Objectives

2.5 Data analysis for two-way tables

- Two-way tables
- Marginal, joint and conditional distributions
- Basic rules in probability (Parts of Section 4.5 in IPS)
- Examples
- Simpson’s paradox
- Applet demonstrating conditional probabilities:
  
  http://onlinestatbook.com/2/probability/conditional-demo.html
Two-way contingency tables

Two-way contingency tables is a method of presenting categorical data which can be easily interpreted. We already came across several examples in Chapter 12, where we looking at the numbers in two different groups.

Example

Recall the Minidoxil/Placebo example considered in Chapter 12.

The spreadsheet where the data is collected would contain 612 rows of the type given in the left.

This is not simple to interpret. Instead when trying to understand the data we write it as a contingency table, which is given below:

<table>
<thead>
<tr>
<th>Improvement</th>
<th>Minidoxil</th>
<th>placebo</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>99</td>
<td>60</td>
<td>159</td>
</tr>
<tr>
<td>No improvement</td>
<td>211</td>
<td>242</td>
<td>463</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>302</td>
<td>612</td>
</tr>
</tbody>
</table>

\[ \hat{p}_M = \frac{99}{310} = 0.32 \]
\[ \hat{p}_P = \frac{60}{302} = 0.2 \]
\[ \hat{p} = \frac{159}{612} = 0.259 \]

Using this table we will define the notion of `marginal', `conditional' and `joint' probabilities and also what statistical independence means.
Proportions and probabilities

This is the same contingency table in Statcrunch (Stat -> Tables -> Contingency -> with summary -> choose Minoxidil and Placebo in Select column(s) -> choose Improve? In Row labels).

We observe that for every cell there are three different proportions associated with it and also a number!
In the next few slides we will explain what each of these terms mean and how they are useful in conveying relationships between categorical variables.

Note: The drug is actually Minodixil – but I can’t spell and the name changes throughout 😊
Marginal probabilities

- The marginal probability is the chance of an event, but completely ignores the influence/effect of other factors.

<table>
<thead>
<tr>
<th></th>
<th>Minidoxil</th>
<th>placebo</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>99</td>
<td>60</td>
<td>159</td>
</tr>
<tr>
<td>No improvement</td>
<td>211</td>
<td>242</td>
<td>463</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>302</td>
<td>612</td>
</tr>
</tbody>
</table>

\[ \hat{p}_M = \frac{99}{310} = 0.32 \quad \hat{p}_P = \frac{60}{302} = 0.2 \quad \hat{p} = \frac{159}{612} = 0.259 \]

Focusing on just hair improvement, we see that 25.9% of the total who received some treatment saw hair improvement, whereas 74.1% did not. These probabilities are **marginal probabilities**, because they do not measure the influence of other variables.

<table>
<thead>
<tr>
<th></th>
<th>Improvement</th>
<th>No Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>( p_I = \frac{159}{612} = 0.259 )</td>
<td>( p_N = \frac{463}{612} = (1 - 0.259) )</td>
</tr>
</tbody>
</table>

The sum of the proportions always add to one.
If we want to understand how one variable may influence another, we need to look at conditional probabilities. Let us return to the Minodixil example. Here we want to understand whether Minodixil increases the chance of hair improvement over the placebo.

<table>
<thead>
<tr>
<th></th>
<th>Minidoxil</th>
<th>Placebo</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>99</td>
<td>60</td>
<td>159</td>
</tr>
<tr>
<td>No improvement</td>
<td>211</td>
<td>242</td>
<td>463</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>302</td>
<td>612</td>
</tr>
</tbody>
</table>

To do this we compare the probabilities in two subgroups, the Minodixil group and the Placebo group. Within each group we calculate the proportion that observed an increase in hair growth after taking the designated treatment.

