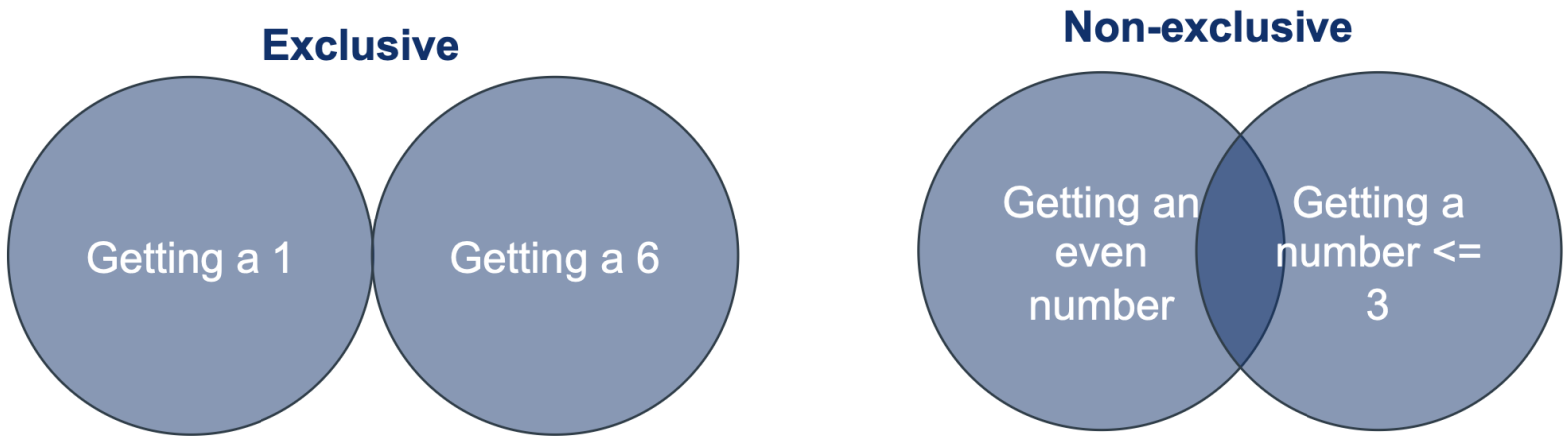
Somewhere in Between (0 < p < 1):
Probability that your favourite team will win

UK Dementia Research Institute

Probability Basics

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

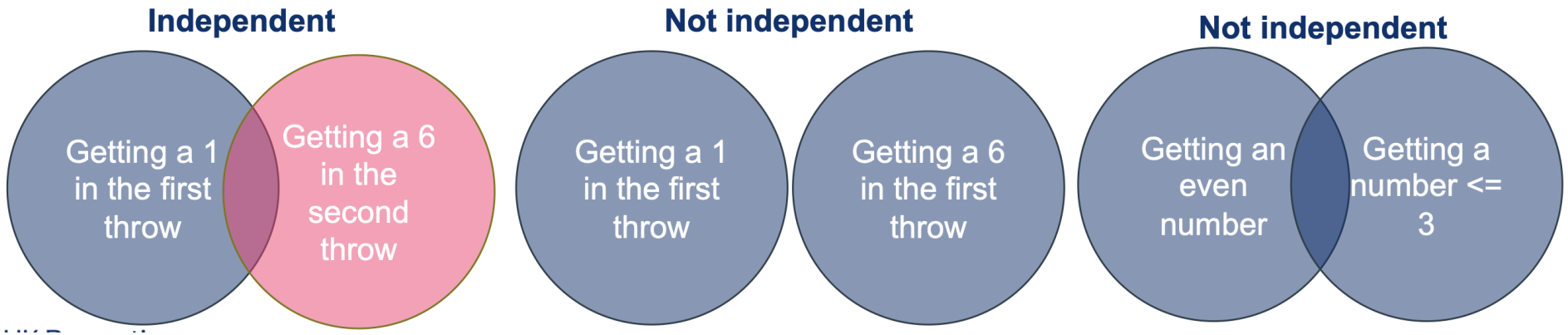
- **Mutually exclusive:** Only one of the events can occur



Probability Basics

Multiplication rule: The probability of two independent events both occurring is the product of their probabilities: P(A 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



