Data Analysis and Statistical Methods
Statistics 651

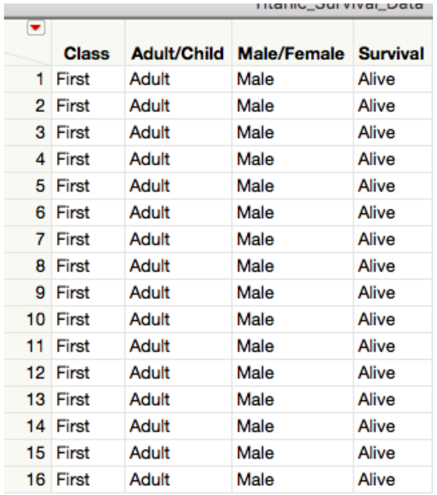
http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 26 (MWF) Tests and CI based on two proportions

Suhasini Subba Rao

Comparing proportions in two populations

- Consider the following data set. It gives the surivival of each person on the titanic (survived/not survive), their gender (male/female), class (first/second/third) and whether they are an adult or child.



Titanic_Survival_Data

	Class	Adult/Child	Male/Female	Survival
1	First	Adult	Male	Alive
2	First	Adult	Male	Alive
3	First	Adult	Male	Alive
4	First	Adult	Male	Alive
5	First	Adult	Male	Alive
6	First	Adult	Male	Alive
7	First	Adult	Male	Alive
8	First	Adult	Male	Alive
9	First	Adult	Male	Alive
10	First	Adult	Male	Alive
11	First	Adult	Male	Alive
12	First	Adult	Male	Alive
13	First	Adult	Male	Alive
14	First	Adult	Male	Alive
15	First	Adult	Male	Alive
16	First	Adult	Male	Alive

- One can ask if gender had an impact on survival. Did females have a higher chance of survival than males?

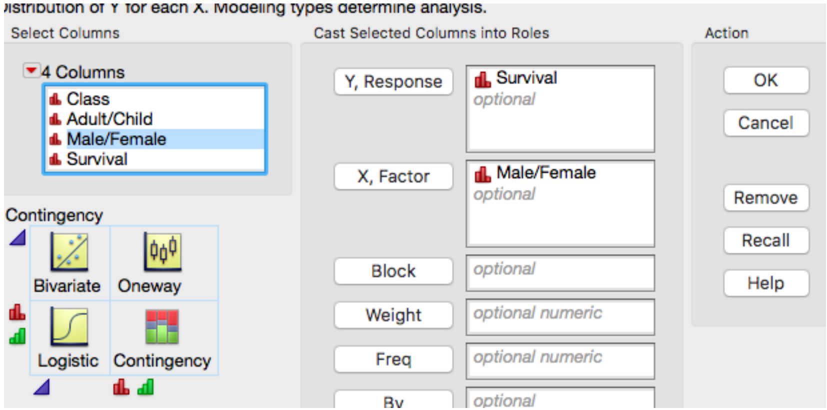
- We articulate this as a hypothesis test. Let p_F denote the survival rate amongst females and p_M the survival rate amongst males.

  The hypotheses of interest is $H_0 : p_F - p_M \leq 0 \ H_A : p_F - p_M > 0$.

- This is a two sample test for proportions, since we are comparing the proportions in two different populations (in this case male and female).

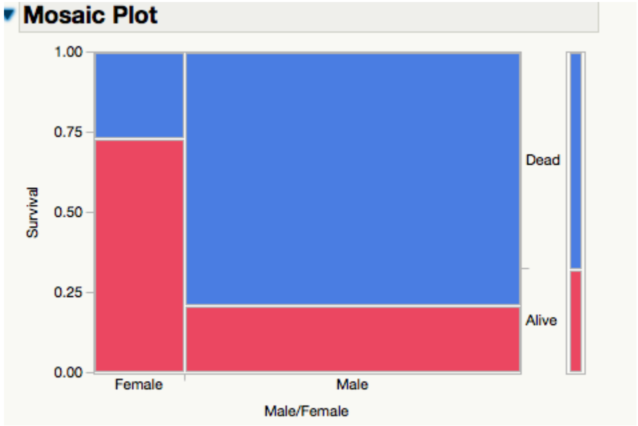
- We explain how to do the test in JMP and interprete the output.

