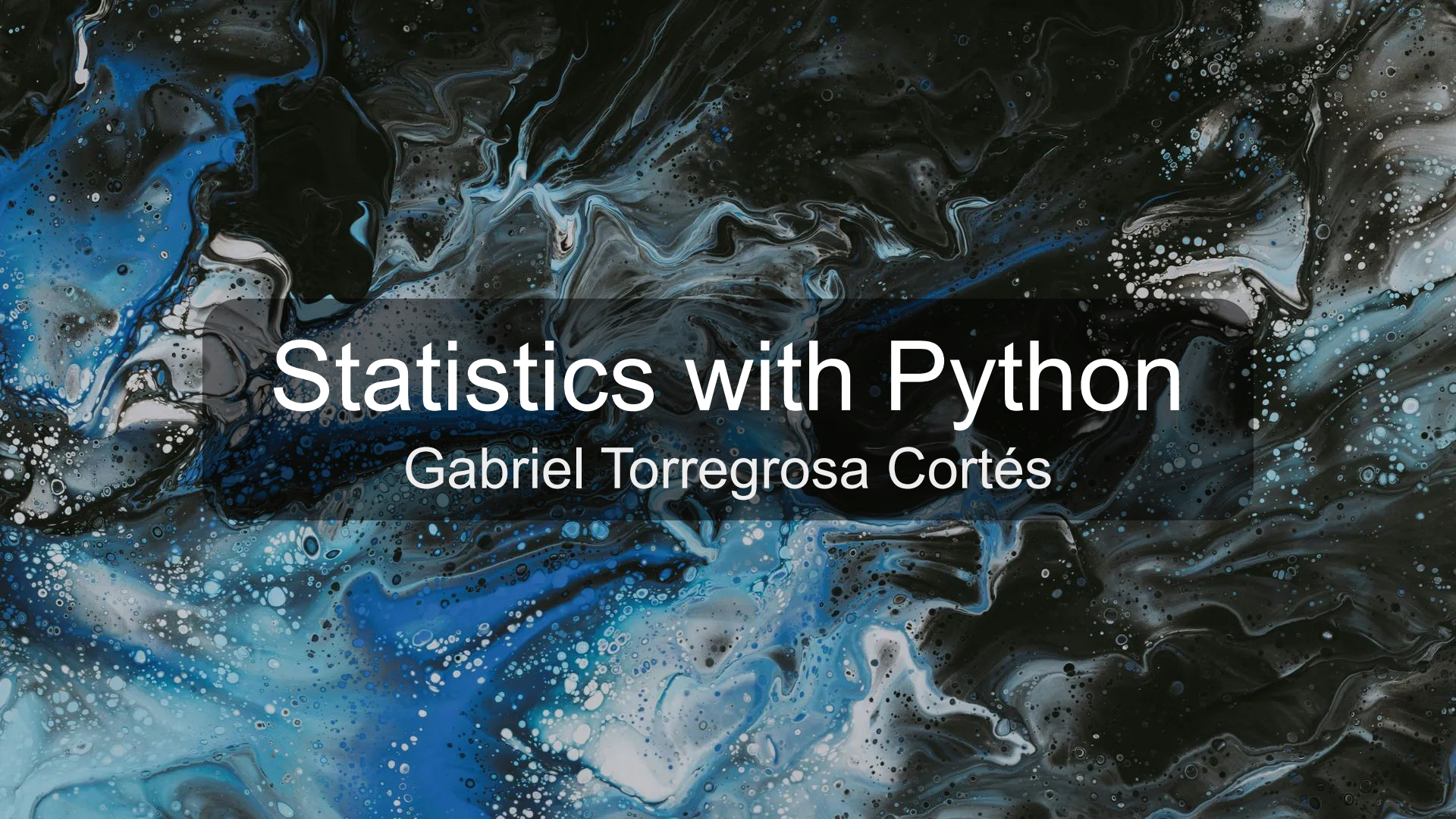

Statistics with Python
Gabriel Torregrosa Cortés

Index

Data Exploration and Description

Index

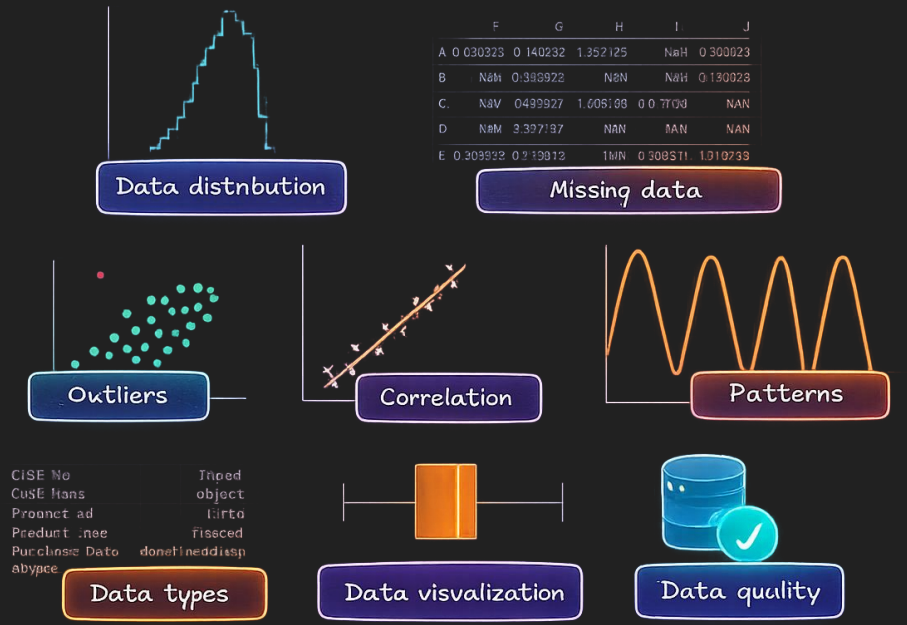
- Exploratory Data Analysis
- Data types
- Plots
- Summary statistics
- Outliers I
- Missing Data

Learning Outcomes:

- How should I start exploring my data?
- How I sumarize my findings?

Exploratory Data Analysis (EDA)

- Statistics Start Before Modelling
- Visualization Reveals Patterns EDA is a process, not a checklist
- Goal: Understand data before formal analysis
- Combine:
  - Visualization
  - Summary Statistics
  - Domain knowledge



From https://medium.com/@WenxinZhang98/exploratory-data-analysis-83bfb1f17dc5

https://medium.com/@WenxinZhang98/exploratory-data-analysis-83bfb1f17dc5

https://dev.to/yankho817/exploratory-data-analysis-edaultimate-guide-174d

EDA Minsdet

- Ask questions before hypothesis
- Expect surprises
- Iterate continuously
- Let the data guide you

What does a typical observation looks like?

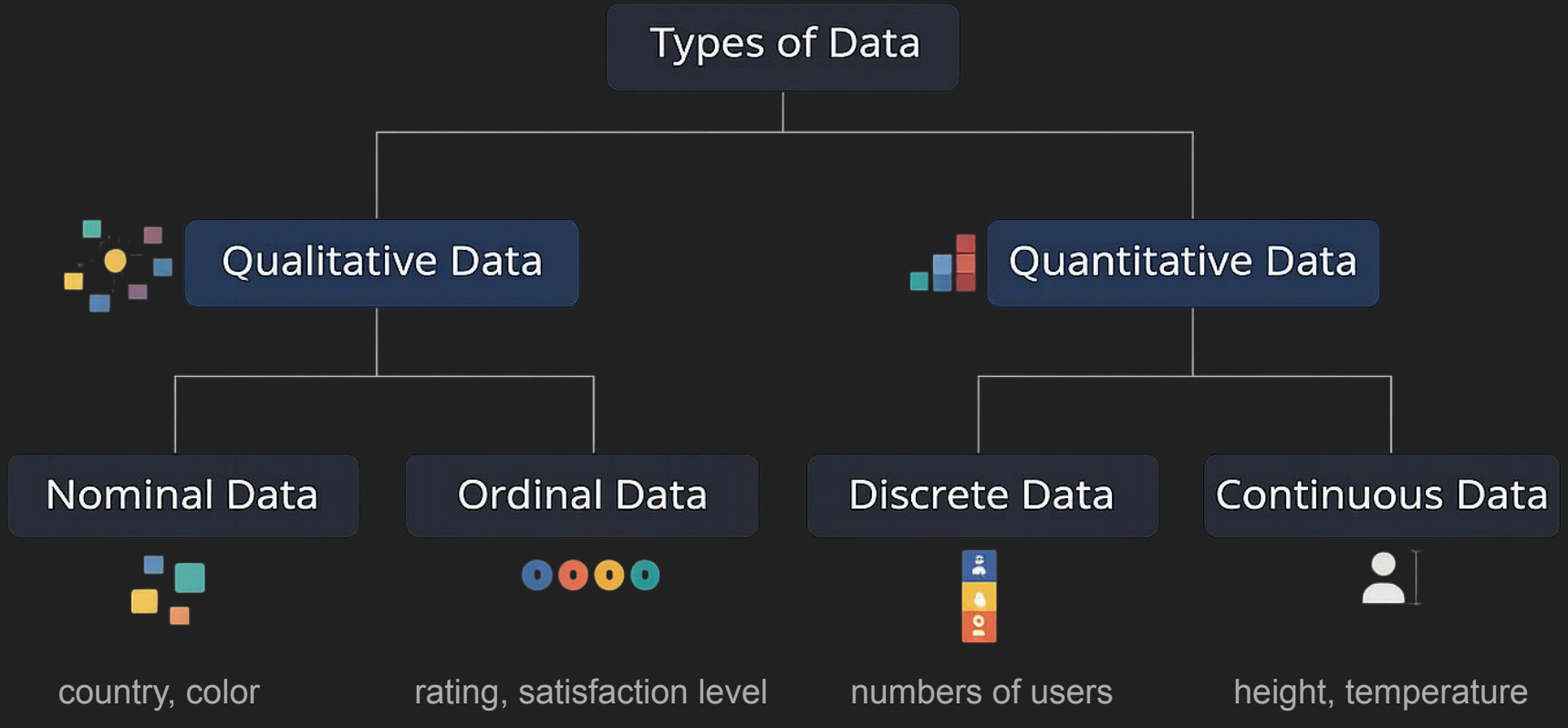
How variable is the data?

Are there unusual values?

Are relationships present?

Are there data quality issues?

Variable Types



Plots

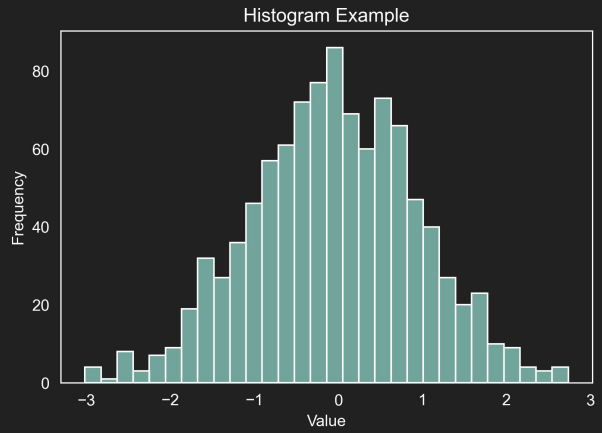
- Visuals reveal patterns numbers can hide

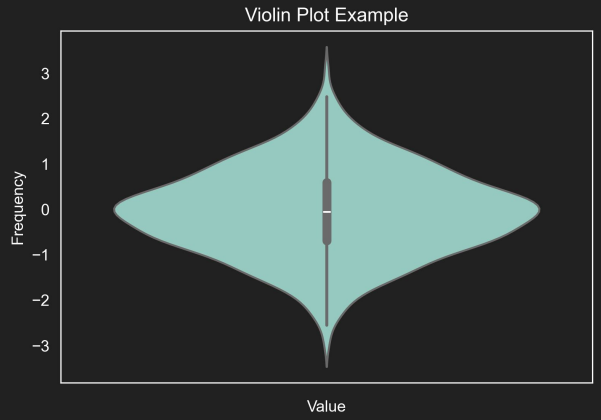
- Help detect:
  - Weird measures
  - Data distributions
  - Relationships between variables

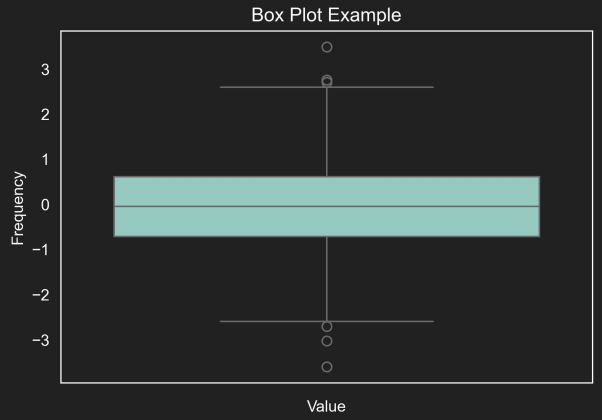
- Essential for communicating the results

Plot: single quantitative variables

We can use **histograms** to visualize single **quantitative variables**. **Violin plots** or **boxplots** can also be used to visualize **continuous variables**.
Careful note: histogram widths and violin kernel widths can hide fine details in the data