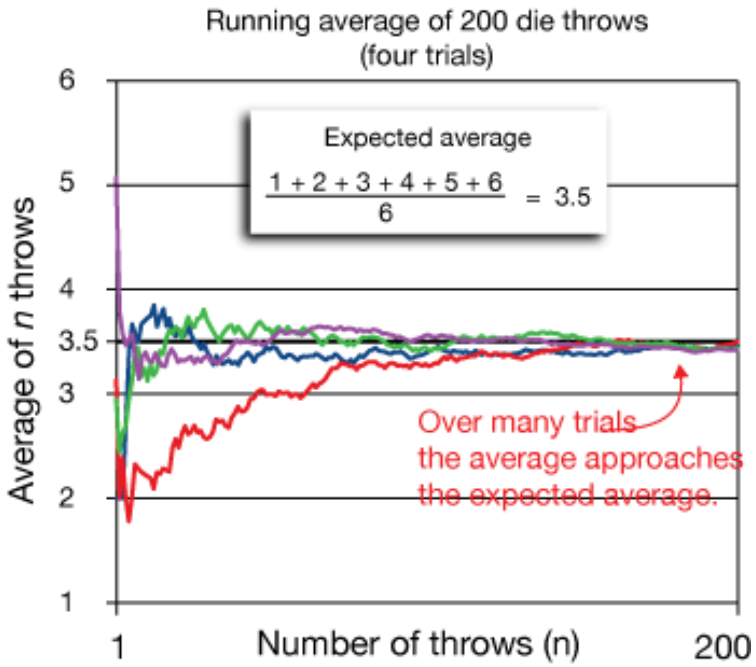
Expected Value & Law of Large Numbers

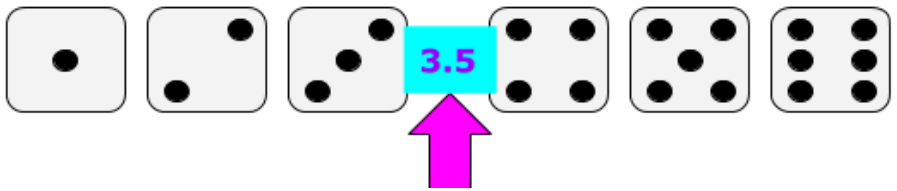
Expected Value (EV): long-run average outcome

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

Law of Large Numbers:

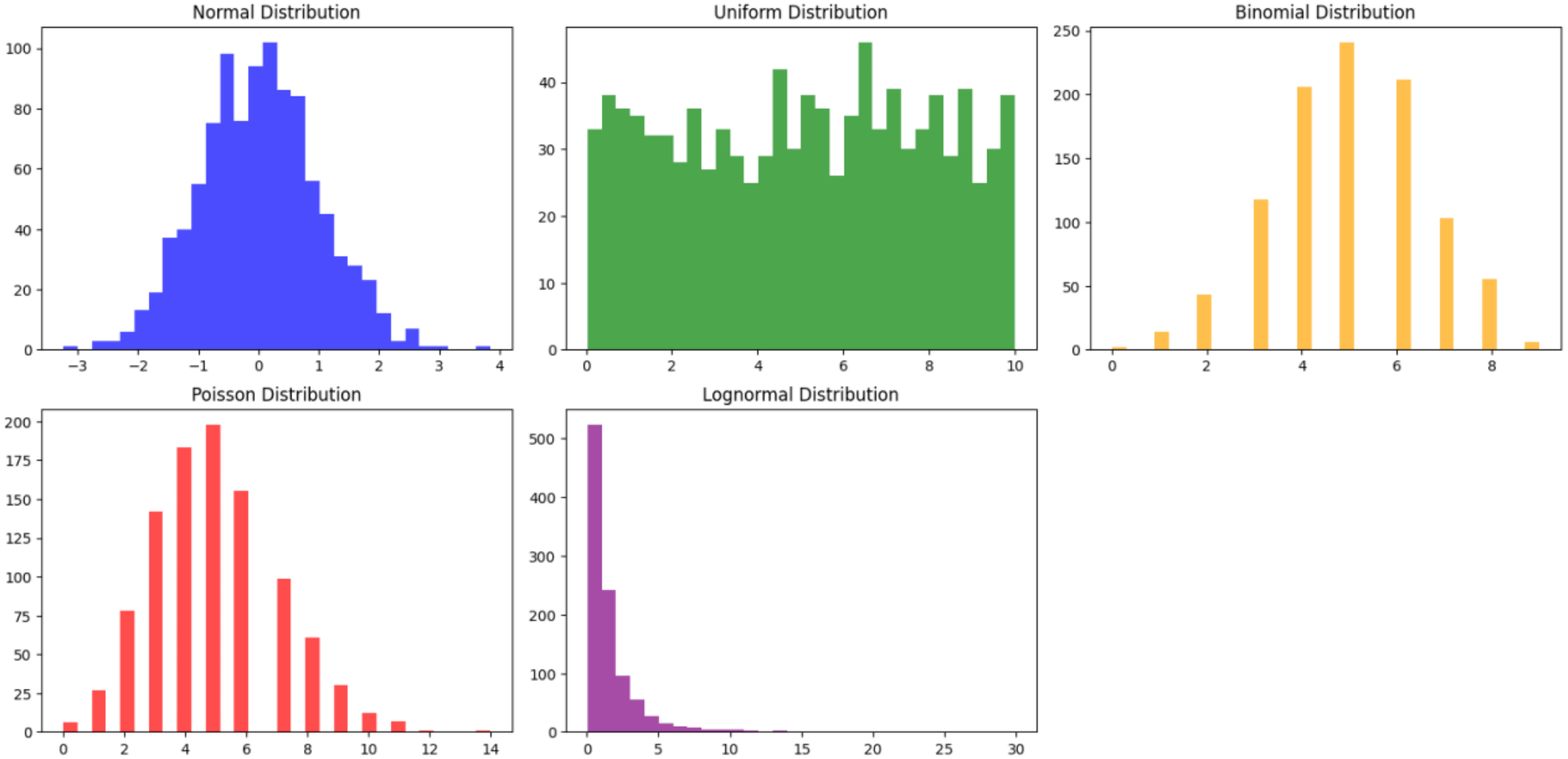
As n → ∞, the sample average → expected value