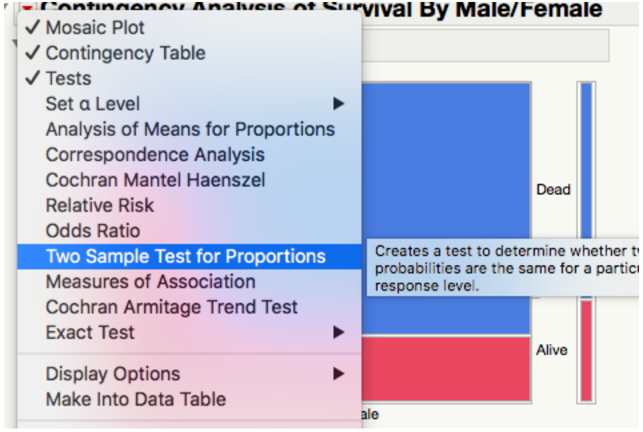
Ensure both the variables of interest, gender and survival described as Nominal.
Go to Analyze > Fit Y by X.
You will see the mosaic plot, similar to the one below.



Y, Response: Survival (optional)

X, Factor: Male/Female (optional)



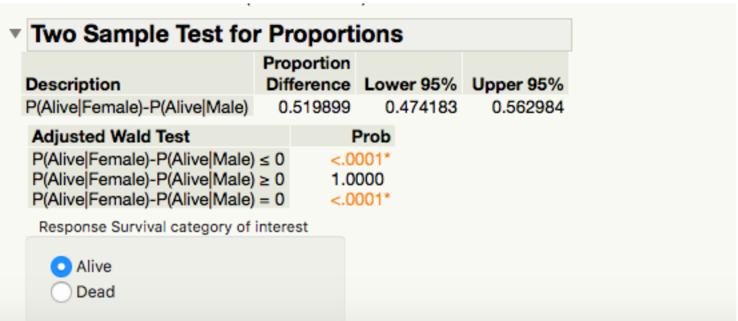


Creates a test to determine whether t
probabilities are the same for a partic
response level.

Press on the red triangle and select Two sample Test for proportions.

Interpreting the output

- This is a snapshot of what you will observe



Two Sample Test for Proportions

Description	Proportion Difference	Lower 95%	Upper 95%
P(Alive|Female)-P(Alive|Male)	0.519899	0.474183	0.562984

Adjusted Wald Test	Prob
P(Alive|Female)-P(Alive|Male) ≤ 0	<.0001*
P(Alive|Female)-P(Alive|Male) ≥ 0	1.0000
P(Alive|Female)-P(Alive|Male) = 0	<.0001*

Response Survival category of interest

● Alive
○ Dead

$$H_0 : p_F - p_M \leq 0 \ H_A : p_F - p_M > 0$$
$$H_0 : p_F - p_M \geq 0 \ H_A : p_F - p_M < 0$$
$$H_0 : p_F - p_M = 0 \ H_A : p_F - p_M \neq 0$$

- The blue button by "Alive" means the test is based on the proportion which had survived. If the blue buttom by "Dead" were highlighted this means the test should be based on the proportion on the titanic that had died (in which case the one-sided test changes direction).

- The sample proportion difference is $\widehat{p}_F - \widehat{p}_M = 0.52$.

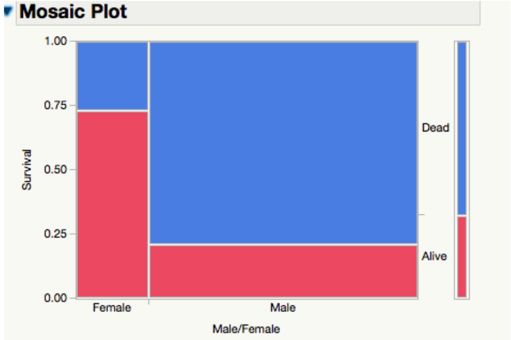
- The output gives the results of both the one-sided tests and the two-sided test.

