Contents

1 Examples...

• Montreal Metropolitan Population by knowledge of official language.
Data collected by Statistics Canada at the 1996 census.
[numbers rounded, so subtotals do not sum exactly to total]

		English?		
		Yes	No	Total
Français?	Oui	1,634,785	1,309,150	2,943,935
	Non	280,205	63,500	343,705
	Total	1,914,990	1,372,650	3,287,645

• Stroke Unit vs. Medical Unit for Acute Stroke in elderly?
Patient status at hospital discharge (BMJ 27 Sept 1980)

		Status at Discharge		
		Independent	Dependent	Total
Unit	Stroke	67	34	101
	Medical	46	45	91
	Total	113(58.9%)	79	192

• Bone mineral density and body composition in boys with distal forearm fractures (J Pediatr 2001 Oct;139(4):509-15)

		Fracture?	
		Yes	No
Overweight?	Yes:	36	14
	No:	64	86
	Total	100	100

• Pour battre Roy, mieux vaux lancer bas ... La Presse, Montreal, 21 Avril 1994
Au cours des vingt matches des séries éiminatoires disputés l'an passé, le Canadien a accordé 51 buts... Des 51 buts alloués par le meilleur gardien au monde...

ont vu la rondelle pénétrer dans la partie .. du filet	Haut	10	(20%)
	Milieu	5	(10%)
	Bas	36	(70%)
		51	(100%)

• Distal radial fractures in young goalkeepers: a case for an appropriately sized soccer ball Br J Sports Med 2001; 35: 409-411. Twenty nine fractures of the distal radius were identified in young goalkeepers (age range 6-15 years) as a direct result of saving the ball. Where ball size was known, 12 of the 15 fractures in children aged 11 years or less occurred as the result of impact with an adult sized ball compared with three when a junior ball was involved. This is statistically significant (p = 0.039). In the 10 children aged 12-15 years, only one fracture involved a junior ball; this is also statistically significant (p = 0.027). [??? JH: do these p-values make sense? see follow-up letter.]

• Are there excess Sharons in genitourinary clinics? BMJ Vol 319 18?25 Dec 1999

Most doctors believe that they can determine the age and social class of a patient merely from hearing their name – but this has not been proved. In the 1990s, paediatricians seldom encounter Hildas or Ethels, and Kylies and Bradleys are yet to call on the services of elderly medicine. Stereotypes abound, but is it true that Camillas are more likely to have private medical insurance than Paulines? Above all, are those "Essex girls" Tracey, Sandra, and Sharon really women of easy virtue? With this in mind we set out to establish whether these names are overrepresented among attenders in departments of genitourinary medicine. In the study period 1462 women aged 16-24 attended our department. The ranking and frequency of girls' names

and the mean age of these patients in genitourinary medicine clinics and their frequency in the population for that age group are shown in the table.

Girls' names most frequently encountered in a Southampton genitourinary medicine clinic

Rank in clinic	Name	Mean age	Total (% of 1462 patients)	National rank*	% of birth cohort*
1	Sarah	21.7	55 (3.8)	1	3.8
2	Emma	20.2	35 (2.4)	4	2.3
3	Kelly	20.9	34 (2.3)	47	0.4
4	Louise	19.6	30 (2.0)	13	1.4
5	Claire	21.5	27 (1.8)	2	2.8
6	Lisa	21.3	26 (1.8)	5	2.2
7	Rachel	21.7	23 (1.6)	12	1.4
8	Clare	22.0	22 (1.5)	15	1.1
9	Michelle	21.1	17 (1.2)	7	1.8
10	Nicola	21.4	16 (1.1)	3	2.6
30	Sharon	22.4	7 (0.48)	17	1.0
35	Tracey	22.8	5 (0.34)	26	0.78
62	Sandra	22.0	1 (0.07)	73	0.25

*Data from Office of Population Censuses and Surveys, 1974 database.

2 More generally...

• **Test of Fit of Multinomial Distribution** of the observed frequencies in a characteristic A, say $A_1, A_2, \ldots A_K$ in a *single* sample of size n to internally- or externally-estimated, or model-based (e.g. Mendelian) frequencies.

	A_1	A_2	...	A_K	Total
frequency	y_1	y_2	...	y_K	n

• **Test of independence of two factors in a cross-classification** of a *single* sample of size n with respect to two characteristics, say A and B.

			B			
			B_1	B_2		Total
A	A_1	:	y_{11}	y_{12}	:	y_{A1}
	A_2	:	y_{21}	y_{22}	:	y_{A2}
	Total	:	y_{B1}	y_{B2}	:	n

• **Comparison of rates determined using "Known-Denominators"** [1]: Fixed /Variable follow-up. Person (P) or Population-Time (PT) denominators. (Cross-sectional study documents *states* rather than *events*)

	Event (or state)	non-event (or state)	Total Persons	or	Total P-T
	"cases" (*numerator*)		D (*Den.*)		D (*Den.*)
"exposed" (1)	c_1		D_1		D_1
not exposed (0)	c_0		D_0		D_0

• **Comparison of rates determined using "estimated-denominators"**: Fixed /Variable follow-up. **Estimated** person or population-time (p-t) denominators.

