CAAP Statistics - Lec06

Jul 14, 2022

Review

- Data manipulation using R
- Reproducible report using R—R Markdown
- Playing with bin width of histogram
- Boxplot and five number summary
- Scatterplot

Learning Objectives

- Definition of Probability
- Probability Distribution
- Sampling from a small population
- Random Variables
- Continuous distributions

Introduction to 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
  - You already know the possible outcome of rolling a dice : 1,2,3,4,5,6
  - However, you don’t know the exact outcome unless you rolled a dice.
- It can be helpful to model a process as random even if it is not truly random.

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.

Note: In more general sense, *Law of large numbers* states that the sample mean converges to the population mean.

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?

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

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?

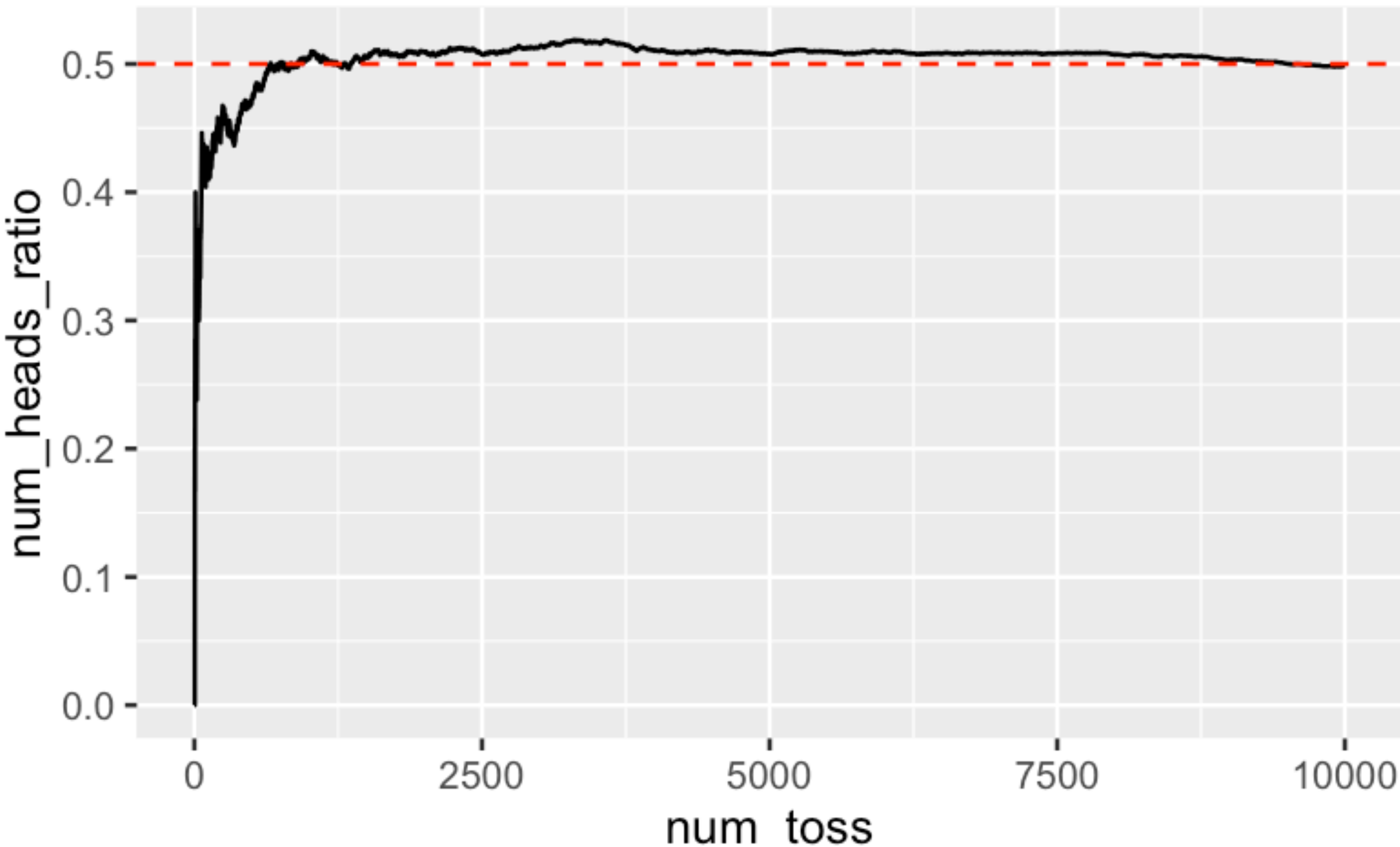
$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- 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*).

Law of large numbers (cont.)



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?

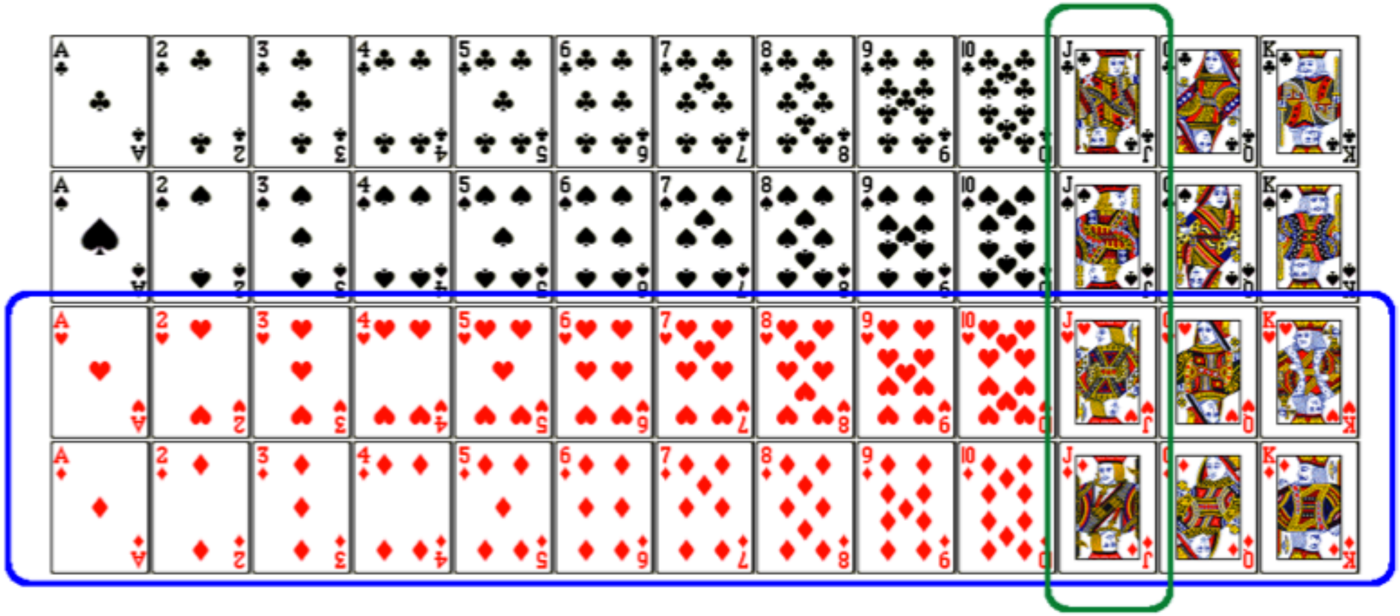
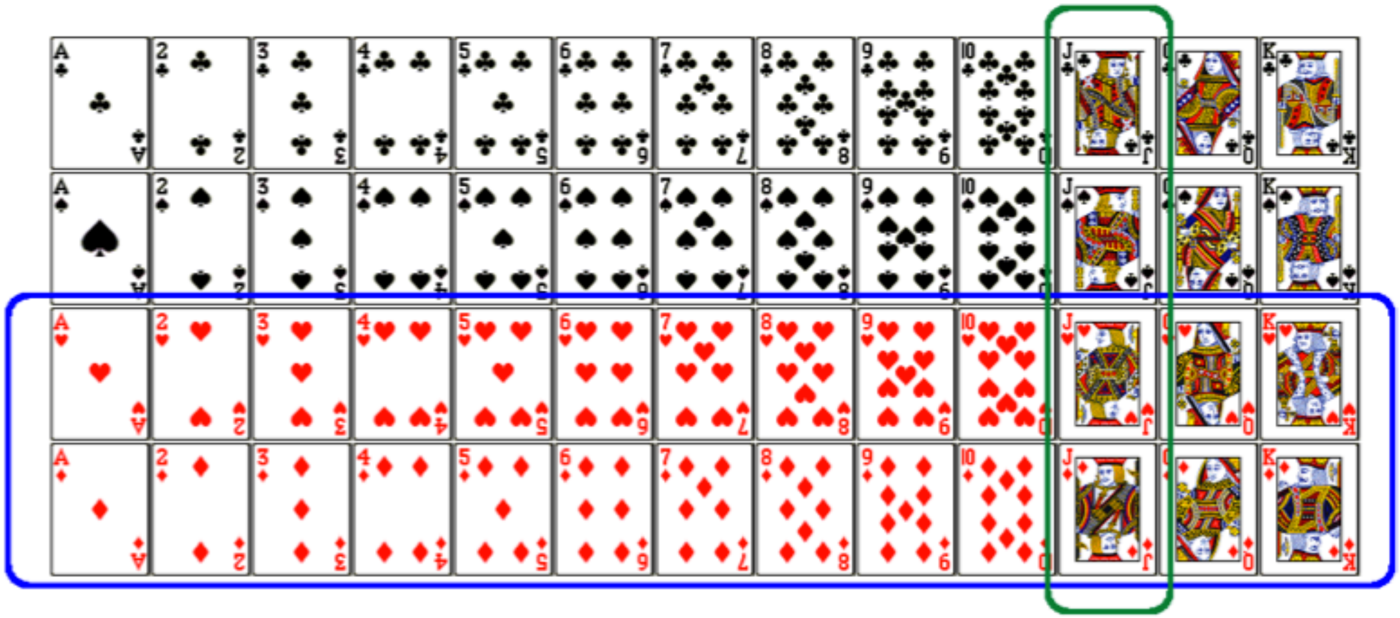


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Union of non-disjoint events

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



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

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

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

Practice

What is the probability that a randomly sampled student thinks marijuana should be legalized or they agree with their parents' political views?