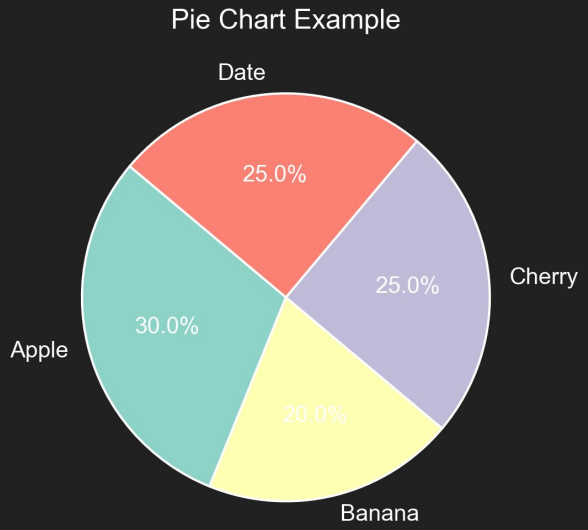


+Raw

+Summarized

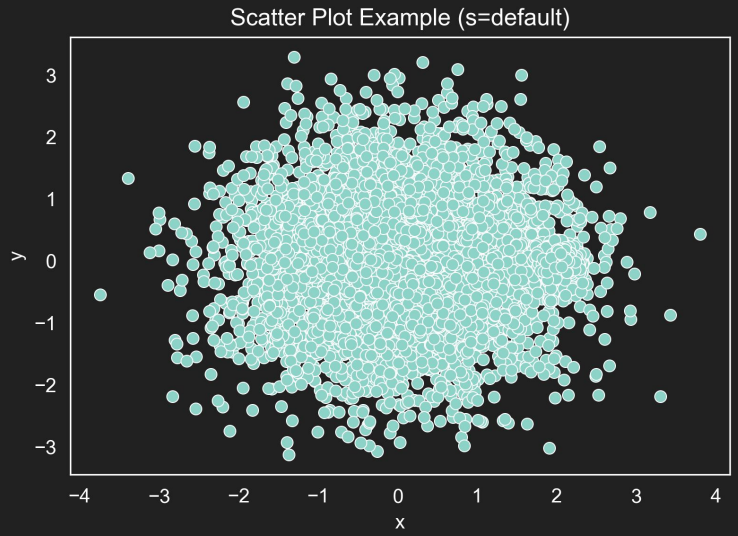
Plot: single qualitative variables

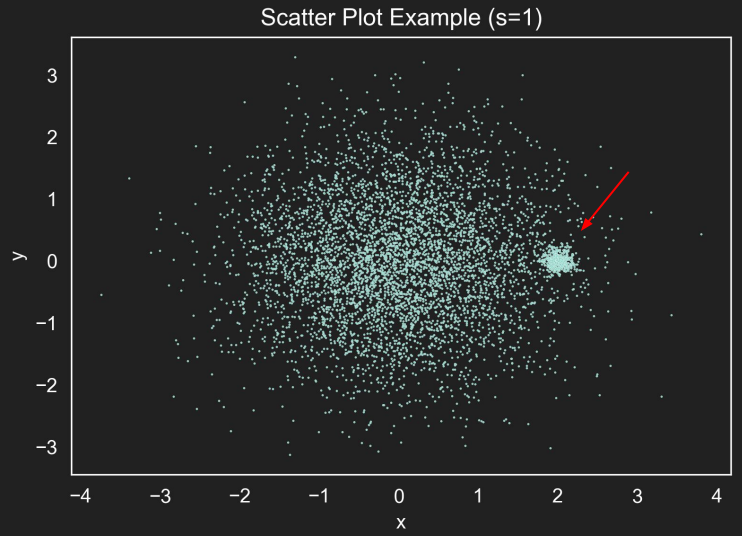
We can use **barplots** or **pie plots** to visualize single **qualitative variables**.



Plot: relations quantitative vs. quantitative

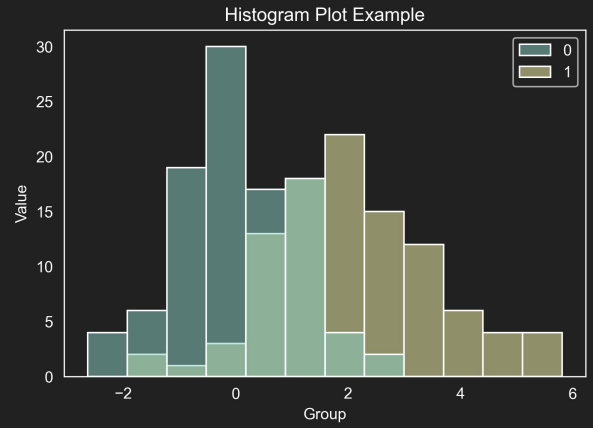
We can use **scatterplots**.
Careful note: dot sizes might hide patterns in the data is the size is not adjusted correctly.

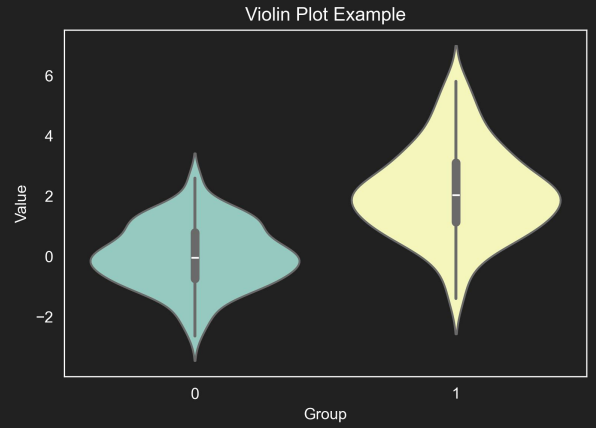


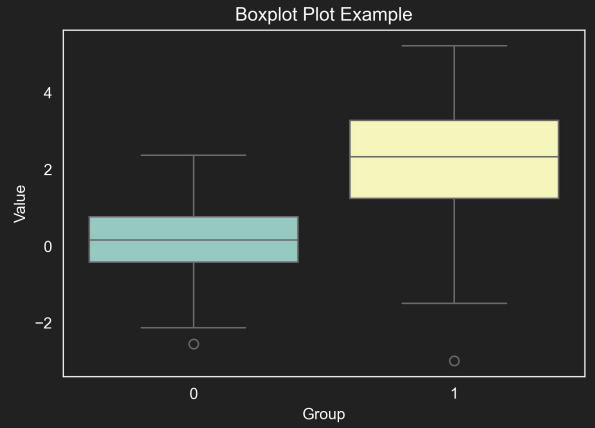


Plot: relations qualitative vs. quantitative

We can use **histograms**, v**iolin plots** or **boxplots** for the quantitative data separated by qualitative variables.





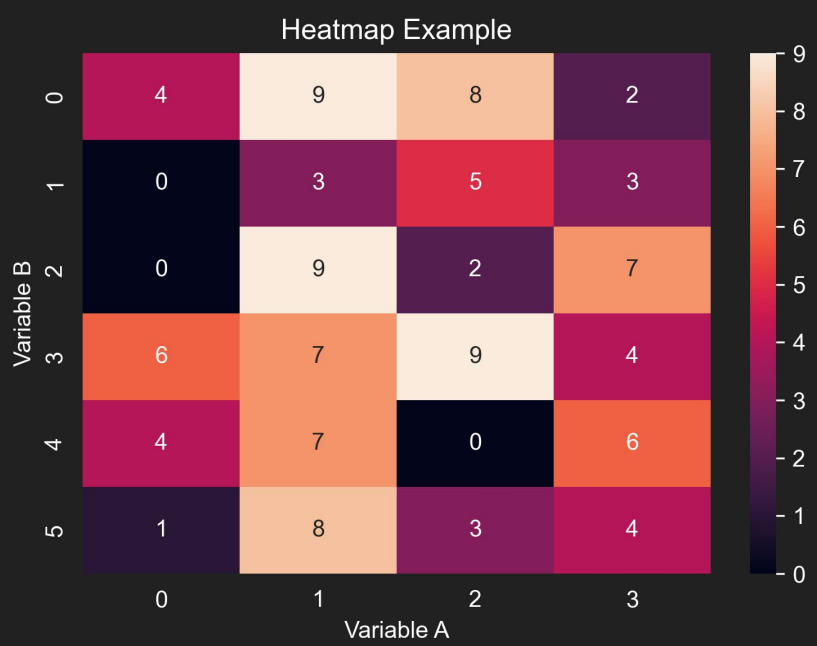


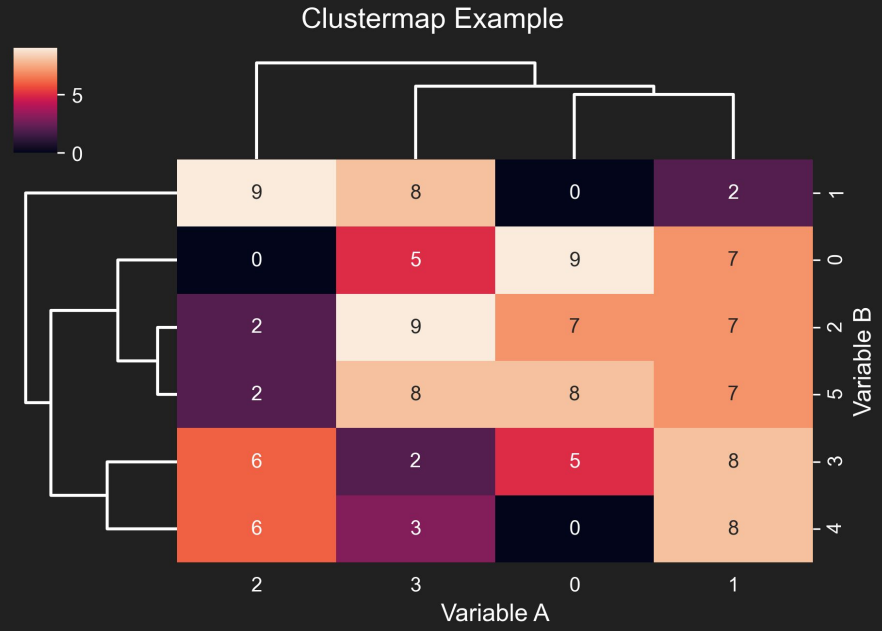
+Raw

+Summarized

Plot: relations qualitative vs. qualitative

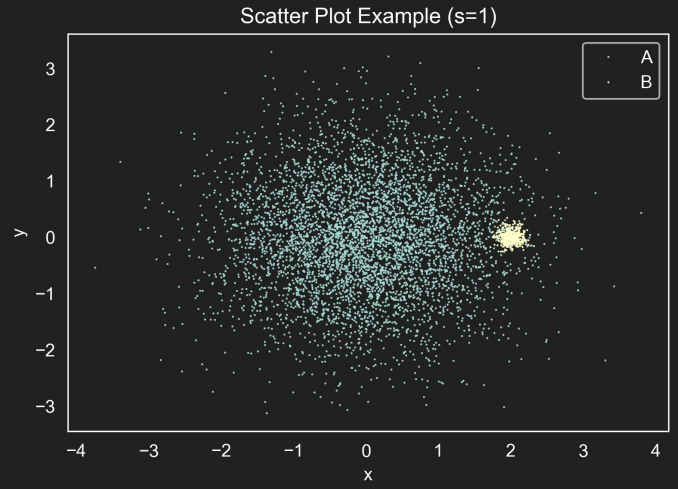
We can use **heatmaps** or **clustermaps**
Careful note: in general, only use clustermap axis with nominal data variables

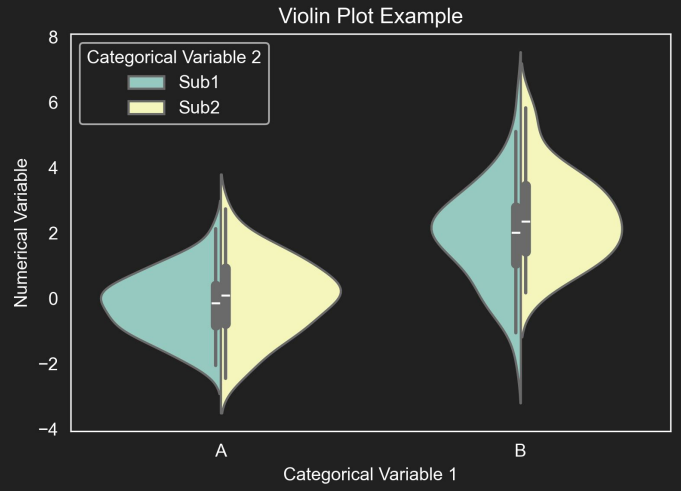




Plot: 3 variable relationships

We can use **colors** on **scatter plots** for 2 numerical and 1 categorical variable
We can use further splitting in **violin plots** or **bar plots** for 1 numerical and 2 categorical variables.
Careful note: try to not overcomplicate plots, always think about the purpose

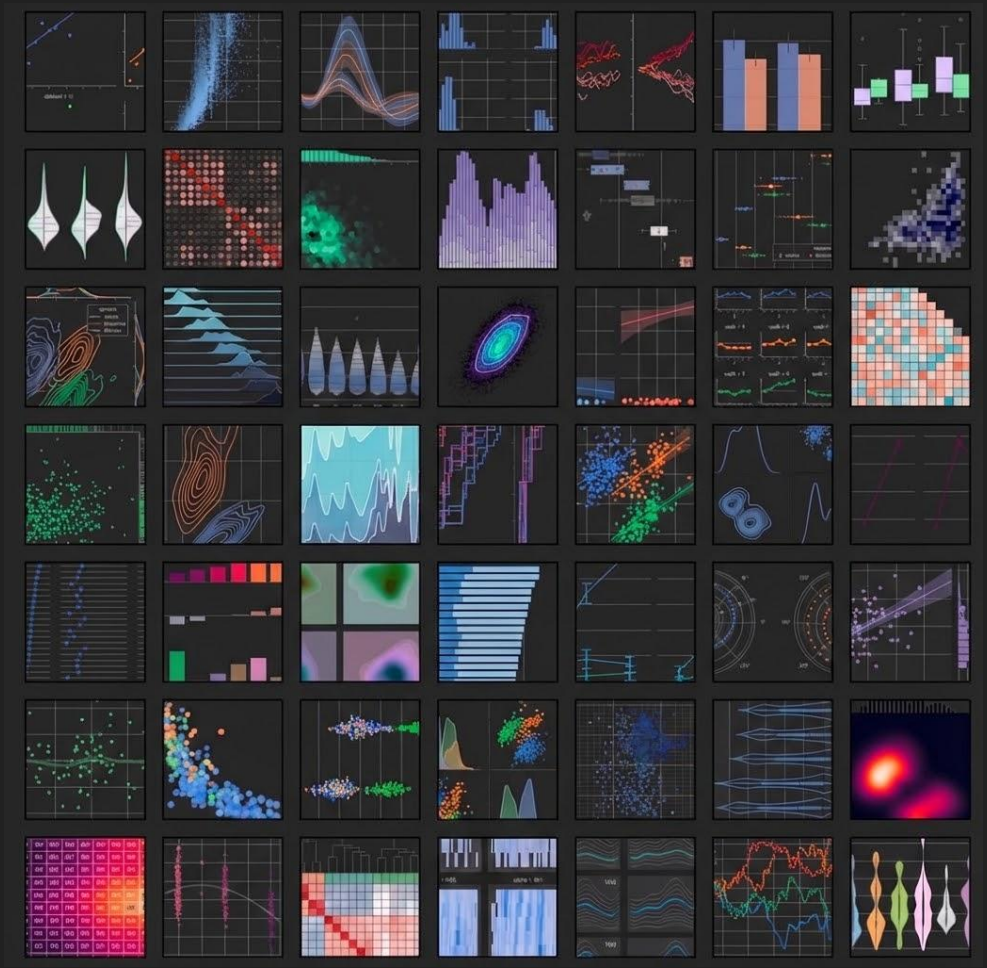




Plot: the plotting zoo

Best practices

- Label axes clearly
- Use appropriate scales
- Use color intentionally
- Always ask: *“What question does this plot answer?”*



https://seaborn.pydata.org/examples/index.html

Summary Statistics

- Connect visual intuition with quantitative interpretation

- Types of summaries:
  - Central tendency
  - Variability
  - Shape
  - Extremeness

