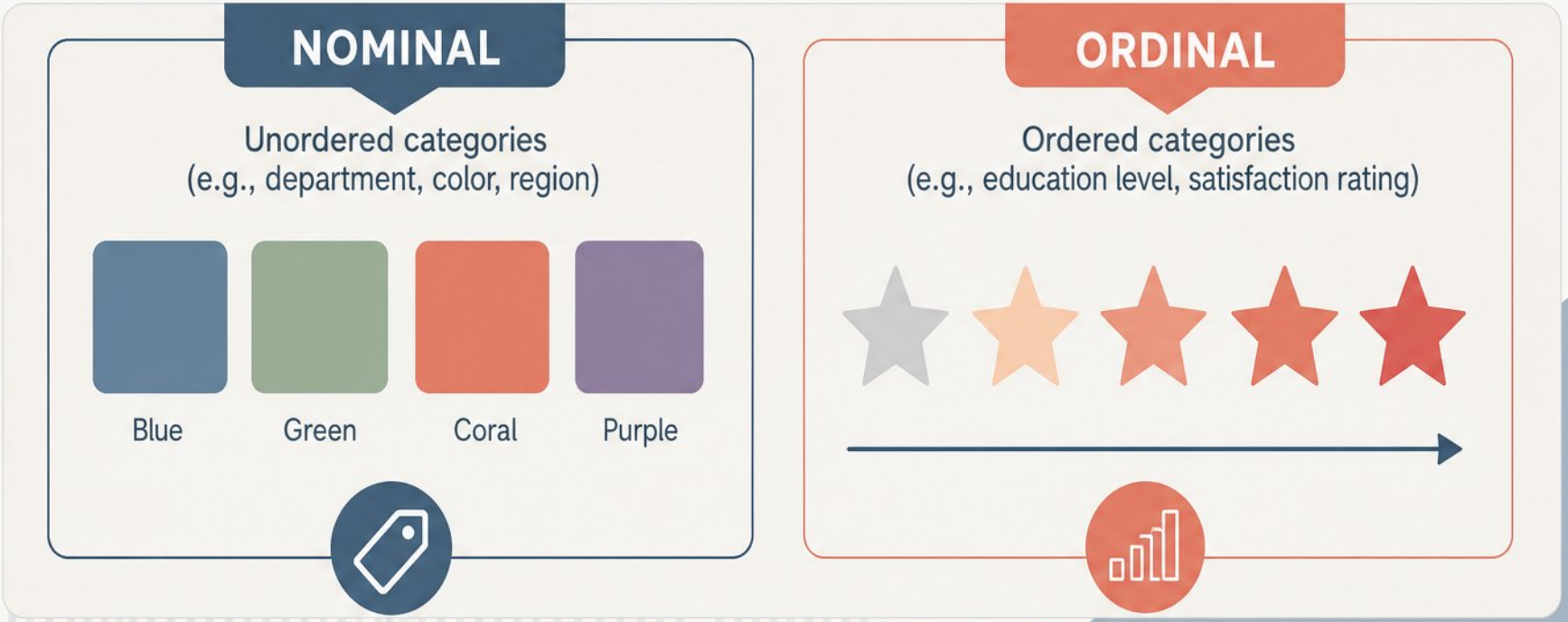
Categorical Data: Counts, Proportions, and Comparisons

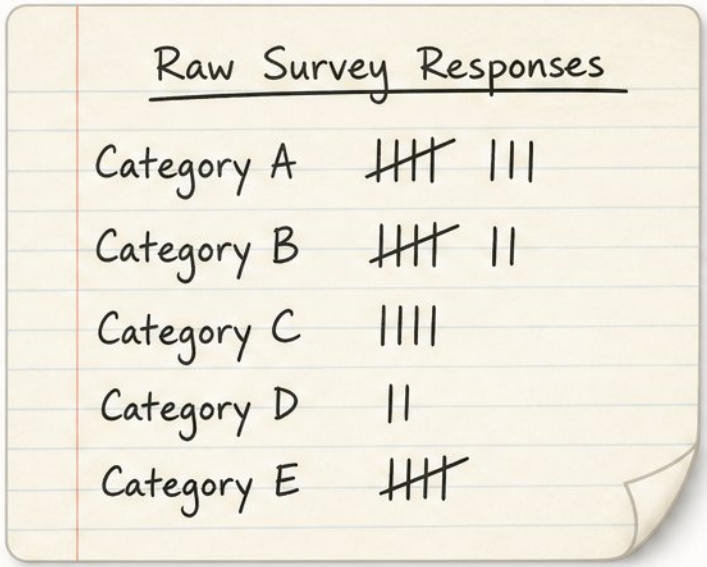
- Data that sorts cases into groups, not measurements
- Essential for segmenting customers, regions, and survey responses
- Core skills: frequency tables, proportions, fair comparisons
- By the end: choose charts and spot misleading claims

What Is Categorical Data?

