Lecture 2: Statistical Inference for Binary Data

Lin Chen

Department of Public Health Sciences
The University of Chicago

Bernoulli random variable

A binary response outcome can be represented by two familiar probability models: the Bernoulli and the binomial

- *Bernoulli distribution:* is often used to describes a binary variable that is either a success (1) or a failure (0).
- *Response variable:* R = 0 or 1
- *Parameter:* $p = P(R = 1)$, "success probability"
- *Likelihood function:* $\underset{\text{Describes p, given that R equals r}}{L(p|R \underline{=} r)} = \overset{\text{Success}}{p^r} \overset{\text{nonsuccess}}{(1-p)^{1-r}}, \quad r = \overset{\text{Response}}{0}$ *or* 1
  - Likelihood function measures the goodness of fit of a statistical model/parameter to a sample of data for given values of the unknown parameters. **Likelihood is a function of model parameters.** $L(p|R = r) = \Pr(R = r|p)$ but the latter is the density function and is a function of data/random variable.
- *Mean:* $\mathrm{E}(R) = 0 \times P(R = 0) + 1 \times P(R = 1) = p$
- *Variance:* $\mathrm{Var}(R) = \mathrm{E}(R^2) - \{\mathrm{E}(R)\}^2 = p(1-p)$

Binomial random variable

- *Binomial:* $Y = k$ count of successes in n independent trials, each with the same probability p of success. This is equivalent to the sum of n independent Bernoulli random variables, i.e.,

$$Y = R_1 + R_2 + \cdots + R_n \sim \textit{Binomial}(n, p) \tag{1}$$

- *Likelihood function:*

$$\mathrm{L}(p) = P(Y = k) = \binom{\overset{\text{\# of Trials}}{n}}{\underset{\text{\# of Event of Interest}}{k}} p^k (1-p)^{(n-k)} \tag{2}$$

- *Mean:* $\mathrm{E}(Y) = np$ (n tries times probability p per try)
- *Variance:* $\mathrm{Var}(Y) = np(1-p)$