- Observe that unlike the t-tests, JMP states the null hypothesis in the three cases.

- Out focus is on the one-sided test $H_0 : p_F - p_M \leq 0$ $H_A : p_F - p_M > 0$, which corresponds to the top option.

  The p-value is less than 0.01%, which seems obvious given the sample size of several thousand and the large difference between the proportions.

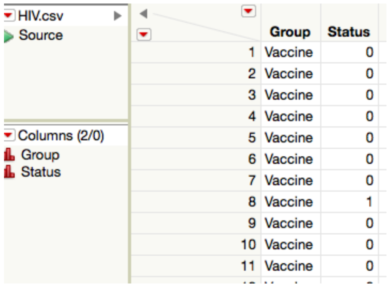
- Therefore there is substantial evidence to suggest that the proportion of females that survived was greater than the proportion of males. The differences seen below in the data cannot be explained by sampling differences.



- The 95% confidence interval for the difference in the survival proportions is $[0.47, 0.57]$.

The Thai HIV Vaccine Trial

In 2006, drug trials in Thailand were done for the vaccine against HIV. The spreadsheet is below.

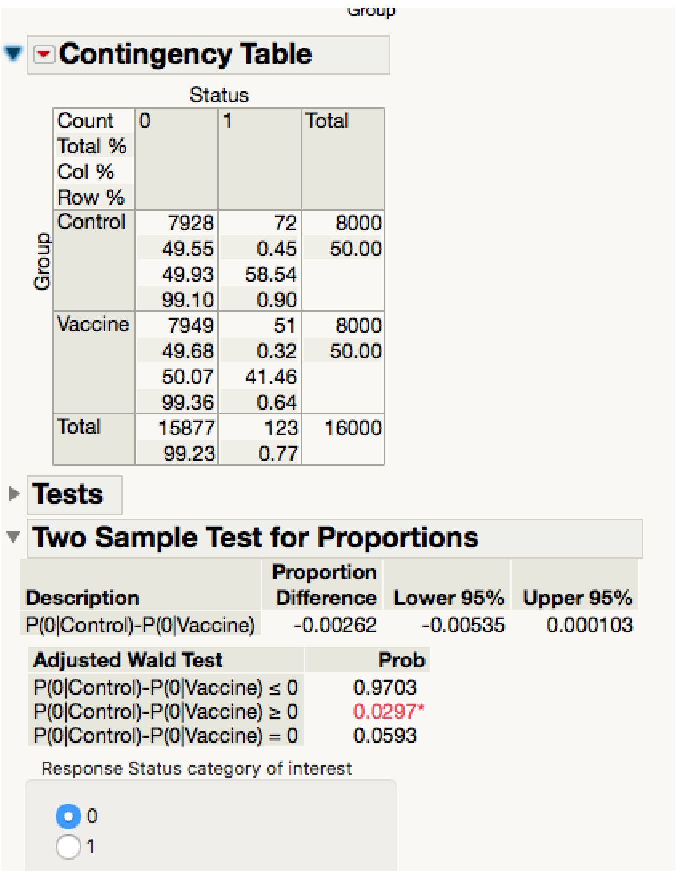


	Group	Status
1	Vaccine	0
2	Vaccine	0
3	Vaccine	0
4	Vaccine	0
5	Vaccine	0
6	Vaccine	0
7	Vaccine	0
8	Vaccine	1
9	Vaccine	0
10	Vaccine	0
11	Vaccine	0

We want to test the hypothesis that the vaccine gave some protection. This would mean that the proportion of people in the entire population who take the vaccine and go on to develop HIV is less than the proportion of the entire population who do not take the vaccine and go on to develop HIV.

We test $H_0 : p_V - p_P \geq 0$ against $H_A : p_V - p_P < 0$ (ie. the people who took the vaccine are at less risk of infection).

The data estimates $\hat{p}_V = 0.0065$ and $\hat{p}_P = 0.009$ and $\hat{p}_V - \hat{p}_P = -0.002625$, is this slight difference (of 0.26%) statistically significant?