- Categorical variables describe groups; numerical variables measure quantities
- **Nominal:** unordered categories (e.g., department, color, region)
- **Ordinal:** ordered categories (e.g., education level, satisfaction rating)
- **Common uses:** customer segments, survey responses, product types



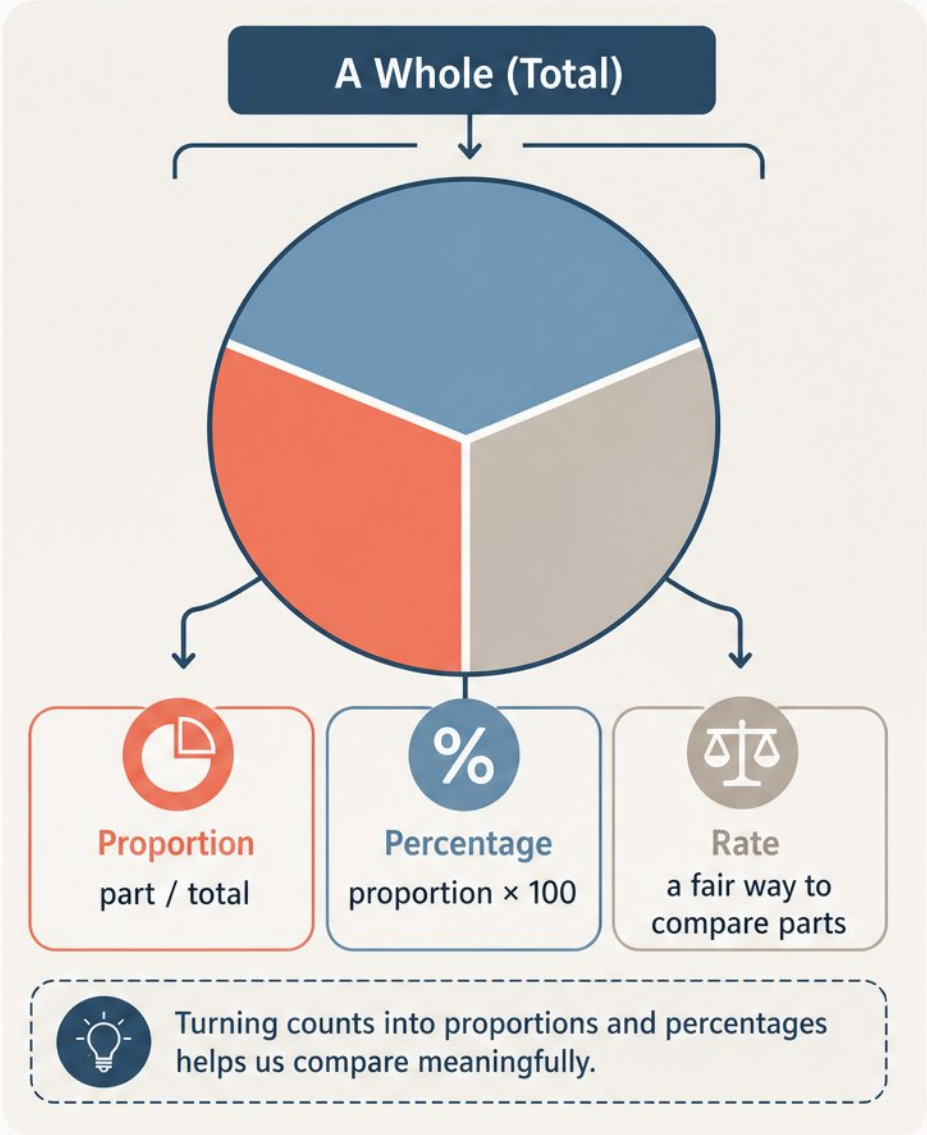
Building Frequency Tables from Raw Data



Frequency Table	
Category	Frequency (Count)

