Two Variables

We often wish to study not merely a single variable, but rather the relation between two variables as these are realized on a range of individuals. We begin with case of nonnumeric variables.

- no numeric var.
- race and hair color,
- sex/gender and word choice,
- syndrome (e.g., Broca's) and symptom (e.g. pronunciation)

**Methods** to study relations between nonnumeric variables:

- cross tables
- side-by-side box diagrams
- side-by-side histograms
- $\chi^2$ test of independence
Tabular Presentation

**Example:** A test is given to students of Dutch from non-Dutch countries.

**Variables:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>area of origin</td>
<td>EUrope, AMerica, AFrica, ASia</td>
</tr>
<tr>
<td>test score</td>
<td>0-40</td>
</tr>
<tr>
<td>sex</td>
<td>Male, Female</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>area</th>
<th>score</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU</td>
<td>22</td>
<td>M</td>
</tr>
<tr>
<td>AM</td>
<td>21</td>
<td>F</td>
</tr>
<tr>
<td>AF</td>
<td>15</td>
<td>F</td>
</tr>
<tr>
<td>AZ</td>
<td>26</td>
<td>M</td>
</tr>
</tbody>
</table>

Here is part of the results.

Three variables, where only score is numeric, & others nominal. Each row is a CASE.

Tables show all the data, which is nice, but can be also be impractical.

Classifying

It is sometimes useful to group numeric values into classes. We'll group score into 0-16 (beginner), 17-24 (advanced beginner), 25-32 (intermediate), and 33-40 (advanced).

<table>
<thead>
<tr>
<th>area</th>
<th>score</th>
<th>sex</th>
<th>score class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>22</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Grouping numerical information into classes loses information. Care!

Reminder:

<table>
<thead>
<tr>
<th>area of origin</th>
<th>EUrope</th>
<th>AMerica</th>
<th>AFrica</th>
<th>ASia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>sex</td>
<td>Male</td>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Cross-Tables**

Cross-TABLES show frequencies of nominal variables by value—values of one var. horizontally against the values of another vertically. Here is ability class by sex:

<table>
<thead>
<tr>
<th>%</th>
<th>SEX</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL_CLASS</td>
<td></td>
<td>---</td>
<td>---</td>
<td>-------</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>beginner</td>
<td></td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>12</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>advanced</td>
<td></td>
<td>12</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>6</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>intermediate</td>
<td></td>
<td>6</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>advanced</td>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Column</td>
<td></td>
<td>23</td>
<td>17</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>57.5</td>
<td>42.5</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**SPSS: Statistics >> Descriptives >> Cross-Tabs**

Each cell shows the number of cases with the combination of values. Upper left shows that 3 beginners are male, etc. Presentation restructured and summed, but original data not lost. Let's examine other ways of visualizing the data, focusing on the differences in the distribution depending on sex. We might also have looked at the distribution depending on area of origin, but we need to choose a perspective.
Side-by-Side Histograms

Visualize information in cross table using HISTOGRAMS, one for male, one female.

This retains all original information as well, in particular, frequencies.

Relative Histograms

Relative histograms, showing percentages hide some data (the frequencies).

“Advanced” column shows that ca. 2/3 were male, 1/3 female—but hides frequencies.
Relative histograms appropriate when rates are significant.
Segmented Bar Charts

SEGMENTED BAR CHARTS show frequencies and ease visual comparison.

At a glance: more men than women were involved, but more women in better classes. Segmented bar charts recommended means of visualizing relation between nominal variables (side-by-side histograms also OK).

Area and Ability

To examine now the relation between area of origin and language proficiency level in the test, we may examine a segmented bar chart (and a cross-table, of course).
But recall the language proficiency was originally a numeric variable, which we classified for convenience. The classification of the numeric scores hides the detail of the scores themselves. There are other ways of analyzing the relation between one numeric and one nonnumeric variable.