Group

Contingency Table

Status

Group	Count Total % Col % Row %	0	1	Total
	Control	7928 49.55 49.93 99.10	72 0.45 58.54 0.90	8000 50.00
	Vaccine	7949 49.68 50.07 99.36	51 0.32 41.46 0.64	8000 50.00
	Total	15877 99.23	123 0.77	16000

Tests

Two Sample Test for Proportions

Description	Proportion Difference	Lower 95%	Upper 95%
P(0|Control)-P(0|Vaccine)	-0.00262	-0.00535	0.000103

Adjusted Wald Test	Prob
P(0|Control)-P(0|Vaccine) ≤ 0	0.9703
P(0|Control)-P(0|Vaccine) ≥ 0	0.0297*
P(0|Control)-P(0|Vaccine) = 0	0.0593

Response Status category of interest

● 0
○ 1

The test we require is the middle option $H_0 : p_V - p_P \geq 0$ against $H_A : p_V - p_P < 0$.

The p-value is 2.97%. The data suggests that the vaccine may have had some protective effect. Though further research was required to see if there was any merit to this claim.

In the next few slides we explain the mechanics behind the test (using Statcrunch).

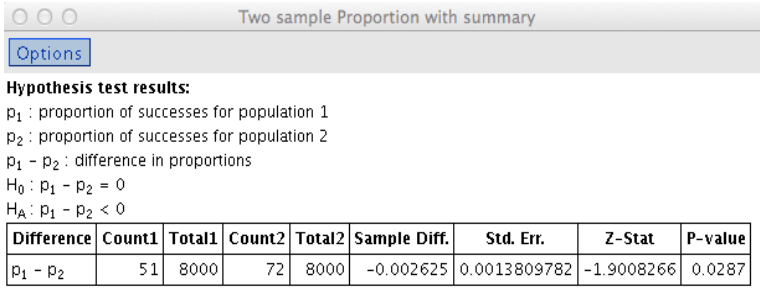
The Thai HIV Vaccine Trials

In 2006, drug trials in Thailand were done for the vaccine against HIV.

	number of people infected	No affected	Sample size
HIV vaccine	51	7949	8000
Placebo	72	7928	8000

We want to test the hypothesis that the vaccine gave some protection. This would mean that the proportion of people in the entire population who take the vaccine and go on to develop HIV is less than the proportion of the entire population who do not take the vaccine and go on to develop HIV. Hence we want to test $H_0 : p_V - p_P \geq 0$ against $H_A : p_V - p_P < 0$ (ie. the people who took the vaccine are at less risk of infection). The data estimates $\hat{p}_V = 0.0065$ and $\hat{p}_P = 0.009$ and $\hat{p}_V - \hat{p}_P = -0.002625$, is this slight difference (of 0.26%) statistically significant?

The analysis



Hypothesis test results:

p_1 : proportion of successes for population 1

p_2 : proportion of successes for population 2

$p_1 - p_2$: difference in proportions

$H_0 : p_1 - p_2 = 0$

$H_A : p_1 - p_2 < 0$

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	51	8000	72	8000	-0.002625	0.0013809782	-1.9008266	0.0287

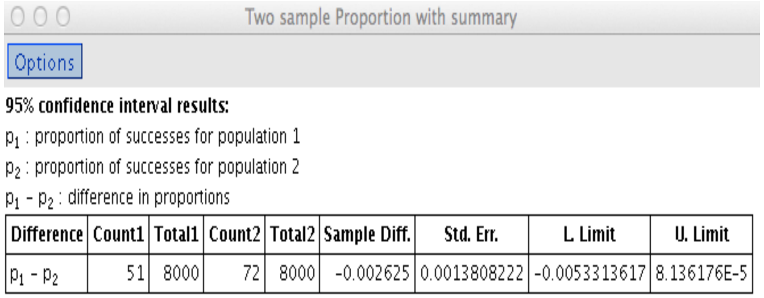
- The main point is that the standard error is relatively small, 0.000138 (we calculate this later). This z-value

$$z = \frac{-0.002625}{0.00138} = -1.90.$$

The area to the LEFT of -1.9 on the z-tables is 2.87%. If we use the significance level of 5% this result is statistically significant, however, if we use 1% as the significance level it's not.

