BIOS 662 Fall 2016

Categorical Data: Contingency Tables

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Contingency Tables

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

$$
\begin{array}{cc|cccc}
 & & & j & & \\
 & & 1 & 2 & \cdots & c \\
\hline
 & 1 & n_{11} & n_{12} & \cdots & n_{1c} \\
 & 2 & n_{21} & n_{22} & \cdots & n_{2c} \\
i & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & r & n_{r1} & n_{r2} & \cdots & n_{rc}
\end{array}
$$

- Notation:

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

Contingency Tables

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

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

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

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

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

Contingency Table: Example I

- A survey of physicians asked about the size of the 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 II

- A case-control study of women was conducted to investigate the relationship between age at which they first gave birth and breast cancer

	Age at first childbirth					
	<20	20-24	25-29	30-34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10245
	1742	5638	3904	1555	626	13465

Contingency Tables

- For the physicians example, H_0 : size of community in which practice is independent of size of community in which reared

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

- 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

- For the breast cancer example, H_0 : distribution of age at first childbirth 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

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

- Consider the breast cancer example
  - The overall proportion of women <20 is

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

  - There are $n_{1\cdot}$ cases, so if H_0 is true we would expect

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

    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

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

that is,

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

Test of Independence

- Under H_0,

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

- Physicians example:

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

$$C_{0.05} = \{X^2 : X^2 > \chi^2_{9,0.95} = 16.92\}$$

Physicians Example

- Expected values

	Practice				
Reared	<5K	5-49K	50-99K	$\geq$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
$\geq$100K	37.7	45.5	48.3	50.5	182
	120	145	154	161	580

Physicians Example cont.

- Calculate the 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 the size of the community in which practice and that of the community in which reared are dependent; the data are consistent with the null hypothesis that size of community in which practice and that in which reared are independent

Breast Cancer Example

- Underlying probabilities

	Age at first childbirth					
	<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

- Null hypothesis

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

- We can use the 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 cont.

- Expected frequencies

	Age at first childbirth					
	<20	20-24	25-29	30-34	≥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 cont.

- Test statistic

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

- Rejection region

$$C_{0.05} = \{X^2 : X^2 > \chi^2_{4,0.95} = 9.49\}$$

- Reject H_0
- Conclude that the age distributions are not the same

Asymptotic Approximation

- Note that the χ^2 distribution for X^2 is an approximation
- The approximation works well 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 $r = c = 2$, one can show that

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

$$= \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{\cdot 1}n_{2\cdot}n_{\cdot 2}}$$

- This is the *Pearson chi-square statistic* we saw for comparing two proportions

Test for Trend

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

Test for Trend

- Suppose the columns correspond to ordered levels of an exposure
- Suppose the rows correspond to disease (yes/no)
- We are interested in detecting alternatives where the probability of disease increases (or decreases) with exposure level
- That is, we are looking for a monotonic dose-response type of relationship

Test for Trend: Example

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

	Duration (days)			
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 this catheter example, ρ_j is the probability of a positive culture given in the j^{th} 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 the categories:

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

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

Test for Trend

- Let

$$[n_1x] \equiv \sum_{j=1}^{c} n_{1j}x_j - \frac{n_{1.}\sum_{j=1}^{c} n_{.j}x_j}{N}$$

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

and

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

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

$$X^2_{\text{trend}} \equiv \frac{[n_1x]^2}{[x^2]p(1-p)}$$

Test for Trend

- Huh?
- Compute the average score in row 1:

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

- Compute the finite-sample expected value under the null:

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

Test for Trend

- Compute the finite-sample variance:

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

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

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

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

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

Test for Trend

- Under H_0,

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

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

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

- Note that here χ^2 has 1 degree of freedom regardless of c

Test for Trend: Example

- Catheter example

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

- $X^2_{\text{trend}} = 6.98$; $p = 0.008$; reject H_0 and conclude that the probability of a positive culture increases with duration of catheterization

Test for Trend: R

```
# Also if  prop.trend.test(c(1,5,5,14),c(47,69,44,90),c(1,2,3,4))
# or       prop.trend.test(c(1,5,5,14),c(47,69,44,90),c(4,3,2,1))
> 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


> prop.trend.test(c(1,5,5,14),c(47,69,44,90),c(1,2,3,6))

        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 6
X-squared = 6.4248, df = 1, p-value = 0.01125
```