- Define categories and tally each occurrence from raw records
- Absolute counts show how many cases fall into each category
- Walkthrough: convert a list of survey responses into a frequency table
- Raw counts alone mislead when group sizes differ—foreshadowing proportions

From Counts to Proportions and Percentages



- Proportion = part / total; percentage = proportion × 100
- Use proportions to compare groups of different sizes fairly
- Choose percentage when part-to-whole story matters
- Avoid raw count comparisons when group totals differ
- Relative frequency reveals patterns counts alone hide

Making Comparisons Fair with Proportions

- Define comparison: which groups, categories, and time period?
- Use proportions to fairly compare groups of different sizes
- Raw counts mislead when baseline group sizes differ
- “More” in absolute numbers does not mean “more likely”
- Rate = (part / total) × 100

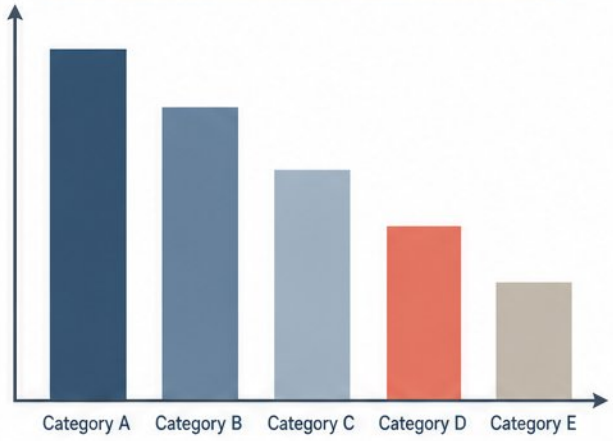
Contingency Tables:

Organizing Two or More Variables

- Cross-tabulation reveals relationships between two categorical variables
- Rows, columns, and cell counts form the basic structure
- Conditional proportions: row %, column %, or total %—pick one
- Choose the proportion that answers your specific comparison question
- Always check if a lurking variable could reverse the conclusion

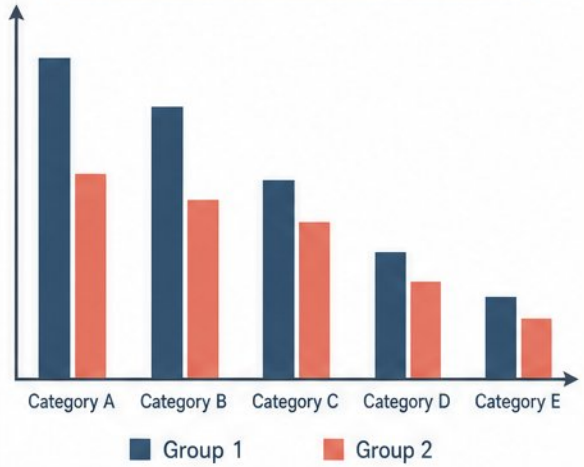
Choosing the Right Chart for Categorical Comparisons

Bar Chart (Good)



Best for comparing categories

Grouped Bar Chart (Good)



Displays two categorical variables side by side

Pie Chart (Not Ideal)

Only for part-to-whole stories with ≤5 slices

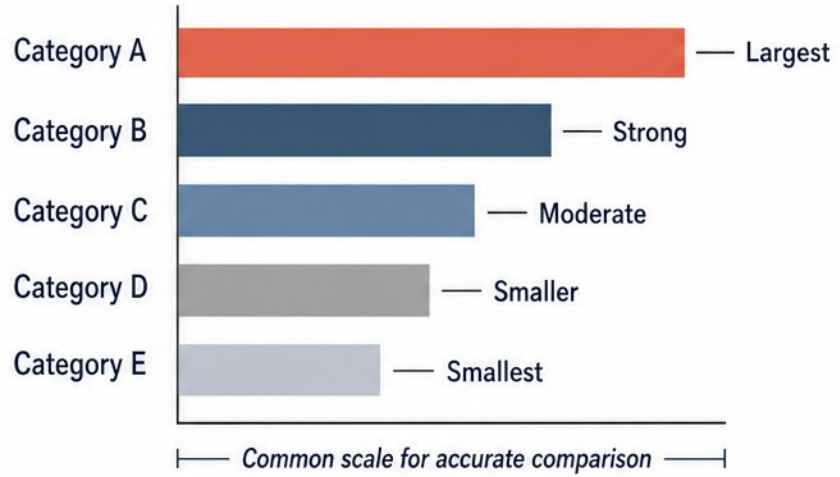
- **Bar charts:** best for comparing categories; use simple, grouped, or horizontal bars per need
- **Grouped/stacked bars:** display two categorical variables side by side
- **Pie charts:** only for part-to-whole stories with ≤5 slices
- **Design rules:** sort bars by value, start y-axis at zero, label clearly

