PO12Q: Introduction to Quantitative Political Analysis II

Week 1 - Worksheet Solutions

Dr Florian Reiche
F.Reiche@warwick.ac.uk

1 | Calculations by Hand

Expected Values:

	Mode of Transport			
Year of Study	Bike	Bus	Car	Total
Fresher	7.2	6.8	6	20
Finalist	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 ≈ 0.16. With N=80 (21/19/19/21) $\chi^2 \approx 0.20$, p ≈ 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. When Fisher appears (and how it compares)
    a. For a=1, b=4, c=4, d=1, N=10 and each expected count is 2.5 (< 5), so the app shows Fisher p in addition to χ^2. This is visible after ticking “Show expected counts table”.
    b. For a=1, b=4, c=4, d=1: $\chi^2 \approx 3.6$ ($p \approx 0.058$), Fisher’s two-sided $p \approx 0.21$ (less significant). Changing to a=1, b=9, c=9, d=1 makes each expected count exactly 5, so the app no longer surfaces Fisher p (by design it appears only when any expected < 5).
    c. Fisher’s exact test is preferred when any expected count is below 5 (small samples, sparse tables, or zeros). It does not rely on the large-sample approximation underlying χ^2, so it gives valid p-values where χ^2 can be liberal or unstable.

2 | Cross-Tabulations in R – Exercises

1. Let is find out whether the completion of primary school influences youth unemployment rates.

a. State the null and directional alternative hypothesis for this test.
b. Create a new variable `primary_fac` using the `primarycom` variable. Cut it into three categories "low", "medium", and "high", cutting `primarycom` at its first quartile, and its mean.

```
summary(wdi$prim_compl)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   37.50   90.71   98.27   92.74  101.50  114.27      83

    wdi <- wdi %>%
      mutate(primary_fac=
               ordered(
                 cut(prim_compl, breaks=c(0,59.868,77.952,135),
                     labels=c("low","medium", "high"))))
```

c. Apply the same procedure to `unemploy`, creating a new variable called `unemp_fac`.

```
summary(wdi$unemploy)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.170   3.555   6.100   7.646   9.908  27.690      25

        wdi <- wdi %>%
          mutate(unemp_fac=
                   ordered(
                     cut(unemploy, breaks=c(0,3.585,8,38),
                         labels=c("low","medium", "high"))))
```

d. Create a cross-tabulation assessing the dependence of youth unemployment on primary completion rate.

```
ex1_table <- with(wdi, table(primary_fac, unemp_fac))
```

e. Test whether the dependence is statistically significant.

```
Xsq <- chisq.test(ex1_table, correct=FALSE)
Xsq
##
##  Pearson's Chi-squared test
##
## data:  ex1_table
## X-squared = 7.0878, df = 4, p-value = 0.1313
```