Categorical Data: Contingency Tables
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-09-29 12:08

Contingency Tables

- Two-way $(r \times c)$ contingency table:

	j			
i	1	2	$\cdots$	c
1	n_{11}	n_{12}	$\cdots$	n_{1c}
2	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

- Notation:

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij} \qquad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$$

Contingency Tables

- Two scenarios where $r \times c$ table arise

  1. Sample from a population and measure two characteristics, say X and Y

$$\Pr[X = i, Y = j] = \pi_{ij} \; ; \; \sum_{i=1}^{r} \sum_{j=1}^{c} \pi_{ij} = 1$$

  2. Each row corresponds to a sample from a different population

$$\sum_{j=1}^{c} \pi_{ij} = 1$$

Contingency Table: Example

- A survey of physicians asked about the size of community in which they were reared and the size of the community in which they practice

	Practice				
Reared	<5k	5-49k	50-99k	100k+	Total
<5k	40	38	32	37	147
5-49k	26	42	35	33	136
50-99k	24	26	34	31	115
100k+	30	39	53	60	182
	120	145	154	161	580

Contingency Table: Example

- A case-control study was conducted to investigate the relationship between age at first birth and breast cancer

	Age at 1st birth					
	<20	20-24	25-29	30-34	$\geq$ 35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10245
	1742	5638	3904	1555	626	13465

Contingency Tables

- Physician's example H_0: size of place of practice is independent of size of place of rearing

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$$

- Test of independence, $X \perp Y$

$$\Pr[X = i, Y = j] = \Pr[X = i] \Pr[Y = j]$$

for $i = 1, \ldots, r$; $j = 1, \ldots, c$

Contingency Tables

- Breast cancer example H_0 : distribution of age at 1st birth is the same for cases and controls

$$H_0 : \pi_{ij} = \pi_{i'j}; j = 1, 2, \ldots, c$$

- Test of homogeneity/association

Test of Independence or No Association

- Under either H_0, the estimated expected frequency in the (i, j) cell is

$$E_{ij} = \frac{n_i . n_{.j}}{N}$$

- Consider breast cancer example
  - If H_0 is true, would expect the proportion of women < 20 to be

$$\frac{n_{11} + n_{21}}{N} = \frac{n_{.1}}{N}$$

  - There are $n_{1.}$ cases, so we would expect

$$E_{11} = n_{1.} \frac{n_{.1}}{N} = \frac{n_{1.} n_{.1}}{N}$$

    cases to be < 20 years old

Test of Independence or Association

- Under H_0, the expected frequency in the (i, j) cell is

$$E_{ij} = \frac{n_{i.} n_{.j}}{N}$$

- Let

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  i.e.

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n_{i.} n_{.j}/N)^2}{n_{i.} n_{.j}/N}$$

Test of Independence

- Under H_0,

$$X^2 \sim \chi^2_{(r-1)(c-1)}$$

- Physician's Example:

$$(r-1)(c-1) = 3 \times 3 = 9$$

$$C_{.05} = \{X^2 : X^2 > \chi^2_{.95,9} = 16.92\}$$

Physician's Example

- Expected values

	Practice				
Reared	<5k	5-49k	50-99k	100k+	Total
<5k	30.4	36.8	39.0	40.8	147
5-49k	28.1	34.0	36.1	37.8	136
50-99k	23.8	28.8	30.5	31.9	115
100k+	37.7	45.5	48.3	50.5	182
	120	145	154	161	580

Physician's Example

- Calculate test statistic

$$X^2 = \frac{(40-30.4)^2}{30.4} + \frac{(38-36.8)^2}{36.8} + \cdots + \frac{(60-50.5)^2}{50.5} = 12.81$$

- Do not reject H_0.
- There is insufficient evidence to conclude that place of practice and place of rearing are dependent; the data are consistent with the null hypothesis that place of practice and place of rearing are independent

Breast Cancer Example

- Underlying probabilities

	Age at 1st birth					
	<20	20-24	25-29	30-34	≥ 35	Total
Case	π_{11}	π_{12}	π_{13}	π_{14}	π_{15}	1
Control	π_{21}	π_{22}	π_{23}	π_{24}	π_{25}	1

Breast Cancer Example

- Null hypothesis

$$H_0 : \pi_{1j} = \pi_{2j} \text{ for } j = 1, 2, 3, 4, 5$$

- Can use same statistic

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(c-1)}$$

Breast Cancer Example

- Expected frequencies

	Age at 1st birth					
	<20	20-24	25-29	30-34	$\geq$ 35	Total
Case	416.6	1348.3	933.6	371.9	149.7	3220
Control	1325.4	4289.7	2970.4	1183.1	476.3	10245
	1742	5638	3904	1555	626	13465

