

2 x 2 tables

Apply a treatment to 20 mice from strains A and B, and observe survival.

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

Question: Are the survival rates in the two strains the same?

Gather 100 rats and determine whether they are infected with viruses A and B.

	I-B	NI-B	
I-A	9	9	18
NI-A	20	62	82
	29	71	100

Question: Is infection with virus A independent of infection with virus B?

Underlying probabilities

Observed data

		B		
		0	1	
A	0	n_{00}	n_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+}
		n_{+0}	n_{+1}	n

Underlying probabilities

		B		
		0	1	
A	0	p_{00}	p_{01}	p_{0+}
	1	p_{10}	p_{11}	p_{1+}
		p_{+0}	p_{+1}	1

Model:

$$(n_{00}, n_{01}, n_{10}, n_{11}) \sim \text{multinomial}(n, (p_{00}, p_{01}, p_{10}, p_{11}))$$

or

$$n_{01} \sim \text{binomial}(n_{0+}, p_{01}/p_{0+}) \text{ and } n_{11} \sim \text{binomial}(n_{1+}, p_{11}/p_{1+})$$

Conditional probabilities

Underlying probabilities

		B		
		0	1	
A	0	p_{00}	p_{01}	p_{0+}
	1	p_{10}	p_{11}	p_{1+}
		p_{+0}	p_{+1}	1

Conditional probabilities

$$\Pr(B = 1 \mid A = 0) = p_{01}/p_{0+}$$

$$\Pr(B = 1 \mid A = 1) = p_{11}/p_{1+}$$

$$\Pr(A = 1 \mid B = 0) = p_{10}/p_{+0}$$

$$\Pr(A = 1 \mid B = 1) = p_{11}/p_{+1}$$

The questions in the two examples are the same!

They both concern: $p_{01}/p_{0+} = p_{11}/p_{1+}$

Equivalently: $p_{ij} = p_{i+} \times p_{+j}$ for all i, j

This is a “composite hypothesis”

2 x 2 table

		B		
		0	1	
A	0	p_{00}	p_{01}	p_{0+}
	1	p_{10}	p_{11}	p_{1+}
		p_{+0}	p_{+1}	1

A different view

$$\boxed{p_{00} \quad p_{01} \quad p_{10} \quad p_{11}}$$

$$H_0: p_{ij} = p_{i+} \times p_{+j} \text{ for all } i, j$$

$$H_0: p_{ij} = p_{i+} \times p_{+j} \text{ for all } i, j$$

$$\text{degrees of freedom} = 4 - 2 - 1 = 1$$

Expected counts

Observed data				Expected counts					
		B				B			
		0	1			0	1		
A	0	n_{00}	n_{01}	n_{0+}	A	0	e_{00}	e_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+}		1	e_{10}	e_{11}	n_{1+}
		n_{+0}	n_{+1}	n			n_{+0}	n_{+1}	n

To get the expected counts **under the null hypothesis** we:

1. Estimate p_{1+} and p_{+1} by n_{1+}/n and n_{+1}/n , respectively. (i.e., MLEs under H_0 .)
2. Turn these into estimates of the p_{ij} .
3. Multiply these by the total sample size, n .

The expected counts

The expected count (assuming H_0) for the “11” cell is the following:

$$\begin{aligned}e_{11} &= n \times \hat{p}_{11} \\ &= n \times \hat{p}_{1+} \times \hat{p}_{+1} \\ &= n \times (n_{1+}/n) \times (n_{+1}/n) \\ &= (n_{1+} \times n_{+1})/n\end{aligned}$$

The other cells are similar.

We can then calculate a χ^2 or LRT statistic as before!

Example 1

Observed data

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

Expected counts

	N	Y	
A	14.5	5.5	20
B	14.5	5.5	20
	29	11	40

$$\chi^2 = \frac{(18-14.5)^2}{14.5} + \frac{(11-14.5)^2}{14.5} + \frac{(2-5.5)^2}{5.5} + \frac{(9-5.5)^2}{5.5} = 6.14$$

$$\text{LRT stat} = 2 \times [18 \log(\frac{18}{14.5}) + \dots + 9 \log(\frac{9}{5.5})] = 6.52$$

P-values (based on the asymptotic χ^2 (df = 1) approximation):
1.3% and 1.1%.

Example 2

Observed data

	I-B	NI-B	
I-A	9	9	18
NI-A	20	62	82
	29	71	100

Expected counts

	I-B	NI-B	
I-A	5.2	12.8	18
NI-A	23.8	58.2	82
	29	71	100

$$\chi^2 = \frac{(9-5.2)^2}{5.2} + \frac{(20-23.8)^2}{23.8} + \frac{(9-12.8)^2}{12.8} + \frac{(62-58.2)^2}{58.2} = 4.70$$

$$\text{LRT stat} = 2 \times [9 \log(\frac{9}{5.2}) + \dots + 62 \log(\frac{62}{58.2})] = 4.37$$

P-values (based on the asymptotic χ^2 (df = 1) approximation):
3.0% and 3.7%.

