- Week 2 –

The Contingency Table and the Chi-Squared Statistic

Professor Eric Beh

School of Mathematical & Physical Sciences
University of Newcastle

Semester 2, 2017

Notation for a Two-Way Contingency Table

Consider

- A random sample of n individuals/units from which two categorical variables are considered
- Two categorical variables, A and B, are cross-classified to form a two-way contingency table, N.
- Let variable A consist of I categories
- Let variable B consist of J categories
- N is of size IxJ

- Denote n_{ij} as the (i, j)th cell frequency
- Denote $n_{i\bullet}$ as the i'th row marginal frequency
- Denote $n_{\bullet j}$ as the j'th column marginal frequency

A/B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	n

We shall consider the case of more than two categorical variables later in this course.

Quetelet's Contingency Tables

One of the first serious categorical data analysts was

Lambert Adolphe Jacques Quetelet
(1796-1874)

His studies on social aspects in France involved the construction of what we now know to be the contingency table. His analysis of these tables was, by current standards, more than superficial, but he did set up the ground word for data analysis that Galton (who developed linear regression analysis) and Pearson (who is a pioneer of categorical data analysis). Their work lives on even today.

What follows are Quetelet's tables, although he did not refer to them as contingency tables.

	1826.	1827.	1828.	1829.	1830.	1831.
Murders in general, -	241	234	227	231	205	266
Gun and pistol, - -	56	64	60	61	57	88
Sabre, sword, stiletto, poniard, dagger, &c.,	15	7	8	7	12	30
Knife, - - - - - - -	39	40	34	46	44	34
Cudgels, cane, &c., -	23	28	31	24	12	21
Stones, - - - - - - -	20	20	21	21	11	9
Cutting, stabbing, and bruising instruments,	35	40	42	45	46	49
Strangulations, - - -	2	5	2	2	2	4
By precipitating and drowning, - - - -	6	16	6	1	4	3
Kicks and blows with the fist, - - - - -	28	12	21	23	17	26
Fire, - - - - - - - -	..	1	..	1	..	..
Unknown, - - - -	17	1	2	..	2	2

He not only speaks of the condition of man at the time, but he also hints at the possibility of being able to model such behaviour.

"I have never failed annually to repeat, that there is a budget which we pay with frightful regularity - it is that of prisons, dungeons, and scaffolds. Now, it is this budget which, above all, we ought to endeavour to reduce; and every year, the numbers have confirmed my previous statements to such a degree, that I might have said, perhaps with more precision ``there is a tribute which man pays with more regularity than that which he owes to nature, or to the treasure of the state, namely, that which he pays to crime". Sad condition of humanity! We might even predict annually how many individuals will stain their hands with the blood of their fellow-men, how many will be forgers, how many will deal in poison, pretty nearly in the same way as we may foretell the annual births and deaths." Quetelet (1842, pg 2)

Lambart Quetelet

Years.	Free Births.		Slave Births.	
	Males.	Females.	Males.	Females.
1813, - -	686	706	188	234
1814, - - - -	802	825	230	183
1815, - -	883	894	221	193
1816, - - -	805	892	325	294
1817, - -	918	927	487	467
1818, - - -	814	832	516	482
1819, - -	810	815	506	509
1820, - - -	881	898	463	464
Total,	6604	6789	2936	2826

An 8x2x2 contingency table cross-classifying the number of births at the Cape of Good Hope (South Africa) between 1813 and 1920 by gender and whether they were “free births” or “slave births”.

More “Early” Contingency Tables

ENFANTS LÉGITIMES. (1824-1825) ENFANTS ILLÉGITIMES.

939 641	garçons.	71 661	garçons
877 931	filles.	68 905	filles.
1,817 572	naissances.	140 566	naissances.

Ce relevé nous conduit à la proposition suivante :
Il naît en France dans l'état de mariage :
51 697 garçons sur 100 000 naissances.

