Data Analytics with R

Sumit Mishra

Institute for Financial Management and Research, Sri City

Probability

11 November 2020

Defining probability

Random processes

- A *random process* is a situation in which we know what outcomes could happen, but we don't know which particular outcome will happen.
- Examples: coin tosses, die rolls, iTunes shuffle, whether the stock market goes up or down tomorrow, etc.
- It can be helpful to model a process as random even if it is not truly random.

MP3 Players › Stories › **iTunes: Just how random is random?**

iTunes: Just how random is random?

By David Braue on 08 March 2007

- **Introduction**
- **Say You, Say What?**
- **A role for labels?**
- **The new random**

Think that song has appeared in your playlists just a few too many times? David Braue puts the randomness of Apple's song shuffling to the test -- and finds some surprising results.

Quick -- think of a number between one and 20. Now think of another one, and another, and another.

Starting to repeat yourself? No surprise: in practice, many series of random numbers are far less random than you would think.

Computers have the same problem. Although all systems are able to pick random numbers, the method they use is often tied to specific other numbers -- for example, the time -- that means you could get a very similar series of 'random' numbers in different situations.

This tendency manifests itself in many ways. For anyone who uses their iPod heavily, you've probably noticed that your supposedly random 'shuffling' iPod seems to be particularly fond of the Bee Gees, Melissa Etheridge or Pavarotti. Look at a random playlist that iTunes generates for you, and you're likely to notice several songs from one or two artists, while other artists go completely unrepresented.

http://www.cnet.com.au/itunes-just-how-random-is-random-339274094.htm

Probability

- There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.
  - $P(A)$ = Probability of event A
  - $0 \leq P(A) \leq 1$

Probability

- There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.
  - $P(A)$ = Probability of event A
  - $0 \leq P(A) \leq 1$
- *Frequentist interpretation:*
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability

- There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.
  - $P(A)$ = Probability of event A
  - $0 \leq P(A) \leq 1$
- *Frequentist interpretation:*
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
- *Bayesian interpretation:*
  - A Bayesian interprets probability as a subjective degree of belief: For the same event, two separate people could have different viewpoints and so assign different probabilities.
  - Largely popularized by revolutionary advance in computational technology and methods during the last twenty years.

Practice

Which of the following events would you be most surprised by?

(a) exactly 3 heads in 10 coin flips

(b) exactly 3 heads in 100 coin flips

(c) exactly 3 heads in 1000 coin flips

Practice

Which of the following events would you be most surprised by?

(a) exactly 3 heads in 10 coin flips

(b) exactly 3 heads in 100 coin flips

(c) *exactly 3 heads in 1000 coin flips*

Law of large numbers

Law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome, $\hat{p}_n$, converges to the probability of that outcome, p.

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$H\ H\ H\ H\ H\ H\ H\ H\ H\ H\ ?$$

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.

$$P(H \text{ on } 11^{th} \text{ toss}) = P(T \text{ on } 11^{th} \text{ toss}) = 0.5$$

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.

$$P(H \text{ on } 11^{th} \text{ toss}) = P(T \text{ on } 11^{th} \text{ toss}) = 0.5$$

- The coin is not "due" for a tail.

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.

$$P(H \text{ on } 11^{th} \text{ toss}) = P(T \text{ on } 11^{th} \text{ toss}) = 0.5$$

- The coin is not "due" for a tail.
- The common misunderstanding of the LLN is that random processes are supposed to compensate for whatever happened in the past; this is just not true and is also called *gambler's fallacy* (or *law of averages*).

Disjoint and non-disjoint outcomes

Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

Disjoint and non-disjoint outcomes

Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

Non-disjoint outcomes: Can happen at the same time.

- A student can get an A in Stats and A in Econ in the same semester.

Union of non-disjoint events

What is the probability of drawing a jack or a red card from a well shuffled full deck?

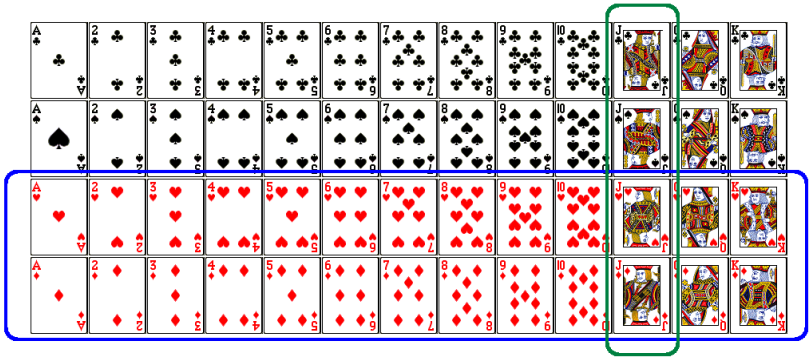
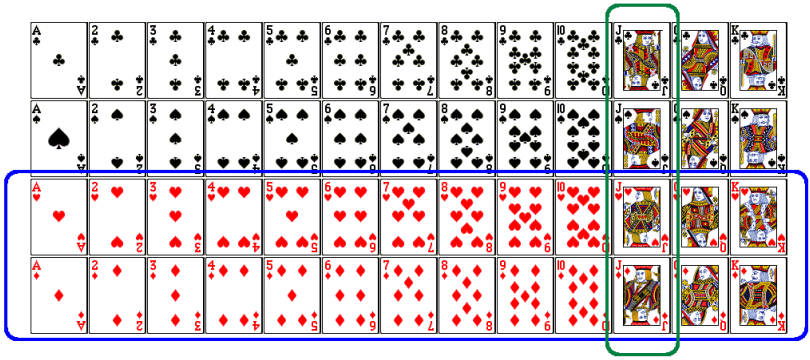


Figure from *http://www.milefoot.com/math/discrete/counting/cardfreq.htm*.

Union of non-disjoint events

What is the probability of drawing a jack or a red card from a well shuffled full deck?



$$P(\textit{jack or red}) = P(\textit{jack}) + P(\textit{red}) - P(\textit{jack and red})$$

$$= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$

Figure from *http://www.milefoot.com/math/discrete/counting/cardfreq.htm*.

Practice

What is the probability that a randomly sampled applicant is a black or a woman in the Bertrand-Mullainathan dataset?

	Gender		
Race	Man	Woman	Total
Black	549	1886	2435
White	575	1860	2435
Total	1124	3476	4870

(a) 0.8

(b) 0.88

(c) $\frac{1886}{2435}$

(d) $\frac{1886}{4870}$

Practice

What is the probability that a randomly sampled applicant is a black or a woman in the Bertrand-Mullainathan dataset?

	Gender		
Race	Man	Woman	Total
Black	549	1886	2435
White	575	1860	2435
Total	1124	3476	4870

(a) 0.8

(b) 0.88

(c) $\frac{1886}{2435}$

(d) $\frac{1886}{4870}$

Recap

General addition rule

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

Note: For disjoint events $P(A \text{ and } B) = 0$, so the above formula simplifies to $P(A \text{ or } B) = P(A) + P(B)$.