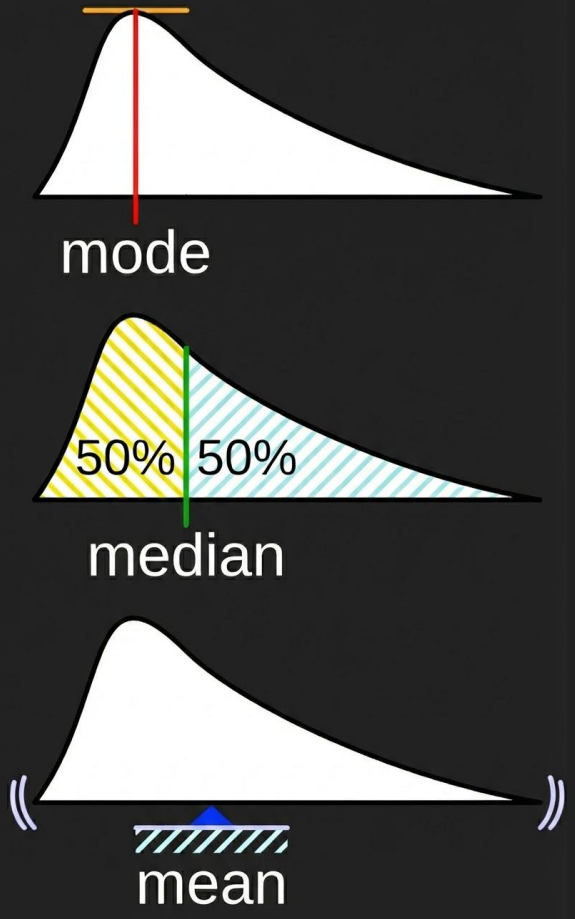
- Robustness to outliers

- If you have many variables, easier to see than individual plots

Summary Statistics: Central tendency

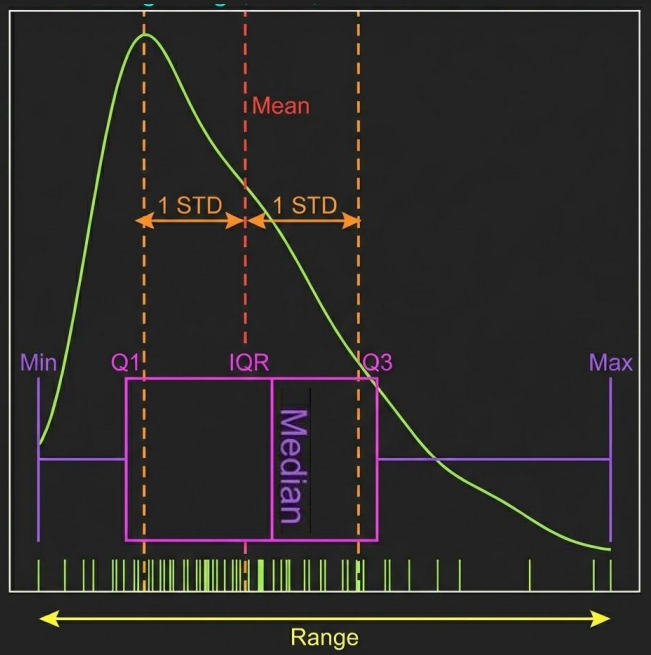
- **Mean:** The average of all observations. Sensitive to outliers.
- **Median:** The middle value when data are ordered. Robust to outliers and skewed distributions.
- **Mode:** The most frequent value in the dataset. Useful for categorical data and multimodal distributions. Not defined for continuous variables*

* You can binarize the data (e.g. histogram) and compute the mode in the discrete data



Summary Statistics: Spread

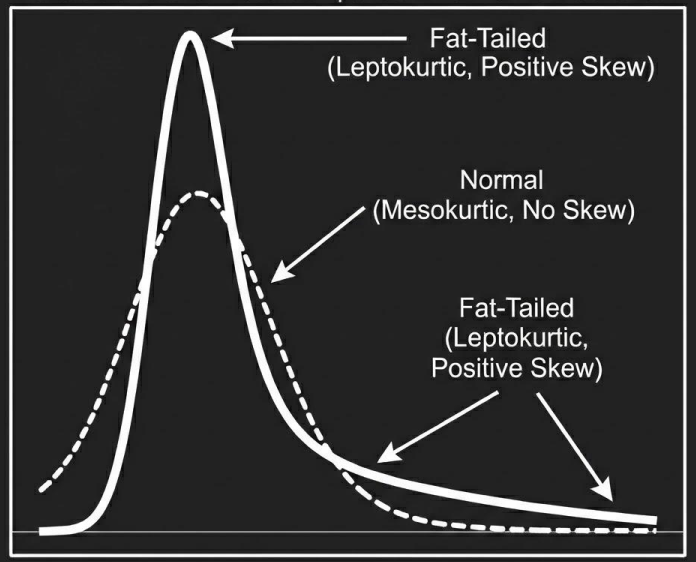
- **Range:** The difference between the maximum and minimum values. Highly sensitive to outliers.
- **Variance and Standard Deviation (STD):** Variance is the average-squared deviation from the mean. Measures overall variability, but is in squared units. STD is the square root of the variance. Represents typical distance from the mean in original units.
- **Percentile** The value below which a given percentage of observations fall.
- **Quartiles:** Values that divide the data into four equal parts. (Q1 = 25%, Q2 = median, Q3 = 75%).
- **Interquartile Range (IQR):** The difference between the 75th and 25th percentiles. Robust measure of spread.



Summary Statistics: Shape

- **Skewness:** Quantifies asymmetry of a distribution. Positive skew: long right tail. Negative skew: long left tail

- **Kurtosis:** Describes tail heaviness and peak sharpness relative to a normal distribution. High kurtosis → heavy tails and extreme values.

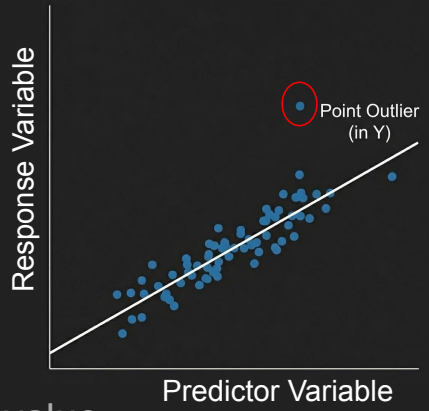


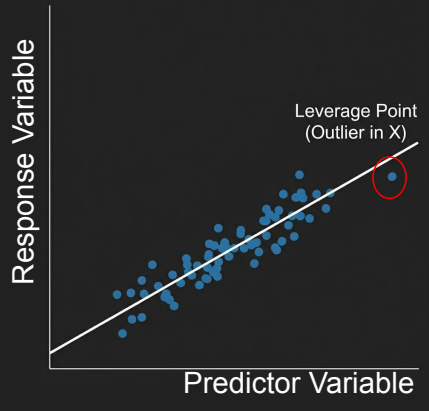
Normal:	Kurtosis=0, Skewness=0
Fat-Tailed:	Kurtosis > 0, Skewness > 0

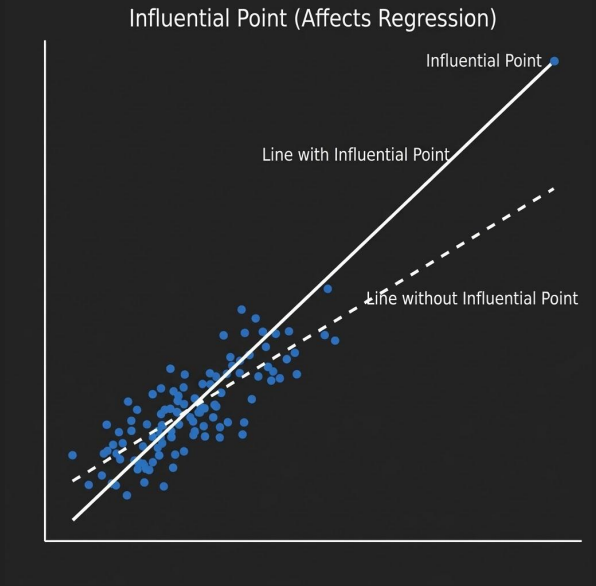
Outliers I

- **Definition:** An observation that deviates markedly from the rest of the data.
- **Types of outliers:**
  - **Point Outlier:** An unusually large or small value in the response variable.
  - **Leverage Point:** An observation with an unusual value in a predictor variable.
- **Influential Point:** An observation that substantially affects statistical results or model estimates.

- **Always make the question:** Is and error or rare-but-real?







Missing data

- **Definition:** A missing value for a measure and some variables.

- **Questions to ask:**

    - Why we have missing data?

    - Can we work with that missing information?

ID	Date	Value_A	Value_B	Category
1	2022-02-17	NaN		NaN
2	2022-02-18	1.5	0.7	Chang
3	2023-03-19	NaN	NaN	Flying
4	2023-04-20	NaN	8.5	Category
5	2023-04-21	2.0	NaN	Cheet
6	2023-05-30	3.2	NaN	Cheet

Association and Correlation

Index

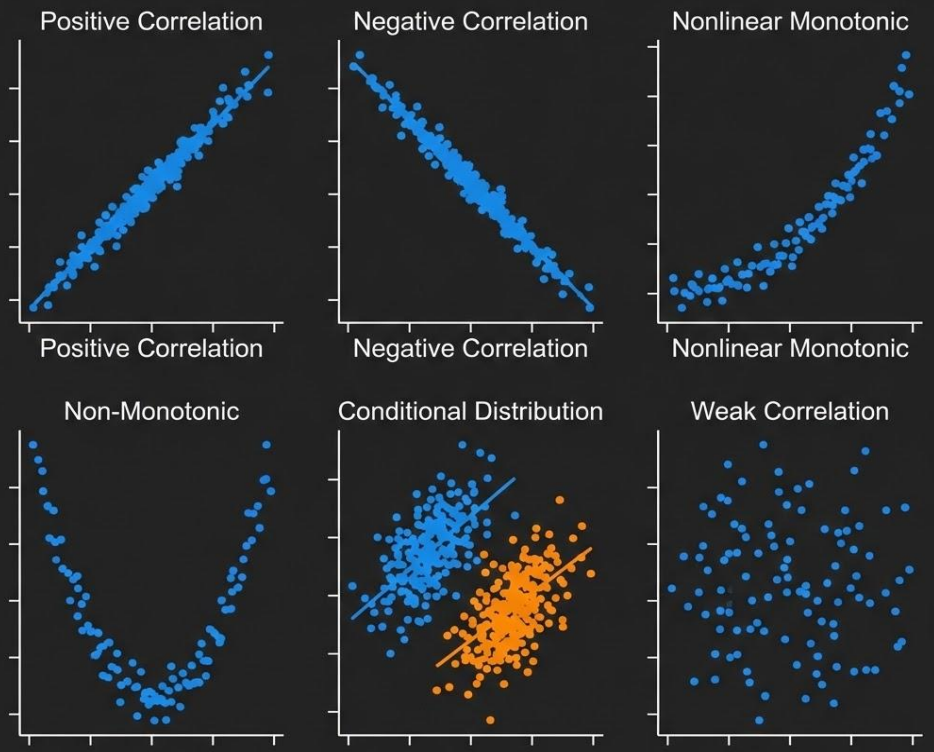
Learning Outcomes:

- Are there relations in my data?
- How I summarize them?
- What I can and I cannot conclude from my summaries?

Types of relations between variables

Two variables are related if knowing one gives information about the other.

- Linear or non-linear
- Monotonic or non-monotonic
- Positive, negative or conditional
- Strong or weak



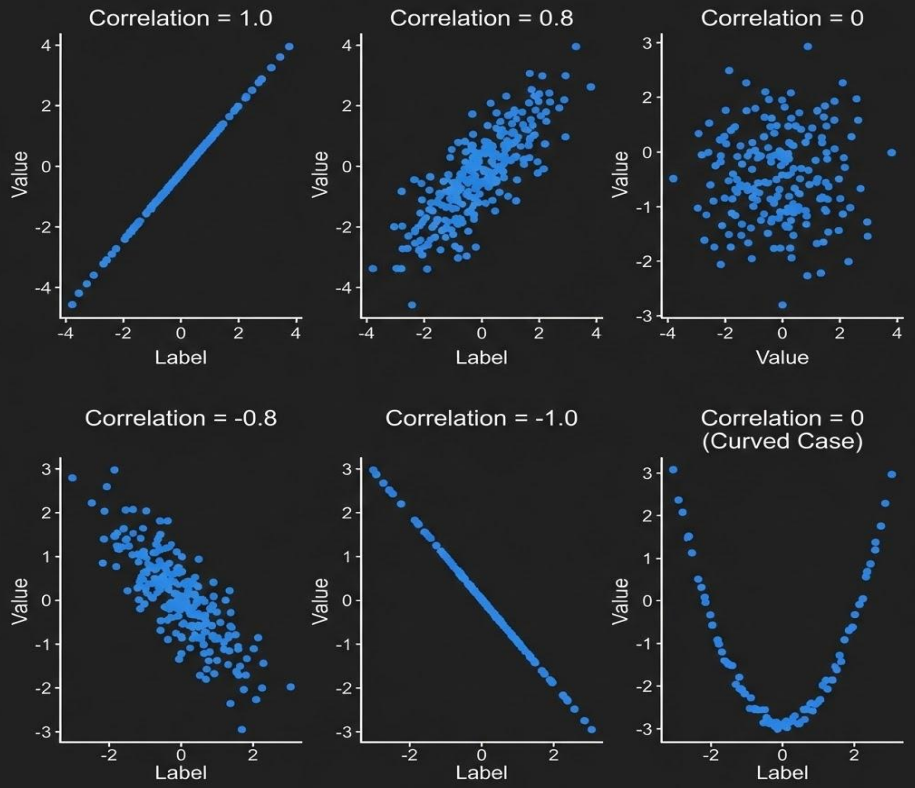
Correlation metrics: Pearson

Pearson correlation is a descriptive measure of **linear association** between **two continuous variables.** It answers the question:

> **Do these two variables move together linearly?**

Properties:

- Measures **linear relationships**
- Ranges from
  - 1: Positive linear relationship
  - 0: No linear relationship
  - -1: Negative linear relationship
- Sensitive to outliers



