Contingency Tables

Basics:

Much of this should be a straight forward review. Let's just do a quick example:

Seat belts in Florida (data from 1988):

Safety equipment in use	Injury		
	Fatal	Non-fatal	Total
None	1,601	165,527	167,128
Seat belt	510	412,368	412,878

H$_0$: $p_1 = p_2$

p$_1$ = proportion of fatalities for people not wearing a seat belt (estimated by $\hat{p}_1$).

p$_2$ = proportion of fatalities for people wearing a seat belt (estimated by $\hat{p}_2$).

H$_1$: $p_1 > p_2$

We calculate the expected values:

Remember:

$$\text{Expected value} = \frac{\text{(Row total)} \times \text{(Column total)}}{\text{(Grand total)}}$$

Safety equipment in use	Injury		
	Fatal	Non-fatal	Total
None	608.28	166,519.72	167,128
Seat belt	1,502.72	411,375.28	412,878
Total	2,111	577,895	580,006

And our we calculate our χ^{2*} using: $$\chi^{2*} = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2{*} = \frac{(1601-608.28)^2}{608.28} + \frac{(165{,}527-166{,}519.72)^2}{166{,}519.72} +$$

$$\frac{(510-1502)^2}{1502.72} + \frac{(412{,}368-411{,}375)^2}{411{,}375{,}28} = 2284.25$$

If we look up our critical value of χ^2 in the table (1 *d.f.*, and $\alpha = .05$), we get:

$$\chi^2_{\text{table}} = \chi^2_{.05,1} = 2.71$$

So we reject H_0 and conclude that seat belts do affect the outcome of a traffic accident.

Somewhere we should have made sure the data deviate from the null hypothesis in the direction of your alternative (or we would have STOPPED)

In other words, we should have made sure that $\hat{p}_1 > \hat{p}_2$.

We note that $\hat{p}_1 = 0.00958$, $\hat{p}_2 = 0.00124$, so we did okay.

Some comments:

The chi-square contingency test can actually be used to test two different hypotheses. We either:

1) Compare proportions (like in the seatbelt example)

- or -

2) Establish independence/dependence

And we would phrase our hypotheses accordingly:

1) H_0: $p_1 = p_2$, where the p's are proportions,
H_1: $p_1 \neq p_2$ (or $p_1 > p_2$, etc.)

2) H_0: Factor 1 and factor 2 are independent
H_1: Factor 1 and factor 2 are dependent

The math and our decision rule is identical(!)

R x K tables work the same way (no changes to the math).

So what about the assumptions?

i) random data

ii) smallest expected value ≥ 5

(There is some debate as to how serious this restriction is, but we'll stick with it).

Relative Risk:

A) Simply, relative risk is the ratio of $\hat{p}_1 / \hat{p}_2$. For instance, suppose we wanted to take another look at our Seat belt safety data from Florida:

Safety equipment in use	Injury Fatal	Non-fatal	Total
None	1,601	165,527	167,128
Seat belt	510	412,368	412,878

Previously, we established that the proportion of fatalities wearing seat belts is 510/412,878 = $\hat{p}_2$ = .001235, and the proportion of fatalities for not wearing a seat belt is 1,601/165,128 = $\hat{p}_1$ = .00958.

So, this means that our relative risk is .009672/.001235 = 7.83. In other words, the risk of dying is 7.8 times higher if one is not wearing a seatbelt than if one is wearing a seatbelt. This makes it sound pretty good for wearing a seatbelt!

Often, when one hears reports in the news that a particular risk is x times something else, they are talking "relative risk". It is important to realize that often the overall chance of something happening is very small. The chance of dying in a car accident, regardless of seatbelt use is quite small.

B) Another example from the old 214 textbook (*Statistics for the Life Sciences*, Samuels et al.)

	Smoking Status	
Birthweight	Smoker	Nonsmoker
Low	237	197
Normal	3,489	5,870
TOTAL	3,726	6,067

(note that the table is arranged the other way from the one above)

$\hat{p}_1$ = 237/3,726 = .064 $\hat{p}_2$ = 197/6,067 = .032

So our relative risk is .064/.032 = 2, indicating that the risk for low birthweight in babies is twice as high for smokers as for non-smokers.

(Incidentally - what do we need to know before we can really conclude this? Chi-square = 52.8, , and *p* < .001, so there is a significant difference here.)

Odds Ratio:

A) This time, one develops the "odds" of something happening. The odds ratio is defined as follows:

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

where theta is the odds ratio. As you might guess, we estimate this using theta-hat and the appropriate p-hat:

$$\hat{\theta} = \frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}}$$

B) For example, using the seat-belt data above, we have:

$$\hat{\theta} = \frac{\frac{0.00958}{1-0.00958}}{\frac{0.001235}{1-0.001235}} = 7.822$$

which is very similar to the relative risk given above.

Relative Risk and Odds ratio:

As it turns out, when $\hat{p}_1$ and $\hat{p}_2$ are both very small, the Relative Risk and the Odds ratio are very close. This is easy to see from the equation above for $\hat{\theta}$. If $\hat{p}_1$ and $\hat{p}_2$ are small, then you wind up dividing $\hat{p}_1$ and $\hat{p}_2$ by 1, which is essentially the Relative Risk.

