10. Categorical Data

Readings: Rosner: Ch. 10
Chihara and Hesterberg: 3.5
OpenIntro Statistics: Ch. 6

R: chisq.test, fisher.test, mcnemar.test, epiR: epi.2by2

Homework: Homework 4 due by noon on October 1
Homework 5 due by noon on October 8

Overview

A) Background and Observational Study Designs
B) Measures of Effect for 2 x 2 Tables of Categorical Data
C) Test of Association
D) Categorical Data with Small Expected Values
E) Paired Samples for Categorical Data

A) Background and Observational Study Designs

- Paired and two-sample t-tests, ANOVA and their nonparametric counterparts are appropriate for examining continuous response or outcome variables.

- Categorical data methods are appropriate for examining categorical response variables.

Example: Prospective study of 10-year incidence of lung cancer in heavy drinkers (≥ 2 drinks per day) and nondrinkers.

Drinking	*Lung Cancer*		
Status	Yes	No	
Heavy	33	1667	1700
Non	27	2273	2300
	60	3940	4000

General 2 × 2 contingency table:

	Disease		
Exposure	Yes	No	
Yes	a	b	$a+b = n_1$
No	c	d	$c+d = n_2$
	$a+c = m_1$	$b+d = m_2$	

Examples of Observational Study Designs

A ***prospective study*** is a study in which a group of disease-free individuals are identified at one point in time, their exposure variables are measured at baseline, and they are then followed over a period of time until some of them develop the disease. The study population in a prospective study is often referred to as a cohort. This type of study is also referred to as a ***prospective cohort study***. Note, you can also have a ***prospective case-control study*** where you select subjects in real time and match controls to cases as cases occur.

A ***retrospective study*** is a study in which two groups of individuals are initially identified: (1) a group that has the disease under study (the cases) and (2) a group that does not have the disease under study (the controls). An attempt is then made to relate their prior health habits (or exposures) to their current disease status. This type of study is also referred to as a ***case-control study***. Note, you can also have ***retrospective cohort studies*** where you select subjects based on past exposure status (e.g., indication of smoking on a past medical record) and look at the outcome today.

A ***cross-sectional study*** is a study in which a study population is ascertained at one point in time. All individuals in the study population are asked about their current disease status and their current or past health habits (or exposures). This type of study is also referred to as a ***prevalence study***, because the prevalence of disease at one point in time is compared between exposed and unexposed individuals.

B. Measures of Effect for 2 × 2 Tables of Categorical Data

	Disease		
Exposure	Yes	No	
Yes	a	b	$a+b = n_1$
No	c	d	$c+d = n_2$
	$a+c = m_1$	$b+d = m_2$	

Let $p_1 = a/n_1$ be the probability of disease among the exposed.

Let $p_2 = c/n_2$ be the probability of disease among the unexposed.

We will explore 3 ways to describe the behavior of p_1 and p_2 between our exposure groups:

B1. risk difference
B2. risk ratio
B3. odds ratio

B1. Risk Difference

- The risk difference is defined as p_1 - p_2 .
- The risk difference is an appropriate effect measure for cohort and cross-sectional studies.
  - In a cohort study, this is the difference in incidence rates between the exposed and unexposed individuals.
  - In a cross-sectional study, this is the difference in the prevalence rates between the exposed and unexposed individuals.
- The standard error for the risk difference is

$$SE(p_1 - p_2) = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{ where } p = \frac{a+c}{n_1+n_2}$$

Example:

Drinking Status	*Lung Cancer* Yes	 No	
Heavy	33	1667	1700
Non	27	2273	2300
	60	3940	4000

B2. Risk Ratio (Special Case of Relative Risk: https://www.ctspedia.org/do/view/CTSpedia/RiskRate)

- The risk ratio is defined as RR=p_1/ p_2.
- The risk ratio is an appropriate effect measure for cohort and cross-sectional studies.
- Assuming a causal effect between the exposure and outcome we can interpret it as: RR=1 means no relationship, RR>1 indicates a risk factor, RR<1 indicates a protective factor.
- The sampling distribution of the natural log of RR is approximately normal and has a standard error of

$$SE[\log(RR)] = \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}}$$

- To find the confidence interval around RR, we can calculate the CI for log(RR) and then exponentiate our two bounds: $\exp\left\{\log(RR) \pm z_{1-\frac{\alpha}{2}} \times SE[\log(RR)]\right\}$