Correlation metrics: Spearman

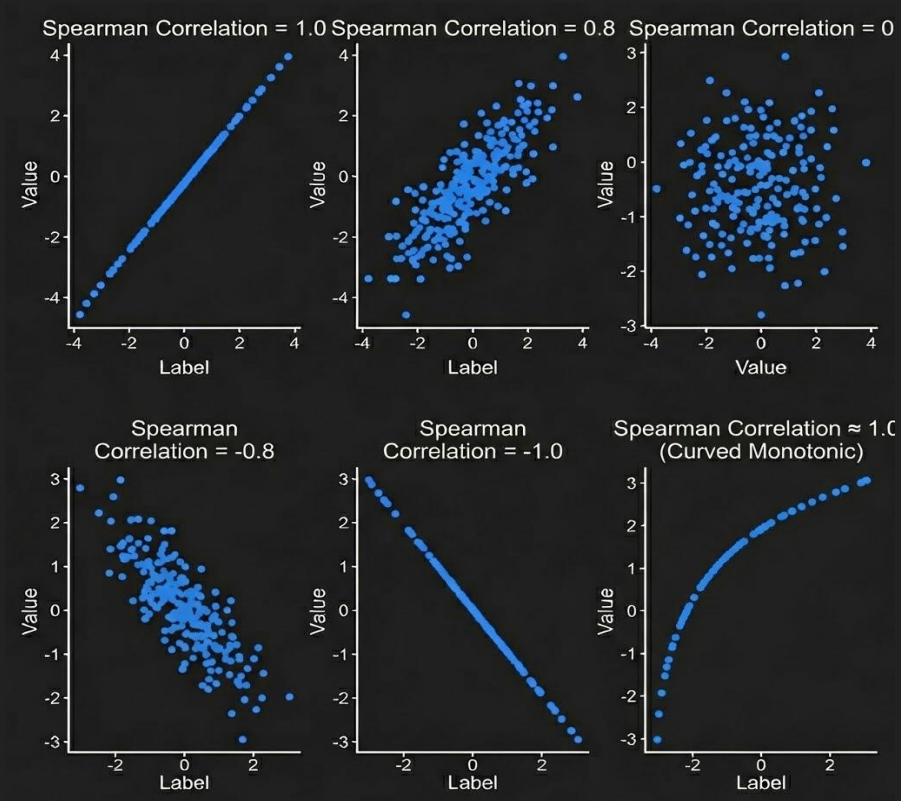
Spearman correlation measures **monotonic association** between two variables.

It answers a different question than Pearson:

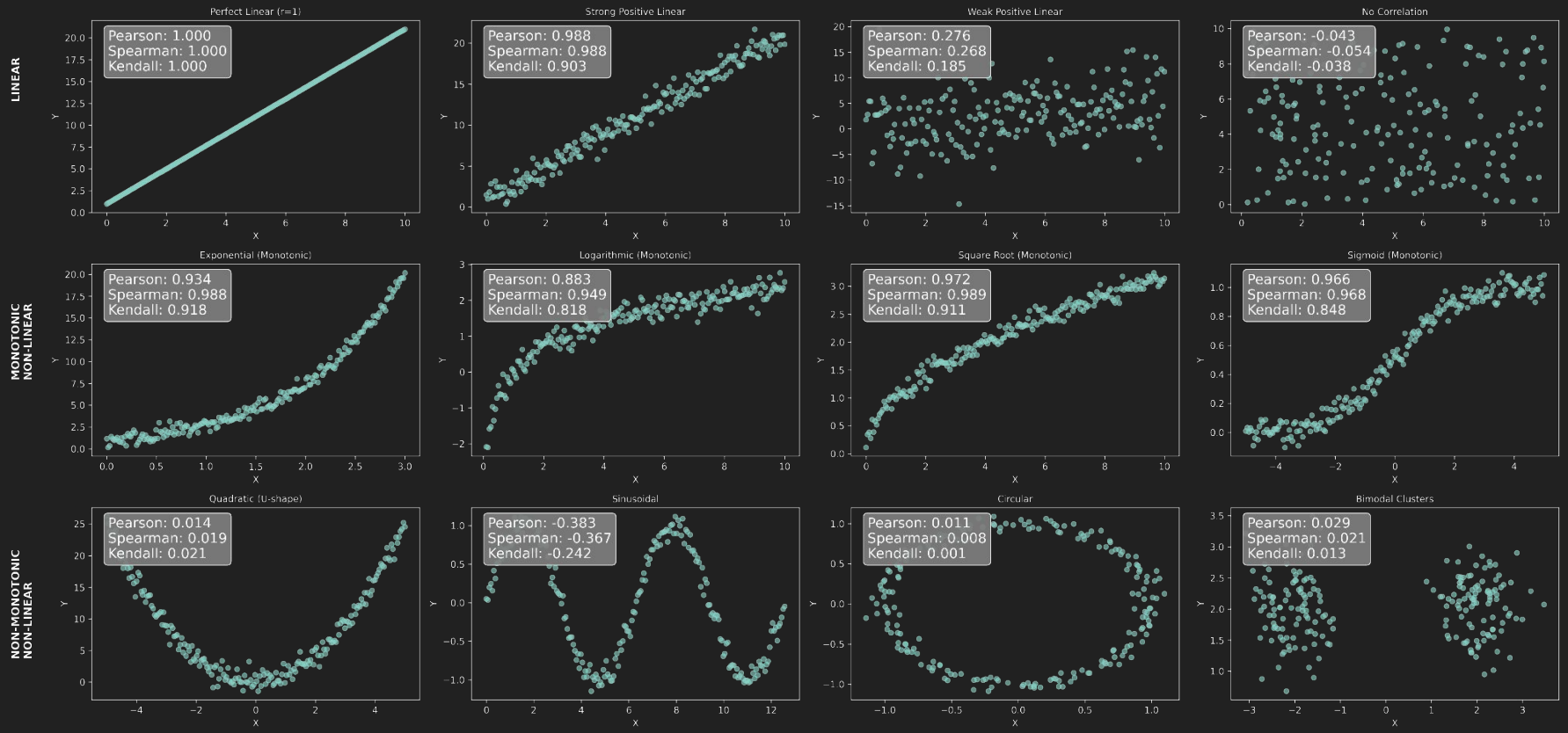
> **As one variable increases, does the other tend to increase or decrease — consistently, even if not linearly?**

Properties:

- Measures **linear relationships**
- Ranges from
  - 1: Positive linear relationship
  - 0: No linear relationship
  - -1: Negative linear relationship
- Insensitive to outliers
- Invariant under invariant transformations



Correlation metrics: Comparison



Correlation metrics: Contingency Tables

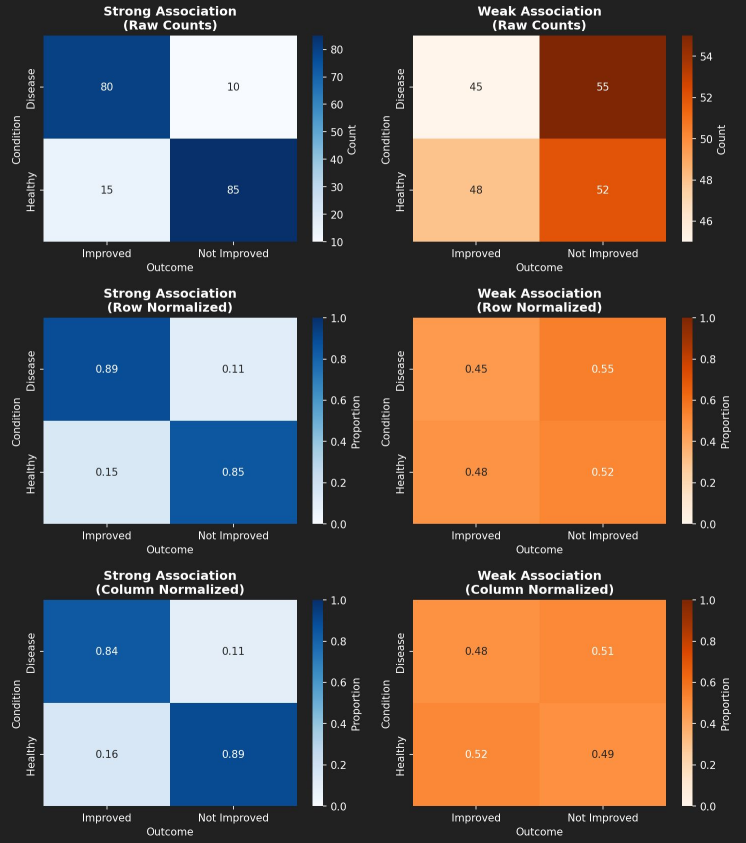
A contingency table summarizes **how often combinations of categorical variables occur**

They are not a single number but a relation between categories

Should look at:

- Rows with different distribution across columns
- Rows with different distribution across rows
- Cells that are unexpectedly largo or small

Uniform colors indicate small gains



Correlation metrics: Cramér’s V

Cramér’s V is a **measure of association strength** between **two categorical variables**.

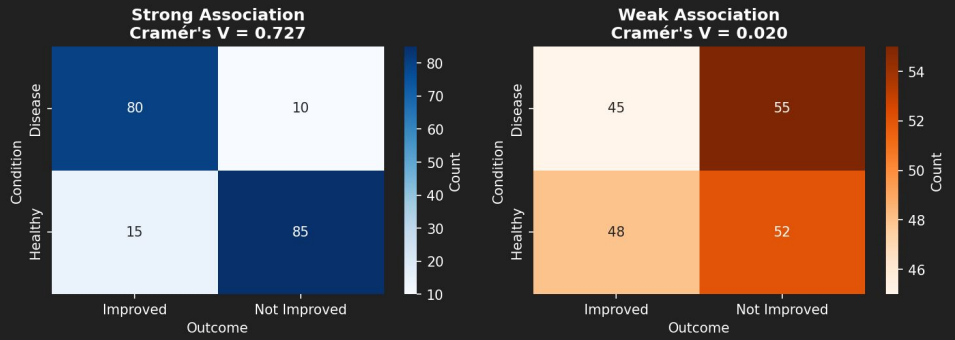
It answers one question:

> **How strongly are these two categorical variables associated in the observed data?**

Cramér’s V is computed from a **contingency table**. So it is a way of summarizing how uneven a table is compared to what uneven the table would look like.

Properties:

- Scale between 0 (no association) and 1 (perfect association)

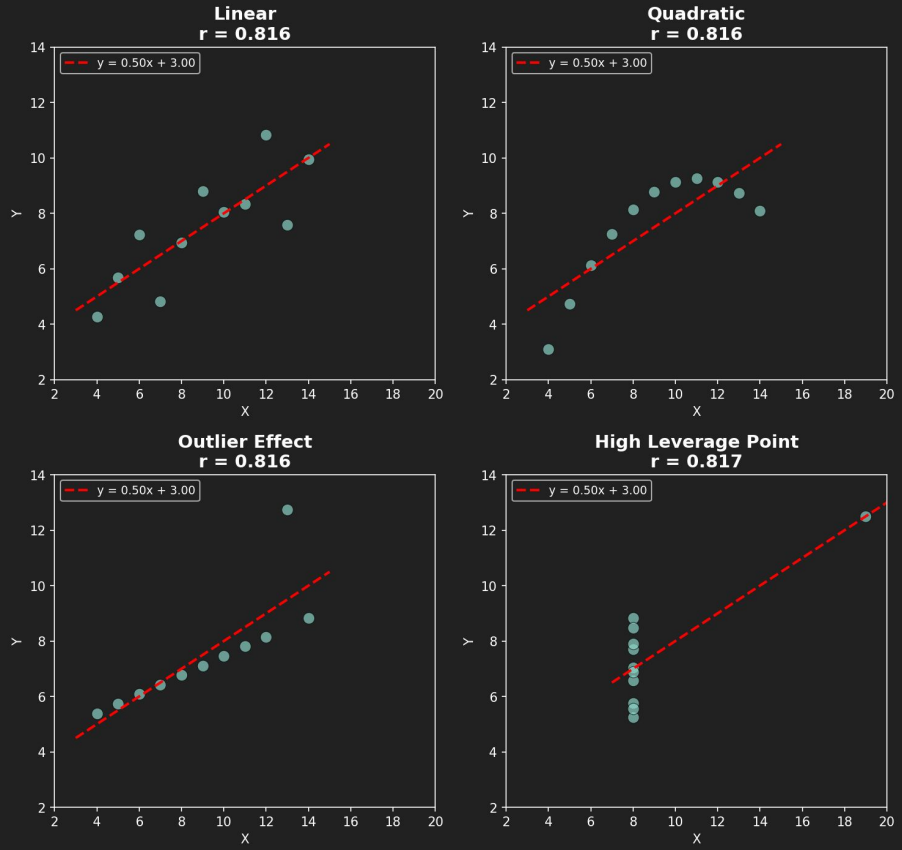


Correlation is not the same as Relationship

Same Parson coefficient, different associations

- Correlation values do not uniquely identify structure.
- Correlation *is* useful — just not as a verdict.
- It is good for:
  - Detecting potential linear or monotone structure
  - Comparing strength across many variable pairs
  - Flagging candidates for deeper analysis

  Summarizing patterns *after visualization*
- **If you report a correlation without a plot, you haven't described the relationship.** Correlation is a shadow of the relationship — useful, but never the object itself.

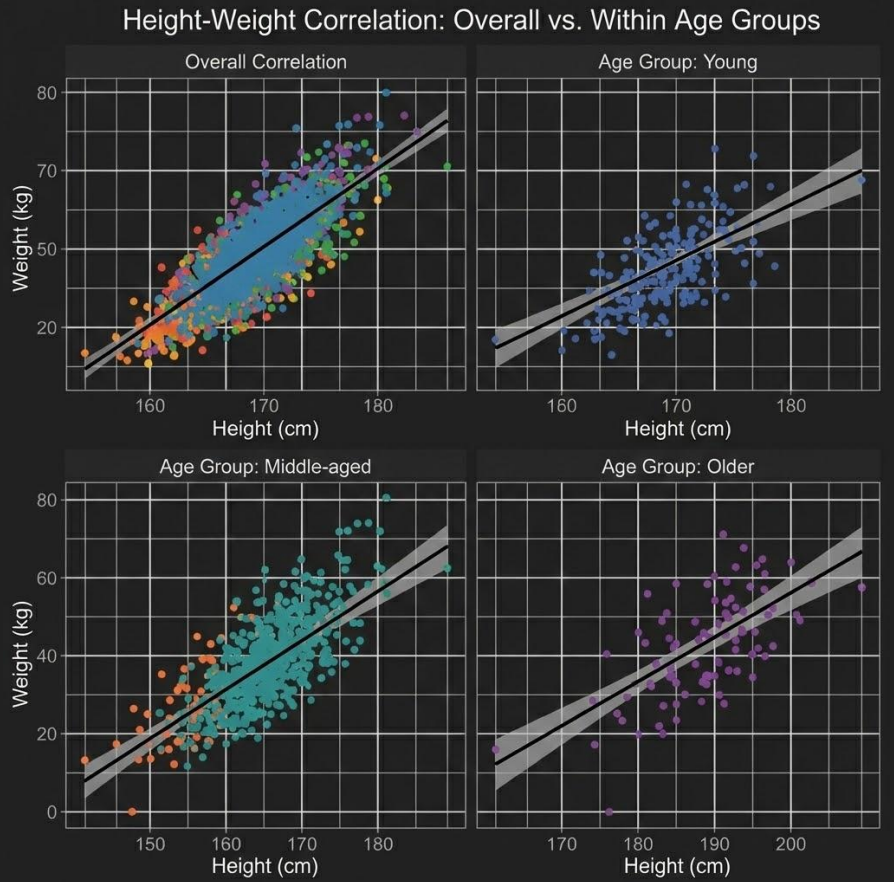


Conditional Correlation

Conditional correlation asks whether two variables still are correlated once we look only at situations where a **third variable is the same**.

e.g. weight and height are correlated overall — are they also correlated when we compare people of the same age?

