Lecture #9

- **Chisquare: Comparisons between proportions or percentages**
  - **Research questions about two or more separate or independent groups**
  - **Research questions about two dependent or correlated groups**

What is proportion & percentage?

- A fraction in which the numerator is included in the denominator (part/total)
- Dimensionless (no units of measurement)
- Values range between 0 and 1
- Can also be expressed as a percentage

Two-way tables

When data is categorical.

Data may be summarized like:

	Category 1	Category 2
Category A	Number in Category A and Category 1	Number in Category A and Category 2
Category B	Number in Category B and Category 1	Number in Category B and Category 2

Two-way tables, cont.

The cells of the table "cross-tabulate" the number of cases having particular joint values of the two distributions.

The **marginal distributions** are the total number of observations for a given category (either summed across rows or columns).

When you use a cross-tab, you want to learn whether or not the rows and columns are related (statistically independent).

	Category 1	Category 2	Row Marginals
Category A	Number in Category A and Category 1	Number in Category A and Category 2	Number in Category A
Category B	Number in Category B and Category 1	Number in Category B and Category 2	Number in Category B
Column Marginals	Number in Category 1	Number in Category 2	Total number of observations

Two-way tables, example

The table to the right is a sample "cross-tab"

Your research hypothesis is that dog ownership and gender are related.

How do you test this hypothesis?

	Dog-Owners	No Pets	Totals
Men	100	400	500
Women	50	450	500
Totals	150	850	1,000

Pearson's Chi-Square Test

- When analysis of categorical data is concerned with two variables, two-way tables (also known as ***contingency tables***) are employed.
- The chi-square test provides a method for testing the association between the row and column variables in a two-way table, ie: to test whether or not there is a relationship between ***two*** categorical (nominal) variables
- Each individual in the sample is classified on two separate variables.
- The null hypothesis H_o states that **there is NO relationship** between the variables (one variable does not vary according to the other variable).
- The alternative H_a claims that some relationship exists. It does not specify the **type** of association.
- The chi-square test is based on a test statistic that measures the divergence of the **observed** data from the values that would be **expected** under the null hypothesis of no association. This requires calculation of expected values based on the data.

Chi-Square test statistic

- The test statistic that makes the comparison is the ***chi-square statistic***
- The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{observed count } - \text{ expected count})^2}{\text{expected count}}$$

The sum is over all $r \times c$ cells in the table

- With (r-1)*(c-1) degree of freedom
- We will reject Ho if the value of the chi-square statistic is too large (why?)

Hypothetical example: to explain expected frequencies

- The totals (marginal frequencies) represent a hypothetical study in which 200 patients receive a treatment and 100 receive a placebo. 75 patients respond positively; the remaining 225 patients respond negatively.
- Proportion responding positively= 75/300, p=0.25
- If treatment is not effective, or response is independent of treatment, we expect:

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

Marginal frequencies

	Treatment	Control	Total
Positive	**50**	**25**	75
Negative	**150**	**75**	225
Total	200	100	300

Basis of Chi-square test

- Based on comparison of “observed-to-expected” frequencies

- If the difference between the observed and expected number is “small”, then there is no relationship

- If difference is “big”, there is likely an association

Comparing 2 proportions; examples

- More interested in comparing 2 or more proportions
  - e.g. Which of two drugs has a higher % of cures ?
  - e.g. Did Hispanic children have a lower % of prenatal care than non-Hispanic children ?

QUESTION: Is there a difference between the proportion of drinkers among cases and controls?

Group 1	Group 2
Disease	No Disease
P1= proportion of drinkers	P2= proportion of drinkrs

Chi-Square Assumptions

- **Random sample data** are assumed.
- **Independence**. Observations must be independent and mutually exclusive.
- **A sufficiently large sample size** is assumed, as in all significance tests. Applying chi-square to small samples exposes the researcher to an unacceptable rate of Type II errors.
- **Adequate cell sizes** are also assumed.
- Data are nominal or ordinal levels.