Test for Trend: SAS

```
data trend;  input culture $1-8 duration count;
cards;
positive 1 1
positive 2 5
positive 3 5
positive 4 14
negative 1 46
negative 2 64
negative 3 39
negative 4 76
;

proc freq data=trend order=data;
   tables culture*duration / chisq nopct norow trend;
   weight count;
```

```
culture     duration

Frequency|
Col Pct  |       1|       2|       3|       4|  Total
---------+--------+--------+--------+--------+
positive |      1 |      5 |      5 |     14 |     25
         |   2.13 |   7.25 |  11.36 |  15.56 |
---------+--------+--------+--------+--------+
negative |     46 |     64 |     39 |     76 |    225
         |  97.87 |  92.75 |  88.64 |  84.44 |
---------+--------+--------+--------+--------+
Total          47       69       44       90      250
```

Test for Trend: SAS, cont.

```
Statistics for Table of culture by duration

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

WARNING: 25% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.


Cochran-Armitage Trend Test
--------------------------
Statistic (Z)       2.6413
One-sided Pr >  Z   0.0041
Two-sided Pr > |Z|  0.0083
```

χ^2 Test of Goodness of Fit

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

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

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

χ^2 Test of Goodness of Fit

- Multinomial: Generalization of binomial from 2 to K categories
- Suppose there are n independent trials, each with K possible outcomes having probabilities $\pi_1, \ldots, \pi_K$
- Let n_i be the number of trials having outcome i, $i = 1, \ldots, K$, such that

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

- Then

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

χ^2 GOF: Genetics Example

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

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

- $H_0 : \pi_1 = 0.25,\ \pi_2 = 0.5,\ \pi_3 = 0.25$
- H_A : at least one of the equalities is false (in which case at least two must be false)
- Suppose in a study the genotypes have frequencies $n_1 = 21,\ \ n_2 = 62,\ \ n_3 = 17$

χ^2 GOF: Genetics Example cont.

- When the expected values are known, in general

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

$K = 3$ for the genetics example

- Under H_0,

$$E_1 = 0.25 \times 100 = 25$$

$$E_2 = 50$$

$$E_3 = 25$$

χ^2 GOF: Genetics Example cont.

- Thus

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

- Because $K = 3$, $\text{df} = 2$, so

$$C_{0.05} = \{X^2 : X^2 \geq \chi^2_{2,0.95} = 5.99\}$$

- Also,

$$\Pr[\chi^2_2 \geq 6.08] = 1 - 0.952 = 0.048$$

- Reject H_0

χ^2 GOF: Genetics Example Using SAS

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

```
The FREQ Procedure

                                              Test    Cumulative    Cumulative
genotype        Frequency     Percent      Percent     Frequency       Percent
------------------------------------------------------------------------------
dominant               21       21.00        25.00            21         21.00
heterozygous           62       62.00        50.00            83         83.00
recessive              17       17.00        25.00           100        100.00

     Chi-Square Test
for Specified Proportions
-------------------------
Chi-Square         6.0800
DF                      2
Pr > ChiSq         0.0478
```

χ^2 GOF: DBP Example

- Diastolic blood pressure (DBP) was measured on a random sample from a population of interest
- It is hypothesized that DBP is normally distributed

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 cont.

- From the sample (before classifying into intervals)

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

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

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

- The expected frequency in the <50 group is

$$14736\,\Phi((50 - 80.7)/12) = 14736\,\Phi[-2.56] = 77.5$$

χ^2 GOF: DBP Example cont.

DBP	Freq	z	$\Phi(z)$	Prob	E
< 50	57	-2.56	0.0053	0.0053	77.5
$[50, 60)$	330	-1.73	0.0423	0.0370	545.3
$[60, 70)$	2,132	-0.89	0.1863	0.1440	2,122.3
$[70, 80)$	4,584	-0.06	0.4767	0.2905	4,280.2
$[80, 90)$	4,604	0.77	0.7808	0.3041	4,481.0
$[90, 100)$	2,119	1.61	0.9461	0.1653	2,435.7
$[100, 110)$	659	2.44	0.9927	0.0466	686.3
≥ 110	251	∞	1	0.0073	107.7
Total	14,736				14,736.0

χ^2 GOF: DBP Example cont.

- 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 the DBP example, $K = 8$ and $S = 2$, so that

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

χ^2 GOF: DBP Example cont.