<table>
<thead>
<tr>
<th></th>
<th>Minidoxil</th>
<th>Placebo</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>$99 \left( \frac{99}{310} = 0.32 \right)$</td>
<td>$60 \left( \frac{60}{302} = 0.2 \right)$</td>
<td>$159 \left( \frac{159}{612} = 0.259 \right)$</td>
</tr>
<tr>
<td>No improvement</td>
<td>$211 \left( \frac{211}{310} = 0.68 \right)$</td>
<td>$242 \left( \frac{242}{302} = 0.8 \right)$</td>
<td>$463 \left( \frac{463}{612} = 0.76 \right)$</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>302</td>
<td>612</td>
</tr>
</tbody>
</table>

We see that 32% percent of the Minoxil group saw an increase in hair growth, whereas only 20% of Placebo group saw an increase. These are called conditional probabilities, because they are conditional on a treatment.
Notation

- To denote a conditional probability we use the notation $P(A|B)$.

<table>
<thead>
<tr>
<th></th>
<th>Minidoxil</th>
<th>Placebo</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>99</td>
<td>60</td>
<td>159</td>
</tr>
<tr>
<td>No improvement</td>
<td>211</td>
<td>242</td>
<td>463</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>302</td>
<td>612</td>
</tr>
</tbody>
</table>

\[ P\left(\text{Hair Improvement} \mid \text{On Minoxidil}\right) = \frac{99}{310} = 0.32 \]

\[ P\left(\text{Hair Improvement} \mid \text{On Placebo}\right) = \frac{60}{302} = 0.2 \]

- In general, we may write $P(A|B)$.

- Where $A$ and $B$ are some events.
Conditional probabilities and independence

- We recall that in Chapter 12 the focus was on comparing the conditional probabilities 
  \( \hat{p}_M = 0.32 \) and \( \hat{p}_P = 0.2 \).

- By using a two sample test on proportions, Minoxidil increased hair growth because 32% was `significantly' larger than 20%.

- **What if these two proportions were the same?**
  - If the proportions in the Minoxidil and Placebo group were the same it would mean that the treatment did not effect the chance of hair growth (for now we ignore that this is a random sample).
  - Further, if the proportions who saw an increase in the minidoxil group and placebo group were the same, then this would be the same as the marginal probability of an increase (regardless of the treatment given).
  - If this were the case, the treatment used is said to be **independent of** hair improvement.

- **Two variables are independent** if the marginal probabilities and conditional probabilities are the same.
  - Put it another way if two variables are independent, then the outcome of one variable (for example treatment given) is not going to change the chances of another variable (for example hair growth).
  - For the Minidoxil example the data suggests a **dependence**.
Example 1 (independence): Gender and grades

- The following pass and fail rate for males and females were observed:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass</td>
<td>108 (108/120 = 0.9)</td>
<td>180 (180/200 = 0.9)</td>
<td>288 (288/320 = 0.9)</td>
</tr>
<tr>
<td>Fail</td>
<td>12 (12/120 = 0.1)</td>
<td>20 (20/200 = 0.1)</td>
<td>32 (32/320 = 0.1)</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>200</td>
<td>320</td>
</tr>
</tbody>
</table>

We see that the proportion of males who passed is the same as the proportion of females who passed (90%), which is the same as the overall proportion who passed. This means that at least for this data set, gender does not influence grades.

Observations:
Do not be fooled by the proportion of passing females (180/320) being greater than the proportion of passing males (108/320). These are joint probabilities, it does not take into account that the number of males and females in the sample are different. This is why we should always compare the conditional probabilities; it takes into account the number in each group.
Example 2 (dependence): Hair and treatment

- Returning to the Minidoxil/Placebo example:

<table>
<thead>
<tr>
<th></th>
<th>Minidoxil</th>
<th>Placebo</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>99 ($\frac{99}{310} = 0.32$)</td>
<td>60 ($\frac{60}{302} = 0.2$)</td>
<td>159 ($\frac{159}{612} = 0.259$)</td>
</tr>
<tr>
<td>No improvement</td>
<td>211 ($\frac{211}{310} = 0.68$)</td>
<td>242 ($\frac{242}{302} = 0.8$)</td>
<td>463 ($\frac{463}{612} = 0.76$)</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>302</td>
<td>612</td>
</tr>
</tbody>
</table>

The conditional probabilities 32% vs 20% (vs the marginal 25.9%) for hair growth are all different, this means that at least for this data set (sample), treatment has an influence on hair growth.