Probability Distributions

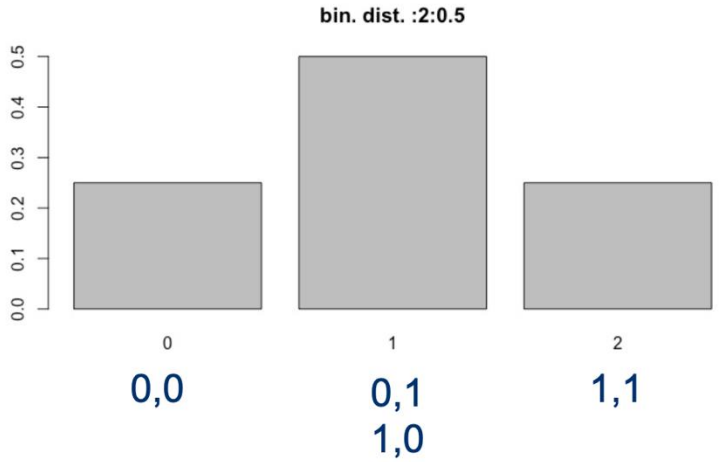
- A **probability distribution** tells us how likely different outcomes are

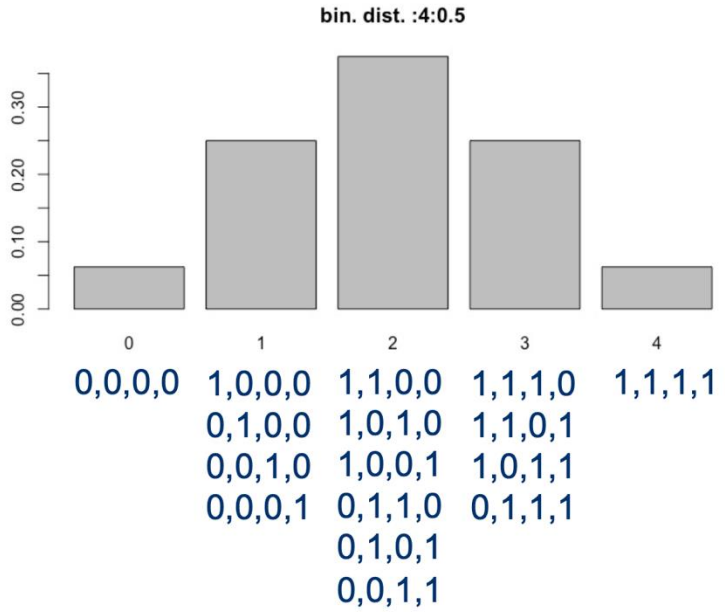


Probability Distributions: common

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

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





Probability Distributions: common

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

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

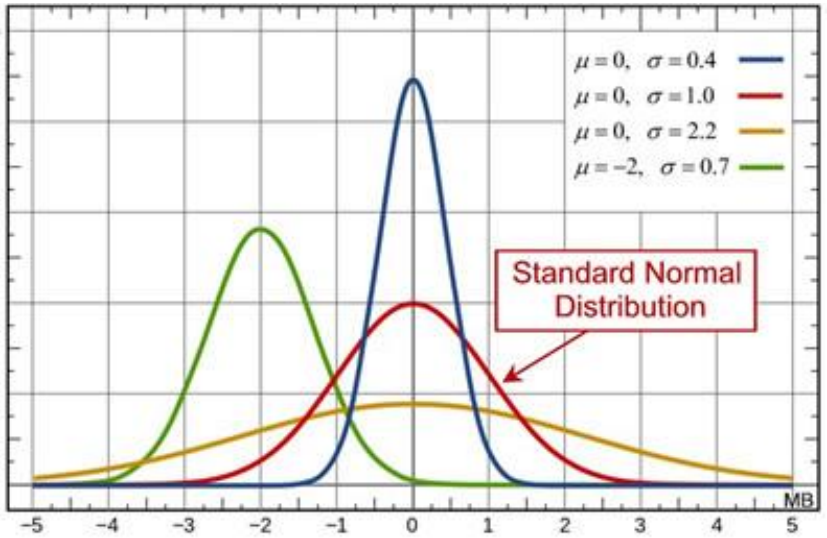
On average, 3.6 people arrive at a booking counter every 10 minutes on weekends. What is the probability that exactly 7 people arrive in 10 minutes?

Probability = 0.0424 (≈ 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. eight, reaction time, blood pressure

