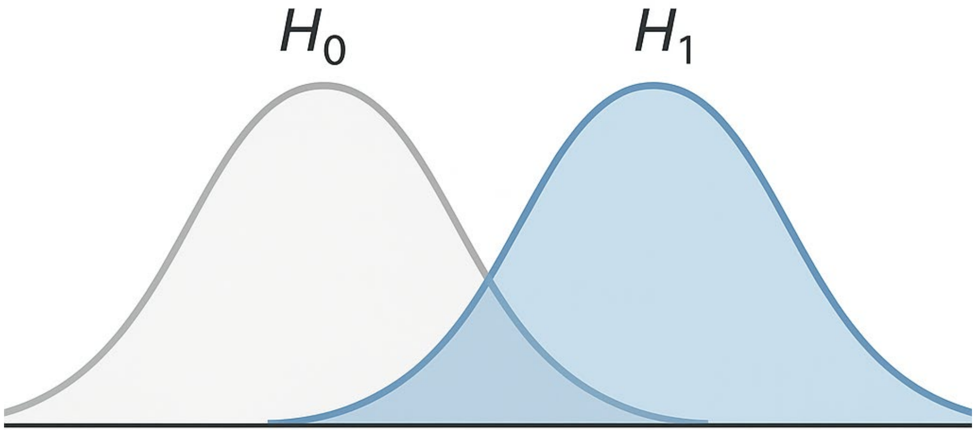
Hypothesis Testing II

- Quick review of last lecture
- Discussion on Multiple Testing & False Discovery
- Exam Review I
- Next lecture, more exam review and a few practice questions/problems

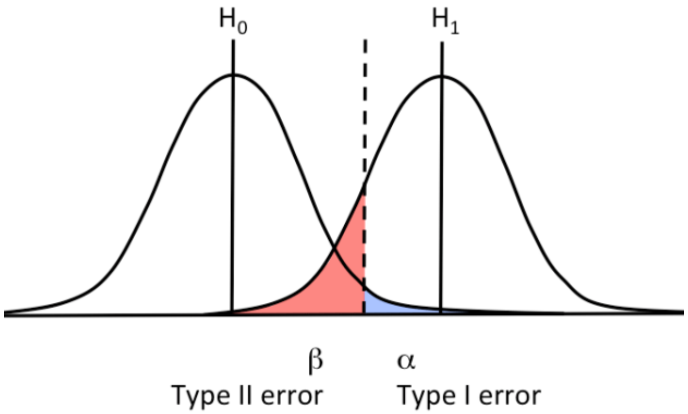
The Decision Framework Intro

- Each curve (H_0 and H_1) shows what your test statistic (like a difference in means) would look like if you ran the experiment many times.
- The x-axis is the value of that test statistic (e.g., the difference between conversion rates in Group A and Group B).
- The y-axis is the probability density of getting that value, assuming H_0 or H_1 is true.
- They are distributions of possible sample outcomes, *not* a histogram of your actual data



