NST 1B Experimental Psychology
Statistics practical 4

$\chi^2$

Rudolf Cardinal & Mike Aitken
1 / 2 March 2005; Department of Experimental Psychology
University of Cambridge

Slides at
pobox.com/~rudolf/psychology
These slides are on the web.
No need to scribble frantically.

pobox.com/~rudolf/psychology
The $\chi^2$ test: for categorical data

(1) “Goodness of fit” test
Categorical data

100 people choose between chocolate and garibaldi biscuits. Every person falls into one of two categories: chocolate or garibaldi. This is categorical data.

If people choose at chance, we’d expect a 50:50 split. — **EXPECTED values** under the hypothesis ‘choose at chance’.

Suppose 65 choose chocolate and 35 choose garibaldi. — **OBSERVED values**.

Do the observed values differ significantly from the expected values? Are the data \(O\) a *good fit* to the ‘model’ \(E\)?

Null hypothesis: observed values do *not* differ from the expected values (the model *is* a good fit to the data).
χ² test with 1 categorical variable, 2 categories

100 people choose between chocolate and garibaldi biscuits

**Expected (E):** 50 chocolate, 50 garibaldi.

**Observed (O):** 65 chocolate, 35 garibaldi.

\[\chi^2 = \sum \frac{(O - E)^2}{E}\]

If O values are close to E values, χ² is small (close to zero).
If O values are very far from E values, χ² is big.
**If χ² is big enough, we will reject the null hypothesis.**

For our biscuit example:

\[\chi^2 = \frac{(65 - 50)^2}{50} + \frac{(35 - 50)^2}{50} = 9\]
Distribution of $\chi^2$ depends on number of degrees of freedom
Degrees of freedom for a goodness-of-fit $\chi^2$ test

100 people choose between chocolate and garibaldi biscuits.  
**Observed (O):** 65 chocolate, 35 garibaldi.  
**Expected (E):** 50 chocolate, 50 garibaldi.

$n = 100$

We have made sure that the expected values add up to the same $n$ as the observed values. Therefore, we lose one $df$.

\[
df = \text{categories} - 1
\]

For our biscuit example: two categories (chocolate, garibaldi), so 1 $df$. We could write  
\[
\chi_1^2 = \frac{(65 - 50)^2}{50} + \frac{(35 - 50)^2}{50} = 9
\]

**Critical value** of $\chi^2$ for 1 $df$ and $\alpha = 0.05$ is 3.84 (from tables). Ours is larger. So we **reject** the null hypothesis (the **model, E**, of a 50:50 split is not a good fit to the data); preferences differed from chance.
$\chi^2$ test with 1 categorical variable, >2 categories

Suppose an Ikea factory makes only rough-hewn pine chair backs, chair seats, and chair legs. We **sample** 50 items at random from the thousands in the warehouse. If they are making items in the correct ratio, there should be one back and one seat for every 4 legs. **Expected (E):** 8.33 backs, 8.33 seats, 33.33 legs. **Observed (O):** 10 backs, 6 seats, 34 legs.

As always,

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Here,

$$\chi^2 = \frac{(10 - 8.33)^2}{8.33} + \frac{(6 - 8.33)^2}{8.33} + \frac{(34 - 33.33)^2}{33.33} = 1$$

$df = \text{categories} - 1 = 3 - 1 = 2$. Critical value of $\chi^2$ for 2 $df$ and $\alpha = 0.05$ is 5.99 (from tables). Ours is **not** larger. So we **don’t** reject the 1:1:4 model; the factory’s doing OK.
Aside: is the $\chi^2$ test one-tailed or two-tailed?

We always test to see if our $\chi^2$ is \textit{bigger} than a critical value ($\chi^2$ is never negative). So the process of testing it is one-tailed.

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

However, the way we \textbf{calculate} $\chi^2$ is inherently two-tailed; $\chi^2$ will get larger whether $O > E$ or $E > O$. So a $\chi^2$ test always performs a \textbf{two-tailed} test on our data. The $\alpha$ in the $\chi^2$ tables is therefore effectively a two-tailed $\alpha$. 
Q. Where does the $\chi^2$ test come from?
A. It’s in the handout, if you’re interested. You don’t need to know. We’ll glance at it briefly.
The binomial distribution: prelude to $\chi^2$

- **Multiple events (trials).** **One of two** things can happen on each event (trial).

