AMS 572 Data Analysis I
Inference on two sample proportions

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

©PF.Kuan

Inference on two proportions

- Small sample sizes
  - Fisher's exact test
- Large sample sizes
  - normal approximation to the binomial
  - χ^2 test

Inference on two proportions

- $X \sim \text{Binomial}(n_1, p_1)$, $Y \sim \text{Binomial}(n_2, p_2)$
- Put data in 2 x 2 table

	Success	Failure	
Sample 1	$n_{11} = x$	$n_{12} = n_1 - x$	n_1
Sample 2	$n_{21} = y$	$n_{22} = n_2 - y$	n_2
	$m = x + y$	$n - m$	n

- Hypotheses

$$H_0 : p_1 = p_2$$

versus

$$H_a : p_1 > p_2 \text{ or } H_a : p_1 < p_2 \text{ or } H_a : p_1 \neq p_2$$

Fisher's Exact Test

- Assume margins m, $n - m$, n_1, n_2 fixed
- Then once we know $n_{11} = x$, the other values n_{12}, n_{21}, and n_{22} are immediately determined
- Under H_0, can show

$$\Pr[X = k | X + Y = m] =$$

Fisher's Exact Test

- To compute p-values, consider all 2 x 2 tables possible given the observed margins
- One tailed p-value: sum the probabilities of the observed table and all tables more extreme than the observed table in the direction of H_a
- Two tailed p-value: sum the probabilities of tables that are more extreme in both directions than the observed table, given the fixed margins

Fisher's Exact Test

1. One tailed test: $H_0 : p_1 = p_2$ vs $H_a : p_1 > p_2$

$$\text{p-value} = p_U = P(X \geq x | X+Y = m) = \sum_{k=x}^{\min(m,n_1)} \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n}{m}}$$

2. One tailed test: $H_0 : p_1 = p_2$ vs $H_a : p_1 < p_2$

$$\text{p-value} = p_L = P(X \leq x | X+Y = m) = \sum_{k=\max(0,m-n_2)}^{x} \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n}{m}}$$

3. Two tailed test: $H_0 : p_1 = p_2$ vs $H_a : p_1 \neq p_2$

$$\text{p-value} = 2\min(p_L, p_U)$$

is the formula presented in your textbook. A pitfall of this approach is that the value can exceed 1. Alternative approaches of defining two-sided p-value can be found in page 92 of Categorical Data Analysis (Agresti) (See the next Example).

Example: The result of a randomized clinical trial for comparing Prednisone and Prednisone+VCR drugs, is summarized below. Test if the success and failure probabilities are the same for the two drugs.

Drug	Success	Failure	Row total
Pred	14	7	$n_1 = 21$
PVCR	38	4	$n_2 = 42$
	m=52	n-m=11	n=63

Solution:
$$\begin{cases} H_0 : p_1 = p_2 \\ H_a : p_1 \neq p_2 \end{cases}$$

$p_L = \sum_{k=\max(0,52-42)}^{14} \frac{\binom{21}{k}\binom{42}{52-k}}{\binom{63}{52}} = \sum_{k=10}^{14} \frac{\binom{21}{k}\binom{42}{52-k}}{\binom{63}{52}} \approx 0.0253$

Note: This is example 9.8 of your textbook. The p-value given in your textbook value is incorrect

If we define two-sided p-value as$= 2(0.0253) = 0.0506$, we do not reject H_0. This p-value is different from the p-value SAS/R output. We will present the method of obtaining exact two sided p-value in the following slides.

1. The two sided p-value is defined as the sum of probabilities for those tables having a test statistic greater than or equal to the value of the observed test statistic.

2. For the Prednisone example, list all possible tables with fixed margins $n_1 = 21, n_2 = 42, m = 52, n - m = 11$. There are 12 possible tables corresponding to possible values of $x = 10, 11, ..., 21$. Note that our observed $x = 14$. Compute $P(X = x | X + Y = 52)$ for each possible x values.

3. The two sided p-value can be obtained by adding all probabilities $\leq P(X = 14 | X + Y = 52)$

```
> prob <- function(x){
choose(21,x)*choose(42,52-x)/choose(63,52)
}
> ### list all the possible values of x for fixed margin ###
> tab <- cbind(10:21,prob(10:21))
> colnames(tab) <- c('x','p(X=x|X+Y=52)')
> tab
        x p(X=x|X+Y=52)
 [1,] 10  5.727859e-07
 [2,] 11  2.405701e-05
 [3,] 12  4.109739e-04
 [4,] 13  3.793605e-03
 [5,] 14  2.113580e-02
 [6,] 15  7.496164e-02
 [7,] 16  1.733488e-01
 [8,] 17  2.622083e-01
 [9,] 18  2.549247e-01
[10,] 19  1.520603e-01
[11,] 20  5.017991e-02
[12,] 21  6.951330e-03
```

1. The two sided p-value is given by
   $\sum_{x \in C} P(X = x | X + Y = 52)$ where
   $C = \{10, 11, 12, 13, 14, 21\}$ which is equal to 0.0323.