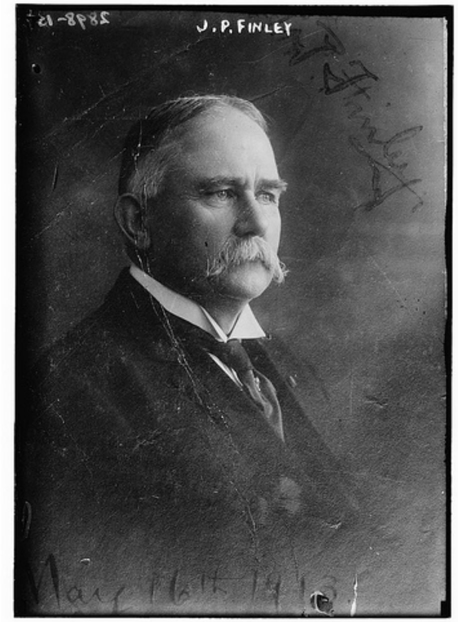
Ce relevé nous conduit à la proposition suivante :
Il naît en France dans l'état de mariage :
50 980 garçons sur 100 000 naissances.

Gender	Legitimate	Illegitimate	Total
Male	939641	71661	1011302
Female	877931	68905	946836
Total	1817572	140566	1958138

Gavarret (1840, pg 93) studied the difference between the proportion of legitimate children that were male with the proportion of illegitimate children that were male between 1824 and 1825. These children were born in France.

Finley's Tornado Data

Month.	Predictions for	Total number.	Number of predictions "favorable for tornadoes."	Fully verified.	Number of predictions "unfavorable for tornadoes."	Fully verified.	Total number made.	Total number fully verified.
March	8 hours	771	43	6	728	721	771	727
April	8 hours	934	25	11	909	906	934	917
May	8 hours	558	10	8	548	542	558	550
May	16 hours	549	22	3	518	511	549	514



John P Finely
(1854 – 1943)

US Army Sergeant John Park Finely was a meteorologist. For four months in 1884 Finley predicted whether or not one or more tornados would occur in each of the eighteen areas of the US he considered, where each daily prediction period lasted eight hours.

While not frequently studied, Finley's "April" data was considered by Goodman and Kruskal (1959, pg 127).

Prediction/Occurrence	Tornado	Not Tornado	Total
Tornado	11	14	25
Not Tornado	3	906	909
Total	14	920	934

Galton and his study of association

Consider two categorical variables, A and B. The simplest quested of such variables is whether they are associated with each other. In a very simple form, we address the hypotheses

H_0: A and B are NOT associated (independent)
H_1: A and B are associated

To help address these hypotheses, we "compare" the observed cell values with the cell values that we would expect to get if the rows and columns are independent.

For the (i, j)th cell, the expected cell frequency (if independence were observed) is

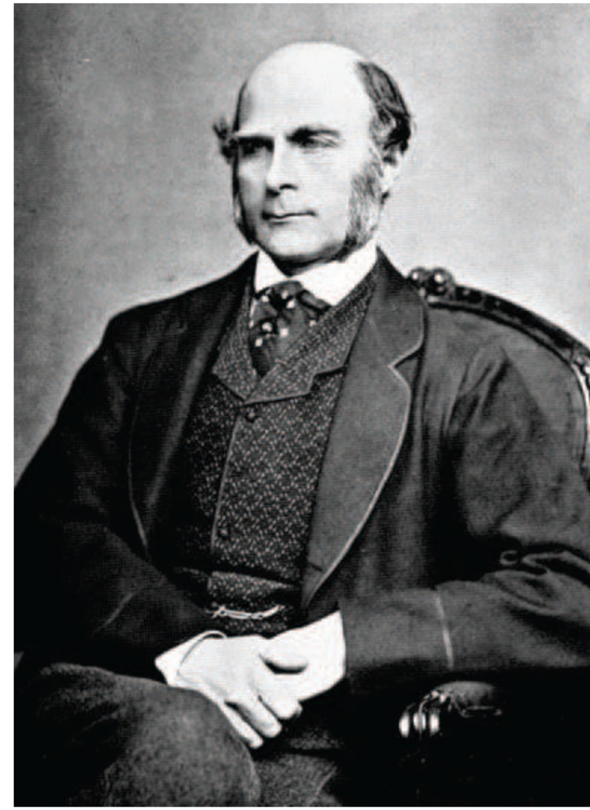
$$\text{Expected (i, j)th count} = \frac{\text{(i'th row total)} \times \text{(j'th column total)}}{\text{sample size}}$$

or more formally

$$E\left(n_{ij}\right) = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

In fact, Galton (1892, pg 174) was the first (know) to state this result. Part of his work involved determining the association of fingerprint characteristics of 105 fraternal (or dizygotic) male twins. One male twin was "earmarked" as twin A and his brother was twin B.