Statistical Analysis of Nominal Data: $\chi^2$

- suitable for nominal data
- compares frequencies in classes
- determine whether there is significant difference in observed freq, expected freq (in respective classes)
- frequent application: test of independence applied to cross-tables

$$\chi^2 = \sum_{i} \frac{(o_i - e_i)^2}{e_i}$$

where $o_i$ observed freq. in class $i$ and $e_i$ expected freq. in class $i$
\( \chi^2 \) Test of Independence

Definition of **CONDITIONAL PROBABILITY**:

\[
P(X|Y) = \frac{P(X, Y)}{P(Y)}
\]

Sometimes we're interested in the probability of a circumstance within a larger group, e.g.,

- the probability that people of strong religious conviction vote for the Christian Democrats \( P(\text{CDA}|\text{Rel.}) \)
- the chance of dyslexia among boys \( P(\text{dyslexia}|\text{sex}=m) \)
- the chance of internet shopping among the young \( P(\text{www-shopper}|(\text{age} \leq 25)) \)

### Statistical Independence

If the condition is irrelevant, then the variables are **STATISTICALLY INDEPENDENT**.

\( X, Y \) are statistically independent iff \( P(X|Y) = P(X) \).

- The chance of rain on Tuesdays \( P(\text{rain}|\text{day=Tues}) = P(\text{rain}) \)
- The chance of a web site being positively evaluated when frames are used.
  \( P(\text{eval}=+|\text{frames-in-www-page}) \)
- The chance of aphasia in right-handed people
  \( P(\text{aphasia}=+|\text{right-handed}) \)

We can check on statistical (independence) simply by counting, and checking whether proportions are the same.
Statistical Independence

The rows of a table show the effect of different conditions on the column variable.

<table>
<thead>
<tr>
<th>Work</th>
<th>Verb</th>
<th>Noun</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 1</td>
<td>32</td>
<td>28</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>Text 2</td>
<td>52</td>
<td>27</td>
<td>21</td>
<td>100</td>
</tr>
<tr>
<td>Totals</td>
<td>84</td>
<td>55</td>
<td>61</td>
<td>200</td>
</tr>
</tbody>
</table>

Where the row, column variables are independent, then the frequencies in the different rows should show about the same proportions (including in the totals).

χ² test of independence test whether these proportions are the same.

Example Application of χ², I.

Wim Vuijk Zicht op interne communicatie, Ch.6
Vuijk compares two versions of a single text, Aanpak Ziekteverzuim, prepared for use in a municipal agency. One version with an introduction stating the purpose of the text, the second without that.
He then asked the question: What is your opinion on the text?

<table>
<thead>
<tr>
<th>W.Intro</th>
<th>'clear'</th>
<th>'unclear'</th>
<th>'other'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

χ² asks: does the variable 'with/without introduction' influence the variable 'clarity'?
Example Application of $\chi^2$, II

**Kos** (thesis under **Bastiaanse**) replicated work by Blumstein, examining phonological errors in different sorts of aphasia.

<table>
<thead>
<tr>
<th></th>
<th>subst</th>
<th>addition</th>
<th>omission</th>
<th>transposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>paarden [pard\ø]</td>
<td>[tard\ø]</td>
<td>[spard\ø]</td>
<td>[ard\ø]</td>
<td>[prad\ø]</td>
</tr>
</tbody>
</table>

She then asked the question: Does the sort of aphasia influence the sort of phonological error?

<table>
<thead>
<tr>
<th></th>
<th>subst</th>
<th>addition</th>
<th>omission</th>
<th>transposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>conduction aphasia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other aphasia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\chi^2$ asks: does 'aphasia sort' influence the (frequency of the values of the variable) 'phonological error sort'?

---

Example Application of $\chi^2$, III

**William Labov** examined variant pronunciations of syllable-final /r/ in American English ([r] vs [ɬ]). New York used to be like Boston, final /r/ is [ɬ], but it started changing in the 1950’s and 1960’s. Labov hypothesized a social basis for the change.

<table>
<thead>
<tr>
<th></th>
<th>/r/ allophones</th>
<th>mixed [r,ɬ]</th>
<th>all cons. [r]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saks</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>Macy’s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.Klein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$N = 68$ 125 71
Example Application of $\chi^2$, III

Labov asked: Does social status influence the pronunciation of final /r/?

<table>
<thead>
<tr>
<th>Social Status</th>
<th>Pronunciation of /r/</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cons. ([r])</td>
<td>vocalic ([ə])</td>
<td>mixed</td>
</tr>
<tr>
<td>high</td>
<td>30</td>
<td>6</td>
<td>32</td>
</tr>
<tr>
<td>medium</td>
<td>20</td>
<td>74</td>
<td>31</td>
</tr>
<tr>
<td>low</td>
<td>4</td>
<td>50</td>
<td>17</td>
</tr>
</tbody>
</table>

$\chi^2$ asks: does 'social status' influence the (frequency of the values of the variable) 'pronunciation of final /r/’?