- Example: flip a coin $n$ times. Possible outcomes on each trial: heads or tails.
- If the coin is fair, $p = q = 0.5$

- How many heads would we expect to see in $n$ coin flips? How likely are we to see 7 heads in 20 flips if the coin is fair? These questions are answered by the **binomial distribution.** *(Details in the handout; you don’t need to know.)*
- As $n$ increases, the binomial distribution starts to look like the **normal distribution.** *(Below: distribution of total number of heads in 1, 2, 10, and 40 coin flips.)*
Logic of $\chi^2$: only for the mathematically inclined!

2 categories, e.g. $n$ coin flips
- $n$ independent events
- Each event can fall into 2 possible categories (H = heads, T = tails).
- Null hypothesis: the probability that any event falls into category H is $p$; probability of category T is $q$ ($= 1 - p$).
- Under null hypothesis, can calculate probability of a certain number of H events and a certain number of T events using the binomial distribution.
- For large $n$, easier to use the normal approximation to the binomial distribution.
- Likelihood can therefore be described by a single $z$ score. Square to ensure $\geq 0$: get $z^2$.
- $\chi^2_1 = z^2$
- So $\chi^2$ with 1 df is the normal approximation to the binomial distribution.

>2 categories, e.g. $n$ die rolls
- $n$ independent events
- $k$ possible categories (A, B, C, ...).
- Null hypothesis: $P(A) = p$; $P(B) = q$; $P(C) = r$ etc.; $p + q + r + ... = 1$.
- Under null hypothesis, can calculate probability of a certain number of A/B/C/... events using the multinomial distribution.
- Approximate multinomial with $k-1$ different normal distributions.
- Square and add. Likelihood is therefore described by a $\sum z^2$ score.
- $\chi^2_2 = z^2_A + z^2_B$ and $\chi^2_{k-1} = \sum_{i=1}^{k-1} z^2_i$
- So $\chi^2$ with $k-1$ df is the normal approximation to the multinomial.
Applying $\chi^2$ to your data
A very simple example indeed…

Reasoning practical.

In Group A, 14 subjects attempted the *Missionaries & Cannibals* problem (or more, but if so their data weren’t reported!). Of those, 2 subjects solved it in under ten minutes, and 12 didn’t. Suppose we had reason to believe that ten minutes was the median time to solve this problem (i.e. that half would solve faster, and half slower or not at all). Are your data consistent with this hypothesis?

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

\[
df = \text{categories} - 1
\]
In Group A, 14 subjects attempted the M&C problem. Of those, 2 subjects solved it in under ten minutes, and 12 didn’t. Suppose we had reason to believe that ten minutes was the median time to solve this problem (i.e. that half would solve faster, and half slower or not at all). Are your data consistent with this hypothesis?

**Observed:** 2 fast, 12 slow.

**Expected under hypothesis:** 7 fast, 7 slow.

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

\[ = \left( \frac{(2 - 7)^2}{7} \right) + \left( \frac{(12 - 7)^2}{7} \right) \]

\[ = 3.571 + 3.571 \]

\[ = 7.14 \]

\[ df = 2 - 1 = 1 \]

For 1 \( df \), critical value of \( \chi^2 \) for \( \alpha = 0.05 \) is 3.84.

Our \( \chi^2 \) is larger: **reject** hypothesis.
The $\chi^2$ test: for categorical data

(2) “Contingency” test
Now suppose we have two categorical variables. Pugh (1983) examined the decisions of US juries in 358 rape trials. Each trial could be classified according to two categorical variables:

- was the defendant found guilty or not guilty?
- did the defence allege the victim was at fault or not?

In the UK, jury research is restricted by the Contempt of Court Act 1981 and this study would have been illegal.

<table>
<thead>
<tr>
<th>Obtained values</th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>153</td>
<td>24</td>
<td>177</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>105</td>
<td>76</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
</tr>
</tbody>
</table>

Did the two variables influence each other?
Was there a contingency between them?
Was the conviction rate different for ‘low-fault’ and ‘high-fault’?
This table is called a contingency table.
If the two variables **did not** influence each other (**null hypothesis**), what values \((E)\) would we expect? **Not this:**

\[
\begin{array}{c|cc|c}
\text{Wrong expected values} & \text{Guilty verdict} & \text{Not guilty verdict} & \text{Total} \\
\hline
\text{Victim portrayed as low-fault} & 89.5 & 89.5 & 179 \\
\text{Victim portrayed as high-fault} & 89.5 & 89.5 & 179 \\
\text{Total} & 179 & 179 & 358 \\
\end{array}
\]

We know that, **overall,**

- the victim was portrayed as low-fault in 49% (177/358) of cases
- the defendant was found guilty in 72% of cases (258/358).