Breast Cancer Example

- Test statistic

$$X^2 = \frac{(320 - 416.6)^2}{416.6} + \cdots + \frac{(406 - 476.3)^2}{476.3} = 130.3$$

- Rejection region

$$C_{.05} = \{X^2 : X^2 > \chi^2_{.95,4} = 9.49\}$$

- Reject H_0
- The age distributions are not the same

Asymptotic Approximation

- Note the χ^2 distribution for X^2 is an approximation
- The approximation works well for if $E_{ij} \geq 5$ for all i, j
- If $E_{ij} < 5$, a generalization of Fisher's exact test can be employed or categories combined

Test of Independence

- For $r = c = 2$, can show

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N}$$

  equals

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{.1}n_{2.}n_{.2}}$$

- *Pearson chi-square statistic*

Test for Trend

- Consider a $2 \times c$
- The χ^2 test for homogeneity does not tell us how the probabilities differ
- Rather, just if they differ
- If the categories of the column variable are ordered, a more powerful test is possible

Test for Trend

- Suppose columns = exposure are ordered
- Rows = disease (yes/no)
- Interested in detecting alternatives where the probability of disease proportional to exposure
- I.e., looking for a monotonic dose-response type relationship

Test for Trend

- Example 7.3 text. Risk of catheter-related infection and the duration of catheterization

	Duration			
Culture	1	2	3	4+
Positive	1	5	5	14
Negative	46	64	39	76
Total	47	69	44	90

- Let ρ_j denote the conditional probability of being in row 1 given in column j
- For catheter example, ρ_j is the probability of pos given in the jth duration category

Test for Trend

- Test

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_c$$

versus

$$H_A : \rho_1 \leq \rho_2 \leq \cdots \leq \rho_c$$

with at least one strict inequality, or

$$H_A : \rho_1 \geq \rho_2 \geq \cdots \geq \rho_c$$

with at least one strict inequality

Test for Trend

- Numerical scores must be assigned to categories:

$$x_j : j = 1, 2, \ldots, c$$

- Example: $x_j = j$
- In breast cancer example, use midrange of age categories

Test for Trend

- Let

$$[n_1 x] \equiv \sum_{j=1}^{c} n_{1j} x_j - \frac{n_{1\cdot} \sum_{j=1}^{c} n_{\cdot j} x_j}{N}$$

$$[x^2] \equiv \sum_{j=1}^{c} n_{\cdot j} x_j^2 - \frac{(\sum_{j=1}^{c} n_{\cdot j} x_j)^2}{N}$$

  and

$$p \equiv \frac{n_{1\cdot}}{N}$$

- Then the chi-square test for trend (p 215 text) is

$$X^2_{trend} \equiv \frac{[n_1 x]^2}{[x^2] p(1-p)}$$

Test for Trend

- Huh? Intuitive development:
- Compute average score

$$\bar{x} \equiv \sum_{j=1}^{c} \frac{n_{1j} x_j}{n_{1.}}$$

- Compute finite-sample expected value under the null

$$E(\bar{x}) \equiv E(x) \equiv \sum_{j=1}^{c} \frac{n_{.j} x_j}{N}$$

Test for Trend

- Compute finite-sample variance

$$V(\bar{x}) \equiv \left(\frac{1-f}{n_{1.}}\right)\left[E(x^2) - E(x)^2\right]$$

where $f = n_{1.}/N$ is the sampling fraction and

$$E(x^2) \equiv \sum_{j=1}^{c} \frac{n_{.j}x_j^2}{N}$$

- Then the chi-square test for trend can equivalently be written

$$X^2_{trend} = \frac{\{\bar{x} - E(\bar{x})\}^2}{V(\bar{x})}$$

Test for Trend

- Under H_0,

$$X^2_{trend} \sim \chi^2_1$$

$$C_\alpha = \{X^2 : X^2 > \chi^2_{1,1-\alpha}\}$$

$$p = \Pr[\chi^2_1 > x^2]$$

- Note degrees of freedom equal 1 regardless of c

Test for Trend

- Catheter Example

	Duration			
Culture	1	2	3	4+
Positive	1	5	5	14
Negative	46	64	39	76
Total	47	69	44	90

- $X^2_{trend} = 6.98$; $p = 0.008$; reject H_0 and conclude the probability of pos culture increases with duration

Test for Trend: R

```
> prop.trend.test(c(1,5,5,14),c(47,69,44,90))

        Chi-squared Test for Trend in Proportions

data:  c(1, 5, 5, 14) out of c(47, 69, 44, 90) ,
 using scores: 1 2 3 4
X-squared = 6.9764, df = 1, p-value = 0.008259
```

