Fisher's Exact test & Log Odds Ratio

ST551 Lecture 22

Charlotte Wickham

2017-11-13

Fisher's Exact test

For 2×2 tables

Setting: two **independent** samples

$Y_1, \ldots, Y_n$ i.i.d from Bernoulli(p_Y)
$X_1, \ldots, X_m$ i.i.d from Bernoulli(p_X)

Null hypothesis: $p_Y = p_X$ *Homegeneity of proportions*

Test statistic: Probability of observed table *conditional on margins*
row sums
column sums

p-value: Sum of probability of all tables *as or more extreme* than observed table.

Data from last time

```
##          ate_breakfast
## cat_dog no yes
##    cats  6   9
##    dogs  6  14
```

$a=6$	$b=9$	$R_1 =$
$c=6$	$d=14$	R_2
C_1	C_2	$N = a+b+c+d$

n_{11}	n_{12}	
n_{21}	n_{22}	

A smaller (in sample size) example

Random sample

```
##          ate_breakfast
## cat_dog no yes
##    cats  4   4
##    dogs  3   4
```

A smaller (in sample size) example with margins

```
##           ate_breakfast
## cat_dog no yes Sum
##    cats  4   4   8
##    dogs  3   4   7
##     Sum  7   8  15
```

What is the probability of these four entries ?
boils to
What is the probability of any one entry? n_{12}

Under null, there is no difference in the probability of eating breakfast between people who prefer cats and people who prefer dogs, what is the probability of observing this exact table, conditioning on the margins?

Working out the test statistic

Under the null $p_Y = p_X = p$.

Conditional on row sums $R_1 = a + b$ and $R_2 = c + d$:

$n_{12} \sim Binomial(R_1, p)$
$n_{22} \sim Binomial(R_2, p)$

But we also need $n_{12} + n_{22} = C_2$.

I.e. What is $P(n_{12} = b \mid n_{12} + n_{22} = C_2)$?

Turns out to be:

$$= \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{N}{b+d}}$$

$\binom{n}{x}$ = n choose x

choose(n, x)

a.k.a Hypergeometric Distribution

Derivation

$$P(n_{12} = b) = \binom{a+b}{b} p^b (1-p)^a$$

$n_{12} \sim \text{Binomial}(R_1, p)$

$$P(n_{22} = d) = \binom{c+d}{d} p^d (1-p)^c$$

$$P(n_{12} + n_{22} = b + d) = \binom{N}{b+d} p^{b+d} (1-p)^{a+c}$$

$$\begin{aligned}
P(n_{12} = b \mid n_{12} + n_{22} = b + d) &= \frac{P(n_{12} = b, n_{12} + n_{22} = b + d)}{P(n_{12} + n_{22} = b + d)} \\
&= \frac{P(n_{12} = b, n_{22} = d)}{P(n_{12} + n_{22} = b + d)} \\
&= \frac{\binom{a+b}{b} p^b (1-p)^a \binom{c+d}{d} p^d (1-p)^c}{\binom{n}{b+d} p^{b+d} (1-p)^{a+c}} \\
&= \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{N}{b+d}}
\end{aligned}$$

Example: Test statistic

```
##          ate_breakfast
## cat_dog no yes Sum
##    cats  4   4   8
##    dogs  3   4   7
##    Sum   7   8  15
```

Under null, probability of seeing this table, given margins:

$$\text{Prob} = \frac{\binom{8}{4}\binom{7}{4}}{\binom{15}{8}} = 0.38$$

Test statistic for observed table

p-values

More extreme? Depends on alternative:

- $p_Y > p_X$, greater n_{12}
- $p_Y < p_X$, smaller n_{12}
- $p_Y \neq p_X$, less likely tables

tables with prob.
less than obs

Example: p-values lower alternative

$H_A: p_Y < p_X$

More extreme tables would have $n_{12} = 3, 2, 1$ or 0

Your turn: Complete the table if $n_{12} = 3$:

	No	Yes	
Cats	5	3	8
Dogs	2	5	7
	7	8	15

Prob =

$$\frac{\binom{8}{3}\binom{7}{5}}{\binom{15}{8}} = 0.18$$

Example: p-values lower alternative

$H_{\cancel{0}A}: p_Y < p_X$

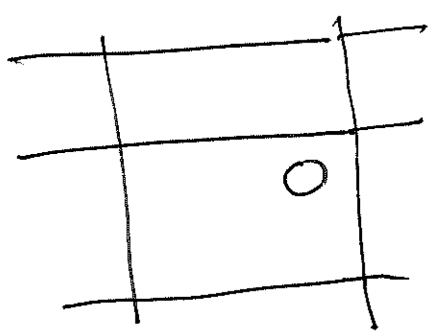
	No	Yes
Cats	5	3
Dogs	2	5

0.18

	No	Yes
Cats	6	2
Dogs	1	6

0.03

Not possible!



