PO91Q: Fundamentals in Quantitative Research Methods

Worksheet Week 7 - Solutions

Dr Florian Linke
Florian.Linke@warwick.ac.uk

1 | Crosstabulations – Calculations by Hand

Expected Values:

	Mode of Transport			
Department	Bike	Bus	Car	Total
PAIS	7.2	6.8	6	20
CIM	10.8	10.2	9	30
Total	18	17	15	50

1. 2.8867
2. 2
3. 0.236, or between 0.90 and 0.10
4. Yes
5. Small effect, big N
    a. For a=210, b=190, c=190, d=210, row 1 "Yes" = 210/(210+190) = 0.525 (52.5%), row 2 "Yes" = 190/(190+210) = 0.475 (47.5%). The gap is about 5 percentage points. The scaled table a=21, b=19, c=19, d=21 has the same row percentages and the same 5-point gap.
    b. With N=800 (210/190/190/210) the app gives roughly $\chi^2 \approx 2.00$, $p \approx 0.16$. With N=80 (21/19/19/21) $\chi^2 \approx 0.20$, $p \approx 0.65$. The effect is the same, but χ^2 is much larger (and p much smaller) at bigger N.
    c. Holding the row-percentage gap fixed keeps the underlying association about the same. As N increases, sampling noise shrinks, so the test statistic χ^2 tends to get larger and the p-value gets smaller. A useful way to think about this is with Cohen's *w*, a standardized effect size for contingency tables that reflects how far the observed cell proportions deviate from what independence would expect. When you scale the table up or down but keep its shape the same, *w* stays about the same, and the rough relationship $\chi^2 \approx N \cdot w^2$ shows why the p-value decreases as N grows: the effect size is unchanged, but you're measuring it more precisely.

6. Same row-percentage gap, different base rates
    a. Example A (balanced): a=60, b=40, c=40, d=60. Row 1 “Yes” = 60%, row 2 “Yes” = 40% (gap 20 points). Column totals are balanced (100/100). Example B (skewed): a=90, b=10, c=70, d=30. Row 1 “Yes” = 90%, row 2 “Yes” = 70% (gap 20 points). Column totals are skewed (160/40).
    b. The app reports $\chi^2 \approx 8.0$ ($p \approx 0.0047$) for Example A and $\chi^2 \approx 12.5$ ($p \approx 0.0004$) for Example B. Even with the same row-percentage gap and the same N, skewed margins change the expected counts, making some $(O-E)^2/E$ terms larger and increasing χ^2. The “Row %” view shows the gap is similar across tables, while the “Column %” view reveals the base-rate imbalance that helps explain the χ^2 difference.
7. Exploring Fisher’s exact test
    a. When at least one expected cell count falls below 5, the app begins to display Fisher’s exact test alongside the χ^2 test. If all expected counts are above 5, only the χ^2 p-value is shown.
    b. In very small samples, the χ^2 and Fisher p-values can differ. Fisher’s test often gives a larger p-value (more conservative) when N is tiny, because it is calculating the *exact* probability under the null rather than relying on the large-sample approximation. As N grows, the two results become nearly identical.
    c. Summary:
        - Trigger: Fisher’s test appears when at least one expected count is less than 5.
        - Preference for Fisher: Researchers prefer Fisher’s exact test in such situations because the χ^2 test relies on large-sample approximations that are unreliable with small expected counts.
        - Preference for χ^2 in large samples: For larger N, the χ^2 test is computationally simpler, and the approximation becomes extremely accurate, so there is little benefit to using Fisher’s test.

2 | Correlation

Linear relationships

1. Increasing n from 50 to 300 makes the sample r cling more tightly to the target (≈0.60), narrows its confidence interval, and yields a much smaller p-value. This shows that larger samples reduce sampling variability and increase power.
2. With a small true effect ($r \approx 0.20$) and $n = 60$, the observed r can bounce widely across seeds and may even flip sign occasionally. This instability comes from high sampling error relative to the weak signal.

Non-linear and non-monotonic (Quadratic)

3. Often yes: r can be near zero despite a clear U-shape. The positive and negative slopes on either side of the minimum/maximum cancel in a linear summary, masking the strong but non-linear association.
4. When Noise is large enough that the two arms blur together (threshold depends on n and the scale), both the straight line and LOESS become close to flat and similarly uninformative.

Non-linear but monotonic (Cubic)

5. Yes. The straight line averages across the curve and often understates the visible relationship (sometimes showing a small slope). LOESS follows the bend and reveals a strong monotonic trend.
6. As Noise increases, r typically declines quickly toward zero because curvature plus added noise erodes linear association. The LOESS curve still shows the underlying shape longer, revealing curvature even when r is small.

Comparing fits (lines)

7. For a curved monotonic pattern, the straight line can imply weak or even misleading direction, while LOESS shows the true increasing pattern with curvature (steeper ends, flatter middle).
8. In Linear mode (truly linear data), LOESS and the least-squares line essentially coincide, especially as n grows, because the best smooth for a straight-line signal is a straight line.

Sampling variability and reproducibility

9. With $r = 0.40$ and $n = 80$, r typically varies noticeably across seeds (often on the order of ± 0.1 or so). With $n = 400$, the variability shrinks substantially (often to just a few hundredths), illustrating how larger n stabilizes estimates.
10. For Cubic with low Noise, r is moderately stable but still more variable than in a truly linear case because the linear summary is sensitive to where the curve is sampled. Stability improves with larger n and lower Noise.

Same r, different story

11. You can tune Cubic (noise) to get a Pearson r similar to a Linear scenario (e.g., both around 0.5). The plots differ: Linear looks elliptical with a constant slope; Cubic is curved. LOESS highlights this curvature, showing why the same r can represent different relationships.

Reflections

12. Pearson's r is adequate when the relationship is approximately linear and residual variation is roughly symmetric and homoscedastic. Prefer a non-linear description when you see curvature or non-monotonic structure that a straight line cannot capture.
13. Justify a non-linear model when LOESS departs clearly from the straight line, residuals show systematic patterns, and a non-linear fit improves predictive performance (e.g., lower residual error, better cross-validated metrics) while matching the visible shape in the plot.

3 | Applied Exercises in R (Core & Going Further)

See RScript in the Online Companion