Test for Trend: SAS

```
proc freq;
  tables sample*duration/chisq nopct norow;
  weight wt;
run;
```

Table of sample by duration

sample	duration				
Frequency|					
Col Pct |	1	2	3	4	Total
pos	1	5	5	14	25
	2.13	7.25	11.36	15.56	
neg	46	64	39	76	225
	97.87	92.75	88.64	84.44	
Total	47	69	44	90	250

Statistics for Table of sample by duration

Statistic	DF	Value	Prob
Chi-Square	3	6.9951	0.0721
Mantel-Haenszel Chi-Square	1	6.9485	0.0084

χ^2 Goodness of Fit

- Goal: assess how well a particular model fits the data
- General form

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df}$$

- E_i computed under H_0
- Rejecting null implies model does not provide an adequate fit to the data

χ^2 Goodness of Fit

- Multinomial: generalization of Binomial from 2 to K categories
- Suppose n independent trials, each with K possible outcomes having probability $\pi_1, \ldots, \pi_K$
- Let n_i be the number of trials having outcome i for $i = 1, \ldots, K$ such that

$$n = \sum_{i=1}^{K} n_i$$

- Then

$$E(n_i) = n\pi_i \text{ and } V(n_i) = n\pi_i(1 - \pi_i)$$

χ^2 GOF: Genetics Example

- Example: Mendelian genetics hypothesizes a particular genotype should occur in the proportions

$$1 : 2 : 1 \text{ (dominant, heterozygous, recessive)}$$

- $H_0 : \pi_1 = .25, \pi_2 = .5, \pi_3 = .25$
- H_A : at least one of the equalities is false
- A survey finds the genotypes $n_1 = 21, n_2 = 62, n_3 = 17$

χ^2 GOF: Genetics Example

- Since expected values known,

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-1}$$

  in general; $K = 3$ for genetics example

- Under H_0, $E_1 = .25 \times 100 = 25$, $E_2 = 50$, $E_3 = 25$

χ^2 GOF: Genetics Example

- Thus

$$X^2 = \frac{(21-25)^2}{25} + \frac{(62-50)^2}{50} + \frac{(17-25)^2}{25} = 6.08$$

- Since $K = 3$, df=2, so

$$C_{.05} = \{X^2 : X^2 \geq \chi^2_{.95,2} = 5.99\}$$

- Reject H_0

χ^2 GOF: Genetics Example using SAS

```
proc freq order=data;
  tables genotype/testp=(25 50 25);
  weight wt;
run;
```

The FREQ Procedure

genotype	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
dominant	21	21.00	25.00	21	21.00
heterozy	62	62.00	50.00	83	83.00
recessiv	17	17.00	25.00	100	100.00

Chi-Square Test
for Specified Proportions

Chi-Square	6.0800
DF	2
Pr > ChiSq	0.0478

Sample Size = 100

χ^2 GOF: DBP Example

- A random sample of diastolic blood pressure was obtained from a population of interest
- It is hypothesized that DBP will be normally distributed

χ^2 GOF: DBP Example

DBP	Frequency
< 50	57
$[50, 60)$	330
$[60, 70)$	2132
$[70, 80)$	4584
$[80, 90)$	4604
$[90, 100)$	2119
$[100, 110)$	659
≥ 110	251
Total	14736

χ^2 GOF: DBP Example

- From the sample

$$\bar{y} = 80.7 \text{ and } s = 12.00$$

- If DBP is normally distributed with $\mu = 80.7$ and $\sigma = 12$, the expected frequency in an interval between a and b is

$$14736\{\Phi[(b - 80.7)/12] - \Phi[(a - 80.7)/12]\}$$

- The expected frequency in the ≤ 50 group is

$$14736\Phi[(50 - 80.7)/12] = 14736\Phi[-2.56] = 76.6$$

χ^2 GOF: DBP Example

DBP	Freq	z	$\Phi(z)$	Prob	E
< 50	57	-2.56	0.0052	0.0052	76.6
$[50, 60)$	330	-1.72	0.0427	0.0375	552.6
$[60, 70)$	2132	-0.89	0.1867	0.1440	2121.9
$[70, 80)$	4584	-0.06	0.4761	0.2894	4264.9
$[80, 90)$	4604	0.77	0.7794	0.3033	4469.4
$[90, 100)$	2119	1.61	0.9463	0.1669	2459.4
$[100, 110)$	659	2.44	0.9927	0.0464	683.8
≥ 110	251		1	0.0073	107.6