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes	36	78	114
Total	47	118	165

(a) (40 + 36 - 78) / 165

(b) (114 + 118 - 78) / 165

(c) 78 / 165

(d) 78 / 188

(e) 11 / 47

Practice

What is the probability that a randomly sampled student thinks marijuana should be legalized or they agree with their parents' political views?

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes	36	78	114
Total	47	118	165

(a) (40 + 36 - 78) / 165

(b) (114 + 118 - 78) / 165

(c) 78 / 165

(d) 78 / 188

(e) 11 / 47

Recap: Addition Rule

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 sum of rolling two dices:

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Figure 3.5: Probability distribution for the sum of two dice.

Probability distributions

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

- The probability distribution for sum of rolling two dices:

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Figure 3.5: Probability distribution for the sum of two dice.

- 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

Practice

In a survey, 52% of respondents said they are Democrats. What is the probability that a randomly selected respondent from this sample is a Republican?

(a) 0.48

(b) more than 0.48

(c) less than 0.48

(d) cannot calculate using only the information given

Practice

In a survey, 52% of respondents said they are Democrats. What is the probability that a randomly selected respondent from this sample is a Republican?

(a) 0.48

(b) more than 0.48

(c) less than 0.48

(d) cannot calculate using only the information given

If the only two political parties are Republican and Democrat, 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. Then (c) is also possible.

Sample space and complements

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

- You flip a coin. What is the sample space? S = {H, T}
- You flip two coins. What is the sample space?

Sample space and complements

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

- You flip a coin. What is the sample space? S = {H, T}
- You flip two coins. What is the sample space?

  S = {HH, HT, TH, TT}

Sample space and complements

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

- You flip a coin. What is the sample space? S = {H, T}
- You flip two coins. What is the sample space?

  S = {HH, HT, TH, TT}

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

- You flip a coin. If we know that it was not head, what is result of flipping coin? { ~~H~~, *T* } Head and tail are *complementary* outcomes.
- You flip two coins. If we know that they are not both heads, what are the possible outcome?

Sample space and complements

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

- You flip a coin. What is the sample space? S = {H,T}
- You flip two coins. What is the sample space?

  S = {HH, HT, TH, TT}

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

- You flip a coin. If we know that it was not head, what is result of flipping coin? { ~~H~~, *T* } Head and tail are *complementary* outcomes.
- You flip two coins. If we know that they are not both heads, what are the possible outcome?

  S = { ~~HH~~, HT, TH, TT }

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.
  >> 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.
  >> 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.
  >> Outcomes of two draws from a deck of cards (without replacement) are dependent.

Practice

In parts (a) and (b), identify whether the events are disjoint, independent or neither (events cannot be both disjoint and independent).

(a) You and a randomly selected student from your class both earn A's in this course.

(b) You and your class study partner both earn A's in this course.

(c) If two events can occur at the same time, must they be dependent?

Practice

In parts (a) and (b), identify whether the events are disjoint, independent or neither (events cannot be both disjoint and independent).

(a) You and a randomly selected student from your class both earn A's in this course. *Independent*

(b) You and your class study partner both earn A's in this course. *Likely to be dependent*

(c) If two events can occur at the same time, must they be dependent? *Independent*

Product rule for independent events

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

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

Product rule for independent events

P(A and B) = P(A) x P(B)

Or more generally, $P(A_1$, and, … and $A_k) = P(A_1) \times \ldots \times P(A_k)$

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

Product rule for independent events

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

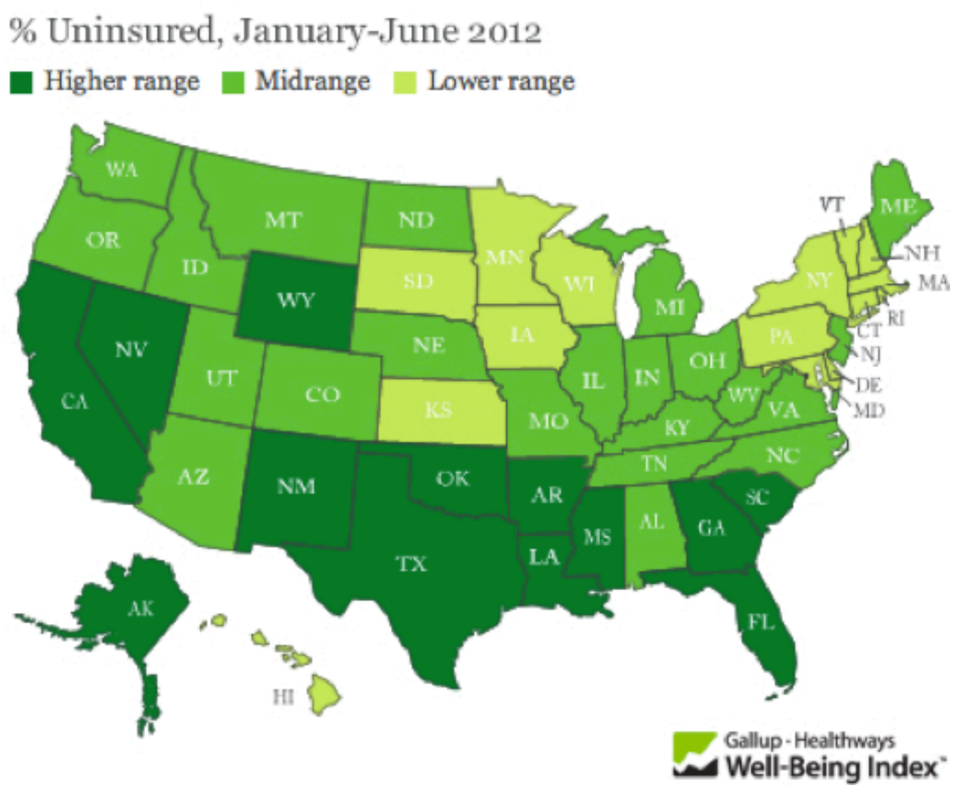
Or more generally, $P(A_1, \text{ and}, \ldots \text{ and } A_k) = P(A_1) \times \ldots \times P(A_k)$

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

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

Practice

A recent Gallup poll suggests that 25.5% of Texans do not have health insurance as of June 2012. Assuming that the uninsured rate stayed constant, what is the probability that two randomly selected Texans are both uninsured?



(a) 25.5^2

(b) 0.255^2

(c) 0.255×2

(d) $(1 - 0.255)^2$

http://www.gallup.com/poll/156851/uninsured-rate-stable-across-states-far-2012.aspx

Practice

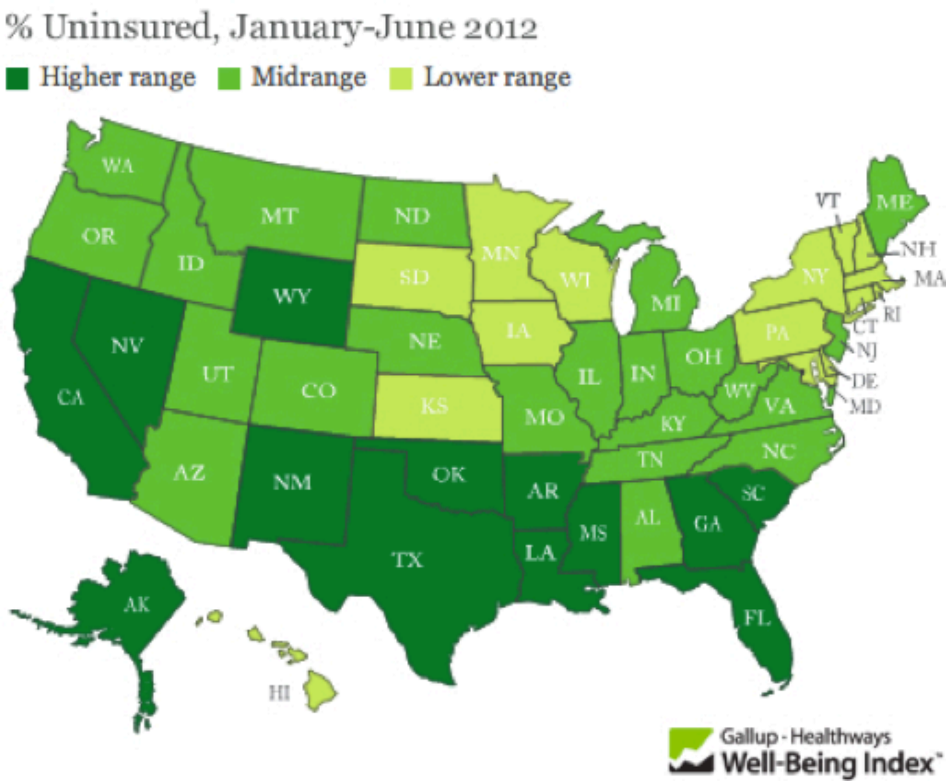
A recent Gallup poll suggests that 25.5% of Texans do not have health insurance as of June 2012. Assuming that the uninsured rate stayed constant, what is the probability that two randomly selected Texans are both uninsured?