	Event (or state)	non-event (or state)	sample of persons	or	sample of p-t
	"cases" (*numerator*)		d (*den.*)		d (*den.*)
"exposed" (1)	c_1		d_1		d_1
not exposed (0)	c_0		d_0		d_0

Generic formula for X^2 Statistic [2]

$$X^2 = \sum \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency},$$

with the summation over all cells.

[1] JH has deliberately avoided the old-fashioned terminology of "cohort studies" and "case-control studies". Think of both as two variants on *the etiologic study*: the etiologic study consists of a case (numerator) series, and *either* known denominator sizes, *or* a denominator series that is used to *estimate* the (relative) sizes of the denominators that constitute the base that gave rise to the cases. The denominators can pertain to a *closed* population (*cohort*) or an *open* population. If the term "known vs estimated denominators" is too clumsy, use the "case series/base series study" as an more informative term than "case-control study."

[2] See Resources for Original Paper by Karl Pearson (1900). Note also the distinction between the *statistic*, X^2 and the *distribution*, χ^2, just like we distinguish between the statistic or random variable Z, and the distribution, $N(0,1)$.

Yates' Continuity-correction

This is sometimes used to reflect the fact that the binomial or multinomial frequencies (counts)t are discrete and that their probabilities are being approximated by intervals (count-0.5, count+0.5). Tail areas based on the uncorrected X^2 are too small: hence the reduction of each absolute deviation, $|observed\ frequency - expected\ frequency|$, by 0.5.

Notes on use of X^2 statistic

- X^2 must be based on observed *counts*, not on percentages or fractions.
- *Short-cut to manually calculate* X^2 from 2×2 table with frequencies $a,\ b,\ c,\ d$, row totals $r_1,\ r_2$, column totals $c_1,\ c_2$, and overall total n:

$$X^2 = n \times \frac{(a \times d\ -\ b \times c)^2}{r_1 \times r_2 \times c_1 \times c_2}.$$

  The formula involves the crossproducts $a \times d$ and $b \times c$. If their ratio (the empirical odds ratio, $or = ad/bc$) is 1, their difference is zero. The direction of the difference in proportions is given by the sign of $a \times d - b \times c$.

  These formulae avoid the decimals involved in O, E, and $O - E$. Presumably this is why books such as Norman & Streiner's *Statistics: the Bare Essentials*, classify X^2 as "non-parametric" or "distribution-free," and put it in the *non-parametric* chapter. After their 1st edition, I told them that the X^2 statistic tests the difference in two *binomial* proportions: how much more parametric or distribution-specific can that be? See the index in the latest edition to see if I convinced them.

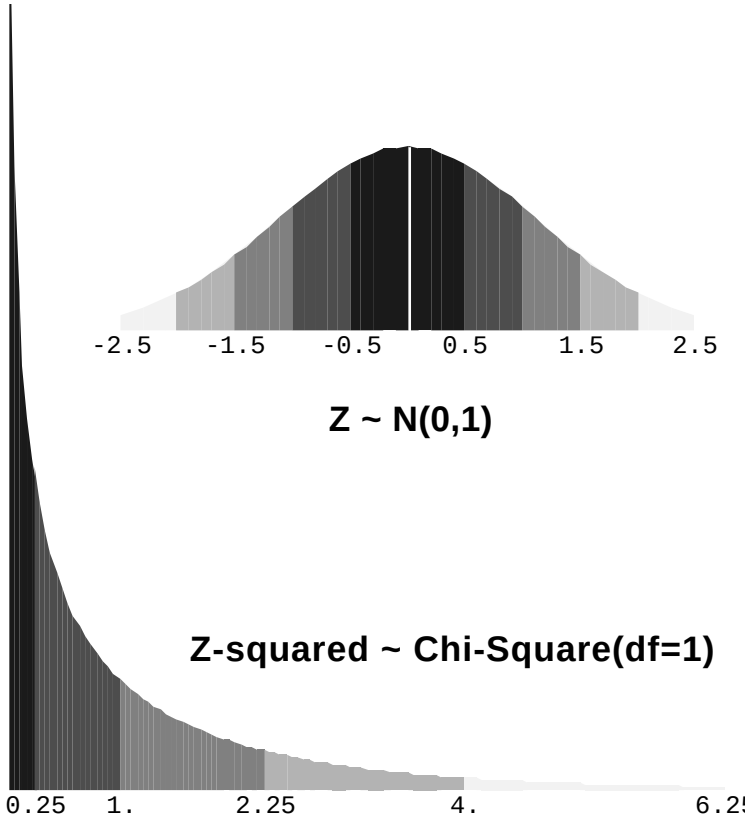
- *Significance and effect size: the* Φ *statistic*: The 'significance' of the observed value of the chi-square statistic X^2 is often given in terms of $P = P[\chi^2 > X^2]$. P does not give any indication of the size of the association, i.e. the effect size, since the value of X^2 is strongly influenced by the sample size n. This can be seen from the above form: for example, even if the difference or ratio between π_1 and π_2 were the same in two different samples sizes, say $n = 87$ and $n = 8700$, the value of X^2 in the case of $n = 8700$ would be 100 times the value in the case of $n = 87$. Think of X^2 as proportional to $n \times (\text{2nd-power})^2/(\text{4th-power}) = n$.
  A more helpful accompanying statistic is Phi (or Cramér's Phi, or Cramér's V). Phi, with its 0-1 range, measures the *correlation* between the variables in the 2×2 table.