**Observations:**
Of course this is only for a sample, to draw any concrete conclusions about the population we have to sure that these differences cannot be explained by random variables. We actually proved that these differences were statistical significant difference in Chapter 12 (and we will do the same thing again in Chapter 14).
Example 3 (very dependent): Oscar and dresses

Here we look at what people wore at the Oscars (sample size 420):

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dress</td>
<td>0 ( \frac{0}{200} = 0.00 )</td>
<td>215 ( \frac{215}{220} = 0.977 )</td>
<td>215</td>
</tr>
<tr>
<td>No Dress</td>
<td>200 ( \frac{200}{200} = 1 )</td>
<td>5 ( \frac{5}{220} = 0.023 )</td>
<td>205</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>220</td>
<td>420</td>
</tr>
</tbody>
</table>

The chance a male wears a dress is 0% (in this sample), whereas the chance female wears a dress is 97.7% (in this sample).

**Observations:** In this example we see there is a very strong dependence between gender and whether they wore a dress or not.

The chance of a person (regardless of gender) wearing a dress is \( \frac{215}{420} = 51.1\% \). However, once we know that the person is female the chance shoots up to 97.7%, on the other hand if we know that person is male the chance drops to 0%.
Joint Probabilities

- A joint probability is the chance of two different events happening together. Look at the Minidoxil example:

<table>
<thead>
<tr>
<th></th>
<th>Minidoxil</th>
<th>Placebo</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>99 ($\frac{99}{612} = 0.16$)</td>
<td>60 ($\frac{60}{612} = 0.098$)</td>
<td>159 ($\frac{159}{612} = 0.259$)</td>
</tr>
<tr>
<td>No improvement</td>
<td>211 ($\frac{211}{612} = 0.34$)</td>
<td>242 ($\frac{242}{612} = 0.395$)</td>
<td>463 ($\frac{463}{612} = 0.76$)</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>302</td>
<td>612</td>
</tr>
</tbody>
</table>

We see that 16% of the sample took Minidoxil and had hair improvement, whereas 9.8% of the sample took the Placebo and had saw an improvement in hair etc. Don't get conditional and joint probabilities mixed up.

The sum of all the joint probabilities is always one.

Formula for calculating joint probabilities in terms of conditionals and marginals.

\[
P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)
\]

In the next few slides we illustrate it.
Dependent events (examples)

- What is the chance you will see a person in the middle of campus (on any given day) is eating an ice-cream?
  - If we have no other information about the day. Then this is a marginal probability. Suppose the marginal probability is 0.2 (20%).
    \[ P(\text{ice cream}) = 0.2 \]

- Suppose you have information about the day. Today the weather is windy and rainy. With this additional information what is the chance of seeing a person eat an ice-cream?
  - When it is rainy and wet, very few people want to eat ice-cream. The chance drop to 0.01 (1%)
  - This is called a conditional probability, because we have reevaluated the chance based on additional information. We write it as
    \[ P(\text{ice cream} | \text{cold weather}) = 0.01 \]

- Suppose the weather is warm, what is the chance?
  - This is a conditional probability, which we write as
    \[ P(\text{ice cream} | \text{hot weather}) = 0.221 \]

- Because weather has an influence on ice cream consumption. We say there is a dependence between weather and ice cream.
If additional information changes the chance of an event, then there is dependence.

- In the previous example, knowledge about the weather changed the likelihood of seeing a person eating an ice cream.
- The type of therapy a person receive may change of their survival (for example chemotherapy or operation in the treatment of cancer).
- Gender may have an influence of whether a person wears trousers or a dress to an occasion.

If an additional piece of information does not change the chance. This two events are said to independent.