(a) 25.5^2

(b) 0.255^2

(c) 0.255×2

(d) $(1 - 0.255)^2$



http://www.gallup.com/poll/156851/uninsured-rate-stable-across-states-far-2012.aspx

Disjoint vs. complementary

Do the sum of probabilities of two disjoint events always add up to 1?

Disjoint vs. complementary

Do the sum of probabilities of two disjoint events always add up to 1?

Not necessarily, there may be more than 2 events in the sample space, e.g. party affiliation.

Disjoint vs. complementary

Do the sum of probabilities of two disjoint events always add up to 1?

Not necessarily, there may be more than 2 events in the sample space, e.g. party affiliation.

Do the sum of probabilities of two complementary events always add up to 1?

Disjoint vs. complementary

Do the sum of probabilities of two disjoint events always add up to 1?

Not necessarily, there may be more than 2 events in the sample space, e.g. party affiliation.

Do the sum of probabilities of two complementary events always add up to 1?

Yes, that's the definition of complementary, e.g. heads and tails

Putting everything together...

If we were to randomly select 5 Texans, what is the probability that at least one is uninsured?}

- If we were to randomly select 5 Texans, the sample space for the number of Texans who are uninsured would be:

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

- We are interested in instances where at least one person is uninsured:

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

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

  S = {0, at least one}

Putting everything together...

Since the probability of the sample space must add up to 1:

P(at least 1 uninsured)

= 1 - *P(none uninsured)*

Putting everything together...

Since the probability of the sample space must add up to 1:

P(at least 1 uninsured)

= 1 - *P(none uninsured)*

= $1 - (1 - 0.255)^5$

Putting everything together...

Since the probability of the sample space must add up to 1:

P(at least 1 uninsured)

$= 1 - P(\text{none uninsured})$

$= 1 - (1 - 0.255)^5$

$= 1 - 0.745^5$

Putting everything together...

Since the probability of the sample space must add up to 1:

P(at least 1 uninsured)

= 1 - *P(none uninsured)*

= $1 - (1 - 0.255)^5$

= $1 - 0.745^5$

= 1 - 0.23

Putting everything together...

Since the probability of the sample space must add up to 1:

P(at least 1 uninsured)

$= 1 - P(\text{none uninsured})$

$= 1 - (1 - 0.255)^5$

$= 1 - 0.745^5$

$= 1 - 0.23$

$= 0.77$

Putting everything together...

Since the probability of the sample space must add up to 1:

P(at least 1 uninsured)

$= 1 - P(\text{none uninsured})$

$= 1 - (1 - 0.255)^5$

$= 1 - 0.745^5$

$= 1 - 0.23$

$= 0.77$

At least 1:

$P(\text{at least one}) = 1 - P(\text{none})$

Practice

Roughly 20% of undergraduates at a university are vegetarian or vegan. What is the probability that, among a random sample of 3 undergraduates, at least one is vegetarian or vegan?

(a) $1 - 0.2 \times 3$

(b) $1 - 0.2^3$

(c) 0.8^3

(d) $1 - 0.8 \times 3$

(e) $1 - 0.8^3$

Practice

Roughly 20% of undergraduates at a university are vegetarian or vegan. What is the probability that, among a random sample of 3 undergraduates, at least one is vegetarian or vegan?

(a) $1 - 0.2 \times 3$

(b) $1 - 0.2^3$

(c) 0.8^3

(d) $1 - 0.8 \times 3$

(e) $1 - 0.8^3$

P(at least 1 from veg)

= 1 - P(none veg)

$= 1 - 0.8^3$

$= 1 - 0.512 = 0.488$

Conditional Probability

Relapse

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below.

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

Marginal probability

What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Marginal probability

What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed) = 48 / 72 ~ 0.67

Joint probability

What is the probability that a patient received the antidepressant (desipramine) and relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Joint probability

What is the probability that a patient received the antidepressant (desipramine) and relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed and desipramine) = 10 / 72 ~ 0.14

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

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

$$P(relapse|desipramine) = \frac{P(relapse \text{ and } desipramine)}{P(desipramine)}$$

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Conditional probability

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

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

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(relapse|desipramine) = \frac{P(relapse \text{ and } desipramine)}{P(desipramine)} = \frac{10/72}{24/72}$$

Conditional probability

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

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

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(relapse|desipramine) = \frac{P(relapse \text{ and } desipramine)}{P(desipramine)} = \frac{10/72}{24/72} = \frac{10}{24}$$

Conditional probability

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

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

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(relapse|desipramine)$$
$$= \frac{P(relapse \text{ and } desipramine)}{P(desipramine)}$$
$$= \frac{10/72}{24/72}$$
$$= \frac{10}{24}$$
$$= 0.42$$

Conditional probability (cont.)

If we know that a patient received the antidepressant (desipramine), what is the probability that they relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Conditional probability (cont.)

If we know that a patient received the antidepressant (desipramine), what is the probability that they relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapse | desipramine) = 10 / 24 ~ 0.42

Conditional probability (cont.)

If we know that a patient received the antidepressant (desipramine), what is the probability that they relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$P(relapse \mid desipramine) = 10 / 24 \sim 0.42$

$P(relapse \mid lithium) = 18 / 24 \sim 0.75$

$P(relapse \mid placebo) = 20 / 24 \sim 0.83$

Conditional probability (cont.)

If we know that a patient relapsed, what is the probability that they received the antidepressant (desipramine)?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Conditional probability (cont.)

If we know that a patient relapsed, what is the probability that they received the antidepressant (desipramine)?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$P(desipramine \mid relapse) = 10 / 48 \sim 0.21$

Conditional probability (cont.)

If we know that a patient relapsed, what is the probability that they received the antidepressant (desipramine)?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(desipramine | relapse) = 10 / 48 ~ 0.21

P(lithium | relapse) = 18 / 48 ~ 0.38

P(placebo | relapse) = 20 / 48 ~ 0.42

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

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

  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

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

  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 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 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 60 / 100 = 0.6.
- The probability that a randomly selected student is a social science major given that they are female is 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 60 / 100 = 0.6.
- The probability that a randomly selected student is a social science major given that they are female is 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 $P(A \mid B) = P(A)$ then the events *A* and *B* are said to be independent.

Independence and conditional probabilities (cont.)

Generically, if $P(A \mid B) = P(A)$ 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 $P(A \mid B) = P(A)$ 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, $P(A \text{ and } B) = P(A) \times P(B)$. Then,

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

Breast cancer screening

- American Cancer Society estimates that about 1.7% of women have breast cancer. *http://www.cancer.org/cancer/cancerbasics/cancer-prevalence*
- Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer. *http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html*
- An article published in 2003 suggests that up to 10% of all mammograms result in false positives for patients who do not have cancer. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940*

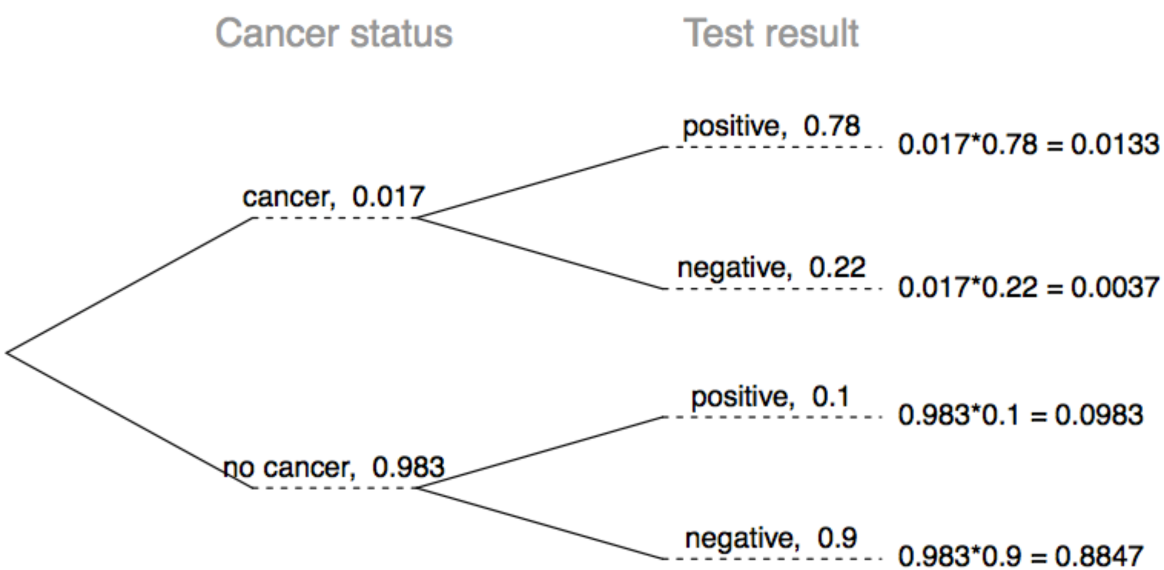
Note: These percentages are approximate, and very difficult to estimate.

