STAT 518 --- Section 4.4 --- Measures of Dependence for Contingency Tables

• **We have seen measures of dependence for two numerical variables: for example, Pearson's and Spearman's correlation coefficient.** (also Kendall's tau)

• **For categorical data summarized in a contingency table, we have seen how to test for dependence between rows and columns.**

• **Suppose we wish to measure the degree (or perhaps nature) of the dependence?**

• **The size of the chi-square test statistic *T* tells us something about the degree of dependence, but it is only meaningful relative to the degrees of freedom.**

Cramér's Contingency Coefficient

• **A more easily interpretable measure of dependence than *T* is obtained by dividing *T* by its maximum possible value (for a given *r* and *c*).**

• **This maximum is** $N(q-1)$

where q = the smaller of r or c

• **The square root of this ratio is called Cramér's coefficient:**

$$V = \sqrt{\frac{T}{N(q-1)}}$$

<u>Interpretations</u>: Cramér's coefficient takes values between __0__ and __1__.

- **A value near 0 indicates** little association between row and column variables

- **A value near 1 indicates** strong dependence between row and column variables

- **Cramér's coefficient is <u>scale-invariant</u>: If the scope of the study were increased such that every cell in the table were multiplied by some constant, Cramér's coefficient remains the same.**

Example 1, Sec. 4.2:

	Score			
	Low	**Marginal**	**Good**	**Excellent**
Private	**6**	**14**	**17**	**9**
Public	**30**	**32**	**17**	**3**

***T* was** 17.29 ***N* was** 128 ***q* is** 2

Cramér's coefficient = $\sqrt{\dfrac{17.29}{128(1)}} = 0.368$

– We conclude there is moderate association between school type and score category.

- **We can easily verify that Cramér's coefficient is unchanged if every cell count were multiplied by 10 (or any number).**

Example 2, Sec. 4.2:

		Snoring Pattern		
		Never	Occasionally	≈Every Night
Heart	Yes	24	35	51
Disease	No	1355	603	416

***T* was** 71.75 ***N* was** 2484 ***q* is** 2

Cramér's coefficient = $\sqrt{\dfrac{71.75}{2484(1)}} = 0.17$

→ A mild association between heart disease and snoring pattern.

The Phi Coefficient

- **While Cramér's coefficient measures the degree of association, it cannot reveal the type of association (positive or negative).**

- **The type of association is only meaningful when the two variables have corresponding categories.**

- **The table must be set up so that the row category ordering "matches" the column category ordering.**

- **Phi is calculated as the** Pearson **correlation coefficient between the row variable and the column variable, if the categories are coded as numbers.**

- **For a 2 × 2 table,** $\text{Phi} = \dfrac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}}$

using

		Column 1	Column 2	
Row	1	a	b	r_1
	2	c	d	r_2
		c_1	c_2	N

Interpretations: **The phi coefficient takes values between** -1 **and** 1**.**

- **A value near 0 indicates** little association between row and column variables

- **A value near +1 indicates** a strong tendency for observations to fall in "alike" categories for both rows and columns

- **A value near –1 indicates** a strong tendency for observations to fall in "unlike" categories for both rows and columns

Example 3 (Page 233-234 data tables):

Table A: Phi = $\dfrac{(28)(7)-(0)(5)}{\sqrt{(28)(12)(33)(7)}} = 0.7035 \rightarrow$ strong tendency for mothers and fathers to have "alike" hair colors

Table B: Phi = $-0.3015 \rightarrow$ moderate tendency for mothers and fathers to have "unlike" hair colors.

Table C: Phi = $-0.0144 \rightarrow$ little association between mothers' and fathers' hair colors

Example 4: **Hair Color / Eye Color:**

Phi = $0.341 \rightarrow$ moderate tendency for people with light eyes to have light hair, and dark eyes to have dark hair.

- **For a 2 × 2 table, Phi equals Cramér's coefficient *V* times the sign of** $(ad-bc)$

Proof: For $r=c=2$, the χ^2 test statistic can be written as $T = \dfrac{N(ad-bc)^2}{r_1 r_2 c_1 c_2}$. So $V = \dfrac{\sqrt{\dfrac{N(ad-bc)^2}{r_1 r_2 c_1 c_2}}}{\sqrt{N(q-1)}}$.

Since $q=2$, $V = \dfrac{\sqrt{(ad-bc)^2}}{\sqrt{r_1 r_2 c_1 c_2}}$

Section 4.6 --- Cochran's Test

- In Sec. 5.8 we learned that a <u>block design</u> is simply an extension of a <u>matched-pairs design</u>.