Probability distributions

A *probability distribution* lists all possible events and the probabilities with which they occur.

- The probability distribution for a toss of a coin:

Event	Head	Tail
Probability	0.5	0.5

Probability distributions

A *probability distribution* lists all possible events and the probabilities with which they occur.

- The probability distribution for a toss of a coin:

Event	Head	Tail
Probability	0.5	0.5

- Rules for probability distributions:
  1. The events listed must be disjoint
  2. Each probability must be between 0 and 1
  3. The probabilities must total 1

Probability distributions

A *probability distribution* lists all possible events and the probabilities with which they occur.

- The probability distribution for a toss of a coin:

Event	Head	Tail
Probability	0.5	0.5

- Rules for probability distributions:
  1. The events listed must be disjoint
  2. Each probability must be between 0 and 1
  3. The probabilities must total 1
- The probability distribution for two tosses of coin:

Event	*HH*	*HT*	*TH*	*TT*
Probability	*0.25*	*0.25*	*0.25*	*0.25*

Practice

In a pre-election survey, 48% of respondents in Bihar said they will vote for the NDA. What is the probability that a randomly selected respondent from this sample is a MGB voter?

(a) 0.52
(b) more than 0.52
(c) less than 0.52
(d) cannot calculate using only the information given

Practice

In a pre-election survey, 48% of respondents in Bihar said they will vote for the NDA. What is the probability that a randomly selected respondent from this sample is a MGB voter?

(a) 0.52

(b) more than 0.52

(c) less than 0.52

(d) *cannot calculate using only the information given*

If the only two coalitions are the NDA and the MGB, then (a) is possible. However it is also possible that some people do not affiliate with a political party or affiliate with a party other than these two coalitions. Then (c) is also possible. However (b) is definitely not possible since it would result in the total probability for the sample space being above 1.

Sample space and complements

Sample space is the collection of all possible outcomes of a trial.

- Toss of a coin $S = \{H, T\}$
- Two tosses of a coin

Sample space and complements

Sample space is the collection of all possible outcomes of a trial.

- Toss of a coin $S = \{H, T\}$
- Two tosses of a coin $S = \{HH, TT, HT, TH\}$

Sample space and complements

Sample space is the collection of all possible outcomes of a trial.

- Toss of a coin $S = \{H, T\}$
- Two tosses of a coin $S = \{HH, TT, HT, TH\}$

Complementary events are two mutually exclusive events whose probabilities that add up to 1.

- If we know that the toss of a coin didn't yield head, what is the outcome? $\{ \cancel{H}, T \} \rightarrow$ head and tail are *complementary* outcomes.
- Two tosses of a coin: if we know that they are not both tails, what are the possible combinations?

Sample space and complements

Sample space is the collection of all possible outcomes of a trial.

- Toss of a coin $S = \{H, T\}$
- Two tosses of a coin $S = \{HH, TT, HT, TH\}$

Complementary events are two mutually exclusive events whose probabilities that add up to 1.

- If we know that the toss of a coin didn't yield head, what is the outcome? $\{ \cancel{H}, T \} \rightarrow$ head and tail are *complementary* outcomes.
- Two tosses of a coin: if we know that they are not both tails, what are the possible combinations? $\{ HH, \cancel{TT}, HT, TH \}$

Independence

Two processes are *independent* if knowing the outcome of one provides no useful information about the outcome of the other.

Independence

Two processes are *independent* if knowing the outcome of one provides no useful information about the outcome of the other.

- Knowing that the coin landed on a head on the first toss does not provide any useful information for determining what the coin will land on in the second toss. $\rightarrow$ Outcomes of two tosses of a coin are independent.

Independence

Two processes are *independent* if knowing the outcome of one provides no useful information about the outcome of the other.

- Knowing that the coin landed on a head on the first toss does not provide any useful information for determining what the coin will land on in the second toss. $\rightarrow$ Outcomes of two tosses of a coin are independent.
- Knowing that the first card drawn from a deck is an ace does provide useful information for determining the probability of drawing an ace in the second draw. $\rightarrow$ Outcomes of two draws from a deck of cards (without replacement) are dependent.

Practice

Economist Priyanka Pandey surveyed a random sample of IIT-BHU students. One of the questions asked in the survey was about the perceptions of academic ability of lower-caste students. 61% upper-caste students felt that academic ability of SC students was lower than others, whereas 46 % SC-ST students felt the same. Which of the below is true?

Perception about caste and caste are most likely

(a) complementary
(b) mutually exclusive
(c) independent
(d) dependent
(e) disjoint

Note: You can read the full article here-

https://www.epw.in/engage/article/Survey-at-an-IIT-Campus-Shows-How-Caste-Affects-Students-Perceptions.

Practice

Economist Priyanka Pandey surveyed a random sample of IIT-BHU students. One of the questions asked in the survey was about the perceptions of academic ability of lower-caste students. 61% upper-caste students felt that academic ability of SC students was lower than others, whereas 46 % SC-ST students felt the same. Which of the below is true?

Perception about caste and caste are most likely

(a) complementary
(b) mutually exclusive
(c) independent
(d) *dependent*
(e) disjoint

Note: You can read the full article here-

https://www.epw.in/engage/article/Survey-at-an-IIT-Campus-Shows-How-Caste-Affects-Students-Perceptions.

Checking for independence

If P(A occurs, given that B is true) = $P(A \mid B) = P(A)$, then A and B are independent.

Checking for independence

If P(A occurs, given that B is true) = $P(A \mid B) = P(A)$, then A and B are independent.

P(demonetization curbed black money) = 0.25

Checking for independence

If P(A occurs, given that B is true) = $P(A \mid B) = P(A)$, then A and B are independent.

P(demonetization curbed black money) = 0.25

P(randomly selected respondent says demonetization curbed black money, given that the resident is a BJP voter) =
P(demonetization curbed black money | BJP) = 0.45

P(demonetization curbed black money | Congress) = 0.15

P(demonetization curbed black money | Others) = 0.20

Checking for independence

If P(A occurs, given that B is true) = $P(A \mid B) = P(A)$, then A and B are independent.

P(demonetization curbed black money) = 0.25

P(randomly selected respondent says demonetization curbed black money, given that the resident is a BJP voter) =
P(demonetization curbed black money | BJP) = 0.45

P(demonetization curbed black money | Congress) = 0.15

P(demonetization curbed black money | Others) = 0.20

P(demonetization curbed black money) varies by party-affiliation, therefore the two variables are mostly likely dependent.

Determining dependence based on sample data

- If conditional probabilities calculated based on sample data suggest dependence between two variables, the next step is to conduct a hypothesis test to determine if the observed difference between the probabilities is likely or unlikely to have happened by chance.
- If the observed difference between the conditional probabilities is large, then there is stronger evidence that the difference is real.
- If a sample is large, then even a small difference can provide strong evidence of a real difference.