Inverting probabilities

When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?

Inverting probabilities

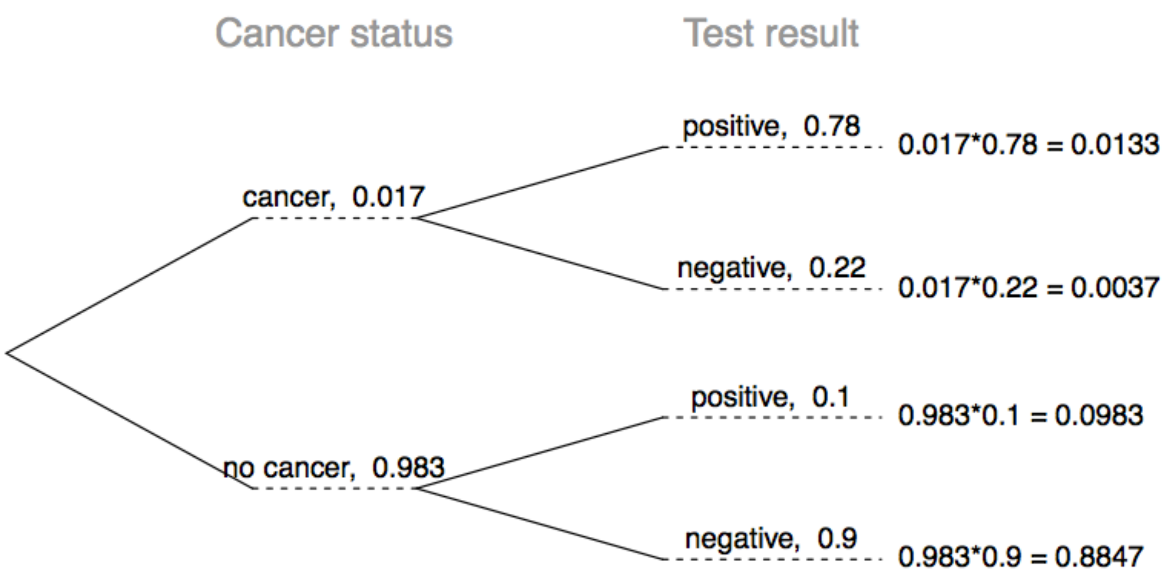
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



Inverting probabilities

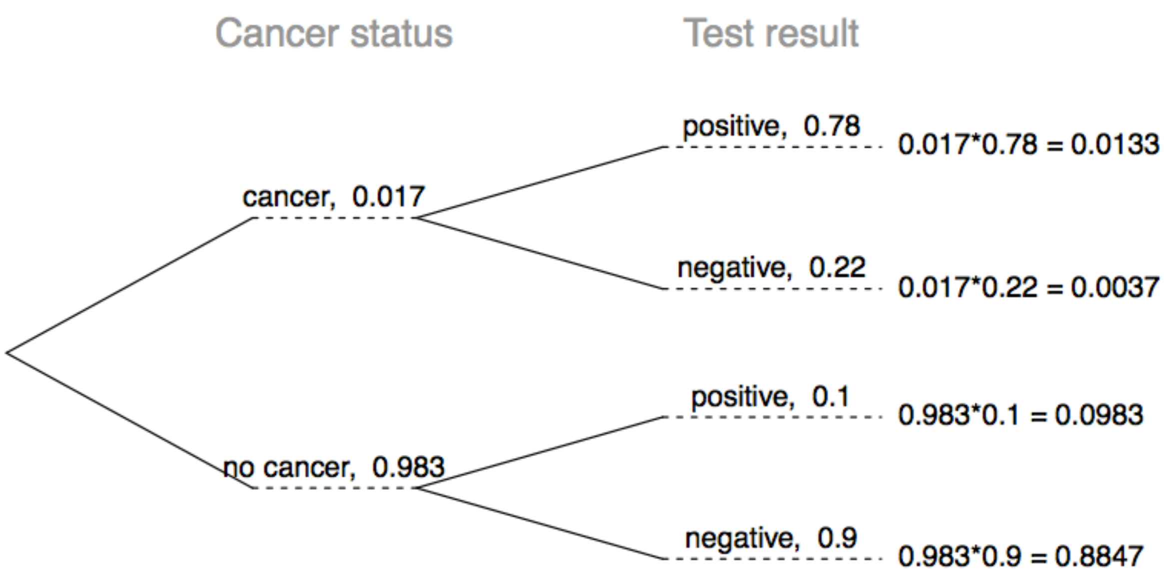
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?

$$P(C|+)$$



Inverting probabilities

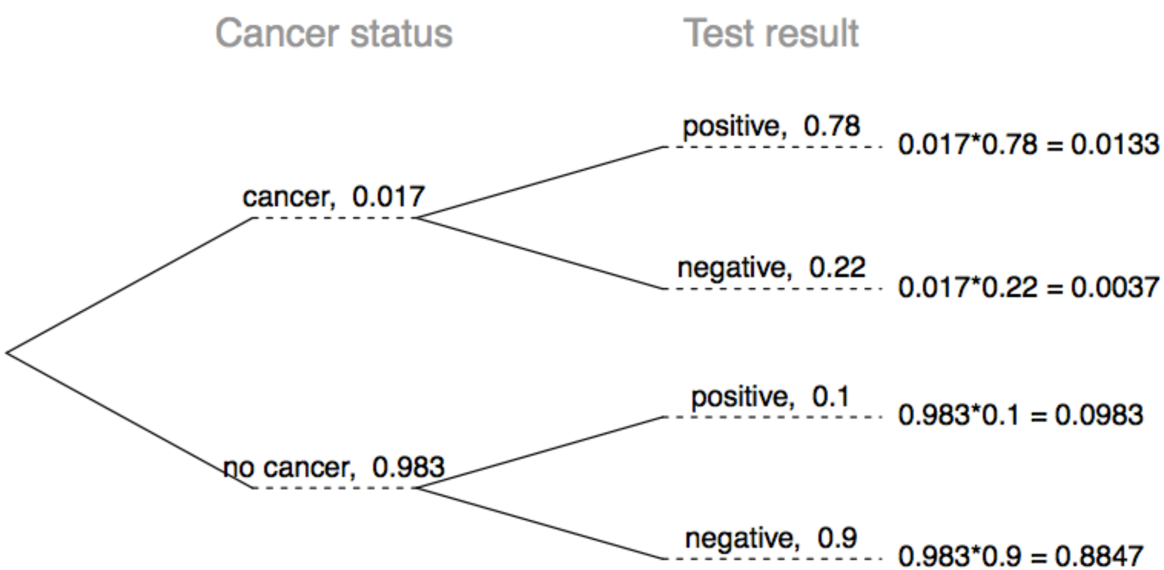
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)}$$

Inverting probabilities

When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?

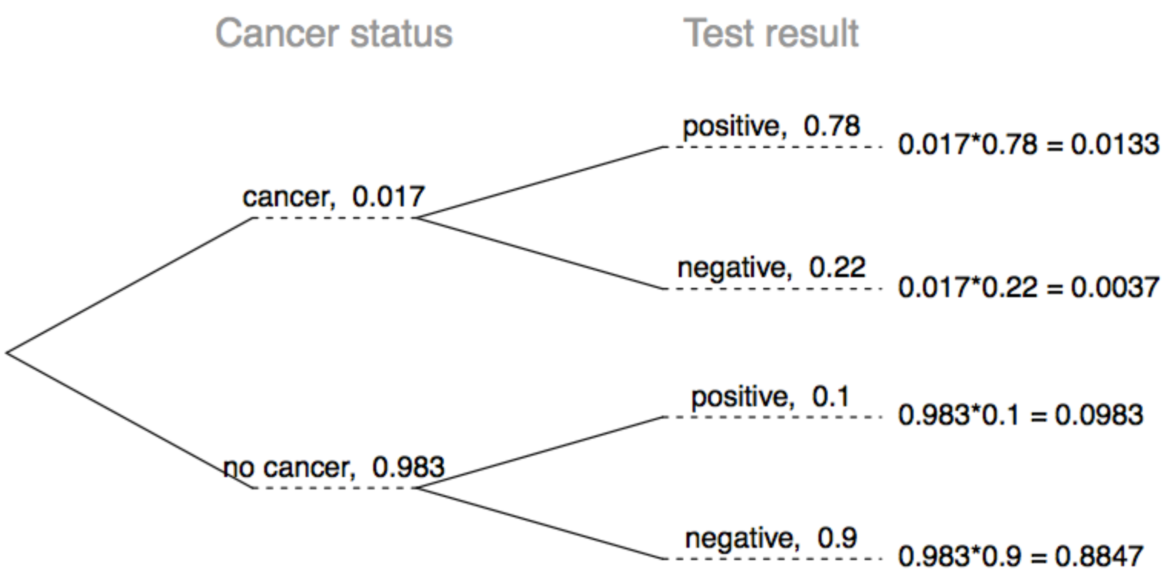


$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)} = \frac{0.0133}{0.0133 + 0.0983}$$

Note: Tree diagrams are useful for inverting probabilities: we are given P(+|C) and asked for P(C|+).