However, if $\hat{p}_1$ and $\hat{p}_2$ are not both small, you can get quite different results.

Here's another example from the same textbook mentioned above:

	Treatment	
	Timolol	Placebo
Angina-free	44	19
Not angina-free	116	128
TOTAL	160	147

Note that what we're really interested in here is the relative risk of having an angina if one takes a placebo as opposed to Timolol. We really should re-arrange the table a bit, but it's not necessary.

What we want is $\hat{p}_1$ = proportion of people taking placebo who have angina = 128/147 = .871, and $\hat{p}_2$ = proportion of people taking timolol who have angina = 116/160 = .725.

So now our Relative Risk = .871/.725 = 1.20

But what happens to our Odds Ratio?

Odds Ratio = .871/(1-.871)/.725/(1-.725) = 2.55

So here we have two different bits of information:

i) The risk of having an angina attack if taking a placebo rather than timolol is 1.2.

ii) But the odds of having an angina attack are actually 2.6 if one is taking a placebo rather than timolol.

A little confusing, no??

Remember:

Relative Risk measures the risk of a nasty outcome compared to some "treatment".

Odds Ratio just compares the odds of one outcome happening over another.

(odds ratio often makes more sense than relative risk: "the relative risk of having dark eyes if one has dark hair.....????)

To finish off this topic (and to confuse things a bit more!!):

a) When can one use RR and OR?

i) OR one can use anytime

ii) RR can only be used in certain circumstances

b) For example, if one goes out and samples 10,000 smokers and 10,000 non-smokers to see if they died of lung cancer, what happens (the proportions are real data based on a study in Britain, but the numbers in the table are made up)?

Disease

Death from lung cancer	Smoker	Non-smoker
Yes	14	1
No	9,986	9,999
TOTAL	10,000	10,000

c) Relative risk of smoker dying from lung cancer as opposed to a non-smoker dying of lung cancer is 14/1 = 14 (technically (14/10,000) / (1/10,000) = 14).

d) Does it make sense to calculate the relative risk of dying from lung cancer if you're a smoker as opposed to dying from something else and if you're a smoker?

the math would be: 14/15 = .933
9986/19985 = .500

and the relative risk would be ..933/500 = 1.87

But it doesn't make any sense!!!

NO - For two reasons:

i) it doesn't really make much sense (think about it)

ii) you can't do it mathematically:

- you selected the number of smokers and non-smokers!! The number of smokers in your study is no longer random, so you can't estimate anything having to do with the number of smokers, at least not directly.

- for example, suppose you picked 5,000 smokers and 10,000 non-smokers. Using the same proportions, our relative risk would now be:

7/8 = .875
4993/14992 = .333

and the RR would be: .875/.333 = 2.627

- notice that it's changed completely! In other words, it depends on our sample sizes, and it SHOULD NOT! But it does, so it's totally useless.

e) But what about the Odds ratio? The Odds ratio is:

$$\text{Odds ratio} = \frac{14 \text{ x } 9{,}999}{1 \text{ x } 9{,}986} = 14$$

(this is a shortcut formula, only usable in 2x2 tables:

$$\hat{\theta} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

where n_{11} is the upper left, n_{12} the upper right, etc.)

almost the same (it is the same if you round it like here). But now a little mathematical fact:

$\hat{p}_1$ = proportion of people who died of lung cancer who smoke

$\hat{p}_1^*$ = proportion of people who smoke if they died of lung cancer

$\hat{p}_2$ = proportion of people who died of lung cancer who didn't smoke

$\hat{p}_2^*$ = proportion of people who smoke if they did not die of lung cancer

As it turns out, the Odds ratio is the same if one uses $\hat{p}_1$ and $\hat{p}_2$ or $\hat{p}_1^*$ and $\hat{p}_2^*$.

So suppose we had the following data (were we select 10,000 people who died of lung cancer and 10,000 people who didn't die of lung cancer):

	Disease		
Death from lung cancer	Smoker	Non-smoker	
Yes	9,333	667	10,000
No	4,997	5,003	10,000

Now we can't calculate the Relative Risk of dying from lung cancer if you're a smoker vs. a non-smoker.

But the Odds ratio still works, and we can easily calculate it:

$$\frac{9{,}333 \text{ x } 5{,}003}{4{,}997 \text{ x } 667} = 14$$

(I used the same proportions as above, just made the samples bigger)

So now we can say that the odds of a smoker dying of lung cancer are 14 times that of a non-smoker dying of lung cancer. It still works!

This is why the Odds ratio is used a lot in medicine and epidemiology - it can be used even when the study wasn't designed to find certain relationships.

Fisher's exact test (or what to do if you violate the assumptions)**:**

Let's start with an example that doesn't come from biology, but a little history of statistics doesn't hurt (this is a famous example from statistics).

Way back when, Fisher came across a lady who claimed to be able to tell if tea or milk were poured into a cup first (the British like to drink milk with their tea). So he put her to the test with eight cups, four of which had milk added first, the other four tea first. He got the following results:

	Lady's guess		
Poured first	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Looking at the results, one can see that the lady seems to have done a little better than "random guessing", but is this result explainable by chance?