Fisher's exact test

Observed data

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

- Assume the null hypothesis (independence) is true.
- Constrain the marginal counts to be as observed.
- What's the chance of getting this exact table?

Hypergeometric distribution

- Imagine an urn with K white balls and $N - K$ black balls.
- Draw n balls **without** replacement.
- Let x = no. white balls in the sample.
- x follows a hypergeometric distribution (with parameters K , N , and n .)

	In urn		
	white	black	
sampled	x		n
not sampled			$N - n$
	K	$N - K$	N

Hypergeometric probabilities

Suppose $X \sim \text{hypergeometric}(N, K, n)$.

[i.e., no. white balls in sample of n , **without replacement** from an urn with K white and $N - K$ black]

$$\Pr(X = x) = \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}}$$

Example:

	In urn			$N = 40, K = 29, n = 20$
	0	1		
sampled	18		20	$\Pr(X = 18) = \frac{\binom{29}{18} \binom{40-29}{20-18}}{\binom{40}{20}} \approx 1.4\%$
not			20	
	29	11	40	

The hypergeometric in R

`dhyper(x, m, n, k)`

`phyper(q, m, n, k)`

`qhyper(p, m, n, k)`

`rhyper(nn, m, n, k)`

In R, things are set up so that

m = no. white balls in urn

n = no. black balls in urn

k = no. balls sampled (without replacement)

x = no. white balls in sample

Back to Fisher's exact test

Observed data

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

- Assume the null hypothesis (independence) is true.
- Constrain the marginal counts to be as observed.
- $\Pr(\text{observed table} \mid H_0) = \Pr(X=18)$ where $X \sim \text{hypergeometric}(N=40, K=29, n=20)$

Fisher's exact test

1. For all possible tables (with the observed marginal counts), calculate the relevant hypergeometric probability.
2. Use that probability as a statistic.
3. **P-value** (for Fisher's exact test of independence) = the sum of the probabilities for all tables having a probability equal to or smaller than that observed.

An illustration

The observed data

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

All possible tables (with these marginals):

20	0	→ 0.00007
9	11	

14	6	→ 0.25994
15	5	

19	1	→ 0.00160
10	10	

13	7	→ 0.16246
16	4	

18	2	→ 0.01380
11	9	

12	8	→ 0.06212
17	3	

17	3	→ 0.06212
12	8	

11	9	→ 0.01380
18	2	

16	4	→ 0.16246
13	7	

10	10	→ 0.00160
19	1	

15	5	→ 0.25994
14	6	

9	11	→ 0.00007
20	0	

Fisher's exact test: Example 1

Observed data

	N	Y	
A	18	2	20
B	11	9	20
	29	11	40

P-value \approx 3.1%

In R: `fisher.test()`

Recall:

χ^2 test: P-value = 1.3%

LRT: P-value = 1.1%

Fisher's exact test: Example 2

Observed data

P-value \approx 4.4%

	I-B	NI-B	
I-A	9	9	18
NI-A	20	62	82
	29	71	100

Recall:

χ^2 test: P-value = 3.0%

LRT: P-value = 3.7%

Summary

Testing for **independence** in a 2 x 2 table:

- A special case of testing a composite hypothesis in a one-dimensional table.
- Can use either the LRT or χ^2 test, as before.
- Can also use Fisher's exact test.
- **I always prefer Fisher's exact test.**

Paired data

Gather 100 rats and determine whether they are infected with viruses A and B.

	I-B	NI-B	
I-A	9	9	18
NI-A	20	62	82
	29	71	100

Underlying probabilities

		B		
		0	1	
A	0	p_{00}	p_{01}	p_{0+}
	1	p_{10}	p_{11}	p_{1+}
		p_{+0}	p_{+1}	1

Another question: Is the rate of infection of virus A the same as that of virus B?

In other words (ur...symbols): Is $p_{1+} = p_{+1}$?

(Equivalently, is $p_{10} = p_{01}$?)

McNemar's Test

$$H_0: p_{01} = p_{10}$$

Under H_0 , the expected counts for cells 01 and 10 are both $(n_{01} + n_{10})/2$.

The χ^2 test statistic reduces to
$$X^2 = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}}$$

For large sample sizes, this statistic has null distribution that is approximately a χ^2 (df = 1).

For the example: $X^2 = (20 - 9)^2 / 29 = 4.17 \longrightarrow P = 4.1\%$.

An exact test

Condition on $n_{01} + n_{10}$.

Under H_0 , $n_{01} \mid n_{01} + n_{10} \sim \text{binomial}(n_{01} + n_{10}, 1/2)$.

In R, use the function `binom.test`.

For the example, $P = 6.1\%$.

Paired data

Paired data

	I-B	NI-B	
I-A	9	9	18
NI-A	20	62	82
	29	71	100

$P = 6.1\%$

Unpaired data

	I	NI	
A	18	82	100
B	29	71	100
	47	153	200

$P = 9.5\%$

Taking appropriate account of the “pairing” is important!