Properties of the binomial distribution

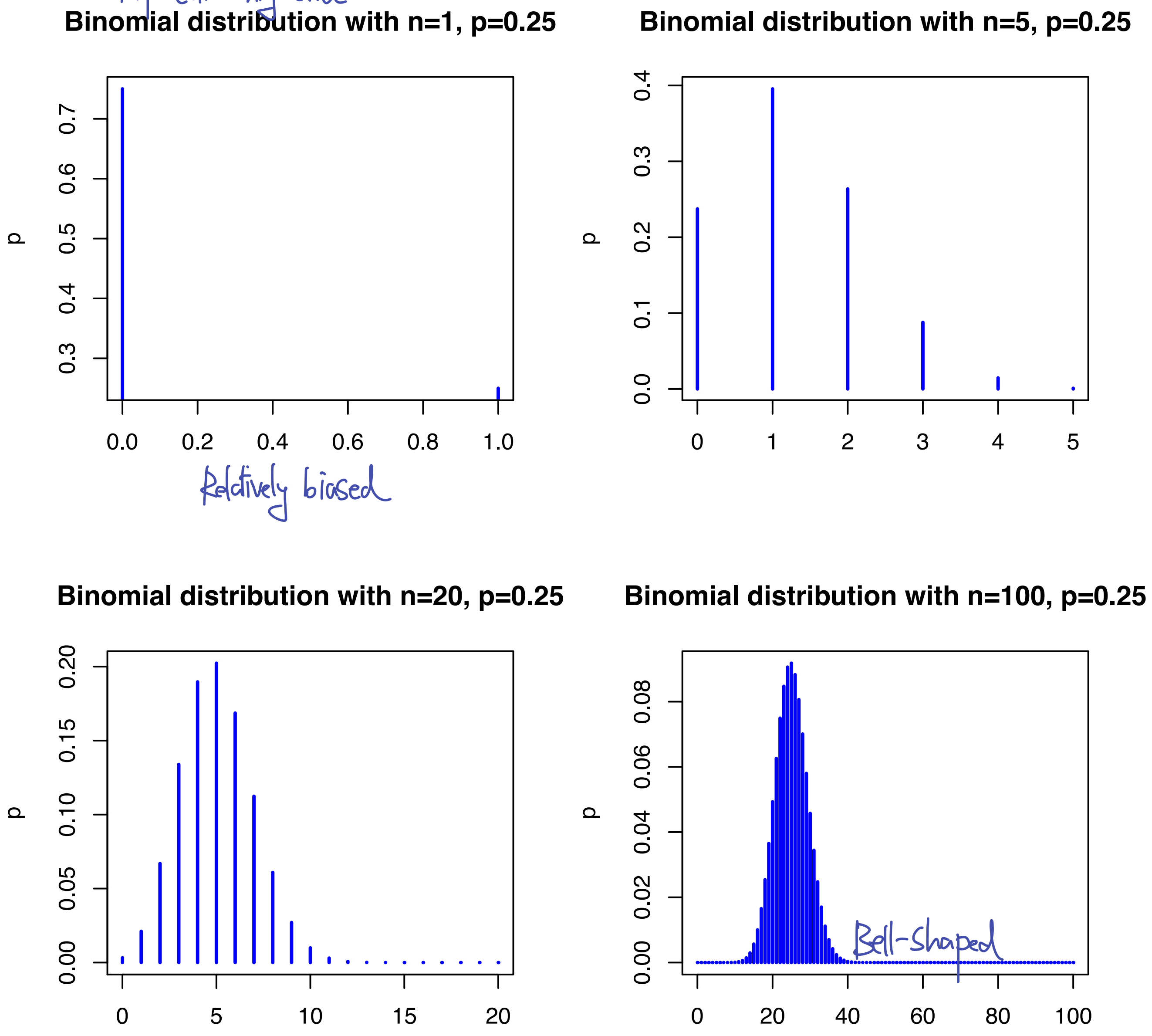
- The binomial distribution is discrete; it can only take on nonnegative integers, namely, $0, 1, ..., n$.
- Exact probabilities for the binomial for a given n and p can be obtained by computing. We typically compute the probability of some k or greater (fewer) successes under some n and p
- The *central limit theorem* says that the probability function of the binomial can be approximated by the probability function of a normal random variable with the same mean and variance when n is large, e.g.,

1. Independent, Random Samples
2. Sufficiently Large Sample Size
3. Approx. Normal Distribution
4. Standard Error

$$Y \sim \overset{\text{Normal Distribution}}{\underline{\mathrm{N}}}(np, np(1-p)) \text{ approximately} \tag{3}$$

Binomial distribution

Flip coin only once



Binomial distribution with n=1, p=0.25

Binomial distribution with n=5, p=0.25

Binomial distribution with n=20, p=0.25

Binomial distribution with n=100, p=0.25

Relatively biased

Bell-Shaped

Approaching Normal Distribution...

Normal approximation to the binomial

Extensive calculation, Inefficient.

↑

When does the Normal approximation apply? Prefer Normal over Binomial

- *Rule of thumb:* CLT works ok for binomial distribution when $np(1-p) \geq 2$ Larger the better.
- Intuitively, it works when the sample size is not small and the probability of event is not too large or small.
- *Standardization:* $\frac{Y-np}{\sqrt{np(1-p)}} \sim \mathrm{N}(0,1)$ *approximately*

With normal approximation, point estimate of p

From a collection of n trials and Y, we estimate the success probability

- *Estimator:* $\hat{p} \equiv \frac{Y}{n}$
- $\hat{p}$ is an unbiased estimator for p: $\mathrm{E}(\hat{p}) = p$
- $\hat{p} \sim \mathrm{N}(p, \frac{p(1-p)}{n})$ approximately when n is large.
- *Standard error of* $\hat{p}$*:* $se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

How to calculate confidence intervals of p

To assess how reliable $\hat{p}$ is, we can construct a *confidence interval* on p. Two ways:

- If we consider p following an approximate normal distribution, the approximate $1-\alpha$ confidence interval (CLT):

  $$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  where $z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ point of the standard normal distribution.
- If we consider p following a binomial distribution, the exact confidence interval is $(p_L,\ p_U)$, where p_L and p_U satisfy $P(Y \geq y \mid p = p_L) = \frac{\alpha}{2}$ and $P(Y \leq y \mid p = p_U) = \frac{\alpha}{2}$. Those probabilities can be calculated based on the probability distribution functions and the cumulative distribution functions of Binomial variables. Not easy to calculate and also tedious, but we have computers!