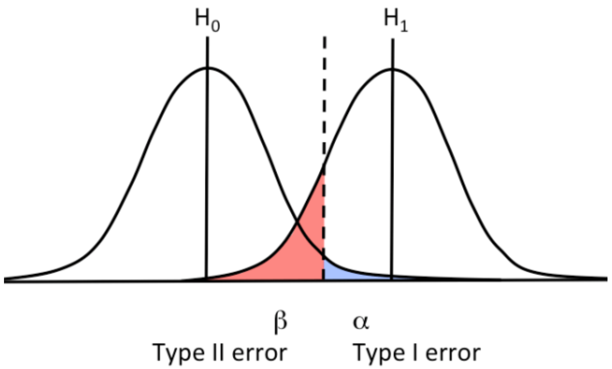
Summary

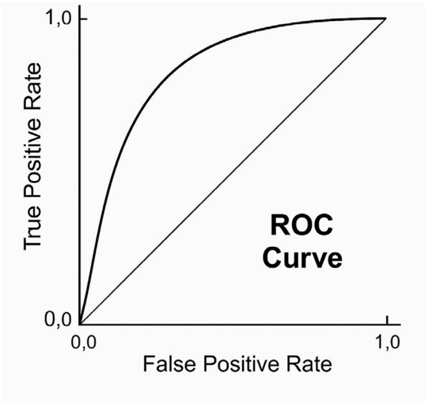
Region	True World	Decision	Probability	Meaning
Left of threshold under H_0	H_0 true	Fail to reject H_0	**$1 - \alpha$**	Correct "no effect" call
Right tail under H_0	H_0 true	Reject H_0	α	False positive
Left tail under H_1	H_1 true	Fail to reject H_0	β	Missed detection
Right of threshold under H_1	H_1 true	Reject H_0	**$1 - \beta$**	Correct detection (power)



ROC: Seeing All Thresholds at Once

- ROC (Receiver Operating Characteristic) curve = plots all thresholds.
- x-axis: False Positive Rate (α).
- y-axis: True Positive Rate (1−β) = Sensitivity.
- Each point = one possible decision threshold.
  - Area Under Curve (AUC): overall ability to separate signal from noise.
- Hypothesis testing → picks one α; ROC → shows all α–β trade-offs.

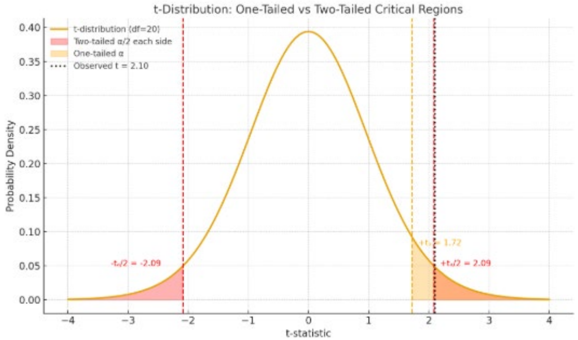


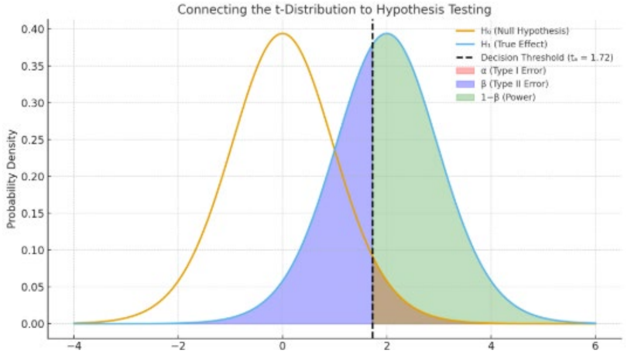


The t-Test: Hypothesis Testing

- The t-test is just one implementation of our hypothesis-testing logic.
- Example: compare two sample means (A vs. B).
- Null (H_0): no difference → $\mu_a = \mu_\beta$.
- Alternative (H_1): difference exists → $\mu_a \neq \mu_\beta$.
- Compute test statistic:

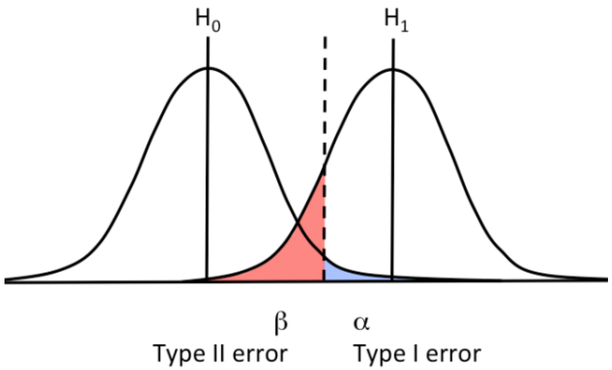
$$t = \frac{\bar{x}_A - \bar{x}_B}{SE_{\text{diff}}}$$





The Problem with Multiple Tests

- If you test enough things, something will always look significant.
- α = 0.05 → 5 % false positives by chance.
- 100 independent nulls → ≈ 5 false "discoveries."