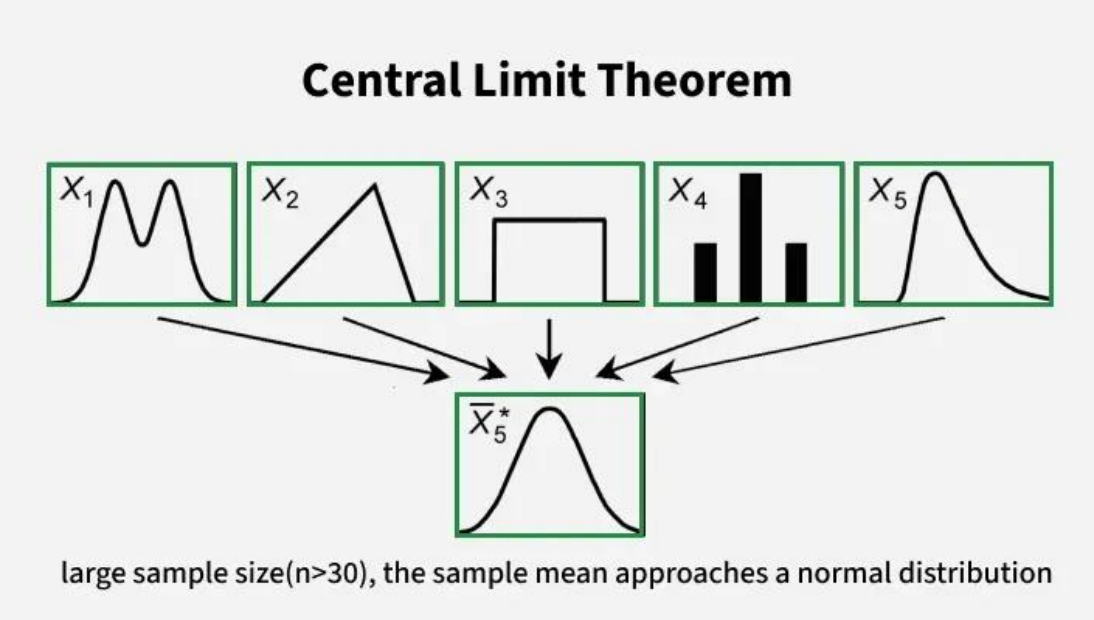


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 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 ↑**
- **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



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

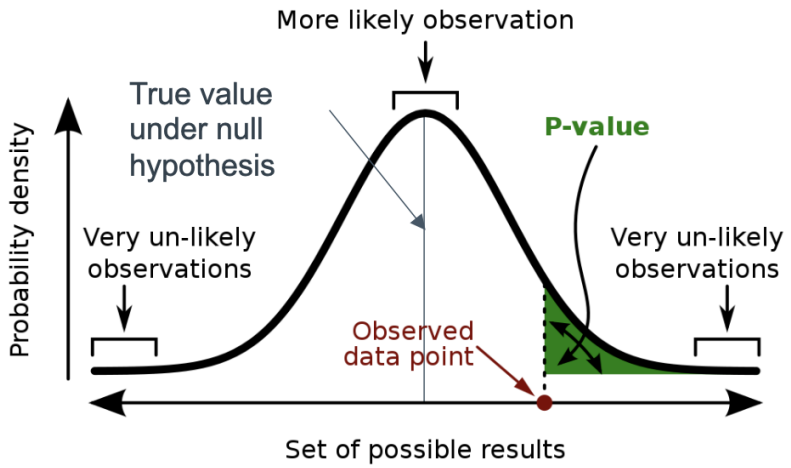
- H0 = *Tasty Beer* fills at least 0.5l into each bottle on average
- H1 = *Tasty Beer* fills less than 0.5l into each bottle on average

How likely are we to observe the sampled values if we assume that H0 is rue (*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 H0, of obtaining a result equal to or more extreme than what was actually observed – P(result | H0).