- For example, the chance of getting an A is 25%. If you know the person is female and chance remains 25%, then you can say grades and gender are independent events i.e.

\[ P(\text{Grade A}|\text{Female}) = P(\text{Grade A}) \]

- On the other hand, if you know that person studied 20 hours per week, then it would change the chance of their attaining an A. We conclude that gender does not influence final grade but the amount of hours studied does.
Connecting marginal, joint and conditional probabilities

- The formula connecting all probabilities is
  \[ P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A) \]

- Example in Texas we know:
  \[ P(\text{Cold weather}) = 0.1 \quad P(\text{Warm weather}) = 0.9 \]

  - Using the previous slide we see this is
  \[
  P(\text{Cold weather and Ice cream}) = P(\text{Ice Cream}|\text{Cold weather}) P(\text{Cold Weather}) \\
  = 0.01 \times 0.1 = 0.001
  \]

  - This chance is extremely small because the chances of cold weather and eating ice cream when it is cold are both extremely small.
Example 1: calculating joint probabilities (minidoxil)

The proportion of people who took minidoxil and saw an improvement is

\[
0.1618 = \frac{\text{no. who took minidoxil and improved}}{\text{total number in study}}
\]

However, this probability can be calculated in different ways. We know that the proportion who took minidoxil is 50.65% and the proportion of the minidoxil group who saw an improvement is 31.94%, thus the joint probability is their product:

\[
0.1618 = \frac{99}{310} \times \frac{310}{612} = \frac{99}{612} = P(\text{Minidoxil and improved})
\]
Multiplicative Rule and independence

- We use the notation $P(\text{Passed}|\text{male})$ to denote the proportion of males who pass (or the probability of passing conditioned on the person being male).

- In some situations it is not easy to directly calculate this. Instead, if we know the marginal and conditional probabilities we use the multiplicative rule:
  \[ P(\text{Male and Passed}) = P(\text{Passed}|\text{male})P(\text{Male}) \]

- If there is no association between two variables (or equivalently they are statistically independent), then the calculation can be written in terms of marginal probabilities. Since $P(\text{Passed}|\text{male}) = P(\text{Passed})$, then we have
  \[ P(\text{Male and Passed}) = P(\text{Passed}|\text{male}) \times P(\text{Male}) = P(\text{Passed})P(\text{Male}) \]

But this probability calculation only holds if passing and gender have no influence on each other. If they do, this calculation is **WRONG** and can lead to misleading conclusions.
Example 2: calculating joint probabilities (gender)

To calculate the proportion of students who were males and passed:

\[
0.2455 = \frac{108}{440} = \frac{\text{males and passed}}{\text{total}}
\]

Again this probability can be calculated using the marginal and the conditional in tandem:

\[
0.2455 = \text{proportion of males who passed} \times \text{proportion of males}
\]

\[
\frac{108}{120} \times \frac{120}{440} = \frac{108}{440}
\]

\[
= P(\text{Male and Passed})
\]

But something interesting happens in this case, the proportion of males who passed is the same as the proportion who passed (independence means the conditional is the same as the marginal) :

\[
0.2455 = \text{proportion who passed} \times \text{proportion of males}
\]

\[
\frac{396}{440} \times \frac{120}{440} = \frac{108}{440}
\]

\[
= P(\text{Male and Passed})
\]
Example 3: calculating joint probabilities (Oscars)

To calculate the proportion who are female and also wore a dress:

\[
0.5119 = \text{proportion who are female and wore a dress} = \frac{215}{440}
\]

Again this probability can be calculated using the marginal and the conditional in tandem:

\[
0.5119 = \text{proportion of females who wore a dress} \times \text{proportion of females}
\]

\[
\frac{215}{220} \times \frac{220}{420} = \frac{215}{420} = \frac{215}{440}
\]

Observe that in this case because dress and gender are dependent we cannot use the product of the marginals to obtain the joint probability, ie.

\[
0.5119 = \text{proportion of females and wore a dress} \neq \text{proportion of females} \times \text{proportion who wore a dress} = 0.5238 \times 0.5119 = 0.2681.
\]
Example 4: Parental smoking

Does parental smoking influence the smoking habits of their high school children?

Summary two-way table:
High school students were asked whether they smoke and whether their parents smoke.

