Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 6 (MWF) Conditional probabilities and associations

Suhasini Subba Rao
Review of previous lecture

• Every random variable has a probability ‘distribution’ associated with it.

• We defined the idea of:

  (i) Mutually exclusive events: $P(A \text{ or } B) = P(A) + P(B)$. In other words, if one ‘event’ happened then the other could not. For example, if you gave birth to one child and it was a boy, then it could not be a girl too (assuming that gender can only be male or female).

  (ii) Conditional probabilities: $P(A|B)$. This is the probability of $A$ being evaluated given the piece of information $B$. For example, you want to evaluate the probability an individual has problems with their lungs, this may be $P(\text{lung problem}) = 0.1$. You then find out that individual smokes, this increases the chance of lung problems, $P(\text{lung problem}|\text{that person smokes}) = 0.3$. This means his smoking
status has an influence on his lung problems or there is a dependency between smoking and lung problems. For contingency tables it is straightforward to calculate conditional probabilities. In general, we use the formula
\[
P(A|B) = \frac{P(A \text{ and } B)}{P(B)}.
\]
(iii) Independent events: Two events $A$ and $B$ are independent if $P(A|B) = P(A)$. In other words, information on $B$ does not influence the event $A$.
Returning to the lung problem example, we know that smoking and lung problems are not independent events since $P(\text{lung problem}|\text{that person smokes}) = 0.3$ whereas $P(\text{lung problems}) = 0.1$. However, it could be that electronic cigarettes have no influence on whether a person has lung problems (I have no idea whether this is true or not), if this were the case knowing a person uses e-cigarettes does not change the chance of their having a lung problem - they are independent
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables events.

• In today’s lecture we use these ideas to do simple probability calculations.
Joint probabilities

- We calculate joint probabilities using marginal and conditional probabilities

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

- Which way you condition depends on what information is available.

- In general, \( P(A|B) \neq P(B|A) \).
Calculating joint probabilities

<table>
<thead>
<tr>
<th>Height</th>
<th>Gender</th>
<th>Height</th>
<th>Gender</th>
<th>Height</th>
<th>Gender</th>
<th>Height</th>
<th>Gender</th>
<th>Height</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5</td>
<td>F</td>
<td>5.9</td>
<td>M</td>
<td>4.9</td>
<td>F</td>
<td>6.2</td>
<td>M</td>
<td>6</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>6.3</td>
<td>M</td>
<td>5.6</td>
<td>F</td>
<td>5.9</td>
<td>M</td>
<td>5.8</td>
<td>F</td>
</tr>
<tr>
<td>5.9</td>
<td>F</td>
<td>5.2</td>
<td>F</td>
<td>5.7</td>
<td>M</td>
<td>5.9</td>
<td>M</td>
<td>6</td>
<td>M</td>
</tr>
<tr>
<td>5.2</td>
<td>F</td>
<td>5.6</td>
<td>F</td>
<td>5.5</td>
<td>M</td>
<td>5.5</td>
<td>F</td>
<td>5.5</td>
<td>F</td>
</tr>
</tbody>
</table>

- We touched on this earlier. But as a reminder $P(A \text{ and } B)$ is the probability that both the events $A$ and $B$ occur.

- To calculate

\[
P(\text{height less than 5.5 and male}) = \frac{\text{number of people less than 5.5 and male}}{\text{number of people}} = \frac{1}{18}
\]
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables

## Probabilities and contingency tables

<table>
<thead>
<tr>
<th></th>
<th>Stroke</th>
<th>No event</th>
<th>Subtotals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>treatment</strong></td>
<td>45</td>
<td>179</td>
<td>224</td>
</tr>
<tr>
<td><strong>control</strong></td>
<td>28</td>
<td>199</td>
<td>227</td>
</tr>
<tr>
<td><strong>Subtotals</strong></td>
<td>73</td>
<td>378</td>
<td>451</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Stroke</th>
<th>No event</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>treatment</strong></td>
<td>45/451=0.099</td>
<td>179/451=0.39</td>
<td>224/451=0.49</td>
</tr>
<tr>
<td><strong>control</strong></td>
<td>28/451=0.062</td>
<td>199/451=0.44</td>
<td>227/451=0.5</td>
</tr>
<tr>
<td><strong>Marginal</strong></td>
<td>73/451=0.16</td>
<td>378/451=0.83</td>
<td>451/451=1</td>
</tr>
</tbody>
</table>

- The edge of the table are the marginal probabilities.
- The center of the table are the joint probabilities.
- Observe that the sum of the joint probabilities is the marginal in each column/row.
• Therefore \( P(\text{stroke and treatment}) = P(\text{stroke}) - P(\text{stroke and control}) \).

