STAT 226 Lecture 27

Section 8.3 Comparing Proportions for Nominal Matched-Pairs Response

Yibi Huang

Example: Coffee Brand Market Share

A survey recorded the brand choice for a sample of buyers of instant decaffeinated coffee. At a later coffee purchase by these subjects, the brand choice was again recorded.

Purchase	High Pt	Taster's	Sanka	Nescafe	Brim	Total
First	171	75	204	36	55	541
	(31.6%)	(13.9%)	(37.7%)	(6.7%)	(10.2%)	
Second	135	82	231	33	60	541
	(25.0%)	(15.2%)	(42.7%)	(6.1%)	(11.1%)	

Question: Do the market shares of the 5 coffee brands change between the two purchases?

Can one test using Pearson's X^2 test, which indicates little evidence of changes between the two purchases (P-value ≈ 0.16).

```
coffeetab = matrix(c(171,75,204,36,55,135,82,231,33,60),
                   nrow=2, byrow=TRUE)
coffeetab
     [,1] [,2] [,3] [,4] [,5]
[1,]  171   75  204   36   55
[2,]  135   82  231   33   60
chisq.test(coffeetab)

    Pearson's Chi-squared test

data:  coffeetab
X-squared = 6.57108, df = 4, p-value = 0.16037
```

Can one test using Pearson's X^2 test, which indicates little evidence of changes between the two purchases (P-value ≈ 0.16).

```
coffeetab = matrix(c(171,75,204,36,55,135,82,231,33,60),
                   nrow=2, byrow=TRUE)
coffeetab
     [,1] [,2] [,3] [,4] [,5]
[1,]  171   75  204   36   55
[2,]  135   82  231   33   60
chisq.test(coffeetab)

    Pearson's Chi-squared test

data:  coffeetab
X-squared = 6.57108, df = 4, p-value = 0.16037
```

Paired data — each customer in the data made two purchases. Cannot regard the two purchases as independent observations — Pearson's X^2 test isn't applicable

Categorical Matched-Pairs Analyses w/ $J > 2$ Categories

Data: n pairs of observations (y_1, y_2)

$$\begin{gathered}(y_{11}, y_{12})\\(y_{21}, y_{22})\\(y_{31}, y_{32})\\\vdots\\(y_{n1}, y_{n2})\end{gathered}$$

Both y_{i1} and y_{i2} are categorical w/ $(J > 2)$ categories

Data are usually summarize as a square $J \times J$ table that the (i, j) cell is

$$n_{ij} = \text{count of pairs w/ } y_1 = i \text{ and } y_2 = j.$$

Example: Coffee Brand Market Share

Data display that reflect the dependence of the two purchases:

First Purchase	Second Purchase: High Pt	Taster's	Sanka	Nescafe	Brim	Total	(%)
High Pt	93	17	44	7	10	171	(31.6%)
Taster's	9	46	11	0	9	75	(13.9%)
Sanka	17	11	155	9	12	204	(37.7%)
Nescafe	6	4	9	15	2	36	(6.7%)
Brim	10	4	12	2	27	55	(10.2%)
Total	135	82	231	33	60	541	(100%)
(%)	(25.0%)	(15.2%)	(42.7%)	(6.1%)	(11.1%)		

Large cell counts on the main diagnal

⇒ Most buyers didn't change their choice

⇒ The two purchases of a buyer are dependent

Population probabilities:

First Purchase	Second Purchase					
	High Pt	Taster's	Sanka	Nescafe	Brim	Total
High Pt	π_{11}	π_{12}	π_{13}	π_{14}	π_{15}	π_{1+}
Taster's	π_{21}	π_{22}	π_{23}	π_{24}	π_{25}	π_{2+}
Sanka	π_{31}	π_{32}	π_{33}	π_{34}	π_{35}	π_{3+}
Nescafe	π_{41}	π_{42}	π_{43}	π_{44}	π_{45}	π_{4+}
Brim	π_{51}	π_{52}	π_{53}	π_{54}	π_{55}	π_{5+}
Total	π_{+1}	π_{+2}	π_{+3}	π_{+4}	π_{+5}	1

Question: Whether the coffee brand market shares change between the two purchases,

$$P(Y_1 = i) = \pi_{i+} = \pi_{+i} = P(Y_2 = i)$$

for $i = 1, \ldots, J$. under which each row marginal probability equals the corresponding column marginal probability, called *marginal homogeneity*.

