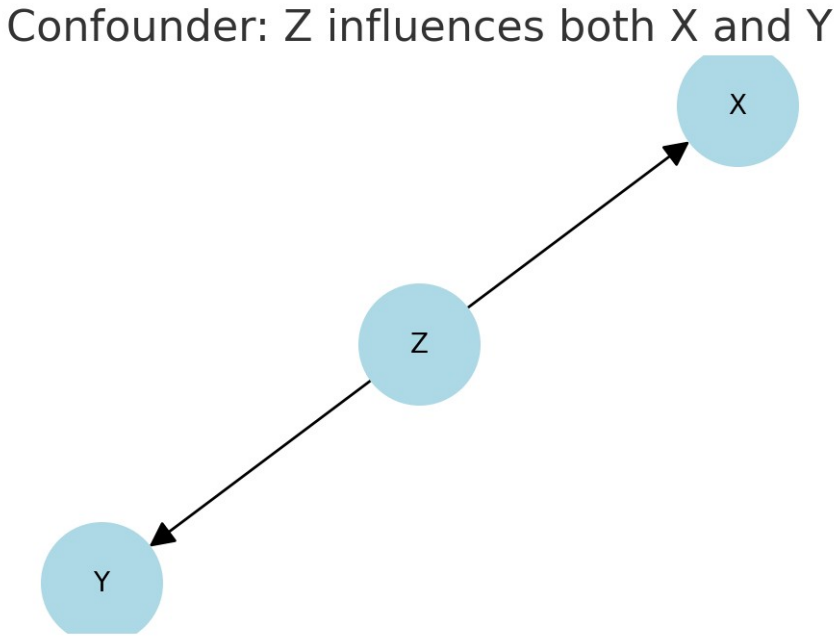
# Exam Review II

# Confounders

- A confounder is a variable that influences both X and Y.
- Creates a spurious association between X and Y.
- Key: If we don’t account for confounders, we may wrongly conclude X causes Y.

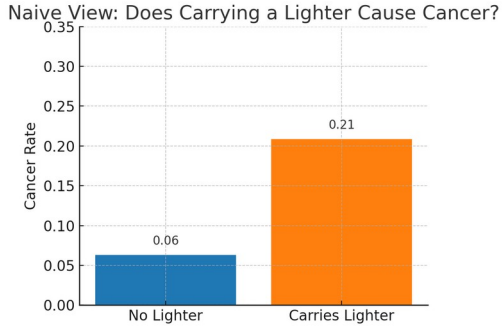


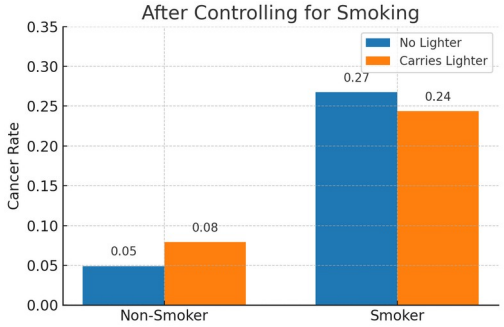
# Confounder Generalizable Logic

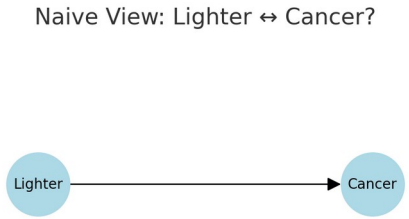
- Confounder must:
  - Be associated with the independent variable (X).
  - Be associated with the dependent variable (Y).
  - Not lie on the causal pathway from X → Y.

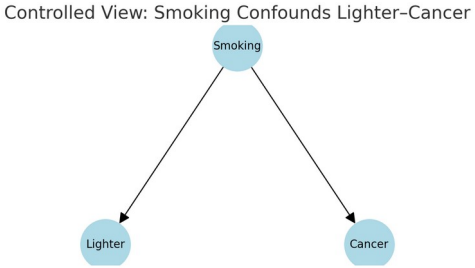
# Confounder Example

- Observed: People who carry lighters are more likely to get lung cancer.
- Hidden confounder: Smoking.
- Smoking → people carry lighters.
- Smoking → higher lung cancer risk.
- So the lighter–cancer correlation is spurious.



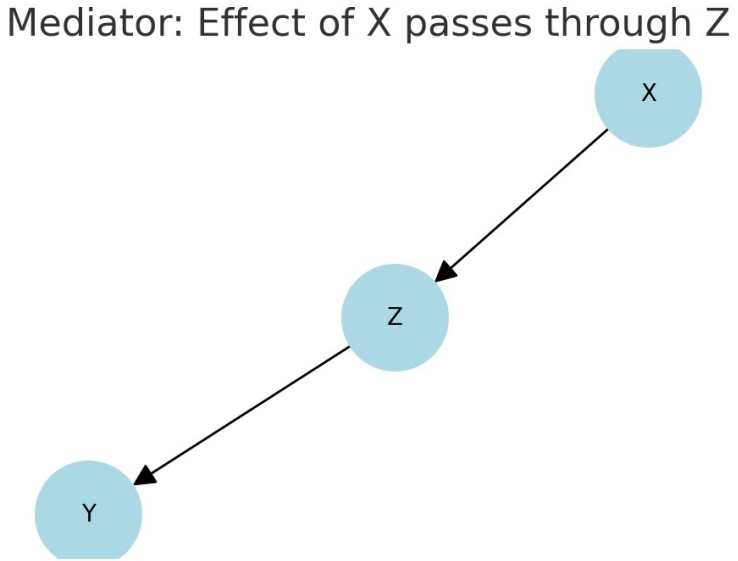



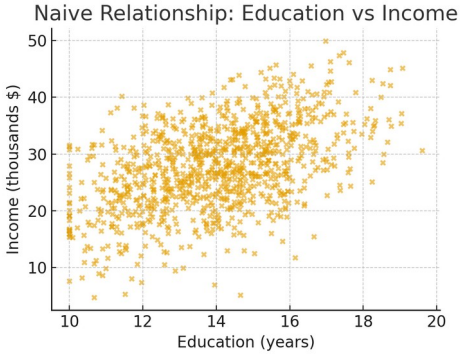
# Mediator

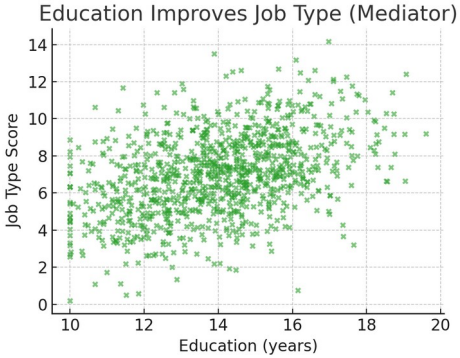
- A mediator explains ***how causality flows*** and usually should **not** be adjusted away if you care about the total effect.
  - Be caused by the independent variable (X).
  - Be associated with the dependent variable (Y).
  - Lie on the causal pathway from X → Y.

