Suppose we are interested in determining if there is an association between heart attacks and diabetes in the US Navajo population.

We could identify a number of Navajos willing to participate in our study and record the heart attack and diabetes status of each. Then, we could ask whether the proportion of diabetics among those that suffered a heart attack was equal to the proportion of diabetics that did not suffer a heart attack. However, since heart attacks might be relatively rare, we'd need a large sample size. We could instead identify Navajos that we know have had heart attacks.

144 Navajos that had suffered heart attacks were identified and a number of characteristics were recorded (age, gender, smoking status, etc.). Following this, 144 Navajos which matched each of the first 144 on every characteristic except heart attack status were identified. The pairs were then tested for diabetes.

	No Heart Attack		
Heart Attack	(Db+)	(Db-)	total
(Db+)	9	37	46
(Db-)	16	82	98
total	25	119	144

Instead of asking about the proportion of events, consider asking about the proportion of events within the pairs.

I have collected pairs in which one person has had a heart attack and the matched person has not.

I would like to test the null that diabetes is not associated with having a heart attack against the alternative that an association exists. (i.e., does diabetes occur (significantly) more often in the heart attack patient ?)

The pairs where both are diabetic or both are not (concordant pairs) provide no information. So, we consider the pairs where one person is diabetic and the other is not (discordant pairs). If there is no association, we expect the diabetes to be found with the heart attack patient about $1/2$ of the time and the diabetes to be found with the non-heart attack patient the other $1/2$ of the time.

Concordant pairs are pairs in which the outcome is the same. In this case, each is a diabetic or each is not.

Discordant pairs are the opposite. The responses of each pair member are opposite that of the other. In this case, one is diabetic and one is not.

	No Heart Attack		
Heart Attack	(Db+)	(Db-)	total
(Db+)	9	37	46
(Db-)	16	82	98
total	25	119	144

Consider the pairs where one person is diabetic and the other is not. Let r represent the number of pairs in which the heart attack victim is the diabetic and s represent the number of pairs in which the non heart attack victim is the diabetic. If the null is true, r and s should be approximately equal.

Under the null,

$$X^2 = \frac{(|r - s| - 1)^2}{r + s}$$

is approximately chi-square (χ^2) distributed with 1 degree of freedom.

	No Heart Attack		
Heart Attack	(Db+)	(Db-)	total
(Db+)	9	37	46
(Db-)	16	82	98
total	25	119	144

For this example,

$$X^2 = \frac{(21-1)^2}{37+16} = 7.55$$

The previous example was an illustration of McNemar's test. Recall that to conduct hypothesis tests on two independent proportions, we can use either a normal approximation to the binomial distribution or the chi square test. Alternatively, when the data are paired, we can use McNemar's test.

Consider one more example of comparing two proportions.

Suppose we want to compare two different chemotherapy protocols for breast cancer after mastectomy. The two treatment groups should be as comparable as possible on other prognostic factors. To accomplish this, we set up a matched study. 621 breast cancer patients are paired. A random member of each pair is assigned to treatment A; the other to treatment B. Survival after 5 years is the measured outcome.

	Outcome		
Treatment	(Survived)	(Died)	total
(A)	526	95	621
(B)	515	106	621
total	1041	201	1242

Note that the 5 year survival rate for treatment A is $526/621 = 0.847$ and for treatment B is $515/621 = 0.829$. You can check that the Yates corrected chi-square statistic is 0.59 with 1 degree of freedom, which is NOT significant.

SINCE THE DATA ARE PAIRED, THE CHI-SQUARE TEST IS NOT APPROPRIATE.

	Outcome Tr. B patient		
Outcome Tr. A patient	(Survived)	(Died)	total
(Survived)	510	16	526
(Died)	5	90	95
total	515	106	621

McNemar's test states that under the null

$$X^2 = \frac{(|r-s|-1)^2}{r+s}$$

is approximately chi-square (χ^2) distributed with 1 degree of freedom.

For this example,

$$X^2 = \frac{(11-1)^2}{5+16} = 4.76$$

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.

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 $\frac{ad}{bc}$ ($\hat{OR} = \frac{ad}{bc}$).

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 estimated to be $\frac{(358)(2745)}{(229)(2492)} = 1.72$

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

Consider two (unrelated) questions that I posed last time:

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

2. Is 1.72 significantly different than 1?

For the 2 x 2 table:

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

The odds ratio OR is estimated by $\hat{OR} = \frac{ad}{bc}$

The estimated standard error of the natural logarithm (ln) of this estimate is given by

$$\hat{se}(ln(\hat{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

An approximate 95 % confidence interval for $ln(\hat{OR})$ is

$$\left(\ln(\hat{OR}) - 1.96\hat{se}[ln(\hat{OR})], \ln(\hat{OR}) + 1.96\hat{se}[ln(\hat{OR})]\right)$$

An approximate 95 % confidence interval for $\hat{OR}$ is

$$\left(e^{\ln(\hat{OR}) - 1.96\hat{se}[ln(\hat{OR})]}, e^{\ln(\hat{OR}) + 1.96\hat{se}[ln(\hat{OR})]}\right)$$

For the C-section data:

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

The odds ratio OR is estimated by $\hat{OR} = \frac{ad}{bc} = 1.72$
and $ln(\hat{OR}) = ln(1.72) = 0.542$

The estimated standard error of the natural logarithm (ln) of this estimate is given by

$$\begin{aligned} \hat{se}(ln(\hat{OR})) &= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \\ &= \sqrt{\frac{1}{358} + \frac{1}{229} + \frac{1}{2492} + \frac{1}{2745}} \\ &= 0.089 \end{aligned}$$

An approximate 95 % confidence interval for $ln(\hat{OR})$ is

$$\left(\ln(\hat{OR}) - 1.96\hat{se}[ln(\hat{OR})], \ln(\hat{OR}) + 1.96\hat{se}[ln(\hat{OR})]\right)$$

which is

$$(0.542 - 1.96(0.089), 0.542 + 1.96(0.089)) = (0.368, 0.716)$$

An approximate 95 % confidence interval for $\hat{OR}$ is

$$\left(e^{0.368}, e^{0.716}\right) = (1.44, 2.05)$$

A similar approach can be taken to estimate the odds ratio following McNemar's test. The estimate of the odds ratio is r/s, not $(ad)/(bc)$. Confidence intervals can also be obtained.

The odds ratio is useful to measure the strenght of association between two dichotomous variables. However, when drawing conclusions based on calculations of the odds ratio (or other quantity for that matter), care must be taken. For example, a sample from one hospital might not provide information about the population at large. When sampling from a restricted population (one hospital), one can often show relationships between variables that do not in fact exist in the general population. This is an illustration of *Berkson's fallacy*.

Suppose for example that you obtain information on 257 hospitalized patients concerning respiratory and circulatory disease. Using this information, you record the following:

	Respiratory Disease		
Circulatory Disease	(Yes)	(No)	total
(Yes)	7	29	36
(No)	13	208	221
total	20	237	257

You could perform a chi-square test and if you did, you would reject the null and conclude (for $\alpha = 0.05$) that there is an association between respiratory and circulatory disease.

If you instead sampled from the population at large and recorded the same information to obtain

	Respiratory Disease		
Circulatory Disease	(Yes)	(No)	total
(Yes)	22	171	193
(No)	202	2389	2591
total	224	2560	2784

you would not reject the null using a chi-square test with $\alpha = 0.05$. The difference occurs because the rates of case prevalence are different in the two populations.