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

(i) Mutually exclusive events: If one event happens it excludes the possibility of the other. If A and B are mutually exclusive then $P(A \text{ or } B) = P(A) + P(B)$. For example, if someone gave birth to one child and it was a boy, then it could not be a girl too (assuming that birth gender can only be male **or** female).

(ii) Conditional probabilities: $P(A|B)$. This is the probability of the event A given the additional piece of information B. For example, you want to evaluate the probability an individual has problems with their lungs: P(lung problem) $= 0.1$. You then find out that individual smokes, this increases the chance of lung problems, P(lung problem|that person smokes) $= 0.3$. This means their smoking status has an influence on his lung problems or there is a $dependency$ between smoking and lung

problems. In general,

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

(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(lung problem|that person smokes) $=$ 0.3 whereas P(lung problems) $=$ 0.1. Thus knowledge that someone smokes changes has an influence on the lung problem. These two events are <u>not</u> independent.

Joint probabilities

- We calculate joint probabilities using marginal and conditional probabilities

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

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

- In general, $P\left(A|B\right) \neq P\left(B|A\right)$.

Independence and joint probabilities

- If A and B are **independent events** then

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

  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.

Probabilities and contingency tables

	Stroke	No event	
treatment	45	179	224
control	28	199	227
Subtotals	73	378	451

We divide by 451 to turn the numbers into probabilities.

	Stroke	No event	Marginal
treatment	45/451=0.099	179/451=0.39	224/451=0.49
control	28/451=0.062	199/451 =0.44	227/451 =0.5
Marginal	73/451=0.16	378/451=0.83	451/451=1

Features in the table

- The outer edge of the table gives the marginal probabilities.
- The center of the table give the joint probabilities between events.
- Observe that the sum of the joint probabilities is the marginal in each column/row.
- Therefore P(stroke and treatment) $=$ P(stroke) - P(stroke and control).
- The conditional probability $P(\text{stroke}|\text{treatment}) = 0.099/0.49 = 0.2$.
- Similarly, we can easily calculate the joint probabilities using the conditionals since P(stroke and treatment) $=$ $P(\text{stroke}|\text{treatment}) \times P(\text{treatment}) = 0.2 \times 0.49 = 0.099$.

- The rules on conditioning a random variable are same as the rules on the marginal. We know that because A and not A are disjoint events, then

$$P(A) + P(\text{not } A) = 1 \text{ therefore } P(A) = 1 - P(\text{not } A).$$

Example: Using this we have that

$$P\left(\text{stroke}\right) \quad = \quad 1 - P\left(\text{no stroke}\right).$$

Similarly, $A|B$ and not $A|B$ are disjoint events therefore

$$P(A|B) + P(\text{not } A|B) = 1 \text{ therefore } P(A|B) = 1 - P(\text{not } A|B).$$

Example: Using this we have that

$$P\left(\text{stroke}|\text{treatment}\right) \quad = \quad 1 - P\left(\text{no stroke}|\text{treatment}\right)$$

However, be very careful

$$P\left(\text{stroke}|\text{treatment}\right) \quad \neq \quad 1 - P\left(\text{stroke}|\text{no treatment}\right).$$

Calculation (to show the above) $P\left(\text{stroke}|\text{treatment}\right) = 0.2$, whereas $P\left(\text{stroke}|\text{no treatment}\right) = 0.12$. It is clear that $0.2 \neq 1 - 0.12 = 0.82$.

- A common mistake is to claim $P\left(\text{stroke}|\text{treatment}\right) = 1 - P\left(\text{stroke}|\text{no treatment}\right)$

  (or in general $P(A|B) = 1 - P(A|B^c)$). Which is clearly <u>wrong</u>.

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(fraternal|non-vegan) = 0.02. Thus based on the above piece of information, P(fraternal|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 catergories, those who are non-vegan and have fraternal twins and those who are vegan and have fraternal twins.

Vegan: Solution

	Vegan	Not Vegan	Marginal
Fraternal Not	P(Frat|V)P(V)	P(Frat|Not V)P(Not V)	
Marginal	0.02	0.98	1

- The question gives the marginal and conditional probabilities i.e. P(fraternal|non-vegan) = 0.02 and P(fraternal|non-vegan) = 0.004.
- Observe that P(not having fraternal|non-vegan) = (1-0.02)
- P(not having fraternal|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).

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

Hair: Solution

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

	Cancer	No Cancer	large
Blonde			0.35
Dark			0.65
Totals			1

- 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 $P(Y = 1|X = D) = 0.005$.