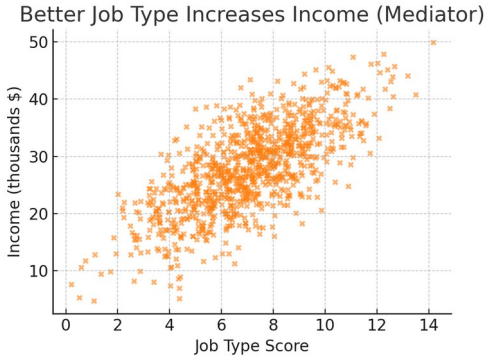


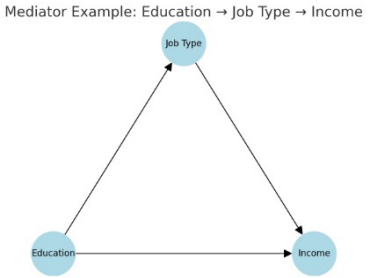
# Mediator Example

- Does higher education lead to higher income — and if so, is the effect direct, or does it work through the kinds of jobs people get?
- Education → better jobs → higher income.
- The effect of education on income is partly explained through job type



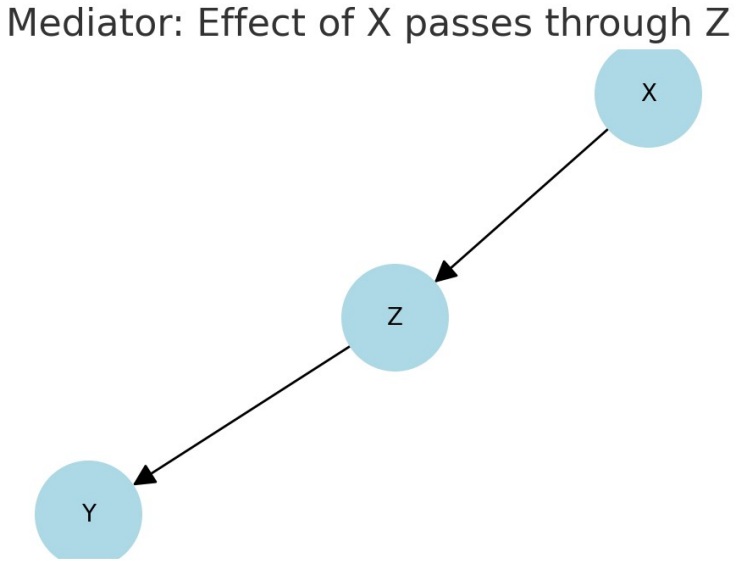



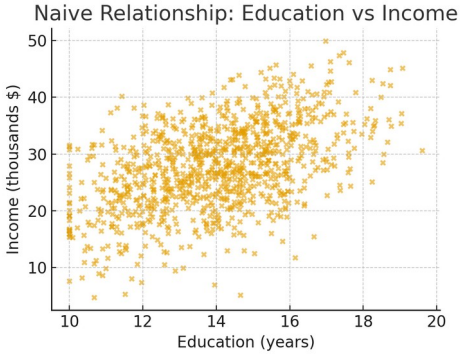
# Mediator

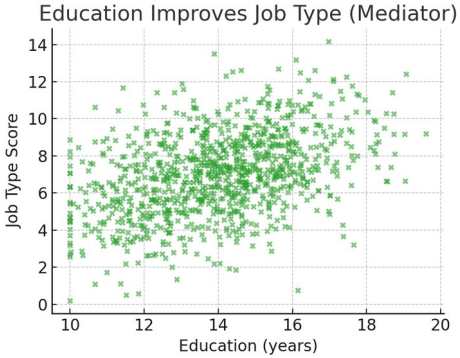
- A mediator explains ***how causality flows*** and usually should **not** be adjusted away if you care about the total effect.
  - Be caused by the independent variable (X).
  - Be associated with the dependent variable (Y).
  - Lie on the causal pathway from X → Y.

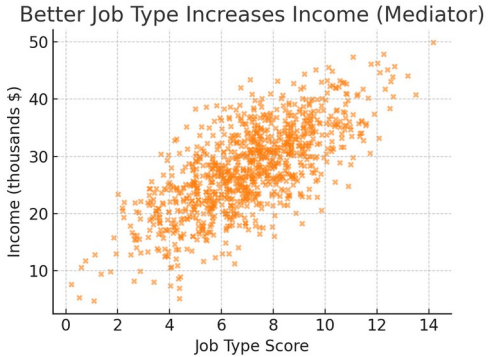


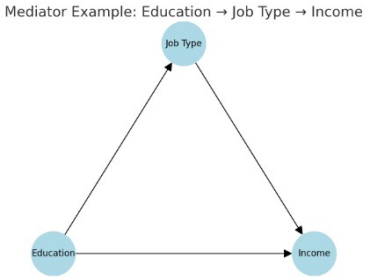
# Mediator Example

- Does higher education lead to higher income — and if so, is the effect direct, or does it work through the kinds of jobs people get?
- Education → better jobs → higher income.
- The effect of education on income is partly explained through job type



