Fisher's Exact Test

Fisher's Exact Test

A woman at the tea party claimed she could tell whether the milk was poured into her tea first (before the tea) or second (after the tea was poured). Fisher devised a test on the spot using 8 tea cups - 4 with the milk poured first and 4 with the milk poured second. The lady and Fisher and all other parties knew this arrangement.

Fisher's Exact Test

A woman at the tea party claimed she could tell whether the milk was poured into her tea first (before the tea) or second (after the tea was poured). Fisher devised a test on the spot using 8 tea cups - 4 with the milk poured first and 4 with the milk poured second. The lady and Fisher and all other parties knew this arrangement.

The interpretation of Fisher's exact test to other testing situations is that it is a conditional test where what you are given are the marginal totals. You might wonder why one would want such a test and p-values since surely your own marginals are not (both) fixed in advanced. However, my recommendation is to use Fisher's test whenever you have small sample sizes: it is reasonably conservative, has been proven to work well, and Fisher is a reliable genius in the murky science that is statistics.

Fisher's Exact Test - example

Here are the results:

	Reality		
Guess	Milk first	Milk second	
Milk first	3	1	4
Milk Second	1	3	4
	4	4	8

Fisher's Exact Test -theory

Fisher's exact test is based on the hypergeometric distribution, typically described as a ball and urn problem. An urn contains A black balls and B white balls. You draw out n without replacement. What is the probability you draw k black balls?

$$P(k) = \frac{\binom{A}{k}\binom{B}{n-k}}{\binom{A+B}{n}}$$

Fisher's Exact Test - example continued...

Now, we just have to figure out what all these balls and colors are. We can do it either way, but suppose the colors are reality - that is A is the 4 cups of tea with the milk really poured first and B is the 4 cups of tea with the milk really poured second. Then n is the 4 cups of tea the lady chose to guess that the milk is poured first (the other 4 have to then be her guess for the milk being poured second.) So what is the probability she could have guessed 3 or more correctly?

Fisher's Exact Test - example continued...

Now, we just have to figure out what all these balls and colors are. We can do it either way, but suppose the colors are reality - that is A is the 4 cups of tea with the milk really poured first and B is the 4 cups of tea with the milk really poured second. Then n is the 4 cups of tea the lady chose to guess that the milk is poured first (the other 4 have to then be her guess for the milk being poured second.) So what is the probability she could have guessed 3 or more correctly?

$$P(3) + P(4) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{17}{70} = .2429$$

Fisher's Exact Test - example continued...

Now, we just have to figure out what all these balls and colors are. We can do it either way, but suppose the colors are reality - that is A is the 4 cups of tea with the milk really poured first and B is the 4 cups of tea with the milk really poured second. Then n is the 4 cups of tea the lady chose to guess that the milk is poured first (the other 4 have to then be her guess for the milk being poured second.) So what is the probability she could have guessed 3 or more correctly?

$$P(3) + P(4) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{17}{70} = .2429$$

Fisher concluded that the lady could not tell which was poured first. The lady, had she been statistically sophisticated, could have complained about the lack of power of the test given the very small sample size.

Simpson's Paradox

Simpson's Paradox

In the 1970's, it was observed that graduate programs at Berkeley, on the whole, admitted significantly fewer women than they did men. See the following table for the results for one time period. The difference is statistically significant.

Simpson's Paradox

In the 1970's, it was observed that graduate programs at Berkeley, on the whole, admitted significantly fewer women than they did men. See the following table for the results for one time period. The difference is statistically significant.

	Applied	Accepted
Men	8442	44%
Women	4321	35%

Simpson's Paradox

In the 1970's, it was observed that graduate programs at Berkeley, on the whole, admitted significantly fewer women than they did men. See the following table for the results for one time period. The difference is statistically significant.

	Applied	Accepted
Men	8442	44%
Women	4321	35%

Test:

$$Z = \frac{0.44 - 0.35}{\sqrt{\frac{(0.44)(0.56)}{8442} + \frac{(0.35)(0.65)}{4321}}} = 9.95$$

Which department was discriminating?

Which department was discriminating?

