Data Analysis and Statistical Methods
Statistics 651

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

https://www.openintro.org/stat/textbook.php?stat_book=os **(Chapter 2)**

Lecture 5 (MWF) The rules of probability

Suhasini Subba Rao

Probability

- Anything which varies within a set of variables due to chance is called a random variable. The random variable is often denoted as X. The notation Ω is called the "sample space" and is the set of all possible outcomes of X.

- A random variable can be discrete i.e. taking any one of the values in the set

$$\Omega = \{a, b, c, d, \ldots\}$$

- Or it can be continuous taking any value in the interval

$$[u_1, u_2] \qquad u_1 \text{ and } u_2 \text{ are numbers.}$$

- We associate a probability to all the possible outcomes.

- If it is discrete, then we write

  $$P(X = a)$$

  to denote the probability of the random variable taking the outcome a.

- If it is continuous, then we write

  $$P(a < X \leq b)$$

  to denote the probability of the random variable taking any number between a and b.

- A probability must lie between zero and one.

Die Toss example

- The throw of a six-sided die gives rise to the possible outcomes $\{1, 2, 3, 4, 5, 6\}$. Observe that the numbering does not have any real ordering (unless we choose it to), in which case X is a categorical random variable.

- If the the die is completely fair then

$$\begin{aligned} P(X = 1) = 1/6, \quad P(X = 2) = 1/6, \quad P(X = 3) = 1/6, \\ P(X = 4) = 1/6, \quad P(X = 5) = 1/6, \quad P(X = 6) = 1/6. \end{aligned}$$

A frequentist method for calculating a probability

- The chance $1/6$ was based on our understanding of a physical system. But it can also be calcuated by repeating an experiment.

- We throw a die n times and count the number of times one arises, we note this as n_1.

- The ratio

  $$\frac{n_1}{n},$$

  "gets" 'closer and closer" (in some sense) to the true chance as n gets larger and larger. In the case the die is fair, this chance is $1/6$.

Proportions and probability

- The proportion of human heights that lie in $[5.5, 5.75]$ is:

$$\frac{\text{Total no. in population with height in the interval } [5.5, 5.75]}{\text{Total no. in population}}.$$

- This proportion can be calculated using the frequentist approach. For each "experiment" we random select a person. Let X_i be one if that person is between $5.5 - 5.75$ and zero otherwise. Suppose n people were sampled. The ratio

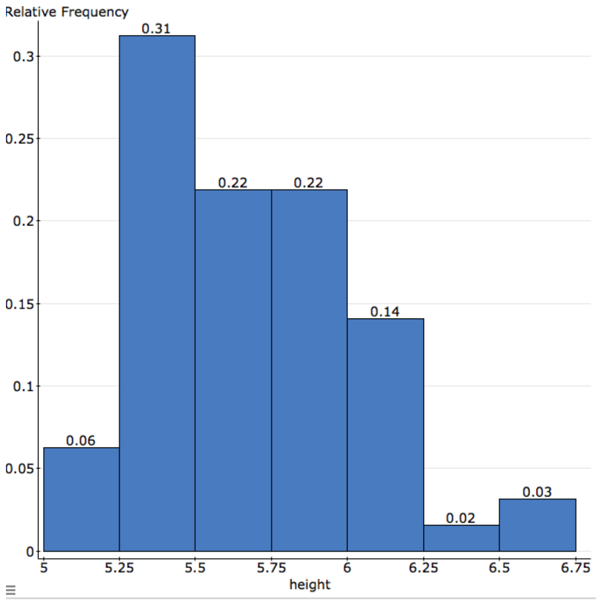
$$\frac{X_1 + \ldots + X_n}{n} = \frac{\text{no. in sample whose height is between } 5.5 - 5.75}{n}$$

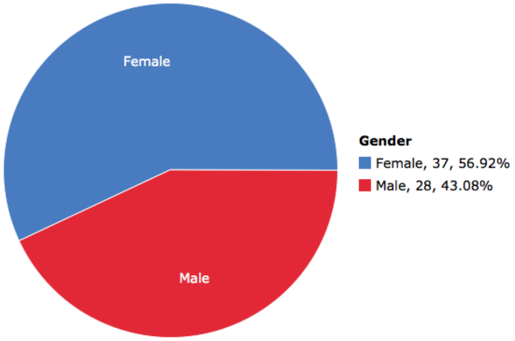
gets 'closer and closer" (in some sense) to the true proportion as the number of people sampled grow.