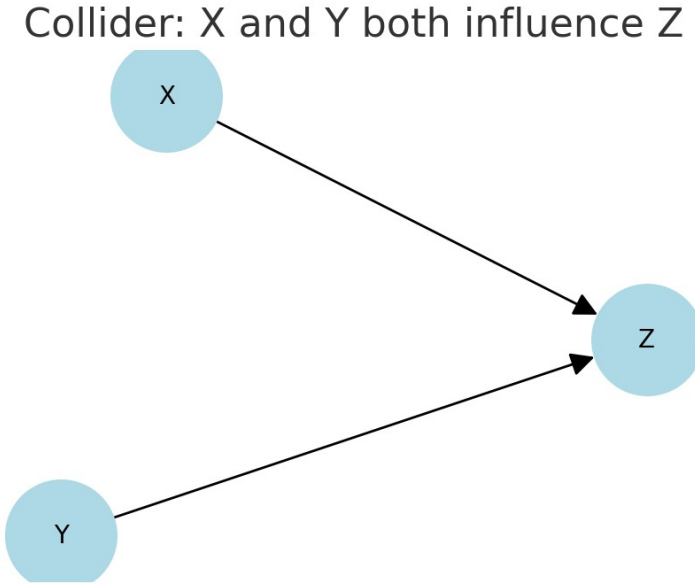



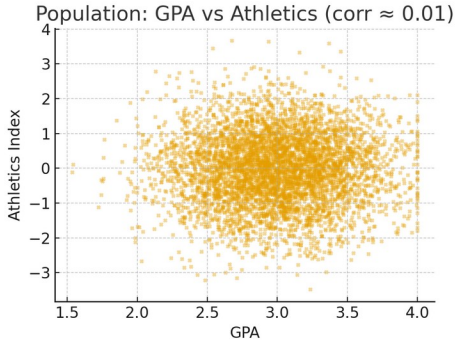
# Collider

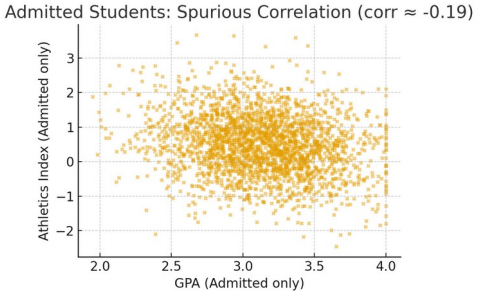
- A collider is a variable that:
  - Is influenced by both X and Y.
  - Conditioning on Z produces an artificial association between X and Y.

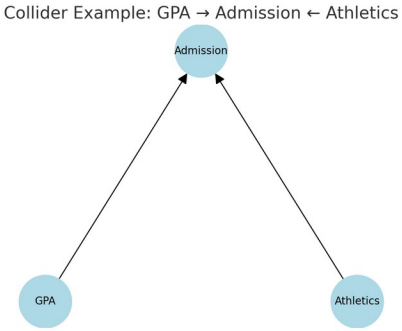


# Collider Example

- Among admitted students, why do we see that higher GPA applicants seem less athletic — is there really a trade-off, or is this a statistical artifact of how admissions decisions are made?
- Why? Admissions depends on both GPA and athletics.
  - –GPA → Admission
  - –Athletics → Admission
- If we condition on “admitted students” (the collider), GPA and athletics appear negatively correlated, even if they aren’t in the full population.
  - –Among admitted students:
    - If a student has a low GPA, they probably got in because of strong athletics.
    - If a student has weak athletics, they probably got in because of a high GPA.
    - This creates an artificial negative correlation between GPA and athletics in the admitted group.



# Summary

## Partial Correlation Outcomes

| Scenario | Naive Correlation | Partial Correlation | Interpretation |
|---|---|---|---|
| **Confounder** (Lighter–Cancer) | ~0.22 | ~0.01 | Effect vanishes → confounder explained the spurious link. |
| **Mediator** (Education–Income) | ~0.50 | ~0.33 | Effect weakens → mediator explains part of the effect. |
| **Collider** (GPA–Athletics) | ~0.01 | ~–0.19 | Effect appears → conditioning on collider creates a false link. |

# Transforming

- What it does: Changes the shape of the distribution.
- Examples: log, square root, reciprocal, Box-Cox, sigmoid.
- Why: Fix skewness, stabilize variance, reveal hidden linearity.
- Effect on correlation: Can change Pearson correlation by altering relationships between variables.
- Critical for some modeling assumptions (mostly statistical)

# Normalizing

- In statistics / ML preprocessing: Often used interchangeably with "scaling," meaning "rescale to a standard range or distribution" (e.g., z-score standardization).
- Changes the **scale** of the data, but preserves the **shape** of the distribution.
- Example: z-score, mean = 0, std = 1.
  - Critical to realize you can normalize/scale over different values (subjects/time)

# Scaling

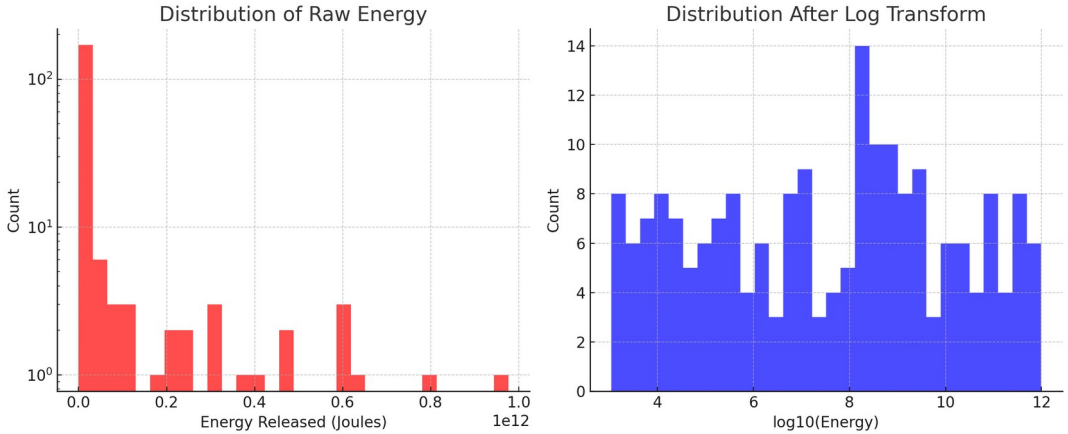
- What it does: Changes the range or unit of a variable, but **not its shape**.
- Examples: min–max scaling to [0,1], dividing by max, unit norm scaling.
- Why: Put different variables on comparable ranges (important for distance-based ML).
- Effect on correlation: Does not change Pearson correlation (linear rescaling leaves r the same).