So our expected values should keep those proportions but have no interrelationship (**contingency**) between the two variables...
\( \chi^2 \) test with 2 categorical variables (contingency test) — 3

<table>
<thead>
<tr>
<th>Obtained values</th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>153</td>
<td>24</td>
<td>177</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>105</td>
<td>76</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
</tr>
</tbody>
</table>

So our expected values should look like this:

<table>
<thead>
<tr>
<th>Expected values</th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>127.559</td>
<td>49.441</td>
<td>177</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>130.441</td>
<td>50.559</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
</tr>
</tbody>
</table>

The row and column totals are now the same as before, so the ‘guilty’ proportion and the ‘low fault’ proportion are the same, but there is **no contingency**: 72\% of defendants in ‘low-fault’ cases are convicted, and so are 72\% of defendants in ‘high-fault’ cases (etc.).

\[
E(row_i, column_j) = \frac{R_iC_j}{n}
\]
Now we can calculate $\chi^2$ exactly as before:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

<table>
<thead>
<tr>
<th>Obtained values</th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>153</td>
<td>24</td>
<td>177</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>105</td>
<td>76</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected values</th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>127.559</td>
<td>49.441</td>
<td>177</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>130.441</td>
<td>50.559</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
</tr>
</tbody>
</table>

$$\chi^2 = \frac{(153 - 127.559)^2}{127.559} + \frac{(105 - 130.441)^2}{130.441} + \frac{(24 - 49.441)^2}{49.441} + \frac{(76 - 50.559)^2}{50.559}$$

$$= 35.9$$

We have made $E$ agree with $O$ as to (1) $n$, (2) proportion low-fault, and (3) proportion guilty, so we have lost 3 $df$ and only have 1 left. In general, for a contingency table,

$$df = (\text{rows} - 1) \times (\text{columns} - 1)$$
χ² test with 2 categorical variables (contingency test) — 5

So we know χ² = 35.9 and df = 1.
Critical value of χ² for 1 df and α = 0.05 is 3.84 (from tables).
Ours is larger. In fact, p < 0.001 (critical value 10.83 for α = 0.001).
So we reject the null hypothesis (the model, E, is not a good fit to the data).

There was a relationship between the defence’s portrayal of the victim and the conviction rate; the defendant was less likely to be convicted if the defence portrayed the victim as being at fault (58% convicted = 105/181) than if they didn’t (86% = 153/177).

<table>
<thead>
<tr>
<th></th>
<th>Obtained values</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Guilty verdict</td>
<td>Not guilty verdict</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Victim portrayed as low-fault</td>
<td>153</td>
<td>24</td>
<td>177</td>
<td></td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>105</td>
<td>76</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Expected values</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Guilty verdict</td>
<td>Not guilty verdict</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Victim portrayed as low-fault</td>
<td>127.559</td>
<td>49.441</td>
<td>177</td>
<td></td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>130.441</td>
<td>50.559</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
<td></td>
</tr>
</tbody>
</table>
Reasoning practical again.

In Group A (who tried *Missionaries & Cannibals*, then the counter-moving problem), 14 subjects attempted the *M&C* problem. Of those, 2 subjects solved it in under ten minutes, and 12 didn’t.

In Group B (who did counter-moving before *M&C*), another 12 subjects attempted the problem. 2 solved it fast, and 10 didn’t.

Did the proportion of people solving the *M&C* problem fast differ across groups?

\[
E(\text{row}_i, \text{column}_j) = \frac{R_i C_j}{n} \quad \chi^2 = \sum \frac{(O - E)^2}{E}
\]

\[
df = (\text{rows} - 1) \times (\text{columns} - 1)
\]
CONTINGENCY EXAMPLE FROM PRACTICAL (2)

<table>
<thead>
<tr>
<th></th>
<th>Obtained values</th>
<th>Slow or not at all</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>2</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>Group B</td>
<td>2</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>22</td>
<td>26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Expected values</th>
<th>Slow or not at all</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>2.154</td>
<td>11.846</td>
<td>14</td>
</tr>
<tr>
<td>Group B</td>
<td>1.846</td>
<td>10.154</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>22</td>
<td>26</td>
</tr>
</tbody>
</table>

\[ E(row_i, column_j) = \frac{R_i C_j}{n} \]

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

\[ = \left( \frac{(2 - 2.154)^2}{2.154} \right) + \left( \frac{(12 - 11.846)^2}{11.846} \right) + \left( \frac{(2 - 1.846)^2}{1.846} \right) + \left( \frac{(10 - 10.154)^2}{10.154} \right) \]

\[ = 0.011 + 0.002 + 0.013 + 0.002 \]

\[ = 0.028 \]

\[ df = (\text{rows} - 1) \times (\text{columns} - 1) = 1 \times 1 = 1 \]