When one test becomes a hundred

- We’ve treated hypothesis testing as a one-off decision — one null, one test, one α.
- But in data science, you rarely stop at one.
- You might run 50 correlations, 100 feature tests, or thousands of model comparisons.

Logic of Multiple Tests/Comparisons

- If there's a 5% chance of a false positive, there's a 95% chance of no false positive.
  - P(no false positive in one test)=1−0.05=0.95
- If you run 40 independent tests, and you want *none* of them to be false positives, you multiply that 0.95 probability across all 40:
  - P(no false positives in 40 tests)=$(0.95)^{40}$
  - $(0.95)^{40} \approx 0.13$
- So there's an 87% chance that at least one of your 40 tests will show up as "significant" purely by random luck.

What should you do?

- Adjust p’s or validate on a hold-out set

Goal	Tool	Effect
You want to **avoid any** false positives (confirmatory research)	**Bonferroni correction:** use α/m	Very strict; reduces false positives but may miss true ones
You want to **balance** discovery vs. caution (exploratory data analysis)	**FDR (Benjamini–Hochberg)**	Keeps proportion of false positives under control; allows some risk
You’re building a **predictive model**	**Use hold-out validation**	Let model performance (not p-values) confirm real signal

Adjusting p-Values

- Family-Wise Error Rate (FWER)
  - Probability of at least one Type I error across all tests
  - Very strict — controls false alarms anywhere
- False Discovery Rate (FDR)
  - Expected proportion of false positives among all significant results
  - More flexible — allows some false alarms but limits their proportion

Family Wise Error Rate

- FWER = P(At least one Type I error among all m tests)
- FWER $=1-(1-\alpha)^m$
- Bonferroni Correction: adjust p-value by dividing α/(number of tests)
  - From the previous example – 0.05/40 = 0.00125
- FWER protects you from *any* false discovery — great for confirmatory research, but often too strict for exploratory data science

False Discovery Rate (FDR)

- FWER tries to make sure you never call anything false, it can be too strict
- Common Method: Benjamini–Hochberg (BH)
  - Rank all p-values smallest → largest
  - Compute threshold line: $(i/m) \times \alpha$
  - Find the largest p_i below that line — everything below it is "significant."

BH Procedure Example

Rank (i)	p-value	BH threshold (i/100 × 0.05)
1	0.0005	0.0005 ✅
2	0.0010	0.0010 ✅
3	0.0020	0.0015 ❌

FWER vs FDR

- Run 10 hypothesis tests, alpha = 0.05

Bonferroni

p_i	Significant?
0.001	✅ yes
0.009	❌ no
0.015	❌ no
…	❌ no

Benjamini-Hochberg

Rank (i)	p_i	Threshold	Significant?
1	0.001	0.005	✅ yes
2	0.009	0.010	✅ yes
3	0.015	0.015	✅ yes
4	0.020	0.020	✅ yes
5	0.032	0.025	❌ no
6–10	…	…	❌ no

Exam Review

Types of Variables

- Nominal/Categorical
- Ordinal
- Boolean
- Discrete
- Continuous
- Datetime

- These are important to know because it will affect how you validate, explore, and ultimately model the data
- For example, continuous data lends itself to regression, but what about nominal?

Examples

Example (CSV or SQL):

`id`	`name`	`age`	`signup_date`
1	Alice	30	2023-01-10
2	Bob	24	2023-02-15

→ Rows = records (people), Columns = variables

→ Schema: `id` (int), `name` (str), `age` (int), `signup_date` (datetime)

```
{
  "user": "alice",
  "age": 30,
  "preferences": {
    "theme": "dark",
    "notifications": true
  }
}
```

```
data = np.array([
    [1.2, 3.5, 5.1],
    [4.4, 0.8, 2.9]
])
```

`timestamp`	`temperature`
2023-01-01 00:00	21.5
2023-01-01 01:00	20.8
2023-01-01 02:00	19.9

The Three Core Goals of EDA

- **Data Quality** – What’s missing, wrong, or suspicious?
- **Data Structure** – How is the data organized? What’s the distribution of variables?
- **Data Insight** – What trends or patterns jump out immediately?