Random variables in a 651 class

- Let Y denote the gender of person in the 651 class (for simplicity assume it is binary). Then it takes the value $\{$Male or Female$\}$.

- X is the height of a student in the class.





Calculating probabilities from the data

(1) The set of all possible outcomes lies in the interval $[5,7]$.

Therefore, $P(5 \leq X \leq 7) = 1$.

- $P(5.75 \leq X < 6.25) = P(5.75 \leq X < 6) + P(6 \leq X < 6.25) = 0.22 + 0.14$.

- It is impossible from this plot to evaluate $P(5.9 \leq X < 6.15)$.

(ii) P(Y = Male) = 0.4308 and P(Y = Female) = 1-0.4308,

Example 2: Stents

An experimental study to see if stents reduce the risk of stroke.

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

- $P(\text{Stroke}) = 73/451 = 0.161.$

- $P(\text{No stroke}) = 378/451 = 0.838.$

- These are called marginal probabilities because they do not take into account the treatment given.

- Definition The marginal probability gives the probability of variable without reference to the values of another variable. Put simply, it gives the probability of an event without additional information.

Mutually exclusive events

- Two events (outcomes) are mutually exclusive, if the occurrence of one event excludes the occurrence of the other event.

- Example 1 Suppose X is the birth gender of a person and Y whether they give birth. The events X=male and Y=give birth are mutually exclusive. Since if the birth gender of a person is male then they cannot give birth.

- Example 2 Define the events $A = [5, 5.5)$ and $B = [5.75, 6)$.

  This means if person's height is in $A = [5, 5.5)$, then their height cannot be in $B = [5.75, 6]$. Similarly, if X is in $B = [5.75, 6)$ then it cannot be in $A = [5, 5.5)$.

  $A = [5, 5.5)$ and $B = [5.75, 6)$ are mutually exclusive events.

- On the other hand, the events $A = [5, 5.5)$ and $C = [5.25, 5.75)$ are not mutually exclusive. If $X = 5.3$, then it is in both A and C.

Mutually exclusive events and probabilities

- **Calculating probabilities** If two events A and B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$ (the probability of either A or B arising).

- Height Example Suppose X is the height of a randomly chosen person.

  If $A = [5, 5.5)$ and $B = [5.75, 6)$, then

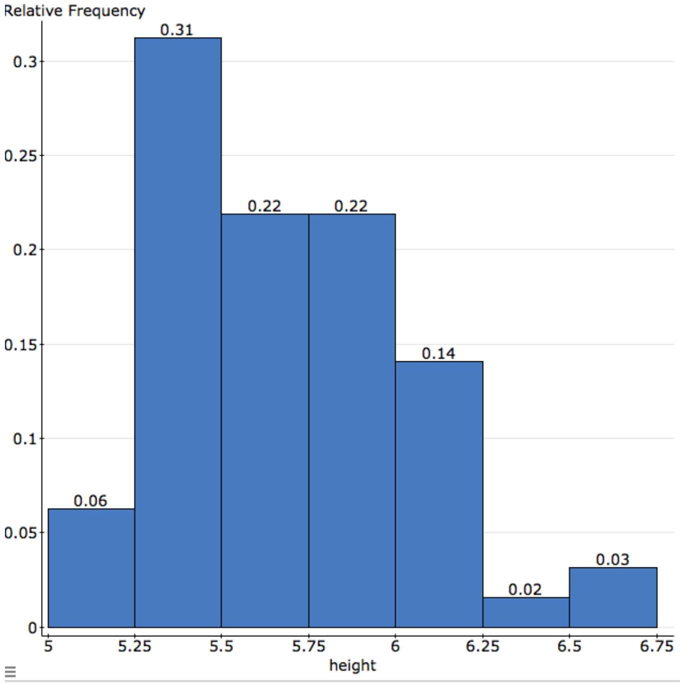
  $P(X \text{ in } A \text{ or } X \text{ in } B) = P(5 \leq X < 5.5 \text{ or } 5.75 \leq X < 6)$.