- Therefore, there is some evidence to suggest that the vaccination may offer some protection against HIV.

- If want to measure the degree of protection we can construct a CI for the difference



95% confidence interval results:

p_1 : proportion of successes for population 1

p_2 : proportion of successes for population 2

$p_1 - p_2$: difference in proportions

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	L. Limit	U. Limit
$p_1 - p_2$	51	8000	72	8000	-0.002625	0.0013808222	-0.0053313617	8.136176E-5

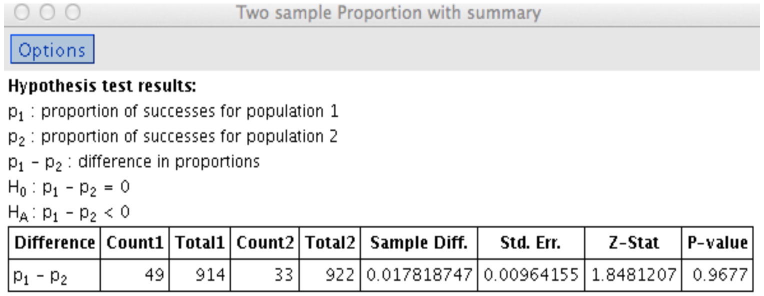
The SA HIV Vaccine Trials

In 2005, drug trials in SA (amonst males) were done for the vaccine against HIV.

	number of people infected	No affected	Sample size
HIV vaccine	49	865	914
Placebo	33	889	922

We want to test the hypothesis that the vaccine gave some protection. This would mean that the proportion of people in the entire population who take the vaccine and go on to develop HIV is less than the proportion of the entire population who do not take the vaccine and go on to develop HIV. Hence we want to test $H_0 : p_V - p_P \geq 0$ against $H_A : p_V - p_P < 0$ (ie. the people who took the vaccine are at less risk of infection). The data estimates $\hat{p}_V = 0.053$ and $\hat{p}_P = 0.035$ and $\hat{p}_V - \hat{p}_P = 0.0178$. It is immediately clear that there is no evidence in the data that the vaccine works (in this cohort).

The analysis



Hypothesis test results:

p_1 : proportion of successes for population 1

p_2 : proportion of successes for population 2

$p_1 - p_2$: difference in proportions

$H_0 : p_1 - p_2 = 0$

$H_A : p_1 - p_2 < 0$

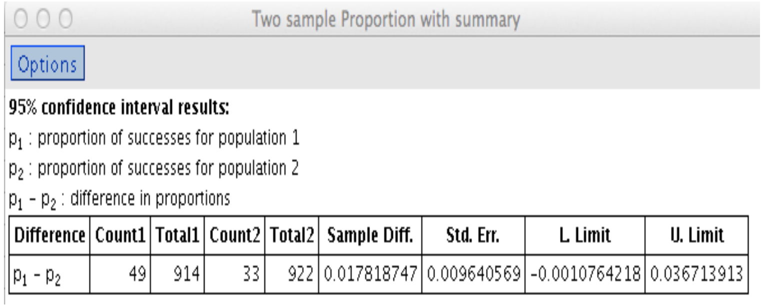
Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	49	914	33	922	0.017818747	0.00964155	1.8481207	0.9677

- The main point $\hat{p}_V - \hat{p}_P = 0.0178$ is positive and we are looking for a significant negative difference. We wanted to do the test. The z-value is

$$z = \frac{0.0178}{0.0096} = 1.848.$$

The area to the LEFT of 1.848 on the z-tables is 96.7%. There is no evidence to against the null. We cannot reject the null.