Inverting probabilities

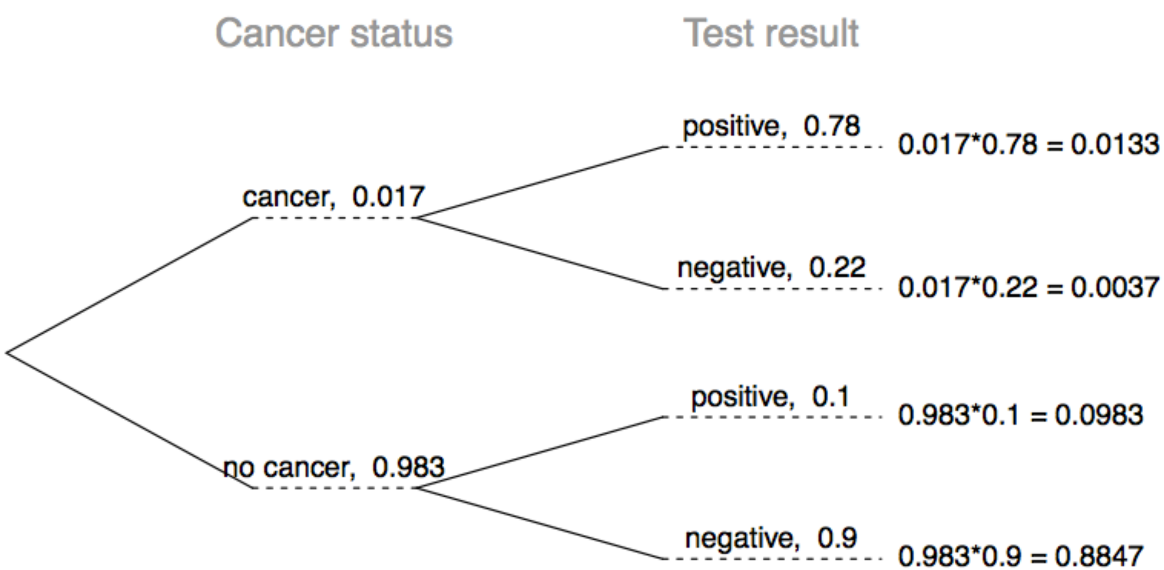
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)} = \frac{0.0133}{0.0133 + 0.0983} = 0.12$$

Inverting probabilities

When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)} = \frac{0.0133}{0.0133 + 0.0983} = 0.12$$

Note: Tree diagrams are useful for inverting probabilities: we are given P(+|C) and asked for P(C|+).

Practice

Suppose a woman who gets tested once and obtains a positive result wants to get tested again. In the second test, what should we assume to be the probability of this specific woman having cancer?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Practice

Suppose a woman who gets tested once and obtains a positive result wants to get tested again. In the second test, what should we assume to be the probability of this specific woman having cancer?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Practice

What is the probability that this woman has cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52

Practice

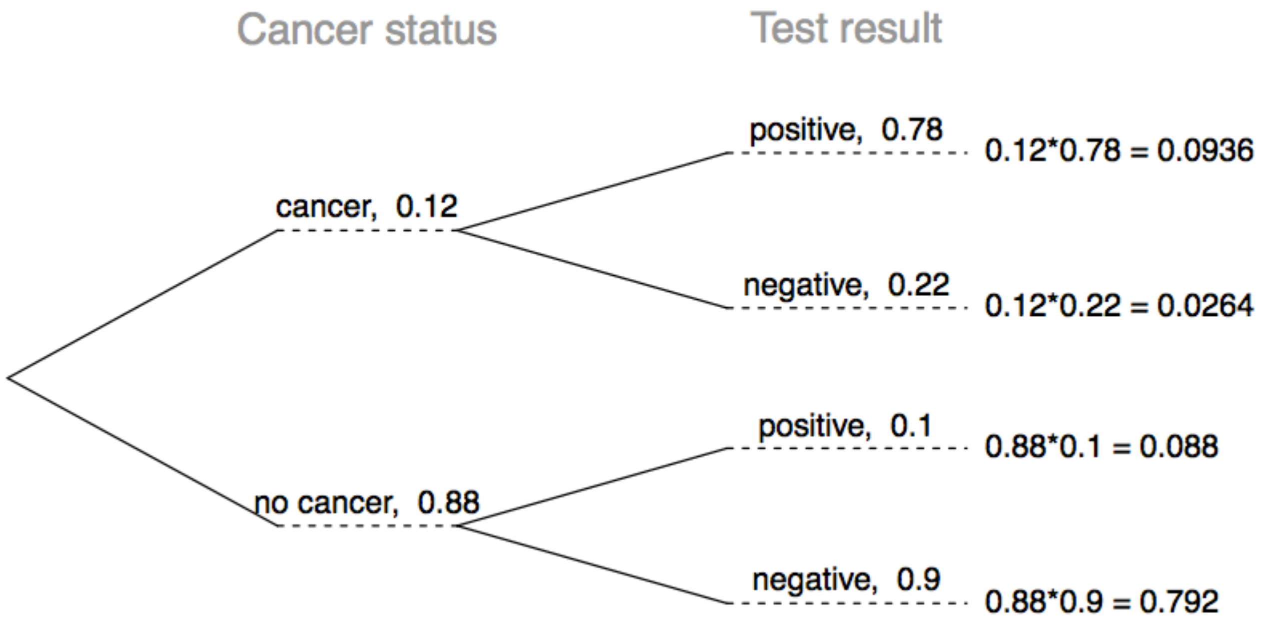
What is the probability that this woman has cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52



Practice

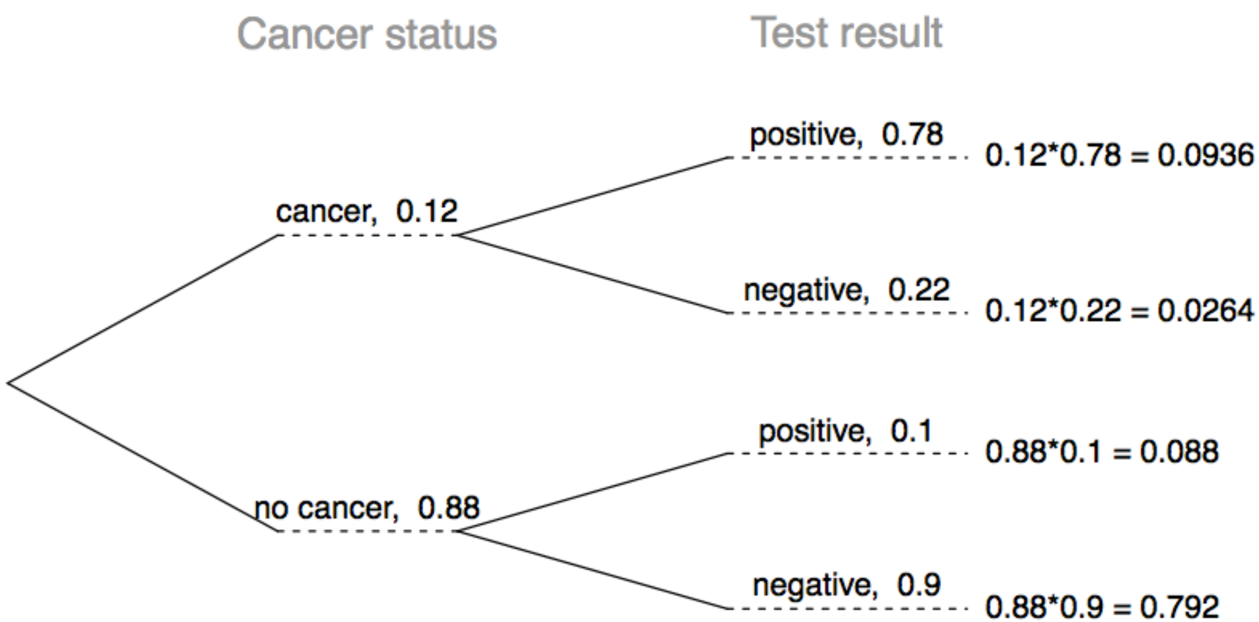
What is the probability that this woman has cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52



$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)} = \frac{0.0936}{0.0936 + 0.088} = 0.52$$

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

$$P(outcome\ A\ of\ variable\ 1 \mid outcome\ B\ of\ variable\ 2)$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