  Since A and B are mutually exclusive then

  $P(X \text{ in } A \text{ or } X \text{ in } B) = P(5 \leq X < 5.5 \text{ or } 5.75 \leq X < 6)$

  $= P(5 \leq X < 5.5) + P(5.75 \leq X < 6)$.

- Mutually exclusive is very natural and we have already used it previously in calculating probabilities. Look at the plot below, each bin is mutually exclusive of the others.



Applying mutually exclusive: Example 2

- Let X denote the height of a person. Suppose that the probability $P(X \leq t)$ is known for all t. We know that students heights lie within $5-7$ feet. Using this information, how can we calculate the probabilities below using $P(X \leq t)$ for any t:

(i) $P(5.5 \leq X < 6.5)$.

(ii) $P(X > 6)$.

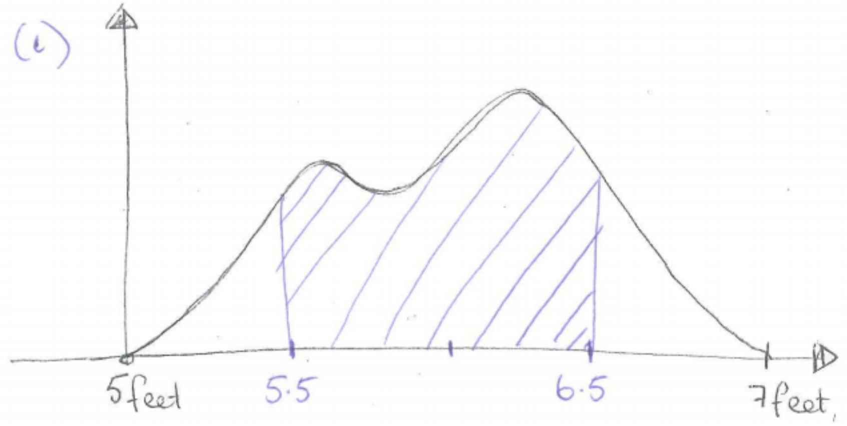
(iii) $P(5 \leq X < 6 \text{ or } 6.5 \leq X < 7)$.

(iv) $P(5 \leq X \leq 6 \text{ or } 5.75 \leq X < 7)$.

(v) $P(5 \leq X < 6 \text{ and } 5.75 \leq X < 7)$.

Solution 2

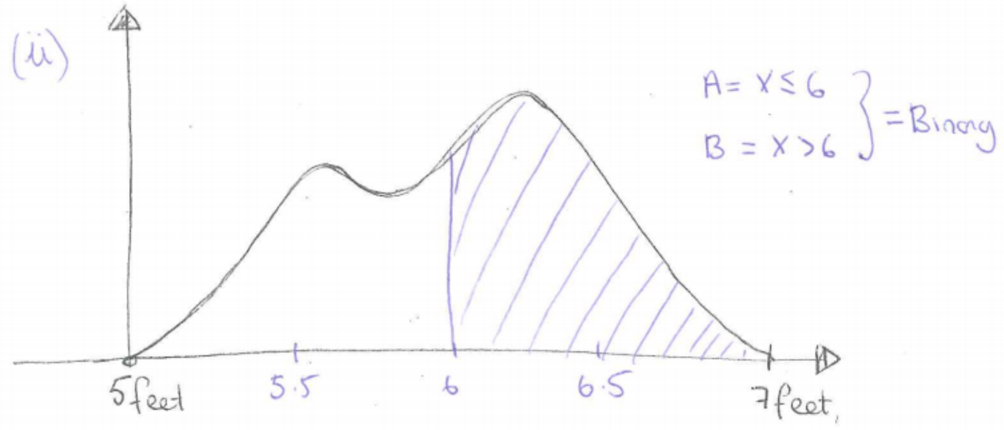
(i) $P(5.5 < X \leq 6.5)$.



