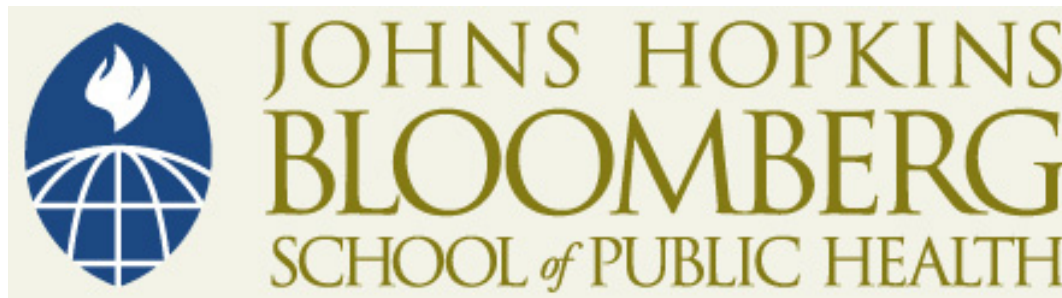


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2008, The Johns Hopkins University and Brian Caffo. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

# Lecture 21

Brian Caffo

Department of Biostatistics  
Johns Hopkins Bloomberg School of Public Health  
Johns Hopkins University

December 19, 2007

# Table of contents

- 1 Table of contents
- 2 Outline
- 3 Fisher's exact test
- 4 The hypergeometric distribution
- 5 Fisher's exact test in practice
- 6 Monte Carlo

- 1 Introduce Fisher's exact test
- 2 Illustrate Monte Carlo version of test

## Fisher's exact test

- Fisher's exact test is “exact” because it guarantees the  $\alpha$  rate, regardless of the sample size
- Example, chemical toxicant and 10 mice

	Tumor	None	Total
Treated	4	1	5
Control	2	3	5
Total	6	4	

- $p_1 =$  prob of a tumor for the treated mice
- $p_2 =$  prob of a tumor for the untreated mice

## Continued

- $H_0 : p_1 = p_2 = p$
- Can't use  $Z$  or  $\chi^2$  because  $SS$  is small
- Don't have a specific value for  $p$

## Fisher's exact test

- Under the null hypothesis every permutation is equally likely

- observed data

```
Treatment : T T T T T C C C C C
Tumor      : T T T T N T T N N N
```

- permuted

```
Treatment : T C C T C T T C T C
Tumor      : N T T N N T T T N T
```

- Fisher's exact test uses this null distribution to test the hypothesis that  $p_1 = p_2$

# Hyper-geometric distribution

- $X$  number of tumors for the treated
- $Y$  number of tumors for the controls
- $H_0 : p_1 = p_2 = p$
- Under  $H_0$ 
  - $X \sim \text{Binom}(n_1, p)$
  - $Y \sim \text{Binom}(n_2, p)$
  - $X + Y \sim \text{Binom}(n_1 + n_2, p)$

## Continued

$$P(X = x \mid X + Y = z) = \frac{\binom{n_1}{x} \binom{n_2}{z-x}}{\binom{n_1+n_2}{z}}$$

This is the hypergeometric pmf

$$P(X = x) = \binom{n_1}{x} p^x (1-p)^{n_1-x}$$

$$P(Y = z - x) = \binom{n_2}{z-x} p^{z-x} (1-p)^{n_2-z+x}$$

$$P(X + Y = z) = \binom{n_1 + n_2}{z} p^z (1-p)^{n_1+n_2-z}$$

## Continued

$$\begin{aligned}P(X = x \mid X + Y = z) &= \frac{P(X = x, X + Y = z)}{P(X + Y = z)} \\ &= \frac{P(X = x, Y = z - x)}{P(X + Y = z)} \\ &= \frac{P(X = x)P(Y = z - x)}{P(X + Y = z)}\end{aligned}$$

Plug in and finish off yourselves

# Fisher's exact test

- More tumors under the treated than the controls
- Calculate an *exact* P-value
- Use the conditional distribution = hypergeometric
- Fixes both the row and the column totals
- Yields the same test regardless of whether the rows or columns are fixed
- Hypergeometric distribution is the same as the permutation distribution given before

Tables supporting  $H_a$ 

- Consider  $H_a : p_1 > p_2$
- P-value requires tables as extreme or more extreme (under  $H_a$ ) than the one observed
- Recall we are fixing the row and column totals
- Observed table

$$\text{Table 1} = \begin{array}{cc|c} 4 & 1 & 5 \\ 2 & 3 & 5 \\ \hline 6 & 4 & \end{array}$$

- More extreme tables in favor of the alternative

$$\text{Table 2} = \begin{array}{cc|c} 5 & 0 & 5 \\ 1 & 4 & 5 \\ \hline 6 & 4 & \end{array}$$

## Calculations

$$\begin{aligned}P(\text{Table 1}) &= P(X = 4 | X + Y = 6) \\ &= \frac{\binom{5}{4} \binom{5}{2}}{\binom{10}{6}} = 0.238\end{aligned}$$

$$\begin{aligned}P(\text{Table 2}) &= P(X = 5 | X + Y = 6) \\ &= \frac{\binom{5}{5} \binom{5}{1}}{\binom{10}{6}} = 0.024\end{aligned}$$

$$P\text{-value} = 0.238 + 0.024 = 0.262$$

```
dat <- matrix(c(4, 1, 2, 3), 2)
fisher.test(dat, alternative = "greater")
```

```
-----output-----
```

```
          Fisher's Exact Test for Count Data
```

```
data:  dat
p-value = 0.2619
alt hypoth: true odds ratio is greater than 1
95 percent confidence interval:
 0.3152217      Inf
sample estimates:
odds ratio
 4.918388
```

- Two sided p-value =  $2 \times$  one sided P-value  
(There are other methods which we will not discuss)
- P-values are usually large for small  $n$
- Doesn't distinguish between rows or columns
- The common value of  $p$  under the null hypothesis is called a nuisance parameter
- Conditioning on the total number of successes,  $X + Y$ , eliminates the nuisance parameter,  $p$
- Fisher's exact test guarantees the type I error rate
- Exact unconditional P-value

$$\sup_p P(X/n_1 > Y/n_2; p)$$

# Monte Carlo

- Observed table  $X = 4$

Treatment : T T T T T C C C C C

Tumor : T T T T N T T N N N

- Permute the second row

Treatment : T T T T T C C C C C

Tumor : T N T N T T N N T T

- Simulated table  $X = 3$
- Do over and over
- Calculate the proportion of tables for which the simulated  $X \geq 4$
- This proportion is a Monte Carlo estimate for Fisher's exact P-value