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

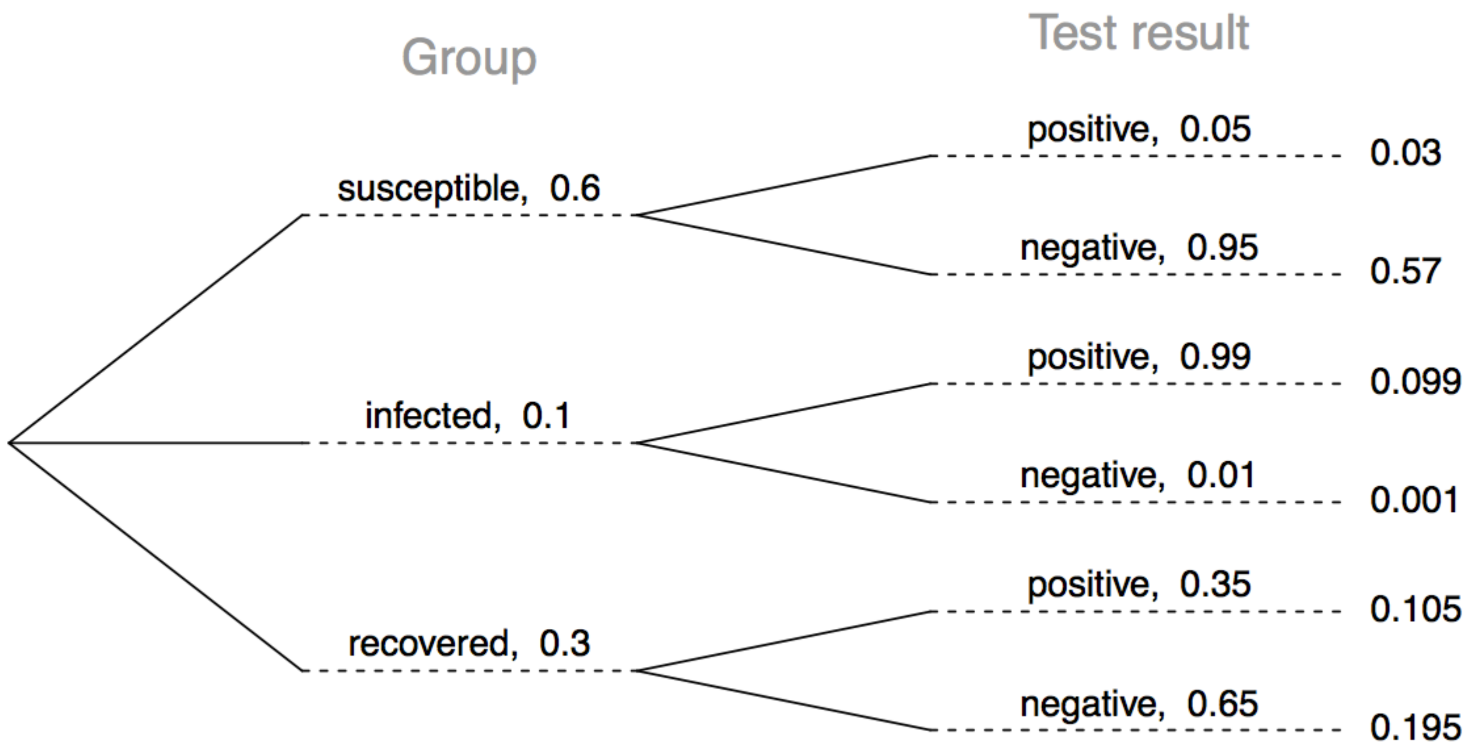
Application activity: inverting probabilities

A common epidemiological model for the spread of diseases is the SIR model, where the population is partitioned into three groups: Susceptible, Infected, and Recovered. This is a reasonable model for diseases like chickenpox where a single infection usually provides immunity to subsequent infections. Sometimes these diseases can also be difficult to detect.

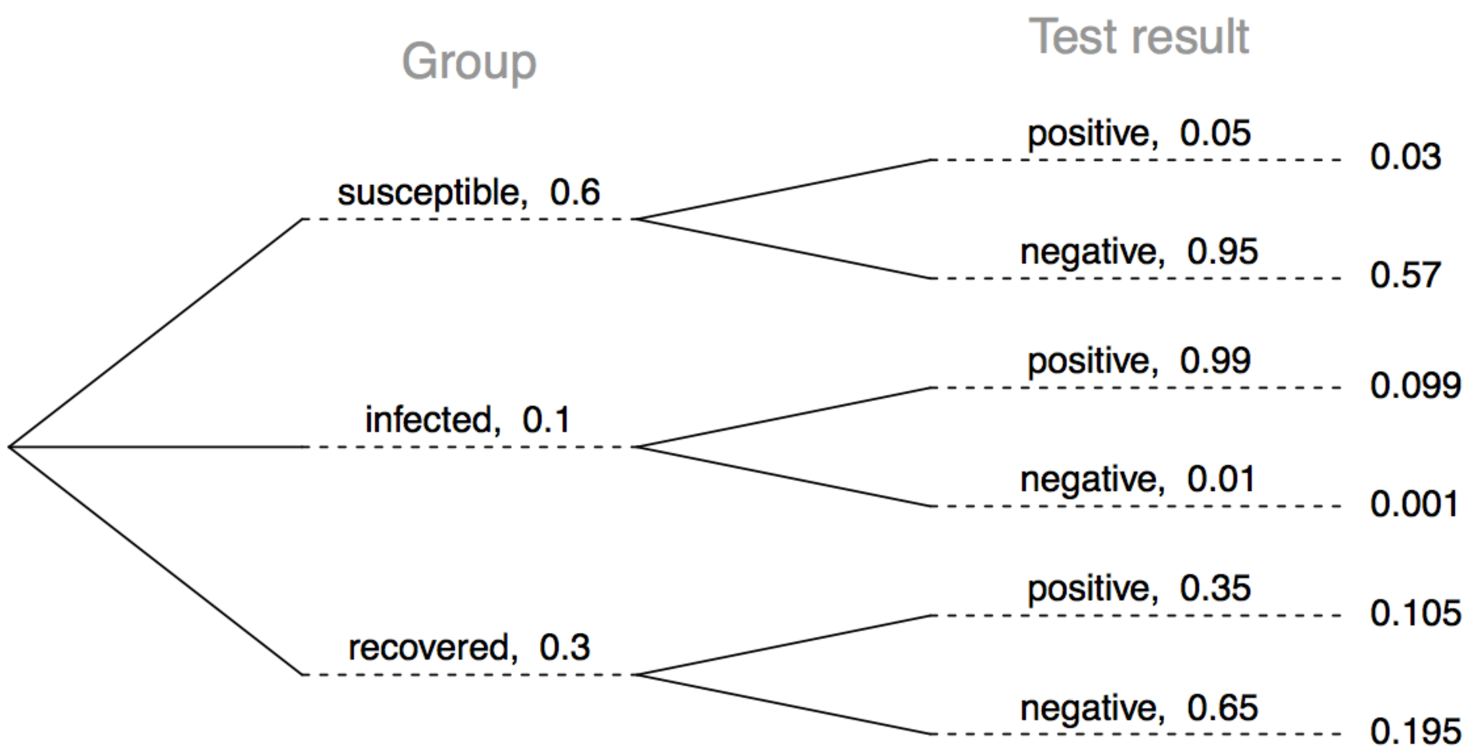
Imagine a population in the midst of an epidemic where 60% of the population is considered susceptible, 10% is infected, and 30% is recovered. The only test for the disease is accurate 95% of the time for susceptible individuals, 99% for infected individuals, but 65% for recovered individuals. (Note: In this case accurate means returning a negative result for susceptible and recovered individuals and a positive result for infected individuals).

Draw a probability tree to reflect the information given above. If the individual has tested positive, what is the probability that they are actually infected?

Application activity: inverting probabilities (cont.)



Application activity: inverting probabilities (cont.)



$$P(inf|+) = \frac{P(inf\ and\ +)}{P(+)} = \frac{0.099}{0.03 + 0.099 + 0.105} \approx 0.423$$

Sampling from a Small Population

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 O, 3 O, 2 O

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

- 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?

1st Draw - 5 O, 3 O, 2 O

2nd Draw - 5 O, 3 O, 2 O

P(2nd chip B | 1st chip B) = 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?

1st Draw - 5 **O**, 3 **O**, 2 **O**

2nd Draw - 5 **O**, 3 **O**, 2 **O**

P(2nd chip B | 1st chip O) = 3 / 10 = 0.3

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

1st Draw - 5 **O**, 3 **O**, 2 **O**

2nd Draw - 5 **O**, 3 **O**, 2 **O**

P(1st chip B) x P(2nd chip B | 1st chip B) = 0.3 x 0.3 = 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.

$$Prob(B \mid B) = Prob(B \mid O)$$

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.

$$Prob(B \mid B) = Prob(B)$$

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?

1st Draw - 5 O, 3 O, 2 O

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?

1st Draw - 5 **O**, 3 **O**, 2 **O**

2nd Draw - 5 **O**, 2 **O**, 2 **O**

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?

1st Draw - 5 **O**, 3 **O**, 2 **O**

2nd Draw - 5 **O**, <u>2</u> **O**, 2 **O**

P(2nd chip B | 1st chip B) = 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?

1st Draw - 5 O, 3 O, 2 O

2nd Draw - 5 O, 2 O, 2 O

P(2nd chip B | 1st chip B) = 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?

1st Draw - 5 O, 3 O, 2 O

2nd Draw - 5 O, 2 O, 2 O

P(2nd chip B | 1st chip B) = 2 / 9 = 0.22

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

1st Draw - 5 O, 3 O, 2 O

2nd Draw - 5 O, 2 O, 2 O

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?