Chi-square Step-by-Step

1) Formulate Hypotheses

2) Calculate row and column totals

3) Calculate row and column proportions

4) Calculate expected frequencies (E_i)

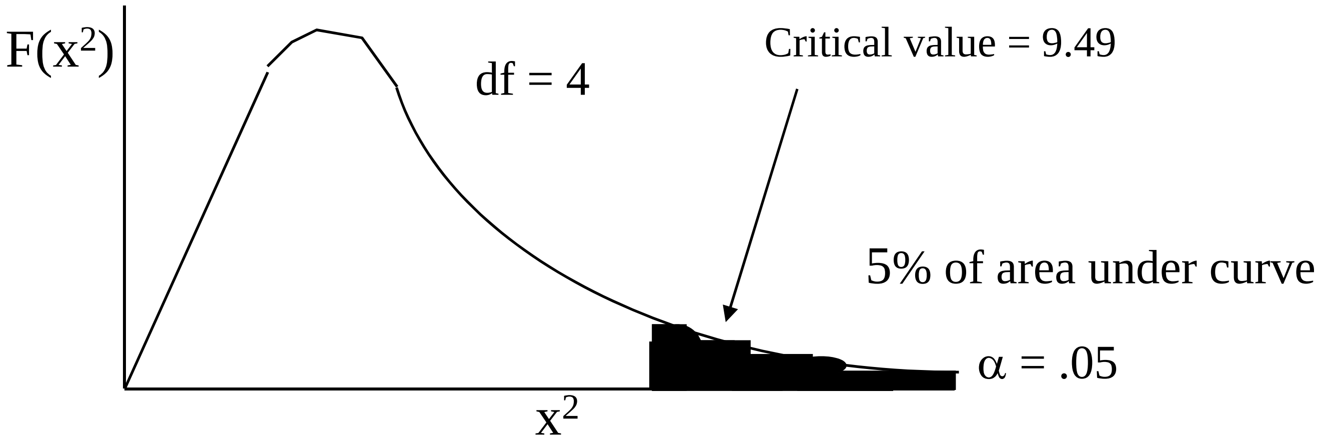
5) Calculate χ^2 statistic

6) Calculate degrees of freedom

7) Obtain Critical Value from table

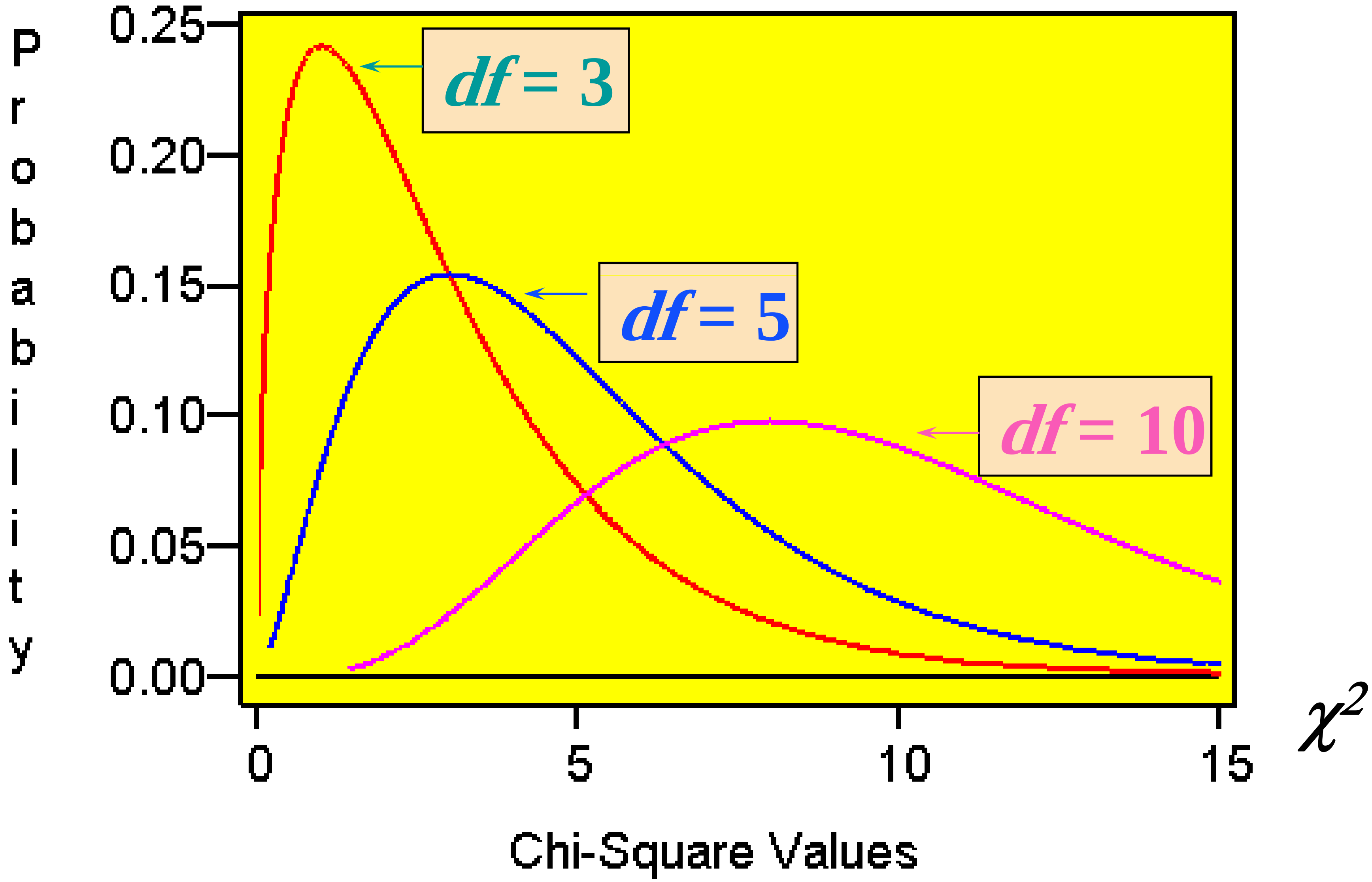
8) Make decision regarding the Null-hypothesis