B children.	A children.			Totals in B children.
	Arches.	Loops.	Whorls.	
Arches . . .	5	12	2	*19*
Loops . . .	4	42	15	*61*
Whorls . . .	1	14	10	*25*
Totals in A children	*10*	*68*	*27*	105



Francis Galton
(1822 – 1911)

"The question, then, was how far calculations from the above table would correspondence with the contents of Table [under independence]. The answer is that it does so admirably. Multiply each of the . . . A totals into each of the . . . B totals, and after dividing the result by [n] . . ."

$$\frac{n_{i\bullet} n_{\bullet j}}{n}$$

Galton (1892, pg 175-176) considered

“The squares that run diagonally from the top at the left, to the bottom at the right, contain the double events, and it is with these that we are now concerned. Are entries in those squares larger or not than the randoms . . . The values of 10x19, 68x61, 27x25, all divided by 105?”

Therefore, Galton only considered comparing the observed cell frequencies with their expected value along the diagonals. That is, he looked only at

$$n_{ii} \quad \text{and} \quad \frac{n_{i\bullet} n_{\bullet i}}{n} \quad \text{for } i = 1, 2, 3$$

B children.	A children.			Totals in B children.
	Arches.	Loops.	Whorls.	
Arches . . .	5	12	2	*19*
Loops . . .	4	42	15	*61*
Whorls . . .	1	14	10	*25*
Totals in A children	*10*	*68*	*27*	105

Note: As we shall see, Pearson (1904) developed the now popular Pearson chi-squared statistic. In doing so, he did discuss the idea of expected cell counts under independence in the same way as Galton, but did not mention Galton in his paper.

Pearson's Chi-squared Statistic

Pearson considered a more general setting than what Galton did and compared all observed cell frequencies with their expected values (under independence) – not just the diagonal elements.

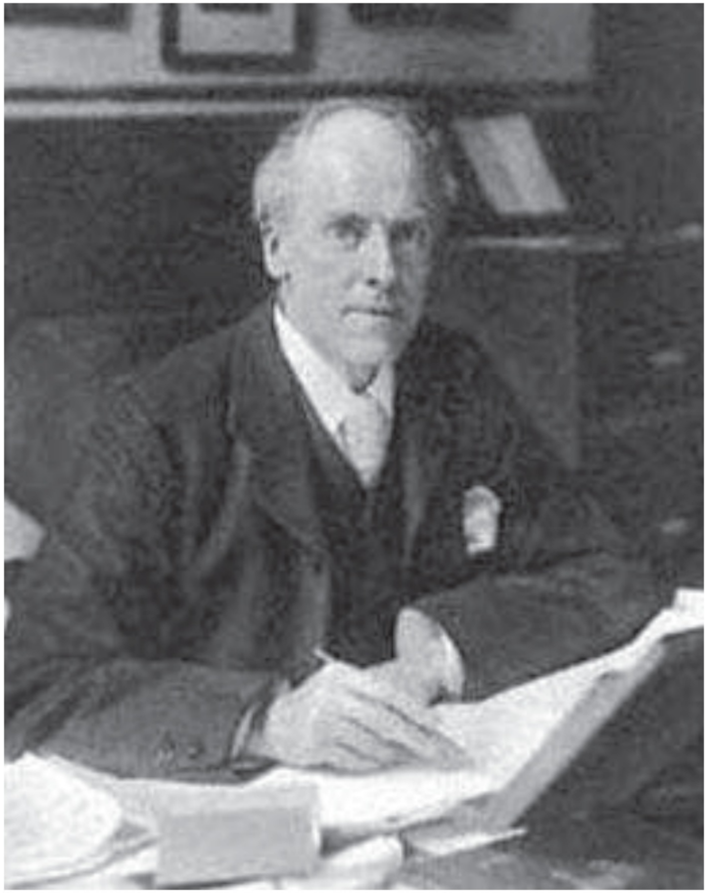
Pearson's approach was to consider looking at the difference between the two:

$$n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}$$

or, equivalently,

$$p_{ij} - p_{i\bullet} p_{\bullet j}$$

Pearson referred to these differences as a cell's *contingency*.