$$\Phi = (X^2/n)^{1/2} = \frac{a \times d - b \times c}{(r_1 \times r_2 \times c_1 \times c_2)^{1/2}}.$$

  Cramer's V can be also used for tables with more than 2 rows/columns.

- The uncorrected version of the 2-sided X-statistic for comparing two proportions gives the *same* p-value as the uncorrected version of the X^2 statistic. ***Exercise:*** check that $Z^2 = X^2$.

  Likewise, the corrected version of the 2-sided Z-test for 2 proportions gives the *same* p-value as the corrected version of the X^2 statistic.



- The chi-square random variable (r.v.) is the square of the $N(0,1)$ r.v. The high "probability density," and rapid change in this density, just to the right of $Z^2 = 0$, is a result of the $Z \to Z^2$ transformation: e.g., the 3.98% of the probability mass between $Z = 0$ and $Z = 0.1$ is transferred to the interval 0^2 to 0.1^2, or 0 to 0.01, a width of 0.01 (an identical amount is transferred from the Z interval -0.1 to 0). The 3.94% of the probability mass between $Z = 0.1$ and $Z = 0.2$ is transferred to the small interval 0.1^2 to 0.2^2, or 0.01 to 0.04, a width of 0.03. There is an identical transfer from the Z interval -0.2 to -0.1, for an 'average" χ^2 density of approximately $2 \times 0.0394/0.03 = 2.6$.

- By construction, despite its 1 '*tail*', the X^2 is a 2-*sided* test, unless one uses X and its sign, and refers it to the Z-table.

- There are other chi-square distributions (with df > 1).

- If use X^2 , it must be based on *all* cells, not just on numerators – unless the more common type of outcome is so much more common that the contribution for these cells is negligible (see below).

- The X^2 test is a *large sample* test i.e. it is always an *approximation*. Since X^2 (1 df) is just Z^2, one is no more exact than the other.

- In t-tests, $n = 30$ is often considered "large enough" for large-sample procedures – it depends on how skewed the distribution of the parent data are and whether the Central Limit Theorem will "Gaussianize" the distribution of the statistic . For 0/1 data, the "real" or "effective" sample sizes are not the denominators but rather the numerators. For a ratio-type comparative parameter, the "*effective sample size*" for binary data is the number of subjects having the less common outcome.

- The "guidelines' (such as they are) about when it is appropriate X^2 are based on "**Expected**" frequencies, **not on the observed frequencies**. One quoted rule [often used by computer programs to generate warning messages] is that the expected numbers in most of the cells should exceed 5 for the X^2 test to be accurate. Thus, the 2×2 table on the left below will generate a "warning"; that on the right will not.

5	2		1	11
66	64		11	1

- IMPORTANT: The regular uncorrected X^2 statistic for a single 2×2 table can be written in a seemingly very different format, as

$$X^2 = \frac{(\ a - E[a|H_0]\ )^2}{Var[a|H_0]} = \frac{(\ a - E[a|H_0]\ )^2}{r_1 \times r_2 \times c_1 \times c_2\ /\ N^3}.$$

  The Variance in the denominator of this statistic can be viewed as arising from a statistical model in which the 2 compared proportions are separate independent random variables, i.e. the 'unconditional' or '2- independent binomials' model . Just like the formula with 4 O's and 4 E's, this format is not as calculator-friendly as the shortcut (integer-only) one. But, this form is key to the testing of an association across several 2×2 tables.

3 Mantel-Haenszel Test Statistic for a *single* 2×2 table

3.1 Preamble

The next section also applies to the understanding of the the logic behind Fisher's exact test, déjà vu. They are especially important when we will need an extension of the above formulation of the X^2 test for a *single* table, to one where we combine evidence over *several* (possibly sparse) 2×2 tables.