The chi-square distribution



- Probability distributions that are continuous, have one mode, and are skewed to the right or positively skewed.
- It is non-negative.
- It is based on degrees of freedom, exact shape varies according to the number of degrees of freedom.
- The critical value of a test statistic in a chi-square distribution is determined by specifying a significance level and the degrees of freedom.

Different chi-square distributions



Chi-square Critical Value Determination

df	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2103	10.5966
3	6.2514	7.8147	9.3484	11.3449	12.8381
4	7.7794	9.4877	11.1433	13.2767	14.8602
.	.	.	.	.	.
.	.	.	.	.	.
8	13.3616	15.5073	17.5346	20.0902	21.9550
.	.	.	.	.	.
.	.	.	.	.	.
30	40.2560	43.7729	46.9792	50.8922	53.6720

In the following contingency table estimate the proportion of drinkers among those who develop Lung Cancer and those without the disease?

		Lung Cancer		Total
		Case	Control	
Drinker	Yes	O11=33	O12=27	R1=60
	No	O21=1667	O22= 2273	R2=3940
Total		C1 = 1700	C2 = 2300	n = 4000

E11=1700(60)/4000=25.5 E12=34.5

E21=1674.5 E22=2265.5

E11=1700(60)/4000=25.5 **E12=34.5**

E21=1674.5 **E22=2265.5**

$$\chi^2_{obs} \sum_{k=1}^{k=4} \frac{(Observed - E\,\mathrm{xp}\,ected)^2}{Expected} =$$

$$\frac{\frac{(33-25.5)^2}{25.5} + \frac{(27-34.5)^2}{34.5} + \frac{(1667-1674.5)^2}{1674.5} + \frac{(2273-2265.5)^2}{2265.5}}{} = 4.0$$

EXAMPLE - Hodgkin's lymphoma

A one year follow-up study was conducted to examine the effect of an experimental drug on mortality in 296 cases of advanced non-Hodgkin's lymphoma. Controls received standard treatment. The data are provided below.

	Died	Survived	
Treatment	9	190	
Control	13	84	

- Calculate the expected counts for the cells in the table above.
- Test to see if the association between mortality outcome and treatment status is statistically significant. Provide the null and alternative hypothesis and an interpretation of the results

EXAMPLE - Hodgkin's lymphoma

STEP 1

State the null and the alternative hypotheses and nominate the significance level, α

H_0: Mortality outcome and treatment status are **INDEPENDENT**

(proportion dying in treatment group is equal to proportion in control group)

H_a: Two variables are **RELATED.**

α **= 0.05**

Reject if computed chisquare>3.84 (from table)