# Transform vs Technique

- **Transform** when:
  - Relationship is monotonic but nonlinear (log, sqrt help).
  - Skew or outliers distort correlation.
  - Model assumptions (e.g., normality, homoscedasticity) need to be met.
  - You want to reveal hidden linear structure.
- **Change Technique** when:
  - Relationship is inherently nonlinear or non-monotonic (U-shape, quadratic).
  - No reasonable transformation improves correlation.
  - Ordinal/rank information matters more than exact values → switch to Spearman/Kendall.
  - Binary or categorical outcomes → logistic regression, classification, or nonparametric methods.
- Rule of thumb: *If a simple monotonic transformation improves linear correlation, transform. If the pattern stays nonlinear or non-monotonic, switch to a different analytic technique.*
- Key Message: Both are valid — judgment depends on your analytic goal

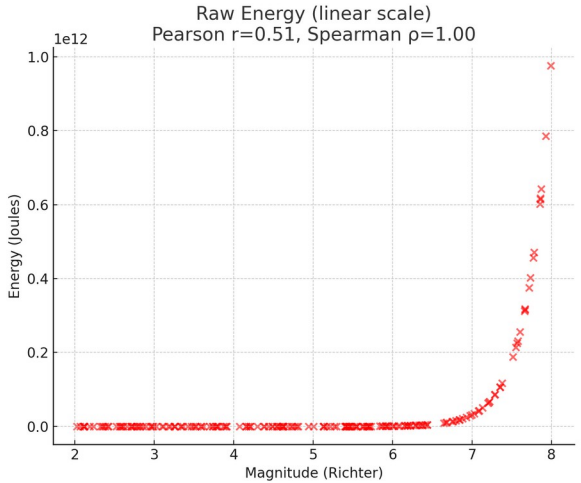
# Log Transform: Visual

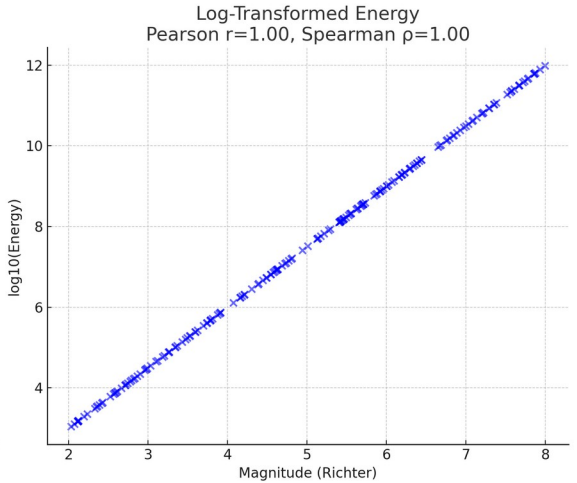
- Suppose you're a seismologist studying earthquakes. The energy released by earthquakes varies enormously: tiny tremors release barely anything, while large quakes release millions of times more energy. If we plot energy on a raw scale, the big ones dominate and the small ones disappear. To compare patterns across all magnitudes, we use a log transform.



# Log Transform Correlation

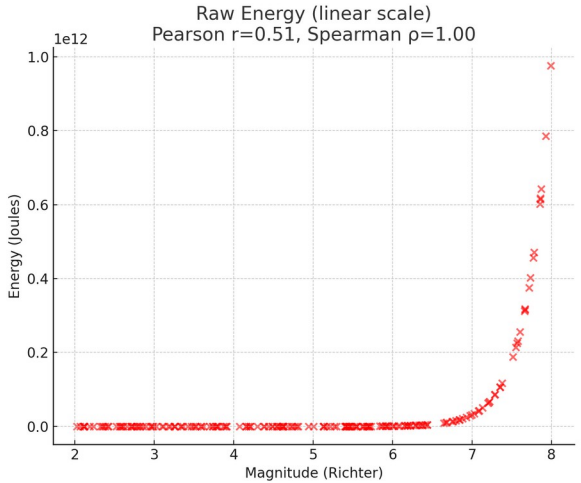
- Raw: Pearson correlation weak or unstable (dominated by outliers)
- After log: Pearson correlation strong, linear, interpretable
- Spearman correlation was always high (monotonic) → log transform brings Pearson in line

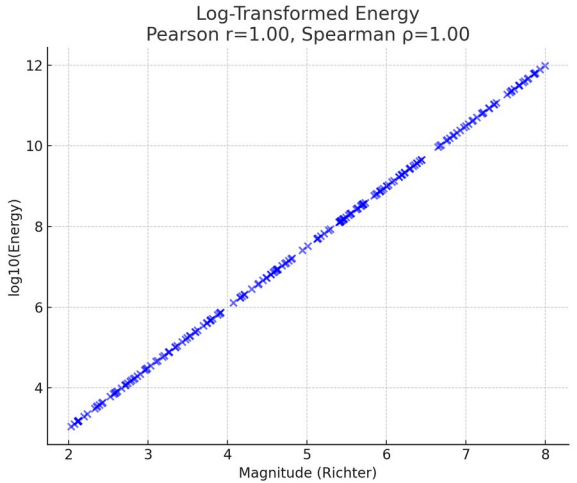


# Log Transform Correlation

- Raw: Pearson correlation weak or unstable (dominated by outliers)
- After log: Pearson correlation strong, linear, interpretable
- Spearman correlation was always high (monotonic) → log transform brings Pearson in line



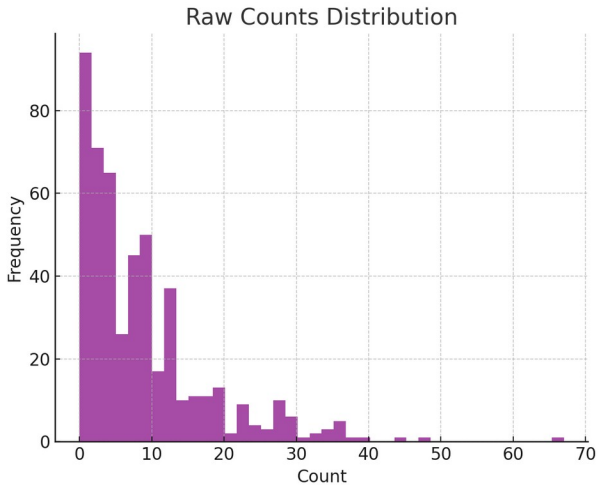



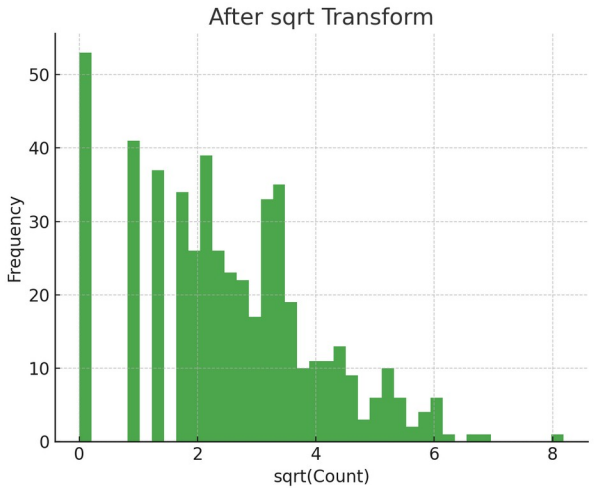
# Square Root Transform: Intro

- Purpose: moderate skew reduction
- Common for counts/frequencies
- Gentle correction for right skew