Data Quality Checks

- Missing values
  - % missing per column
  - Patterns of missingness (spot visually)
- Outliers
  - Context vs. error (deer antler/gender example)
    - Will female deer have entlers?
- Duplicates
  - Exact vs near-duplicate records

Missingness

- 1. **MCAR** – Missing Completely At Random Definition: The probability of a value being missing is unrelated to the data (observed or unobserved).
  - Example: A lab tech accidentally drops a test tube and loses the blood sample → the missingness is random and unrelated to patient characteristics.
  - Consequence: Safe to analyze the remaining data — no systematic bias, though you lose power.
- 2. **MAR** – Missing At Random Definition: Missingness depends on observed data but not the missing value itself.
  - Example: Older participants are less likely to respond to a digital survey → missingness depends on age (observed), but not directly on the unreported values.
  - Consequence: Can be handled if you condition on the related observed variables.
- 3. **MNAR** – Missing Not At Random Definition: Missingness depends on the missing value itself.
  - Example: People with higher incomes are less likely to report their income → the probability of missingness depends on the true (unobserved) value.
  - Consequence: Very tricky — requires domain assumptions or specialized models.

Handling Missingness

- **Drop rows/columns (listwise deletion)**
  - Easy but can waste data
  - Risk of bias if missingness is not MCAR
  - *Example: Drop all rows missing Age → smaller dataset*
- **Simple imputation (mean, median, mode)**
  - Fills gaps with a single summary statistic
  - Can shrink variance, distort distributions
  - *Example: Replace missing Age with median Age*
- **Forward/backward fill (time series)**
  - Carries forward last known value or fills with next value
  - Assumes stability between measurements
  - *Example: Missing stock price on Tuesday filled with Monday's*
- **Model-based imputation**
  - Use regression, k-NN, or ML model to predict missing values
  - More powerful but requires assumptions and computation
  - *Example: Predict missing Age using Income and Education*

