Statistical Formulas Reference

These tools help you find patterns and relationships in data without needing a math background. This handout shows the formulas each tool uses and explains what they mean in everyday language.

Tool 1 - Correlate (numeric data)

What it does

This tool takes columns of numbers, like test scores, word counts, or survey ratings, and measures how strongly two columns move together. If one goes up, does the other tend to go up too? Or down? The result is a single number between -1 and +1 called a **correlation coefficient**.

The Two Formulas Used

① Pearson r (for normally distributed numbers)

$$r = \text{sum}[(x_i - x_mean)(y_i - y_mean)] / \text{sqrt}[\text{sum}(x_i - x_mean)^2 * \text{sum}(y_i - y_mean)^2]$$

In plain English: subtract each value from its column average, multiply the pairs together, then divide by how spread-out each column is. The result tells you how tightly the two columns track each other.

② Spearman rho (for ranked or non-normal data)

```
rho = Pearson r applied to the rank-orders of x and y
```

Instead of using the raw numbers, the tool first converts each value to its rank within its column: the lowest value becomes 1, the next lowest becomes 2, and so on. Pearson r is then calculated on those ranks rather than the original numbers. This matters because Pearson r assumes your data is normally distributed and that the relationship between the two variables is linear. Spearman rho only requires that the relationship is monotonic (consistently going in one direction), making it more appropriate for ordinal scales like Likert responses (Strongly Agree to Strongly Disagree), ranked lists, or any data with outliers that would distort a straight-line fit.

Reading the result

Value	Meaning
+1.0	Perfect positive relationship (both always move up together)
+0.5	Moderate positive relationship
0.0	No relationship
-0.5	Moderate negative relationship
-1.0	Perfect negative relationship (one up = the other down)

Tool 2 - Survey Correlate (categorical / multiple-choice data)

What it does

When your data is in categories, like 'Agree / Neutral / Disagree' or 'Male / Female / Non-binary', you cannot use Pearson r. This tool instead uses **Chi-squared (chi^2)** to test whether two survey questions are related, and **Cramer's V** to measure how strong that relationship is.

The Two Formulas Used

① Chi-squared test (chi^2)

```
chi^2 = sum (Observed - Expected)^2 / Expected
```

For every cell in a cross-tab table, the tool asks: how many responses did we actually see, and how many would we expect if the two questions were completely unrelated? A large chi^2 means the categories are not independent: they co-occur more (or less) than chance.

② Cramer's V (effect size)

```
V = sqrt( chi^2 / (n * (k - 1)) )
```

chi^2 grows with sample size, making it hard to compare across studies. Cramer's V fixes this by dividing by the number of responses (n) and the number of categories (k). Like r, V ranges from 0 (no association) to 1 (perfect association).

Quick interpretation guide

Cramer's V	Association Strength
0.00 - 0.05	Negligible
0.05 - 0.10	Small
0.10 - 0.20	Medium
0.20 - 0.30	Large
0.30+	Very large

Tool 3a - Theme Correlate: Phi Coefficient (open-ended text)

What it does

This tool finds themes in free-text responses and measures how strongly any two themes co-occur using the **phi coefficient**, a correlation measure for binary data (present/absent). It also reports whether the co-occurrence is statistically significant using chi-squared, and uses Jaccard similarity as a supporting measure. Unlike the confidence-based tool below, phi is symmetric: it treats both themes equally and can detect negative associations (themes that tend to avoid each other).

The Formulas Used

① TF-IDF (identifies distinctive words per response)

```
TF-IDF(word, response) = TF(word) × IDF(word) where TF = (times word appears in this response) / (total words in response) and IDF = log( total responses / responses containing word )
```

TF rewards words that appear often in a specific response. IDF penalizes words that appear in almost every response (like 'the' or 'very'). Together, TF-IDF surfaces the signature words of each theme.

② Jaccard Similarity (overlap between two themes)

```
Jaccard(A, B) = |A and B| / |A or B| = (responses with both) / (responses with either)
```

Jaccard measures what fraction of the responses that mention either theme also mention both. It ranges from 0 (no overlap) to 1 (identical sets of responses). It is used here as a simple overlap check alongside the more powerful phi coefficient.

③ Phi Coefficient (correlation for binary co-occurrence)

```
phi = (a*d - b*c) / sqrt[ (a+b)(c+d)(a+c)(b+d) ] where: a = both themes present b = only Theme A present c = only Theme B present d = neither theme present
```

Phi is essentially Pearson r applied to two yes/no variables. It ranges from -1 to +1. A positive phi means the themes tend to appear together; a negative phi means they tend to avoid each other (mentioning one makes the other less likely). The tool flags pairs with phi above 0.10 as weakly associated, above 0.20 as moderate, and above 0.35 as strong.

④ Chi-squared significance test

$$chi^2 = n * phi^2$$

This tests whether the co-occurrence could have happened by chance. A low p-value (typically below 0.05) means the association is unlikely to be random. Phi tells you the strength; chi-squared tells you whether to trust it.

Tool 3b - Theme Correlate: Association Rule Confidence (open-ended text)

What it does

This tool also finds themes in free-text responses, but measures relationships using **directional confidence** from data mining rather than a symmetric correlation. Instead of asking 'do these themes co-occur?', it asks 'when Theme A appears, how often does Theme B also appear?' The direction matters: A predicting B is a different (and often stronger) claim than B predicting A. This makes it useful for finding sub-themes: topics that almost always appear inside a larger topic but not vice versa.

The Formulas Used

① TF-IDF and Global Importance Score (same as Tool 3a, plus theme ranking)

```
Importance = mean TF-IDF × sqrt(responses containing the word)
```

The square root factor rewards words that are spread broadly across many responses, not just distinctive within one. This helps surface themes that are both characteristic and common rather than narrow and rare.

② Directional Confidence

```
Confidence(A → B) = P(A and B) / P(A) = (responses with both A and B) / (responses with A)
```

This asks: of all responses that mention Theme A, what fraction also mention Theme B? A confidence of 0.80 means that in 80% of responses where A appears, B appears too. Crucially, Confidence(A→B) and Confidence(B→A) are usually different numbers. If A→B is high but B→A is low, then A is almost always accompanied by B, but B often appears without A: B is the broader topic, A is the sub-theme.

Example

Suppose students write about their research process. Theme A = 'finding sources' appears in 40 responses. Theme B = 'library databases' appears in 35 of those same 40 responses. Confidence(A→B) = 35/40 = **87.5%**. But suppose 'library databases' appears in 50 responses total, so Confidence(B→A) = 35/50 = **70%**. Finding sources almost always involves library databases, but library databases come up in other contexts too: 'finding sources' is the sub-theme, 'library databases' is the broader one.

All three tools compute statistics locally in your browser: no data is sent to any server. No math background required to use them; this sheet is provided for transparency and reference.