Example:

Drinking Status	*Lung Cancer* Yes	No	
Heavy	33	1667	1700
Non	27	2273	2300
	60	3940	4000

B3. Odds Ratio (Another special case of Relative Risk)

- The odds ratio is defined as $\frac{p_1/_{1-p_1}}{p_2/_{1-p_2}} = \frac{ad}{bc}$.
- The odds ratio is an appropriate effect measure for cohort, cross-sectional, and case-control studies.
- Interpreted as the odds in favor of an outcome for the exposed group are *OR* times the odds in favor for the unexposed group.
- The odds ratio will approximate the risk ratio for rare diseases (“low” incidence).
- The standard error for the log(OR) is (with confidence intervals for OR derived by exponentiating the bounds)

$$SE[\log(OR)] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Example:

Drinking Status	*Lung Cancer* Yes	 No	
Heavy	33	1667	1700
Non	27	2273	2300
	60	3940	4000

C. Test of association (chi-squared test)

Each of the effect measures will have a respective associated confidence interval (and method for obtaining it). Regardless of the study design we can test the hypothesis of no association between exposure and disease using one (and the same) test statistic. It should be noted that the CI for an effect measure will sometimes be at odds with the p-value for the test statistic, particularly when the results are marginally significant. This is a result of a mismatch between the ways in which the CI and the test statistic are formulated.

The test statistic used for this purpose is based on the assumption of independence between the row and column variables, *exposure* (A) (vs. *no exposure*, A^c) and *disease* (B) (vs. *no disease*, B^c). Our H_0 is that there is no association between exposure and disease.

Recall that when two events A and B are *independent* P(A and B) = P(A) x P(B). The expected *number* of events (A and B) is then *N* x P(A) x P(B), where *N* is the total number of events (A, A^c, B, B^c). Our table of observed data can then be represented as:

	Disease		
Exposure	Yes	No	
Yes	O_{11}	O_{12}	n_1
No	O_{21}	O_{22}	n_2
	m_1	m_2	N

The expected number of events in a cell above is N x (n_i/N) x (m_j/N) = (n_i x m_j)/N. We can then create a table for what we would expect to observe:

	Disease		
Exposure	Yes	No	
Yes	E_{11}	E_{12}	n_1
No	E_{21}	E_{22}	n_2
	m_1	m_2	N

To summarize how much the entries in the observed table deviate from the assumption of independence, we combine the information in the two tables in this way (from Karl Pearson):