Cochran, 1954, was the first to propose combining evidence from 2×2 tables. His aim was to combine a small number of 'large' tables, and he did not anticipate that this technique could also be used to combine a *large* number of quite '*small*' 2×2 tables (each one with quite sparse information), with the combination of data from n matched pairs as the limiting case. Thus, he was a wee bit careless about variances. *It took the now famous Mantel-Haenszel*[3] *paper of 1959 to make a variance correction that for 'large 'tables was trivial, but for matched pair tables, was critical.*

`SAS` and others rightly acknowledge Cochran's role in the test statistic, calling it the 'Cochran-Mantel-Haenszel' or 'CMH' statistic. (Indeed 'CMH' is the option one uses with the PROC FREQ to obtain the summary measures and the overall test statistic). This formulation is also the one most commonly used for the log-trank test used to compare two survival curves.

Most consider that the biggest legacy of the "M-H" paper is the Mantel-Haenszel *summary measure (point estimate) of Odds Ratio*. We will come back a little later to this issue of combining data from 2×2 tables.

3.2 Conditional vs. Unconditional?

In the "2 separate binomials" model, the only marginal totals that are fixed ahead of time are the two sample sizes. In most instances, this model reflects reality. The only exception I know of is the design exemplified by the psychophysics study of the lady tasting tea. If she is told that there are 4 cups where the tea is poured first, and 4 where it is poured second, then she will arrange her responses so that there are 4 of each. Thus, in this instance,

[3]Note the correct spelling. JH has proposed that questions on the correct spelling of the statistician and epidemiologist in this partnership, together with questions such as "What is the colour of the covers of the textbooks by Breslow and Day?"; "Complete the partnerships: Doll and ????; Rothman and ?????????;" etc could form a simple screening test to tell bona fide from pretend epidemiologists. JH remembers the correct spelling of Haenszel by remembering that the one letter it doesn't have is the letter t!

both the row totals and the column totals are "fixed" ahead of time, and so it makes sense that the (frequentist) inference be limited to the (only) 5 possible data tables that have all margins fixed.

This 'conditioning on all four marginal totals' is the statistical model behind Fisher's exact test, and indeed Fisher used the tea-tasting example to explain it. But this test is now used for data situations where one cannot – at least ahead of time – consider both sets of marginal totals fixed. For example, in the food sensitivity study, from the answers given, it appears that the subjects were not told that there were 3 three injections of extract and nine of diluent, but the authors used the conditional test anyway.

Many of the reasons put forward for using the conditional test based on all margins fixed (i.e. the hypergeometric model, with only one random variable) involve practicality rather than adherence to a coherent set of inferential principles. They mostly have to do with one of the following 'supposed' difficulties (a) using the normal approximation when the expected numbers are low (b) the fact that there are two parameters, but one is only interested in their difference, or ratio, or odds ratio, and so the 'remaining' parameter is just a 'nuisance' (c) how to order or rank the tables by their degree of evidence against H_0. For example, in a 2×2 table with $n_0 = 23$, and $n_1 = 24$ (as in the bromocryptine and infertility study) , there are theoretically $24 \times 25 = 600$ possible tables. However, if one – after the fact – restricts the analysis to only those tables where the total number of "successes" is 12 (12 pregnancies) , then there are only 13 possible tables (see notes and Excel spreadsheet for Fisher's exact test). And, by reducing the problem from a 2-dimensional one to a 1-dimensional one, is also becomes possible to more easily rank the tables by their degree of evidence against H_0, something that is supposedly more difficult when the tables are simultaneously arrayed along both dimensions. (d) a fourth reason, which I will illustrate with the Marvin Zelen "Marbles in the Folger's Coffee Can" model, is that, after the fact, it is much easier to empirically – and heuristically – demonstrate a low p-value using the single random variable, conditional (hypergeometric), model than it is with the '2-separate binomials' model.

In fact there are many ways to circumvent these objections without having to 'condition' on all margins, and there is still a considerable debate, much of it philosophical, on this 100 years after analyses of 2×2 tables were first introduced. However, since we often combine information from data arranged as matched pairs or 'finely stratified ' strata, we do need to consider this one setting where conditioning is the 'right thing to do.' In the example here, there will only be 1 large table, so the difference will not be important. But when we come to matched pairs, the implications are large.

3.3 Details

In the conditional model, with both margins fixed, there is only one cell entry that can vary independently. Without loss of generality, we focus on the frequency in the 'a' cell. Then, under the null hypothesis,

$$a \sim \text{Hypergeometric}[r_1,\ r_2,\ c_1,\ c_2]$$

Thus,

$$E[a|H_0] = \{r_1 \times c_1\}/n; \quad Var[a|H_0] = \{r_1 \times r_2 \times c_1 \times c_2\}/\{n^2(n-1)\}.$$

Under the null, the expectation is the same under the conditional and the unconditional models. *Note however the difference in the variance*: under the conditional model it is different, since it uses $n^2(n-1)$ rather than n^3. This reflects the different (*narrower*) pattern of variation in the frequency in the 'a' (and consequently in the other 3) cell(s) if all margins are fixed (*vs.* what would happen if the lady were not told "4 where the milk was added 1st; 4 where it was added 2nd").