1st Draw - 5 O, 3 O, 2 O

2nd Draw - 5 O, 2 O, 2 O

P(2nd chip B | 1st chip B) = 2 / 9 = 0.22

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

1st Draw - 5 O, 3 O, 2 O

2nd Draw - 5 O, 2 O, 2 O

P(1st chip B) x P(2nd chip B | 1st chip B) = 0.3 x 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.

$$Prob(B \mid B) \neq Prob(B)$$

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) (4/52) x (3/52)

(b) (3/52) x (4/52)

(c) (4/52) x (3/51)

(d) (4/52) x (2/51)

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) (4/52) x (3/52)

(b) (3/52) x (4/52)

(c) (4/52) x (3/51)

(d) (4/52) x (2/51)

P(ace, then 3) = (4 / 52) x (4 / 51) ~ 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/observations 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 \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 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.

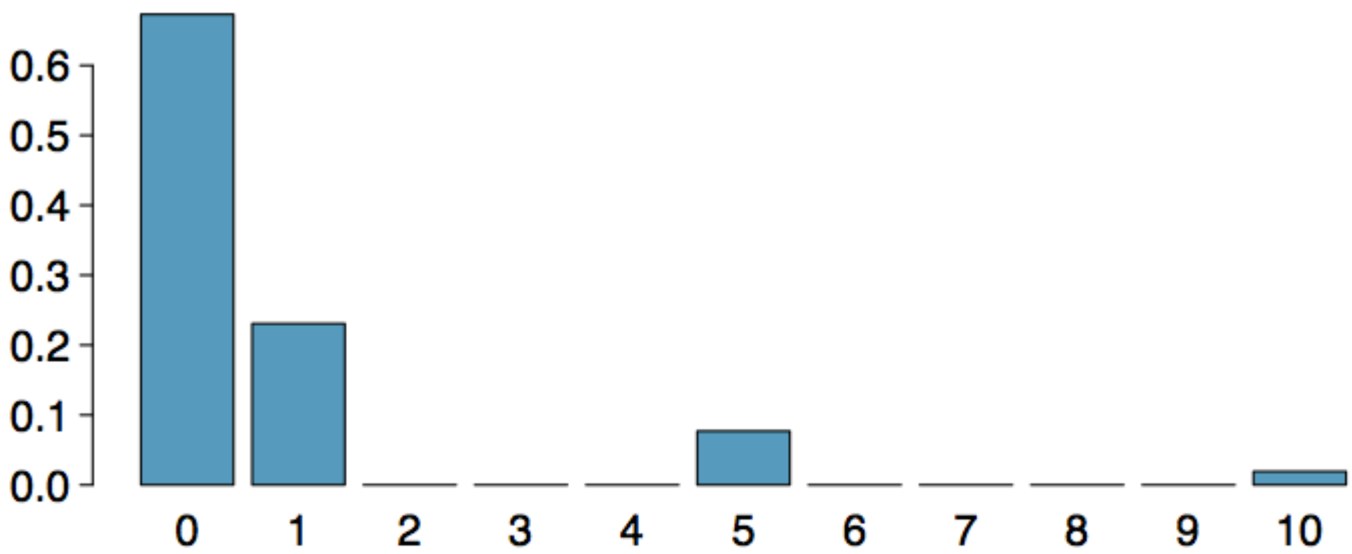
Expected value of a discrete random variable

In a game of cards you win $1 if you draw a heart, $5 if you draw an ace (including the ace of hearts), $10 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	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

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	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		

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	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$
				$SD(X) = \sqrt{3.4246} = 1.85$

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 chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics 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 chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

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

Linear Combination

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 Combination

- 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.

Linear combinations

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and chemistry homework for the week if you have 5 statistics and 4 chemistry homework problems assigned?

Linear combinations

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and chemistry homework for the week if you have 5 statistics and 4 chemistry homework problems assigned?

$$V(S + S + S + S + S + C + C + C + C)$$
$$= V(S) + V(S) + V(S) + V(S) + V(S) + V(C) + V(C) + V(C) + V$$
$$= 5 \times V(S) + 4 \times V(C)$$
$$= 5 \times 1.5^2 + 4 \times 2^2$$
$$= 27.25$$

Looking for a complete explanation? There's a discussion in the OpenIntro Forums with a detailed explanation for why we aren't squaring the 5 and 4 in the equation above. [link]

Practice

A casino game costs $5 to play. If you draw first a red card, then you get to draw a second card. If the second card is the ace of hearts, you win $500. If not, you don't win anything, i.e. lose your $5. What is your expected profits (or losses) from playing this game? Remember: profit (or loss) = winnings - cost.

(a) a loss of 10¢

(b) a loss of 25¢

(c) a loss of 30¢

(d) a profit of 5¢

Practice

A casino game costs \$5 to play. If you draw first a red card, then you get to draw a second card. If the second card is the ace of hearts, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits (or losses) from playing this game? Remember: profit (or loss) = winnings - cost.

(a) a loss of 10¢

(b) a loss of 25¢

(c) a loss of 30¢

(d) a profit of 5¢

Event	*Win*	*Profit: X*	$P(X)$	$X \times P(X)$
Red, A♥	*500*	*500 - 5 = 495*	$\frac{25}{52} \times \frac{1}{51} = 0.0094$	$495 \times 0.0094 = 4.653$
Other	*0*	*0 - 5 = -5*	$1 - 0.0094 = 0.9906$	$-5 \times 0.9906 = -4.953$
				$E(X) = -0.3$

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:

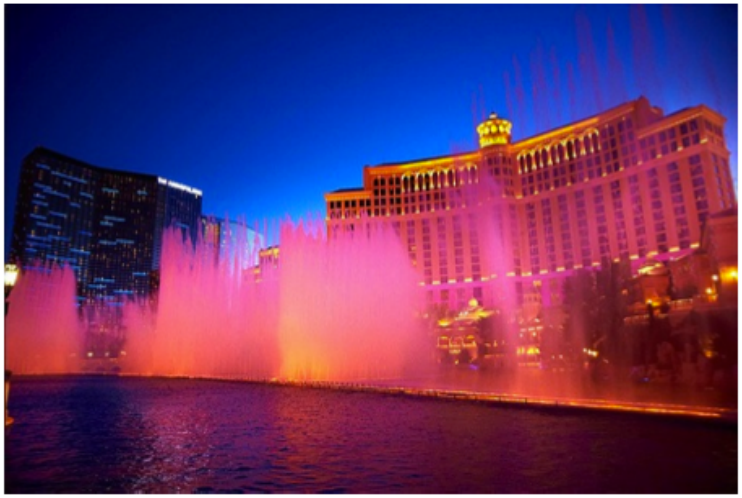


Image by Moyan_Brenn on Flickr
http://www.flickr.com/photos/aigle_dore/5951714693

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

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 company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

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).

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

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

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 ($5X$).

$$\begin{aligned} E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\ &= 5 \times E(X) = 5 \times 2,154 = \$10,770 \end{aligned}$$

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

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).

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770
\end{aligned}
$$

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

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

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 ($5X$).

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770
\end{aligned}
$$

$$
\begin{aligned}
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 132^2 = \$87,120
\end{aligned}
$$

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

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 ($5X$).

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770
\end{aligned}
$$

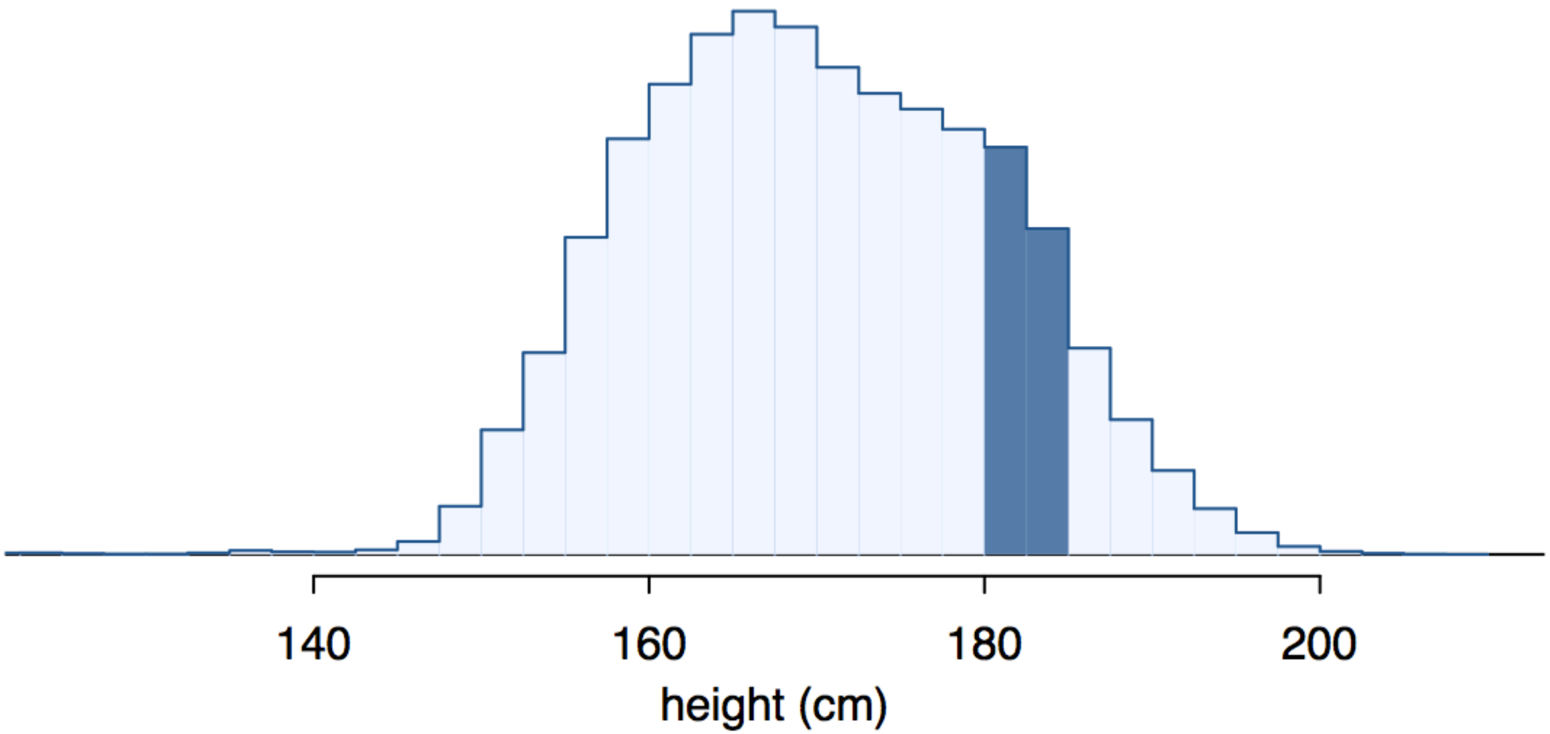
$$
\begin{aligned}
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 132^2 = \$87,120
\end{aligned}
$$