$$\mathrm{X}^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi_1^2$$

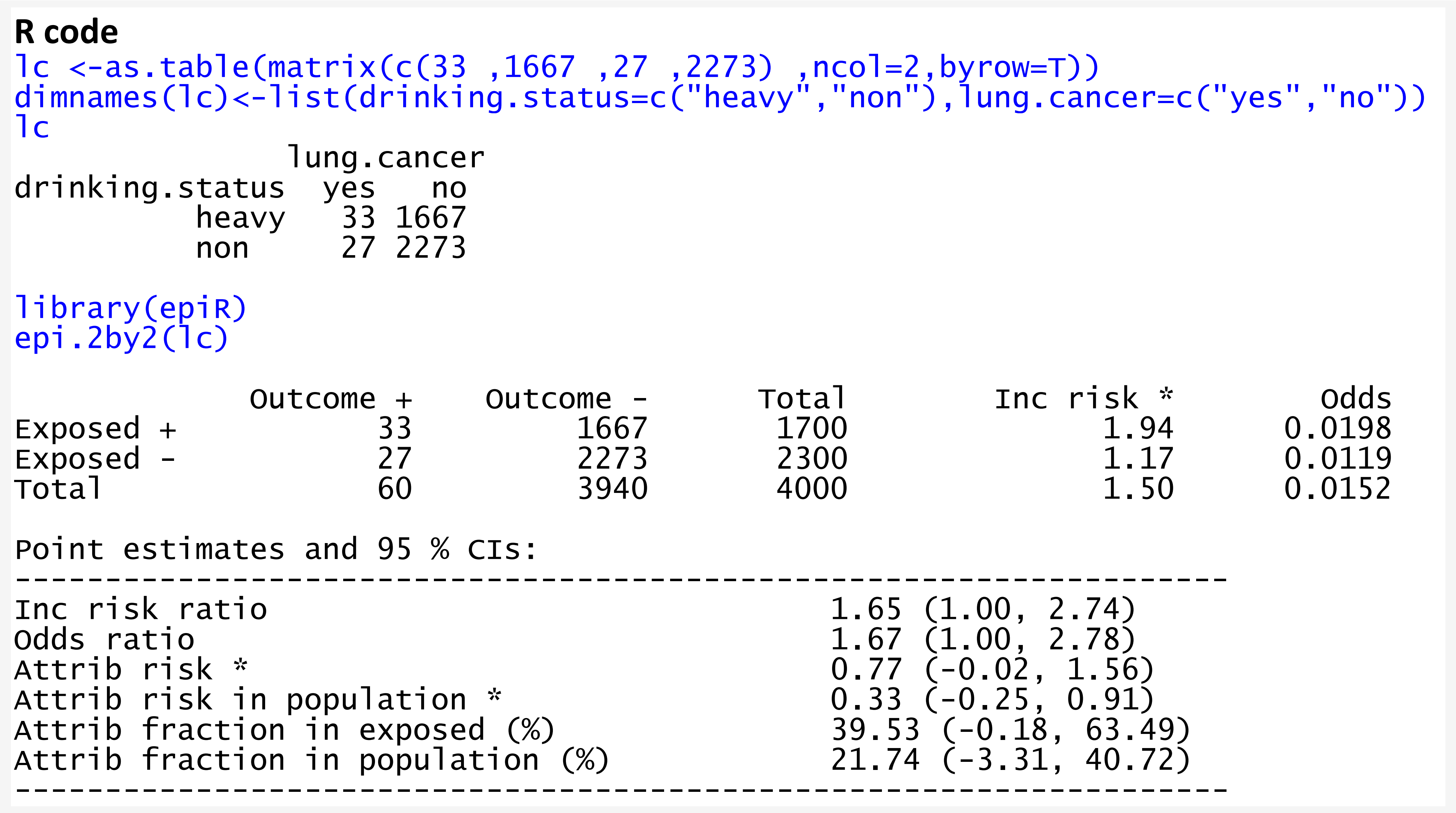
The χ^2 distribution with 1 degree of freedom is the square of a standard normal, i.e. $\chi_1^2 = Z^2$. One assumption we must make for this test to be valid is that *all of the expected values must be greater than or equal to 5*.

The Yates-corrected version of this test applies the correction for continuity, because we are applying continuous distributions to discrete data:

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(\left|O_{ij} - E_{ij}\right| - 0.5\right)^2}{E_{ij}} \sim \chi_1^2$$

The χ^2 test of independence (no association) can be generalized to an *r* x *c* table (vs. a 2 x 2 table). The double subscripts on *O* and *E* run instead from 1 to *r* and 1 to *c*, respectively. Likewise for the double sum, $\sum_{i=1}^{r}\sum_{j=1}^{c}$. The resulting statistic is a χ^2 with (*r*-1) x (*c*-1) df. Note that there are families of tests that can be applied to large tables depending on the pattern or hypothesis of interest. For more information, see: Agresti, A. *Categorical Data Analysis*, Wiley, 2002.

Let's take a look at how we can use R to obtain these effect measures, their CI, and statistics for testing independence.



R code

```
lc <-as.table(matrix(c(33 ,1667 ,27 ,2273) ,ncol=2,byrow=T))
dimnames(lc)<-list(drinking.status=c("heavy","non"),lung.cancer=c("yes","no"))
lc
               lung.cancer
drinking.status  yes   no
          heavy   33 1667
          non     27 2273

library(epiR)
epi.2by2(lc)

             Outcome +    Outcome -      Total        Inc risk *       Odds
Exposed +           33         1667       1700              1.94     0.0198
Exposed -           27         2273       2300              1.17     0.0119
Total               60         3940       4000              1.50     0.0152

Point estimates and 95 % CIs:
-------------------------------------------------------------------
Inc risk ratio                               1.65 (1.00, 2.74)
Odds ratio                                   1.67 (1.00, 2.78)
Attrib risk *                                0.77 (-0.02, 1.56)
Attrib risk in population *                  0.33 (-0.25, 0.91)
Attrib fraction in exposed (%)               39.53 (-0.18, 63.49)
Attrib fraction in population (%)            21.74 (-3.31, 40.72)
-------------------------------------------------------------------
```