EXAMPLE - Hodgkin's lymphoma

$$Expected Count = \frac{(Row Total)(Column Total)}{Grand Total}$$

E(9)	=	(199 x 22)/296	= 14.79
E(190)	=	(199x274)/296	= 184.21
E(13)	=	(97x22)/296	= 7.21
E(84)	=	(97x274)/296	= 89.79

Chi-Square Test

Expected counts are printed below observed counts

	died	survived	Total
Treatment	9 (**14.79**)	190 (**184.21**)	**199**
Control	13 (**7.21**)	84 (**89.79**)	**97**
Total	**22**	**174**	296

EXAMPLE - Hodgkin's lymphoma

STEP 2

Decide which test to use and obtain test statistic

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2 = \frac{(9-14.79)^2}{14.79} + \frac{(13-7.21)^2}{7.21} + \frac{(190-184.21)^2}{184.21} + \frac{(84-89.79)^2}{89.79}$$

= 2.267 + 4.651 + 0.182 + 0.373 = 7.473

EXAMPLE - Hodgkin's lymphoma

STEP 3 **Check the assumptions and conditions**

STEP 4 **Obtain the p-value**

d.f. = (r-1) x (c -1) =1

Test statistic is between 6.63 and 7.88

0.01 < P-value < 0.005

STEP 5 **Formulate and apply a decision rule**

STEP 6 **State the conclusion**

EXAMPLE - DVT

- Prevention of deep venous thrombosis (DVT) is a critical issue in patients undergoing total hip replacement surgery.
- Orthopaedic surgeons recognise the importance of prophylaxis in the management of their patients but do not agree on an optimal method.
- Three different prophylactic measures are to be compared for the prevention of a proximal DVT after total hip replacement surgery.
- Three independent groups of patients (n= 85, 75 and 80 respectively) undergoing total hip replacements were given different prophylactics.
- After surgery, it was noted whether patients had complications from proximal DVT or not.
- The results are presented in the following contingency table. (expected frequencies are shown in brackets below their corresponding observed frequencies).

EXAMPLE - DVT

	Group 1	Group 2	Group 3	Total
No complications	76	71	69	216
		(67.5)	*(72)*	
DVT complications	9	4	11	24
	(8.5)	*(7.5)*	*(8)*	
Total:	85	75	80	

Does this provide statistically significant evidence of a relationship between risk of DVT complications and type of prophylactic measure?

EXAMPLE - DVT

STEP 1

State the null and the alternative hypotheses and nominate the significance level, α

H_0: Risk of DVT and type of prophylactic measure are **INDEPENDENT**

H_a: Two variables are **RELATED.**

α **= 0.05**

Reject if computed chisquare>5.99 (from table)

EXAMPLE - DVT

STEP 2 **Decide which test to use and obtain test statistic**

	Group 1	Group 2	Group 3	Total
No complications	76	71	69	216
		(67.5)	*(72)*	
DVT complications	9	4	11	24
	(8.5)	*(7.5)*	*(8)*	
Total:	85	75	80	240

Expected count = (216 * 85)/240 = 76.5

EXAMPLE - DVT

STEP 2 **Decide which test to use and obtain test statistic**

	Group 1	Group 2	Group 3	Total
No complications	76	71	69	216
	(76.5)	*(67.5)*	*(72)*	
DVT complications	9	4	11	24
	(8.5)	*(7.5)*	*(8)*	
Total:	85	75	80	

$$\chi^2 = \frac{(76-76.5)^2}{76.5} + \frac{(71-67.5)^2}{67.5} + \frac{(69-72)^2}{72} + \frac{(9-8.5)^2}{8.5} + \frac{(4-7.5)^2}{7.5} + \frac{(11-8)^2}{8} =$$

$$0.003 + 0.181 + 0.125 + 0.029 + 1.633 + 1.125 = \mathbf{3.097}$$

EXAMPLE - DVT

STEP 3 **Check the assumptions and conditions**

STEP 4 **Obtain the p-value (or determine critical value(s))**

d.f. = (r-1) x (c -1) = (2-1) x (3-1) = 2

Test statistic is between 3.79 and 4.61

0.10 < P-value < 0.15

STEP 5 **Formulate and apply a decision rule**

STEP 6 **State the conclusion**

How many patients are enough? Calculating sample size

- Effect size : magnitude of the difference(s) to be detected
  - more patients needed to detect smaller difference
- Alpha error :risk of concluding treatment is effective when it is not (protection from falsely-positive statistical test)
- Beta error : (protection from a falsely-negative statistical test)

What if *n* is too small

- Increase *n*.
- If *categories* > 2, combine categories.
- Use a correction factor.
- Use another test.