# Square Root Transform: Visual

- Histogram of variable before and after square root transform
- Distribution or scale changes visibly



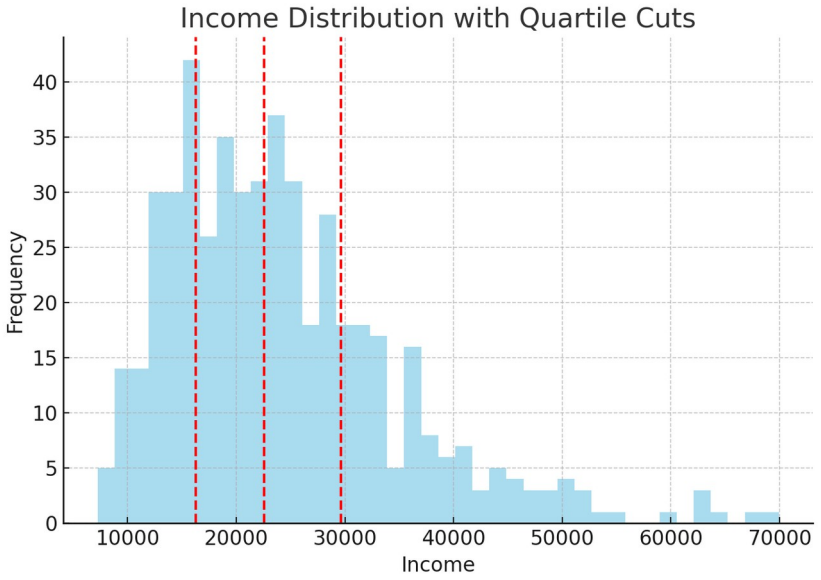



# Thresholding/Quantiling

- Splits continuous data into quantiles or bins
- Useful for skewed data, group comparisons
- Risk: arbitrary cutoffs, information loss
- Continuous ages -> binary: Age > 65 = 1, else 0, or smaller: Ages 0 > 40 = 1, Ages >45 <65 = 2, etc.
- Pearson weakens with thresholding
- Point-biserial/chi square correlation may be more appropriate

# Thresholding/Quantile Example

- Suppose we want to understand how income relates to whether someone defaults on a loan.
- If we use income as a continuous predictor, Pearson correlation might be weak.
- But if we chop the income distribution into quantiles, the relationship becomes clear:
  - default rates are much higher in the lowest group and drop off in the higher groups.



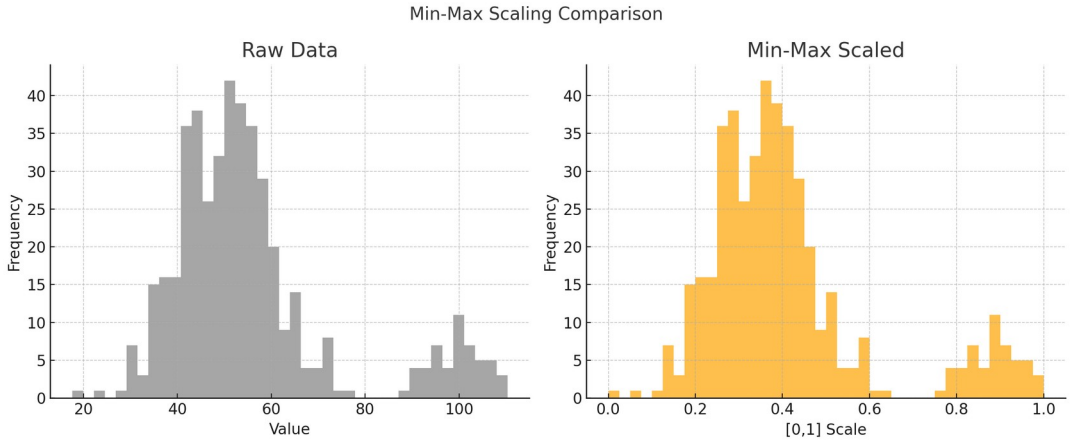
# Transformations

| Transformation | Effect on Data | When to Use |
|---|---|---|
| **Log** | Compresses large values, reduces right skew, stabilizes variance. | Data spanning multiple orders of magnitude (income, population, energy). |
| **Square Root** | Gentle compression, reduces moderate skew. | Count data (Poisson-like), event frequencies. |
| **Reciprocal (1/x)** | Flips and compresses large values, spreads out small values. | Inverse relationships (time → rate, speed, diminishing returns). |
| **Box-Cox** | Optimizes λ to best normalize skew. | Positive data where you want automated transformation. |
| **Yeo-Johnson** | Like Box-Cox, but works with zero/negative values. | Skewed data that may include negatives. |
| **Sigmoid** | Squashes to [0,1], emphasizes mid-range differences. | Probability mapping, logistic regression prep. |
| **Thresholding/Binning** | Converts continuous → categorical. | Interpretability, policy thresholds, noisy predictors. |

Transformations change the distributions of variables, use with caution!

# Min-Max Scaling

- Rescales values to [0,1].
- Preserves shape of distribution, compresses outliers into narrow bands.
- Often used in neural networks and when absolute range matters.
- Histogram of income before and after min-max scaling (same shape, but squashed into [0,1]).