Designing Charts to Communicate Clearly

BEFORE: Confusing

Hard to read. 3D, exploded slices and many colors create confusion and distortion.

AFTER: Clear

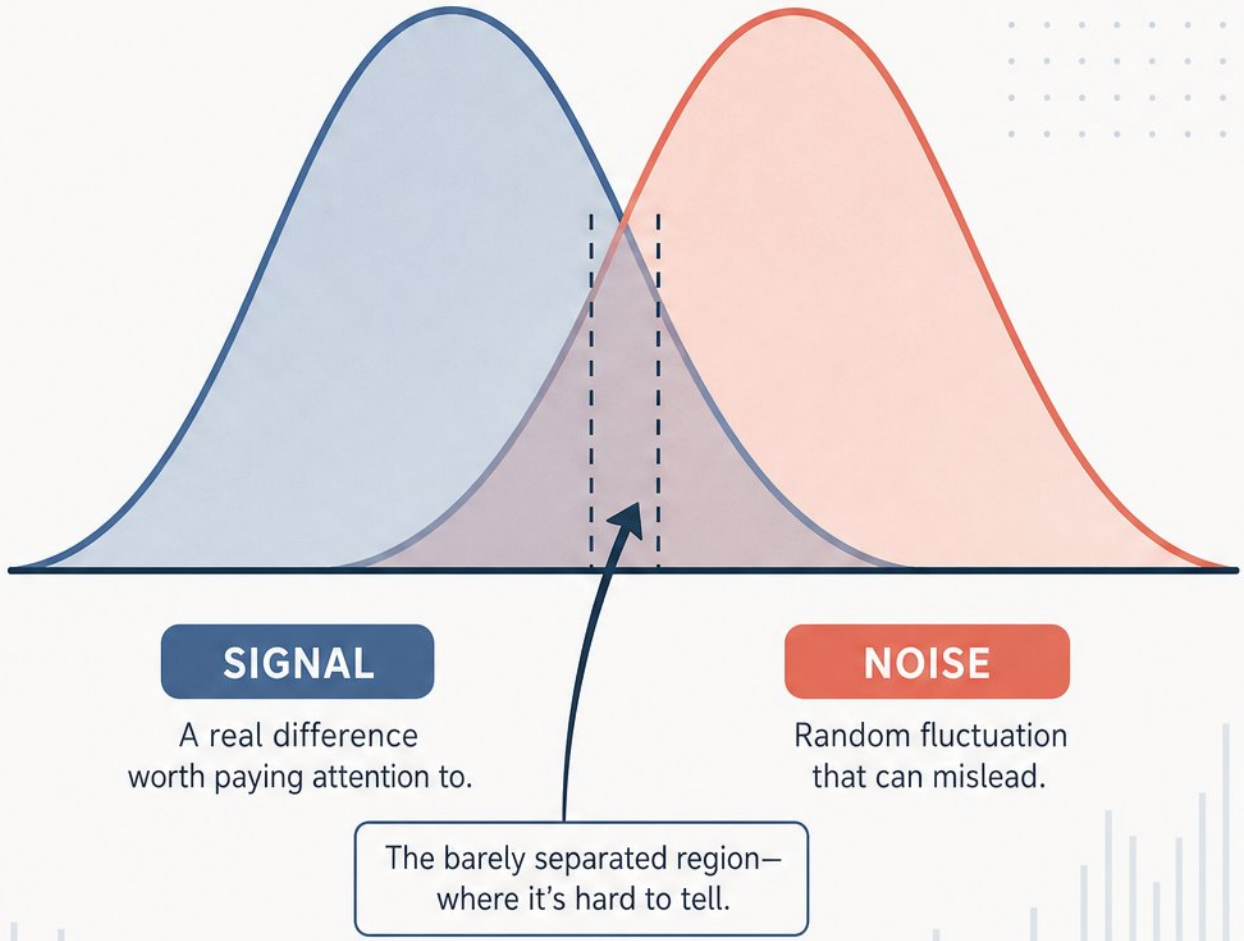


Easy to read. Flat design, limited colors, direct labels, and a common scale improve clarity and honesty.