- If want to measure the degree of difference we can construct a CI for the difference



95% confidence interval results:

p_1 : proportion of successes for population 1

p_2 : proportion of successes for population 2

$p_1 - p_2$: difference in proportions

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	L. Limit	U. Limit
$p_1 - p_2$	49	914	33	922	0.017818747	0.009640569	-0.0010764218	0.036713913

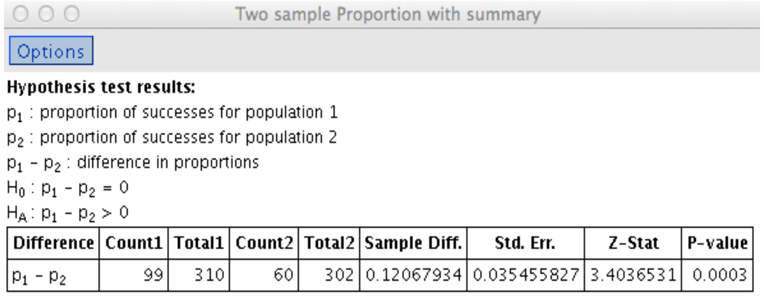
Hair remedies

The FDA approved the drug Minodixil as a remedy for male pattern baldness. They did a study and this is what they found:

	new hair growth	no hair growth	Sample size
Minodixil	99	211	310
Placebo	60	242	302

Let π_M be the probability a person has new hair growth and uses Minodixil and π_P be the probability a person has new hair growth and and doesn't use Minodixil. Our estimates are $\hat{p}_M = 0.32$ and $\hat{p}_P = 0.2$. Use this data to test $H_0 : p_M - p_P \leq 0$ against the alternative $H_A : p_M - p_P > 0$. The data gives an estimated difference $\hat{p}_M - \hat{p}_P = 0.12$, is the 12% difference seen in the data statistically significant?

The analysis



Hypothesis test results:

p_1 : proportion of successes for population 1

p_2 : proportion of successes for population 2

$p_1 - p_2$: difference in proportions

$H_0 : p_1 - p_2 = 0$

$H_A : p_1 - p_2 > 0$

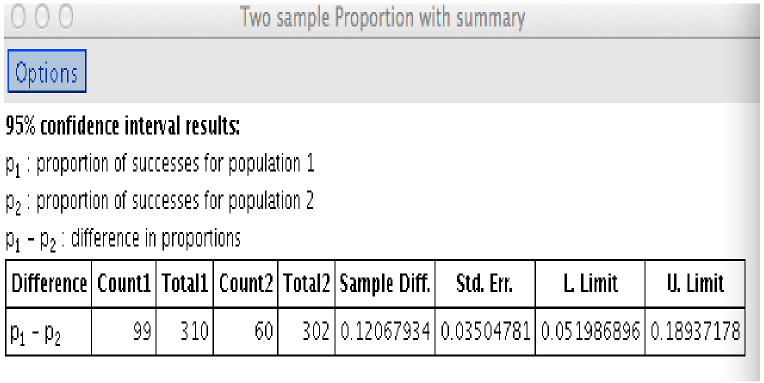
Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	99	310	60	302	0.12067934	0.035455827	3.4036531	0.0003

- The main point $\hat{p}_M - \hat{p}_P = 0.12$ is positive and we are looking for a *significant* positive difference. We do a two sample test on proportions. The z-value is

$$z = \frac{0.12}{0.035} = 3.40.$$

The area to the RIGHT of 3.40 on the z-tables is 0.03%. As this is less than 5%, there IS evidence that minidoxil reduces hair loss.

- If want to measure how much better Minodoxil is over the placebo we can construct the CI



95% confidence interval results:

p_1 : proportion of successes for population 1

p_2 : proportion of successes for population 2

$p_1 - p_2$: difference in proportions

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	L. Limit	U. Limit
$p_1 - p_2$	99	310	60	302	0.12067934	0.03504781	0.051986896	0.18937178

- Compare the standard errors of the test statistic with those used in constructing the confidence interval and we see that they are different.

  This is because like the one sample proportion procedures, they are constructed under different conditions. We explain why below.