- **Aggregation:** pooling all data together to compute a single summary measure. Assumes:
  - All observations are comparable
  - Differences between groups do not matter
  - Mixing doesn't distort meaning.
- **Stratification:** separating the data by condition and computing measures within each group. It compares like with like, instead of mixing different situations.



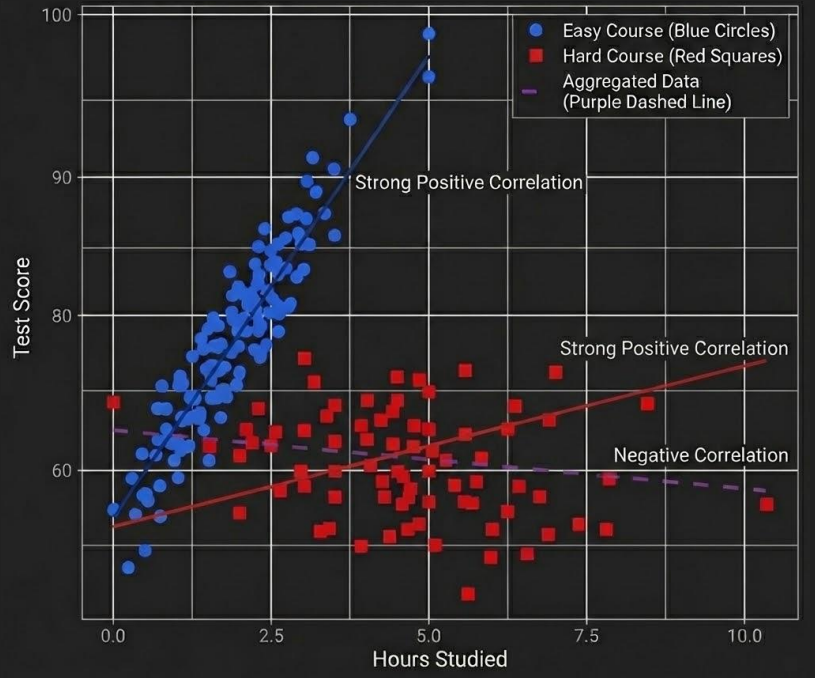
Simpson’s paradox

Simpson’s paradox happens when aggregation hides patterns that stratification reveals.

e.g. **Overall**, test score and hours of study appear negatively related.**Within each difficulty level**, more study is associated with higher scores.

Simpson’s paradox shows that aggregation can mislead — **not that stratification is always correct.**

Always ask: am I mixing different situations?



Correlation is not Causality

Simpson's paradox happens when aggregation hides patterns that stratification reveals.

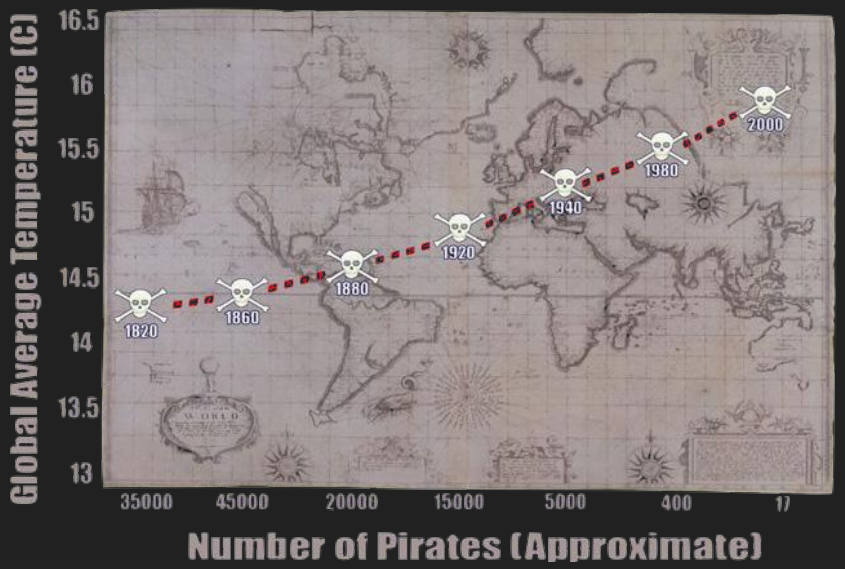
e.g. There is an inverse monotonic correlation in the temperature of the sea and the number of pirates in the ocean.

What if we do conditioning on time as the third variable? We would be answering:

"Do pirates and temperature have a trend *within the same year*?"

The causal question is:

"If I changed the number of pirates, would temperature change?"



"Correlation needs context, design and humility"

So why we bother with correlations at all?

Correlation is useful when you want to:

- compare many relationships
- track changes
- summarize consistently
- communicate succinctly
- feed later analysis

Correlation is *not* useful when:

- you haven’t plotted
- the relationship is clearly nonlinear
- the number will be misread as causal

Humans are great at: spotting patterns, seeing direction, noticing outliers

Humans are *terrible* at: comparing subtle strength differences, being consistent across plots, remembering magnitudes

Two plots can *look* similar but differ meaningfully in strength. Correlation gives you:

- a stable scale
- a reference point
- a way to say “this is stronger than that”

Not truth — **comparability**

Study Design

Index

Learning Outcomes:

- What I should be aware before during and after designing a statistical experiment?

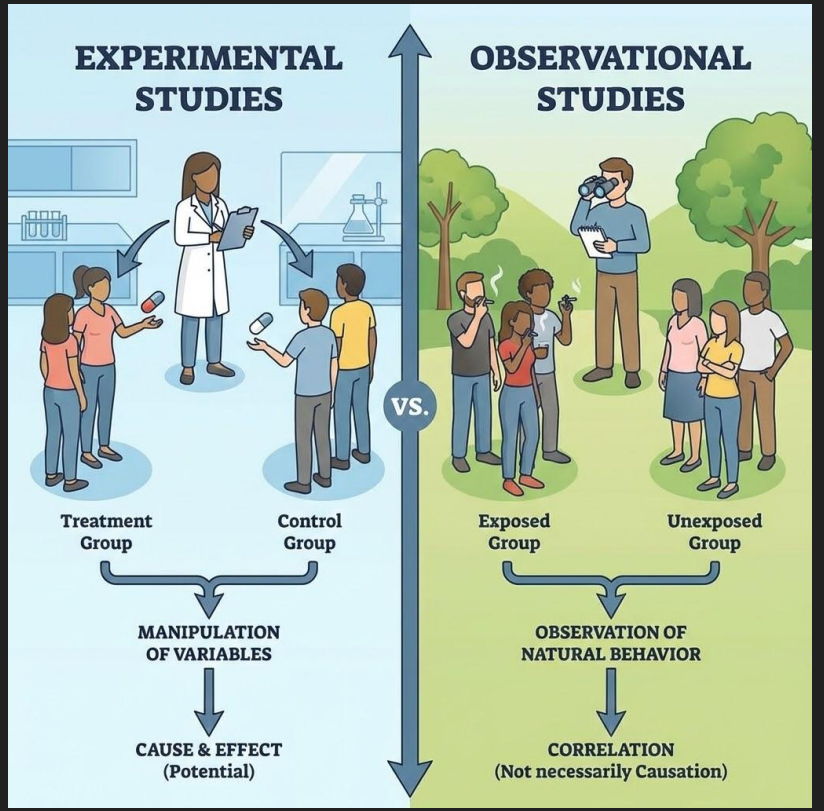
Experimental vs. Observational Studies

Experimental studies

- Assignment is under control
- Groups are comparable *by design*
- Randomization breaks many spurious associations

Observational Studies

- No control over assignment
- Groups may differ in many hidden ways
- Correlation and conditioning must do more work

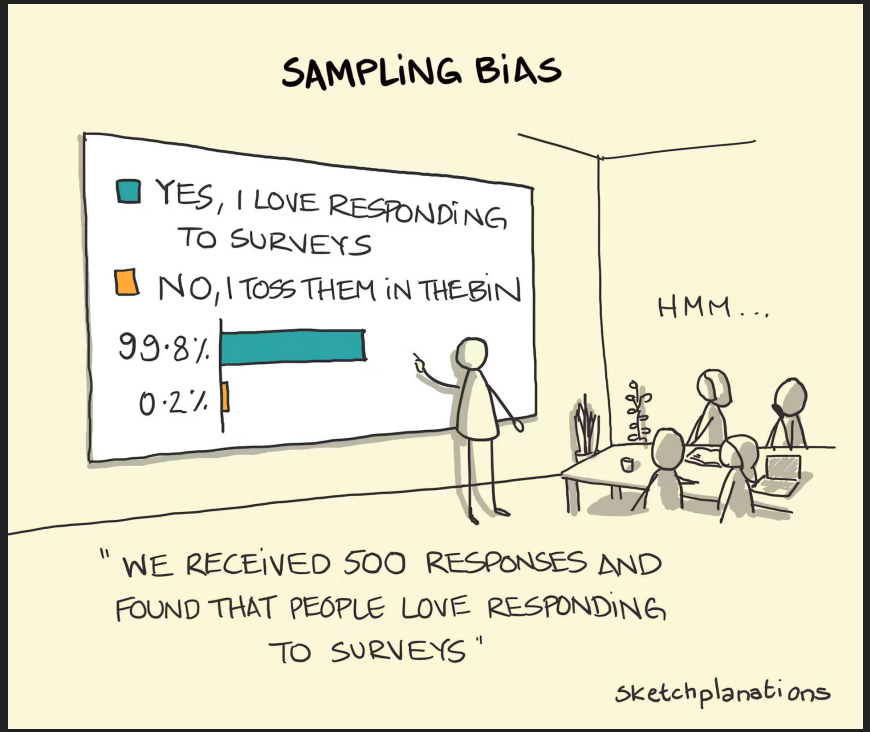


Bias

Bias is **systematic distortion**, not random noise. It does not average out with more data.

Common types:

- Selection bias (who appears in the data)
- Measurement bias (how variables are recorded)
- Survivorship bias (who remains observable)
- Reporting bias (what is missing or censored)



Randomization and Blocking

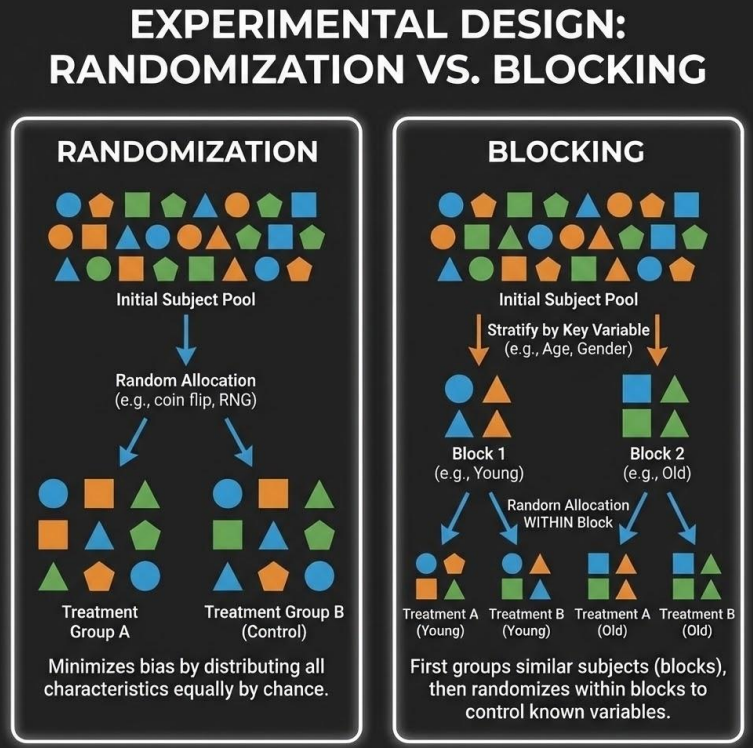
Randomization and **Blocking** are techniques to balance out **some preconditioning** known and unknown biasing factors.

Randomization assigns conditions by chance, not choice.

Blocking groups similar units **before randomization**.