If *categories* > 2, combine categories An example: Combining categories

- With three habitat categories, expected frequencies are too small in 2 cells.
- Therefore, combine habitats B and C.

Habitat	***Males***	***Females***	***Total***
A	**30**	**34**	**64**
B	**55**	**25**	**80**
C	**3**	**1**	**4**
Total	**88**	**60**	**148**

Habitat	***Males***	***Females***	***Total***
A	**30**	**34**	**64**
B/C	**58**	**26**	**84**
Total	**88**	**60**	**148**

Use another test: Fisher's Exact Test

- This test can be used for 2 by 2 tables when the number of cases is too small to satisfy the assumptions of the chi-square.
  - Total number of cases is <20 or
  - The expected number of cases in any cell is <1 or
  - More than 25% of the cells have expected frequencies <5.

PNEUMONIA COMPLICATION 480.00-486.99 * CIRRHOSIS OR CHRONIC LIVER 571 Crosstabulation

			CIRRHOSIS OR CHRONIC LIVER 571		
			ABSENT	PRESENT	Total
PNEUMONIA COMPLICATION 480.00-486.99	ABSENT	Count	870	5	875
		Expected Count	867.5	7.5	875.0
		% within PNEUMONIA COMPLICATION 480.00-486.99	99.4%	.6%	100.0%
		% within CIRRHOSIS OR CHRONIC LIVER 571	94.1%	62.5%	93.8%
	PRESENT	Count	55	3	58
		Expected Count	57.5	.5	58.0
		% within PNEUMONIA COMPLICATION 480.00-486.99	94.8%	5.2%	100.0%
		% within CIRRHOSIS OR CHRONIC LIVER 571	5.9%	37.5%	6.2%
Total		Count	925	8	933
		Expected Count	925.0	8.0	933.0
		% within PNEUMONIA COMPLICATION 480.00-486.99	99.1%	.9%	100.0%
		% within CIRRHOSIS OR CHRONIC LIVER 571	100.0%	100.0%	100.0%

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.545[b]	1	.000		
Continuity Correction[a]	8.674	1	.003		
Likelihood Ratio	6.842	1	.009		
Fisher's Exact Test				.010	.010
Linear-by-Linear Association	13.531	1	.000		
N of Valid Cases	933				

a. Computed only for a 2x2 table

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is .50.

PNEUMONIA COMPLICATION 480.00-486.99 * SEX Crosstabulation

			SEX		
			MALE	FEMALE	Total
PNEUMONIA COMPLICATION 480.00-486.99	ABSENT	Count	197	678	875
		% within SEX	90.0%	95.0%	93.8%
	PRESENT	Count	22	36	58
		% within SEX	10.0%	5.0%	6.2%
Total		Count	219	714	933
		% within SEX	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.197[b]	1	.007		
Continuity Correction[a]	6.364	1	.012		
Likelihood Ratio	6.492	1	.011		
Fisher's Exact Test				.010	.008
Linear-by-Linear Association	7.189	1	.007		
N of Valid Cases	933				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.61.

Expected number

NEUMONIA COMPLICATION 480.00-486.99 * CIRRHOSIS OR CHRONIC LIVER 571 Crosstabulation

Count

		CIRRHOSIS OR CHRONIC LIVER 571		
		ABSENT	PRESENT	Total
PNEUMONIA COMPLICATION 480.00-486.99	ABSENT	870	5	875
	PRESENT	55	3	58
Total		925	8	933

PNEUMONIA COMPLICATION 480.00-486.99 * CIRRHOSIS OR CHRONIC LIVER 571 Crosstabulation

			CIRRHOSIS OR CHRONIC LIVER 571		
			ABSENT	PRESENT	Total
PNEUMONIA COMPLICATION 480.00-486.99	ABSENT	Count	870	5	875
		Expected Count	867.5	7.5	875.0
	PRESENT	Count	55	3	58
		Expected Count	57.5	.5	58.0
Total		Count	925	8	933
		Expected Count	925.0	8.0	933.0