This led to new understanding of the social motivation for language change.

Example Application of $\chi^2$, IV

$\chi^2$ Test of Independence

Frequent “literary” application of $\chi^2$: authorship studies.

Examples:

- Aristotle: *Rhetoric to Alexander* (Kenny)
- Hamilton, Madison and Jay: *Federalist Papers*  
  cf. Welling’s WWW site:  
  http://www.let.rug/ welling/usa/federalist.html  
  Who was “Publius”?
- W.F.Hermanns war diaries (Dorlein and Zwarts)  
  editing/rewriting much later?
Literary Application of $\chi^2$

We check whether the distribution of a style variable is (proportionally) similar in two authors using $\chi^2$.

1. **Preparation**
   1. Identify quantifiable characteristic of style (literary)
   2. Show that style element characterizes writer (or genre, or epoch, or writer at a particular period on a particular subject, etc.)

2. **Apply** $\chi^2$
   1. Identify null hypothesis (normally that style and text variables are independent)
   2. Test likelihood of sample
   3. If sample unlikely (low p-value), then style and text are dependent.

3. **Combine (1) and (2)**: If $H_0$ is rejected, we conclude that texts differ in style. If style characterize authors (1), we conclude that texts have different authors.

Kenny on Aristotle

**Procedure**

1. Identify characteristic of style (literary)
   Syntax: which part of speech at end of sentence (last word): verb, noun or other

2. Show that it identifies writer (or genre, or epoch, or subject s, etc.) unavailable!

3. Identify null (authorship) hypothesis
distribution (of part of speech at sentence-end) same in questionable work and similar work definitely by Aristotle

4. Test likelihood of null hypothesis
data on (i) *Rhetoric to Alexander* and (ii) *Rhetoric A*
Kenny’s Sample

First 100 sentences of each text:

<table>
<thead>
<tr>
<th>Work</th>
<th>Verb</th>
<th>Noun</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Rhetoric A</em></td>
<td>32</td>
<td>28</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td><em>Rhetoric to Alexander</em></td>
<td>52</td>
<td>27</td>
<td>21</td>
<td>100</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>84</td>
<td>55</td>
<td>61</td>
<td>200</td>
</tr>
</tbody>
</table>

This is a CROSS-TABLE or a $2 \times 3$ CONTINGENCY TABLE

Note generality: test whether distribution of variable values sentence-ending-type is affected by distribution of author.

Calculating $\chi^2$

Problem: what are expected frequencies?
Solution: use existing data (assume that both samples from same population), then relative frequencies (%’s) should be same.

<table>
<thead>
<tr>
<th>Work</th>
<th>Verb</th>
<th>Noun</th>
<th>Other</th>
<th>Total</th>
<th>Verb</th>
<th>Noun</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Rhetoric A</em></td>
<td>32</td>
<td>28</td>
<td>40</td>
<td>100</td>
<td>42</td>
<td>27.5</td>
<td>30.5</td>
</tr>
<tr>
<td><em>Rhetoric to Alexander</em></td>
<td>52</td>
<td>27</td>
<td>21</td>
<td>100</td>
<td>42</td>
<td>27.5</td>
<td>30.5</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>84</td>
<td>55</td>
<td>61</td>
<td>200</td>
<td>42</td>
<td>27.5</td>
<td>30.5</td>
</tr>
</tbody>
</table>

expected freq. = \[ \frac{\text{row-total} \times \text{col.-total}}{\text{grand-total}} \]
Applying $\chi^2$

$$\chi^2 = \sum_{i} \left( \frac{o_i - e_i}{e_i} \right)^2$$

where

- $o_i$ observed freq. in class $i$
- $e_i$ expected freq. in class $i$

In this case $i$ ranges over the cells of the table.

Calculating $\chi^2$

$$\chi^2 = \sum_{i} \frac{(o_i - e_i)^2}{e_i}$$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>$(O - E)^2$</th>
<th>$(O - E)^2 / E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>27.5</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>28</td>
<td>27.5</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>32</td>
<td>42</td>
<td>100</td>
<td>2.38</td>
</tr>
<tr>
<td>52</td>
<td>42</td>
<td>100</td>
<td>2.38</td>
</tr>
<tr>
<td>40</td>
<td>30.5</td>
<td>90.25</td>
<td>2.95</td>
</tr>
<tr>
<td>21</td>
<td>30.5</td>
<td>90.25</td>
<td>2.95</td>
</tr>
<tr>
<td>$\sum$</td>
<td></td>
<td>10.68</td>
<td></td>
</tr>
</tbody>
</table>