<table>
<thead>
<tr>
<th></th>
<th>Student smokes</th>
<th>Student does not smoke</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both parents smoke</td>
<td>400</td>
<td>1380</td>
<td>1780</td>
</tr>
<tr>
<td>One parent smokes</td>
<td>416</td>
<td>1823</td>
<td>2239</td>
</tr>
<tr>
<td>Neither parent smokes</td>
<td>188</td>
<td>1168</td>
<td>1356</td>
</tr>
<tr>
<td>Total</td>
<td>1004</td>
<td>4371</td>
<td>5375</td>
</tr>
</tbody>
</table>

Marginal distribution for the categorical variable “parental smoking”:
The row totals are used and re-expressed as percent of the grand total.

<table>
<thead>
<tr>
<th></th>
<th>Both parents smoke</th>
<th>One parent smokes</th>
<th>Neither parent smokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of Students</td>
<td>33.1%</td>
<td>41.7%</td>
<td>25.2%</td>
</tr>
</tbody>
</table>

The percents are displayed in a bar graph.
Comparing the conditional probabilities with the marginal probabilities we see that from this sample there is a 39.84% chance a student smokes if both parents smoke whereas overall there appears to be a 18.68% chance of a student smoking. This difference between the conditional and marginal probabilities means that for this sample there is a dependence between whether a student smokes and whether their parents smoked. In Chapter 14 we use this data to infer dependencies over the entire population of students.

Contingency table results:
Rows: Parent
Columns: None

<table>
<thead>
<tr>
<th></th>
<th>Student smokes</th>
<th>Does not smoke</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Both smoke</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>1380</td>
<td>1780</td>
<td></td>
</tr>
<tr>
<td>(22.47%)</td>
<td>(77.53%)</td>
<td>(100.00%)</td>
<td></td>
</tr>
<tr>
<td>(39.84%)</td>
<td>(31.57%)</td>
<td>(33.12%)</td>
<td></td>
</tr>
<tr>
<td>(7.442%)</td>
<td>(25.67%)</td>
<td>(33.12%)</td>
<td></td>
</tr>
<tr>
<td>332.5</td>
<td>1448</td>
<td>1824.5</td>
<td></td>
</tr>
<tr>
<td><strong>One smokes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>416</td>
<td>1823</td>
<td>2239</td>
<td></td>
</tr>
<tr>
<td>(18.58%)</td>
<td>(81.42%)</td>
<td>(100.00%)</td>
<td></td>
</tr>
<tr>
<td>(41.43%)</td>
<td>(41.71%)</td>
<td>(41.66%)</td>
<td></td>
</tr>
<tr>
<td>(7.74%)</td>
<td>(33.92%)</td>
<td>(41.66%)</td>
<td></td>
</tr>
<tr>
<td>418.2</td>
<td>1821</td>
<td>2239.2</td>
<td></td>
</tr>
<tr>
<td><strong>None smoke</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>188</td>
<td>1168</td>
<td>1356</td>
<td></td>
</tr>
<tr>
<td>(13.86%)</td>
<td>(86.14%)</td>
<td>(100.00%)</td>
<td></td>
</tr>
<tr>
<td>(18.73%)</td>
<td>(26.72%)</td>
<td>(25.23%)</td>
<td></td>
</tr>
<tr>
<td>(3.498%)</td>
<td>(21.73%)</td>
<td>(25.23%)</td>
<td></td>
</tr>
<tr>
<td>253.3</td>
<td>1103</td>
<td>1356.3</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>4371</td>
<td>5375</td>
<td></td>
</tr>
<tr>
<td>(18.68%)</td>
<td>(81.32%)</td>
<td>(100.00%)</td>
<td></td>
</tr>
<tr>
<td>(100.00%)</td>
<td>(100.00%)</td>
<td>(100.00%)</td>
<td></td>
</tr>
<tr>
<td>(18.68%)</td>
<td>(81.32%)</td>
<td>(100.00%)</td>
<td></td>
</tr>
</tbody>
</table>
Example 5: Twins and Independent Events

- You read the following in a local newspaper:
  - The probability of a couple having fraternal twins (non-identical twins) is 1 in 60 (1/60). Tom and Jerry, a couple from Forth Worth, had a set of non-identical twins in 2012 and recently gave birth to another set of non-identical twins. The chance of this happening is 1 in 3600! How amazing is this!