Karl Pearson (1857 – 1936)

If all of the contingency's are zero then there is *complete independence* between the two categorical variables.

Note: Pearson used the word **compartment** while we now use the word **cell**

Suppose we consider the Galton expectations and tie this in with Pearson's idea of a contingency. The hypothesis

H_0: A and B are NOT associated (independent)
H_1: A and B are associated

can be more formally expressed by

$$H_0 : n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

$$H_1 : n_{ij} \neq \frac{n_{i\bullet} n_{\bullet j}}{n}$$

Quantitatively, Pearson proposed the following statistic as a single measure of the strength of the association between the rows and columns of the contingency table

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n} \right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}$$

Here X^2 is a chi-squared random variable with $(I - 1)(J - 1)$ degrees of freedom.

Several more succinct expressions of this statistic can be derived. For example . . .

Suppose we express the above null and alternative hypothesis as

$$H_0 : p_{ij} = p_{i\bullet} p_{\bullet j}$$
$$H_1 : p_{ij} \neq p_{i\bullet} p_{\bullet j}$$

Then an equivalent expression for Pearson's chi-squared statistic is

$$X^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(p_{ij} - p_{i\bullet} p_{\bullet j}\right)^2}{p_{i\bullet} p_{\bullet j}}$$

Here X^2 is also a chi-squared random variable with (I – 1)(J – 1) degrees of freedom.

Alternative Expression

$$
\begin{aligned}
X^2 &= n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij}-p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}} \\
&= n\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\frac{p_{ij}^2-2p_{ij}p_{i\bullet}p_{\bullet j}+p_{i\bullet}^2p_{\bullet j}^2}{p_{i\bullet}p_{\bullet j}}\right] \\
&= n\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\frac{p_{ij}^2}{p_{i\bullet}p_{\bullet j}}-2p_{ij}+p_{i\bullet}p_{\bullet j}\right] \\
&= n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i\bullet}p_{\bullet j}}-2+1\right] \\
&= n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i\bullet}p_{\bullet j}}-1\right] \\
&= n\left[\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)-1\right]
\end{aligned}
$$

Selikoff's Asbestos Data

Occupational Exposure (yrs)	None	Asbestos grade Grade 1	Diagnosed Grade 2	Grade 3	Total
0-9	310	36	0	0	346
10-19	212	158	9	0	379
20-29	21	35	17	4	77
30-39	25	102	49	18	194
40+	7	35	51	28	121
Total	575	366	126	50	1117

Beh and Smith (2011)

```
> selikoff.dat<-matrix(c(310, 212, 21, 25, 7, 36, 158, 35, 102,
+                35, 0, 9, 17, 49, 51, 0, 0, 4, 18, 28),nrow = 5)
> dimnames(selikoff.dat) <- list(paste(c("0-9", "10-19", "20-29",
+                "30-39", "40+")), paste(c("None", "Grade 1",
+                "Grade 2", "Grade 3")))
> selikoff.dat
      None Grade 1 Grade 2 Grade 3
0-9    310      36       0       0
10-19  212     158       9       0
20-29   21      35      17       4
30-39   25     102      49      18
40+      7      35      51      28
>
```

Occupational Exposure (yrs)	None	Asbestos grade Grade 1	Diagnosed Grade 2	Grade 3	Total
0-9	310	36	0	0	346
10-19	212	158	9	0	379
20-29	21	35	17	4	77
30-39	25	102	49	18	194
40+	7	35	51	28	121
Total	575	366	126	50	1117

Selikoff's Asbestos Data

```
> chisq.test(selikoff.dat)

        Pearson's Chi-squared test

data:  selikoff.dat
X-squared = 648.8115, df = 12, p-value < 2.2e-16

Warning message:
In chisq.test(selikoff.dat) : Chi-squared approximation may be
incorrect
>
```

P-values by simulation may also be obtained – using the Monte-Carlo method. R calculates the Monte-Carlo p-value of a contingency table.

```
> chisq.test(selikoff.dat, simulate.p.value = T)

        Pearson's Chi-squared test with simulated p-value (based
        on 2000 replicates)

data:  selikoff.dat
X-squared = 648.8115, df = NA, p-value = 0.0004998
>
```

The simulated (Monte-Carlo) p-value is obtained by randomly generating many hundreds, or thousands, of contingency tables. By default, we have calculated 2000 contingency tables. We could also simulate 10000 tables and obtain the Monte-Carlo p-value by considering

```
chisq.test(selikoff.dat, simulate.p.value=T, B=10000)
```

The algorithm used to randomly generate the contingency tables is that of Patefield (1981). The R function that is used is

```
r2dtable(n, rr, cc)
```

where

- `n` is the number of randomly generated tables
- `rr` is the vector of row totals
- `cc` is the vector of column totals.