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

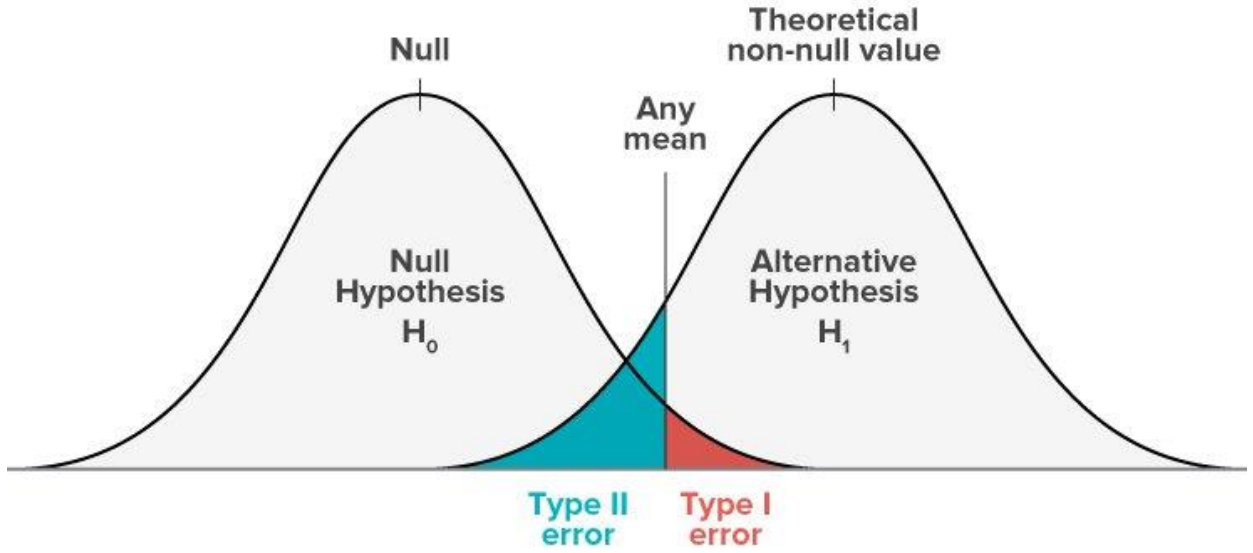
Probability = α (significance level, usually 0.05)

- **Type II error (false negative):** fail to reject H_0 when H_1 is true

Probability = β

- **Power = 1 − β** = 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:

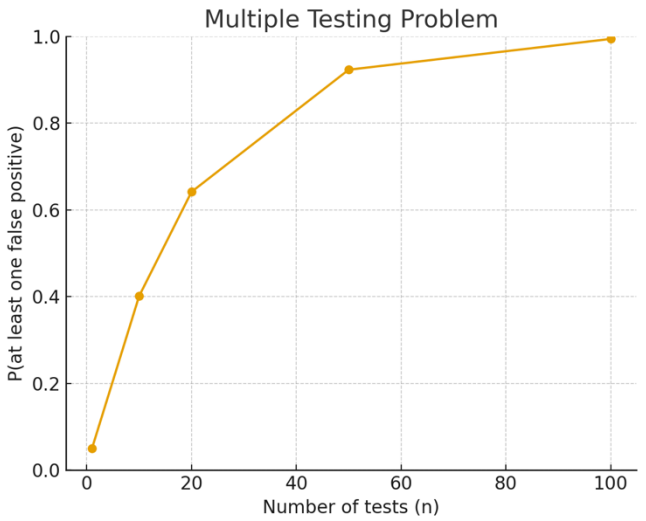
10 tests → 40% chance of ≥1 false positive