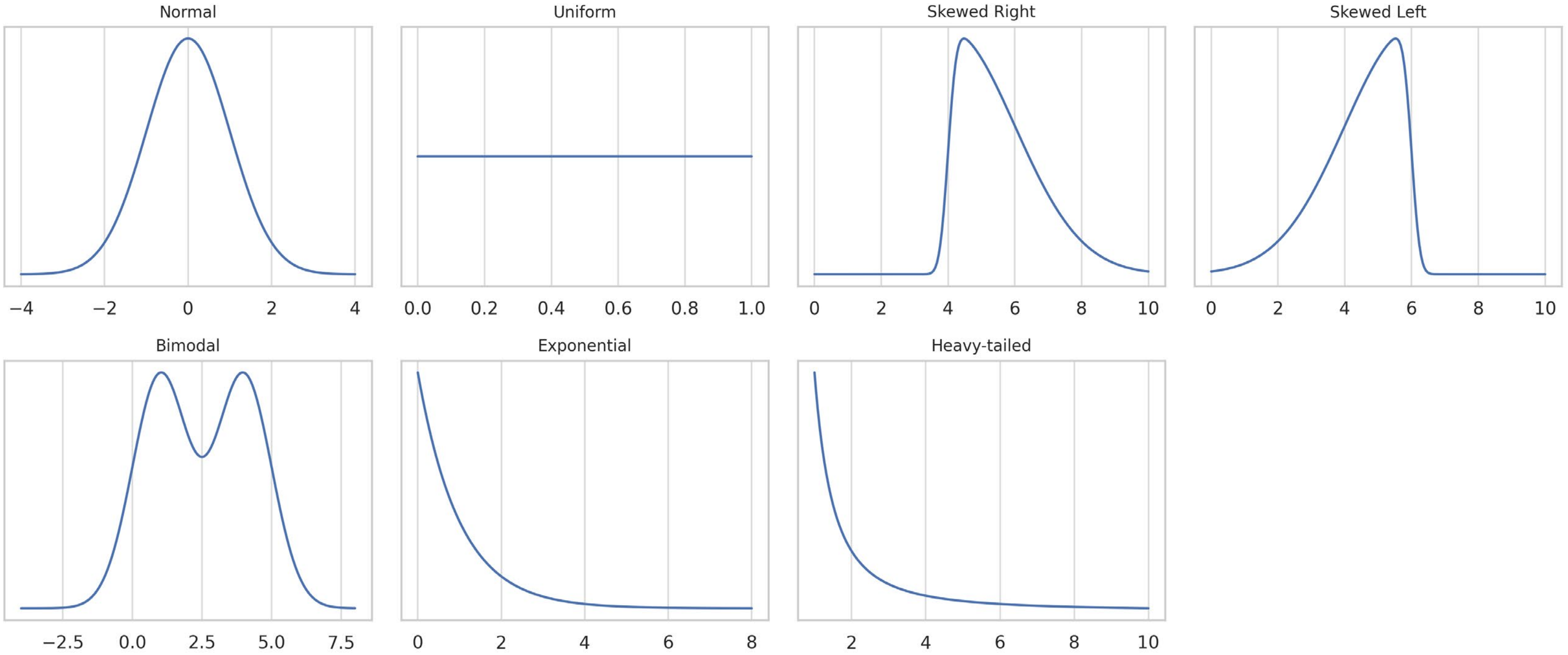
Outliers

- Errors (measurement/data entry)– Typos, sensor glitches, unit mismatches
  - e.g., Height = 300 cm
- Contextual– Unusual only in certain situations
  - e.g., 30°C in winter
- Natural Extremes– Rare but valid tail values
  - e.g., very tall athlete
- Multivariate– Odd combinations of features
  - e.g., Math = 100, English = 5
- Sampling/Processing Artifacts– Wrong population or merge error
  - e.g., dog weights in human dataset

Data Distribution Importance

- The shape of a variable's distribution affects the summaries, statistical tests, and models you can use.
- Always visualize distributions — numbers alone can hide skew, multimodality, or outliers.
- Common shapes: normal, uniform, skewed, bimodal, exponential, heavy-tailed.
- Skewed data may need transformations (log, square root) before modeling.
- Multimodal patterns often indicate distinct subgroups in your data.

Data Structure - Distributions



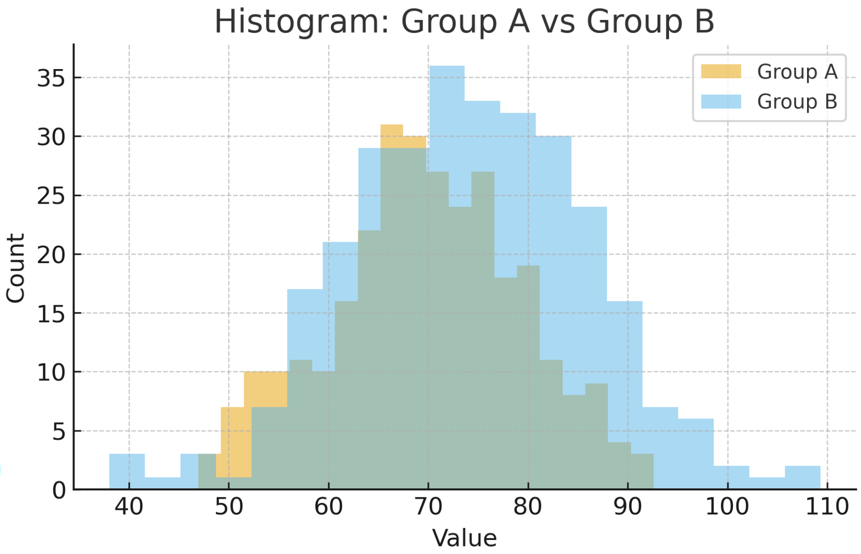
Histogram

- Generally your first stop when visualizing data (can even apply to time series data)
- 1 main parameter: number of bins: more bins → higher resolution