For example, to find `rr` and `cc` of `selikoff.dat`:

```
> rr = apply(selikoff.dat, 1, sum)
> cc = apply(selikoff.dat, 2, sum)
>
> rr
  0-9 10-19 20-29 30-39   40+
  346   379    77   194   121
>
> cc
   None Grade 1 Grade 2 Grade 3
    575     366     126      50
```

We can use `r2dtable` to randomly generate 2, say, contingency tables with the same marginal frequencies:

```
> r2dtable(2, rr, cc)
[[1]]
     [,1] [,2] [,3] [,4]
[1,]  172  124   32   18
[2,]  200  117   47   15
[3,]   44   20    8    5
[4,]  100   63   25    6
[5,]   59   42   14    6

[[2]]
     [,1] [,2] [,3] [,4]
[1,]  160  117   50   19
[2,]  202  129   36   12
[3,]   44   20   10    3
[4,]  107   65   16    6
[5,]   62   35   14   10
```

You can use the apply function to confirm the row and column totals of these 2 tables are the same as `selikoff.dat`

Some Properties

There are a few important things to note about the chi-squared statistic which are sometimes overlooked

- When there is complete **independence** in the contingency table (so that each and every one of Pearson's contingency's is zero) then the **chi-squared statistic** will also be **zero**.

  I must concede that this may not necessarily be overly surprising (nor, I concede a point that is often overlooked).

  However, note that a contingency table does not consist of contingency's. Instead, what we refer to as a contingency table is in fact a table of joint frequencies.

What follows are points that are often overlooked and have repercussions on how to use Pearson's chi-squared statistic

Some Properties

- The **chi-squared statistic** remains **unchanged** even if the **rows** and/or **columns** are **interchanged**, or swapped (Pearson was aware of this)
- The **magnitude** of the **chi-squared statistic** is **dependent** on the **sample size**, n, selected. Therefore, for a large enough sample size, it is possible to ALWAYS conclude that an there exists a statistically significant association between the rows and columns, even if the association is very weak.
- In fact

$$0 \leq X^2 \leq n[\min(I, J) - 1]$$

There are a variety of ways of assessing the strength of the association that removes the impact of the sample size. We shall look at them shortly.

- It may seem surprising at first that, in its day, while the classic variance and least squares was known, why didn't Pearson simply consider the sum-of-squares of the contingency's:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2 \qquad ?$$

To answer this question, suppose we consider n_{ij} to be a Poisson random variable so that

$$E(n_{ij}) = Var(n_{ij}) = \frac{n_{i\bullet}n_{\bullet j}}{n}$$

If there are issues concerning the stability of the expectation/variance equality there are ways in which we can deal with this.

Then normalising the cells leads to

$$Z_{ij} = \frac{n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}}{\sqrt{\frac{n_{i\bullet}n_{\bullet j}}{n}}} \sim N(0,1)$$

of which the sum-of-squares is the chi-squared statistic.

(You may remember from STAT2010 that the sum-of-squares of normally distributed variables gives a chi-squared random variable)

On Dealing with the Sample Size

Pearson's phi-squared statistic

One obvious way of dealing with the impact of the sample size on Pearson's chi-squared statistic is to simply divide it by n:

$$\phi^2 = \frac{X^2}{n} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(p_{ij} - p_{i\bullet} p_{\bullet j}\right)^2}{p_{i\bullet} p_{\bullet j}}$$

Pearson (1904, pg 6) referred to this as the *mean-squared contingency*. These days its also called *Pearson's phi-squared statistic*.