• The conditional probability \( P(\text{stroke|treatment}) = 0.099/0.49 = 0.2 \).

• Similarly, we can easily calculate the joint probabilities using the conditionals since \( P(\text{stroke and treatment}) = P(\text{stroke|treatment}) \times P(\text{treatment}) = 0.2 \times 0.49 = 0.099 \).

• Things to observe

\[
P(\text{stroke|treatment}) = 1 - P(\text{no stroke|treatment})
\]

BUT \( P(\text{stroke|treatment}) \neq 1 - P(\text{stroke|no treatment}) \).

Example: Recall that \( P(\text{stroke|treatment}) = 0.2 \), whereas \( P(\text{stroke|no treatment}) = 0.12 \). It is clear that \( 0.2 \neq 1 - 0.12 = 0.82 \).
Independence and joint probabilities

• If $A$ and $B$ are independent events then

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = P(B).$$

Rearranging the above gives the identity

$$P(A \text{ and } B) = P(A)P(B)$$

(the same result holds for random variables). $P(A)$ and $P(B)$ are often called marginal probabilities.

• If $A$ and $B$ are independent, then calculation of the joint probability is straightforward.
Example: Fraternal Twins

- It is thought that the chance of having fraternal twins depends on several factors including ethnicity and diet (for example the chance of someone from the Yoruba’s - a group of people in South West Nigeria is as much as 100 out of 1000 live births).

- We are given the following information:
  - It is known that vegans have a fifth of the chance of non-vegans to have fraternal twins.
  - The number of fraternal twins born to non-vegans is 20 in 1000 live births (thus \( P(\text{fraternal}|\text{non-vegan}) = 0.02 \)). Thus based on the above piece of information, \( P(\text{fraternal}|\text{vegan}) = 0.004 \).
  - The proportion of vegans in this country is 2%.

- Based on this information, what is the probability a person (we have no
information on their diet) has fraternal twins?\(^1\)

\(^1\)Hint: split the chance of having fraternal twins into two categories, those who are non-vegan and have fraternal twins and those who are vegan and have fraternal twins.
Vegan: Solution 1

- The events of being vegan or not being vegan are mutually exclusive events. Therefore, the event of having fraternal twins can be split up into two mutually exclusive events. Those that have fraternal twins and are not vegans and those that have fraternal twins and being vegan. Therefore, using the additive property of mutually exclusive events we have

\[
P(\text{fraternal twins}) = P(\text{fraternal twins and vegan}) + P(\text{fraternal twins and not vegan}).
\]

- We use \( P(A \text{ and } B) = P(A|B)P(B) \), where we define the two events 
  \( A = \{\text{fraternal twin}\} \) and \( B = \{\text{vegan}\} \).
• We are given that \( P(\text{fraternal | vegan}) = 0.004 \) and that the chance of being vegan is \( P(\text{vegan}) = 0.02 \). Therefore using the above formula we have \( P(\text{fraternal twins and vegan}) = 0.0004 \times 0.02 = 0.000008 \).

• Using a similar argument we have \( P(\text{fraternal twins and not vegan}) = 0.02 \times 0.98 = 0.0196 \).

• Therefore

\[
P(\text{fraternal twins})
= P(\text{fraternal twins and vegan}) + P(\text{fraternal twins and not vegan})
= 0.000008 + 0.0196 = 0.019608.
\]
Vegan: Solution 2

<table>
<thead>
<tr>
<th></th>
<th>Vegan</th>
<th>Not Vegan</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraternal Not</td>
<td>$P(\text{Frat}</td>
<td>\text{Vegan})P(\text{Vegan})$</td>
<td>$P(\text{Frat}</td>
</tr>
<tr>
<td>Marginal</td>
<td>0.02</td>
<td>0.98</td>
<td>1</td>
</tr>
</tbody>
</table>

- The question gives the marginal and conditional probabilities i.e. $P(\text{fraternal}|\text{non-vegan}) = 0.02$ and $P(\text{fraternal}|\text{non-vegan}) = 0.004$.