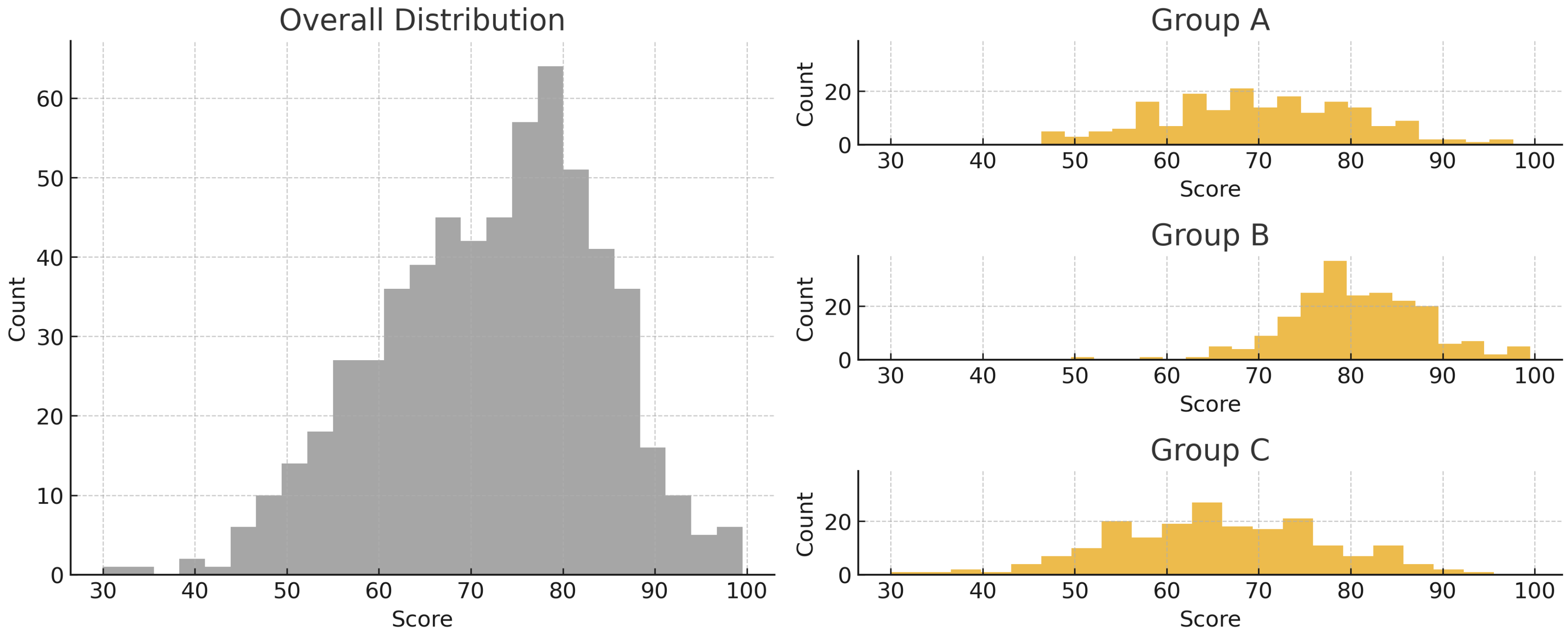
$$\text{bin\_size} = \frac{\max(x) - \min(x)}{\text{number of bins}}$$

$$\text{edges} = \min(x), \min(x) + \text{bin\_size}, \ldots, \max(x)$$

$$\text{center}_i = \frac{\text{edge}_i + \text{edge}_{i+1}}{2}$$



Facets/Small Multiples



Overall Process (Iterative)

- **Practical Steps for EDA**
- **Start with structure**
  - Identify IDs, categorical vs numerical variables
  - Check ordering (time, grouping)
- **Check data quality**
  - Look for duplicates (exact, key, near-duplicates)
  - Summarize missingness (% overall, by subgroup)
  - Identify outliers (errors vs real extremes)
- **Explore distributions**
  - Plot histograms, boxplots, violin plots
  - Compare distributions across groups (facets / small multiples)
  - Watch for skewness, multimodality
- **Investigate relationships**
  - Cross-tabulations, grouped summaries
  - Scatterplots for pairs of numeric variables
  - Split patterns by subgroup (e.g., Group A vs Group B)
- **Iterate & document**
  - Clean obvious errors (e.g., impossible ages)
  - Re-check after cleaning — new issues may emerge
  - Keep notes: *what you saw, what you changed, why*