$$P(X \leq 6.5) = P(X \leq 5.5) + P(5.5 < X \leq 6.5)$$

$$\Rightarrow P(5.5 < X \leq 6.5) = P(X \leq 6.5) - P(X \leq 5.5)$$

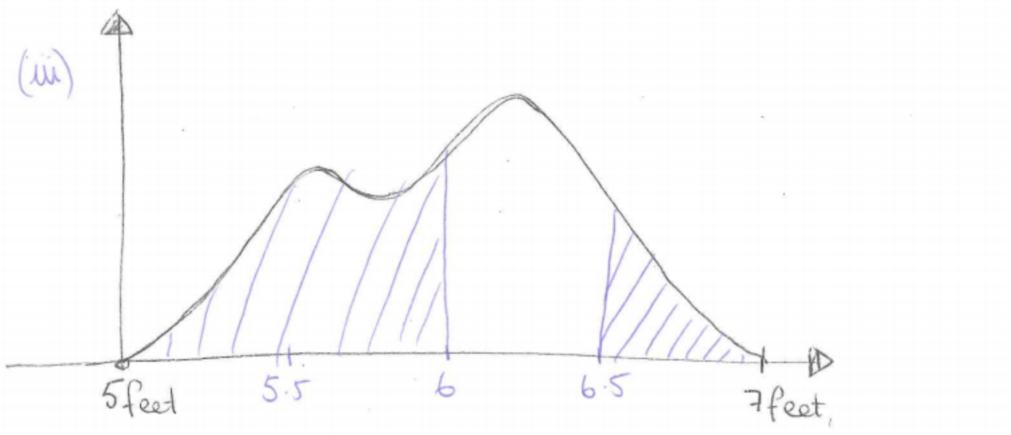
(ii) $P(X > 6) = 1 - P(X \leq 6)$.



$P(X \leq 6 \text{ or } X > 6) = P(X \leq 6) + P(X > 6) = 1$

$P(X > 6) = 1 - P(X \leq 6)$

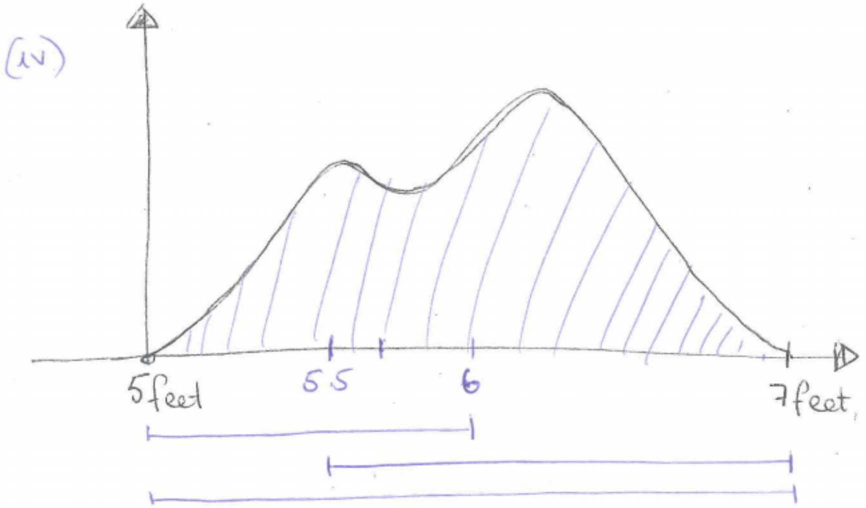
(iii) $P(5 < X \leq 6 \text{ or } 6.5 < X \leq 7)$.



$P(5< X\leq 6 \ \underline{\text{or}}\ 6.5< X \leq 7) = P(5< X \leq 6) + P(6.5< X \leq 7)$

$= \left[P(X \leq 7) - P(X \leq 6.5)\right] + \left[P(X \leq 6) - P(X \leq 5)\right]$.