- Observe that $P(\text{not having fraternal}|\text{non-vegan}) = (1-0.02)$

- $P(\text{not having fraternal}|\text{non-vegan}) = (1-0.004)$. 
Example: Hair color

Let $X$ be the colour of a women’s hair, it can be either blonde or dark. It is known that the probability of drawing a women with blonde hair is $0.35$ ($P(X = B) = 0.35$) and the probability of drawing a women with dark hair is $0.65$ ($P(X = D) = 0.65$). Let $Y$ indicate whether a women has skin cancer (it can take two values $Y = 1$ means the women has skin cancer and $Y = 0$ means the women does not have skin cancer). It is known that the probability a women has skin cancer given that she is blonde is $0.01$ ($P(Y = 1|X = B) = 0.01$) and the probability a women has skin cancer given that she is has dark hair is $0.005$ ($P(Y = 1|X = D) = 0.005$). Calculate

- $P(Y = 1 \text{ and } X = B)$ (probability women is blonde and has skin cancer).
- $P(Y = 1 \text{ and } X = D)$ (probability women has dark hair and has skin cancer).
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables.

- $P(Y = 1)$ (probability women has skin cancer with no information on hair colour).
Hair: Solution 1

To answer the question we use the formula $P(A \text{ and } B) = P(B|A)P(A)$.

- $P(Y = 1 \text{ and } X = B) = P(Y = 1|X = B)P(X = B) = 0.01 \times 0.35$.
- $P(Y = 1 \text{ and } X = D) = P(Y = 1|X = D)P(X = D) = 0.005 \times 0.65$.

- We observe that the events \{Y = 1 \text{ and } X = B\} and \{Y = 1 \text{ and } X = D\} are mutually exclusive event (a women cannot be both blonde and dark haired). Moreover, we can decompose

$$P(Y = 1) = P(\{Y = 1 \text{ and } X = B\}\text{ or }\{Y = 1 \text{ and } X = D\})$$

$$= P(\{Y = 1 \text{ and } X = B\}) + P(\{Y = 1 \text{ and } X = D\})$$

$$= 0.01 \times 0.35 + 0.005 \times 0.65.$$
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables

**Hair: Solution 2**

We can simply use a contingency table (with probabilities instead of numbers) to calculate the probabilities

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>No Cancer</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blonde</td>
<td></td>
<td></td>
<td>0.35</td>
</tr>
<tr>
<td>Dark</td>
<td></td>
<td></td>
<td>0.65</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

- We see from the table the proportion of blonde women is \( P(X = B) = 0.35 \) and \( P(X = D) = 0.65 \). Furthermore, \( P(Y = 1|X = B) = 0.01 \) and \( P(Y = 1|X = D) = 0.005 \).
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables

• Using this we have

\[ P(Y = 1 \text{ and } X = B) = 0.01 \times 0.35 = 0.0035 \]

\[ P(Y = 1 \text{ and } X = D) = 0.005 \times 0.65 = 0.00325. \]

• Finally, to calculate the probability that someone has skin cancer regardless of whether she is blonde or dark

\[ P(Y = 1) = 0.0035 + 0.00325 = 0.00675. \]
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables

**Hair: Solution through graphics**

\[
P(\text{Skin cancer} \mid \text{Blonde hair}) = \frac{\text{Red Area}}{\text{Blonde Area}} = 0.01
\]

\[
P(\text{Skin cancer} \mid \text{Dark hair}) = \frac{\text{Blue Area}}{\text{Dark Area}} = 0.005
\]

\[
P(\text{Skin cancer}) = \text{Red Area} + \text{Blue Area} = P(\text{Skin cancer} \mid \text{Blonde hair}) \times P(\text{Blonde hair}) + P(\text{Skin cancer} \mid \text{Dark hair}) \times P(\text{Dark hair})
\]
Tragedies that can arise when calculating probabilities incorrectly

• 10 years ago a solicitor called Sally Clark had two babies, unfortunately both those babies died before they were 3 months old.

• It was thought that the first baby had died of SID syndrome (Sudden Infant Death), after the second death it was also assumed to be SID too.

• But then suspicions were raised. Police thought that the odds of two SID deaths in a row were small. Sally Clark was put on trial.

• She was convicted and given a life sentence.
• The most damming piece of evidence against her was that the odds of two babies dying of SID syndrome was 5 in 10 million. This piece of evidence was given by a paediatrician called Roy Meadow.

