Chi-squared tests of independence/homogeneity for two-way tables

SECTION 2.4

Example: *Lighting the Way to Nearsightedness*

Survey of ***n* = 479** children. → Independence
Those who slept with nightlight or in fully lit room before age 2 had higher incidence of nearsightedness (myopia) later in childhood.

TABLE 2.3 Nighttime Lighting in Infancy and Eyesight

Slept with:	No Myopia	Myopia	High Myopia	Total
Darkness	155 (90%)	15 (9%)	2 (1%)	172
Nightlight	153 (66%)	72 (31%)	7 (3%)	232
Full Light	34 (45%)	36 (48%)	5 (7%)	75
Total	342 (71%)	123 (26%)	14 (3%)	479

3 x 3

conditional %
on light cat's

Note: Study *cannot prove* sleeping with light actually *caused* myopia in more children. WHY?

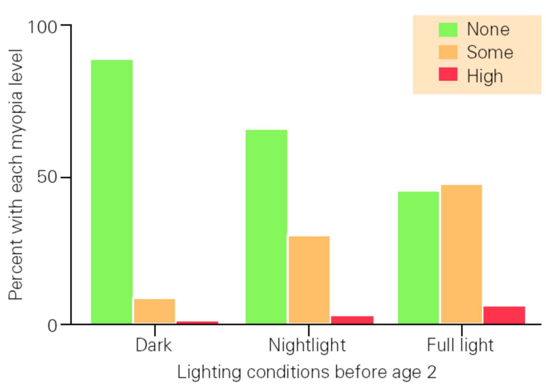


FIGURE 2.3 Bar chart for myopia and nighttime lighting in infancy

```
> test <- chisq.test(dat$x, dat$y, correct=FALSE)
Warning message:
In chisq.test(dat$x, dat$y, correct = FALSE) :
  Chi-squared approximation may be incorrect
> test

   Pearson's Chi-squared test

data:  dat$x and dat$y
X-squared = 58.374, df = 4, p-value = 6.368e-12

> test$expected
       dat$y
dat$x        High      None     Some
  Dark  5.027140 122.80585 44.16701
  Full  2.192067  53.54906 19.25887
  Night 6.780793 165.64509 59.57411
> test$residuals   # Components of sum in test stat
       dat$y
dat$x          High        None        Some
  Dark  -1.35011886  2.90514293 -4.38877137
  Full   1.89653070 -2.67146792  3.81477800
  Night  0.08418089 -0.98250048  1.60989895
```

→ to p-value

<5 → so not large enough

Why is it giving us a warning??

Verify these calculations

→ Expected cell counts are < 5 (?)

χ^2 df = 4

H_0: independence ←

H_A: not independent

$\dfrac{\text{row total} \cdot \text{column total}}{\text{table total}}$

Pearson resids = $\dfrac{\text{Obs} - \text{Exp}}{\sqrt{\text{Exp}}}$

Pearson $X^2 = \sum_{\text{all cells}} \left(\dfrac{\text{Obs}-\text{Exp}}{\sqrt{\text{Exp}}}\right)^2$

Example: Nicotine Patch

Double-blind randomized experiment (1994) where 240 smokers were randomly assigned to either a nicotine patch or placebo patch (see case study for details):

	Quit	Didn't	Total	% Quit
Nicotine	56	64	120	46%
Placebo (baseline)	24	96	120	20%
Total	80	160	240	33%

Find and interpret all summary measures for these data.

Conduct a chi-squared test of independence for these data.

Conduct a test for difference in proportions for these data.

H_0: no difference in quit prop. across treatment categories

2×2

Nicotine Example: Step 1

Population: For the nicotine patch example, our hypotheses are about the hypothetical behavior of *all* smokers with a desire to quit, *if* given nicotine patch compared with *if* given placebo patch similar to those in the study (not a random sample).

Null hypothesis (H_0): In the population of smokers who want to quit, there is no association between patch type and whether or not someone quits smoking.

Alternative hypothesis (H_a): In this population, there *is* an association between patch type and whether or not someone quits smoking.

} Independence

Nicotine Example: Step 2

Observed counts	Quit	Didn't	Total	% Quit
Nicotine	56	64	120	46%
Placebo (baseline)	24	96	120	20%
Total	80	160	240	**33%**

What to expect if no relationship?

Note that 80/240 = 1/3 (or 33%) quit smoking overall.