20 tests → 64%

50 tests → 92%

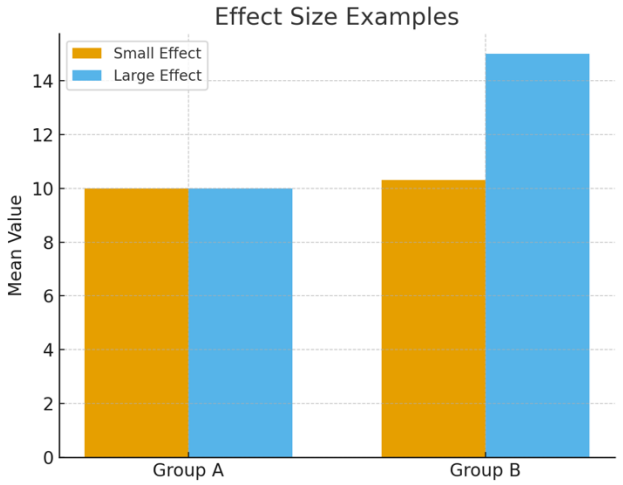
- **Solutions: Bonferroni correction**
- Other methods: FDR, Holm–Bonferroni

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



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)

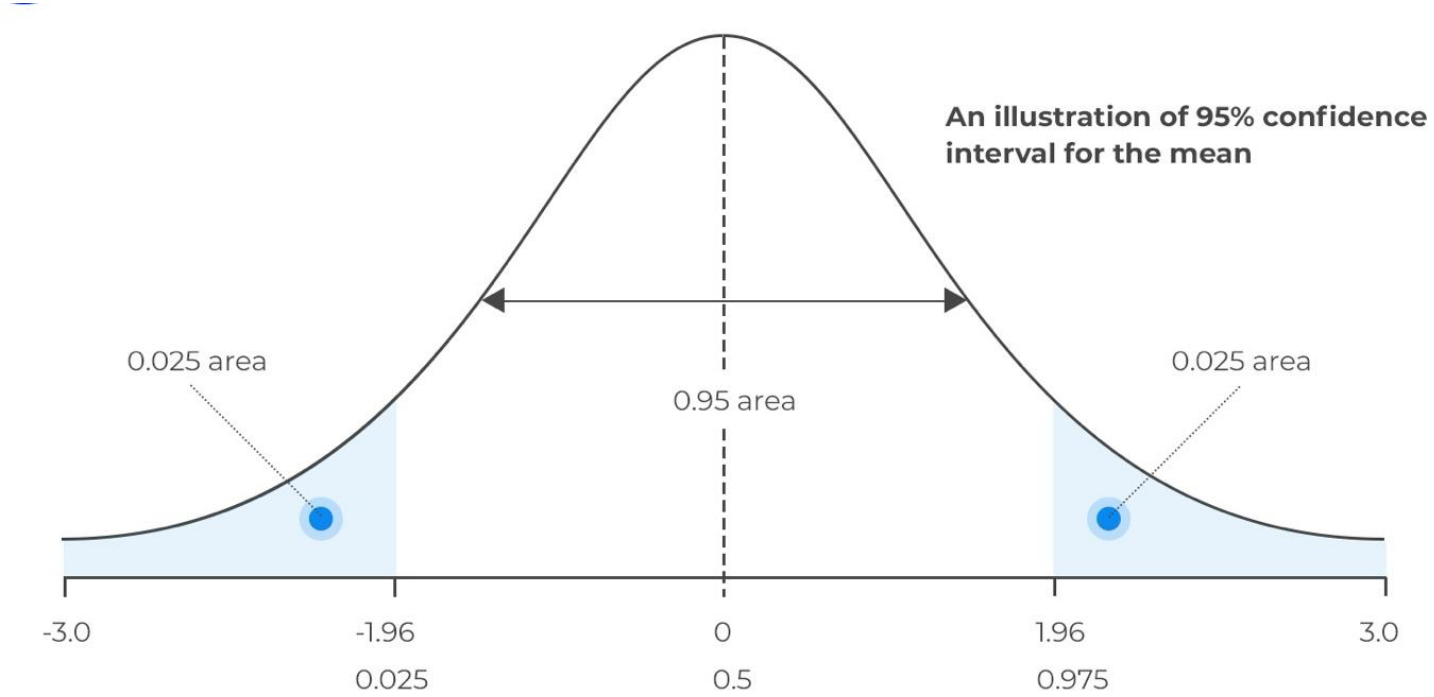


Confidence Interval

A 95% confidence interval means we are using a method that captures the true value 95% of the time.