It does not solve:

- measurement bias (bad sensors, bad labels)
- attrition bias (who drops out)
- noncompliance (people not following assignment)



Randomization and Blocking

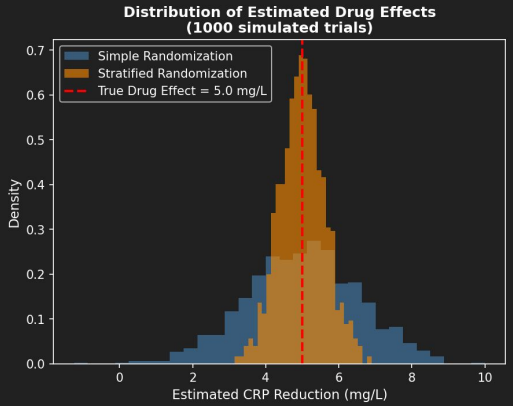
e.g. Scenario: Testing a new anti-inflammatory drug's effect on CRP (C-reactive protein) levels

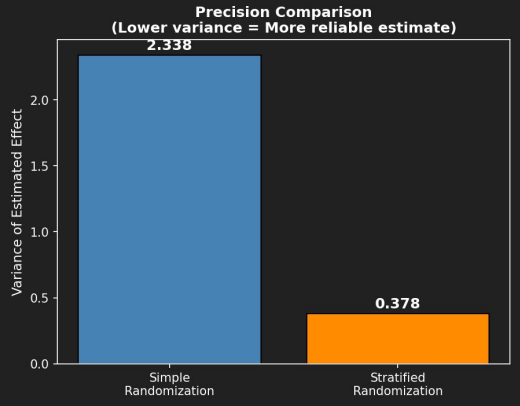
- **Parameter A (confounding):** Disease severity (Mild/Moderate/Severe) - affects baseline CRP
- **Treatment:** Drug vs Placebo
- **Outcome:** CRP reduction (mg/L)

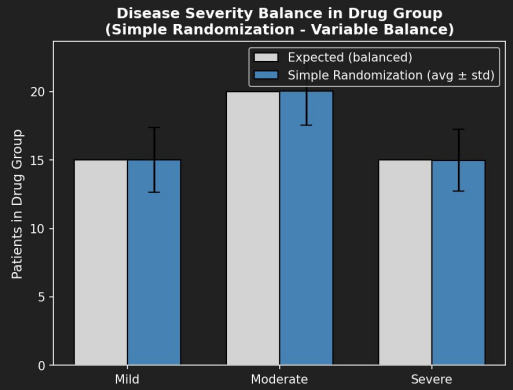
Comparisons:

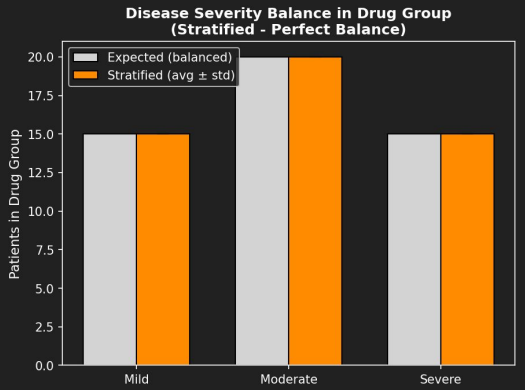
- Simple Randomization: ignores disease severity, leading to potential imbalance
- Stratified Randomization: ensures equal severity distribution in drug/placebo groups

Key insight: Stratified randomization (blocking) reduce variance and require smaller sample sizes.ç









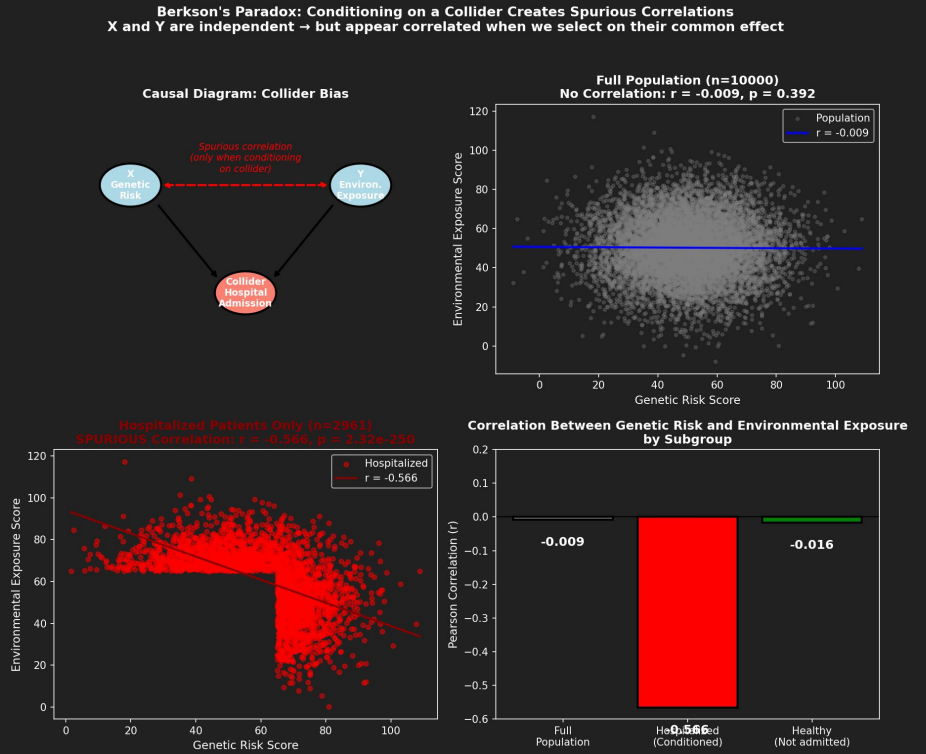
Confounders

A **confounder** is a type of bias coming by a variable that is related to multiple variables of interest that mix distinct situations together.

- Produces the Simpson's paradox
- Misleading correlations
- Unstable conditional relationships

Confounding can be mitigated by randomization/blocking by conditioning and for other statistical techniques (we will see them later).

Care note: Conditioning on confounders that depend on the tested variables can generate spurious correlations.



Outliers II: Trimming vs Winsorizing

Trimming removes extreme observations entirely.

e.g. drop top and bottom 1% of values

Pros

- simple
- limits extreme influence
- transparent if documented

Cons

- reduces sample size
- can bias estimates
- throws away real data
- silently changes the population

Winsorizing keeps all observations but caps extreme values.

e.g. values above 99th percentile → set to 99th percentile

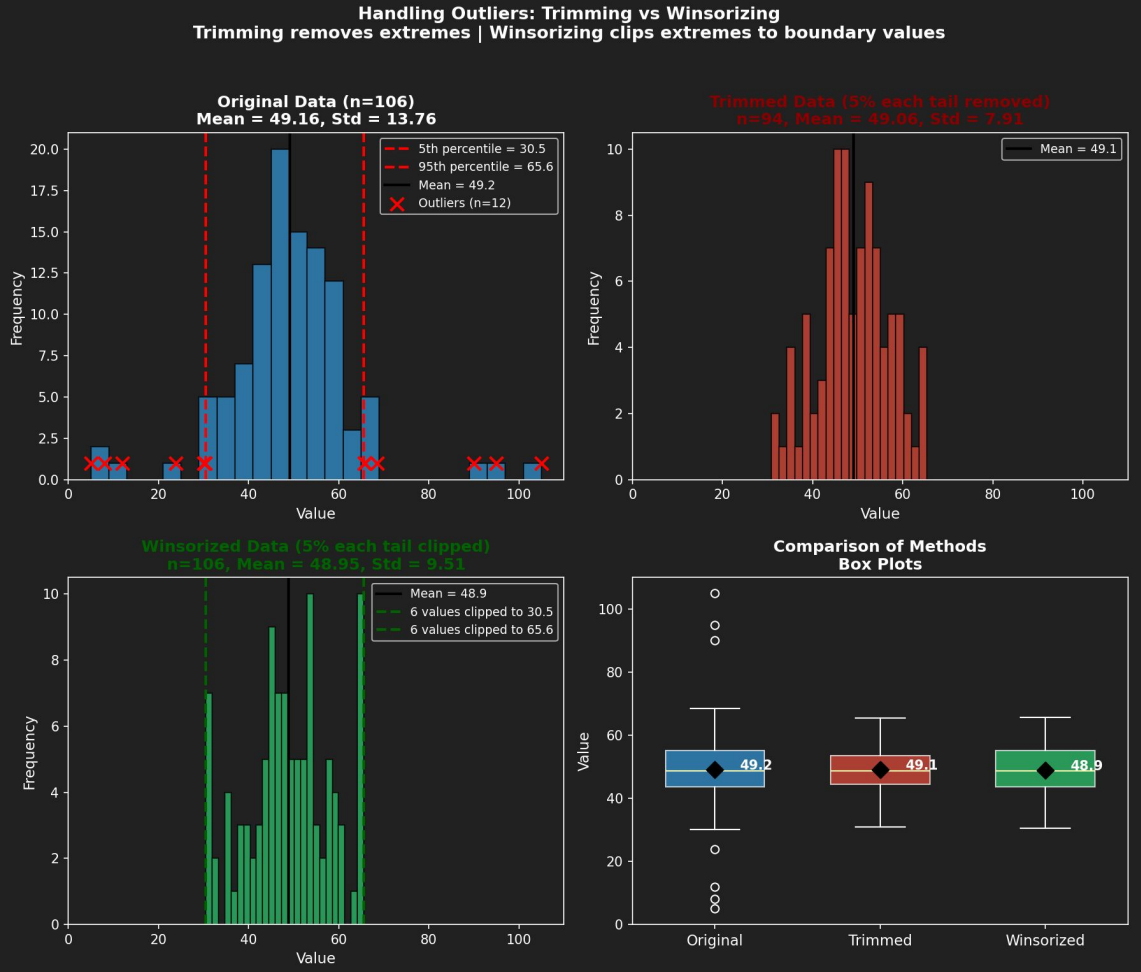
Pros

- keeps sample size
- limits leverage of extremes
- often more stable than trimming

Cons

- alters real values
- hides true extremes
- still arbitrary if unjustified

Outliers II: Trimming vs Winsorizing



Pre-registering Analysis Plans and Cleaning

Pre-registration means deciding what you will do *before* seeing the answers.

That includes:

- which variables matter
- which comparisons you will report
- how you will clean the data
- how you will handle outliers
- which analyses are **confirmatory vs exploratory**

Patterns discovered in exploratory analysis must be validated on data that did not influence their discovery.

Audit Trails and Data Provenance

Data provenance: “Where did this data come from?”

Data provenance means knowing:

- how the data were collected
- by whom
- under what conditions
- with which instruments
- and with which known limitations

If you don’t know this:

- you don’t know what the data can support
- no amount of statistics can rescue it

Audit trails: “What happened to the data?”

An **audit trail** is a complete record of:

- filtering
- exclusions
- transformations
- recording
- imputations
- modeling decisions

This should answer:

How did raw data become results?

Probability Foundations

Index

Learning Outcomes:

- What is probability?
- How I can understand the data generated from a probabilistic perspective?

Sample Spaces

The **sample space** is the set of all possible outcomes of an experiment.

We name it usually as

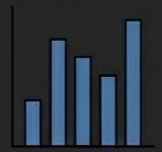
Ω

- Coin toss ⇨ {H, T}

- Die roll ⇨ {1,2,3,4,5,6}

- Exam score ⇨ [0,10]

- number of proteins ⇨ {0,1,2,...}



Random Variable

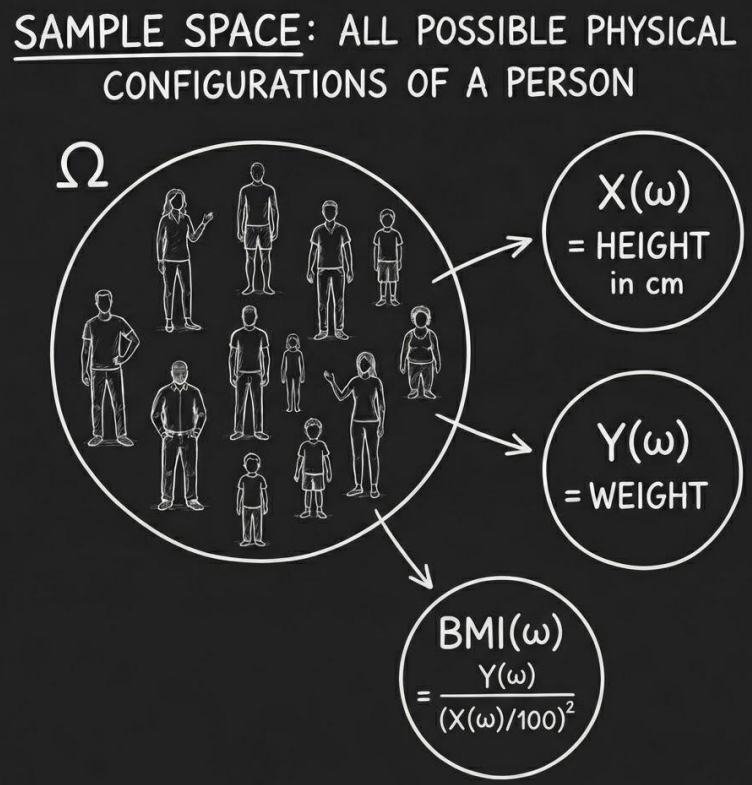
A **random variable** is a function mapping outcomes of the sample space to a number

$$X : \Omega \to \mathbb{R}$$

A sample space can be associated with many random variables

A transformation of a random variable is another random variable

What we measure in the real world are realizations of random variables defined on an underlying probability model



Probability

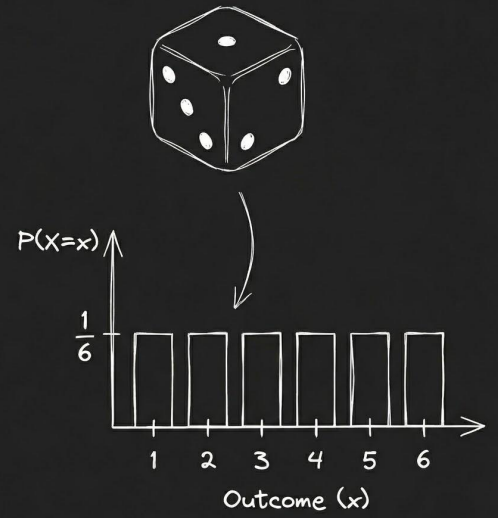
Probability measures how likely an outcome is to occur, on a scale from 0 to 1.

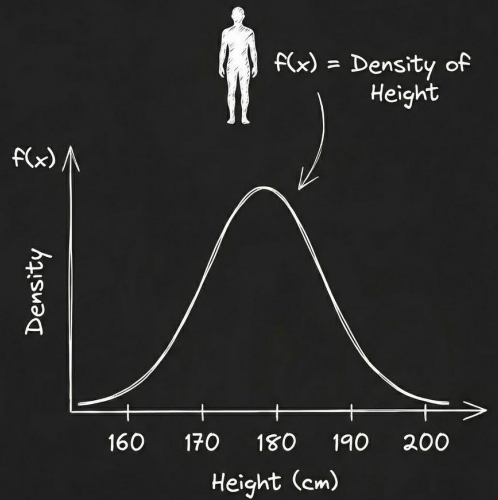
If the random variable is discrete we define its **probability mass function**

$$P(X = x)$$

If continuous we define its **probability density** (f)

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$





Probability Summary Statistics

As in the datasets, in a probability distribution we can compute also summarizing statistics.