Determining dependence based on sample data

- If conditional probabilities calculated based on sample data suggest dependence between two variables, the next step is to conduct a hypothesis test to determine if the observed difference between the probabilities is likely or unlikely to have happened by chance.
- If the observed difference between the conditional probabilities is large, then there is stronger evidence that the difference is real.
- If a sample is large, then even a small difference can provide strong evidence of a real difference.

We saw that P(callback | Black) = 6.4% and P(callback | White) = 9.7%. Under which condition would you be more convinced of a real difference between the proportions of callbacks for African-American and White CVs? $n = 50$ or $n = 5,000$

Determining dependence based on sample data

- If conditional probabilities calculated based on sample data suggest dependence between two variables, the next step is to conduct a hypothesis test to determine if the observed difference between the probabilities is likely or unlikely to have happened by chance.
- If the observed difference between the conditional probabilities is large, then there is stronger evidence that the difference is real.
- If a sample is large, then even a small difference can provide strong evidence of a real difference.

We saw that P(callback | Black) = 6.4% and P(callback | White) = 9.7%. Under which condition would you be more convinced of a real difference between the proportions of callbacks for African-American and White CVs? $n = 50$ or $n = 5,000$

$n = 5,000$

For two independent events, the product rule is:

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

Or more generally, $P(A_1 \text{ and } \cdots \text{ and } A_k) = P(A_1) \times \cdots \times P(A_k)$

For two independent events, the product rule is:

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

Or more generally, $P(A_1 \text{ and } \cdots \text{ and } A_k) = P(A_1) \times \cdots \times P(A_k)$

You toss a coin twice, what is the probability of getting two tails in a row?

For two independent events, the product rule is:

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

Or more generally, $P(A_1 \text{ and } \cdots \text{ and } A_k) = P(A_1) \times \cdots \times P(A_k)$

You toss a coin twice, what is the probability of getting two tails in a row?

$$P(\text{T on the first toss}) \times P(\text{T on the second toss}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Practice

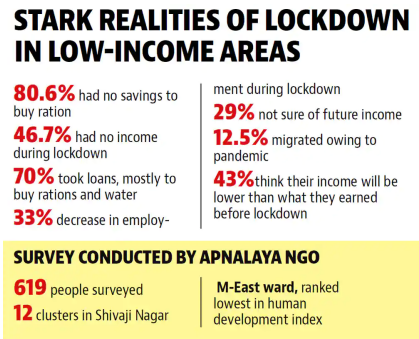
A recent survey suggests that 80.6% of Shivaji Nagar had no saving to buy ration during the lockdown. Assuming that the saving rate stayed constant, what is the probability that two randomly selected persons from Shivaji Nagar had no savings?

(a) 80.6^2

(b) 0.806^2

(c) 0.806×2

(d) $(1 - 0.806)^2$



https://www.hindustantimes.com/mumbai-news/with-no-income-during-lockdown-people-in-slums-forced-to-borrow-money-for-water-survey/story-px8fdUIWQQTXbjBCJPBhlK.html

Practice

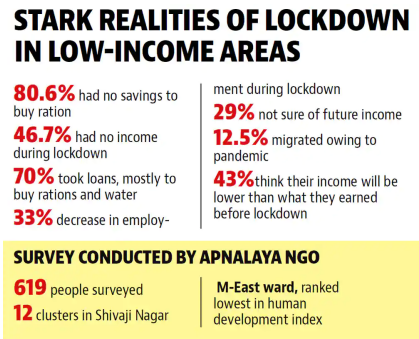
A recent survey suggests that 80.6% of Shivaji Nagar had no saving to buy ration during the lockdown. Assuming that the saving rate stayed constant, what is the probability that two randomly selected persons from Shivaji Nagar had no savings?

(a) 80.6^2

(b) 0.806^2

(c) 0.806×2

(d) $(1 - 0.806)^2$



https://www.hindustantimes.com/mumbai-news/with-no-income-during-lockdown-people-in-slums-forced-to-borrow-money-for-water-survey/story-px8fdUIWQQTXbjBCJPBhlK.html

Putting everything together...

If we were to randomly select 5 households from Shivaji Nagar, what is the probability that at least one had no saving?

- If we were to randomly select 5 households, the sample space for the number of households without savings would be:

$$S = \{0, 1, 2, 3, 4, 5\}$$

- We are interested in instances where at least one household is without saving:

$$S = \{0, 1, 2, 3, 4, 5\}$$

- So we can divide up the sample space into two categories:

$$S = \{0, \textit{at least one}\}$$

Conditional probability

Dataset : `photoclassify`
What it contains: 1822 photos from a photo-sharing website
Objective: figure out whether a photo is about fashion or not. Method:

Each photo gets two classifications (the variable is called `mach_learn`):

1. `predfashion`
2. `prednot`

Each photo also gets categorized (the variable is called `truth`):

1. `fashion`
2. `not`

Dataset : `photoclassify`
What it contains: 1822 photos from a photo-sharing website
Objective: figure out whether a photo is about fashion or not. Method:

Each photo gets two classifications (the variable is called `mach_learn`):

1. `predfashion`
2. `prednot`

Each photo also gets categorized (the variable is called `truth`):

1. `fashion`
2. `not`

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

Figure: Contingency table summarizing the `photo_classify` data set.

What is the probability that a photo is not actually about fashion?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

Marginal probability

What is the probability that a photo is not actually about fashion?

mach_learn		truth		
		fashion	not	Total
	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

Marginal probability

What is the probability that a photo is not actually about fashion?

mach_learn	truth		
	fashion	not	Total
pred_fashion	197	22	219
pred_not	112	1491	1603
Total	309	*1513*	*1822*

P(not fashion) = $\frac{1513}{1822} \approx 0.83$

Joint probability

What is the probability that a photo was about fashion and the ML method correctly classified it as one?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

Joint probability

What is the probability that a photo was about fashion and the ML method correctly classified it as one?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	*197*	22	219
	pred_not	112	1491	1603
	Total	309	1513	*1822*

P(truth and predicted) = $\frac{197}{1822} \approx 0.11$

Conditional probability

The conditional probability of the outcome of interest A given condition B is calculated as

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

Conditional probability

The conditional probability of the outcome of interest A given condition B is calculated as

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

$$P(truth|predicted) = \frac{P(truth \text{ and } predicted)}{P(predicted)}$$

`mach_learn`

	truth		
	fashion	not	Total
pred_fashion	197	22	219
pred_not	112	1491	1603
Total	309	1513	1822

Conditional probability

The conditional probability of the outcome of interest A given condition B is calculated as

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

`mach_learn`

	truth		
	`fashion`	`not`	Total
`pred_fashion`	197	22	219
`pred_not`	112	1491	1603
Total	309	1513	1822

$$P(\textit{truth}|\textit{predicted}) = \frac{P(\textit{truth and predicted})}{P(\textit{predicted})} = \frac{197/1822}{219/1822}$$