Department	Men		Women	
	# Applicants	% Accepted	# Applicants	% Accepted
A	825	62%	108	82%

Which department was discriminating?

Department	Men		Women	
	# Applicants	% Accepted	# Applicants	% Accepted
A	825	62%	108	82%
B	560	63%	25	68%

Which department was discriminating?

Department	Men		Women	
	# Applicants	% Accepted	# Applicants	% Accepted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%

Which department was discriminating?

Department	Men		Women	
	# Applicants	% Accepted	# Applicants	% Accepted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%

Which department was discriminating?

Department	Men		Women	
	# Applicants	% Accepted	# Applicants	% Accepted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%

Which department was discriminating?

Department	Men		Women	
	# Applicants	% Accepted	# Applicants	% Accepted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Which department was discriminating?

Department	Men		Women	
	# Applicants	% Accepted	# Applicants	% Accepted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

The Mantel Haenszel Test

The Mantel Haenszel Test

- The Mantel Haenszel Test allows you to combine the results of several Chi-Square tables in one analysis and thus avoid the problem of pooling results from different populations.

The Mantel Haenszel Test

- The Mantel Haenszel Test allows you to combine the results of several Chi-Square tables in one analysis and thus avoid the problem of pooling results from different populations.

- It is a test of independence of two binary categories conditional on a third category (strata, such as department.)

The Mantel Haenszel Test

- The Mantel Haenszel Test allows you to combine the results of several Chi-Square tables in one analysis and thus avoid the problem of pooling results from different populations.
- It is a test of independence of two binary categories conditional on a third category (strata, such as department.)
- It works well if the marginals in each table are large OR if there are a lot of tables.

The Mantel Haenszel Test

The Mantel Haenszel Test

1. Choose one square for the analysis. The upper left hand square will work. Use this square in all your tables.

The Mantel Haenszel Test

1. Choose one square for the analysis. The upper left hand square will work. Use this square in all your tables.
2. Calculate the Excess for this square. That is, calculated the observed minus the expected values for this square.

The Mantel Haenszel Test

1. Choose one square for the analysis. The upper left hand square will work. Use this square in all your tables.
2. Calculate the Excess for this square. That is, calculated the observed minus the expected values for this square.
3. Calculate the variance as $\frac{R_1 R_2 C_1 C_2}{T^2(T-1)}$ where R_1 and R_2 are the row totals, C_1 and C_2 are the column totals, and T is the total total.

The Mantel Haenszel Test

1. Choose one square for the analysis. The upper left hand square will work. Use this square in all your tables.
2. Calculate the Excess for this square. That is, calculated the observed minus the expected values for this square.
3. Calculate the variance as $\frac{R_1 R_2 C_1 C_2}{T^2(T-1)}$ where R_1 and R_2 are the row totals, C_1 and C_2 are the column totals, and T is the total total.
4. Sum the excesses. The standard error is the square root of the sum of the variances. Use this to form a standard normal test statistic.

Mantel Haenszel Test – continued

The set up is that you will have multiple tables like the following:

	Column 1	Column 2	
Row 1	$n_{1,1}$	$n_{1,2}$	R_1
Row 2	$n_{2,1}$	$n_{2,2}$	R_2
	C_1	C_2	T

Mantel Haenszel Test – continued

The set up is that you will have multiple tables like the following:

	Column 1	Column 2	
Row 1	$n_{1,1}$	$n_{1,2}$	R_1
Row 2	$n_{2,1}$	$n_{2,2}$	R_2
	C_1	C_2	T

If you want the pooled odds ratio, find $\frac{\sum \frac{n_{1,1}n_{2,2}}{T}}{\sum \frac{n_{1,2}n_{2,1}}{T}}$ where the sum is over all the individual tables.