The test statistic using this conditional variance can be computed as a Z statistic

$$Z = X = \{a - E[a|H_0]\}/SD_{condn'l}[a|H_0],$$

which has the same form as the critical ratios used in the z-test for proportions or means, or as the more traditional squared form

$$X^2_{MH} = \{a - E[a|H_0]\}^2/Var_{condn'l}[a|H_0].$$

Note: The Mantel-Haenszel (MH) test does not use the continuity correction with the $\{a - E[a]\}$. Part of the justification for this is that when the point estimate of the odds ratio falls at the null, i.e. when $a\ timesd = b \times c$, so that $E[a|H_0] = a$, it would be good if the test statistic also had a value of zero. A continuity correction would force the test-statistic to have a positive value even when the "observed a" = "expected frequency under the null" !

3.4 Example

In our stroke vs. medical unit example above, the marginal totals were $r_1 = 101,\ r_2 = 91,\ c_1 = 113,\ c_2 = 79$, so $n = 192$. These yield the "excess in the a cell" of

$$67 - (101 \times 113)/192 = 67 - 59.44 = 7.56,$$

and conditional variance

$$\{101 \times 91 \times 113 \times 79\}/\{192^2 \times 191\} = 11.6528,$$

giving

$$X^2_{MH} = 7.56^2/11.6528 = 4.9.$$

4 2×2, "2×1", "1×2", and "1×1" Tables

2×2	Samples reasonably equal in size, two types of outcome: common e.g. outcomes in trial of stroke vs. medical unit

In italic: *"Observed" frequencies*

In **bold**: **"Expected" frequencies** under H_0: proportions not different
(split the outcomes across 2 samples in ratio of $n_1 : n_2$)

	Outcome		Total
	BAD	GOOD	Persons
sample 1	bad_1	$good_1$	n_1
sample 2	bad_2	$good_2$	n_2
sample 1	$\mathbf{bad_1}$	$\mathbf{good_1}$	n_1
sample 2	$\mathbf{bad_2}$	$\mathbf{good_2}$	n_2

$$X^2 = \frac{(bad_1 - \mathbf{bad_1})^2}{\mathbf{bad_1}} + \frac{(good_1 - \mathbf{good_1})^2}{\mathbf{good_1}}$$

$$X^2 + \frac{(bad_2 - \mathbf{bad_2})^2}{\mathbf{bad_2}} + \frac{(good_2 - \mathbf{good_2})^2}{\mathbf{good_2}}.$$

2×1	2 Samples large and of the same order of magnitude, BAD outcome uncommon : e.g. leukemias and breast cancers

	Outcome		Total Persons or
	BAD	GOOD	Person-Time*
sample 1	bad_1	$good_1$	n_1 or PT_1
sample 2	bad_2	$good_2$	n_2 or PT_2
sample 1	$\mathbf{bad_1}$	$\mathbf{good_1}$	
sample 2	$\mathbf{bad_2}$	$\mathbf{good_2}$	

$$X^2 = \frac{(bad_1 - \mathbf{bad_1})^2}{\mathbf{bad_1}} + \text{Minimal contribution}$$

$$X^2 + \frac{(bad_2 - \mathbf{bad_2})^2}{\mathbf{bad_2}} + \text{Minimal contribution.}$$

* If P-T denominator, there is no "GOOD Outcomes" column.

See Armitage & Berry section 4.10. We will revisit this 2×1 tables, and the 1×1 table, when computing effect measures for Incidence rates.

1×2	1 sample ; two types of outcome common: e.g. male and female births with specific timing of conception

	Outcome		Total
	BAD	GOOD	Persons
sample	*bad*	*good*	n
sample	**bad**	**good**	n

$$X^2 = \frac{(bad - \mathbf{bad})^2}{\mathbf{bad}} + \frac{(bad - \mathbf{bad})^2}{\mathbf{bad}}.$$

"**Expected**" numbers of outcomes under H_0: proportion not different from EXTERNAL proportion. We use EXTERNAL proportion, based on LARGE amount of data (e.g. national data), or theoretical, i.e. model-based, expectation, to calculate the expected split of outcomes. If use *internal* comparison, then we have full 2×2 table.

1×1	1 sample , BAD outcome uncommon e.g. 78 cancers observed in Alberta study, 83.5 expected.

	Outcome		Total Persons
	BAD	GOOD	or person-Time*
sample	*bad*	*MOST*	n or PT
sample	**bad**	**MOST**	n

$X^2 = \frac{(bad - \mathbf{bad})^2}{\mathbf{bad}}$.

"Expected" number of outcomes under H_0: rate not different from **External** rate (use External rate to calculate expected number of BAD events)

This

$$X^2 = \frac{\{observed - \mathbf{expected}\}^2}{\mathbf{expected}}$$

is the square of the familiar large sample Z statistic

$$\{y - \mu_0\} \ / \ {\mu_0}^{1/2}$$

used to test the deviation of a Poisson count y from a null expectation μ_0.