If there is no difference in the effect of patch type, we *expect* to see 1/3 of each type quit. So, we would expect:

Expected counts	Quit	Didn't	Total	% Quit
Nicotine	40	80	120	**33%**
Placebo (baseline)	40	80	120	**33%**

Nicotine Example: Step 2

O **= Observed count** in each cell = actual sample data

E **= Expected count** (if null is true) in each cell =

$$\frac{(Row\ total)(Column\ total)}{Total\ sample\ size}$$

Note: Only need to compute E for one cell; others determined by totals.

	Quit	Did not quit	Total
Nicotine	**56** **(120)(80)/240** **= 40**	**64** **(120)(160)/240** **= 80**	**120**
Placebo	**24** **(120)(80)/240** **= 40**	**96** **(120)(160)/240** **= 80**	**120**
Total	**80**	**160**	**240**

Nicotine Example: Step 2

Data conditions:

- Sample is representative of the population of smokers with a desire to quit similar to those in the sample. ✓
- All expected counts are greater than or equal to 5. ✓

How far are observed numbers who quit from what we expect if there is no difference for patch types?

$\frac{(56-40)^2}{40} = \frac{256}{40} = 6.4$	$\frac{(64-80)^2}{80} = \frac{256}{80} = 3.2$
$\frac{(24-40)^2}{40} = \frac{256}{40} = 6.4$	$\frac{(96-80)^2}{80} = \frac{256}{80} = 3.2$

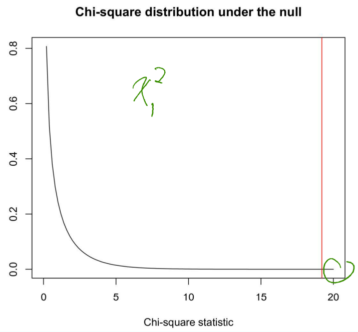
$\Rightarrow \chi^2$ = 6.4+3.2 +6.4+3.2 = 19.2

Does this value indicate a strong relationship in the population?? On to Step 3…

Nicotine Example: Step 3

The p-value is the *probability* of seeing a chi-squared statistic of 19.2 *or greater* in a sample of size 240, *assuming there is no relationship* between patch type and ability to quit smoking.

The area under the curve to the right of 19.2… pretty much zero.



Nicotine Example: Step 3

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 19.2, df = 1, p-value = 1.177e-05
```

p-value = 1.177 × 10^{-5} = 0.00001177

Nicotine Example: Step 4

from χ^2 w/ 1 df

For the nicotine patch example: The p-value of 0.00001177 is much less than 0.05, so

the relationship is statistically significant.

we reject the null hypothesis.

Each of the two statements above are equivalent (you only need to say one).

Nicotine Example: Step 5

Conclusion: There is significant evidence that there is a relationship between type of patch worn and the ability to quit smoking if we were to give nicotine or placebo patches to the entire population of smokers similar to those in the sample.

Note: Because this was a well-designed *randomized experiment*, we have evidence that using a nicotine patch *causes* the probability of quitting to increase.

Swedish Fish Example

Work through in Rstudio.

Inference on Contingency Tables

SUMMARY

Exact Inference

1. Exact binomial inference for a single binary variable (one proportion):
   - Use binomial distribution to calculate p-value
   - Invert test to obtain confidence interval
   - R: `binom.test()`
2. Exact inference for 2x2 tables: permutation in 217
   a. Randomization (simulation-based) test (not in book)
      - Uses simulation to approximate an exact p-value
      - Can be generalized to other scenarios, e.g., 1 x *c* table
      - R: functions in the `mosaic` library
   b. Fisher's Exact Test (2.6.1-2)
      - Calculates p-value using hypergeometric distribution
      - R: `fisher.test()`

Asymptotic Inference

3. Asymptotic inference for 2x2 tables using a normal approximation via CLT:
   a. Difference in proportions (2.2.1)
      R: `prop.test()`
   b. Relative risk (2.2.3)
      R: `relrisk()` (in `mosaic` library)
   c. Odds ratio (2.3)
      R: `oddsRatio()` (in `mosaic` library)
4. Asymptotic inference for *I x J* tables using a chi-squared distribution approximation to distribution of test statistic:
   a. Pearson chi-squared test of independence/homogeneity (2.4.1)
      R: `chisq.test()`
   b. Likelihood ratio test (2.4.2) (no built-in R function)