15. Contingency Tables

In this chapter, we compare proportions of units (subjects) categorized according to the levels of two factors (grouping variables) each with multiple levels. The book tends present the material by comparing frequencies (counts) in each group, but it may be easier to think about comparing proportions because they remind us of probabilities. We use the same examples as in the book to give you both perspectives.

Here are some typical situations that we'll encounter:

- Collect a sample from one population and categorize each unit according to the levels of 2 factors. The outcome from one unit does not affect the outcome from another (i.e., observations are assumed independent). Are the factors related in some way or are they independent?
- Collect samples from 2 or more populations and categorize each according to the levels of a single factor. Again, the observations are assumed independent, both within samples and between samples. Does the distribution of counts (or proportions) among the various factor levels differ across the populations from which the samples were obtained? That is, are the distributions homogeneous across populations?
- Collect samples from 2 populations, but now the units from one sample are paired with those of the other; the observations cannot be assumed to be independent across samples. Then, as above, each sample is categorized according to the levels of some factor. Does the distribution of counts (or proportions) for this factor's

levels differ across the populations from which the samples were obtained?

As we go through this chapter, try to recognize these different situations. It turns out that the first two situations may be analyzed using the same method. The pairing in the third situation will require a modification of the procedure used for the first two situations.

15.1. The Chi-Square Test

15.1.1. 2x2 tables

E.g., Bicycles, Helmets, and Head Injuries.

Here, a sample of 793 subjects is cross-classified according to the levels of 2 factors. Should we wear a helmet, or is wearing a helmet independent of head injury in accidents?

Data in S-Plus

helmet	injury	count
yes	yes	17
yes	no	130
no	yes	218
no	no	428

H_0: For the population of bicyclists involved bicycle accidents, the proportion of head injuries among those wearing helmets is the same as the proportion of injuries among those not wearing helmets. That is, there is no association between head injuries and wearing a helmet (i.e., they're independent).

H_A: The proportion of injuries differs between those wearing a helmet compared to those not wearing a helmet. That is, wearing a helmet is associated with head injury in bicycle accidents, in some way (i.e., they're dependent).

Statistics>Data Summaries>Crosstabulations…

```
     *** Crosstabulations ***
Call:
crosstabs(formula = count ~ injury + helmet, data = headhurts,
na.action = na.fail, drop.unused.levels = T)
793 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
injury |helmet
       |yes    |no     |RowTotl|
-------+-------+-------+-------+
yes    | 17    |218    |235    |
       |0.072  |0.928  |0.3    |
       |0.116  |0.337  |       |
       |0.021  |0.275  |       |
-------+-------+-------+-------+
no     |130    |428    |558    |
       |0.233  |0.767  |0.7    |
       |0.884  |0.663  |       |
       |0.164  |0.540  |       |
-------+-------+-------+-------+
ColTotl|147    |646    |793    |
       |0.19   |0.81   |       |
-------+-------+-------+-------+
Test for independence of all factors
     Chi^2 = 28.2555 d.f.= 1 (p=1.063122e-007)
     Yates' correction not used
```

If our sample size were large enough, we might use the frequentist definition of probability to assume the above proportions describe the population probability distribution across the different levels of the factors. If we could assume this, we could simply multiply marginal probabilities (e.g. P(helmet)*P(injury), etc.) to see if these values are the same as joint probabilities (e.g. P(helmet and injury), etc.). If they are the same, we know the helmet and injury factors are independent. If they differ, the probability of head injury depends on whether you wear a helmet or not (i.e., helmet and

injury factors are dependent). Because we will not assume the proportions are probabilities (but it may help to think about them that way), any differences between the product of marginal proportions and the corresponding joint proportion may just be due to sampling variability. We need to decide when these differences are not due just to chance alone.

Idea: multiply marginal proportions to get joint proportions that would be expected IF the factors were independent (null). Then, compare the expected proportions to observed proportions and determine how much they differ.

Equivalently, we could get expected counts by multiplying expected proportions by the total sample size:

- Let O_i be the observed count in each cross-classified category, i. In this case, we have i = 1 to 4 because we have the cross-classified categories: yes/yes, yes/no, no/yes, no/no
- Let E_i be the expected count (assuming independence of factors) in the same categories.
    - e.g. yes/yes: (235/793)*(147/793)*793=235*147/793

The degree to which the observed counts differ from the expected counts is an indication of the degree of departure from the null hypothesis. We'll use the chi-square test statistic,

X^2, and its associated distribution to quantify this degree of departure in the form of a p-value.