5 Tests of Association — Tables with > 2 rows and/or > 2 columns

The global X^2-statistic, $X^2 = \sum(O-E)^2/E$, with summation over all $r \times c$ cells, is a natural extension of the one for the 2×2 table. Under the null hypothesis of independence of the row and column variables, it (asymptotically) has a chi^2 distribution with $(r-1) \times (c-1)$ degrees of freedom.[4]

It is **appropriate** to use it as a test of the independence of the 2 variables against a *global* alternative, i.e., that there is *some – unspecified –* non-independence. Early examples were to check for evidence of any relationship between laterality of hand and laterality of eye (measured by astigmatism, acuity of vision, etc.) in 413 subjects crossclassified into a 3×3 table. [e.g. Woo, Biometrika 2A 79-148]

One can think of it as a test of the similarity of the multinomial profiles in each row (column).

[4] Fewer if parameters are estimated.

It is **inappropriate** when the **rows or columns have a natural ordering**.

In such instances, omnibus X^2 tests (H_0: identical response profiles) with large *df* are seldom of interest, since the alternative hypothesis (profiles are not identical) is so broad, and **the global X^2 test is invariant to the ordering of the rows and columns**.

More often, a specific alternative hypothesis is of interest; omnibus tests penalize one for looking in *all* directions, when in fact one's focus is narrower, and one wishes to pick up a specific 'signal.' The next 2 examples (> 2 ordered *response* categories in each of 2 groups; binary responses in > 2 ordered *exposure* categories) are a more fruitful step in this direction.

6 Analyzing data from ORDERED categories

Using a global (2 df) chi-square test for the following 2×3 table ignores the **ordered** nature of the responses.

example 1: ordered *response* categories

	Quality of sleep before elective operation			
Patients given...	Bad	Reasonably Good	Good	Total
Triazolam	2	17	12	31
Placebo	8	15	8	31
Total	10	32	20	62

See article by L. Moses L et al, NEJM 311 442-448 1984. (also published as Chapter in Medical Uses of Statistics by J Bailar and F Mosteller.

example 2: ordered *response* categories

	Outcome after 2 to 7 days of Rx*				
Patients given...	1(good)	2	3	4 (poor)	Total
Clotrimazole	6	3	1	0	10
Placebo	1	0	0	9	10
Total	7	3	1	9	20

* in 20 patients with chronic oral candidiasis.

Any dichotomization of outcomes loses information and statistical power. Moses et al. suggest using the Mann-Whitney U test (also known as the Wilcoxon Rank sum test) to take account of ordered nature of response categories.

example 1: ordered *exposure* categories

Distribution of subjects with polluted-water exposure-related symptoms

("Sx") among Competitors and Employees. Relative Risk (RR) According to Number of Falls in the Water.[5]

Groups of subjects	No. *with* Sx	% *with* Sx	No. *without*	Total	RD	RR	OR
Employees*	8	20%	33	41	0	1.0	1.0
Competitors							
0-10 falls	15	44%	19	34	24%	2.3	3.3
11-20 falls	9	45%	11	20	25%	3.5	3.4
21-30 falls	10	71%	4	14	51%	3.7	10.3
>30 falls	10	100%	0	10	80%	5.1	∞

* Reference group; RD Risk Difference; RR Risk Ratio; OR Odds Ratio.

Any dichotomization of exposure loses information and statistical power.

Authors correctly used X^2 test for trend, yielding $X^2_{1df} = 25.3, P = 10^{-6}$.

JH got 24.58 with the "spacing" 0, 5, 15, 25 and 40.

SAS[6], using the "Cochran-Armitage Trend Test," with same spacing, gives a Z statistic of -4.969 ($Z^2 = 24.69$).

The entire variation among the 5 proportions in the table (ignoring ordering) is approximately $X^2_{4df} = 27$, but *it is almost all explained by the exposure gradient.*

In smaller datasets, even if the overall X^2 is not significant, the trend portion can be. In this e.g., there was such a strong relationship that even the overall test was significant. The same is true in the next example (dealing with birth date and sporting success), where again the sample sizes are large and the signal strong.

PS: If you look up Armitage and Berry, you will find another X^2 [Eqn. 12.2]. This value, calculated as the *difference* between the trend and the overall X^2 statistics, can be is used to test if there is serious non-linear variation over and above the linear trend.

[5]Data from "Health Hazards Associated with Windsurfing on Polluted Water " AJPH 76 690-691, 1986 – research conducted at the Windsurfer Western Hemisphere Championship held over 9 days in August 1984. During the championships, the same single-menu meals were served to both competitors and employees.

[6]`[* PROC FREQ; TABLES falls*sick /TREND; ]`

6.1 Test for trend in (response) proportions

from A&B section12.2.