Example 1: binary outcome

Time magazine reported the result of a telephone poll of 800 adult Americans. The question posed of the Americans who were surveyed was: "Should the federal tax on cigarettes be raised to pay for health care reform?" The results of the survey were summarized in Table 1.

Table 1

Group	Yes	No	Total
Smokers	41	154	195
Non - Smokers	351	254	605

Example 1: Estimation of proportion for one group

41 out of 195 smokers voted "yes" to the question "Should the federal tax on cigarettes be raised to pay for health care reform?"

- $\hat{p} = \frac{y}{n} = \frac{41}{195} = 0.2103$
- $\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0292$
- Approximate 95% CI: $(0.1531,\ 0.2675)$
- Exact 95% CI: $(0.1553,\ 0.2742)$
- We can use Stata to directly calculate these from Y and n values:
    - Stata command for the estimate $\hat{p}$, $\text{se}(\hat{p})$ and the approximate 95% CI is "`cii proportion` n y". Note that `ci` means confidence interval, the second `i` means immediate command no use of variable, "," means options, "`wald`" means using normal approximation).

      Confidence interval input

      total, success

      ```
      cii proportions 195 41, wald
      ```

      Z-test: Normal Approximation
    - Stata command for the estimate $\hat{p}$, $\text{se}(\hat{p})$ and the exact 95% CI (the default is exact binomial calculation):

      ```
      cii proportions 195 41
      ```

Here is the execution of the commands

```
. cii proportions 195 41, wald

                                                         Normal Distribution
                                                         -- Binomial Wald ---
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        195    .2102564     .029181        .1530627    .2674501

. cii proportions 195 41

                                                         Binomial Distribution
                                                         -- Binomial Exact --
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        195    .2102564     .029181        .1553102    .2742331
```

The exact confidence interval is not symmetric.

If data is loaded, the command `ci proportions smoke`, where `smoke` is the 0/1 variable for each of the 195 individuals, would give the same answer.

Example 1: Testing for one proportion

Historical data suggested that only 10% or less of the smokers voted yes for raising cigarette tax. Is the current data consistent with historical data the significance level 0.05?

- $H_0 : p = 0.1,\ H_1 : p \neq 0.1$
- Calculate p - value: $z = \frac{\overset{45/195}{\hat{p}} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.21-0.1}{\sqrt{\frac{0.1\cdot0.9}{195}}}$ = 5.132 ·Including Po : Assuming null is true.

  p-value = $P(|Z| \geq |z|) < 0.001$
- In Stata, use the immediate function for `prtest` (`prtesti`) followed by $n\ \hat{p}\ p_0$ to test for one proportion using normal approx.

\# of obs. prop null

```
. prtesti 195 .21025641 0.1
```

$\longrightarrow n, \hat{p}, p_0$

Probability 45/195

```
One-sample test of proportion                   x: Number of obs =      195
------------------------------------------------------------------------------
             |       Mean   Std. Err.                 [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .2102564    .029181                   .1530627    .2674501
------------------------------------------------------------------------------
    p = proportion(x)                                         z =    5.1322
Ho: p = 0.1
     Ha: p < 0.1                Ha: p != 0.1                 Ha: p > 0.1
 Pr(Z < z) = 1.0000        Pr(|Z| > |z|) = 0.0000          Pr(Z > z) = 0.0000
```

The above test is based on normal approximation, to obtain the exact p-value, use Stata `bitesti` followed by $n\ y\ p_0$:

```
. bitesti 195 41 0.1

        N     Observed k    Expected k    Assumed p    Observed p
------------------------------------------------------------------
      195          41           19.5        0.10000       0.21026

  Pr(k >= 41)             = 0.000004  (one-sided test)
  Pr(k <= 41)             = 0.999998  (one-sided test)
  Pr(k <= 3 or k >= 41)   = 0.000006  (two-sided test)
```

Binomial

→ n, y, p_0 Input different than "pr".