2. NOTE: In SAS, you can actually infer the one sided Fisher's exact p-value from the output below. The "left-sided" p-value corresponds to the alternative that the obtained frequency of cell (1,1) of SAS 2×2 table, that is 4 is the maximal value. In other words, the proportion of failure in PVCR is lower than proportion of failure in Prednisone, equivalently the proportion of success in Prednisone is lower than proportion of success in PVCR

```
                Fisher's Exact Test
        Cell (1,1) Frequency (F)          4
        Left-sided Pr <= F            0.0254
        Right-sided Pr >= F          0.9958
```

SAS Code

```
Data trial;
input drug $ outcome$ count;
datalines;
pred S 14
pred F 7
PVCR S 38
PVCR F 4
;
run;

proc freq data=trial;
tables drug*outcome/chisq;
weight count;
run;
```

SAS Output

```
              The FREQ Procedure

          Table of drug by outcome

      drug      outcome

      Frequency
      Percent
      Row Pct
      Col Pct   F        S         Total

      PVCR            4       38         42
                   6.35    60.32     66.67
                   9.52    90.48
                  36.36    73.08

      pred            7       14         21
                  11.11    22.22     33.33
                  33.33    66.67
                  63.64    26.92

      Total          11       52         63
                  17.46    82.54    100.00
```

SAS Output

```
                        Statistics for Table of drug by outcome

            Statistic                    DF      Value      Prob

            Chi-Square                    1      5.5070     0.0189
            Likelihood Ratio Chi-Square   1      5.2010     0.0226
            Continuity Adj. Chi-Square    1      3.9788     0.0461
            Mantel-Haenszel Chi-Square    1      5.4196     0.0199
            Phi Coefficient                     -0.2957
            Contingency Coefficient              0.2835
            Cramer's V                          -0.2957

            WARNING: 25% of the cells have expected counts less
                     than 5. Chi-Square may not be a valid test.


                            Fisher's Exact Test

                    Cell (1,1) Frequency (F)       4
                    Left-sided Pr <= F          0.0254
                    Right-sided Pr >= F         0.9958

                    Table Probability (P)       0.0211
                    Two-sided Pr <= P           0.0323

                        Sample Size = 63
```

R Code and Output

```
> obs <- matrix(c(14,7,38,4),nrow=2,byrow=T)
> fisher.test(obs)

Fisher's Exact Test for Count Data

data:  obs
p-value = 0.03232
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.03998108 1.00410538
sample estimates:
odds ratio
 0.2166817
```

Inference on two proportions, large samples

- If X and Y are large, we can use the normal distribution
- Let X and Y be the number of successes sample 1 and sample 2, respectively.
- Point estimators $\hat{p}_1 = X/n_1$, $\hat{p}_2 = Y/n_2$,
  $\hat{p}_1 - \hat{p}_2 = X/n_1 - Y/n_2$
- The CLT shows that if n_i is large

Inference on two proportions, large samples

- If samples are independent and p_i known for $i = 1, 2$, it follows

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1)$$

- This approximation is good if $n_i p_i (1 - p_i) \geq 10$ for $i = 1, 2$

Inference on two proportions, large samples

- If samples are independent and p_i unknown for $i = 1, 2$, Slutsky/CLT imply

$$Z^* = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0, 1)$$

  for sufficiently large n_1 and n_2 (rule of thumb: $n_i \hat{p}_i (1 - \hat{p}_i) \geq 10$ for $i = 1, 2$)

Confidence interval for $p_1 - p_2$ for large samples

$$1 - \alpha = P\left(-z_{\frac{\alpha}{2}} \le Z^* \le z_{\frac{\alpha}{2}}\right)$$

$$= P\left(\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2}S \le p_1 - p_2 \le \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2}S\right)$$

where $S = \sqrt{\dfrac{\widehat{p}_1(1 - \widehat{p}_1)}{n_1} + \dfrac{\widehat{p}_2(1 - \widehat{p}_2)}{n_2}}$