• Roy Meadow calculated the probability as follows:
  
  – Let $X_i$ denote whether the $i$th baby dies of a cot death with $X_i = 1$ if it dies and $X_i = 0$ if it does not. It is generally believed that the probability of SID for an affluent mother (such as Sally Clark) is $P(X_i = 1) \approx 0.0007$ (about 7 in 10000).
  
  – We are interested in the probability that baby 1 and baby 2 both have SIDS. Formally we write this as $P(X_1 = 1 \text{ and } X_2 = 1)$.
  
  – In his evidence Roy Meadow used
    
    $P(X_1 = 1 \text{ and } X_2 = 1) = P(X_1 = 1) \times P(X_2 = 1)$. With $P(X_1 = 1) \times P(X_2 = 1) \approx 5/(10^7)$. 
Based on this argument, Roy Meadow said that the probability that two children dying of SID syndrome is so small that it is unlikely the children died naturally. This was the most damming piece of evidence against Sally Clark and lead to her conviction.

There is a fundamental problem with Roy Meadow’s derivation. This caught the notice of the Royal Statistical Society, and eventually lead to Sally Clark’s conviction being quashed. What is it?
The problem with Roy Meadow’s derivation

• Suppose that $X_1$ is the first baby in a family and $X_2$ is the second child in a family. Then $P(X_1 = 1 \text{ and } X_2 = 1) = P(X_1 = 1) \times P(X_2 = 1)$ is calculated on the assumption that $X_1$ and $X_2$ are independent random variables.

• This is quite an incredible assumption to make when the individuals concerned are brothers! It does not take into account any genetic abnormalities etc. which could easily arise.

• So this incredibly small probability was calculated on the assumption that the random variables were independent.
• We recall if they are not independent events then

\[
P(X_1 = 1 \text{ and } X_2 = 1) = P(X_2 = 1|X_1 = 1) \times P(X_1 = 1).
\]

It seems likely that \( P(X_1 = 1|X_2 = 1) \) is larger than the marginal \( P(X_1 = 1) \). Knowledge of a sibling SID is likely to increase the risk of subsequent siblings. Using the correct calculation would have increase the chance of two siblings SID.

• The Royal Statistics Society took the unprecedented step of writing to the Lord Chancellor to object to the way this probability had been calculated saying it was inaccurate.

• Sally Clark conviction was quashed based on this and another piece of evidence. Sadly she died in 2007 (http://en.wikipedia.org/wiki/Sally_Clark).
Distributions and probabilities

• We have looked at probabilities and how they should be calculated.

• Given a random variables we often want to calculate the probability of an event. To do this we will often assume that the random variable comes from a known distribution.

• Now we look at well known distributions which are used to describe the frequency/distribution of data.
Discrete random variables

• Until now we have focussed on the idea of distributions for continuous random variables, such as heights, weight etc. In all these examples the random variable can take any number in an interval, say any number in the interval $[2, 3]$.

• Recall that the other type of random variables are for categorical data (or discrete numbers), such a gender of a person, major of a person is taking, number of children a person may have.

In these cases, a random variable can only take discrete values. For example, if we mapped $\{\text{female, male}\} \rightarrow \{0, 1\}$ (so female $= 1$ and male $= 0$). If $X$ is the gender of a randomly chosen person, getting the outcome 0 and 1 makes sense but the outcome 1.5 makes no sense at all.
If $X$ is the number of children a person had then getting the outcome 0, 1, 2, 3 makes sense, but getting a number inbetween 2 and 3 (say 2.4) makes no sense.

- In all these cases the random variables are called discrete.

- For discrete random variables we no longer use the density function (the area under the graph), we use, instead discrete distributions.

- This where we allocate to each discrete outcome (say male/female, number of children etc.) a probability.
Number of children

Suppose 30 people are questioned about the number of children they have:

<table>
<thead>
<tr>
<th>No. of kids</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>22</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>probability</td>
<td>0.7</td>
<td>0.2</td>
<td>0</td>
<td>0.04</td>
<td>0</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Suppose $X$ is number of children of a randomly chosen person in this group. Below is the discrete distribution of $X$. 
Useful distributions

There are many distributions which are of interest to us. We will consider in this course just a few of them.

• For discrete random variables.
  – The Binomial distribution.

• For continuous random variables,
  – The normal (Gaussian) distribution.
  – The $\chi^2$-distribution.
  – The $t$-distribution.
  – The $F$-distribution.