Compare Binary Data Between two groups

Review

Compare means | Compare medians

Setting	Parametric Test	Nonparametric Test
Matched pairs	Paired t-test (one sample t-test on the differences)	Wilcoxon signed rank test
Two independent samples	Two-sample t-test	Wilcoxon Rank Sum test

Data File

ID	group	outcome
1	1	611
2	1	621
3	1	614
4	1	593
5	1	593
6	1	653
7	1	600
8	1	554
9	1	603
10	1	569
11	2	635
12	2	605
13	2	638
14	2	594
15	2	599
16	2	632
17	2	631
18	2	588
19	2	607

Review

Data File

ID	group	outcome
1	1	1
2	1	1
3	1	0
4	1	0
5	1	0
6	1	1
7	1	0
8	1	1
9	1	0
10	1	1
11	2	0
12	2	0
13	2	0
14	2	1
15	2	0
16	2	0
17	2	1
18	2	1
19	2	0

Compare means | Compare medians

Setting	Parametric Test	Nonparametric Test
Matched pairs	Paired t-test (one sample t-test on the differences)	Wilcoxon signed rank test
Two independent samples	Two-sample t-test	Wilcoxon Rank Sum test

Chi-square Test or Fisher exact Test

Risk

Goal: Compare risk of PVD in smokers and non-smokers?

Example: cross-sectional study of peripheral vascular disease (PVD) in Scottish men

	PVD		
Cigarette Smoker	Yes	No	Total
Yes	15	1712	1727
No	41	3188	3229
Total	56	4900	4956

Risk is the Prob(disease/exposure to risk factor)

Measure Association of Exposure and Disease

- **Risk Difference**
- **Risk Ratio (Relative Risk)** ✓
- **Odds Ratio** ✓

Calculate Risks

	PVD		
Cigarette Smoker	Yes	No	Total
Yes	15	1712	1727
No	41	3188	3229
Total	56	4900	4956

$$\begin{aligned} p_S &= \text{probability of PVD in smokers} \\ &= \text{risk of PVD in smokers} \end{aligned}$$

$$\widehat{p}_S = 15/1727 = 0.009$$

$$\begin{aligned} p_{NS} &= \text{probability of PVD in non-smokers} \\ &= \text{risk of PVD in non-smokers} \end{aligned}$$

$$\widehat{p}_{NS} = 41/3229 = 0.013$$

Risk Difference

Risk Difference

$$\Delta \quad = \quad p_S - p_{NS}$$

In our example: $\widehat{\Delta} \quad = \quad \widehat{p}_S - \widehat{p}_{NS} \quad = \quad -0.004$

The magnitude of the difference (exposed vs. non-exposed) in risk often is interpreted in relation to the “baseline" (non-exposed) risk

Risk Ratio

- Risk ratio (relative risk)

$$\lambda = \frac{\text{risk of disease in exposed subjects}}{\text{risk of disease in non-exposed subjects}}$$

	PVD		
Cigarette Smoker	Yes	No	Total
Yes	15	1712	1727
No	41	3188	3229
Total	56	4900	4956

$$\lambda = \frac{p_S}{p_{NS}} = \frac{15/1727}{41/3229} = 0.68$$

Interpretation: We estimate that the risk (probability) of PVD in smokers is 68% of the risk of PVD in non-smokers

Odds

- Risk is a probability that quantifies the likelihood that an event occurs

- We may instead be interested in the likelihood of an event in relation to how unlikely it is

- Such a quantity is known as the odds of an event:

$$\text{Odds} = \frac{\text{Risk}}{1 - \text{Risk}}$$

- Example: If a disease occurs in 25% of adults, we can also say adults have an odds of $1/3$ of developing the disease.

Odds Ratio

- As relative risk is used to compare the risk of two groups, we can use the relative odds to compare the odds of two groups

- Relative odds is more commonly referred to as the odds ratio:

$$\psi = \frac{\text{Odds for Group 1}}{\text{Odds for Group 2}}$$

- An odds ratio is much harder to interpret than a relative risk

- However, an odds ratio is used much more often than a relative risk

Computing Odds Ratio from a 2 by 2 Table

- Thus, for a general 2 x 2 table of exposure and disease

	Disease		
Risk Factor	Yes	No	Total
Exposed	a	b	a+b
Non-exposed	c	d	c+d
Total	a+c	b+d	n

the odds ratio of disease for exposed versus non-exposed is

$$\widehat{\psi} = \frac{ad}{bc}$$

$$\begin{aligned}\widehat{\psi} &= \frac{\frac{a}{a+b} \Big/ \frac{b}{a+b}}{\frac{c}{c+d} \Big/ \frac{d}{c+d}} \\ &= \frac{a/b}{d/d} \\ &= \frac{ad}{bc}\end{aligned}$$

Example

	PVD		
Cigarette Smoker	Yes	No	Total
Yes	15	1712	1727
No	41	3188	3229
Total	56	4900	4956

- Risk difference = 15/1727 – 41/3229 = - 0.004
- Relative risk = $\frac{15/1727}{41/3229}$ = 0.6840
- Odds Ratio = $\frac{15\times 3188}{41\times 1712}$ = 0.68127

Chi-squared & Fisher’s Exact Test

Test Association of Exposure and Disease

2 × 2 table		
Exposed	Disease	
	Yes	No
Yes		
No		

Chi-squared Test

- We can also view our hypothesis test as:

$$H_o \; : \; \psi = 1 \quad \rightarrow \text{disease is not associated with exposure}$$
$$H_a \; : \; \psi \neq 1 \quad \rightarrow \text{disease is associated with exposure}$$