$100(1-\alpha)\%$ large samples CI for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}S$$

Hypothesis test on p_1 and p_2 for large samples

- Suppose $H_0 : p_1 - p_2 = \Delta$

- Test statistic

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2 - \Delta}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1) \text{ under } H_0$$

- For $\Delta = 0$, this reduces to $H_0 : p_1 = p_2$. One may use the pooled proportion in the denominator

and the test statistic

Hypothesis test on p_1 and p_2 for large samples

1. One tailed $H_0 : p_1 - p_2 = \Delta$ vs $H_1 : p_1 - p_2 > \Delta$
   - At significance level α, reject H_0 in favor of H_a if $Z_0 \geq z_\alpha$
   - Alternatively, if p-value= $P(Z_0 \geq z_0 | H_0) \leq \alpha$, reject H_0

2. One tailed $H_0 : p_1 - p_2 = \Delta$ vs $H_1 : p_1 - p_2 < \Delta$
   - At significance level α, reject H_0 in favor of H_a if $Z_0 \leq -z_\alpha$
   - Alternatively, if p-value= $P(Z_0 \leq z_0 | H_0) \leq \alpha$, reject H_0

3. Two tailed $H_0 : p_1 - p_2 = \Delta$ vs $H_1 : p_1 - p_2 \neq \Delta$
   - At significance level α, reject H_0 in favor of H_a if $|Z_0| \geq z_{\alpha/2}$
   - Alternatively, if p-value= $P(|Z_0| \geq |z_0| | H_0) = 2P(Z_0 \geq |z_0| | H_0) \leq \alpha$, reject H_0

Example 1. A random sample of Democrats and a random sample of Republicans were polled on an issue. Of 200 Republicans, 90 would vote yes on the issue; of 100 democrats, 58 would vote yes. Let p_1 and p_2 denote respectively the proportions of all Democrats or all Republicans who would vote yes on this issue.

(a) Construct a 95% confidence interval for $p_1 - p_2$

(b) Can we say that more Democrats than Republicans favor the issue at the 1% level of significance? Report the p-value.

Inference on two proportions: χ^2 test

- Alternative test of $H_0 : p_1 = p_2$ is χ^2 test
- Recall 2 x 2 table

	Success	Failure	
Sample 1	$n_{11} = x$	$n_{12} = n_1 - x$	n_1
Sample 2	$n_{21} = y$	$n_{22} = n_2 - y$	n_2
	$m_1 = x + y$	$m_2 = n - m_1$	n

- It can be shown that under H_0, the statistic

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_1 n_2 m_1 m_2} \sim \chi_1^2$$

- Rejection region for $H_a : p_1 \neq p_2$

$$C_\alpha = \{X^2 : X^2 \geq \chi_{1,\alpha,U}^2\}$$

Inference on two proportions: χ^2 test

- Aka "Pearson" chi-square statistic
- Equivalent form

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - En_{ij})^2}{En_{ij}}$$

where $E(n_{ij}) = n_i m_j / n$

- We will see this again for $r \times c$ tables

Inference on two proportions: Summary

- For small samples, use Fisher's Exact Test
- For large samples and two tailed test $H_a : p_1 \neq p_2$, use χ^2 or Z test, i.e., rejection region

$$C_\alpha = \{X^2 : X^2 > \chi^2_{1,\alpha,U}\} \text{ or } C_\alpha = \{z : |z| > z_{\alpha/2}\}$$

- For large samples and one tailed test $H_a : p_1 < p_2$ or $H_a : p_1 > p_2$, use Z test, i.e., rejection region

$$C_\alpha = \{z : z < -z_\alpha\} \text{ or } C_\alpha = \{z : z > z_\alpha\}$$

Matched or Paired Observations

- In some studies, subjects occur naturally in pairs or matches; e.g., twins or response under two conditions
- If we want to compare binary responses in matched pairs, the assumption of independence is violated