This is the $\chi^2$ value.
Degrees of Freedom

In a table, degrees of freedom are not \( \#\text{cells} - 1 \)

<table>
<thead>
<tr>
<th>Work</th>
<th>Verb</th>
<th>Noun</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhetoric A</td>
<td>x</td>
<td>y</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Rhetoric to Alexander</td>
<td></td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>84</td>
<td>55</td>
<td>61</td>
<td>200</td>
</tr>
</tbody>
</table>

Once \( x \) and \( y \) are known, other cells known.
For example, in Rhet-A, “other” must be \( 100 - (x + y) \).
\( \text{dF} = 2 \)

For tables in general, \( \text{dF} = (\#\text{rows} - 1) \times (\#\text{cols} - 1) \)

Applying \( \chi^2 \) tables

\[
\sum \frac{(O_i - E_i)^2}{E_i} = 10.68
\]

\( \text{dF} = 2, \chi^2 = 10.68 \)

How likely is \( H_0 \) (refer to \( \chi^2 \) tables)?

If \( p \leq 0.005 \), should we assume that the two works are by the same author?
Visualizing the $\chi^2$ Question

$\chi^2$ is asking whether freq. distributions are comparable. Relative Histogram:

Since distributions comparable, bars in each cluster should be about the same height. OK here instead of segmented bar chart because of emphasis on relative frequency.

Modifying $\chi^2$

It can be useful to group var. values
Example: you’ve tested language attitudes among foreigners living in the Netherlands
attitude: positive, neutral, negative
nationality: American, English, French, German, Maroccan, Turkish, . . .

—maybe more revealing to group nationality into N-Am., Euro., and Other
—depends on research question!
Yates’s Correction

When investigating $2 \times 2$ tables (4 cells), with small frequencies, the common $\chi^2$ measure overstates the improbability of the sample. **YATES’S CORRECTION** must be applied in this case

$$\chi^2 = \sum_{i} \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

where $o_i, e_i$ observed, expected freq. in class $i$
If the correction is not applied, results may be reported as significant which are not.
It makes little difference when numbers are large (all cell frequencies above 30), and is not needed for tables larger that $2 \times 2$.
Consider also **FISHER’S EXACT TEST** when numbers are small.

Avoid Overuse of $\chi^2$

$\chi^2$ not very sensitive, so use numerical variables where appropriate.
Example: Instead of testing:

*This course was (check one)*

very useful [ ] useful [ ] not useful [ ] completely useless [ ]

Try so-called **Likert scales**

*On a scale of 1-7, this course was*

very useful (1) ______________________ completely useless (7)
Likert Scales

On a scale of 1-7, this course was
very useful (1) _________________ completely useless (7)

- use at least 5 points
- allow a “center” (use 1 through odd number)
- be consistent, keeping “positive” answers on one side
  —e.g., always low (as above)
- compare using t-test or ANOVA!

χ² Goodness of Fit. Since χ² compares observations to expectations, we can apply it to check on how well observations fit a theoretical distribution. **Example:** coin toss, results heads (h) or tails (t), series of 100 tosses recorded.

H₀ coin fair, dist. 50-50 and Hₐ coin unfair, dist. not 50-50

Critical region: let’s reject H₀ only if 95% certain
dF: since there are two variable values (frequencies in respective classes), there is one dF

cf. χ² table, e.g., M&M Tabel G (p.642), P_{df=1}(χ² ≤ 3.84) = 0.95
If χ² > 3.84, then we’re 95% certain that coin is bad.
\[ \chi^2 \text{ Application to Coin Toss} \]

\[ \chi^2 \text{ “Goodness of Fit” test: If } \chi^2 > 3.84, \text{ then we’re 95\% certain that coin is bad.} \]

\[ \chi^2 = \sum_{\forall i} \left( \frac{|o_i - e_i|}{e_i} \right)^2 \]

where \( o_i, e_i \) observed, expected freq. in class \( i \). Expected is 50.

\[ \chi^2 \text{ Calculating } \chi^2 \text{ Test of Independence} \]