R code cont.

```
X2 test statistic: 3.895 p-value: 0.048
 Wald confidence limits
 * Outcomes per 100 population units

chisq.test(lc ,correct=F)

     Pearson's Chi-squared test

data:  lc
X-squared = 3.8947, df = 1, p-value = 0.04844

chisq.test(lc)

     Pearson's Chi-squared test with Yates' continuity correction

data:  lc
X-squared = 3.3927, df = 1, p-value = 0.06548
```

Conclusion:

D. Categorical Data - Small Expected Values

Fisher's Exact test of H_0: $p_1 = p_2$, small independent samples

The normal theory method discussed for comparing two binomial proportions yields appropriate p-values but requires that the normal approximation to the binomial distribution be valid, which is not always the case for small sample sizes.

When comparing two proportions with sample sizes too small to use the χ^2 methods, Fisher's exact test can be used. It is the two-sample (conditional) analog to the exact one-sample binomial test and gives exact p-value results for any 2x2 table but is only necessary for tables with small *expected* values, i.e. where the standard χ^2 test is not applicable.

The two tests can provide similar results: Fisher's exact p-value $\rightarrow \chi^2$ test p-value for large samples. The p-values for the Yates-corrected χ^2 and the Fisher's exact test are usually very close. The conditional nature of the Fisher's exact test is such that the p-values are larger than those for the uncorrected χ^2 test.

The hypergeometric distribution on which Fisher's exact test is based is very discrete (far from continuous). For large samples Fisher's exact test is computationally intensive.

Example: Suppose a retrospective study is done on the deaths of all men aged 50-54 in a specific county over a 1-month period. Of the 35 men who died of cardiovascular disease, 5 were on a high salt diet before they died. Of the 25 men who died of other causes, 2 were on a high salt diet. Is there an association between a high salt diet and CVD?

Salt	*Death*		
Levels	CVD	Other	
High	5	2	7
Low	30	23	53
	35	25	60

Are the expected values too small for the χ^2 test?

Expected values in each cell:

Salt	*Death*		
Levels	CVD	Other	
High			7
Low			53
	35	25	60

Procedure for Fisher's Exact test:

1. Write down all possible tables with the same marginals (assume that the margins of the table are fixed: e.g. numbers on high and low salt diets, numbers with CVD and other death)

- Rearrange the rows and columns of the observed table so that the smaller row total is in the first row and the smaller column total is in the first column.
- Start with the table with 0 in the (1,1) cell. Other cells are determined by the fixed row and column margins (e.g. the 1,2 element must be a+b)
- Construct the next tables by increasing the (1,1) cell by 1 and re-computing the other cells based on the fixed marginals
- Continue increasing and decreasing the cells by 1 until one of the cells is 0, at which point all possible tables with the given marginals have been enumerated. Commonly, each table is referred to by its (1,1) element. e.g. the 1st table is called the "0 table"

2. Calculate the probability of each table occurring (given the marginals):

a	b	a+b
c	d	c+d
a+c	b+d	

$$P(a,b,c,d) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

This is the exact probability of observing a table with cells *a, b, c,* and *d.* It is the probability mass function for the *hypergeometric distribution*, one of the many discrete distributions used in statistics (for more information, see Chapter 10, Section 3 in Rosner text).

3. To test H_0: $p_1 = p_2 = p$ vs. H_1: $p_1 \neq p_2$, calculate the probability of obtaining a table as extreme or more extreme than the observed table:

Two-sided p-value:

$2 \times \min\{P(0) + P(1) \ldots + P(a);\ P(a) + P(a+1) + \ldots + P(k);\ 0.5\}$

where the *observed* table is the *a* table and the *last* table enumerated is the *k* table.