Suppose that, in a $K \times 2$ contingency table the K groups fall into a natural order. They may correspond to different values, or groups of values, of a quantitative variable like age; or they may correspond to qualitative categories, such as severity of a disease, which can be ordered but not readily assigned a numerical value. The usual X^2_{K-1} test is designed to detect differences between the K proportions – without taking the 'ordering' of the rows into account. It is an 'omnibus' test and is unchanged even if we interchange the order of the columns. More specifically one might ask whether there is a significant trend in these proportions from group 1 to group K. Let us assign a quantitative variable, X, to the K groups. If the definition of groups uses such a variable, this can be chosen to be X. If the definition is qualitative, X can take integer values from 1 to K. The notation is as follows:

		Frequency			Proportion
Group	X	Pos.	Neg.	Total	positive
1	x_1	r_1	$n_1 - r_1$	n_1	p_1
2	x_2	r_2	$n_2 - r_2$	n_2	p_2
...					
K	x_K	r_K	$n_K - r_K$	n_K	p_K
All		R	$N - R$	N	$P(= R/N)$

The χ^2_1 **statistic for trend**, X^2_1, which forms part of the overall X^2, can be computed as follows:

$$X^2_{1df} = \frac{N\{N \times \Sigma r_i x_i - R \times \Sigma n_i x_i\}}{R \times (N - R) \times \{N \times \Sigma n_i x_i^2 - (\Sigma n_i x_i)^2\}}$$

From SAS

```
1 line of data for each individual
PROC FREQ DATA= ...  ;
TABLES falls*sick /TREND;

if enter a variable (say "number" to
indicate how many persons has each
exposure/response pattern,
then syntax is...

PROC FREQ DATA= ;
TABLES falls*sick / TREND;
WEIGHT number;
```

From Stata

```
input falls ill number
0 0 33
0 1 8
5 0 19
5 1 15
15 0 11
25 1 10
40 0 0
15 1 9
25 0 4
40 1 10
end
tabodds ill falls [freq=number]
```

Example: Birth date and sporting success

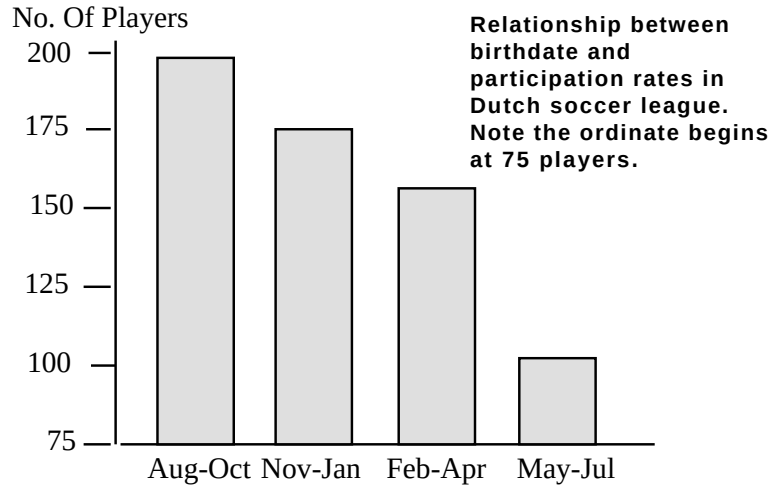
Ad Dudink, Faculty of Psychology, University of Amsterdam, 1018 WB 3 Amsterdam, The Netherlands

Scientific Correspondence in Nature, Vol. 368, 14 April, 1994 p592

Sir - I have found a significant relationship between birth date and success in tennis and soccer. In the Netherlands and England, players born early in the competition year are more likely to participate in national soccer leagues. The high incidence[7] of elite athletes born in the first quarter of the competition year can be explained by the effects of age-group position.

In organized sport. talent is considered predominantly in terms of physical skills. and the influence of social and psychological factors is often ignored or underestimated(1). Various studies have investigated the psychological characteristics of elite athletes(2), but none has looked for an effect of age. I discovered a strikingly skewed distribution of the dates of birth of 12- to l6-year-old tennis players in the top rankings of the Dutch youth league. Half of a sample of 60 tennis players were born in the first 3 months of the year.

This discovery led me to consider the distribution of the dates of birth of professional soccer players. In the Netherlands, there are two leagues comprising a total of 36 clubs. I found a striking difference between participation rates of those born in August and July. The Dutch soccer competition year starts on the first of August. A chi-square test indicates that the distribution is not uniform (P¡0.001); and a regression analysis demonstrates a clear linear relationship between month of birth and number of participants. The dates of birth of 621 players, compiled into quarters, are shown in the figure. This relationship cannot be attributed to the distribution of births in the Netherlands, as this is highly uniform.



