Data Analysis and Statistical Methods
Statistics 651

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

Lecture 6 (MWF) Conditional probabilities and looking for associations

Suhasini Subba Rao
Review of previous lecture

- Every random variable has a probability ‘distribution’ associated with to 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.

(iii) Independent events: Two events $A$ and $B$ are independent if $P(A|B) = P(A)$ (later we will show that this is the same as saying $P(A \text{ and } B) = P(A)P(B)$). In other words, knowledge of $B$ does not exert an influence on 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 events.

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

Remember that mutually exclusive and independence are totally different. In fact if two events are mutually exclusive they cannot be independent.

- **Why mutually exclusive event are not independent**

Let $X$ be the gender of a randomly selected person. There are two possible outcomes $M$ or $F$. Let $A = \{F\}$ (the event that the randomly selected person is female) and $B = \{M\}$ (the event that the randomly selected person is male). Clearly $A$ and $B$ are mutually exclusive, which means if one event occurs the other cannot (eg. if $X = F$, then $X \neq M$ and visa versa).

It is clear that $P(A|B) = P(X = F|X = M) = 0$. However $P(X = F) = 1/2$. Therefore since $P(X = F|X = M) \neq P(X = F)$, hence $A$ and $B$ are not independent!
Independent events

Independent events are completely different. If $A$ and $B$ are independent, then event $A$ has no influence on event $B$. For example, the age of a randomly selected adult has no influence on their height.
Example: Contingency tables and evaluating probabilities

Psychologists wanted to investigate whether there was dependence between height and how bossy someone was (aka Do short men have a Napoleon complex). Often we see data tabulated as follows.

<table>
<thead>
<tr>
<th></th>
<th>short</th>
<th>medium</th>
<th>large</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>bossy</td>
<td>90</td>
<td>155</td>
<td>55</td>
<td>300</td>
</tr>
<tr>
<td>not bossy</td>
<td>110</td>
<td>445</td>
<td>145</td>
<td>700</td>
</tr>
<tr>
<td>Totals</td>
<td>200</td>
<td>600</td>
<td>200</td>
<td>1000</td>
</tr>
</tbody>
</table>

Calculate from the data:

- The probability \( P(\text{bossy}) \), the probability \( P(\text{short}) \), the probability \( P(\text{bossy and short}) \) (this is known as a joint probability that is the probability of two events happening), the probability \( P(\text{bossy | short}) \).
Solution

- \( P(\text{bossy}) = \frac{300}{1000}, \ P(\text{short}) = \frac{200}{1000}. \)

- \( P(\text{bossy and short}) = \frac{90}{1000}. \)

- There are two (equivalent) ways to calculate \( P(\text{bossy} \mid \text{short}) \).

  This first way is to focus on the subpopulation of Short guys and calculate the probability of being bossy when it is known you are a short guy, this is \( P(\text{bossy} \mid \text{short}) = \frac{90}{200}. \)

  The second way is to use the formula \( P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}. \) Using this we have \( P(\text{bossy} \mid \text{short}) = \frac{P(\text{bossy and short})}{P(\text{short})} = \frac{90}{1000} \times \frac{1000}{200} = \frac{90}{200}. \)

  Both methods will always lead to the same probability.
Calculating conditional from joint probabilities

As shown on the previous slide, the conditional probability can be written in terms of the joint probability - $P(A|B) = P(A \text{ and } B)/P(B)$. We illustrate this using the example from the previous lecture:

<table>
<thead>
<tr>
<th>Height</th>
<th>Gender</th>
<th></th>
<th>Height</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5</td>
<td>F</td>
<td></td>
<td>5.9</td>
<td>M</td>
</tr>
<tr>
<td>5.9</td>
<td>M</td>
<td></td>
<td>4.9</td>
<td>F</td>
</tr>
<tr>
<td>6.2</td>
<td>M</td>
<td></td>
<td>6</td>
<td>M</td>
</tr>
<tr>
<td>5.9</td>
<td>F</td>
<td></td>
<td>5.2</td>
<td>F</td>
</tr>
<tr>
<td>5.7</td>
<td>M</td>
<td></td>
<td>5.3</td>
<td>F</td>
</tr>
</tbody>
</table>

- We will calculate $P(\text{male})$, $P(\text{height less than 5.5 and male})$ and $P(\text{height less than 5.5} | \text{male})$.

- Recall how $P(B)$ is calculated:

\[
P(B) = \frac{\text{number of occurrences of event B}}{\text{total number of occurrences}}
\]
Example:

\[ P(\text{male}) = \frac{\text{number of males}}{\text{number of people}} = \frac{9}{18} \]

\[ P(\text{height less than 5.5 and male}) = \frac{\text{number of people less than 5.5 and male}}{\text{number of people}} = \frac{6}{18} \]

- Recall how the probability \( P(A|B) \) is calculated.

\[ P(A|B) = \frac{\text{number of occurrences of event A and B}}{\text{number of occurrences of B}} \]

Example:

\[ P(\text{height less than 5.5}|M) = \frac{\text{number of males who are less than 5.5}}{\text{number of males}} = \frac{1}{9}. \]
Hence we identified all males (there were 9) and in this subgroup counted all males less than 5.5.