Since there are only two (symmetric) classes, we ask when is

\[ 2 \times (o - e)^2 / e = 3.84? \]

\[
\begin{align*}
2 \times (o - e)^2 / e & \geq 3.84 \\
(o - 50)^2 / 50 & \geq 1.92 \\
(o - 50)^2 & \geq 96 \\
o^2 - 100o + 2500 & \geq 96 \\
o^2 - 100o & \geq -2404 \\
o & \leq 40
\end{align*}
\]

If we see \( \leq 40 \) heads (or \( \geq 60 \)), then we’re 95\% certain that coin is dishonest.
Verifying Calculations (\(\chi^2\) Test of Indep.)

\[
\frac{(36 - 50)^2}{50} \approx 3.84 \\
\frac{(14)^2}{50} \approx 3.84 \\
\frac{192}{50} \approx 3.84 \\
3.92 > 3.84
\]

If we use binomial distribution, then we obtain:

\[
SE = \sqrt{p(1 - p)100} = \sqrt{0.5(0.5)100} = \sqrt{0.25100} = \sqrt{25} = 5
\]

The 95% CI is 50 ± (2 × SE) = (40, 60). The numeric test is similar. But \(\chi^2\) still used to check more complicated apparatus, such as roulette wheels.

\(\chi^2\) Test requirements

1. each \(o_i\) falls into exactly one class
2. outcomes for \(N\) independent observations
3. sample \(N\) large

ad(1) no ambiguous classification, no differing classifications
ad(2) study specific. Tense use e.g. shouldn’t study consecutive verbs, but every 100th
ad (3) each class has \(\geq\) 5 elements (expected)
Puzzle about Relations

Here is data from two hospitals about patient survival after surgery. All of the patients who underwent surgery are included in the table. 'Survived' means that the patient was alive six weeks after the surgery.

<table>
<thead>
<tr>
<th></th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deceased</td>
<td>63 (3%)</td>
<td>16 (2%)</td>
</tr>
<tr>
<td>Survived</td>
<td>2037</td>
<td>784</td>
</tr>
<tr>
<td>Total</td>
<td>2100</td>
<td>800</td>
</tr>
</tbody>
</table>

Which hospital would you choose for surgery?

Hidden Variables

With only the original information, Hospital B appears better.

But appearances can deceive. Consider:

<table>
<thead>
<tr>
<th></th>
<th>Good Condition</th>
<th>Poor Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Deceased</td>
<td>6 (1%)</td>
<td>57 (3.8%)</td>
</tr>
<tr>
<td>Survived</td>
<td>594</td>
<td>1433</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>1500</td>
</tr>
</tbody>
</table>

Notate bene In both categories of patients, hospital A has better survival rates. But it attracts more difficult patients (perhaps because it’s better?) So it’s overall survival rate is worse.
### Simpson’s Paradox

Terminology: when we take all patients together, we **AGGREGATE** the values of the variable ‘condition’.

**Simpson’s Paradox**: The distribution of a nominal variable can be reversed under aggregation.

We blame this on **HIDDEN VARIABLES**—a variable which is important, but neglected in analysis. In this case ‘condition of patients’.

---

### Toward Causality

Terminology: **STUDIES** record variables as they are naturally found; **EXPERIMENTS** manipulate situations in order to record variables. **Very difficult** to show cause-effect relationship without experiment

Hidden variables always possible

**Example**: Does smoking cause cancer?
Example: Does smoking cause cancer?

- Smokers suffer lung cancer ten times more often than nonsmokers
- Measure strength of smoking by cigarettes/day, and incidence of lung cancers in percentages (for all population groups)
  \[ r = 0.716, \quad r^2 = 0.51 \]
  Amount of smoking

But the tobacco industry noted that there might be hidden variables:

- Genetic predisposition toward smoking and lung cancer?
  —Later passive smoking linked to cancer, too
- “Unhealthy lifestyle” contributes crucially?

Very difficult to show cause-effect relationship without experiment

Given highly correlating phenomena \( p, p' \), it could be that

- \( p \) causes \( p' \) (or vice versa);
- Both are caused by another, hidden \( c \) (genetic predisposition); or
- There is a crucial confounding factor ("unhealthy lifestyle")
**Toward Causality**

**EXPERIMENT** needed to systematically control potential causes and confounding factors.

“Future oriented” **LONGITUDINAL STUDIES** follow cases chosen for possible common dependence and/or possible. Methodology course (Ensink, Stowe): design of experiments, data acquisition

... and for the smokers: animal experiments have proven the link between smoking and cancers (in animals)

**Next Topics**

Nonparametric Tests — Mann-Whitney, Wilcoxon