Conditional probability

The conditional probability of the outcome of interest A given condition B is calculated as

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

`mach_learn`

	truth		
	fashion	not	Total
pred_fashion	197	22	219
pred_not	112	1491	1603
Total	309	1513	1822

$$\begin{aligned} & P(\textit{truth}|\textit{predicted}) \\ &= \frac{P(\textit{truth and predicted})}{P(\textit{predicted})} \\ &= \frac{197/1822}{219/1822} \\ &= \frac{197}{219} \end{aligned}$$

Conditional probability

The conditional probability of the outcome of interest A given condition B is calculated as

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

`mach_learn`

	truth		
	fashion	not	Total
pred_fashion	197	22	219
pred_not	112	1491	1603
Total	309	1513	1822

$$\begin{aligned} & P(\textit{truth}|\textit{predicted}) \\ &= \frac{P(\textit{truth and predicted})}{P(\textit{predicted})} \\ &= \frac{197/1822}{219/1822} \\ &= \frac{197}{219} \\ &= 0.9 \end{aligned}$$

Conditional probability (cont.)

If we know that a particular photo is about fashion, what is the probability that the machine learning classifier correctly predicted it as one?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

Conditional probability (cont.)

If we know that a particular photo is about fashion, what is the probability that the machine learning classifier correctly predicted it as one?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

$$P(\textit{predicted} | \textit{truth}) = \frac{P(\textit{predicted and truth})}{P(\textit{truth})}$$

Conditional probability (cont.)

If we know that a particular photo is about fashion, what is the probability that the machine learning classifier correctly predicted it as one?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

$$
\begin{aligned}
& P(\textit{predicted} | \textit{truth}) \\
&= \frac{P(\textit{predicted and truth})}{P(\textit{truth})} \\
&= \frac{197/1822}{309/1822}
\end{aligned}
$$

Conditional probability (cont.)

If we know that a particular photo is about fashion, what is the probability that the machine learning classifier correctly predicted it as one?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

$$
\begin{aligned}
& P(\textit{predicted} | \textit{truth}) \\
&= \frac{P(\textit{predicted and truth})}{P(\textit{truth})} \\
&= \frac{197/1822}{309/1822} \\
&= \frac{197}{309}
\end{aligned}
$$

Conditional probability (cont.)

If we know that a particular photo is about fashion, what is the probability that the machine learning classifier correctly predicted it as one?

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

$$
\begin{aligned}
& P(\textit{predicted} | \textit{truth}) \\
&= \frac{P(\textit{predicted and truth})}{P(\textit{truth})} \\
&= \frac{197/1822}{309/1822} \\
&= \frac{197}{309} \\
&= 0.64
\end{aligned}
$$

General multiplication rule

- Earlier we saw that if two events are independent, their joint probability is simply the product of their probabilities. If the events are not believed to be independent, the joint probability is calculated slightly differently.

General multiplication rule

- Earlier we saw that if two events are independent, their joint probability is simply the product of their probabilities. If the events are not believed to be independent, the joint probability is calculated slightly differently.
- If A and B represent two outcomes or events, then

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

Note that this formula is simply the conditional probability formula, rearranged.

General multiplication rule

- Earlier we saw that if two events are independent, their joint probability is simply the product of their probabilities. If the events are not believed to be independent, the joint probability is calculated slightly differently.
- If A and B represent two outcomes or events, then

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

  Note that this formula is simply the conditional probability formula, rearranged.
- It is useful to think of A as the outcome of interest and B as the condition.

Independence and conditional probabilities

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

Independence and conditional probabilities

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is

Independence and conditional probabilities

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is $\frac{60}{100} = 0.6$.

Independence and conditional probabilities

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is $\frac{60}{100} = 0.6$.
- The probability that a randomly selected student is a social science major given that they are female is

Independence and conditional probabilities

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is $\frac{60}{100} = 0.6$.
- The probability that a randomly selected student is a social science major given that they are female is $\frac{30}{50} = 0.6$.

Independence and conditional probabilities

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is $\frac{60}{100} = 0.6$.
- The probability that a randomly selected student is a social science major given that they are female is $\frac{30}{50} = 0.6$.
- Since $P(SS|M)$ also equals 0.6, major of students in this class does not depend on their gender: P(SS | F) = P(SS).

Independence and conditional probabilities (cont.)

Generically, if $P(A|B) = P(A)$ then the events A and B are said to be independent.

Independence and conditional probabilities (cont.)

Generically, if $P(A|B) = P(A)$ then the events A and B are said to be independent.

- Conceptually: Giving B doesn't tell us anything about A.

Independence and conditional probabilities (cont.)

Generically, if $P(A|B) = P(A)$ then the events A and B are said to be independent.

- Conceptually: Giving B doesn't tell us anything about A.
- Mathematically: We know that if events A and B are independent, $P(A \textit{ and } B) = P(A) \times P(B)$. Then,

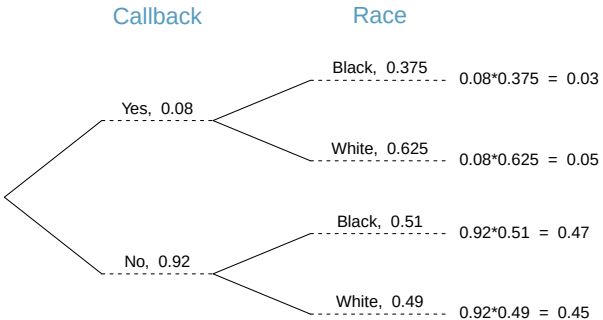
$$P(A|B) = \frac{P(A \textit{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Inverting probabilities

When an applicant applies for a job, there are two possibilities: callback, and no callback. If the candidate has a Black-sounding name, what is the probability that she got a callback?

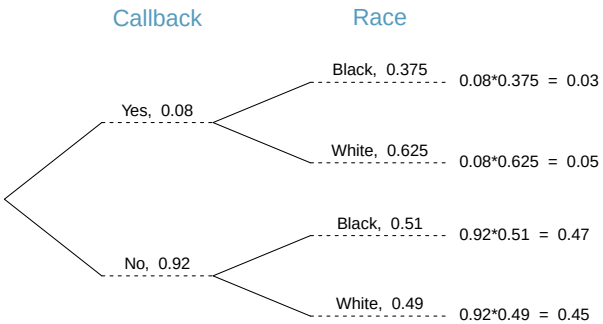
Inverting probabilities

When an applicant applies for a job, there are two possibilities: callback, and no callback. If the candidate has a Black-sounding name, what is the probability that she got a callback?