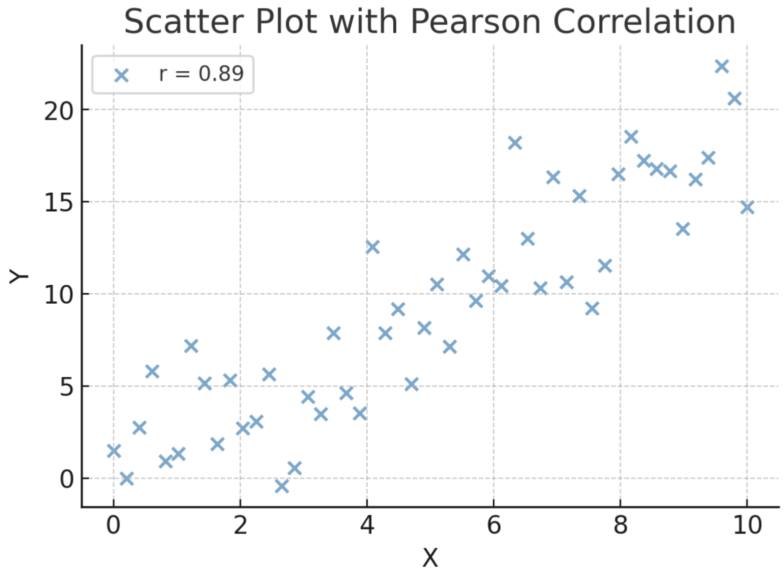
Correlation

- Correlation is a **statistical measure of association** that describes how strongly and in what direction two variables are related.
- Correlation ≠ Causation

Pearson's Correlation Coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$



Spearman’s cont’d.

Spearman's Rank Correlation Calculation (ρ = 0.50)

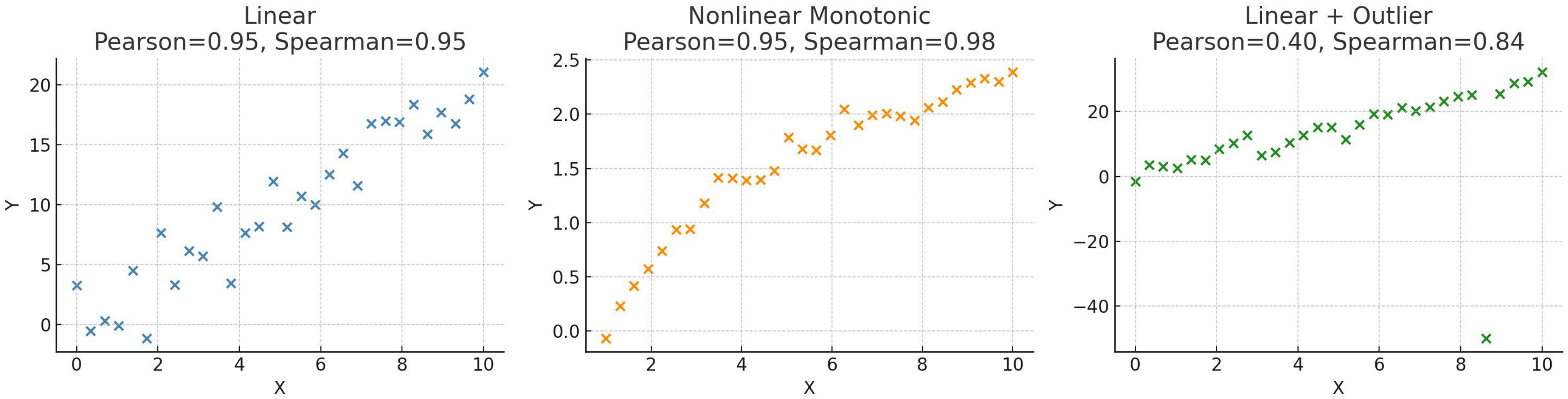
X	Rank X	Y	Rank Y	d = RankX - RankY	d^2
10.0	1.0	15.0	1.0	0.0	0.0
20.0	2.0	40.0	4.0	-2.0	4.0
30.0	3.0	25.0	2.0	1.0	1.0
40.0	4.0	50.0	5.0	-1.0	1.0
50.0	5.0	35.0	3.0	2.0	4.0