PARTICIPATION RATES IN ENGLISH SOCCER LEAGUES

We also inspected the distribution of the dates of birth of English football players in league clubs in the period 1991-92 (3). Birth dates for all players were tabulated by month and compiled into quarters. The results (table) show the significant effect of date of birth on participation rate of soccer players within each of the national leagues, indicating that. as in the Netherlands, significantly more football players are born in the first quarter of the competition year (which starts in September in England).

There is a known relationship between date of birth and educational achievement(5). implying that the younger children in any school year group are at a disadvantage compared to the older children. Children who participate in sports are also placed in age groups, and my results imply many athletes in organized sports may never get a fair chance because of this method of classification. Very little attention has been drawn to this problem. One of the few studies done in this area analysed the dates of birth of young Canadian hockey players in the 1983-84 season(6). Players possessing a relative age advantage (born in the months lanuary-June) were more likely to participate in minor hockey and more likely to play for top teams than players in July-December.

More than 20 years ago, this journal published an article concerning the relationship between season of birth and cognitive development(7). The authors

[7] JH: In epidemiology, we would say 'prevalence'.

attributed this relationship to a fault in the British educational system. A similar relationship was found5 in the Netherlands. Despite this, no action was undertaken to change the educational system. One can only hope that this will not he the case for sports.

	Players in birthdate quarters					Statistics	
League	Sep-Nov	Dec-Feb	Mar-May	Jun-Aug	Total	Chi-Square	Sig. Level
FA premier	288	190	147	136	761	75.5	P<0.0001
Division 1	264	169	154	147	734	48.47	P<0.0001
Division 2	251	168	123	131	673	61.11	P<0.0001
Division 3	217	169	121	102	609	52.38	P<0.0001
Total	1020	696	545	516	2777	230.77	P<0.0001

References: 1 Dudink A Fur J High Ability 1, 144-150 (1990). 2 Dudink A & Bakker. F. Ned. Tschr. Psychol 48. 55 -69 (1993). 3 Rollin J, Rothmans Football Yearbook 1992-93 (Headline. London. 1992). 4 Shearer E, Educ Res 10. 51-56 (1967) 5 Doornbos K. [Date of birth and scholastic performance (Wolters-Noordhoff, Groningen. 1971). 6 Barnsley RH. & Thompson AH. Can. J. Behav. Sci 20. 167-176 (1988). 7 Williams. Ph.. Davies P., Evans, R & Ferguson, N. Nature 228. 1033-1036 (1970).

—

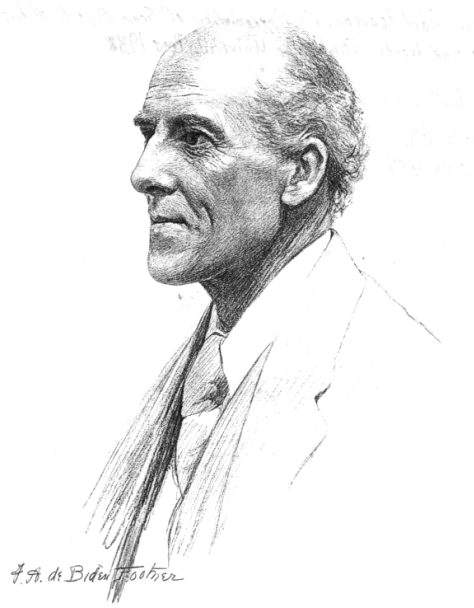
For an example of an analysis of seasonal variation, see the article by H T Srensen et al. Does month of birth affect risk of Crohn's disease in childhood and adolescence? p 907 BMJ VOLUME 323 20 OCTOBER 2001 bmj.com (copy of article, and associated dataset, on course 626 website).

6.2 Test for trend in (response) proportions: other (regression) approaches

The above method is designed to detect linear trends in proportions, and is similar to using a binomial regression model with the `identity` link.

Other options are binomial regression models with the `log` or `logit` link – the latter fits an S-shaped response curve.

All approaches require an X on a numerical scale.



http://www.york.ac.uk/depts/maths/histstat/people/

Illustration III.

In the case of runs of colour in the throws of the roulette-ball at Monte Carlo, I have shown * that the odds are at least 1000 millions to one against such a fortnight of runs as occurred in July 1892 being a random result of a true roulette. I now give χ^2 for the data printed in the paper referred to, *i. e.*:

4274 Sets at Roulette.

Runs	1	2	3	4	5	6	7	8	9	10	11	12	Over 12
Actual ...	2465	945	333	220	135	81	43	30	12	7	5	1	0
Theory ...	2137	1068	534	267	134	67	33	17	8	4	2	1	0

From this we find $\chi^2 = 172\ 43$, and the improbability of a series as bad as or worse than this is about $14{\cdot}5/10^{30}$! From this it will be more than ever evident how little chance had to do with the results of the Monte Carlo roulette in July 1892.

Top: Karl Pearson (1857-1936). Bottom: From Pearson's 1900 paper.