Conclusion: We reject the null and the data is inconsistent with historical data. Statistical Significance

Ex 1: Comparing two population proportions

Group	Yes	No	Total
Smokers	41	154	195
Non - Smokers	351	254	605

- *Estimation of* $p_1 - p_2$

  difference in proportions

  - *Estimate:* $\hat{p}_1 - \hat{p}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2} = 41/195 - 351/605 = -0.3699$ Mean
  - $\text{se}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.0354$ S.E.
  - Approximate $1-\alpha$ confidence interval:
    $\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \cdot \text{se}(\hat{p}_1 - \hat{p}_2)$
  - 95% CI on difference in proportions: (-0.4393, -0.3005)
- In Stata, you may do the calculation using `display` or `di`

```
. display 41/195-351/605
-.36990888

. display sqrt((41*154)/(195^3)+(351*254)/(605^3))
.03541373

etc
```

Comparing two proportions – Hypothesis test

Group	Yes	No	Total
Smokers	41	154	195
Non - Smokers	351	254	605

Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that the two populations - smokers and non-smokers- differ significantly with respect to their opinions?

- *Hypothesis testing:*
  - $H_0 : p_1 = p_2$ vs. $H_1 : p_1 \neq p_2$
  - Under H_0, $p_1 = p_2 = p$, $\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})$, using "pooled estimate" of p (estimate under H_0), $\hat{p} = \frac{y_1+y_2}{n_1+n_2}$, $\mathrm{se}(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$

    Combined information from both samples
    To obtain a single overall estimate.
  - $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = -8.9859$
  - The absolute value of Z statistic is large. P-value is very small. H_0 is rejected at the 0.05 significance level.

```
. di (41/195-351/605)/sqrt((392/800)*(408/800)*(1/195+1/605))
-8.985901
. di 2*normal(-8.985901)
2.566e-19
```

Two Sample

Or use `prtest` immediate command to perform a normal approximation test, with `prtesti` followed by $n_1\ p_1\ n_2\ p_2$:

```
. prtesti 195 0.2103 605 0.5802

Two-sample test of proportions                    x: Number of obs =      195
                                                  y: Number of obs =      605
------------------------------------------------------------------------------
             |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |      .2103   .0291832                      .1531019    .2674981
           y |      .5802   .0200647                      .5408739    .6195261
-------------+----------------------------------------------------------------
        diff |     -.3699   .0354154                      -.439313   -.300487
             |  under Ho:   .0411655    -8.99   0.000
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =  -8.9857
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0000         Pr(|Z| > |z|) = 0.0000          Pr(Z > z) = 1.0000
```

Continuing Ex 1: Comparing two proportions using a 2 by 2 contingency table

When H_1 is two-sided, an equivalent test statistics is z^2. Z^2 follows a χ_1^2 distribution. Chi-Square

Group	1. Yes	2. No	Total
1. Smokers	41(O_{11})	154 (O_{12})	195 (r_1)
2. Non - Smokers	351(O_{21})	254(O_{22})	605(r_2)
Total	392(c_1)	408(c_2)	800(n)

observed - expected

$X^2 = \sum_{i=1}^{i=2} \sum_{j=1}^{j=2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, where $E_{ij} = \frac{r_i c_j}{n}$ which is the estimated expected count in the i^{th} row and the j^{th} column under the null. Under $H_0 : p_1 = p_2$, X^2 follows a χ_1^2 distribution.

This test statistic is refered to as *Pearson's χ^2 -statistic*. Note that $X^2 = Z^2$ where Z is the Wald statistic we computed on page 15.

In Stata, you may use `tabi 41 154 \ 351 254, chi` to test for dependency of row and column binary variables. The four numbers are all counts in the 2x2 table (not the totals). The option `chi` will show results for the χ^2 test and `exp` will provide the expected frequencies/counts.