χ^2 GOF: DBP Example

- H_0: data are from a normal distribution
- H_A: data are not from a normal distribution
- Rejection region

$$C_\alpha = \{X^2 : X^2 \geq \chi^2_{K-S-1,1-\alpha}\}$$

  where S= number of parameters estimated

- For DBP example, $K = 8$ and $S = 2$ such that

$$C_{0.05} = \{X^2 : X^2 \geq \chi^2_{5,0.95} = 11.07\}$$

χ^2 GOF: DBP Example

- GOF test statistic

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \frac{(57 - 76.6)^2}{76.6} + \cdots = 361.8$$

- Reject H_0

- Read Section 6.6.4 of text w/r/t distribution of X^2 when parameter estimation necessary

χ^2 GOF: Mendel Example

- Fisher examined Mendel's experiments (Table 6.9 text)

Experiment	X^2	DF
3:1 Ratios	2.14	7
2:1 Ratios	5.17	8
Bifactorial	2.81	8
Gametic ratios	3.67	15
Trifactorial	15.32	26
Total	29.11	64

- If $X_1^2, \ldots X_n^2$ independent χ^2 w/ $m_1, \ldots, m_n$ df, then $\sum_i X_i^2 \sim X_m^2$ where $m = \sum_i m_i$

$$p = \Pr[\chi^2_{64} > 29.11] = 0.9999474$$

Measurement of Agreement

- Kappa statistic
- Example: Adults were asked to rate their weight as underweight, normal, overweight, or obese

  There weight was then measured.

Measure of Agreement: Example

	Measured				
Self	Under	Normal	Over	Obese	Total
Under	462	178	0	0	640
Normal	72	2868	505	2	3447
Over	0	134	2086	280	2500
Obese	0	0	59	809	868
Total	534	3180	2650	1091	7455

Measure of Agreement: Kappa

- The χ^2 test for independence will reject H_0
- If we want to measure agreement, we might take the proportion on the diagonal:

$$p_a = \frac{462 + 2868 + 2086 + 809}{7455} = 0.835$$

- However there would be some agreement by chance even if the two classifications were independent

Measure of Agreement: Kappa

- Under independence,

$$E_{11} = (640)(534)/7455 = 45.8$$

$$E_{22} = 1470.3, E_{33} = 888.7, E_{44} = 127.0$$

- Therefore, by chance we expect 2531.8 agreements

$$p_c = \frac{2531.8}{7455} = 0.340$$

Measure of agreement: Kappa

- Let

  p_a = observed prop of agreement

  p_c = expected prop of agreement

- Kappa test statistic

$$\kappa = \frac{p_a - p_c}{1 - p_c}$$

- κ is a chance adjusted measure of agreement

Measure of Agreement: Kappa

- Note

$$\frac{-p_c}{1-p_c} \leq \kappa \leq 1$$

$\kappa = 0$ if agreement is totally by chance

$\kappa = 1$ iff there is perfect agreement

κ	Interpretation
(.8,1]	Excellent
(.6,.8]	Substantial
(.4,.6]	Moderate
(.2,.4]	Slight
$< .2$	Poor

Measure of Agreement: Kappa

- Under $H_0 : \kappa = 0$,

$$Var(\kappa) = \frac{p_c + p_c^2 - N^{-3} \sum_{i=1}^{r} (n_{i.}^2 n_{.i} + n_{i.} n_{.i}^2)}{N(1 - p_c)^2}$$

- For moderate sample sizes,

$$z = \frac{\kappa}{\sqrt{Var(\kappa)}} \sim N(0, 1)$$

  under H_0

Measure of Agreement: Kappa

- Example revisited:

$$\kappa = \frac{.835 - .34}{1 - .34} = 0.75$$

  indicating substantial agreement

- Compute variance of κ

$$Var(\kappa) = \frac{.34 + .34^2 - .2631}{7455(.66)^2} = 5.93 \times 10^{-5}$$

- Therefore

$$z = \frac{.75}{.00769} = 97.6$$

Kappa: SAS/R

- SAS

```
proc freq order=data;
  tables self*measured/agree nopct norow nocol;
  test kappa;
  weight wt;
```

```
    Simple Kappa Coefficient
--------------------------------
Kappa                     0.7502
ASE                       0.0065
95% Lower Conf Limit      0.7374
95% Upper Conf Limit      0.7629

     Test of H0: Kappa = 0

ASE under H0              0.0077
Z                        97.6540
One-sided Pr >  Z         <.0001
Two-sided Pr > |Z|        <.0001
```

- R: library(irr); kappa2()