	No	Yes
Cats	7	1
Dogs	0	7

0.001

p-value $= 0.381 + 0.183 + 0.03 + 0.001 = 0.595$

Example: p-values greater alternative

$H_{\cancel{0}A}: p_Y > p_X$

More extreme tables would have $n_{12} = 5, 6, 7$ or 8

Your turn: What are the more extreme tables?

	no	yes	
cat	3	5	8
dog	4	3	7
	7	8	

2	6
5	2

1	7
6	1

0	8
7	0

p-value $= \boxed{0.381} + 0.305 + 0.091 + 0.009 + 0 = 0.786$

$\uparrow$ to 3 d.p.

Example: p-values two-sided alternative

Which tables are less likely than that observed?

All of them

$$\text{p-value} = \sum_{\text{all tables}} P(\text{table} \mid \text{margins}) = 1$$

$$\text{less likely: } P(\text{table} \mid \text{margins}) \le P(\underbrace{\text{obs. table} \mid \text{margins}}_{0.38})$$

In R

Careful with direction (rows permuted in R)

```
x <- xtabs(~ cat_dog + ate_breakfast, data = class_data)
fisher.test(x, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.397
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3796782       Inf
## sample estimates:
## odds ratio
##   1.535689
```

$P_{row1} < P_{row2}$ $\quad$ $P_{row2} > P_{row1}$

Sampling for binomial proportions

- **Multinomial** Obtain a sample of N units, and cross classify according to the grouping variable G and the binary response variable Y.
  - Can estimate all probabilities, $P(Y_i = 1)$, $P(G_i = 1)$, $P(Y_i = 1|G_i = 0)$ etc.
- **Two-sample** Obtain two samples, one of size n from group 1 ($G_i = 0$), and one of size m from group 2 ($G_i = 1$).
  - Can't estimate $P(G_i = 1)$ or $P(Y_i = 1)$ but can estimate $P(Y_i = 1|G_i = 0)$.

Retrospective studies

Sometimes, particularly if an outcome is very rare, it's hard to observe any successes ($Y_i = 1$), with Multinomial or Binomial sampling.

E.g.

L NL

Response variable Y = struck by lightning (Yes/No)

G NG

Grouping variable G = regular golfer (Yes/No)

Lightning strikes are very rare events, so even with rather large samples we may not observe anybody who has been struck.

In a **retrospective sample** we get s units where $Y_i = 1$ and r units where $Y_i = 0$ then classify them to the groups G_i.

Example

	$Y_i = 0$	$Y_i = 1$	**Total**
$G_i = 0$	a	b	R_1
$G_i = 1$	c	d	R_2
Total	$C_1 = r$	$C_2 = s$	$N = r + s$

fixed by
study design

Retrospective studies

We can no longer estimate $P(Y_i = 1 \mid G_i = 1)$ (e.g. P(getting hit by lightening | regular golfer) = P(L|G)) since we don't have a representative sample of golfers.

Means we can't estimate:

- Risk difference: $P(Y_i = 1 \mid G_i = 1)$ - $P(Y_i = 1 \mid G_i = 0)$
- Relative risk: $\frac{P(Y_i=1 \mid G_i=1)}{P(Y_i=1 \mid G_i=0)}$

We can still estimate $P(G_i = 1 \mid Y_i = 1)$, which means we can estimate the odds ratio.

$$\frac{P(L \mid G)/(1 - P(L \mid G))}{P(L \mid NG)/(1 - P(L \mid NG))} = \frac{P(G \mid L)/(1 - P(G \mid L))}{P(G \mid NL)/(1 - P(G \mid NL))}$$

Example

	$Y_i = 0$	$Y_i = 1$	**Total**
$G_i = 0$	a	b	R_1
$G_i = 1$	c	d	R_2
Total	$C_1 = r$	$C_2 = s$	$N = r + s$

Estimate of $P(G_i = 1 | Y_i = 1) = d/C_2$

Estimate of $P(G_i = 1 | Y_i = 0) = c/C_1$

Estimate of Odds(G|L) $= \frac{d/C_2}{1-d/C_2} = \frac{d}{b}$

Estimate of Odds(~~L|G~~) $= \frac{c/C_1}{1-c/C_1} = \frac{c}{a}$

(G|NL)

Estimate of Odds ratio: $\frac{ad}{bc}$

Properties of the sample odds ratio

Sample odds ratio

$$\hat{\omega} = \frac{ad}{bc}$$

The log of this estimate is asymptotically Normal

$$\log(\hat{\omega}) \dot{\sim} N\left(\log(\omega), \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)$$

Log Odds ratio test

oddsratio is one

Leads to test of $H_0 : \omega = 1$ using statistic

$$Z = \frac{\log(\hat{\omega}) - \log(1)}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}} = \frac{\log(\hat{\omega})}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

Under H_0 has an approximate $\mathrm{N}(0, 1)$ distribution.

$(1 - \alpha)100\%$ confidence interval for $\log(\omega)$:

$$\log(\hat{\omega}) \pm z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

(Exponentiate both endpoints to get CI for ω)

Which test?

All three tests, Chi-square, Fisher's Exact, and Log Odds ratio can be used in all sampling schemes (Multinomial, two sample Binomial and Retrospective).

Which probabilities can be estimated depends on sampling scheme.