The normality result

- The p-values and confidence intervals constructed above were derived under the assumption that the estimate for the difference in proportions is normally distributed

$$(\hat{p}_1 - \hat{p}_2) \overset{\mathcal{D}}{\rightarrow} \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}\right).$$

- The above result holds ‘roughly’ true if the number of successes and failures in both groups is greater than 5 (this basically ensures the distributions are not too asymmetric about the mean of the distribution - just as in the one sample case).

- We then can almost plug and chug. There is however, one problem p_1 and p_2 are unknown.

- These need to be estimated - but the eagle eyed may have noticed that the standard errors are different for testing and the confidence intervals.

- Importantly, we do not use the t-distribution in any of the calculations, we only use the normal distribution.

The standard error for constructing the test

- Let us return to the hair remedy example. We want to test H_0 : $p_M - p_P \leq 0$ against $H_A : p_M - p_P > 0$.

	new hair growth	no hair growth	Sample size
Minodixil	99	211	310
Placebo	60	242	302

- Remember we want to calculate the chance of the data giving a difference of 12% (0.12) when placebo and Minodixil have exactly the same effect of hair. The standard error is

$$\sqrt{\frac{p_M(1-p_M)}{310} + \frac{p_P(1-p_P)}{302}}$$

but the placebo and Minidoxil have same effect ($p_M = p_P = p$), then

$$\sqrt{\frac{p_M(1-p_M)}{310} + \frac{p_P(1-p_P)}{302}} = \sqrt{p(1-p)\left(\frac{1}{310} + \frac{1}{302}\right)}.$$

- Now we need to find the 'best estimator' of p. The larger the sample size the the more reliable (and better) and estimator.
- Since under the null there is NO difference between the placebo or Minodixil we can 'pool' the data.

	new hair growth	no hair growth	Sample size
Minodixil	99	211	310
Placebo	60	242	302
Pooled	159	453	612

- If there is no gain from using Minodixil, the proportion of of people who notice a difference by simply massaging the head would be $\hat{p} = 159/612 = 0.260$. We use this as our best estimator of p (under the null). The standard error under the null:

$$s.e = \sqrt{0.26 \times 0.74 \left(\frac{1}{310} + \frac{1}{302} \right)} = 0.0354.$$

- This is quite important. If we have more information about the data we need to pool it to obtain a better estimator - as we have done above.

The standard error for constructing confidence intervals

- We return to the normality result:

$$(\hat{p}_1 - \hat{p}_2) \overset{\mathcal{D}}{\rightarrow} \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}\right).$$

- Using this result the 95% CI for the difference in proportions is

$$\left[\widehat{p}_1 - \widehat{p}_2 \pm 1.96 \times \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}\right].$$

  Application to hair remedy data:

$$\left[0.12 \pm 1.96 \times \sqrt{\frac{p_M(1-p_M)}{310} + \frac{p_P(1-p_P)}{302}}\right].$$

- But we do not know p_M or p_P. Unlike testing we do not make any assumptions about it under the null. Therefore we simply replace p_M and p_P with their estimators $\widehat{p}_M = 0.32$ and $\widehat{p}_P = 0.2$ to give

$$\left[0.12 \pm 1.96 \times \sqrt{\frac{0.32 \times 0.68}{310} + \frac{0.2 \times 0.8}{302}}\right] = [0.051, 0.189] = [5.1, 18.9]\%.$$

Relative Risk

- In medical data and other applications the relative risk is often considered. We illustrate this through the HIV example:

	number of people infected	Sample size	
HIV vaccine	51	8000	0.006375
Placebo	72	8000	0.009

- We may ask how much more risk is there is taking the Placebo over the vaccine, this is best measured by the ratio

$$RR = \frac{0.009}{0.006375} = 1.4.$$

- Ie. The data suggests that you are 1.4 times more likely to develop HIV if you don't take the vaccine than if you do.

- However, caution needs to be used when interpreting 1.4. 1.4 has been calculated from the $sample$. This is an estimate of the relative risk based on the $population$. Therefore confidence intervals need to obtained for the population RR. If these were constructed, one found that the interval is wide with the left hand side end overlapping 1. This would be mean we have to be very cautious about interpreting the factor 1.4 as a gain in using a vaccine.