Test of Marginal Homogeneity

We will estimate $\pi_{i+} - \pi_{+i}$ by

$$d_i = \widehat{\pi}_{i+} - \widehat{\pi}_{+i} = \frac{n_{i+}}{n} - \frac{n_{+i}}{n}, \quad \text{for} \quad i = 1, \ldots, J.$$

To test $(\pi_{1+}, \pi_{2+}, \ldots, \pi_{J+}) = (\pi_{+1}, \pi_{+2}, \ldots, \pi_{+J})$, we use all of

$$\mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{J-1} \end{pmatrix} = \begin{pmatrix} \widehat{\pi}_{1+} - \widehat{\pi}_{+1} \\ \widehat{\pi}_{2+} - \widehat{\pi}_{+2} \\ \vdots \\ \widehat{\pi}_{(J-1)+} - \widehat{\pi}_{+(J-1)} \end{pmatrix}$$

It's redundant to include $d_J$since

$$\sum_{i=1}^{J} d_i = \sum_{i=1}^{J} \widehat{\pi}_{i+} - \sum_{i=1}^{J} \widehat{\pi}_{+i} = 1 - 1 = 0.$$

Wald Test of Marginal Homogeneity

One can show that $\sqrt{n}(\mathbf{d} - \mathrm{E}(\mathbf{d}))$ has an asymptotic multivariate normal distribution with the covariance matrix $\mathbf{V}$ with the elements below.

$$V_{ab} = n\,\mathrm{Cov}(d_a, d_b) = -(\pi_{ab} + \pi_{ba}) - (\pi_{a+} - \pi_{+a})(\pi_{b+} - \pi_{+b}) \quad \text{for } a \neq b$$

$$V_{aa} = n\,\mathrm{Var}(d_a) = \pi_{a+} + \pi_{+a} - 2\pi_{aa} - (\pi_{a+} - \pi_{+a})^2$$

Wald statistic for testing the H_0 of marginal homogeneity is

$$W = n\mathbf{d}^T\widehat{\mathbf{V}}^{-1}\mathbf{d}$$

which has an approx. chi-squared distribution w/ df $= J - 1$. Here $\widehat{V}$ is the estimate of the covariance matrix V that π_{i+}, π_{+i} and π_{ab} are estimated by

$$\widehat{\pi}_{i+} = \frac{n_{i+}}{n}, \quad \widehat{\pi}_{+i} = \frac{n_{+i}}{n}, \quad \text{and} \quad \widehat{\pi}_{ab} = \frac{n_{ab}}{n}.$$

Score Test of Marginal Homogeneity

The score test estimates the covariance matrix $\mathbf{V}$ under the H_0 of marginal homogeneity: $\pi_{i+} = \pi_{+i}$ using the matrix $\widehat{\mathbf{V}}_0$ with the elements below

$$\widehat{V}_{ab0} = -(\widehat{\pi}_{ab} + \widehat{\pi}_{ba}) = -\frac{n_{ab} + n_{ba}}{n} \quad \text{for } a \neq b$$

$$\widehat{V}_{aa0} = \widehat{\pi}_{a+} + \widehat{\pi}_{+a} - 2\widehat{\pi}_{aa} = \frac{n_{a+} + n_{+a} - 2n_{aa}}{n}$$

Score statistic for testing the H_0 of marginal homogeneity is

$$n\mathbf{d}^T\widehat{\mathbf{V}}_{\mathbf{0}}^{-\mathbf{1}}\mathbf{d}$$

which has an approx. chi-squared distribution w/ df $= J - 1$. Here $\widehat{V}_0$ is the estimate of the covariance matrix V_0 that π_{i+}, π_{+i} and π_{ab} are estimated by

$$\widehat{\pi}_{i+} = \frac{n_{i+}}{n}, \quad \widehat{\pi}_{+i} = \frac{n_{+i}}{n}, \quad \text{and} \quad \widehat{\pi}_{ab} = \frac{n_{ab}}{n}.$$

Coffee Brand Market Share Data in R

```
coffee = read.table(
  "http://www.stat.ufl.edu/~aa/cat/data/Coffee.dat",
  header=TRUE)
```

```
# purchase = 1 for first purchase
# purchase = 0 for second purchase
   person purchase y
1       1        1 1
2       1        0 1
3       2        1 1
4       2        0 1
5       3        1 1
6       3        0 1
(...)
     person purchase y
1079    540        1 5
1080    540        0 5
1081    541        1 5
1082    541        0 5
```