One-sided p-value:

To test H_0: $p_1 = p_2 = p$ vs. H_1: $p_1 < p_2$, the p-value = $P(0) + P(1) + \ldots + P(a)$

To test H_0: $p_1 = p_2 = p$ vs. H_1: $p_1 > p_2$, the p-value = $P(a) + P(a+1) + \ldots + P(k)$

Example salt and CVD:

0	7	7
25	28	53
25	35	60

Prob: 0.017

1	6	7
24	29	53
25	35	60

0.105

2	5	7
23	30	53
25	35	60

0.252

…

7	0	7
18	35	53
25	35	60

0.001

$$P(0) = \frac{7!53!25!35!}{60!0!7!25!28!} = 0.017$$

$$P(1) = \frac{7!53!25!35!}{60!1!6!24!29!} = 0.105$$

$$\vdots$$

$$P(7) = \frac{7!53!25!35!}{60!7!0!18!35!} = 0.001$$

R code

```
cvd <-as.table(matrix(c(5 ,2 ,30 ,23),ncol=2,byrow=T))
dimnames(cvd)<-list(salt=c("high","low"),death=c("yes",
"other"))
cvd

      death
salt    yes other
  high    5     2
  low    30    23

fisher.test(cvd)

       Fisher's Exact Test for Count Data

data:  cvd
p-value = 0.6882
alternative hypothesis: true odds ratio is not equal to
1
95 percent confidence interval:
  0.278957 21.620483
sample estimates:
odds ratio
  1.897126
```

Recall Observed:

	CVD	Other	
High Salt	5	**2=a**	7
Low Salt	30	23	53
	35	25	60

p-value = (This is what R does; other programs will do it slightly differently.)

sum of table probabilities for tables with probabilities less than or equal to the probability of the observed table:

0.017, 0.105, 0.252, **0.312**, 0.214, 0.082, 0.016, 0.001

= 1 – 0.3118225 = 0.6881775

Conclusion:

R code – p-value using hypergeometric pmf

```
dhyper(0, 7, 53, 25)
[1] 0.0174117
dhyper(1, 7, 53, 25)
[1] 0.1050706
dhyper(2, 7, 53, 25) # prob observed table
[1] 0.2521695
dhyper(3, 7, 53, 25) # prob of table with greater prob than observed table
[1] 0.3118225
dhyper(4, 7, 53, 25)
[1] 0.214378
dhyper(5, 7, 53, 25)
[1] 0.0818534
dhyper(6, 7, 53, 25)
[1] 0.01604969
dhyper(7, 7, 53, 25)
[1] 0.00124467
1-dhyper(3, 7, 53, 25)
[1] 0.6881775 # two-sided p-value from R
```

E. Categorical Data - Paired Samples: McNemar's Test for paired (or matched) data

Example: A study was done to compare two chemotherapy regimens. Subjects were matched by age and stage of disease. A random member of each pair received treatment A and the other was assigned to treatment B. The patients were followed for 5 years with survival as the outcome variable. There were 1242 patients total, and 621 pairs.

Treatment A: Survival rate = 526/621 = 0.847 (84.7%)
Treatment B: Survival rate = 515/621 = 0.829 (82.9%)

Is this (small) difference in survival significant? We might first set up the usual 2x2 contingency table appropriate for a χ^2 analysis:

Treat-ment	*Survive*		
	Yes	No	
A	526	95	621
B	515	106	621
	1041	201	1242

If we analyzed this with a χ^2 test, it would not be significant.
But, is this the correct analysis?

The χ^2-test is valid only if the two samples are *independent.* Since the patients have been *matched on age and stage*, the groups assigned to treatment A and treatment B *are not independent*. What should we do?

We set up a different type of 2x2 contingency table with the matched pair as the unit of observation. Here are the results for the matched pairs data:

Outcome Trt A Patient	*Outcome Trt B Patient*		
	Survive	Die	
Survive >5 yrs	510	**16**	526
Die in 5 yrs	**5**	90	95
	515	106	621

Pairs where the outcomes differed are known as *discordant pairs* (bold numbers). In these discordant pairs we see that more pairs showed the patient on treatment A surviving and the matched patient on treatment B dying than the reverse.