Can we use a chi-squared statistic? NO! (all categories have expected values < 5)

But we can generate exact probabilities!

But before we do:

H_0: The lady can not tell the difference between pouring tea or milk first.

H_1: The lady can tell if milk or tea is poured first (note that this is actually a directional hypothesis)

$\alpha = .05$

What we do is calculate the probability of having gotten that result or "worse" if our null hypothesis is that the lady can't tell the difference. Our marginal totals are fixed for this test (they can't change - more on that later).

What is "worse"?

4	0	
0	4	(the lady did perfectly)

So now we just need to calculate the probability of getting this and the actual result we got.

This uses something called the "hypergeometric distribution", which is a bit similar to the binomial. Here's how you make it work:

The number of ways the lady could guess milk first in 3 out of 4 cups with milk poured first is:

$$\binom{4}{3} = \frac{4!}{3!1!} = 4$$

The number of ways the lady could guess milk first in 1 out of 4 cups with tea poured first is:

$$\binom{4}{1} = \frac{4!}{1!3!} = 4$$

And the number of ways she can guess milk first in four out of 8 cups:

$$\binom{8}{4} = \frac{8!}{4!4!} = 70$$

Notice that what we're doing is calculating the probability of getting the result she got and then dividing by the total number of possible outcomes one can get.

So now we just multiply the first two quantities and divide by the last:

4 x 4 = 16   16/70 = 0.229

Pr{of getting the above result} = 0.229

Now we repeat all this for the table with the "worse result":

Pr{of getting even worse result} = .014

add these up and we get 0.229 + 0.014 = 0.243 ($= p$).

So the probability of the lady guessing is .243, or a little less than ¼. We conclude that we do not have any evidence that she can tell the difference (our $p > \alpha$).

Some more details:

Here's a formula (for 2 x 2 tables):

$$\frac{\text{\# of ways of getting result in first cell} \times \text{\# of ways of getting result in second cell (row wise)}}{\text{Probability of getting result in column 1 total}}$$

where: # of ways of getting result in first cell = Binom (row 1 total, first cell count)

# of ways of getting result in second cell = Binom (row 2 total, second cell count)

# of ways of getting result in column totals = Binom(grand total, column 1 totals)

(you're basically going row by row, using the first column and the total column to get your binomial coefficients).

As you can see, this gets messy real fast. Not only do you have to calculate the probability of getting the table you got (the above formula), but also the probability of getting each worse table.

Fisher’s exact test is almost always done on a computer, particulary for 3 x 3 tables and higher (the formula above will need to be modified).

You should know how to do a very simple test, sort of like the Tea-drinker example. See also section 24.10 in your text.

As presented, Fisher’s test is actually directional. One can use Fisher’s test in a non-directional setting, but then we’d have to calculate all the tables in the “opposite direction” of our result or worse. For our tea drinker, we’d have to include tables:

$$\begin{array}{cc} 1 & 3 \\ 3 & 1 \end{array} \quad \text{and} \quad \begin{array}{cc} 0 & 4 \\ 4 & 0 \end{array}$$

(add up the probabilities of these to our total)

Let's do one more example (since this is sort of confusing). We'll do example 24.16 in the text. Note that all the stuff about antilogs isn't necessary, particularly if you're using R. We'll do this the same way as above:

We're looking at two species of snail, on of which is found in rapidly moving water (species 1).

We want to find out if there is a difference in how well these snails can resist moving water, and we get the following results:

		Outcome		
		resisted	yielded	Total
Species	1	12	7	19
	2	2	9	11
	Total	14	16	30

H_0: Snails are equally successful in resisting current
H_1: Snail species 1 is better at resisting current (directional)

$\alpha = .05$

Which tables are worse than our result? First, remember that our marginal totals are fixed. Why? We picked 19 snails for species 1, and 11 snails for species 2. Our null hypothesis implies that the numbers in our column total can't change (if there is no effect, then the probability of resisting is the same regardless of species, and we can use the column total to estimate that which implies that *column totals do not change.*

So here are the tables that would be worse than what we got.

$$\begin{array}{cc} 13 & 6 \\ 1 & 10 \end{array} \qquad \begin{array}{cc} 14 & 5 \\ 0 & 11 \end{array}$$

So now all we have to do (ha, ha) is figure out the probability of getting these three tables and add these probabilities up.

Probability of getting the result we got:

$$\frac{\binom{19}{12}\binom{11}{2}}{\binom{30}{14}} = 0.019057$$

(we could stop here - why??)

Probability of getting the two worse tables:

$$\frac{\binom{19}{13}\binom{11}{1}}{\binom{30}{14}} = 0.002053 \quad \text{and} \quad \frac{\binom{19}{14}\binom{11}{0}}{\binom{30}{14}} = 0.000080$$

Adding up all these probabilities, we get:

$$.019057 + .002053 + .000080 = .021189$$

Since our p-value is $\leq \alpha$, we conclude that snail 1 is better able to resist our water current. That's as much as we'll do with Fisher's exact test.