Row by Row...

```
. tabi 41 154 \ 351 254, chi exp

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
+--------------------+

           |          col
       row |          1           2 |     Total
-----------+------------------------+----------
         1 |         41         154 |       195
           |       95.5        99.5 |     195.0
-----------+------------------------+----------
         2 |        351         254 |       605
           |      296.4       308.6 |     605.0
-----------+------------------------+----------
     Total |        392         408 |       800
           |      392.0       408.0 |     800.0

          Pearson chi2(1) =  80.7464   Pr = 0.000
```

STATA:
Doesn't get imported

A disadvantage of χ^2 test is that it provides only a significance measure (p-value), and does not provide the direction of association effects, nor the magnitude of effects.

Example 2: Comparing K proportions: 2 by K contingency table

Row Column

This is the example we discussed in Lecture 1. The numbers of cuttings surviving for each of the four combinations of planting time and length of cutting.

Ultimate condition of the cutting	Planted at once		Planted in spring		Total
	Short	Long	Short	Long	
Alive	107	156	31	84	378
Dead	133	84	209	156	582
Total	240	240	240	240	960

predictor are #of columns

- $H_0 : p_1 = p_2 = ... = p_K$ ($K = 4$ here) dof: $\boxed{\overset{2-1}{(\# r-1)} \times \overset{4-1}{(\# C-1)}}$

  ex) $1 \times 3 = 3$
- $X^2 = \sum_{i=1}^{i=2} \sum_{j=1}^{j=K} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, X^2 has a χ^2_{K-1} – distribution under the null hypothesis that there is no association between the survival status (row variable) and two explanatory variables (column variable).

In Stata, you may use `tabi` followed by the cell values in the contingency table. The option `col` will provide column frequencies (each column adds up to 100%). You may also use the option `row` for row frequencies.

Row by Row...

```
. tabi 107 156 31 84 \ 133 84 209 156, col chi2
```

Specify

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |                          col
       row |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |       107        156         31         84 |       378 
           |     44.58      65.00      12.92      35.00 |     39.38 
-----------+--------------------------------------------+----------
         2 |       133         84        209        156 |       582 
           |     55.42      35.00      87.08      65.00 |     60.62 
-----------+--------------------------------------------+----------
     Total |       240        240        240        240 |       960 
           |    100.00     100.00     100.00     100.00 |    100.00 

          Pearson chi2(3) = 141.0527   Pr = 0.000
```

Example 3: Association test for r by c contingency table

In the dataset "Popular Kids," students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below:

Goals	Grade 4	Grade 5	Grade 6	Total
Good grades	49	50	69	168
Athletic ability	19	22	28	69
Popularity	24	36	38	98
Total	92	108	135	335

Data source: Chase, M.A and Dummer, G.M. (1992), "The Role of Sports as a Social Determinant for Children," Research Quarterly for Exercise and Sport, 63, 418-424.

- H_0: no association between row variable and column variable
- H_1 : there is association between row variable and column variable
- Test statistic $X^2 = \sum_{i=1}^{i=r}\sum_{j=1}^{j=c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, under H_0, X^2 follows $\chi^2_{(r-1)(c-1)}$.

In Stata:

```
. tabi 49 50 69 \ 19 22 28 \ 24 36 38, col chi

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+
```

dof: $2\times2=4$

```
           |               col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        49         50         69 |       168 
           |     53.26      46.30      51.11 |     50.15 
-----------+---------------------------------+----------
         2 |        19         22         28 |        69 
           |     20.65      20.37      20.74 |     20.60 
-----------+---------------------------------+----------
         3 |        24         36         38 |        98 
           |     26.09      33.33      28.15 |     29.25 
-----------+---------------------------------+----------
     Total |        92        108        135 |       335 
           |    100.00     100.00     100.00 |    100.00 