- ϕ^2 ranges from 0 (complete independence) to min(I, J) – 1 (complete dependence)
- Its magnitude is **independent** of the sample size

On Dealing with the Sample Size

If $X^2 \sim \chi^2_\alpha(df)$ then, for $c > 0$, $cX^2 \sim \text{Gamma}\left(\frac{df}{2}, \frac{2}{c}\right)$

Thus, since $c = \frac{1}{n} \quad (> 0)$

$$\phi^2 = \frac{1}{n} X^2 \sim \text{Gamma}\left(\frac{\min(I, J) - 1}{2}, \frac{2}{n}\right)$$

Therefore

$$E(\phi^2) = \frac{\min(I, J) - 1}{n}$$

$$\text{Var}(\phi^2) = 2\frac{\min(I, J) - 1}{n^2}$$

$$\text{Skew}(\phi^2) = 2\sqrt{\frac{2}{\min(I, J) - 1}}$$

Problem:
Different sized contingency tables will yield different upper bounds for quantifying the association. This poses problems when comparing the association structure of variables between two contingency tables of different sizes.

On Dealing with the Sample Size

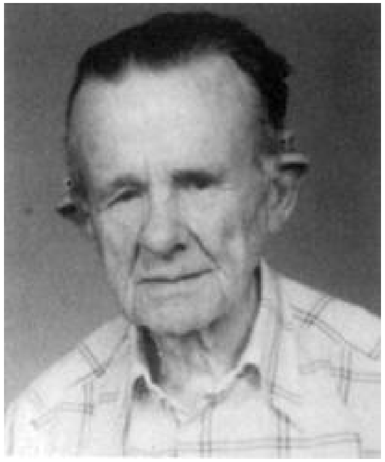
Cramer's Coefficient

One obvious way of dealing with the varying maximum possible phi-squared statistic is to divide by this upper bound. Cramer (1946, pages 282 and 443) proposed the following coefficient:

$$V = \sqrt{\frac{\phi^2}{\min(I, J) - 1}}$$

Therefore

- When there is complete independence between the categorical variables, V = 0
- When there is complete dependence between the categorical variables, V = 1.



Harald Cramér
(1893 – 1985)

On Dealing with the Sample Size

Tchouproff's contingency coefficient (Tchouproff, 1919)

Alexander Alexandrovich Chuprov
(1874 – 1926)

An alternative way to adjust the size of the contingency table when using Pearson’s phi-squared statistic is to consider

$$t = \sqrt{\frac{\phi^2}{(I-1)(J-1)}}$$

When there is complete independence between the rows and columns t = 0. However . . . Liebetrau (1983, page 13) points out the maximum value of Tchouproff’s coefficient is

$$t_{max} = \sqrt[4]{\frac{\min(I, J)-1}{\max(I, J)-1}}$$

Therefore, when a square contingency table is being analysed, the maximum that the Tchouproff coefficient takes is 1. However, if there is a large difference between the number of rows in the contingency table and the number of columns t_{max} *becomes much less than t.*

On Dealing with the Sample Size

Pearson's contingency coefficient

An alternative approach to adjusting Pearson's (1904) mean squared contingency, and one that eliminates the size of the table is to consider the following contingency coefficient

$$p = \sqrt{\frac{\phi^2}{1+\phi^2}}$$

When the contingency table consists of two row categories and two column categories, we obtain equivalent p, t and V values. Liebetrau (1983, page 16) also specifies the variance of these measures. We shall not consider this issue.

However, it is important to note that the maximum value of p is

$$p_{max} = \sqrt{\frac{\min(I, J)-1}{\min(I, J)}}$$

See, for example, Liebetrau (1983, page 14)

On Dealing with the Sample Size

Sakoda's contingency coefficient

For example, the maximum Pearson contingency coefficient for a 2x2 contingency table is

$$p_{\max} = \frac{1}{\sqrt{2}}$$

Therefore, one may amend Pearson's contingency coefficient such that

$$p^* = \frac{p}{p_{\max}} = \sqrt{\frac{\phi^2}{1+\phi^2}\,\frac{\min(I, J)}{\min(I, J)-1}}$$

and is called Sakoda’s (1977) contingency coefficient

Unlike many of the other simple measures of association, Sakoda's coefficient ranges from 0 to 1 for all sample sizes and all sized contingency tables. Its interpretation is a natural one when considering the association between two variables; a zero coefficient indicates perfect independence while a coefficient of one reflects perfect dependence.