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval
$\bar{x}$ = sample mean
z = confidence level value
s = sample standard deviation
n = sample size



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)

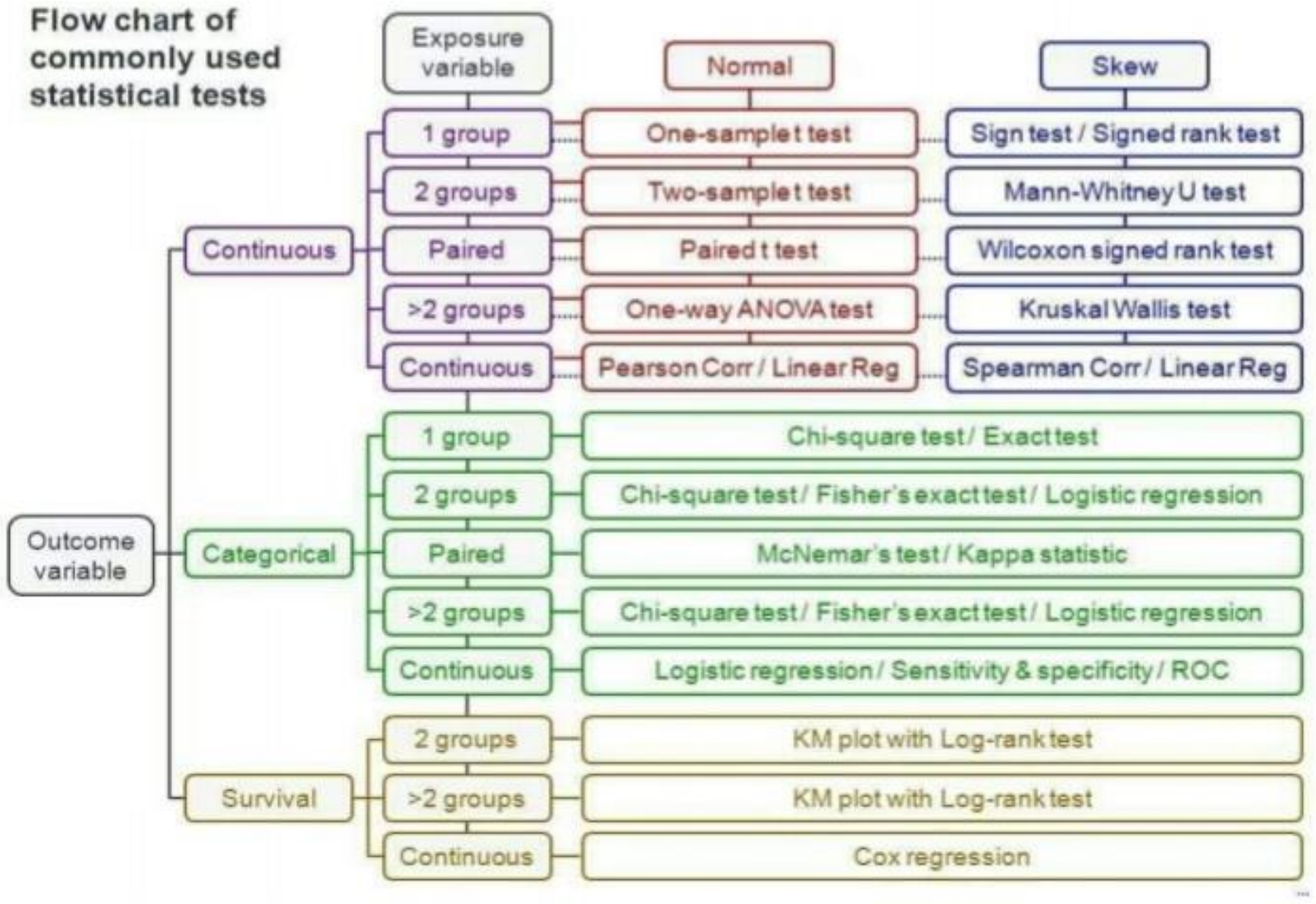
The T-test

- **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 / √n)