- Use color purposefully: highlight key categories, avoid decorative palettes that confuse
- Never use 3D effects on bar or pie charts; keep aspect ratios honest
- Annotate directly on the chart when possible rather than relying solely on legends
- Position along a common scale is the most accurate visual encoding for comparisons
- Default to bar charts for flexibility; use pie charts only for ≤5 meaningful part-to-whole slices

Interpreting Differences: Signal or Noise?

- - Distinguish a real pattern from random fluctuation
- - Small samples produce unstable percentages—check the n
- - Practical significance: is the difference large enough to act on?
- - Statistical significance: is the difference unlikely by chance?

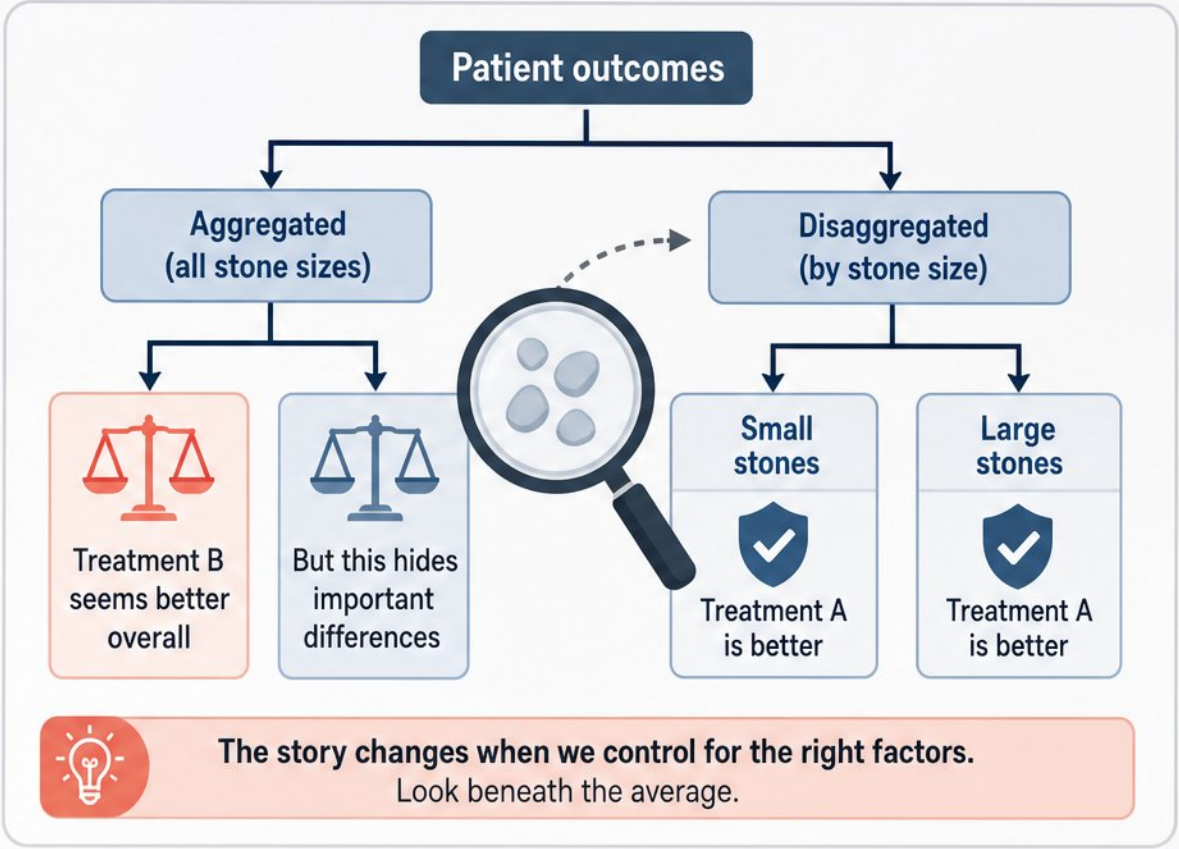


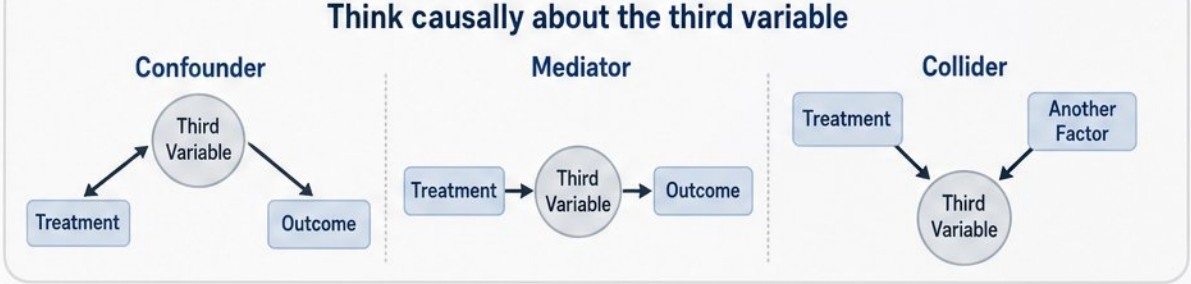
Simpson's Paradox: When Aggregated Data Misleads