          Pearson chi2(4) =   1.5126   Pr = 0.824
```

The preference of goals do not differ much among students in grades 4-6.

Fail to reject H_0

Back to 2x2 Tables: Odds and odds ratio

We want to concentrate on 2x2 tables for a bit and focus on an effect measure called the odds ratio

	outcome		
Exposure	Yes	No	Total
Exposed	a	b	$a+b$
Non-exposed	c	d	$c+d$
	$a+c$	$b+d$	n

prevalence v.s. odds ↘

transformation probability

for exposed

- *Odds of a success:* $\boxed{\frac{p}{1-p}}$, can be estimated by $\frac{\hat{p}}{1-\hat{p}} = \frac{a/(a+b)}{b/(a+b)} = a/b$

for unexposed c/d

- odds $\frac{p}{1-p} \in [0,\infty)$, small values associated with lower probability: odds of 0.5 = 33% probability, odds of 1 = 50%, odds of 3 = 75%
- We can compute odds of success among exposed and non-exposed and contrast these.

Odds ratio

Odds, like probability, is about one group. Odds ratio compare the ratio of odds for **two groups**.

- *Odds ratio:* $\phi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$, which is the ratio of the odds of a success in one set of binary data relative to the other.
- Knowing two of the three (ϕ, p_1, p_2), one would be able to estimate the other.
- The odds ratio is a measure of the extent to which two success probabilities differ:
    - $\phi = 1 \iff p_1 = p_2$;
    - $\phi > 1 \iff p_1 > p_2$;
    - $\phi < 1 \iff p_1 < p_2$.

Inference about odds ratio

Group	Yes	No
Smokers	41 (a)	154(b)
Non - Smokers	351(c)	254(d)

Woolf's method – an approximated CI for odds ratio (OR)

- Odds ratio comparing smoker versus non-smokers: $\hat{\phi} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = (a/b)/(c/d) = \frac{ad}{bc} = 0.1927$ (cross-product ratio)
- Approximate $\text{se}(\log(\hat{\phi})) \approx \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = 0.1941$ → $OR = \frac{p^1/1-p^1}{p^2/1-p^2}$
- 95% CI for log OR
  $\log(\phi) : \log(0.1927) \pm 1.96 \times 0.1941 = (-2.0272,\ -1.2664)$

  $\log OR = \left(\frac{p^1}{1-p^1}\right) - \left(\frac{p^2}{1-p_2}\right)$
- 95% CI for OR $\phi : (e^{-2.0272},\ e^{-1.2664}) = (0.1317,\ 0.2818)$ – exponentiating the lower and upper bounds for log OR CI
- The 95% CI for ϕ does not cover 1 and smaller than 1, which indicates that the evidence that the odds of yes is smaller among smokers is certainly significant at the 5% level.

Stata has several modules for producing commonly used statistics in epidemiology. One of these is the `cs` (cohort studies) command. The option or produces the odds ratio and option `woolf` produces the asymptotic (approximated) confidence interval (we use `cs` because it also gives risk difference w/ci, more discussions later)

Approx. OR with cohort, c.c. also appropriate.

```
. csi 41 351 154 254, or woolf

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        41         351  |        392
        Noncases |       154         254  |        408
-----------------+------------------------+------------
           Total |       195         605  |        800
                 |                        |
            Risk |  .2102564    .5801653  |        .49
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.3699089       |   -.4393185   -.3004992
      Risk ratio |         .3624078       |    .2738095    .4796744
 Prev. frac. ex. |         .6375922       |    .5203256    .7261905
 Prev. frac. pop |         .1554131       |
      Odds ratio |         .1926592       |     .131699    .2818363 (Woolf)
                 +-------------------------------------------------
                               chi2(1) =    80.75  Pr>chi2 = 0.0000
```

Summary

Basic Methods for Binary Outcome Data

- Methods for binary outcome data include familiar tools from elementary statistics (exact vs normal approx. of CIs and tests for one proportion or difference in two proportions)
- Tests for difference in proportions $(H_0 : p_1 = p_2 = 0)$ and tests of association (χ^2 test or testing for OR being 1) in tables yield same results (are identical or asymptotically equivalent test in many cases).
- The odds ratio is another measure of association, will be used
  2 probability
  extensively as metric in binary data modeling
- Extension to effect of multiple factors (the predictor of interest and covariates) on the odds ratio available (next)