How to choose your statistical test



Power & Sample Size

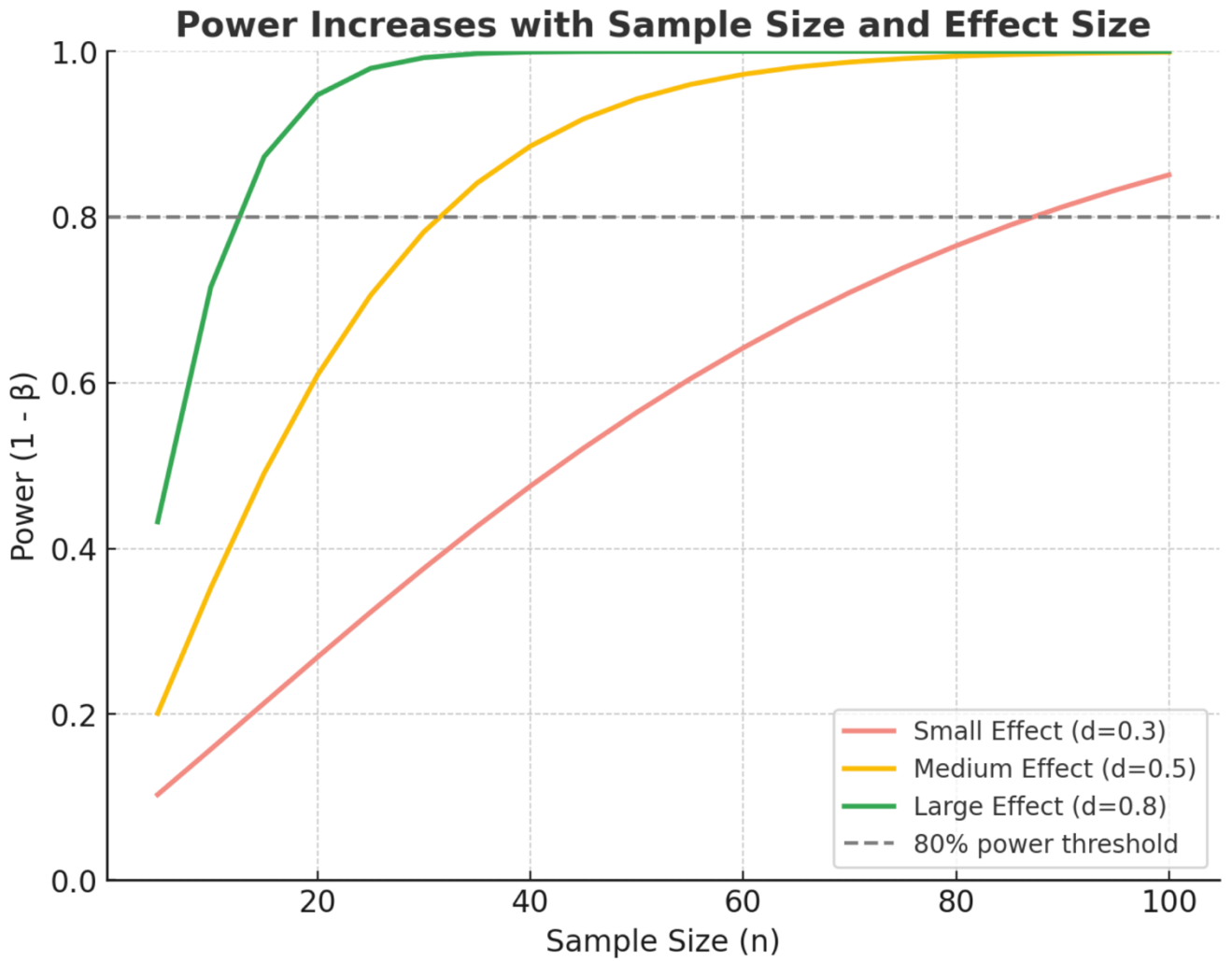
- **Power = 1 − β**
  - Probability of detecting a true effect
- Depends on:
  - **Effect size** (bigger effects easier to detect)
  - **Sample size (n)** (more data → less noise)
  - **Significance level (α)**

- **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



Take-home Messages

- **Visualize your data first** – plots tell you more than tables
- **P-values ≠ 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

Let's Code

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Phillip Press	Viola Ceriani	Duru Okay	Yossra Serroukh	Harsimran Kaur	Jenna Lin	Muireann Hogan	Lucia Jing
Margarida Neves	Anya Kaur Haddon	Chloe Kam	Rebekah Bourne	Luca Pastorello	Mustafa Gunaydin	Edona Bajrami	Zishan Lin
Olivia Pownall	Anais Chia	Wenbo Liao	Hattie Oliver	Luoyuan Zhang	Laura Raklec	Francesca Murley-Holme	Stephanie Sun
Charlotte Yu	Shobana Chandrashekar	Yunyi Gao	Thishany Kuganeswaran	Ellie Carre	Hanna Altessaid	Ruben Thiagarajah-Fernar	Neera Gahir
Chun Hei Leung	Katie Hay	Anas Saleem	Hanyue Pang	Veronika Shevchenko	Aitsilja Dabasinskaite	Felix Varenne	Chuyi Zhang
Andrea Fan	Nina Jeffrey	Chi U Chau	Zeinab Ben Halim	Amina Bououdine	Ema Ferra	Ruofan CAO	Lili Yassin
Adelina Shahata	Asma Abdullahi	YINUO Wang	Isabella Coloru	Krystal Tan	Temi Lana	Xinrui Fan	Keya Tanwani
Tanaka Udugama Jal	Vea Bley	Mari Hronska	Lucas Yebra Garcia	Sarah Kurbanov			

https://github.com/Sandoretal/Module_3/tree/main/tutorial_2

UK Dementia Research Institute