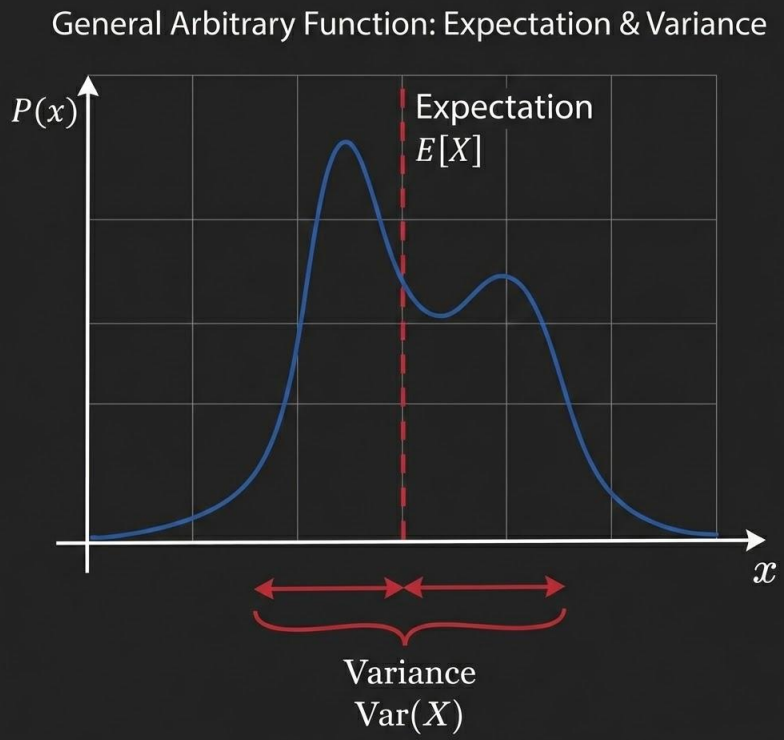
Among the most common are the:

- Expectation (“mean”)

$$\mathbb{E}[X]$$

- Variance

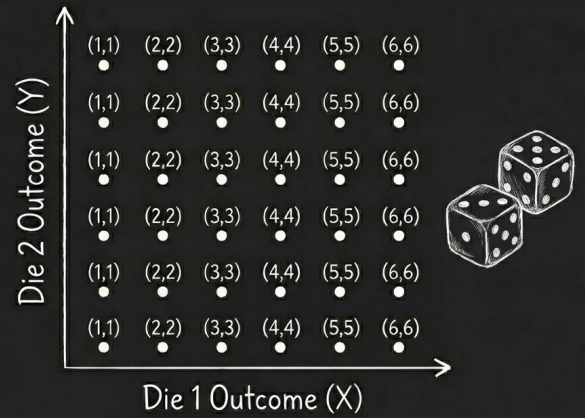
$$\mathrm{Var}[X]$$



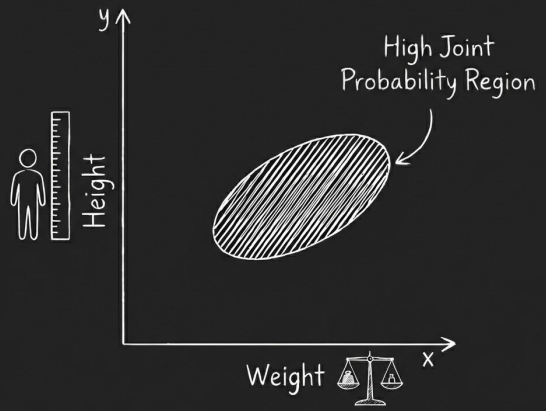
Joint probability

Joint probability measures the probability that two (or more) events occur together.

$$P(X, Y)$$



$$P(X=x, Y=y) = p(x,y)$$



$$P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x,y)\, dy\, dx$$

Conditional Probability and Independence

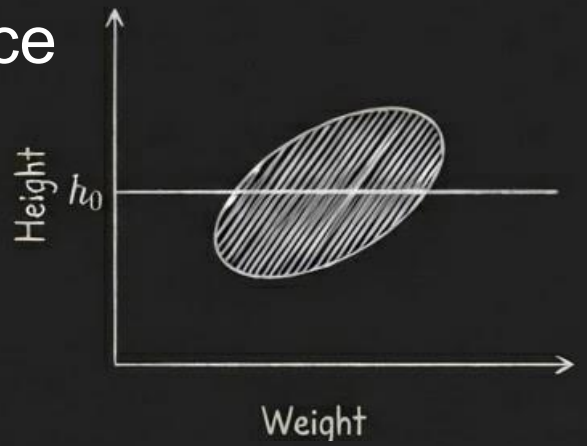
The conditional probability measures the probability of a random variable conditional on other random variable

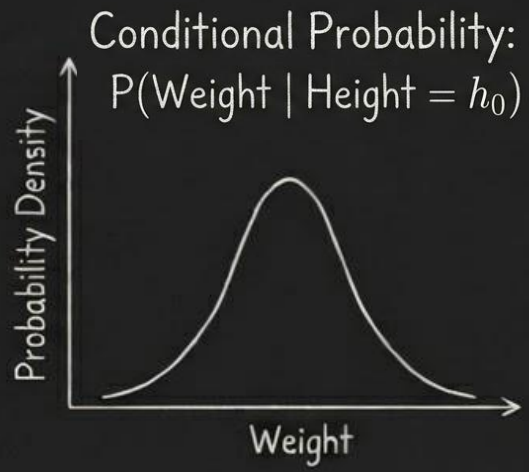
$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

Two random variables are **independent** if the probability of one variable does not depend on the other

$$P(X|Y) = P(X)$$

e.g. Knowing the result of one die does not give us any clue of the result of the other





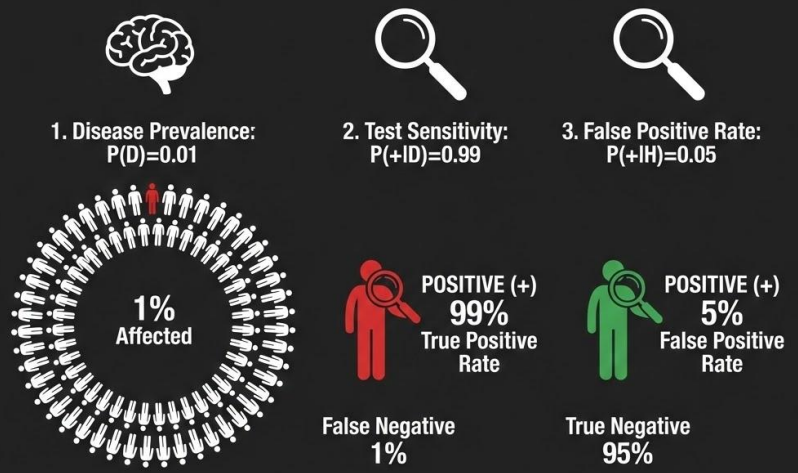
Bayes Rule

Bayes' rule expresses one conditional probability in terms of the reverse conditional probability and the marginal probabilities.

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \qquad P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(Y|X)P(X) = P(X|Y)P(Y)$$

e.g. What is the probability that I actually have a disease if I give positive in a test?

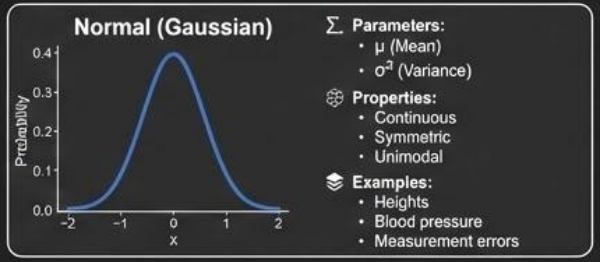


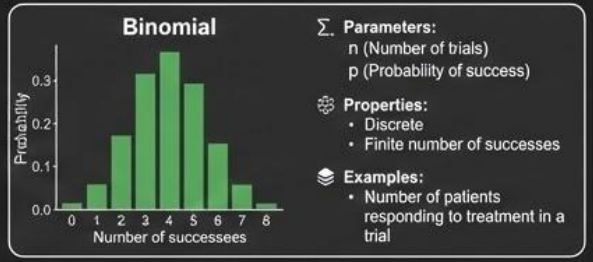
$$P(D|+) = \frac{P(+|D) \cdot P(D)}{P(+)} = \frac{0.99 \cdot 0.01}{0.0594} = \frac{0.0099}{0.0594} \approx 0.167$$

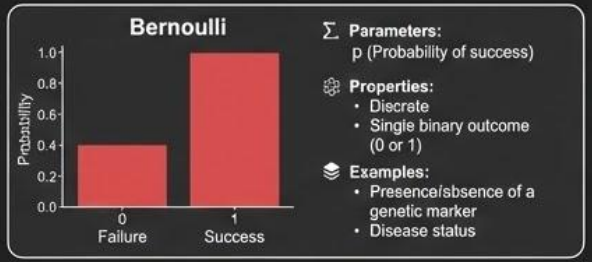
Common distributions

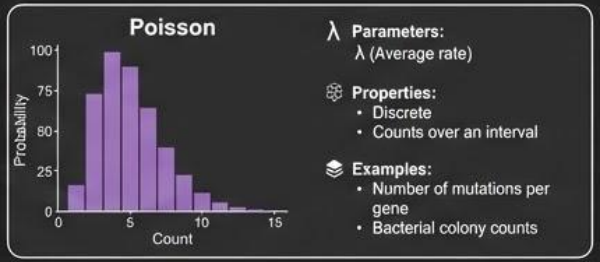
Each distribution describes a different phenomenon of life

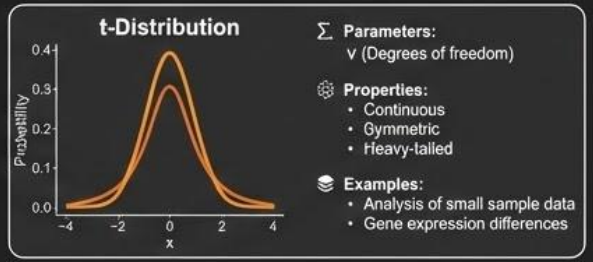
Each is characterized by a different set of parameters

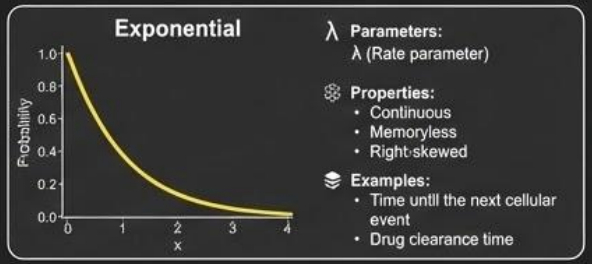


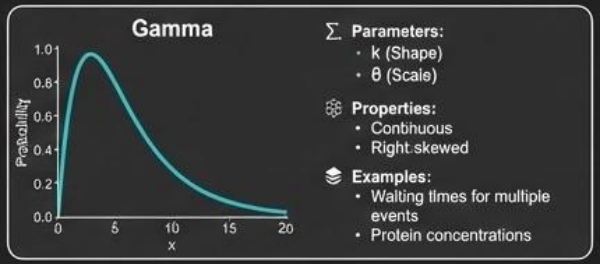












Estimation, Confidence Intervals and Error Propagation

Index

- Motivation
- Point Estimates and Estimators
- Law of Large Numbers (LLN)
- Confidence Intervals (CI)
    - Analytic
    - Central Limit Theorem (CLT)
    - Bootstrapping
- Error Propagation with Bootstrapping
- Robust Estimators
- Correlated Inputs

Learning Outcomes:

- What is the best estimate to describe my data, and how uncertain is that estimate?

Motivation

We have data that was produced according to a probability distribution.

e.g. the height of a person can be described by a normal distribution

$$P(\text{height}|\mu,\sigma)$$

I can measure many people's height.

- How can I know using my data the parameters of my distribution?
- With what confidence can I trust these parameters based on my data?

$$P(\mu,\sigma|\{\text{heights}\})$$

Point Estimates and Estimators

EXAMPLE CONTEXT: Survival Experiment
Objective: Interested in the probability of survival.
Model: Bernoulli distribution for data.
Parameter: "p" = True probability of survival.

ESTIMATOR: The Rule or Formula

produces an estimate from a sample.

Analogy: Like a recipe for baking a cake.

Example (Formula):
$\hat{p} = \text{mean}\{survived\}$

Nature: Abstract, General Rule

ESTIMATE: The Specific Value

obtained by applying the estimator to a specific sample.

Analogy: Like the actual slice of cake you eat.

Example (Numerical):
$\hat{p} = 0.8$ (e.g., $80/100$)

Nature: Concrete, Specific Number

Apply to Sample

Why the Sample Mean ($\hat{p}$) is a Good Estimator for p

Population (Bernoulli Distribution)

p

True Probability of Survival (unknown)

↓

Sample Data

1, 0, 1, 1, 0, ...

↓

Estimator: Sample Mean ($\hat{p}$)

$$\hat{p} = \frac{\text{Sum of outcomes}}{n}$$

⇨

Why It's Good: Expected Value

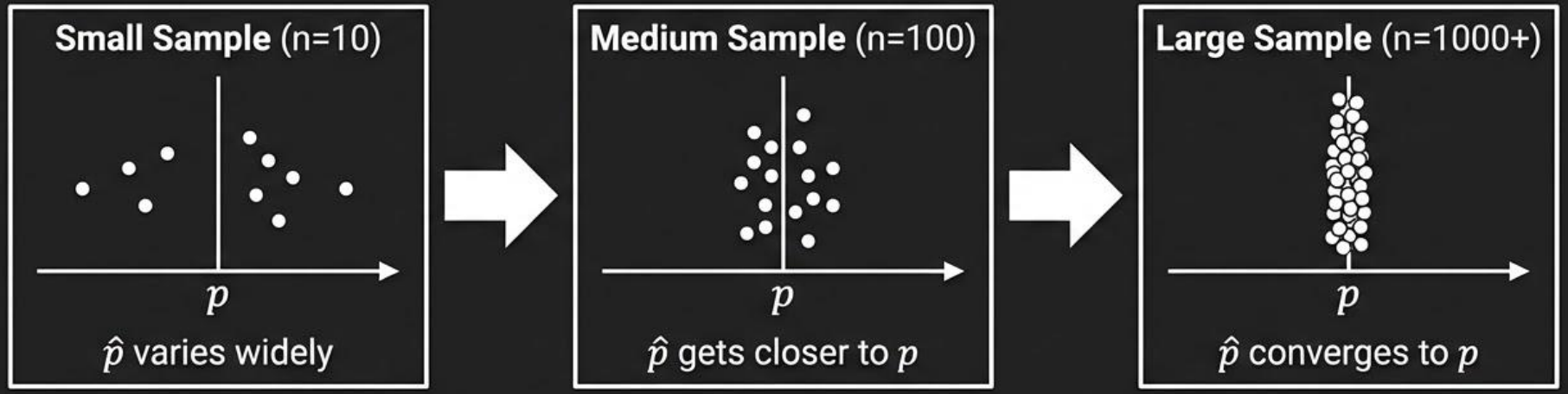
$E(\text{Bernoulli Trial}) = 1*p + 0*(1\text{-}p) = p$

$E(\hat{p}) = E(\text{mean}\{\text{survived}\}) = p$

The expected value of the sample mean equals the true parameter p, making it an unbiased estimator.