General Procedure:

- Ignore concordant pairs (pairs where the patients had the same outcome), focus on if the discordant pairs occur in equal frequencies
- Let p = P(patient on Treatment A lived given that the paired patients had different outcomes), i.e. the probability of a discordant pair
- Test H_0: $p = ½$

As usual for proportions, we can test the hypothesis using exact or normal approximation methods:

- Large sample: for n_D = number of discordant pairs ≥ 20 we can use a normal theory test
- Small samples: for $n_D < 20$ use exact binomial test

These tests are equivalent to a one-sample method for paired qualitative data, *the Sign test.*

Notation:
n_D = number of discordant pairs
n_A = number of discordant pairs where Treatment A patient lived (or Treatment B patient lived, we just need to choose one for calculations)

Large Sample: for $n_D \geq 20$, with correction for continuity

$\mathrm{X}^2 = \frac{\left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2}{n_D/4} \sim \chi_1^2$, under H_0: p=0.5

p-value = $P(\chi_1^2 > \mathrm{X}^2)$

Small Sample: for $n_D < 20$ - Find the exact binomial probabilities:

If $n_D < 20$ then exact binomial probabilities rather than the normal approximation must be used to compute the p-value. H_0 will be rejected if n_A is very large or very small, or if the associated p-value is less than the pre-specified α-level. The exact p-value is given by:

$$\text{If } n_A < \frac{n_D}{2}: \quad p = 2 \times \sum_{k=0}^{n_A} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$$

$$\text{If } n_A > \frac{n_D}{2}: \quad p = 2 \times \sum_{k=n_A}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$$

$$\text{If } n_A = \frac{n_D}{2}: \quad p = 1.0$$

If n_D = # discordant pairs, we expect $n_A = n_D/2$ of them to be of the type where “treatment” A shows benefit and “treatment” B does not.

Example: $n_D = 21$, $n_A = 5$

Outcome Trt A Patient	*Outcome Trt B Patient*		
	Survive	Die	
Survive >5 yrs	510	**16**	526
Die in 5 yrs	**5**	90	95
	515	106	621

Large sample approach:

$$\mathrm{X}^2 = \frac{\left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2}{{n_D}/{4}} = \frac{\left(\left|5 - \frac{21}{2}\right| - \frac{1}{2}\right)^2}{{21}/{4}} = 4.76 \textit{ or } \mathrm{X}^2 = \frac{\left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2}{{n_D}/{4}} = \frac{\left(\left|16 - \frac{21}{2}\right| - \frac{1}{2}\right)^2}{{21}/{4}} = 4.76$$

This results in $P(\chi_1^2 > \mathrm{X}^2) = P(\chi_1^2 > 4.76)$ = 1-pchisq(4.76, df=1) = 0.029

An alternative approach using the connection between $\chi_1^2 = Z^2$:

$n_A \sim \mathrm{Bin}\left(n_D, \frac{1}{2}\right)$ under H_0.

$$Z = \frac{\left(\left|n_A - E[n_A]\right| - \frac{1}{2}\right)^2}{\sqrt{V[n_A]}} = \frac{\left(\left|5 - \frac{21}{2}\right| - \frac{1}{2}\right)^2}{\sqrt{21\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = 2.18 \rightarrow Z^2 = 2.18^2 = 4.76 \text{ and } P(\chi_1^2 > 4.76) = 0.029.$$

Small sample approach:

$$\begin{aligned}\text{p-value} &= 2\times[P(n_A=0)+p(n_A=1)+\cdots+P(n_A=5)]\\ &= 2\times\left[\binom{21}{0}0.5^0(1-0.5)^{21}+\cdots+\binom{21}{5}0.5^5(1-0.5)^{16}\right]\\ &= 0.0266037\end{aligned}$$

Conclusion: ?

R Code

```
chemopairs <-as.table(matrix(c(510 ,16 ,5 ,90),ncol=2,byrow=T))
dimnames(chemopairs)<-list(tmtA=c("survive","5yr"),tmtB=c("survive","5yr"))
chemopairs

         tmtB
tmtA      survive 5yr
  survive     510  16
  5yr           5  90

mcnemar.test(chemopairs, correct=F) # no continuity correction

	McNemar's Chi-squared test

data:  chemopairs
McNemar's chi-squared = 5.7619, df = 1, p-value = 0.01638

mcnemar.test(chemopairs, correct=T) # with continuity correction

	McNemar's Chi-squared test with continuity correction

data:  chemopairs
McNemar's chi-squared = 4.7619, df = 1, p-value = 0.0291

2*pbinom (5, 21, 0.5) #P(X<=5 Type A discordant pairs out of 21 discordant pairs if H0:
prob of Type A vs Type B discordant pairs is 0.5 vs. H1: prob not equal to 0.5)
[1] 0.0266037
```

Stats Humor

The risk I took was calculated, but *man*, am I bad at math.

☺