For 1 \( df \), critical value of \( \chi^2 \) for \( \alpha = 0.05 \) is 3.84. **Retain** null hypothesis.
Out of interest, last year (more data, too!), there was an effect:

\[ E(\text{row}_i, \text{column}_j) = \frac{R_i C_j}{n} \]

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

\[ = \left( \frac{(9 - 22.5)^2}{22.5} \right) + \left( \frac{(42 - 28.5)^2}{28.5} \right) + \left( \frac{(36 - 22.5)^2}{22.5} \right) + \left( \frac{(15 - 28.5)^2}{28.5} \right) \]

\[ = 8.1 + 6.395 + 8.1 + 6.395 \]

\[ = 28.99 \]

\[ df = (\text{rows} - 1) \times (\text{columns} - 1) = 1 \times 1 = 1 \]

For 1 \( df \), critical value of \( \chi^2 \) for \( \alpha = 0.05 \) is 3.84. **Reject** null hypothesis.
Assumptions of the $\chi^2$ test: IMPORTANT

The $\chi^2$ test is simple to use, but it is perhaps the most commonly misused statistical test. There are several ways to cock up. The test assumes:

- **Independence of observations.** In the examples so far, one person’s chocolate/garibaldi choice didn’t affect another’s; one court case didn’t affect another. If this isn’t true, can’t use a $\chi^2$ test.
  - **Mustn’t** analyse data from several subjects when there are multiple observations per subject. Need one observation per subject. **Most common cock-up?**
  - Can analyse data from *only* one subject — then all observations are equally independent — but conclusions only apply to that subject.

- **Normality.** There shouldn’t be any very small expected frequencies, or the data won’t approximate a normal distribution (required by the underlying maths). Rule of thumb: **no $E$ value less than 5.** (Possible to go <5 under some special circumstances — see handout — but can never have an $E$ value of 0 or you can’t calculate $\chi^2$!)

- **Inclusion of non-occurrences.** (See next slide.)
Include non-occurrences!

Suppose we ask 20 men and 20 women whether they supported the sale of alcohol in petrol stations. 17 men say yes; 11 women say yes. Do men’s preferences differ from women’s?

This is **wrong**; it omits information about nay-sayers.

<table>
<thead>
<tr>
<th>Obtained values</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support booze</td>
<td>17</td>
<td>11</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 1.29, \text{NS} \]

This is correct:

<table>
<thead>
<tr>
<th>Obtained values</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes to booze</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>No</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 4.29, p < 0.05 \]

Should be easy to understand. The first (incorrect) table could equally represent this situation, which represents a completely different pattern of male/female preference:
Don’t analyse proportions! Analyse the actual numbers.

<table>
<thead>
<tr>
<th>Obtained values</th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>153</td>
<td>24</td>
<td>177</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>105</td>
<td>76</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>358</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 35.9 \]

**WRONG: Observed values as proportions of total**

<table>
<thead>
<tr>
<th></th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>42.7%</td>
<td>6.7%</td>
<td>49.4%</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>29.3%</td>
<td>21.2%</td>
<td>50.6%</td>
</tr>
<tr>
<td>Total</td>
<td>72.1%</td>
<td>27.9%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**WRONG: Expected values as proportions of total**

<table>
<thead>
<tr>
<th></th>
<th>Guilty verdict</th>
<th>Not guilty verdict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim portrayed as low-fault</td>
<td>35.6%</td>
<td>13.8%</td>
<td>49.4%</td>
</tr>
<tr>
<td>Victim portrayed as high-fault</td>
<td>36.5%</td>
<td>14.1%</td>
<td>50.6%</td>
</tr>
<tr>
<td>Total</td>
<td>72.1%</td>
<td>27.9%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Analysing percentages pretends that you had exactly \( n = 100 \) events. Unless you did, the answer will be wrong. If \( n > 100 \), your calculated \( \chi^2 \) will be too low (as here); if \( n < 100 \), your calculated \( \chi^2 \) will be too high.
We suggest you practise with example questions (booklet §1–5, 7) and the past exam questions (§6).

Please note that the description of the exam on p85 is wrong — the NST IB exam has changed this year. Papers 1 & 2: each three hours, six essays; Paper 3 (‘written practical paper’): 1.5h, one stats question (no choice), one experimental design question (choice: one from three).

See www.psychol.cam.ac.uk
   → Teaching Resources
   → Examination Details
   → NST 1B 2005 exam details