- Using this we have

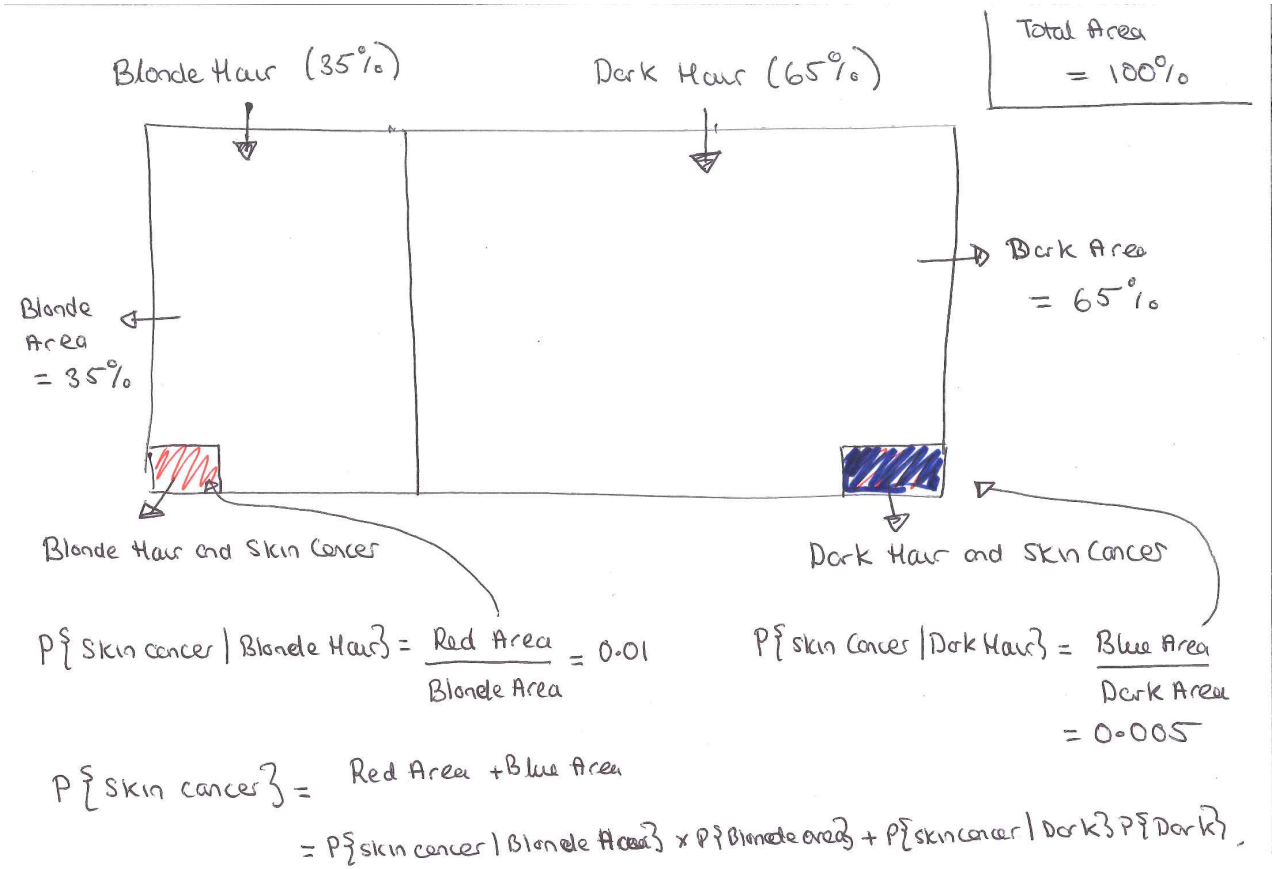
$$P(Y = 1 \text{ and } X = B) = 0.01 \times 0.35 = 0.0035$$

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

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

$$P(Y = 1) = 0.0035 + 0.00325 = 0.00675.$$

Hair: Solution through graphics



Tragedies that can arise when calculating probabilities incorrectly

- 20 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 ith 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
$P(X_1 = 1 \text{ and } X_2 = 1) = P(X_1 = 1) \times P(X_2 = 1) \approx 5/(10^7)$.

- 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 fundemental problem with Roy Meadow's derivation. This caught the notice of the Royal Statisical 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 Statisical 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`).

Additional comments

- It is very important to understand that even if $5/(10^7)$ was the correct probability of both of her children dying of SIDS, this probability gives no information what so ever on the probability of her children dying in a more sinster fashion. It is tempting but completely **wrong** to say

$$P(\text{the deaths were sinster}) = 1 - 5/(10^7).$$