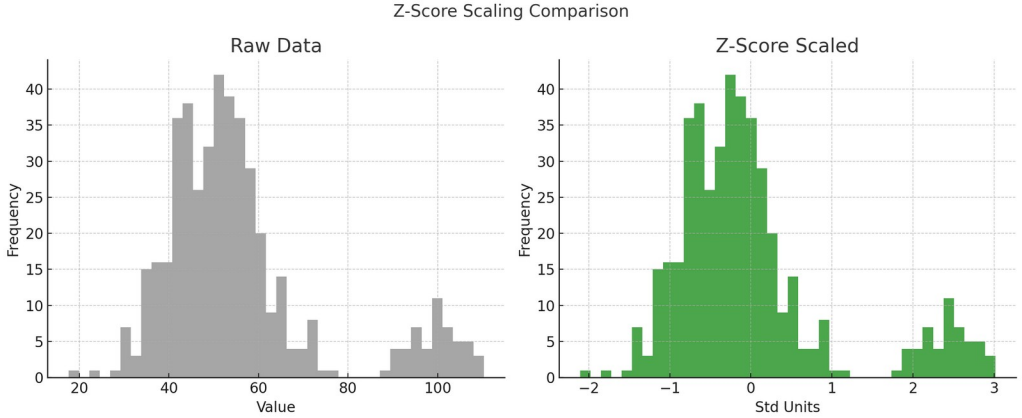



$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

What are the units?

# Z-Score Standardization

- Centers data at mean = 0, scales to standard deviation = 1.
- Distributions keep shape but are re-centered and rescaled.
- Useful for comparability when units differ (e.g., height vs weight).
- Visual: Histogram showing shift to mean 0, spread 1. Overlay normal curve.




$$x' = \frac{x - \mu}{\sigma}$$

What are the units?

# Different Types of Scaling

| Method | Equation | Effect on Data | When to Use |
|---|---|---|---|
| **Min–Max** | $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$ | Rescales to [0,1]; shape preserved. | Neural nets, when bounded inputs are useful; visualization. |
| **Z-Score** | $x' = \frac{x - \mu}{\sigma}$ | Centers mean = 0, std = 1; shape preserved. | When comparing across different units (height vs weight). |
| **Robust Scaling** | $x' = \frac{x - \text{median}}{IQR}$ | Median = 0, spread = 1 (IQR units). | Heavy-tailed or outlier-heavy data. |
| **Group vs Global Scaling** | Apply formulas per-group vs full dataset. | Changes interpretability: global = compare across groups; group-level = compare fairly within groups. | |

# Transforms vs. Scaling

| Aspect | Transformations | Scaling/Normalization |
|---|---|---|
| **Goal** | Change the *shape* of the distribution (reduce skew, linearize relationships, stabilize variance). | Change the *scale* of the variable for comparability. |
| **Examples** | Log, sqrt, reciprocal, Box-Cox, sigmoid, thresholding/binning. | Min–Max, Z-score, Robust scaling. |
| **Effect on Relationships** | Can change correlation values (esp. Pearson), reveal hidden linearity, reduce outlier influence. | Pearson correlation unchanged (linear rescaling); Spearman unchanged. |
| **Units** | May change units (e.g., log-income = log dollars). | Removes or redefines units (e.g., SD units, IQR units, or unitless [0,1]). |
| **When to Use** | Skewed data, heavy tails, nonlinear relationships, inverse effects, categorical thresholds. | Distance-based models (kNN, clustering), gradient descent models (NNs), mixed-unit datasets. |

# Wide vs Long Format

Wide vs Long Format Example

Wide Format

| Student | Math | English | Science |
|---|---|---|---|
| A | 90 | 85 | 95 |
| B | 80 | 88 | 78 |

Long Format

| Student | Subject | Score |
|---|---|---|
| A | Math | 90 |
| A | English | 85 |
| A | Science | 95 |
| B | Math | 80 |
| B | English | 88 |
| B | Science | 78 |

# Joining Basics - Keys

- Joins use keys (IDs).
- Primary key
  - A **unique identifier** for rows in a table
  - Each value appears **only once** in that table.
- Foreign key
  - A column in one table that **refers to a primary key in another table**.
  - Values can repeat (many students can take the same exam)

| StudentID (PK) | Name | GradeLevel |
|---|---|---|
| 101 | Alice | 10 |
| 102 | Bob | 11 |
| 103 | Carol | 10 |

| ExamID | StudentID (FK) | Exam | Score |
|---|---|---|---|
| 1 | 101 | Math | 90 |
| 2 | 102 | Math | 80 |
| 3 | 101 | English | 85 |

# Joining Basics – Types of Joins

- Inner Join
  - Returns only rows with keys present in both tables.
- Left Join
  - Returns all rows from the left table, with matching rows from the right.
  - Missing matches in the right table become NULL.
- Right Join
  - Returns all rows from the right table, with matching rows from the left.
  - Missing matches in the left table become NULL.
- Full Outer Join
  - Returns all rows from both tables.
  - Where no match exists, fills with NULL.
- Cross Join (Cartesian Product)
  - Every row in the left table combines with every row in the right.
  - Rarely used, but useful for generating combinations
- **If keys don't align → missing values or extra rows.**

📘 Table 1: Student Roster

| StudentID | Name | GradeLevel |
|---|---|---|
| 101 | Alice | 10 |
| 102 | Bob | 11 |
| 103 | Carol | 10 |
| 104 | David | 12 |

📘 Table 2: Exam Scores

| StudentID | Exam | Score |
|---|---|---|
| 101 | Math | 90 |
| 102 | Math | 80 |
| 105 | Math | 85 |

# Label Encoding

- Assigns **arbitrary** integers to categories.
  - Example: red=0, blue=1, green=2.
  - Fast and simple.
  - Dangerous if the model interprets the numbers as ordered.
  - Use only when categories are nominal and the model doesn't assume order (tree-based models are okay).

# Ordinal Encoding

- Numbers are assigned to respect a real, meaningful order.
  - Example: shirt sizes Small=1, Medium=2, Large=3.
  - Encodes the rank information.
  - Doesn’t capture distances (Medium isn’t necessarily “twice” Small).
  - Use when categories are truly ordinal.