Notice we already did the chi-square test in S-Plus above.

The chi-square test is an approximate test based on large sample sizes. A rule of thumb:

- expected counts should be greater than 1 in each cross-classified category.
- no more than 20% of cross-classified cells should have expected counts less than 5.

That it is an approximate test should be obvious when you consider that we are working with categorical data but using a continuous distribution (chi-square is a continuous distribution).

The approximation may be questionable in a 2x2 (1df) situation. An attempt to improve the test is the

Yates Correction:

Splus Says: Statistics>Compare Samples>Counts and Proportions>Chi-square Test…

```
        Pearson's chi-square test with Yates' continuity correction

data:  headhurts2
X-square = 27.2018, df = 1, p-value = 0
```

Same conclusion as before.

In the above example, a single sample was classified according to two factors, and the test was for independence among the factor variables.

Another test which is exact (no approximating going on) is Fisher's exact test. But, it is computationally intensive and S-Plus will not do it for a total sample size greater than 200.

15.1.2. rxc tables

The above e.g. was a 2x2 table (2 levels of a factor by 2 levels of another). But, the chi-square method can be extended to the more general case of rxc (rows by columns), where the

levels of one factor are associated with table rows and the levels of the other factor with the columns.

E.g. Death certificate status across hospitals

Statistics>Data Summaries>Crosstabulations…

```
     *** Crosstabulations ***
Call:
crosstabs(formula = count ~ hospital + cert.status, data = death,
na.action = na.fail, drop.unused.levels = T)
575 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
hospital|cert.status
        |accurat|no.chng|recode |RowTotl|
--------+-------+-------+-------+-------+
A       |157    | 18    | 54    |229    |
        |0.686  |0.079  |0.236  |0.4    |
        |0.369  |0.290  |0.614  |       |
        |0.273  |0.031  |0.094  |       |
--------+-------+-------+-------+-------+
B       |268    | 44    | 34    |346    |
        |0.775  |0.127  |0.098  |0.6    |
        |0.631  |0.710  |0.386  |       |
        |0.466  |0.077  |0.059  |       |
--------+-------+-------+-------+-------+
ColTotl|425    |62     |88     |575    |
        |0.74   |0.11   |0.15   |       |
--------+-------+-------+-------+-------+
Test for independence of all factors
     Chi^2 = 21.52346 d.f.= 2 (p=0.00002119536)
     Yates' correction not used
```

In this example, we can identify two (conceptual) populations: death certificates in hospital A, and death certificates in hospital B. We are interested in how the distribution of proportions (think probability) across the death certificate status may differ across hospitals. That is, we are wondering if

the proportions in each status level are the same (homogeneous) for each hospital. Once more, we are testing for independence of the status factor and hospital factor. When the levels of one factor can be identified with different populations as here, the chi-square test is often called a "test for homogeneity" (of status distribution across hospital populations). In the first e.g. (helmets/injury), we had one population classified according to two factors. In this case, the test is often called a "test for independence" among factors. In either case, the test is carried out in the same way. In both cases, we may generally say that we are testing the null hypothesis of no association between factors.

H_0: Within each category of death certificate status, the proportions of death certificates are identical for Hospitals A and B.

15.2. McNemar's Test (paired samples)

In the hospital example, the subjects (death certificates) were in no way paired across hospitals. It was fairly clear that the samples of death certificates from each hospital could be assumed to be independent. But, sometimes, units may be paired across samples.

E.g., Heart Attack and Diabetes among Navajos

We want to know whether heart attacks are associated with diabetes. Now, we know that there may be several factors that contribute to heart attacks and diabetes (e.g., age, gender), but we want to control for these effects and try to isolate the association between heart attack and diabetes. Here, 144 subjects having heart attacks were age/gender matched with 144 subjects not having heart attacks. We assume that one pair of subjects does not influence any other pair (independence across pairs).

