Recall: Hypothesis testing on Proportions

Let p be the true proportion of times that an event occurs in a population. Suppose we would like to test

$$H_0 : p = p_0 \text{ against } H_A : p \neq p_0,$$

We collect a sample of size n from the population of interest. Under the null, the estimate of the standard error of $\hat{p}$ takes the form

$$\sqrt{\frac{p_0(1-p_0)}{n}}$$

The appropriate test statistic is

$$Z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

If the null is true ($p = p_0$), this statistic is approximately standard normal for n large (we defined how large n needs to be last time).

An approximate $100(1-\alpha)$ % confidence interval for p:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Suppose we would like to test

$$H_0 : p_1 = p_2 \text{ against } H_A : p_1 \neq p_2,$$

We collect a sample of size n_1 from the first population and a sample of size n_2 from the second population. Under the null, the estimate of the standard error of the difference $p_1 - p_2$ takes the form

$$\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$

The appropriate test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

If n_1 and n_2 are large, this statistic is approximately standard normal.

A few lectures ago, we considered the effectiveness of bike helmets in preventing head injury. In particular, we considered two random samples: one of size 147 from a population of people that wear helmets and the other of size 646 from a population of people that do not wear helmets. We record that 17 of the 147 suffered a serious head injury and 218 of the 646 suffered a serious head injury. We wanted to know if the proportion of serious head injuries was the same in the two populations.

Recall the evaluated test statistic was

$$z = \frac{(0.116 - 0.337)}{\sqrt{0.296(1-0.296)\left[\frac{1}{147} + \frac{1}{646}\right]}} = -5.3$$

$$\begin{aligned} \text{p-value} &= P(Z \leq -5.3) + P(Z \geq 5.3) \\ &= 5.8 \cdot 10^{(-8)} + 5.8 \cdot 10^{(-8)} \\ &= 1.16 \cdot 10^{(-7)} \end{aligned}$$

The null was rejected at significance level $\alpha = 0.01$.

Another way to approach the same question is to consider a random sample of 793 bike riders and classify the riders using two questions:

1. Do you wear a helmet ?
2. Have you suffered a serious head injury ?

	Wearing Helmet		
Head Injury	(Y)	(N)	total
+ (Y)	17	218	235
- (N)	130	428	558
total	147	646	793

What would we expect this table to look like if the null was true ?

	Wearing Helmet		
Head Injury	(Y)	(N)	total
+ (Y)	NA	NA	235
- (N)	NA	NA	558
total	147	646	793

If the null was true, then the proportion of people suffering head injuries would be the same in the two populations (those that wear helmets and those that do not wear helmets). The proportion of people suffering head injuries is $\frac{235}{793} = 0.296$; and the proportion of people not suffering head injuries is $\frac{558}{793} = 0.704$

As a result, if the null is true, then for the 147 people wearing helmets, we would expect 29.6 % (43.6) of them to suffer a head injury and 70.4 % (103.4) of them to be free of head injury. Similar reasoning applies to the the 646 people wearing helmets. So, if the null is true, we'd expect the table to look like the one below:

	Wearing Helmet		
Head Injury	(Y)	(N)	total
+ (Y)	43.6	191.4	235
- (N)	103.4	454.6	558
total	147	646	793

How far is this expected table from the observed table ?

	Wearing Helmet		
Head Injury	(Y)	(N)	total
+ (Y)	43.6	191.4	235
- (N)	103.4	454.6	558
total	147	646	793

	Wearing Helmet		
Head Injury	(Y)	(N)	total
+ (Y)	17	218	235
- (N)	130	428	558
total	147	646	793

You could think about summing the squared differences between the four cells in the two tables:
$(17 - 43.6)^2 + (218 - 191.4)^2 + (130 - 103.4)^2 + (428 - 454.6)^2$.

Under the null,

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

is approximately chi-square (χ^2) distributed with $(2 - 1) \cdot (2 - 1)$ degrees of freedom.