- **Question**: Was the probability calculated correctly?
- **Answer**: By using the multiplicative rule (based on marginal and conditional probabilities) we have:

\[
P(\text{First set and Second set}) = P(\text{First set}) \times P(\text{Second set given| First set})
\]

\[
= \frac{1}{60} \times P(\text{Second set given| First set})
\]

The newspaper makes the assumption that the conditional probability is the same as the marginal probability. This is ONLY true if the chance of getting twins are independent events, that is there is no genetic factor playing a role. Unless you really have some idea of the biology you should not make this assumption.

In fact, there is a genetic element to fraternal twins and \( P(\text{Second set given| First set}) > \frac{1}{60} \)

The chance of having two sets of fraternal twins is greater/larger than 1 in 3600!
Example 6: Lottery!

- You read the following article in a newspaper:
  - The Davidson family from East Texas have good reason to celebrate, they have won the lottery not once, but twice this year! The chance of winning it once is $10^{-5}$, but the chance of winning twice is extremely small $(10^{-5} \times 10^{-5} = 10^{-10})$. What a lucky family!

- Question: Was the probability calculated correctly?

- Answer: The probability was calculated as:

$$P(\text{First win and Second win}) = P(\text{First Win}) \cdot P(\text{Second Win} | \text{First Win})$$

$$= \frac{1}{10^5} \times \frac{1}{10^5} = \frac{1}{10^{10}}$$

Assuming there is no dependence between the first and second win of a lottery (which is suppose to be completely fair and favoring everyone equally) seems reasonable. So the calculation is correct. Note, this chance is so small, the family may be investigated for criminal activities!
Example 7: Marathons

- You read the following article in a newspaper:
  - The Smith family have good reason to celebrate. The Smith Siblings, Peter and Jane, are both amateur long distance runners and both won the New York Marathon this year! The chance of an amateur winning a major event is $10^{-3}$, so the chance that two in the family winning a major event is one in a million ($10^{-3} \times 10^{-3} = 10^{-6}$). What a lucky family!

- **Question:** Is the calculation correct?

- **Answer:**
  
  \[
P(\text{Peter win and Jane win}) = P(\text{Peter Win}) \cdot \frac{P(\text{Jane win} \mid \text{Peter Win})}{\text{chance of peter winning given that his sister has already won}}
  \]
  
  \[
  = \frac{1}{1000} \times P(\text{Jane Win} \mid \text{Peter Win})
  \]

- The last part of the calculation was based on the assumption that there is no dependence between Peter and Jane winning. Again genetic factors may favor athleticism in a family and it is likely that the chance of Peter winning given that his sister has won is greater than the chance of Peter winning without this additional information ie. $P(\text{Jane Win} \mid \text{Peter Win}) > \frac{1}{1000}$

- It is likely the chance is greater than that calculated.
Example 8: Wrong calculations

- There is a tendency that probabilities do not matter. However, it is important to understand how to calculate probabilities. Wrong calculations can have a terrible effect on peoples lives.
- The following is a real and rather tragic story of a probability that was not calculated properly.
- In 1996 Sally Clark had her first child, Christopher, who tragically died at less than 3 months old.
- In 1998 Sally Clark had a second child who also died 8 weeks after being born.
- After the death of her second child Sally Clark was arrested and charged with the murder of her two children.
- We will look at the most damning piece of evidence against her, which lead to her conviction, and how this piece of evidence was obtained.
What the prosecution said

- The most damning piece of evidence against Clark was given by the pediatrician Prof. Roy Meadows.

- He claimed that for an affluent family like the Clarks, the probability of a cot death (SIDS – sudden death syndrome) was 1 in 8543, therefore the probability of two cot deaths (this is $P(\text{death of first and second child})$)- so we need to use the multiplicative rule) is $1 \times 8543 = 1/(73\text{ million}) = 1.37 \times 10^{-8}$

- These very small odds, lead to her conviction.

- If we are to write her story as a hypothesis test it is $H_0$: Innocent against $H_A$: Guilty.