Inverting probabilities

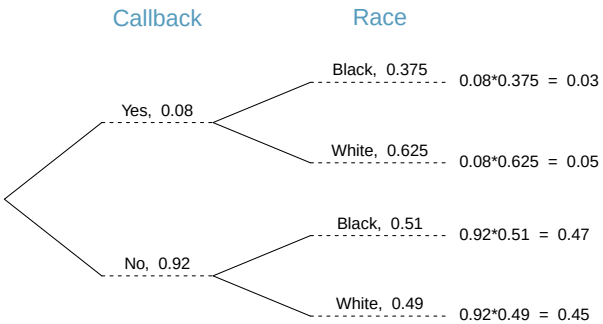
When an applicant applies for a job, there are two possibilities: callback, and no callback. If the candidate has a Black-sounding name, what is the probability that she got a callback?



$P(Callback|Black)$

Inverting probabilities

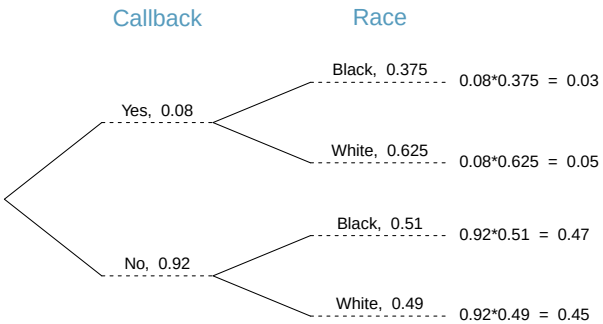
When an applicant applies for a job, there are two possibilities: callback, and no callback. If the candidate has a Black-sounding name, what is the probability that she got a callback?



$$P(\textit{Callback}|\textit{Black}) = \frac{P(\textit{Callback and Black})}{P(\textit{Black})}$$

Inverting probabilities

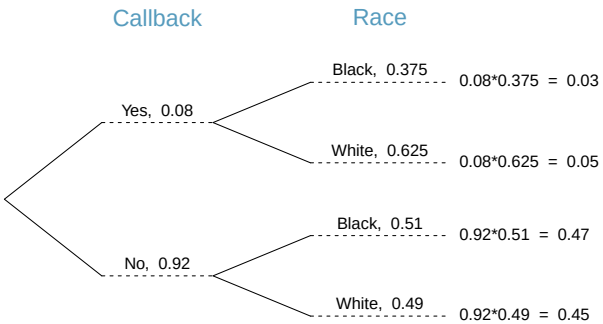
When an applicant applies for a job, there are two possibilities: callback, and no callback. If the candidate has a Black-sounding name, what is the probability that she got a callback?



$$
\begin{aligned}
&P(\textit{Callback}|\textit{Black}) \\
&= \frac{P(\textit{Callback and Black})}{P(\textit{Black})} \\
&= \frac{0.03}{0.03 + 0.047}
\end{aligned}
$$

Inverting probabilities

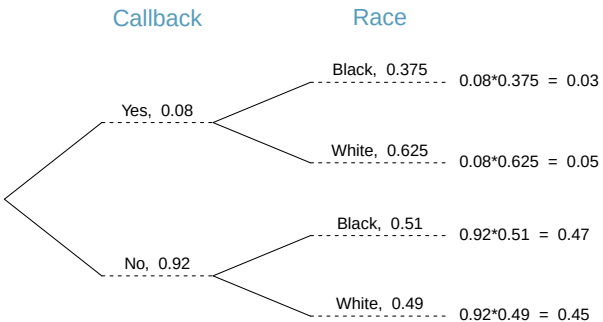
When an applicant applies for a job, there are two possibilities: callback, and no callback. If the candidate has a Black-sounding name, what is the probability that she got a callback?



$$
\begin{aligned}
&P(\textit{Callback}|\textit{Black}) \\
&= \frac{P(\textit{Callback and Black})}{P(\textit{Black})} \\
&= \frac{0.03}{0.03 + 0.047} \\
&= 0.06
\end{aligned}
$$

Inverting probabilities

When an applicant applies for a job, there are two possibilities: callback, and no callback. If the candidate has a Black-sounding name, what is the probability that she got a callback?



$$
\begin{aligned}
&P(\textit{Callback}|\textit{Black}) \\
&= \frac{P(\textit{Callback and Black})}{P(\textit{Black})} \\
&= \frac{0.03}{0.03 + 0.047} \\
&= 0.06
\end{aligned}
$$

Note: Tree diagrams are useful for inverting probabilities: we are given $P(\textit{Black}|\textit{Callback})$ and asked for $P(\textit{Callback}|\textit{Black})$.

Bayes' Theorem

- The conditional probability formula we have seen so far is a special case of the Bayes' Theorem, which is applicable even when events have more than just two outcomes.

Bayes' Theorem

- The conditional probability formula we have seen so far is a special case of the Bayes' Theorem, which is applicable even when events have more than just two outcomes.
- *Bayes' Theorem:*

$$P(\textit{outcome } A_1 \textit{ of variable } 1 \mid \textit{outcome } B \textit{ of variable } 2) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

where $A_2, \cdots, A_k$ represent all other possible outcomes of variable 1.

Application activity: Inverting probabilities

Stores Murugan, Saravana, and Nilgiris have 50, 75, and 100 employees and, respectively, 50, 60, 70 percent of these are women. Resignations are equally likely among all employees, regardless of sex. One employee resigns, and this happens to be a woman. What is the probability that she works in Nilgiris store?

Application activity: Inverting probabilities (cont.)

Start figuring out the probability of resignation from each store.

Application activity: Inverting probabilities (cont.)

Start figuring out the probability of resignation from each store.
$P(W_1|M) = 0.5$, $P(W_2|S) = 0.6$, $P(W_3|N) = 0.7$.
$P(M) = 2/9$, $P(S) = 1/3$, $P(N) = 4/9$.

Application activity: Inverting probabilities (cont.)

Start figuring out the probability of resignation from each store.
$P(W_1|M) = 0.5$, $P(W_2|S) = 0.6$, $P(W_3|N) = 0.7$.
$P(M) = 2/9$, $P(S) = 1/3$, $P(N) = 4/9$.

$$P(N|W_3) = \frac{P(N \text{ and } W_3)}{P(W_3)} = \frac{0.311}{0.111 + 0.2 + 0.311} = 0.5$$

Sampling from a small population

Sampling with replacement

When sampling *with replacement*, you put back what you just drew.

Sampling with replacement

When sampling *with replacement*, you put back what you just drew.

- Imagine you have a bag with 5 red, 3 blue and 2 orange chips in it. What is the probability that the first chip you draw is blue?

5 🔴 , 3 🔵 , 2 🟠

Sampling with replacement

When sampling *with replacement*, you put back what you just drew.

- Imagine you have a bag with 5 red, 3 blue and 2 orange chips in it. What is the probability that the first chip you draw is blue?

5 🔴 , 3 🔵 , 2 🟠

$$Prob(1^{st} \text{ chip } B) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$$

Sampling with replacement

When sampling *with replacement*, you put back what you just drew.