For this example, the value of the test statistic is

$$x_2 = \frac{(17 - 43.6)^2}{43.6} + \frac{(218 - 191.4)^2}{191.4} + \frac{(130 - 103.4)^2}{103.4} + \frac{(428 - 454.6)^2}{454.6} = 28.32$$

p-value: $P(\chi_1^2 \geq 28.32) = 1.028 \cdot 10^{(-7)}$

The null is rejected and we conclude that there is an association between helmet wearing and suffering of a serious head injury.

Note: Since we are using discrete observations to estimtate a continuous distribution, a continuity correction could be applied which might make the approximation of the test statistic a little better. Yates proposed such a correction.

Under the null,

$$X^2 = \sum_{i=1}^{4} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

is approximately chi-square (χ^2) distributed with $(2 - 1) \cdot (2 - 1)$ degrees of freedom.

In the example above, the value of this corrected statistic is 27.27 and the pvalue is $1.769 * 10^{(-7)}$.

In practice, you will often see this correction applied to 2 x 2 tables.

Notes:

The chi-square test is used when a random sample of some population is obtained and then split into categories.

The chi-square test shown here can be extended to rxc tables.

The chi-square test is based on an approximation that works best when n is large. To make sure approximation is valid, no cell should have expected count less than 1 and no more than 20 % of the cells should have expected counts less than 5.

When n is not large enough, Fisher's exact test can be applied. It gives you the exact p-value. A bit lower on power, but, again, the p-value is exact.

The *New England Journal of Medicine* published a study in 2000 (Volume 343 (2): 78 - 85) entitled "Environmental and Heritable Factors in the Causation of Cancer".

In this study, thousands of twins were obtained. The study was done in part to see if there are certain types of cancer in which the concordance is higher among monozygotic (identical) twins than among dizygotic. If identified, the genetic effects in such cancers are likely to be important.

For one part of the study, 708 twin pairs were considered where at least one of the twins has lung cancer. Out of the 708 twin pairs, 248 were monozygotic. For 669 of the twin pairs, only one twin suffered from lung cancer; for the remaining 39 pairs, both suffered from lung cancer.

Determine if there is an association between occurrence of disease and amount of genome shared (twin status).

Q: Does this article suggest that the role of genes in cancer is as important (or more important) than other factors (e.g. environmental, behavioral) or does it suggest that genes play only a minor role in cancer development ?

Thus far, we have only considered 2 x 2 tests. What about comparing factors that can be classified into more than two categories ?

The chi-square test can be extended to compare more than two proportions.

Suppose we wish to study the relationship between age at which person gives birth to first child and the development of breast cancer. Information on 13,465 women is given below.

	Age at first birth					
Disease status	< 20	$20 - 24$	$25 - 29$	$30 - 34$	≥ 35	total
(Breast Cancer +)	320	1206	1011	463	220	3220
(Breast Cancer -)	1422	4432	2893	1092	406	10,245
total	1742	5638	3904	1555	626	13,465

As before, we need to calculate the expected table and see if this observed table is "very unusual" in comparison.

The proportion of people with breast cancer is 3220/13465 = 0.2391 and the proportion of people with no breast cancer is 10245/13465 = 0.7609. If age at first birth has no impact on the proportions, then for the 1742 people with first birth under the age of 20, we expect approximately 24 % ($1742 \cdot 0.2391 = 416.6$) to have breast cancer and 76 % of them ($1742 \cdot 0.7609 = 1325.4$) to be free of breast cancer.