Law of Large Numbers

The Law of Large Numbers states that as the sample size (n) grows, the sample mean ($\hat{p}$) approaches the true population mean (p) with high probability.

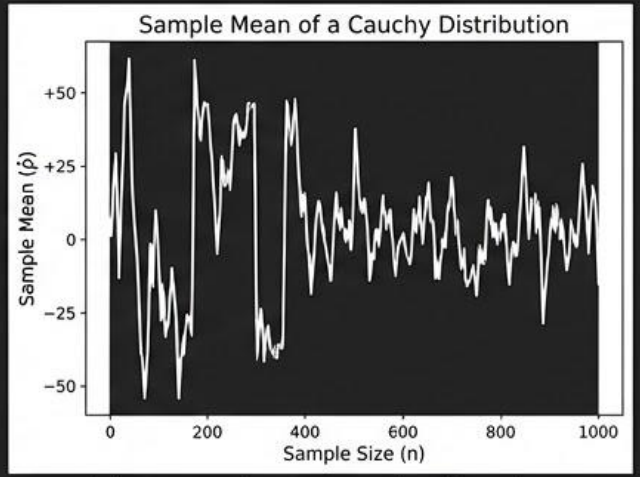


Many estimators are defined in terms of the sample mean, so this convergence is quite ubiquitous in most of the estimators you will find!

Law of Large Numbers

The Law of Large Numbers (LLN) relies on specific assumptions. It can fail or be inapplicable when these are violated:

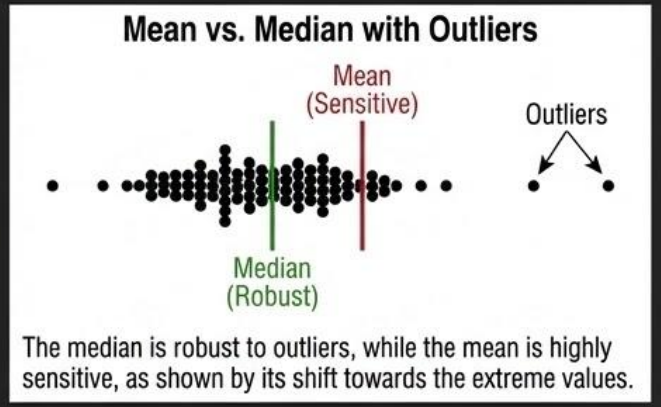
- **Non-Independent Samples (Correlation):** If observations are correlated (e.g., time series), convergence is slower or may not occur.
- **Non-Identical Distributions (Non-Stationarity):** If the underlying distribution changes over time, there is no single 'true mean' to converge to.
- **Heavy-Tailed Distributions (Infinite Mean/Variance):** For distributions like the Cauchy, the theoretical mean is undefined. The sample mean will fluctuate wildly and never converge, regardless of sample size.



The sample mean of a Cauchy distribution does not converge.

Robust Estimators

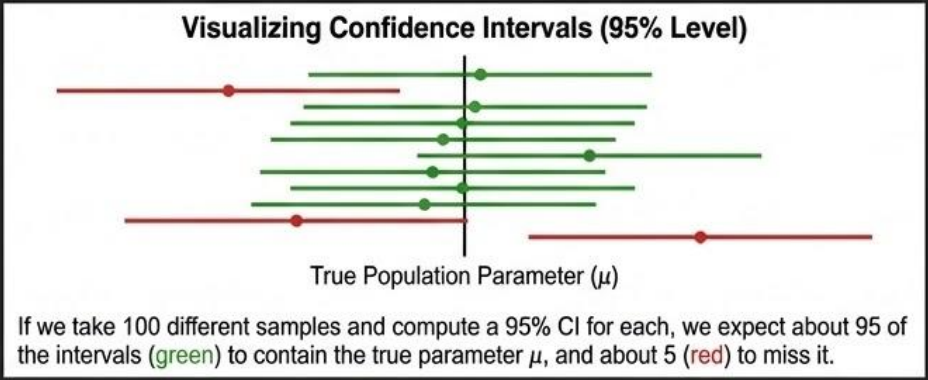
- Estimators that are not heavily influenced by outliers or deviations from the assumed distribution.
- Provide a more stable and representative measure of the population parameter when data is not perfectly well-behaved (e.g., heavy tails, outliers).
- Examples:
  - **Median:** A robust alternative to the Mean for central tendency.
  - **Interquartile Range (IQR):** A robust alternative to the Standard Deviation for spread.
  - **Trimmed Mean:** Removes a percentage of the most extreme values before calculating the mean.



A confidence interval is valid for the parameter defined by the model.
If the model poorly represents reality, the interval may not answer the scientific question of interest.

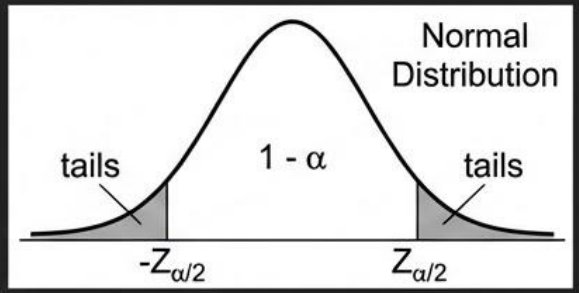
Confidence Intervals

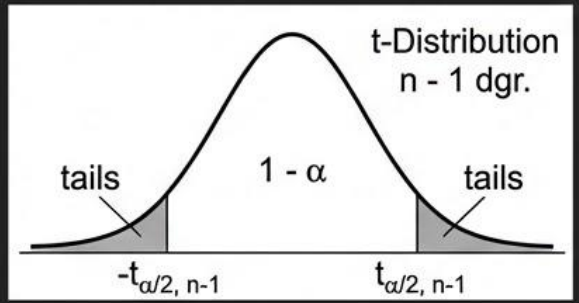
- A **confidence interval (CI)** is a range of values, derived from sample statistics, that is likely to contain the true population parameter.
- The **confidence level** (e.g., 95%, 99%) represents the long-run proportion of such intervals that would contain the true parameter if we were to repeat the sampling process many times.
- It is **not** the probability that the true parameter falls within a specific calculated interval. The true parameter is fixed; the interval is what varies from sample to sample.
- Constructed using the point estimate (e.g., sample mean $\bar{x}$) and the margin of error, which depends on the standard error and the chosen confidence level (often using the normal or t-distribution, based on the CLT).



Confidence Intervals: Analytic Confidence Intervals

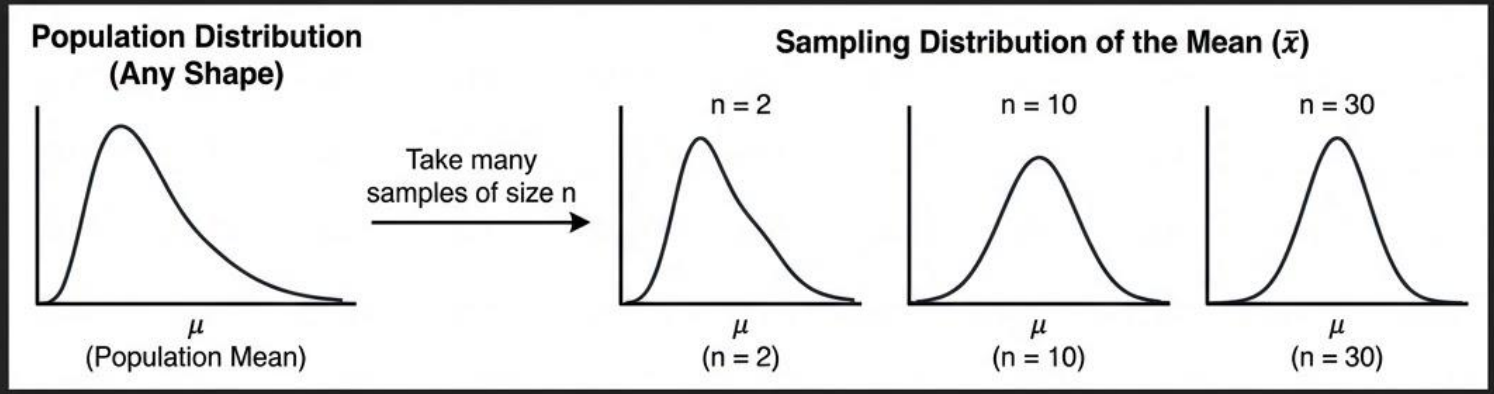
- Methods based on theoretical distributions (e.g., Normal, t-distribution).
- Requires assumptions about the population distribution and sample size.
- Common approaches for:
  - Population mean (μ) with known variance (σ^2).
  - Population mean (μ) with unknown variance (s^2).
  - Population proportion (p).





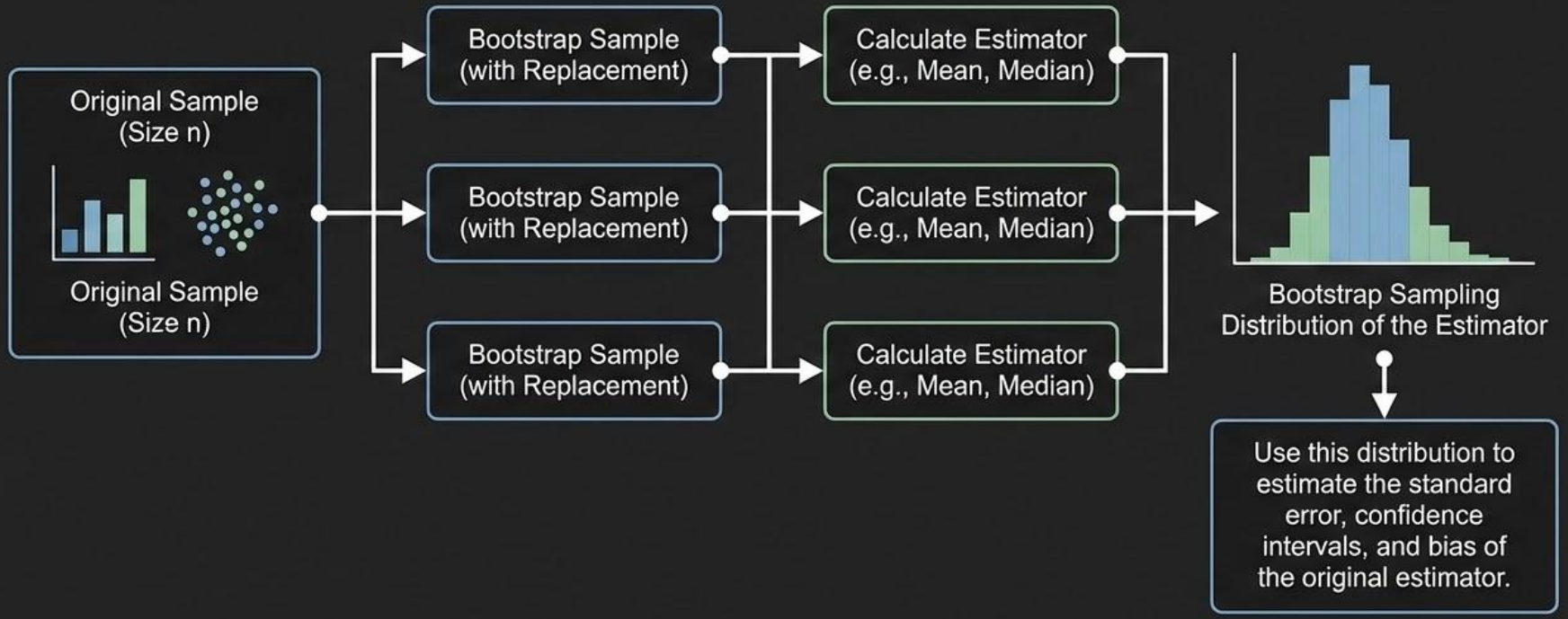
Confidence Intervals: Central Limit Theorem

- The Central Limit Theorem states that the sampling distribution of the sample mean will approach a **normal distribution** as the sample size (n) increases, regardless of the shape of the population distribution.
- This holds true for any underlying distribution, as long as it has a finite variance.
- Key Condition: The sample size must be sufficiently large (often n ≥ 30 is a rule of thumb).



The CLT explains why the normal distribution is so prevalent in statistics and allows us to make inferences about the population mean even when the population distribution is unknown.

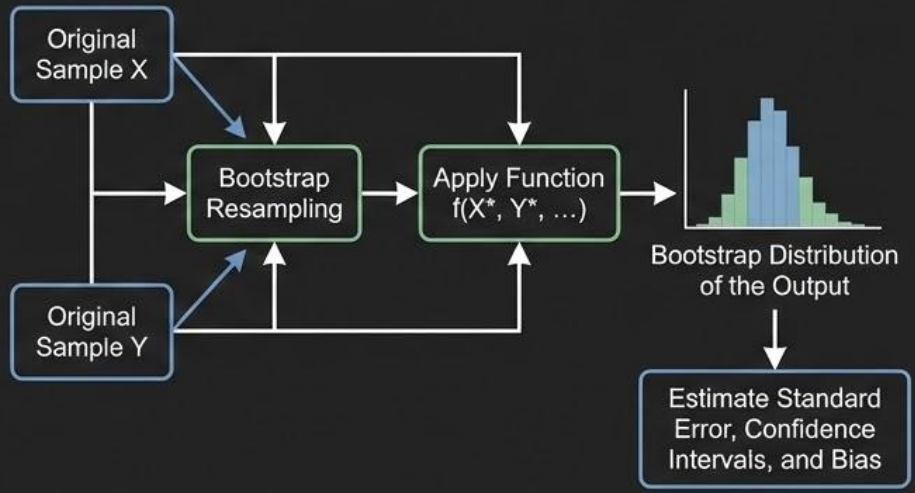
Confidence Intervals: Bootstrapping



Error Propagation with Bootstrapping

A Non-Parametric Approach for Complex Functions

- When the function is complex or the input distributions are unknown, analytic methods can be difficult.
- Bootstrapping simulates the sampling distribution of the inputs to estimate the uncertainty of the function output.
- This method is non-parametric and flexible, applicable to a wide range of functions and estimators.
- Develops a functality of motivation.
- It is robust and doesn't rely on the Central Limit Theorem.



Assumption of Independence

- A crucial assumption for many statistical methods (e.g., t-tests, linear regression).
- Observations must not influence each other; knowing one outcome shouldn't predict another.
- Violating this assumption can lead to:
  - **Biased** estimates,
  - Underestimated standard errors,
  - Invalid confidence intervals and p-values.
- Common sources of dependence:
  - Time series data (autocorrelation),
  - Clustered data (e.g., students in classrooms),
  - Repeated measures on the same subjects.