- We shall go through some of the problems with the calculation and interpretations with this probability.
  - The first major problem was with interpretation. The press (in their usual inability to understand what a probability meant, understood 1 in 73 million as the probability of Sally Clark being innocent). By now, we should understand it has the probability of observing the evidence given that she is innocent. Probability of innocence requires a more sophisticated calculation (and assumptions).
The second major problem was the calculation of the probability itself.

The chance 1 in 8453 that was used is controversial because the general figure is 1 in 3200. Roy Meadows reduced the odds based on Sally Clark’s affluence, but did not take into account that she had two boys, which increases the odds of a cot death.

The most controversial part of his testimony was the calculation of the probability, which gave these tiny odds and was the main reason for Sally Clark’s conviction.

In 2001, the Royal Statistical Society wrote a damning report criticizing this calculation. We look at this calculation and it’s fatal flaws.

By using the multiplicative rule we have

\[ P(\text{First baby dies and Second baby dies}) = P(\text{First baby dies})P(\text{Second baby dies|First baby dies}) \]

We want to calculate this probability under the assumption that they died from a Cot death. We know over the probability a child dies of a cot death (over the general population) is 1 in 8453, therefore using this

\[ P(\text{First baby dies and Second baby dies}) = (1/8453) \times P(\text{Second baby dies|First baby dies}). \]
In order to get the second probability, Roy Meadows assumed that the second probability was equal to its marginal

\[ P(\text{First and Second baby dies}) = P(\text{First dies}) \times \underbrace{P(\text{Second baby dies}|\text{First dies})}_{P(\text{Second baby dies})} \]

We recall that the conditional probability is only equal to its marginal if two variables are independent. This means the probability that the first baby dying is completely independent of the second child dying.

However, to make this rather strong assumption we are assuming that the first and second child share not factors in common. But in the case of Sally Clark the two children were brothers.

The assumption of statistical independence in this case seems to completely incredulous when applied to siblings!

The RSS wrote this in their report, it lead to a second appeal and in 2003 Sally Clark was released.

Sadly she never recovered from the experience and died in 2007.
Relationship between categorical variables

The **marginal distributions** summarize each categorical variable separately. But the **two-way table** actually provides information about the **relationship** between the categorical variables.

The cells of a two-way table represent the *simultaneous occurrence* of a given level of one categorical factor with a given level of the other categorical factor. We can ask whether certain combinations of the variables are generally more (or less) likely to occur together than other combinations.

In the smoking example, we can see that “student smokes” and “neither parent smokes” are relatively unlikely to occur together (even when accounting for the smaller sizes of these groups). More likely are the combinations that “student smokes” and “one or both parents smoke”. We would like to analyze such relationships statistically which we do in Chapter 14.
Association does not mean cause

- We have defined the idea of association between two variables, and in the next section we consider tests. But the associations we see does not necessarily mean that one variable cause the other.
- These are confounding variables, which we need to investigate in order to come up with causal relationships between two variables.
- We illustrate this in the next example.
- See IPS, page 519 for more information.
Simpson’s paradox

An association or comparison that holds for all of several groups can reverse direction when the data are combined (aggregated) to form a single group. This reversal is called **Simpson’s paradox**.

**Example:**

**Hospital death rates**

But when the patient condition is taken into account, we see that Hospital A admits more patients in a poor condition, which is what gives the appearance of a worse survival record.

<table>
<thead>
<tr>
<th></th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patients in good condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Died</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Survived</td>
<td>594</td>
<td>592</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>% surv.</td>
<td>99.0%</td>
<td>98.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patients in poor condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Died</td>
<td>57</td>
<td>8</td>
</tr>
<tr>
<td>Survived</td>
<td>1443</td>
<td>192</td>
</tr>
<tr>
<td>Total</td>
<td>1500</td>
<td>200</td>
</tr>
<tr>
<td>% surv.</td>
<td>96.2%</td>
<td>96.0%</td>
</tr>
</tbody>
</table>

On the surface (from the combined results), Hospital B would seem to have a slightly better record of survival.

Here, Patient Condition is called a “hidden” variable.
Accompanying problems associated with this Chapter

- Quiz 17
- Homework 9