Data in S-Plus

attack	diabetes	count
yes	yes	46
yes	no	98
no	yes	25
no	no	119

Statistics>Data Summaries>Crosstabulations…

```
     *** Crosstabulations ***
Call:
crosstabs(formula = count ~ diabetes + h.attack, data = death,
na.action = na.fail, drop.unused.levels = T)
288 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
diabetes|h.attack
        |yes    |no     |RowTotl|
--------+-------+-------+-------+
yes     | 46    | 25    |71     |
        |0.648  |0.352  |0.25   |
        |0.319  |0.174  |       |
        |0.160  |0.087  |       |
--------+-------+-------+-------+
no      | 98    |119    |217    |
        |0.452  |0.548  |0.75   |
        |0.681  |0.826  |       |
        |0.340  |0.413  |       |
--------+-------+-------+-------+
ColTotl|144    |144    |288    |
        |0.5    |0.5    |       |
--------+-------+-------+-------+
```

Notice the above table suggests that we have 288 independent observations. We don't; we really only have 144 independent pairs of observations. The idea is that, by pairing, we control for the effects of extraneous factors (age, gender) and get a more precise look at the relationship between factors of interest (heart attack, diabetes). We hope the pairing more than makes up for the fewer pieces of independent information than if we had sampled independently (no pairing, but bigger total sample size). The chi-square test assumes independence among observations and is not appropriate without modification to the paired case.

Alternative table:

attack	no attack		total
	diabetes	no diabetes	
diabetes	9	37	46
no diabetes	16	82	98
total	25	119	144

H_0:

H_A:

Which pairs in the table give information on the departure from the null hypothesis?

Concordant Pairs and Discordant Pairs.

Splus
Data
9 37
16 82

Statistics>Compare Samples>Counts and Proportions>McNemar's Test...

```
     McNemar's chi-square test with continuity correction

data:  navajo
McNemar's chi-square = 7.5472, df = 1, p-value = 0.006
```

McNemar's test is also a large sample approximation (even with the continuity correction). What's large?

15.3. Odd's Ratio

Another way to test association between factors in a contingency table is to use the odds ratio (OR). We'll discuss this using 2x2 tables with factors that we will generically call Disease and Exposure. We assume independent observations from one cross-classified sample.

If we had information on the entire population of interest we could construct a table of (population) probabilities as follows

	Exposed	Not Exposed	
Disease	P(D and E)	P(D and E-)	P(D+)
No Disease	P(D- and E)	P(D- and E-)	P(D-)
	P(E)	P(E-)	1

The population odds ratio is then

$$OR = \frac{\frac{P(D|E)}{1-P(D|E)}}{\frac{P(D|E-)}{1-P(D|E-)}} = \frac{\frac{\frac{P(D \cap E)}{P(E)}}{\frac{P(D^- \cap E)}{P(E)}}}{\frac{\frac{P(D \cap E^-)}{P(E^-)}}{\frac{P(D^- \cap E^-)}{P(E^-)}}} = \frac{P(D \cap E) * P(D^- \cap E^-)}{P(D^- \cap E) * P(D \cap E^-)}$$

We could do the same with a table of sample proportions (i.e., don't assume the sample size is large enough to invoke frequentist definition of probability).

Sample cross-classified counts:

	Exposed	Not Exposed	
Disease	a	b	a + b
No Disease	c	d	c + d
	a + c	b + d	n

$$OR == \frac{\frac{a}{n}}{\frac{c}{n}} = \frac{ad}{bc}$$

If exposure does not affect disease, then we expect OR to be about 1 (in the population case, it would be 1). In the sample case, we expect sampling variability, so we would like to have more than just a point estimate of OR. We need to find a(n) (approximate) distribution of OR so we get an idea of the precision of the estimated OR. Then we can construct confidence intervals, etc.

Once again (thanks to some version of the CLT), we have a large sample theoretical result to help us out:

- $\ln(OR) \sim N(\ln(OR), (s.e.(\ln(OR)))^2)$
- $s.e.(\ln(OR))=\sqrt{1/(a+.5) + 1/(b+.5) + 1/(c+.5) + 1/(d+.5)}$

So, an approximate $(1-\alpha)*100\%$ CI for ln(OR) is

- $[\ln(OR) - z_{\alpha/2}s.e.(OR), \ln(OR) + z_{\alpha/2}s.e.(OR)]$

Most of us don't like to think in terms of logarithms, so we exponentiate to get a(n) (approximate) $(1-\alpha)*100\%$ CI for OR

- $[\exp\{\ln(OR) - z_{\alpha/2}s.e.(OR)\}, \exp\{\ln(OR) + z_{\alpha/2}s.e.(OR)\}]$

(Note: ln(OR) is the mean and median of the distribution of ln(OR), but OR is the median of the distribution of OR, at least approximately)

E.g. Bicycles revisited.

Here, we view a head injury as "disease" and not wearing a helmet as "exposed". You may want to switch columns in bicycle table to ease confusion (or just relabel counts). When we do this we get:

15.4. Berkson's Fallacy

Again, reading is FUNdamental.