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$d_i = \text{rank}(x_i) - \text{rank}(y_i)$$
$$n = \text{number of pairs}$$

$$\rho = \text{Pearson}(\,\text{rank}(X),\, \text{rank}(Y)\,)$$

Spearman vs Pearson



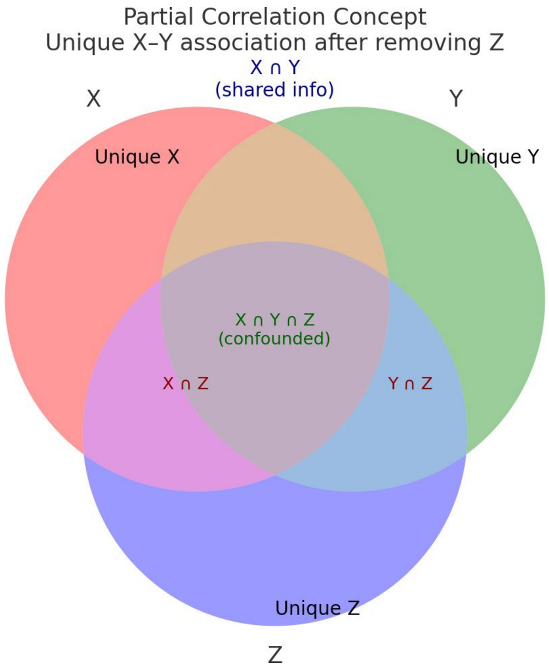
Kendall's Derivation

Obs	X	Y
A	1	12
B	2	15
C	3	14
D	4	10

Pair	Compare X	Compare Y	Result
(A, B)	1 < 2	12 < 15	Concordant
(A, C)	1 < 3	12 < 14	Concordant
(A, D)	1 < 4	12 > 10	Discordant
(B, C)	2 < 3	15 > 14	Discordant
(B, D)	2 < 4	15 > 10	Concordant
(C, D)	3 < 4	14 > 10	Concordant

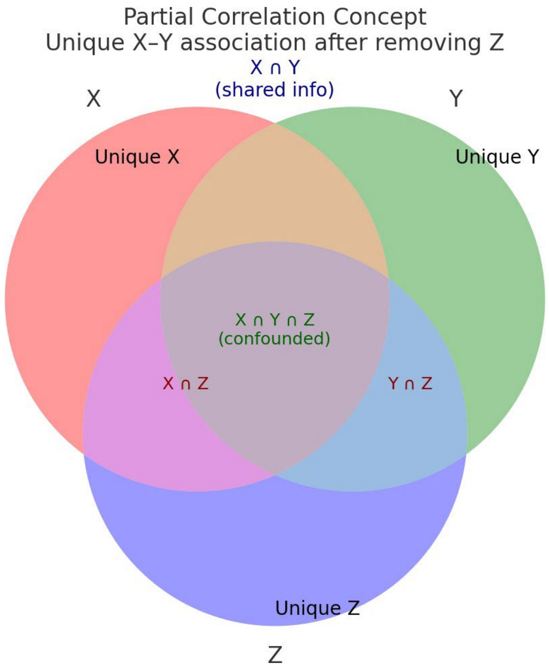
$$\tau = \frac{C - D}{\binom{n}{2}} = \frac{4 - 2}{6} = \frac{2}{6} = 0.333$$

Concept and Formula



$$r_{XY\cdot Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

Concept and Formula



$$r_{XY\cdot Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$