Converting Data to Wide-Format

```
library(reshape2)
coffee.w = dcast(coffee, person ~ purchase, value.var="y")
head(coffee.w)
  person 0 1
1      1 1 1
2      2 1 1
3      3 1 1
4      4 1 1
5      5 1 1
6      6 1 1
colnames(coffee.w)[2:3] = c("y2","y1")
head(coffee.w)
  person y2 y1
1      1  1  1
2      2  1  1
3      3  1  1
4      4  1  1
5      5  1  1
6      6  1  1
```

```r
# wide format to 2-way table
tab = xtabs(~y1+y2, data=coffee.w); tab
   y2
y1    1   2   3   4   5
  1  93  17  44   7  10
  2   9  46  11   0   9
  3  17  11 155   9  12
  4   6   4   9  15   2
  5  10   4  12   2  27
```

$\widehat{\pi}_{ab} = n_{ab}/n$ can be obtained as follows.

```r
ptab = prop.table(tab); ptab
   y2
y1         1        2        3        4        5
  1 0.171904 0.031423 0.081331 0.012939 0.018484
  2 0.016636 0.085028 0.020333 0.000000 0.016636
  3 0.031423 0.020333 0.286506 0.016636 0.022181
  4 0.011091 0.007394 0.016636 0.027726 0.003697
  5 0.018484 0.007394 0.022181 0.003697 0.049908
```

$\widehat{\pi}_{a+} = n_{a+}/n$

```
py1 = prop.table(margin.table(tab, "y1"))
py1
y1
      1       2       3       4       5
0.31608 0.13863 0.37708 0.06654 0.10166
```

$\widehat{\pi}_{+a} = n_{+a}/n$

```
py2 = prop.table(margin.table(tab, "y2"))
py2
y2
     1      2      3      4      5
0.2495 0.1516 0.4270 0.0610 0.1109
```

Sample Covariance Matrix for Wald Statistic in R

$$\widehat{V}_{ab} = -(\widehat{\pi}_{ab} + \widehat{\pi}_{ba}) - (\widehat{\pi}_{a+} - \widehat{\pi}_{+a})(\widehat{\pi}_{b+} - \widehat{\pi}_{+b}) \quad \text{for } a \neq b$$

$$\widehat{V}_{aa} = \widehat{\pi}_{a+} + \widehat{\pi}_{+a} - 2\widehat{\pi}_{aa} - (\widehat{\pi}_{a+} - \widehat{\pi}_{+a})^2$$

```
J = dim(tab)[1]              # J = 5 for Coffee Data
V = array(dim=c(J-1,J-1)) # creating a (J-1)x(J-1) empty array
for(a in 1:(J-1)){
  for(b in 1:(a-1)){
    V[a,b] = - (ptab[a,b]+ptab[b,a]) - (py1[a]-py2[a])*(py1[b]-py2[b])
    V[b,a] = V[a,b]
  }
  V[a,a] = py1[a] + py2[a] - 2*ptab[a,a] - (py1[a]-py2[a])^2
}
V # Sample covariance matrix calculated
        [,1]      [,2]     [,3]      [,4]
[1,]  0.2174 -0.047198 -0.10943 -0.024399
[2,] -0.0472  0.119980 -0.04131 -0.007322
[3,] -0.1094 -0.041311  0.22856 -0.032995
[4,] -0.0244 -0.007322 -0.03299  0.072058
```

Wald Statistic for Marginal Homogeneity

Wald statistic: $W = n\mathbf{d}^T\widehat{\mathbf{V}}^{-1}\mathbf{d}$. Recall $\mathbf{d} = \begin{pmatrix} \widehat{\pi}_{1+} - \widehat{\pi}_{+1} \\ \widehat{\pi}_{2+} - \widehat{\pi}_{+2} \\ \vdots \\ \widehat{\pi}_{(J-1)+} - \widehat{\pi}_{+(J-1)} \end{pmatrix}$

```
n = sum(tab)  # n = number of customers (pairs)
d = py1[1:(J-1)] - py2[1:(J-1)]
Wald = n*t(d) %*% solve(V, d);
Wald  # output is a 1x1 matrix
      [,1]
[1,] 12.58
Wald = as.numeric(Wald); Wald      # Convert the matrix to a number
[1] 12.58
pchisq(Wald, df=J-1, lower.tail=F) # Wald P-value
[1] 0.01354
```

Wald statistic is 12.5771 with df $= 4$, P-value $= 0.0135$, giving some evidence of changes in market shares between the two purchases.