In general,

The expected value in the (1,1) cell:

$$\frac{\text{first row total} \times \text{first column total}}{\text{grand total}} = \frac{3220(1742)}{13,465} = 416.6$$

The expected value in the (1,2) cell:

$$\frac{\text{first row total} \times \text{second column total}}{\text{grand total}} = \frac{3220(5638)}{13,465} = 1348.3$$

... verify the next few

The expected value in the (2,1) cell:

$$\frac{\text{second row total} \times \text{first column total}}{\text{grand total}} = \frac{10245(1742)}{13,465} = 476.3$$

... verify the next few

The expected value in the (2,5) cell:

$$\frac{\text{second row total} \times \text{fifth column total}}{\text{grand total}} = \frac{10245(626)}{13,465} = 476.3$$

The expected table is:

	Age at first birth					
Disease status	< 20	$20-24$	$25-29$	$30-34$	≥ 35total	
(Breast Cancer +)	416.6	1348.3	933.6	371.9	149.7	3220
(Breast Cancer -)	1325.4	4289.7	2970.4	1183.1	476.3	10,245
total	1742	5638	3904	1555	626	13,465

As before, we need to figure out if the observed table is "unusual" compared to this expected table.

Under the null,

$$X^2 = \sum_{i=1}^{r \cdot c} \frac{(O_i - E_i)^2}{E_i}$$

is approximately chi-square (χ^2) distributed with $(r-1) \cdot (c-1)$ degrees of freedom.

Here, $\Sigma_{i=1}^{r \cdot c} \frac{(O_i - E_i)^2}{E_i}$

$$= \frac{(320-416.6)^2}{416.6} + \frac{(1206-1348.3)^2}{1348.3} + \ldots + \frac{(406-476.3)^2}{476.3}$$
$$= 130.3$$

Consider another example (2 x 2). We want to figure out if electronic fetal monitoring during labor affects the frequency of caesarean sections. 5824 infants are randomly sampled. Out of these, 2850 were monitored and 2974 were not. Results of C-sections are below.

	Monitored		
C section	(Yes)	(No)	total
(Yes)	358	229	587
(No)	2492	2745	5237
total	2850	2974	5824

The expected table is

	Monitored		
C section	(Yes)	(No)	total
(Yes)	287.25	299.75	587
(No)	2562.75	2674.25	5237
total	2850	2974	5824

and the value of the test statistic is

$$\begin{aligned} x_2 &= \frac{(358-287.25)^2}{287.25} + \frac{(229-299.75)^2}{299.75} \\ &+ \frac{(2492-2562.75)^2}{2562.75} + \frac{(2745-2674.25)^2}{2674.25} \\ &= 37.95 \end{aligned}$$

The null is rejected with $p < 0.001$ and we conclude that there is an association between monitoring and C-sections.

We don't know (using tests on proportions) how strong the association is.

Recall the odds ratio.

If an event takes place with probability p, the odds in favor of the event are $\frac{p}{1-p}$ to 1. $p = \frac{1}{2}$ implies 1 to 1 odds; $p = \frac{2}{3}$ implies 2 to 1 odds.

In this class, the **odds ratio** (OR) is the odds of disease among exposed individuals divided by the odds of disease among unexposed.

$$OR = \frac{P(disease|exposed)/(1-P(disease|exposed))}{P(disease|unexposed)/(1-P(disease|unexposed))}$$

Note that the OR is sometimes defined alternatively as

$$OR_{alt} = \frac{P(exposure|disease)/(1-P(exposure|disease))}{P(exposure|nondiseased)/(1-P(exposure|nondiseased))}$$

Note that these definitions are equivalent (we showed that in an earlier lecture).

	Exposure		
Disease	(Yes)	(No)	total
(Yes)	a	b	a+b
(No)	c	d	c+d
total	a+c	b+d	n

We also showed that the odds ratio (if table above was obtained from n independent observations) could be estimated by $\frac{ad}{bc}$

Back to the C-section example...

	Monitored		
C section	(Yes)	(No)	total
(Yes)	358	229	587
(No)	2492	2745	5237
total	2850	2974	5824

The odds ratio here is $\frac{(358)(2745)}{(229)(2492)} = 1.72$

The odds of being delivered by C-section are 1.72 times greater for fetuses that are being monitored.

Consider two (unrelated) questions:

1. Does this imply that monitoring causes a condition which requires C-sections more often ?

2. Is 1.72 significantly different than 1?