- Imagine you have a bag with 5 red, 3 blue and 2 orange chips in it. What is the probability that the first chip you draw is blue?

5 🔴 , 3 🔵 , 2 🟠

$$Prob(1^{st} \text{ chip } B) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$$

- Suppose you did indeed pull a blue chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

Sampling with replacement

When sampling *with replacement*, you put back what you just drew.

- Imagine you have a bag with 5 red, 3 blue and 2 orange chips in it. What is the probability that the first chip you draw is blue?

5 🔴 , 3 🔵 , 2 🟠

$$Prob(1^{st} \text{ chip } B) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$$

- Suppose you did indeed pull a blue chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴 , 3 🔵 , 2 🟠

Sampling with replacement

When sampling *with replacement*, you put back what you just drew.

- Imagine you have a bag with 5 red, 3 blue and 2 orange chips in it. What is the probability that the first chip you draw is blue?

5 🔴 , 3 🔵 , 2 🟠

$$Prob(1^{st} \text{ chip } B) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$$

- Suppose you did indeed pull a blue chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴 , 3 🔵 , 2 🟠
2^{nd} draw: 5 🔴 , 3 🔵 , 2 🟠

Sampling with replacement

When sampling *with replacement*, you put back what you just drew.

- Imagine you have a bag with 5 red, 3 blue and 2 orange chips in it. What is the probability that the first chip you draw is blue?

5 🔴 , 3 🔵 , 2 🟠

$$Prob(1^{st} \text{ chip } B) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$$

- Suppose you did indeed pull a blue chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴 , 3 🔵 , 2 🟠
2^{nd} draw: 5 🔴 , 3 🔵 , 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } B) = \frac{3}{10} = 0.3$$

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴 , 3 🔵 , 2 🟠

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 3 🔵, 2 🟠

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 3 🔵, 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } O) = \frac{3}{10} = 0.3$$

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 3 🔵, 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } O) = \frac{3}{10} = 0.3$$

- If drawing with replacement, what is the probability of drawing two blue chips in a row?

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 3 🔵, 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } O) = \frac{3}{10} = 0.3$$

- If drawing with replacement, what is the probability of drawing two blue chips in a row?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 3 🔵, 2 🟠

Sampling with replacement (cont.)

- Suppose you actually pulled an orange chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 3 🔵, 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } O) = \frac{3}{10} = 0.3$$

- If drawing with replacement, what is the probability of drawing two blue chips in a row?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 3 🔵, 2 🟠

$$Prob(1^{st} \text{ chip } B) \cdot Prob(2^{nd} \text{ chip } O | 1^{st} \text{ chip } B) = 0.3 \times 0.3$$

$$= 0.3^2 = 0.09$$

Sampling with replacement (cont.)

- When drawing with replacement, probability of the second chip being blue does not depend on the color of the first chip since whatever we draw in the first draw gets put back in the bag.

$$Prob(B|B) = Prob(B|O)$$

- In addition, this probability is equal to the probability of drawing a blue chip in the first draw, since the composition of the bag never changes when sampling with replacement.

$$Prob(B|B) = Prob(B)$$

- *When drawing with replacement, draws are independent.*

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴 , 3 🔵 , 2 🟠

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 2 🔵, 2 🟠

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠

2^{nd} draw: 5 🔴, 2 🔵, 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } B) = \frac{2}{9} = 0.22$$

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

$$1^{st} \text{ draw: } 5 \text{ 🔴}, 3 \text{ 🔵}, 2 \text{ 🟠}$$
$$2^{nd} \text{ draw: } 5 \text{ 🔴}, 2 \text{ 🔵}, 2 \text{ 🟠}$$
$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } B) = \frac{2}{9} = 0.22$$

- If drawing without replacement, what is the probability of drawing two blue chips in a row?

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠
2^{nd} draw: 5 🔴, 2 🔵, 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } B) = \frac{2}{9} = 0.22$$

- If drawing without replacement, what is the probability of drawing two blue chips in a row?

1^{st} draw: 5 🔴, 3 🔵, 2 🟠
2^{nd} draw: 5 🔴, 2 🔵, 2 🟠

Sampling without replacement

When drawing *without replacement* you do not put back what you just drew.

- Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 🔴 , 3 🔵 , 2 🟠

2^{nd} draw: 5 🔴 , 2 🔵 , 2 🟠

$$Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } B) = \frac{2}{9} = 0.22$$

- If drawing without replacement, what is the probability of drawing two blue chips in a row?

1^{st} draw: 5 🔴 , 3 🔵 , 2 🟠

2^{nd} draw: 5 🔴 , 2 🔵 , 2 🟠

$$Prob(1^{st} \text{ chip } B) \cdot Prob(2^{nd} \text{ chip } B | 1^{st} \text{ chip } B) = 0.3 \times 0.22$$

$$= 0.066$$

Sampling without replacement (cont.)

- When drawing without replacement, the probability of the second chip being blue given the first was blue is not equal to the probability of drawing a blue chip in the first draw since the composition of the bag changes with the outcome of the first draw.

$$Prob(B|B) \neq Prob(B)$$

Sampling without replacement (cont.)

- When drawing without replacement, the probability of the second chip being blue given the first was blue is not equal to the probability of drawing a blue chip in the first draw since the composition of the bag changes with the outcome of the first draw.

$$Prob(B|B) \neq Prob(B)$$

- *When drawing without replacement, draws are not independent.*

Sampling without replacement (cont.)

- When drawing without replacement, the probability of the second chip being blue given the first was blue is not equal to the probability of drawing a blue chip in the first draw since the composition of the bag changes with the outcome of the first draw.

$$Prob(B|B) \neq Prob(B)$$

- *When drawing without replacement, draws are not independent.*
- This is especially important to take note of when the sample sizes are small. If we were dealing with, say, 10,000 chips in a (giant) bag, taking out one chip of any color would not have as big an impact on the probabilities in the second draw.

Practice

In most card games cards are dealt without replacement. What is the probability of being dealt an ace and then a 3? Choose the closest answer.

(a) 0.0045
(b) 0.0059
(c) 0.0060
(d) 0.1553

Practice

In most card games cards are dealt without replacement. What is the probability of being dealt an ace and then a 3? Choose the closest answer.