Sample Covariance Matrix for Score Statistic:

$$\widehat{V}_{ab0} = -(\widehat{\pi}_{ab} + \widehat{\pi}_{ba}), \quad \widehat{V}_{aa0} = \widehat{\pi}_{a+} + \widehat{\pi}_{+a} - 2\widehat{\pi}_{aa}$$

```
V0 = array(dim=c(J-1,J-1))
for(i in 1:(J-1)){
  for(j in 1:(i-1)){
    V0[i,j] = - (ptab[i,j]+ptab[j,i])
    V0[j,i] = V0[i,j]
  }
  V0[i,i] = py1[i] + py2[i] - 2*ptab[i,i]
}
```

Score statistic: $W_0 = n\mathbf{d}^T\widehat{\mathbf{V}}_\mathbf{0}^{-1}\mathbf{d}$

```
Score = as.numeric(n*t(d) %*% solve(V0, d)); Score
[1] 12.29135
pchisq(Score, df=J-1, lower.tail=F)
[1] 0.01531125
```

Score statistic is 12.2913 with df $= 4$, P-value $= 0.0153$, giving some evidence of changes in market shares between the two purchases.

`mantelhaen.test()` Does Score Test of Marginal Homogeneity

```
mantelhaen.test(xtabs(~purchase + y + person, data=coffee))


	Cochran-Mantel-Haenszel test


data:  xtabs(~purchase + y + person, data = coffee)
Cochran-Mantel-Haenszel M^2 = 12.2913, df = 4, p-value = 0.015311
```

```
with(coffee, mantelhaen.test(purchase, y, person))


	Cochran-Mantel-Haenszel test


data:  purchase and y and person
Cochran-Mantel-Haenszel M^2 = 12.2913, df = 4, p-value = 0.015311
```

Observe the CMH statistic `M^2 = 12.2913` is exactly the score statistic we computed.

Testing the Change in One Category (1)

As Wald & Score tests indicate changes in market share between purchases, least one of 5 brands must have $\pi_{i+} \neq \pi_{+i}$.

First	Second Purchase						
Purchase	High Pt	Taster's	Sanka	Nescafe	Brim	Total	(%)
High Pt	93	17	44	7	10	171	(31.6%)
Taster's	9	46	11	0	9	75	(13.9%)
Sanka	17	11	155	9	12	204	(37.7%)
Nescafe	6	4	9	15	2	36	(6.7%)
Brim	10	4	12	2	27	55	(10.2%)
Total	135	82	231	33	60	541	(100%)
(%)	(25.0%)	(15.2%)	(42.7%)	(6.1%)	(11.1%)		

To test the change for a given brand, e.g., High Pt, we can combine the other categories and use the methods of Section 8.1.

First	2nd Purchase	
Purchase	High Pt	Other
High Pt	93	78
Other	42	328

First Purchase	2nd Purchase: High Pt	2nd Purchase: Other	
High Pt	93	78	
Other	42	328	
			541

McNemar's test

$$\frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{78 - 42}{\sqrt{78 + 42}} \approx 3.286$$

P-value ≈ 0.00071.

```
2*pnorm(3.386, lower.tail=FALSE)
[1] 0.00070919384
```

95% CI for $\pi_{1+} - \pi_{+1}$

$$\begin{aligned}\widehat{\pi}_{1+} - \widehat{\pi}_{+1} \pm 1.96\text{SE} &= \frac{n_{12} - n_{21}}{n} \pm 1.96\frac{1}{n}\sqrt{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{n}} \\ &= \frac{78 - 42}{541} \pm 1.96\frac{1}{541}\sqrt{78 + 42 - \frac{(78 - 42)^2}{541}} \\ &= 0.0665 \pm 0.0393 = (0.0272, 0.1058)\end{aligned}$$

The brand share of High Pt. dropped 2.7% to 10.6% between the two purchases, with 95% confidence.