- Instead of each of a <u>pair</u> of similar subjects receiving one of <u>two</u> treatments, we have each of a <u>block</u> of similar subjects receiving one of *c* treatments.

- When the measurements can be ranked (ordinal or stronger data), we have studied nonparametric analyses of both <u>paired</u> and <u>blocked</u> designs.

- When the measurements are binary, we have studied nonparametric analyses of <u>paired</u> designs.

Recall:

Data	Design: Paired	Design: Blocks
Binary	McNemar's	
Ordinal/Stronger	Sign or Signed-Rank	Friedman or Quade

Cochran's

- Now we study block designs with binary measurements. The data are arranged as:

$$
\begin{array}{c|cccc|c}
\text{Blocks} & \multicolumn{4}{c|}{\text{Treatments}} & \\
 & 1 & 2 & \cdots & c & \\
\hline
1 & X_{11} & X_{12} & \cdots & X_{1c} & R_1 \\
2 & X_{21} & X_{22} & \cdots & X_{2c} & R_2 \\
\vdots & & & \vdots & & \vdots \\
r & X_{r1} & X_{r2} & \cdots & X_{rc} & R_r \\
\hline
 & C_1 & C_2 & \cdots & C_c & N
\end{array}
$$

- Since the data are binary, all X_{ij} are either: 0 or 1

Hypotheses of Cochran's Test:

H$_0$: $p_1 = p_2 = \cdots = p_c$ within each block

(The c treatments are equally effective)

where p_j = probability of success (i.e., "1") for treatment j

H$_1$: $p_i \neq p_j$ for some treatments i and j

Development of Cochran's Test Statistic

• **Note that for large r, by the Central Limit Theorem, the j-th column sum C_j =** $\sum_{i=1}^{r} X_{ij}$ is approximately normal

$$\Rightarrow \frac{C_j - E(C_j)}{\sqrt{var(C_j)}} \stackrel{\cdot}{\sim} N(0,1)$$

and so $$\sum_{j=1}^{c}\left[\frac{C_j - E(C_j)}{\sqrt{var(C_j)}}\right]^2 = \sum_{j=1}^{c}\frac{[C_j - E(C_j)]^2}{var(C_j)} \stackrel{\cdot}{\sim} \chi^2_c$$

we estimate E(C_j) by $\bar{C} = \frac{1}{c}\sum_{j=1}^{c} C_j = \frac{N}{c}$

and estimate var(C_j) by $\sum_{i=1}^{r} \hat{p}_i(1-\hat{p}_i) \approx \frac{c}{c-1}\sum_{i=1}^{r}\frac{R_i}{c}\left(1-\frac{R_i}{c}\right)$

$$= \sum_{i=1}^{r}\frac{R_i}{c-1}\,\frac{(c-R_i)}{c}$$

since under H$_0$, p = probability of success is the same for all treatments within a block $\Rightarrow$ p in each row is estimated by proportion of successes in that row: $\frac{R_i}{c}$

So the test statistic is

$$T = \sum_{j=1}^{c}\frac{\left(C_j - \frac{N}{c}\right)^2}{\sum_{i=1}^{r}\frac{R_i(c-R_i)}{c(c-1)}} = \frac{c(c-1)\sum_{j=1}^{c} C_j^2 - (c-1)N^2}{cN - \sum_{i=1}^{r} R_i^2}$$

• By estimating E(C_j) and var(C_j), we lose 1 degree of freedom, so the null distribution is χ^2 with $c-1$ d.f.

• We reject H_0 when T is excessively large.

Decision rule: Reject H_0 if $T > \chi^2_{1-\alpha, c-1}$

• The P-value is found through interpolation in Table A2 or using R.

Note: For $c = 2$ treatments, Cochran's Test is equivalent to McNemar's Test.

Example: We test whether three rock climbs are equally easy. Five climbers attempted each of the three climbs, and their outcomes were recorded as 0 (failure) or 1 (success).

Data:

Climber \ Climb	1	2	3	
1	1	1	0	2
2	1	0	1	2
3	0	0	1	1
4	0	1	1	2
5	1	0	1	2
	3	2	4	9 ← N

H_0: $p_1 = p_2 = p_3$ for each climber (climbs are equally easy)

H_1: $p_i \neq p_j$ for some climbs i, j (difference in ease among the climbs)

Test statistic

$$T = \frac{3(3-1)(3^2+2^2+4^2)-(3-1)(9^2)}{(3)(9)-(2^2+2^2+1^2+2^2+2^2)} = 1.2$$

Decision Rule and Conclusion:

Reject H_0 if $T > \chi^2_{.95,2} = 5.99$. Since $1.2 \not> 5.99$, we fail to reject H_0. The climbs may be equally easy.

P-value $\approx .549$ from R.