Contingency Table

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Outline

- Fisher's exact test
- The hypergeometric distribution
- Fisher's exact test in practive
- Monte Carlo

Fisher's exact test

- Fisher's exact test is "exact" because it guarantees the α 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=probability of a tumor for the treated mice
- p_2=probability of a tumor for the untreated mice

Continued

- $H_0 : p_1 = p_2 = p$
- Can not use Z or χ^2 because sample size is small
- Do not know 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 null hypothesis that $p_1 = p_2$

Hyper-geometric distribution

- X: number of tumors in the treated group
- Y: number of tumors in the control group
- $H_0 : p_1 = p_2 = p$
- Under H_0
    - $X \sim Binom(n_1, p)$
    - $Y \sim Binom(n_2, p)$
    - $X + Y \sim Binom(n_1 + n_2, p)$

Continued

$$P(X = x | X + Y = z) = \frac{\binom{n_1}{x}\binom{n_2}{z-x}}{\binom{n_1+n2}{z}}$$

This is the hypergeometric pmf

Proof

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

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

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

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

Fisher's exact test

- More tumors under the treated than the controls
- Calculate an *exact* p-value
- Use the conditional distribution $=$ hypergeometric
- Fixed both the row and the column totals
- Hypergeometric distribution is the same as the permutation distribution given before

Tables supporting H_a

- Consider $Ha : p_1 > p_2$
- p-value requires table as extreme or more extreme (under H_a) than the one observed
- Recall we are fixing the row and column totals
- Observed table Table 1 =

4	1	5
2	3	5
6	4	

- More extreme in favor of the alternative Table 2 =

5	0	5
1	4	5
6	4	

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-value $= 0.238 + 0.024 = 0.262$

R code

```
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
```

Notes

- Two sided p-value $= 2 \times$ one sided p-value
- p-values are usually large for small n
- Does not 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

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