Other Contingency Based Measures

One may derive a number of other measures of association based on Pearson's contingency

$$p_{ij} - p_{i\bullet}p_{\bullet j}$$

Some less well known measures for an IxJ contingency table, as summarised by Marcotorchino (1984), include

Belson's statistic $$B = n^2 \sum_{i=1}^{I}\sum_{j=1}^{J}\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2$$

Jordan's statistic $$J = n \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2$$

Variation of Squares $$V = n^2 \sum_{i=1}^{I}\sum_{j=1}^{J}\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)\left(p_{ij} + p_{i\bullet}p_{\bullet j}\right)$$

- They are all zero when the categorical variables are independent
- B and J are at least zero. V can be negative.
- There is no known distributional property of these measures

Example 1: Selikoff's Asbestos Data

Occupational Exposure (yrs)	None	Asbestos grade Grade 1	Diagnosed Grade 2	Grade 3	Total
0-9	310	36	0	0	346
10-19	212	158	9	0	379
20-29	21	35	17	4	77
30-39	25	102	49	18	194
40+	7	35	51	28	121
Total	575	366	126	50	1117

```
                Value   MC.P-value
Chi-sq       648.8115   0
A.chi-sq    1219.7951   0
Belson     41400.1323   0
Jordan         5.0550   0
Var.sq     63297.9966   0
Phi2           0.5809   0
Sakoda         0.6999   0
Tschuprow      0.2200   0
Cramer         0.4400   0
```

We can calculate the Monte-Carlo p-values of each measure of association. 1000 contingency tables were randomly generated using the `r2dtable` function in R

Example 1:

Galton's Fingerprint Data

B children.	A children.			Totals in B children.
	Arches.	Loops.	Whorls.	
Arches . . .	5	12	2	*19*
Loops . . .	4	42	15	*61*
Whorls . . .	1	14	10	*25*
Totals in A children	*10*	*68*	*27*	105

```
            Value MC.P-value
Chi-sq     11.1699 0.031
A.chi-sq   18.8401 0.096
Belson     48.0305 0.195
Jordan      0.0496 0.275
Var.sq    146.9029 0.113
Phi2        0.1064 0.031
Sakoda      0.3798 0.031
Tschuprow   0.1631 0.031
Cramer      0.2306 0.031
```

There might indeed be a reason why the study of these has not continued

While the magnitude of the values is not the same as X^2, they give identical Monte-Carlo p-values

Next Week

- Next week (Week 3) we shall look at other measures of association for two dichotomous categorical variables
- In doing so we shall also explore the use of measures and issues for 2x2 tables including

  - Tetrachoric correlation

  - odds ratios and its variations

- *Week 4* – measures of association for IxJ tables
- *Week 5* – scoring methods for categorical variables (reciprocal averaging, eigen-decomposition, SVD)

References

- Beh EJ and Smith D 2011 Real world occupational epidemiology, Part 2: A visual interpretation of statistical significance. *Archives of Environmental & Occupational Health*, 66, 245 – 248.
- Cramer H 1946 *Mathematical Methods of Statistics*. Princeton University Press
- Finley JP 1884 Tornado predictions. *The American Meteorological Journal* 1, 85 – 88.
- Galton F 1892 *Finger Prints*. Macmillan
- Gavarret J 1840 *Principes generaux de statistique medicale ou developpement des regles qui doivent presider a son emploi*. Bechet Jeune et Labe.
- Goodman LA and Kruskal WH 1954 Measures of association for cross classifications II: Further discussions and references. *Journal of the American Statistical Association 54, 123 – 163*
- Liebetrau AM 1983 *Measures of Association. Sage Publications*
- Marcotorchino, F 1984 *Utilisation des Comparaisons par Paires en Statistique des Contingencies: Partie II*, Report # F 071, Etude du Centre Scientifique, IBM, France.
- Patefield, WM 1981 Algorithm AS 159: An efficient method of generating random RxC tables with given row and column totals. *Applied Statistics* 30, 91-97.
- Pearson K 1904 On the theory of contingency and its relation to association and normal correlation. *Drapers Memoirs. Biometric Series Vol. 1. London.*
- Quetelet A 1842 *A Treatise on Man and the Development of His Faculties. William and Robert Chambers.*
- Sakoda JM 1977 Measures of association for multivariate contingency tables. *Proceedings of the American Statistical Association Social Statistics Section, 777 780.*
- Tchuprow AA 1919 On the mathematical expectation of the moments of frequency distribution. *Biometrika 12, 185*– 210.

References