(a) 0.0045
(b) 0.0059
(c) *0.0060*
(d) 0.1553

$$P(\textit{ace then } 3) = \frac{4}{52} \times \frac{4}{51} \approx 0.0060$$

Random variables

Random variables

- A *random variable* is a numeric quantity whose value depends on the outcome of a random event
  - We use a capital letter, like X, to denote a random variable
  - The values of a random variable are denoted with a lowercase letter, in this case x
  - For example, $P(X = x)$
- There are two types of random variables:
  - *Discrete random variables* often take only integer values
    - Example: Number of credit hours, Difference in number of credit hours this term vs last
  - *Continuous random variables* take real (decimal) values
    - Example: Cost of books this term, Difference in cost of books this term vs last

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the *expected value* (mean),
  and it is a weighted average of the possible outcomes

$$\mu = E(X) = \sum_{i=1}^{k} x_i \, P(X = x_i)$$

Expected value of a discrete random variable

In a game of cards you win ₹10 if you draw a heart, ₹50 if you draw an ace (including the ace of hearts), ₹100 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

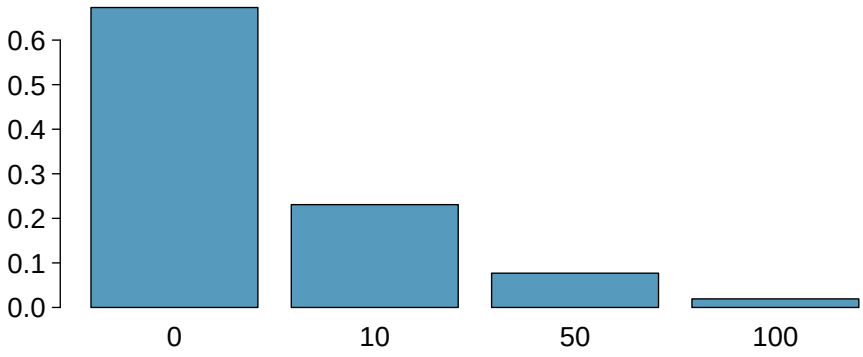
Expected value of a discrete random variable

In a game of cards you win ₹10 if you draw a heart, ₹50 if you draw an ace (including the ace of hearts), ₹100 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	X	$P(X)$	$X \cdot P(X)$
Heart (not ace)	10	$\frac{12}{52}$	$10 \times \frac{12}{52}$
Ace	50	$\frac{4}{52}$	$50 \times \frac{4}{52}$
King of spades	100	$\frac{1}{52}$	$100 \times \frac{1}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{420}{52} \approx 8.1$

Expected value of a discrete random variable (cont.)

Below is a visual representation of the probability distribution of winnings from this game:



Variability

We are also often interested in the variability in the values of a random variable.

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
10	0.23	2.31	900.00	207.69
50	0.08	3.85	100.00	7.69
100	0.02	1.92	3600.00	69.23
0	0.67	0.00	1600.00	1076.92
		$E(X) = 0.81$		

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
10	0.23	2.31	900.00	207.69
50	0.08	3.85	100.00	7.69
100	0.02	1.92	3600.00	69.23
0	0.67	0.00	1600.00	1076.92
		$E(X) = 0.81$		$V(X) = 1361.5$

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
10	0.23	2.31	900.00	207.69
50	0.08	3.85	100.00	7.69
100	0.02	1.92	3600.00	69.23
0	0.67	0.00	1600.00	1076.92
		$E(X) = 0.81$		$V(X) = 1361.5$ $SD(X) = \sqrt{1361.5} = 36.9$

Linear combinations

- A *linear combination* of random variables X and Y is given by

$$aX + bY$$

where a and b are some fixed numbers.

Linear combinations

- A *linear combination* of random variables X and Y is given by

$$aX + bY$$

  where a and b are some fixed numbers.
- The average value of a linear combination of random variables is given by

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each economics homework problem. This week you have 5 statistics and 4 economics homework problems assigned. What is the total time you expect to spend on statistics and economics homework for the week?

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each economics homework problem. This week you have 5 statistics and 4 economics homework problems assigned. What is the total time you expect to spend on statistics and economics homework for the week?

$$
\begin{aligned}
E(S + S + S + S + S + Ec + Ec + Ec + Ec) &= 5 \times E(S) + 4 \times E(Ec) \\
&= 5 \times 10 + 4 \times 15 \\
&= 50 + 60 \\
&= 110 \; min
\end{aligned}
$$

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.

Note: If the random variables are not independent, the variance calculation gets a little more complicated and is beyond the scope of this course.

Calculating the variance of a linear combination

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each economics problem. What is the standard deviation of the time you expect to spend on statistics and economics homework for the week if you have 2 statistics and 2 economics homework problems assigned? Suppose that the time it takes to complete each problem is independent of another.

Calculating the variance of a linear combination

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each economics problem. What is the standard deviation of the time you expect to spend on statistics and economics homework for the week if you have 2 statistics and 2 economics homework problems assigned? Suppose that the time it takes to complete each problem is independent of another.

$$
\begin{aligned}
V(S + S + E + E) &= V(S) + V(S) + V(E) + V(E) \\
&= 2 \times (V(S) + V(E)) \\
&= 2 \times (1.5^2 + 2^2) \\
&= 12.5
\end{aligned}
$$

Practice

A casino game costs $5 to play. If the first card you draw is red, then you get to draw a second card (without replacement). If the second card is the ace of clubs, you win $500. If not, you don't win anything, i.e. lose your $5. What is your expected profits/losses from playing this game? Remember: profit/loss = winnings - cost.

(a) A profit of 5¢

(b) A loss of 10¢

(c) A loss of 25¢

(d) A loss of 30¢

Practice

A casino game costs \$5 to play. If the first card you draw is red, then you get to draw a second card (without replacement). If the second card is the ace of clubs, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits/losses from playing this game? Remember: profit/loss = winnings - cost.

(a) A profit of 5¢

(b) *A loss of 10¢*

(c) A loss of 25¢

(d) A loss of 30¢

Event	*Win*	*Profit: X*	$P(X)$	$X \times P(X)$
Red, A♣	*500*	*500 - 5 = 495*	$\frac{26}{52} \times \frac{1}{51} = 0.0098$	$495 \times 0.0098 = 4.851$
Other	*0*	*0 - 5 = -5*	$1 - 0.0098 = 0.9902$	$-5 \times 0.9902 = -4.951$
				$E(X) = -0.1$

Fair game

A *fair* game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Fair game

A *fair* game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Do you think casino games in Vegas cost more or less than their expected payouts?

Fair game

A *fair* game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Do you think casino games in Vegas cost more or less than their expected payouts?

If those games cost less than their expected payouts, it would mean that the casinos would be losing money on average, and hence they wouldn't be able to pay for all this:

Image by Moyan_Brenn on Flickr *http://www.flickr.com/photos/aigle_dore/5951714693*.