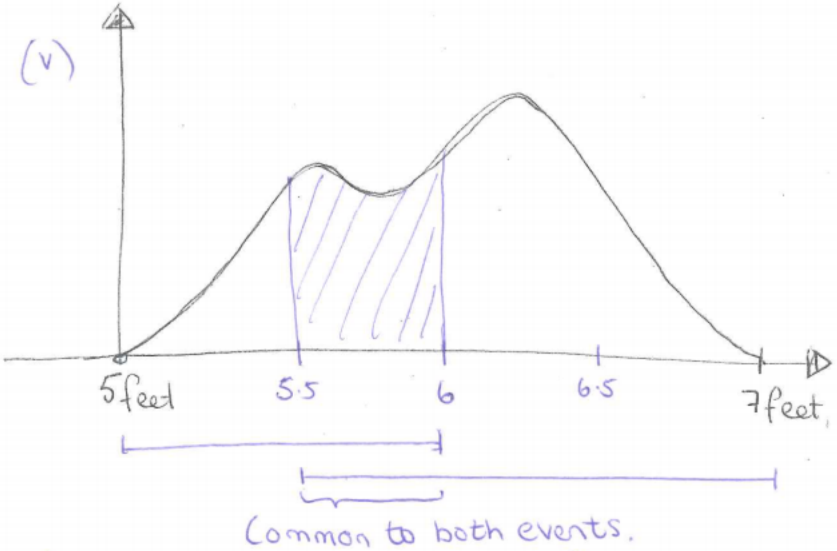
(iv) $P(5 < X \leq 6 \text{ or } 5.5 < X \leq 7)$.



$P(5 < X \leq 6 \ \underline{\underline{\text{or}}}\ 5.5 < X \leq 7) = P(5 < X \leq 7) = 1$

(no heights less than 5 feet or more than 7 feet).

(v) $P(5 < X \leq 6 \text{ and } 5.5 < X \leq 7)$.



$$P(5 < X \leq 6 \ \underline{\text{and}}\ 5.5 < X \leq 7) = P(5.5 < X \leq 6)$$

$$= P(X \leq 6) - P(X \leq 5.5)$$

Conditional probabilities: Stents

Is there a relationship between the occurence of strokes in a patient and whether they received treatment with stents?

In order to analyze this relationship, we need to introduce the notion of a *conditional probability*. This is a probability calculated using only subpopulation of a population.

- A conditional probability is the probability of an event given that we already have some (possibly partial) information about.
- In the example about it is the probability of the occurence of a stroke, given that the patient has received treatment for the stroke (or amongst patients who receive stents treatment). We denote this probability as

  $$P(\text{stroke}|\text{stents treatment}).$$

- In order to determine if stents have an effect it must be compared with the proportion of patients who do not receive stents treatment:

$$P(\text{stroke}|\text{stents treatment}).$$

- We must compare the two probabilities $P(\text{stroke}|\text{stents treatment})$ and $P(\text{stroke}|\text{stents treatment})$.

Analysis based on the experimental data

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

- $P(\text{stroke}|\text{stent}) = 45/224 = 0.2$
- $P(\text{stroke}|\text{control}) = 28/227 = 0.12$
- What does the difference suggest. Why should we be careful about drawing conclusions about the population of at risk stroke patients based on just on the difference of 0.2 and 0.12 in this sample?

Conditional probability: Heights and gender of 18 people

Height	5.5	5.9	4.9	6.2	6	5.9	5.2	5.7	5.3
Gender	F	M	F	M	M	F	F	M	F

Height	5	6.3	5.6	5.9	5.8	5.9	6	5.6	5.5
Gender	M	M	F	M	F	M	M	F	F

- We randomly select a person. Let X denote their height and Y their gender. Calculate

(i) The marginal: $P(X \leq 5.5)$.
(ii) The conditional: $P(X \leq 5.5|Y = M)$.
(iii) The conditional: $P(X \leq 5.5|Y = F)$.

- What do we observe? Is there a difference in the distribution of male and female heights.

Conditional probability

- Calculating conditional probabilities for categorical data is straightforward.

- However, sometimes the variable on which we condition is not always categorical. In this case we use the formula:

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

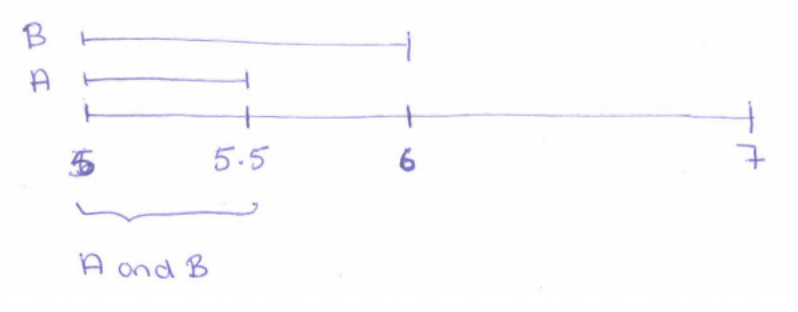
- Keep in mind that $P(A \text{ and } B)$ means that an observation has to be in *both* sets A and B.