- GOF test statistic

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \frac{(57 - 77.5)^2}{77.5} + \cdots = 348.3$$

- Reject H_0

- Read Section 6.6.4 of the text about the distribution of X^2 when parameter estimation is necessary

χ^2 GOF: Mendel Example

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

Experiment	X^2	DF	p
3:1 Ratios	2.14	7	0.95
2:1 Ratios	5.17	8	0.74
Bifactorial	2.81	8	0.95
Gametic ratios	3.67	15	0.999
Trifactorial	15.32	26	0.95
Total	29.11	64	

- If $X_1^2, \ldots X_n^2$ are independent χ^2 with $m_1, \ldots, m_n$ degrees of freedom, then

$$\sum_i X_i^2 \sim \chi_m^2 \quad \text{where} \quad m = \sum_i m_i$$

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

Measurement of Agreement

- Example: Adults were asked to rate their weight as underweight, normal, overweight, or obese; their weight was then measured

	Measured				
Self-report	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,

$$E_{11} = (640)(534)/7455 = 45.84$$

$$E_{22} = 1470.35, \quad E_{33} = 888.67, \quad E_{44} = 127.03$$

- Therefore we expect 2531.89 agreements just by chance

$$p_c = \frac{2531.89}{7455} = 0.340$$

Measure of Agreement: Kappa

- Let

  p_a = observed proportion of agreement

  p_c = expected proportion of agreement

- Kappa statistic

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

- κ is a chance-adjusted measure of agreement

Measure of Agreement: Kappa

- Note
$$\frac{-p_c}{1-p_c} \le \kappa \le 1$$
$\kappa = 0$ if agreement is totally by chance

  $\kappa = 1$ if and only if there is perfect agreement

- There are various categorizations (or guidelines) of values of κ; the best known is by Landis & Koch (1977). For instance, 0.41-0.60 is classified as "moderate agreement and 0.61-0.80 as "substantial agreement".

- My opinion is that the interpretation of κ needs to be context-dependent.

Measure of Agreement: Kappa

- Under $H_0 : \kappa = 0$,

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

- For moderate sample sizes,

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

under H_0

Measure of Agreement: Kappa

- Weight example revisited:

$$\kappa = \frac{0.835 - 0.34}{1 - 0.34} = 0.75$$

which Landis & Koch would classify as indicating substantial agreement

- Compute variance of κ

$$\text{Var}(\kappa) = \frac{0.34 + 0.34^2 - 0.2631}{7455(0.66)^2} = 5.901 \times 10^{-5}$$

- Therefore

$$z = \frac{0.75}{0.0077} = 97.654$$

Kappa: 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
```

Kappa: R

- R has kappa statistics in various libraries. The one below is in the vcd library. Note the uppercase K in Kappa()

```
> table<-matrix(c(462,178,0,0,72,2868,505,2,0,134,2086,280,
+   0,0,59,809),nrow=4,byrow=T)

> Kappa(table)
               value         ASE
Unweighted 0.7501581 0.006509663
Weighted   0.8109287 0.013309715

> confint(Kappa(table))

Kappa                lwr       upr
  Unweighted 0.7373994 0.7629169
  Weighted   0.7848422 0.8370153
```

Kappa: R cont.

- R also has a function kappa2() in the library irr; but it requires the data in a different format – an $n \times 2$ matrix with each row corresponding to an observation (a pair)

```
> col1<-c(rep('under',640),rep('normal',3447),rep('over',2500),
+   rep('obese',868))

> col2<-c(rep('under',462),rep('normal',178),rep('under',72),
+   rep('normal',2868),rep('over',505),rep('obese',2),
+   rep('normal',134),rep('over',2086),rep('obese',280),
+   rep('over',59),rep('obese',809))

> tab2<-cbind(col1,col2)
```

Kappa: R cont.

```
> rbind(tab2[1:5,],tab2[4081:4090,])
      col1     col2
 [1,] "under"  "under"
 [2,] "under"  "under"
 [3,] "under"  "under"
 [4,] "under"  "under"
 [5,] "under"  "under"
 [6,] "normal" "over"
 [7,] "normal" "over"
 [8,] "normal" "over"
 [9,] "normal" "over"
[10,] "normal" "over"
[11,] "normal" "obese"
[12,] "normal" "obese"
[13,] "over"   "normal"
[14,] "over"   "normal"
[15,] "over"   "normal"
```

Kappa: R cont.

```
> kappa2(tab2)
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 7455
   Raters = 2
    Kappa = 0.75

        z = 97.7
  p-value = 0

> agree(tab2)
 Percentage agreement (Tolerance=0)

 Subjects = 7455
   Raters = 2
  %-agree = 83.5
```