• We can see from the above example that

\[
P(\text{height less than 5.5}|M) = \frac{\text{number of males who are less than 5.5}}{\text{number of males}}
\]

\[
= \frac{P(\text{height less than 5.5 and male})}{P(\text{male})} = \frac{1/18}{9/18} = \frac{1}{9}.
\]

• To summarise, what the above is saying is that

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

Rearranging the above gives us \( P(A \text{ and } B) = P(A|B)P(B) \).
• **Application** If \( A \) and \( B \) are **independent events** then \( P(A \text{ and } B) = P(A)P(B) \).

However, in the example given above height and gender are not independent.

In general \( P(A \text{ and } B) \) can be difficult to evaluate. But since \( P(A \text{ and } B) = P(A|B)P(B) \), it can be evaluated if \( P(A|B) \) and \( P(B) \) are known.

We note that if \( A \) and \( B \) are independent then \( P(A \text{ and } B) = P(A|B)P(B) = P(A)P(B) \) (since \( B \) does not exert any influence on \( A \)).
**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).

- Example: Here we want to calculate the chance of getting fraternal twins.
  - 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%.
What is the probability someone (regardless of them being vegan or not) has fraternal twins?

• 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.
Solution

• 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

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

• On the previous slides we gave the formula \( P(A \text{ and } B) = P(A|B)P(B) \), which we know use. We define the two events \( A = \{\text{fraternal twin}\} \) and \( B = \{\text{vegan}\} \). We are given that
P(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

$$P(\text{fraternal twins}) = P(\text{fraternal twins and vegan}) + P(\text{fraternal twins and not vegan}) = 0.000008 + 0.0196 = 0.019608.$$
Example

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

- $P(Y = 1$ and $X = B)$ (probability women is blonde and has skin cancer).
- $P(Y = 1$ 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 cancer).

• (a little harder) $P(Y = 1)$ (probability women has skin cancer, regardless of hair colour).
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables

**Solution using probabilities**

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

\[
P(Y = 1) = P(\{Y = 1 \text{ and } X = B\} \text{or} \{Y = 1 \text{ and } X = D\})
= P(\{Y = 1 \text{ and } X = B\}) + P(\{Y = 1 \text{ and } X = D\})
= 0.01 \times 0.35 + 0.005 \times 0.65.
\]
Solution using data

We can look at the above example through the lens of data. Suppose 10K women were sampled. Using the above probabilities, the numbers would look like (on average!):

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>No Cancer</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blonde</td>
<td>35</td>
<td>3465</td>
<td>3,500</td>
</tr>
<tr>
<td>Dark</td>
<td>32.5(!)</td>
<td>6467.5</td>
<td>6,500</td>
</tr>
<tr>
<td>Totals</td>
<td>67.5</td>
<td>9932.5</td>
<td>10,000</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, $P(Y = 1|X = B) = 35/3500 = 0.01$ and $P(Y = 1|X = D) = 32.5/6500 = 0.005$. 
• Further we see that

\[ P(Y = 1 \text{ and } X = B) = \frac{35}{10,000} = \frac{35}{3,500} \times \frac{3,500}{10,000} = 0.0035 \]

and

\[ P(Y = 1 \text{ and } X = D) = \frac{32.5}{10,000} = \frac{32.5}{6,500} \times \frac{6,500}{10,000} = 0.00325. \]

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

\[ P(Y = 1) = \frac{67.5}{10,000} = \frac{35}{10,000} + \frac{32.5}{10,000}. \]
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables

\[ P(\text{Skin Cancer} | \text{Blonde Hair}) = \frac{\text{Red Area}}{\text{Blonde Area}} = 0.01 \]

\[ P(\text{Skin Cancer} | \text{Dark Hair}) = \frac{\text{Blue Area}}{\text{Dark Area}} = 0.005 \]

\[ P(\text{Skin Cancer}) = \frac{\text{Red Area} + \text{Blue Area}}{\text{Total Area}} = 0.005 \]

\[ = P(\text{Skin Cancer} | \text{Blonde Hair}) \times P(\text{Blonde Hair}) + P(\text{Skin Cancer} | \text{Dark Hair}) \times P(\text{Dark Hair}) \]
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 a cot death, after the second death it was also assumed to be a cot death.

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

• She was convicted and give a life sentence.

• The most damming piece of evidence against her was that the odds of two babies dying of a cot death 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 cot death for an affluent person (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 die from a cot death. Formally we write this as $P(X_1 = 1 \text{ and } X_2 = 1)$.

- Roy Meadow supposed $P(X_1 = 1 \text{ and } X_2 = 1) = P(X_1 = 1) \times P(X_2 = 1)$. Now $P(X_1 = 1) \times P(X_2 = 1) \approx 5/(10^7)$.

- Based on this argument, Roy Meadow said that the probability that two children dying of a cot death 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...
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables
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, if the dependence was taken into account (and this hard to do) it is likely to be a lot smaller.

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 $P(X_1 = 1)$
Lecture 6 (MWF) Evaluating conditional probabilities, and checking for associations (dependence) between variables

\[ P(X_2 = 1 \mid X_1 = 1) \] (that is the probability the second child dies given the first child dies) is far larger than \( \frac{5}{10 \times 10^6} \).

- The Royal Statistical 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 \{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 in between 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.