Relative Risk or Odd Ratio

or

$$H_o \; : \; \psi = 1 \quad \rightarrow \text{disease and exposure are independent}$$
$$H_a \; : \; \psi \neq 1 \quad \rightarrow \text{disease and exposure are dependent}$$

- We can compare these hypotheses using a Chi-squared test of association

Chi-squared Test of Association

- Thus, we have two 2x2 tables:

Observed:

	Disease		
Risk Factor	Yes	No	Total
Exposed	$a = O_{11}$	$b = O_{01}$	$a + b$
Non-exposed	$c = O_{10}$	$d = O_{00}$	$c + d$
Total	$a + c$	$b + d$	n

Expected:

	Disease		
Risk Factor	Yes	No	Total
Exposed	$(a+b)(a+c)/n$ $= E_{11}$	$(a+b)(b+d)/n$ $= E_{01}$	$a + b$
Non-exposed	$(c+d)(a+c)/n$ $= E_{10}$	$(c+d)(b+d)/n$ $= E_{00}$	$c + d$
Total	$a + c$	$b + d$	n

Example

Vitamin C	Developed Cold Yes	No	Total
Yes	17	122	139
No	31	109	140
Total	48	231	279

Calculate Expected Counts

$$E_{11} = 23.9 \qquad E_{01} = 115.1$$

$$E_{10} = 24.1 \qquad E_{00} = 115.9$$

How to construct the test ?

Chi-Squared Test

Hypothesis: H_0: there is no association H_1: there is an association

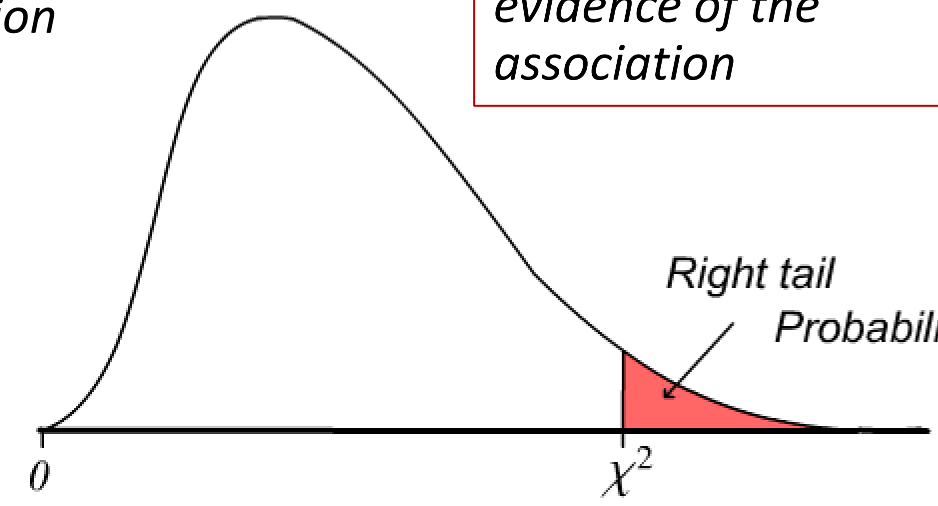
Test statistic:

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

Larger value of the statistic indicates more evidence of the association

Follows the chi-square distribution with degrees of freedom:

(# rows – 1)x(# cols – 1)



Fisher's Exact Test

- The p-value for a chi-squared test of association is approximate; the approximation does poorly in "small" samples

- A common rule-of-thumb for "small" is if any of the (expected) cell counts is ≤ 5

- In small samples, we do not use a large-sample approximation, but compute the exact p-value

- This approach was first proposed by R.A. Fisher, hence the name Fisher's Exact Test

Motivating Example

- Example: A retrospective case/control study among men aged 50-54 who died over a 1-month period; their dietary habits were ascertained from a close relative

	Salt Intake		
Disease	High	Low	Total
CVD	5	30	35
No CVD	2	23	25
Total	7	53	60

How to construct the test ?

Fisher's Exact Test

Risk Factor	Disease: Yes	Disease: No	Total
Exposed	?	?	a+b
Non-exposed	?	?	c+d
Total	a+c	b+d	n

- Fisher's Exact Test fixes the row and column totals (margins)
- Once we know one of the cells (a, b, c, d), we know them all because the margins are fixed

  (a) Determine every possible table that would still lead to the same margins as those observed

  (b) Compute the probability that each of the tables would be observed

  (c) p-value $=$ sum of probabilities corresponding to our observed table and any tables more extreme than our observed table

Analysis Using R

```
> riskratio.wald(M,rev = c("both"))
$data
       cold
vitc      N  Y Total
  N     109 31   140
  Y     122 17   139
  Total 231 48   279

$measure
    risk ratio with 95% C.I.
vitc  estimate     lower     upper
   N 1.0000000        NA        NA
   Y 0.5523323 0.3209178 0.9506203

$p.value
    two-sided
vitc midp.exact fisher.exact chi.square
   N         NA           NA         NA
   Y 0.02951602   0.03849249 0.02827186
```

```
> oddsratio.wald(M,rev = c("both"))
$data
       cold
vitc      N  Y Total
  N     109 31   140
  Y     122 17   139
  Total 231 48   279

$measure
    odds ratio with 95% C.I.
vitc  estimate     lower     upper
   N 1.0000000        NA        NA
   Y 0.4899524 0.2569419 0.9342709

$p.value
    two-sided
vitc midp.exact fisher.exact chi.square
   N         NA           NA         NA
   Y 0.02951602   0.03849249 0.02827186
```

Paired Data

- **McNemar**'s test for paired binary data

- Use ***exact*** McNemar's test for small data

- **R code**
  - https://www.statology.org/mcnemars-test-r/
  - https://yuzar-blog.netlify.app/posts/2022-02-20-mcnemar/