# One Hot vs Dummy Encoding

One-Hot Encoding (All Columns)
→ Multicollinearity Risk

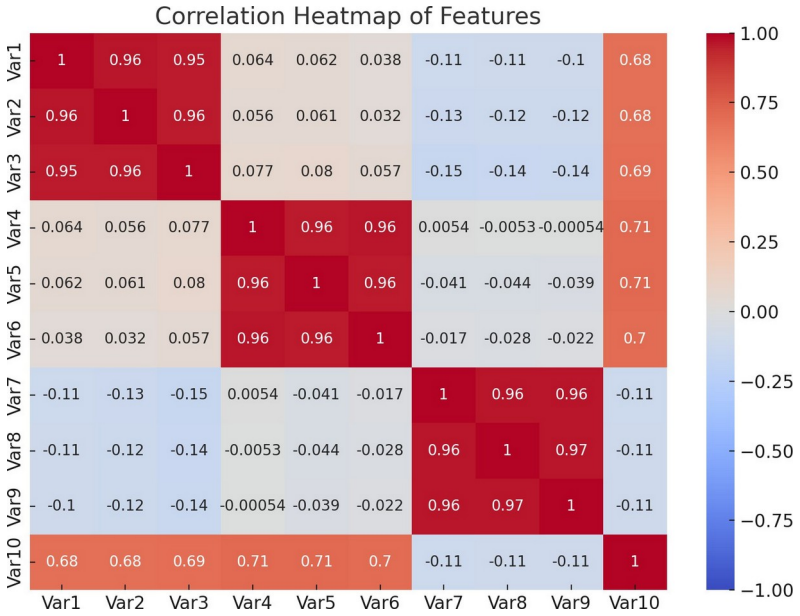
| Student | Color | Red | Blue | Green |
|---|---|---|---|---|
| A | Red | 1 | 0 | 0 |
| B | Blue | 0 | 1 | 0 |
| C | Green | 0 | 0 | 1 |

Dummy Encoding (Drop Red)
→ No Redundancy

| Student | Color | Blue | Green |
|---|---|---|---|
| A | Red | 0 | 0 |
| B | Blue | 1 | 0 |
| C | Green | 0 | 1 |

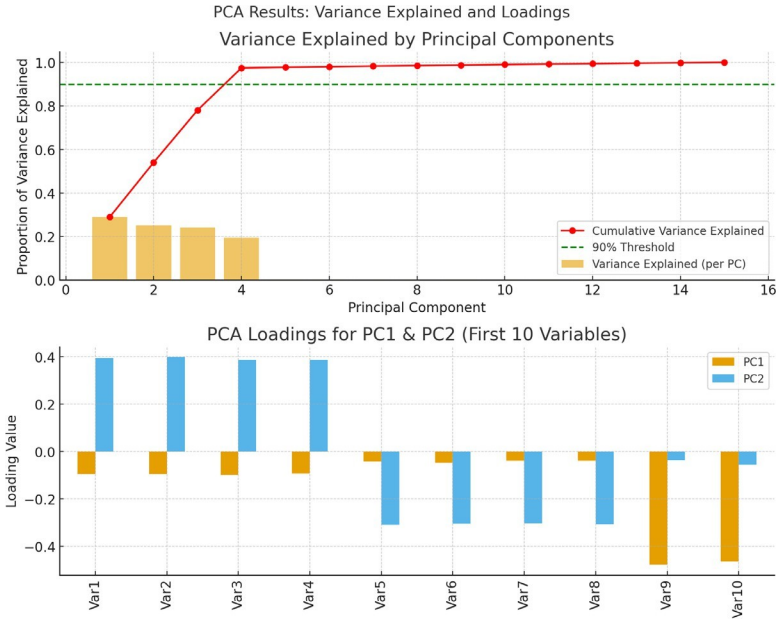
- One-Hot Encoding (Red, Blue, Green all included) → multicollinearity risk because one column can be perfectly predicted from the others.
  - If you know Red and Blue, you can always infer Green.
  - That's perfect multicollinearity.
  - In linear models (regression, logistic regression), this makes the design matrix singular (can't invert), so the model parameters can't be estimated uniquely.
- Dummy Encoding (drop Red as baseline) → removes redundancy, no multicollinearity.

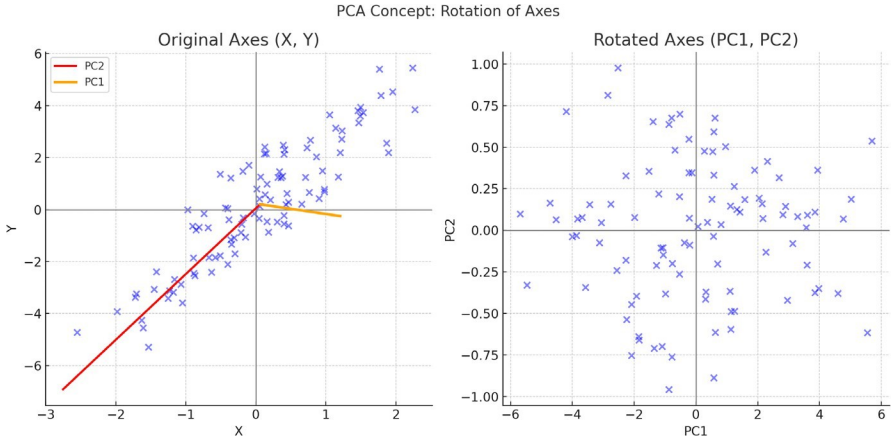
# Toy Example




- A correlation heatmap of 10 synthetic variables.
- Clear clusters of strong correlations (Var1–Var3, Var4–Var6, Var7–Var9), plus a mixed variable (Var10).
- We have redundancy. PCA can simplify this

# PCA Interpretation



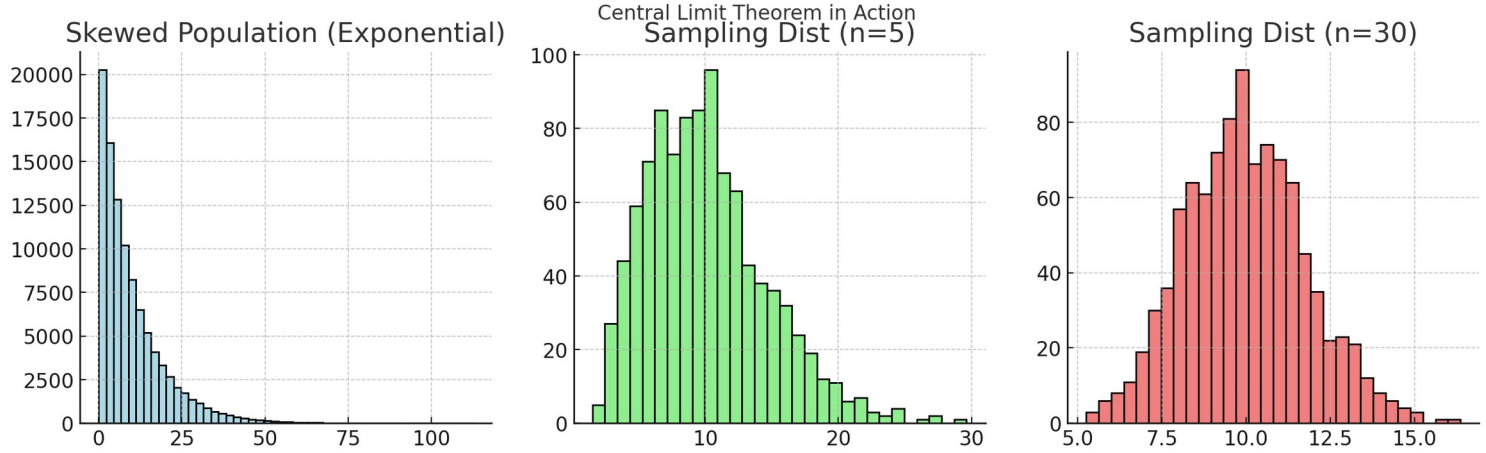

Reduce the feature space from 15 to 4 variables and capture almost 100% of the variance

# Central Limit Theorem

As sample size ↑:

- Sampling distribution gets narrower.
- Shape approaches normal, even if population not normal.