- Trend in combined data reverses or vanishes within subgroups
- UC Berkeley case: 44% men vs 35% women admitted overall
- By department, women had equal or higher admission rates
- Women applied more to competitive departments with low admit rates
- Never trust an overall proportion without checking subgroups

Simpson's Paradox: Real-World Cases and Causal Thinking

- **Kidney stones:** Treatment B seemed better overall, but A was better for both small and large stones
- **COVID-19:** More deaths among vaccinated? Adjusting for age and risk reversed the story
- **Aggregation hides confounders** like stone size or age; controlling for them changes conclusions
- **Causal diagrams:** Is a third variable a confounder, mediator, or collider?





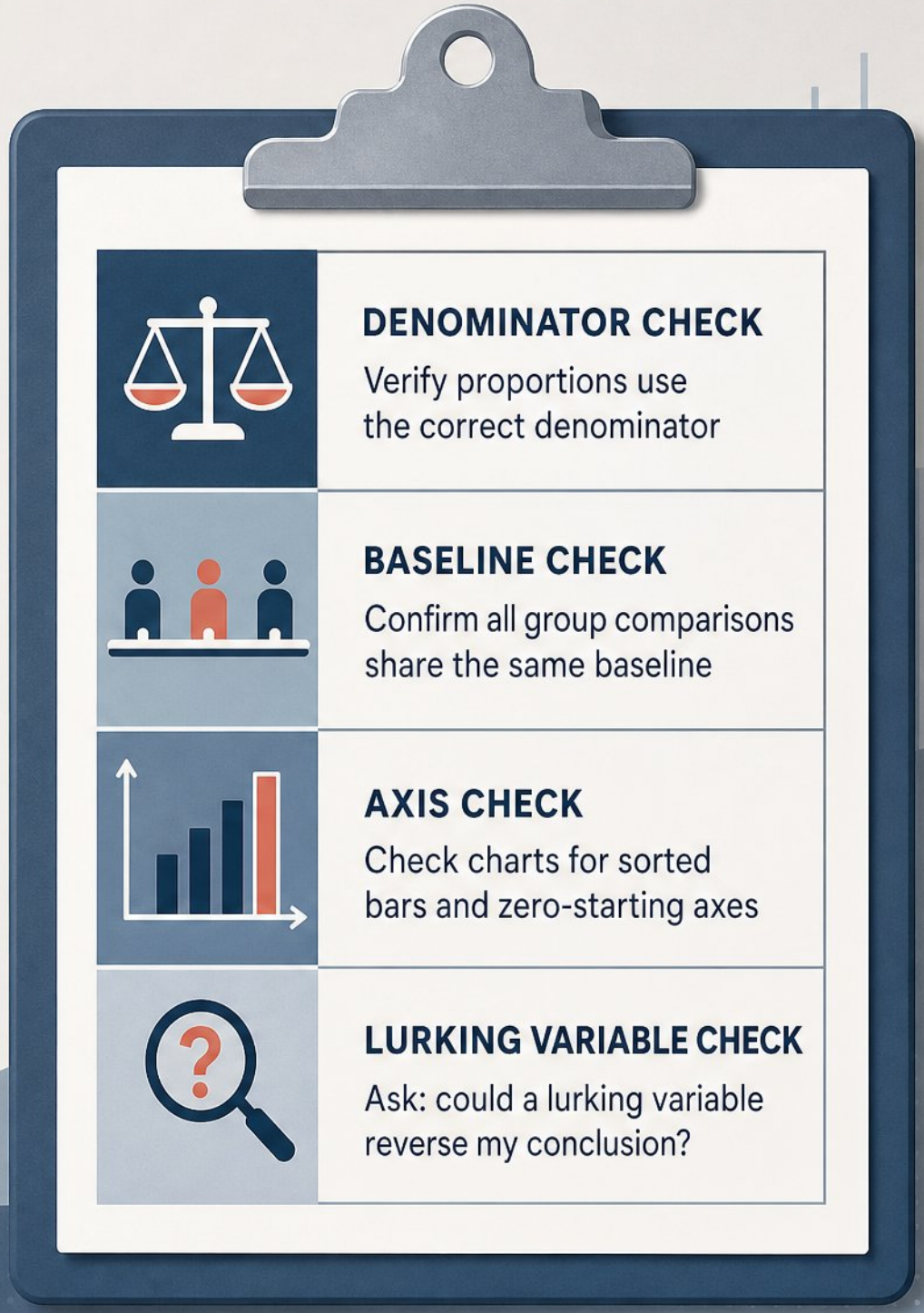
Avoiding Common Errors in Categorical Comparisons