- Whereas $P(A \text{ or } B)$ means that an observation can be in either one of the sets.

- Example 1 (height) Suppose X is the height of a randomly selected person and we know that X lies in the interval $[5,6]$, then what is the probability X lies in $[5,5.5]$? We write this probability as:

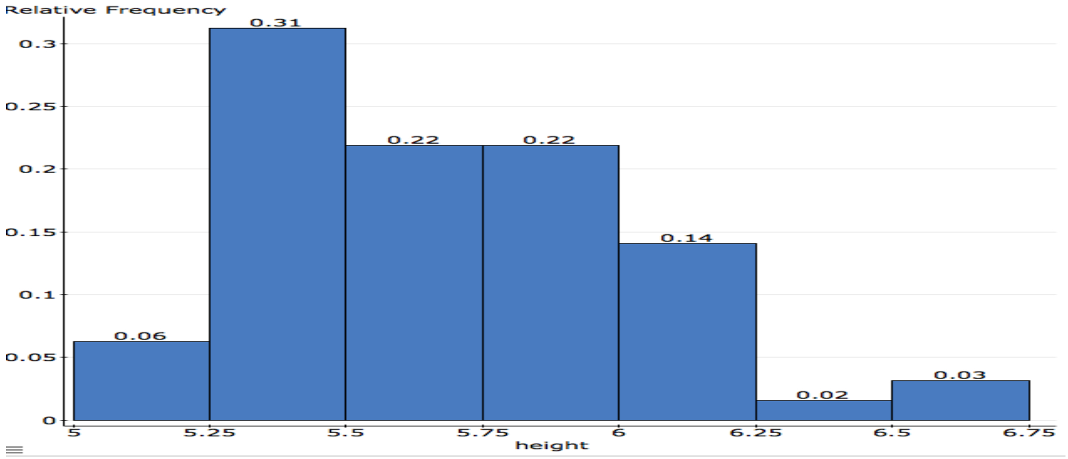
$$\frac{\text{No. whose height is in } [5,5.5]}{\text{No. whose height is in } [5,6]} = P(5 \leq X \leq 5.5 | 5 \leq X \leq 6)$$

- **Answer**: Using the formula we calculate the probability. Let $A = 5 \leq X \leq 5.5$ and $B = 5 \leq X \leq 6$. Using



We have

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



$$P(5 \leq X \leq 5.5|5 \leq X \leq 6) = \frac{0.06 + 0.31}{0.06 + 0.31 + 0.22 + 0.22} = \frac{0.37}{0.81} = 0.42.$$

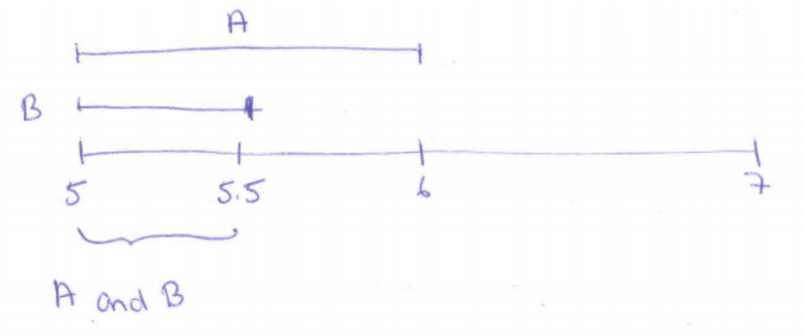
- We see that $P(5 \leq X \leq 5.5|5 \leq X \leq 6) > P(5 \leq X \leq 5.5)$.

  In this example, additional information about the height *increased* the chance.

- Additional information will *not* always increase the chance. For example, $P(X > 6|5 \leq X \leq 6) = 0 < P(5 \leq X \leq 5.5)$.