NEUMONIA COMPLICATION 480.00-486.99 * CIRRHOSIS OR CHRONIC LIVER 571 Crosstabulatio

			CIRRHOSIS OR CHRONIC LIVER 571		
			ABSENT	PRESENT	Total
PNEUMONIA COMPLICATION 480.00-486.99	ABSENT	Count	870	5	875
		Expected Count	867.5	7.5	875.0
		% within PNEUMONIA COMPLICATION 480.00-486.99	99.4%	.6%	100.0%
		% within CIRRHOSIS OR CHRONIC LIVER 571	94.1%	62.5%	93.8%
	PRESENT	Count	55	3	58
		Expected Count	57.5	.5	58.0
		% within PNEUMONIA COMPLICATION 480.00-486.99	94.8%	5.2%	100.0%
		% within CIRRHOSIS OR CHRONIC LIVER 571	5.9%	37.5%	6.2%
Total		Count	925	8	933
		Expected Count	925.0	8.0	933.0
		% within PNEUMONIA COMPLICATION 480.00-486.99	99.1%	.9%	100.0%
		% within CIRRHOSIS OR CHRONIC LIVER 571	100.0%	100.0%	100.0%

Test of paired proportions

- Analogous to paired t-test, but binary rather than continuous outcome:

Test of paired proportions, *Example*

- Johnson and Johnson (*NEJM 287: 1122-1125, 1972*) selected 85 Hodgkin's patients who had a sibling of the same sex who was free of the disease and whose age was within 5 years of the patient's…they presented the data as….

	Tonsillectomy	**None**
Hodgkin's	**41**	**44**
Sib control	**33**	**52**

chi-square=1.53 (NS)

Test of paired proportions, Example

- But several letters to the editor pointed out that those investigators had made an error by ignoring the pairings. These are not independent samples because the sibs are paired…better to analyze data like this:

Case	**Control**	
	Tonsillectomy	**None**
Tonsillectomy	**37**	**7**
None	**15**	**26**

Chi-square=2.91 (p=.09)

Test of paired (matched) proportions, Example

Study of the relationship between diabetes and MI

Match each MI case to an MI control based on age and gender.

Ask about history of diabetes to find out if diabetes increases your risk for MI.

Test of paired (matched) proportions, Example

Just the discordant cells are informative!

MI controls

MI cases	Diabetes	No Diabetes	
Diabetes	**9**	**37**	**46**
No diabetes	**16**	**82**	**98**
	25	**119**	**144**

Which cells are informative?

Test of paired (matched) proportions, Example

MI cases	MI controls: Diabetes	MI controls: No Diabetes	
Diabetes	9	37	46
No diabetes	16	82	98
	25	119	144

The question is: *among the discordant pairs*, what proportion are discordant in the direction of the case vs. the direction of the control. If more discordant pairs "favor" the case, this indicates that diabetes increases the risk of MI

McNemar's Test: generally

controls

cases	**exp**	**No exp**
exp	**a**	**b**
No exp	**c**	**d**

- In this situation, the McNemar test for paired proportions is appropriate with 1 df

$$\chi_1^2 = \left(\frac{|b-c|-1}{\sqrt{b+c}}\right)^2 = \frac{(b-c-1)^2}{b+c}$$

Test of paired (matched) proportions, Example

MI controls

MI cases	**Diabetes**	**No Diabetes**	
Diabetes	**9**	**37**	**46**
No diabetes	**16**	**82**	**98**
	25	**119**	**144**

$$\chi_1^2 = 26.8, p < 0.001$$

Paired Binary Data

Example: measured a binary response pre and post treatment. This is an example of **paired binary data**. One way to display these data is the following:

	Correct	Incorrect	Total
Baseline	166	429	595
Month 3	276	319	595
Total	442	748	1190

Q: Can't we simply use X^2 Test to assess whether this is evidence for an increase in knowledge?

A: NO!!! The X^2 tests assume that the rows are **independent** samples. In this design it is the same 595 people at Baseline and at 3 months.

Paired Binary Data

For paired binary data we display the results as follows:

		Time 2	
		0	1
Time 1	0	a	b
	1	c	d

Paired Binary Data

For paired binary data we display the results as follows:

Baseline	Month 3		
	Cor	Inc	Total
Cor			166
Inc			429
Total	276	319	595

The End