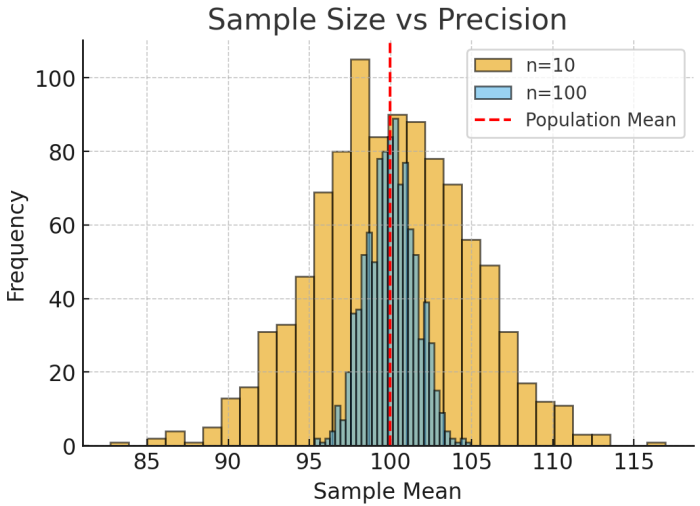
The law that makes inference possible



# Bigger Samples = More Precision

n = 10 → wide variability.

n = 100 → narrower variability.



# What is a Confidence Interval?

A range of plausible values for a population parameter.
Based on sampling variability.
Example: “Average sleep = 6.8 hours (95% CI: 6.3 – 7.3).”
Key message: Don’t trust just a point estimate.

# Example: Go / No-Go Logic for ML

**Example scenario:**

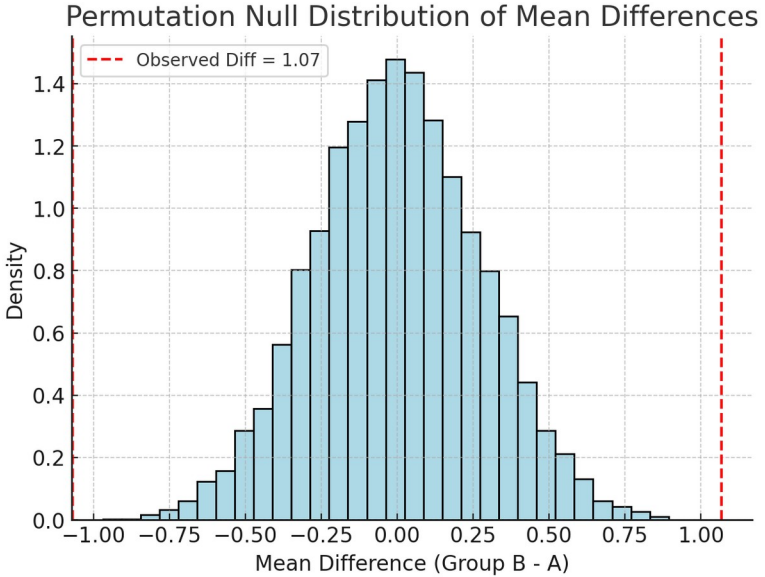
You're evaluating whether to build a churn prediction model.

| Feature | Estimated Effect | 95% CI | Interpretation |
|---|---|---|---|
| Calls to support | +0.25 | (0.10, 0.40) | Stable signal → predictive |
| Region | +0.03 | (–0.15, +0.21) | Unclear → weak |
| Age | –0.01 | (–0.05, +0.03) | No signal |
| Random noise feature | 0.00 | (–0.10, +0.10) | Overfitting risk |

Takeaway: Focus only on variables with **tight, directional CIs**.

# Build the Null Distribution

Collect all shuffled differences → histogram.
Centered around 0 (since under null, no difference).
Mark the observed difference with a vertical line.



# What a p-value Means

p-value = measure of data–model compatibility.
Small p → data are less compatible with the null.
Large p → data are plausible under the null.
**Not** a probability the null is true.
**Not** evidence strength on a graded scale.

# What Is an Effect Size?

Quantifies the magnitude of a relationship or difference.

Independent of sample size.

Types:

- Mean difference
- Correlation (r, $R^2$)
- Standardized difference (Cohen's d
- Odds ratio or risk ratio

$$d = \frac{\bar{X}_2 - \bar{X}_1}{s_p}$$

| d | Magnitude |
|---|---|
| 0.2 | Small |
| 0.5 | Medium |
| 0.8 | Large |

# Summary

Classical stats: Is there an effect?
Data science: Is it big enough to matter, predict, or act on?
Effect size connects the two.

| Concept | Question it answers | DS connection |
|---|---|---|
| p-value | "Is there signal beyond chance?" | Detects presence |
| Effect size | "How strong is the signal?" | Measures predictability |
| CI | "How stable is that estimate?" | Quantifies reliability |

# Example Questions

Which of the following best describes an **ordinal** variable?

**A.** Favorite color

**B.** Temperature in Fahrenheit

**C.** Customer satisfaction: Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied

**D.** ZIP code

You receive a dataset where the income variable is heavily right-skewed.

**a)** What kind of transformation would you apply to normalize the distribution?

**b)** How would you check whether the transformation worked?

Which correlation method is best for detecting a **nonlinear but monotonic** association between two variables?

A. Pearson correlation

B. Spearman correlation

C. Kendall’s tau

D. Point-biserial correlation

You perform a left join between two tables and discover that the number of rows has **doubled**.
**Explain what likely caused this. How would you fix it?**