Matched or Paired Observations: McNemar's test

▶ The data are of the form

	Condition 2 response	
	Yes	No
Condition 1 response		
Yes	a	b
No	c	d

▶ Let A, B, C, D be the random variables corresponding to the observed counts a, b, c, d

$$A + B + C + D = n, \text{ and } p_A + p_B + p_C + p_D = 1$$

Then $(A, B, C, D) \sim$

Matched or Paired Observations: McNemar's test

- Response rate under condition 1 is $p_1 = p_A + p_B$
- Response rate under condition 2 is $p_2 = p_A + p_C$
- Consider $H_0 : p_1 = p_2$ (no difference between response rate between the two conditions)

$$p_1 = p_2 \Leftrightarrow$$

- The test statistic for comparing p_1 to p_2 (or equivalently p_B to p_C) can be derived using the conditional distribution of

$$B|B + C = m \sim B\left(m, p = \frac{p_B}{p_B + p_C}\right)$$

Thus $H_0 : p_1 = p_2 \Leftrightarrow H_0 : p = 1/2$

- This is the McNemar's test

McNemar's Test

1. One tailed test: $H_0 : p_1 = p_2$ vs $H_a : p_1 > p_2$
   or $H_0 : p = 1/2$ vs $H_a : p > 1/2$

   $$\text{p-value} = p_U = P(B \geq b | B + C = m) = \sum_{i=b}^{m} \binom{m}{i} \left(\frac{1}{2}\right)^m$$

2. One tailed test: $H_0 : p_1 = p_2$ vs $H_a : p_1 < p_2$
   or $H_0 : p = 1/2$ vs $H_a : p < 1/2$

   $$\text{p-value} = p_L = P(B \leq b | B + C = m) = \sum_{i=0}^{b} \binom{m}{i} \left(\frac{1}{2}\right)^m$$

3. Two tailed test: $H_0 : p_1 = p_2$ vs $H_a : p_1 \neq p_2$
   or $H_0 : p = 1/2$ vs $H_a : p \neq 1/2$

   $$\text{p-value} = 2\min(p_L, p_U)$$

   is the formula presented in your textbook. A pitfall of this
   approach is that the value can exceed 1. Alternative
   approach is by summing all the probabilities
   $\leq P(B = b | B + C = m)$ (Similar to a two tailed exact

Example: A preference poll of a panel of 75 voters was conducted before and after a TV debate during the campaign for the 1980 presidential election between Jimmy Carter and Ronald Reagan. Test whether there was a significant shift from Carter as a result of the TV debate.

	Preference after	
Preference before	Carter	Reagan
Carter	28	13
Reagan	7	27

SAS Code

```
Data election;
input before $ after $ count;
datalines;
Carter Carter 28
Carter Reagan 13
Reagan Reagan 27
Reagan Carter 7
 ;
run;

proc freq data=election;
exact agree;
tables before*after/agree;
weight count;
run;
```

SAS Output

```
                    The FREQ Procedure

             Table of before by after

         before      after

         Frequency
         Percent
         Row Pct
         Col Pct   Carter   Reagan     Total

         Carter        28       13        41
                    37.33    17.33     54.67
                    68.29    31.71
                    80.00    32.50

         Reagan         7       27        34
                     9.33    36.00     45.33
                    20.59    79.41
                    20.00    67.50

         Total         35       40        75
                    46.67    53.33    100.00
```

SAS Output

```
Statistics for Table of before by after

            McNemar's Test

    Statistic (S)          1.8000
    DF                          1
    Asymptotic Pr >  S     0.1797
    Exact        Pr >= S   0.2632  <---

            The SAS System
          The FREQ Procedure

Statistics for Table of before by after

        Simple Kappa Coefficient

    Kappa (K)              0.4700
    ASE                    0.1003
    95% Lower Conf Limit   0.2734
    95% Upper Conf Limit   0.6666

        Test of H0: Kappa = 0

    ASE under H0           0.1140
    Z                      4.1225
    One-sided Pr >  Z      <.0001
    Two-sided Pr > |Z|     <.0001

    Exact Test
    One-sided Pr >=  K     <.0001
    Two-sided Pr >= |K|    <.0001
```

R Code and Output

```
> obs <- matrix(c(28,13,7,27),nrow=2,byrow=T)
> rownames(obs) = colnames(obs) <- c('Carter','Reagan')
> names(dimnames(obs)) <- c('Before','After')
> mcnemar.test(obs)

McNemar's Chi-squared test with continuity correction

data:  obs
McNemar's chi-squared = 1.25, df = 1, p-value = 0.2636

> mcnemar.test(obs,correct=FALSE)

McNemar's Chi-squared test

data:  obs
McNemar's chi-squared = 1.8, df = 1, p-value = 0.1797
```

$$\chi^2 = \frac{(b-c)^2}{(b+c)}$$

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

R Code and Output

```
> library(exact2x2)
> exact2x2(obs,paired=TRUE,alternative='greater')

Exact McNemar-type test

data:  obs
b = 13, c = 7, p-value = 0.1316
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.7920039        Inf
sample estimates:
odds ratio
  1.857143
```