Mantel Haenszel Test – same-sex twin example

Sex	zygosity	NRH	RH	Excess	Variance
Male	MZ	46	431	$46 - \frac{(477)(75)}{932}$ $= 7.6$	$\frac{(477)(455)(75)(857)}{((932^2)(931))}$ $= 17.25$
Female	MZ	29	426		

Mantel Haenszel Test – same-sex twin example

Sex	zygosity	NRH	RH	Excess	Variance
Male	MZ	46	431	$46 - \frac{(477)(75)}{932} = 7.6$	$\frac{(477)(455)(75)(857)}{((932^2)(931))} = 17.25$
Female	MZ	29	426		
Male	DZ	73	743	$73 - \frac{(816)(123)}{1577} = 9.4$	$\frac{(816)(761)(123)(1454)}{((1577^2)(1576))} = 27.96$
Female	DZ	50	711		
Totals				17.0	45.21

Mantel Haenszel Test – same-sex twin example

Sex	zygosity	NRH	RH	Excess	Variance
Male	MZ	46	431	$46 - \frac{(477)(75)}{932} = 7.6$	$\frac{(477)(455)(75)(857)}{((932^2)(931))} = 17.25$
Female	MZ	29	426		
Male	DZ	73	743	$73 - \frac{(816)(123)}{1577} = 9.4$	$\frac{(816)(761)(123)(1454)}{((1577^2)(1576))} = 27.96$
Female	DZ	50	711		
Totals				17.0	45.21

$$Z = 17.0/\sqrt{45.21} = 2.53 \text{ with one-sided p-value } = 0.0057.$$

This analysis is bad: it ignores obvious pairing in the data.

McNemar's Test

McNemar's test provides a way to handle paired binary response data.

McNemar's Test

McNemar's test provides a way to handle paired binary response data.

The data come from a study by Johnson and Johnson. It was a follow up study on whether tonsillectomies increase the risk of contracting Hodgkin's disease. They collected data from 85 sibling pairs: one of the siblings had Hodgkin's and the other sibling was within 5 years of age but was disease free. The original paper analyzed the data with the following 2×2 table:

McNemar's Test

	Tonsillectomy	No tonsillectomy
Hodgkin's	41	44
Control	33	52

McNemar's Test

	Tonsillectomy	No tonsillectomy
Hodgkin's	41	44
Control	33	52

which lead to a χ^2 statistic of 1.53 which is quite insignificant. However, this analysis ignores the pairing of the data. Ignoring the pairing inflates the variance and leads to a lack of power. The analysis that addresses the pairing is McNemar's test which analyzes the following 2×2 table instead:

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

The null hypothesis in this case is the tonsillectomy's are just as likely in the Hodgkin's group as in the control group.

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

The null hypothesis in this case is the tonsillectomy's are just as likely in the Hodgkin's group as in the control group. Since there are equal numbers of Hodgkin's patients and controls, this would mean the second row total is the same as the second column total.

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

The null hypothesis in this case is the tonsillectomy's are just as likely in the Hodgkin's group as in the control group. Since there are equal numbers of Hodgkin's patients and controls, this would mean the second row total is the same as the second column total. Equivalently, the first row total should be the same as the first column total.

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

The null hypothesis in this case is the tonsillectomy's are just as likely in the Hodgkin's group as in the control group. Since there are equal numbers of Hodgkin's patients and controls, this would mean the second row total is the same as the second column total. Equivalently, the first row total should be the same as the first column total. These equalities will hold exactly when the off-diagonal entries are equal.

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

The null hypothesis in this case is the tonsillectomy's are just as likely in the Hodgkin's group as in the control group. Since there are equal numbers of Hodgkin's patients and controls, this would mean the second row total is the same as the second column total. Equivalently, the first row total should be the same as the first column total. These equalities will hold exactly when the off-diagonal entries are equal. Currently they are in proportion 7:15. Is that statistically equal?

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

McNemar's Test

Hodgkin's patient	Disease-free Control Sibling	
	No tonsillectomy	Tonsillectomy
No tonsillectomy	37	7
Tonsillectomy	15	26

The test is a Chi-Square test with 1 degrees of freedom but it is calculated as:

$$\chi^2 = \frac{(n_{1,2} - n_{2,1})^2}{n_{1,2} + n_{2,1}} = \frac{(7-15)^2}{7+15} = 2.91$$

which has a p-value of 0.09 which is suggestive of tonsillectomies being associated with Hodgkin's disease.