- Example 2 (height) Calculate the the probability X is between 5 and 6 feet given that we know that person's height is between 5-5.5 feet ($P(5 < X \leq 6|5 < X \leq 5.5)$)?

- **Answer** Using the formula we calculate the probability. Let $A = 5 \leq X \leq 6$ and $B = 5 \leq X \leq 5.5$. Using



We have

$$\begin{aligned} & P(5 < X \leq 6|5 < X \leq 5.5) \\ = & P\left(A|B\right) \\ = & \frac{P(B)}{P(B)} = 1. \end{aligned}$$

Independence and conditional probabilities

- **Definition** Suppose that we have two events A and B. The events A and B are independent of each other if $P(A|B) = P(A)$. This means the event B has no influence what so ever on the chance of event A occurring.

- Let us return to the example of heights and gender:

- Are height and gender independent variables? In other words, does information about the gender of a person give additional information about their height.

  Intuitively this seems to be the case. We also have empirical evidence:

- We showed in a previous example: $P(5 \leq X \leq 5.5|Y = \text{ female}) > P(5 \leq X \leq 5.5|Y = \text{ male})$.

Which means that $P(5 \leq X \leq 5.5|Y = \text{ female}) \neq P(5 \leq X \leq 5.5)$ This implies that gender and height are not independent variables. There is an *association* between the two random variables.

- Random variables X and Y are independent if for any outcome of X and and outcome of Y, $P(X = \text{event A}|Y = \text{event B}) = P(X = \text{event A})$.

Mutually exclusive events are dependent!

- **Example** Let A denote the event a height is in the interval $[5, 5.5]$ and B denote the event a height is in the interval $[6, 6.5]$. Then

$$P(X \text{ in } [5, 5.5]|X \text{ in } [6, 6.5]) = 0 \quad \neq \quad P(X \text{ in } [5, 5.5]).$$

A and B are **not** independent events. In other words, information about Z gives us additional information about Y.

Examples: Independent events

Define the random variables.

- X is the ozone level in a town.
- Y are the number of hair dressers in a town.
- Z is the temperature in a town.
- It is unlikely that the number of hair dressers in a town has an association with the towns temperature. Therefore

$$P(Z \in [25, 28]\text{Celsius}|Y = 10) = P(X \in [25.5, 28.6]).$$

- On the other hand, it is well known that the temperature does have an influence on ground ozone

$$P(X \in [40,60]\text{ppm}|Z \in [25,28]\text{Celsius}) \neq P(X \in [40,60]\text{ppm}).$$

Take care on what you see

A study was done in the early 90s to see if the mortality rates between left and right handed people were the same. To do this the psychologists collected the death records of 2000 people who died in May, 1990, in Southern California. They rang the families of all these people and asked whether the were left or right handed. They also categorized the people who died into those above 60 and below 60 years old.

This is the data they collected.

- Out of the 2000, 400 were left handed.
- Out of the 2000, 150 were left handed and died below 60.
- Out of the 2000, 300 were right handed and died below 60.

(a) A person dies below 60?

(b) A person dies below 60 given that they are left handed?

(c) Does the data suggest there is an association/dependence between left handedess and early mortality?

Solution

It instructive to summarize the data as a contingency table:

	Died before 60	Died after 60	totals
Left	150	250	400
Right	300	1300	1600
Totals	450	1550	2000

(a) The chance a person dies below 60.

Calculate the total proportion who died before 60

$$P(\text{before } 60) = \frac{450}{2000} = 22.5\%$$

(b) A person dies below 60 given that they are left handed.

Focus on people who are **only** left handed. There are 400 of these. Of these, 150 died before 60. Therefore the proportion of left handed people in the sample who died before 60 is

$$P(\text{before 60}|\text{ left handed}) = \frac{150}{400} = \frac{3}{8} = 37.5\%$$

Using the same argument the proportion of right handed people in the sample who died before 60 is

$$P(\text{before 60}|\text{ right handed}) = \frac{300}{1600} = \frac{3}{16} = 18.75\%.$$

(c) As these proportions are different, it suggests there is some sort of dependence/association between lefthandedness and early mortality.

Why the dependence?

- Can we conclude from this data that being left handed **increases** the chance of early mortality?

- This statement is a causal statement, it asks whether being left handed **causes** early mortality.