- ✗ Ask key questions before claiming a categorical difference
- ✗ Never compare raw counts when group totals differ
- ✗ One number can hide opposite subgroup patterns
- ✗ Pie charts fail for precise comparisons; use bars
- ✗ Small samples produce unreliable percentages

Building Your Comparison Workflow

Peer Review Checklist and Quality Assurance

- Verify proportions use the correct denominator
- Confirm all group comparisons share the same baseline
- Check charts for sorted bars and zero-starting axes
- Avoid 3D effects and distorted aspect ratios
- Ask: could a lurking variable reverse my conclusion?

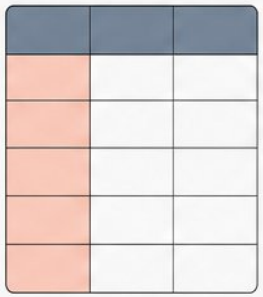


Practice Cases and Self-Assessment

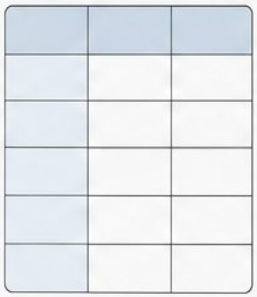
- Two to three scenarios: choose counts vs. proportions, then build tables and charts
- Spot and correct misleading pre-built tables and graphs in a short exercise
- Self-assessment rubric mapped directly to each training objective
- Key takeaways summary and links to further practice resources

Wrong vs. Corrected Table

Misleading Table



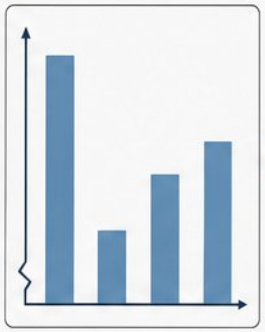
Corrected Table



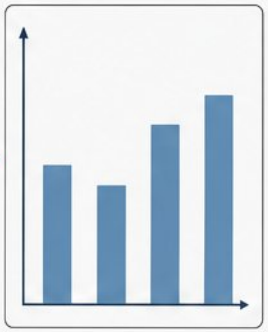
Look for inconsistent formatting, missing context, and choices that overemphasize certain categories.

Misleading Chart vs. Proper Chart

Misleading Chart



Proper Chart



Check axes, scales, and chart types. Ensure the visual tells the truth about the data.

Rubric Scorecard

Training Objective	Self-Assessment		

Rate your confidence for each objective and focus on areas for improvement.

PersonWise.

PersonWise • AI digital-human course platform

Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/efd186b1/public