Simplifying random variables

Random variables do not work like normal algebraic variables:

$$X + X \neq 2X$$

Simplifying random variables

Random variables do not work like normal algebraic variables:

$$X + X \neq 2X$$

$$\begin{aligned} E(X+X) &= E(X) + E(X) \\ &= 2E(X) \end{aligned}$$

$$\begin{aligned} Var(X+X) &= Var(X) + Var(X) \text{ (assuming independence)} \\ &= 2\ Var(X) \end{aligned}$$

$$E(2X) = 2E(X)$$

$$\begin{aligned} Var(2X) &= 2^2\ Var(X) \\ &= 4\ Var(X) \end{aligned}$$

Simplifying random variables

Random variables do not work like normal algebraic variables:

$$X + X \neq 2X$$

$$\begin{aligned} E(X+X) &= E(X) + E(X) \\ &= 2E(X) \end{aligned}$$

$$\begin{aligned} Var(X+X) &= Var(X) + Var(X) \text{ (assuming independence)} \\ &= 2\ Var(X) \end{aligned}$$

$$E(2X) = 2E(X)$$

$$\begin{aligned} Var(2X) &= 2^2\ Var(X) \\ &= 4\ Var(X) \end{aligned}$$

$E(X+X) = E(2X)$, but $Var(X+X) \neq Var(2X)$.

Adding or multiplying?

A university in Sri City has 5 cars for transportation on campus. Historical data show that annual maintenance cost for each car is on average ₹10,000 with a standard deviation of ₹100. What is the mean and the standard deviation of the total annual maintenance cost for these vehicles?

Adding or multiplying?

A university in Sri City has 5 cars for transportation on campus. Historical data show that annual maintenance cost for each car is on average ₹10,000 with a standard deviation of ₹100. What is the mean and the standard deviation of the total annual maintenance cost for these vehicles?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

Adding or multiplying?

A university in Sri City has 5 cars for transportation on campus. Historical data show that annual maintenance cost for each car is on average ₹10,000 with a standard deviation of ₹100. What is the mean and the standard deviation of the total annual maintenance cost for these vehicles?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$E(X_1 + X_2 + X_3 + X_4 + X_5) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5)$$

Adding or multiplying?

A university in Sri City has 5 cars for transportation on campus. Historical data show that annual maintenance cost for each car is on average ₹10,000 with a standard deviation of ₹100. What is the mean and the standard deviation of the total annual maintenance cost for these vehicles?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$\begin{aligned} E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\ &= 5 \times E(X) = 5 \times 10,000 = ₹50,000 \end{aligned}$$

Adding or multiplying?

A university in Sri City has 5 cars for transportation on campus. Historical data show that annual maintenance cost for each car is on average ₹10,000 with a standard deviation of ₹100. What is the mean and the standard deviation of the total annual maintenance cost for these vehicles?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 10,000 = ₹50,000 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5)
\end{aligned}
$$

Adding or multiplying?

A university in Sri City has 5 cars for transportation on campus. Historical data show that annual maintenance cost for each car is on average ₹10,000 with a standard deviation of ₹100. What is the mean and the standard deviation of the total annual maintenance cost for these vehicles?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 10,000 = ₹50,000 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 100^2 = ₹50,000
\end{aligned}
$$

Adding or multiplying?

A university in Sri City has 5 cars for transportation on campus. Historical data show that annual maintenance cost for each car is on average ₹10,000 with a standard deviation of ₹100. What is the mean and the standard deviation of the total annual maintenance cost for these vehicles?

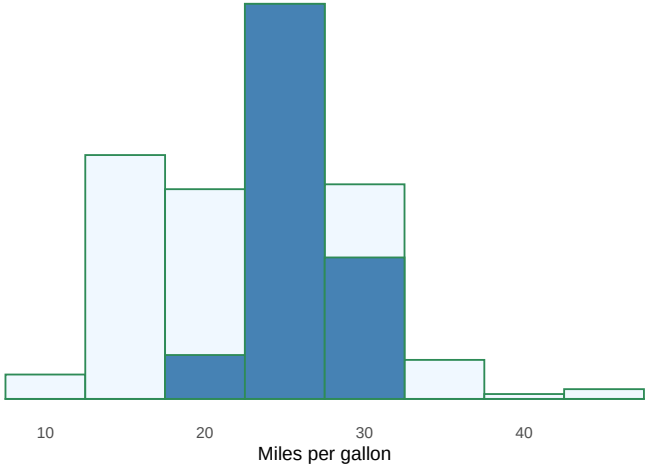
Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 10,000 = ₹50,000 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 100^2 = ₹50,000 \\
SD(X_1 + X_2 + X_3 + X_4 + X_5) &= \sqrt{50,000} = 223.6
\end{aligned}
$$

Continuous distributions

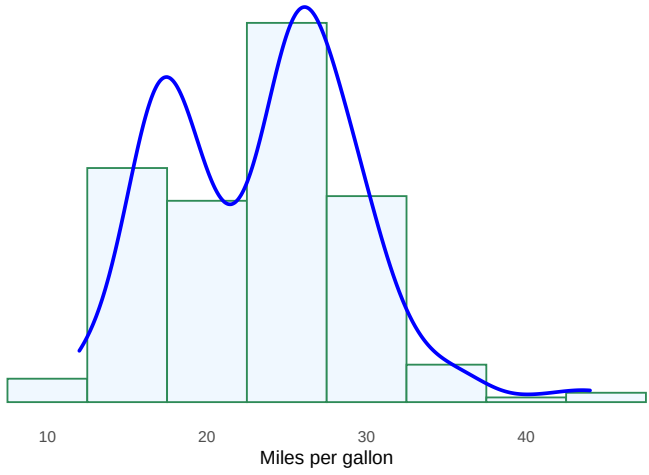
Continuous distributions

- Below is a histogram of the distribution of the variable `hwy` from the dataset `mpg`.
- The proportion of data that falls in the shaded bins gives the probability that a randomly sampled car mileage is between 20 and 30 miles per gallon.



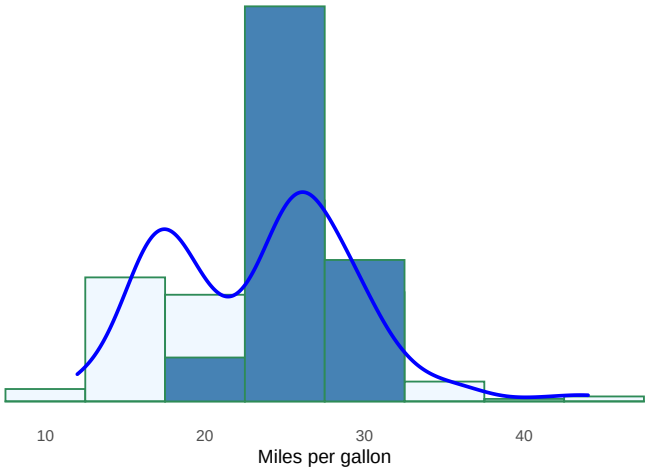
From histograms to continuous distributions

Since miles per gallon is a continuous numerical variable, its *probability density function* is a smooth curve.



Probabilities from continuous distributions

Therefore, the probability that a randomly sampled car has mileage between 20 mpg and 30 mpg can also be estimated as the shaded area under the curve.



By definition...

Since continuous probabilities are estimated as “the area under the curve”, the probability of a car having a mileage of exactly 25 mpg (or any exact value) is defined as 0.