$$
SD(X_1 + X_2 + X_3 + X_4 + X_5) = \sqrt{87,120} = 295.16
$$

Continuous Distributions

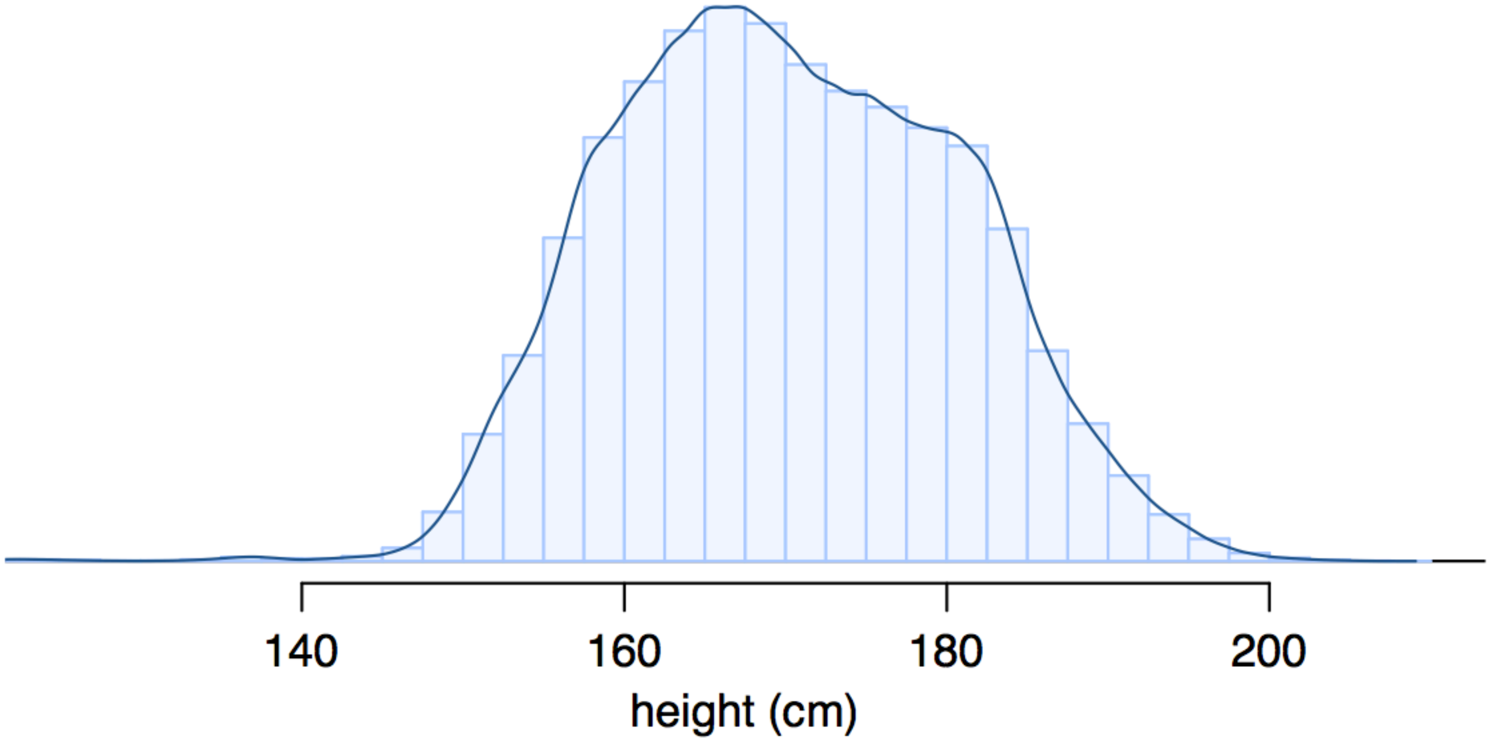
Continuous distributions

- Below is a histogram of the distribution of heights of US adults.
- The proportion of data that falls in the shaded bins gives the probability that a randomly sampled US adult is between 180 cm and 185 cm (about 5'11" to 6'1").



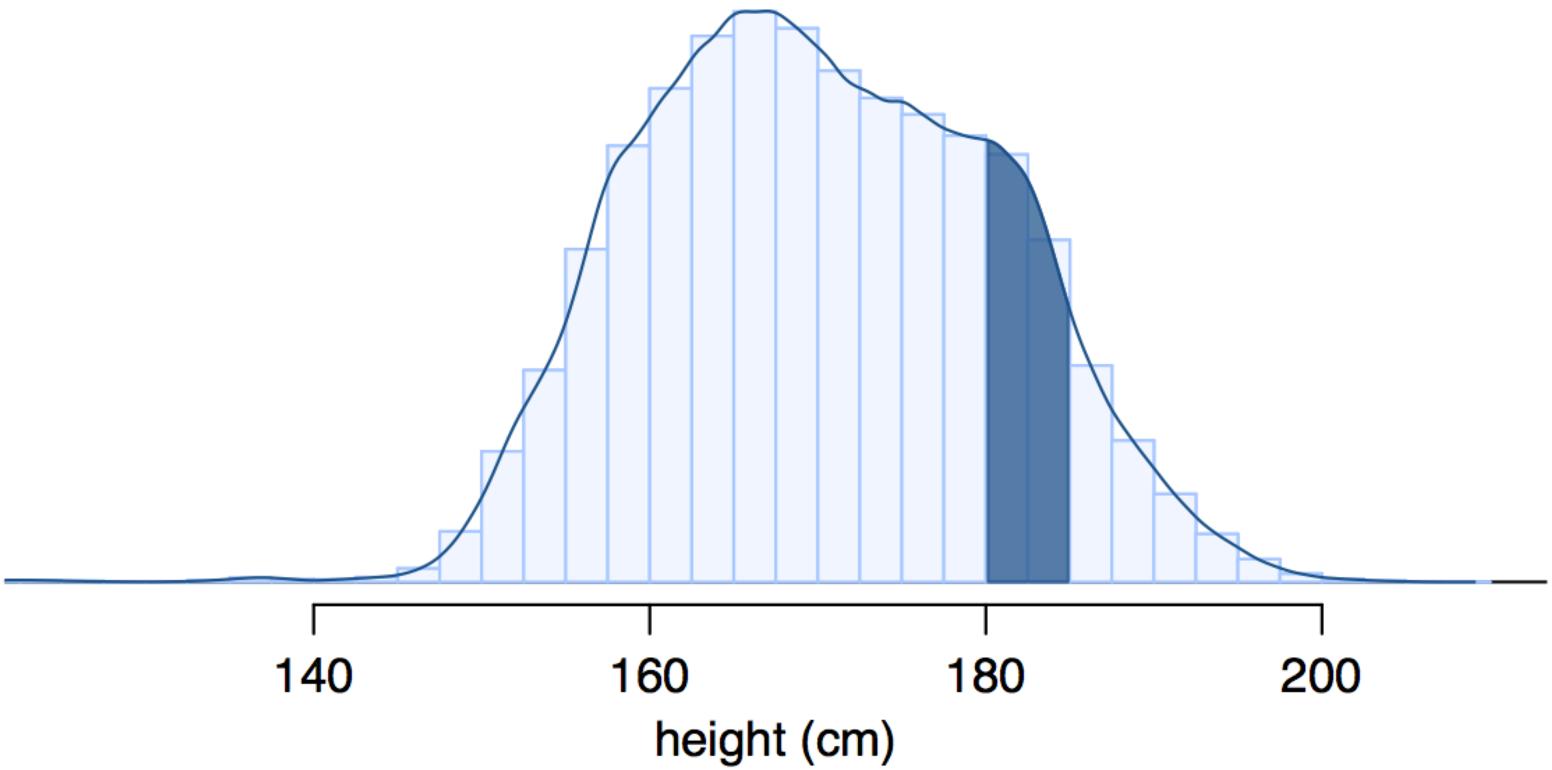
From histograms to continuous distributions

Since height is a continuous numerical variable, its probability density function is a smooth curve.



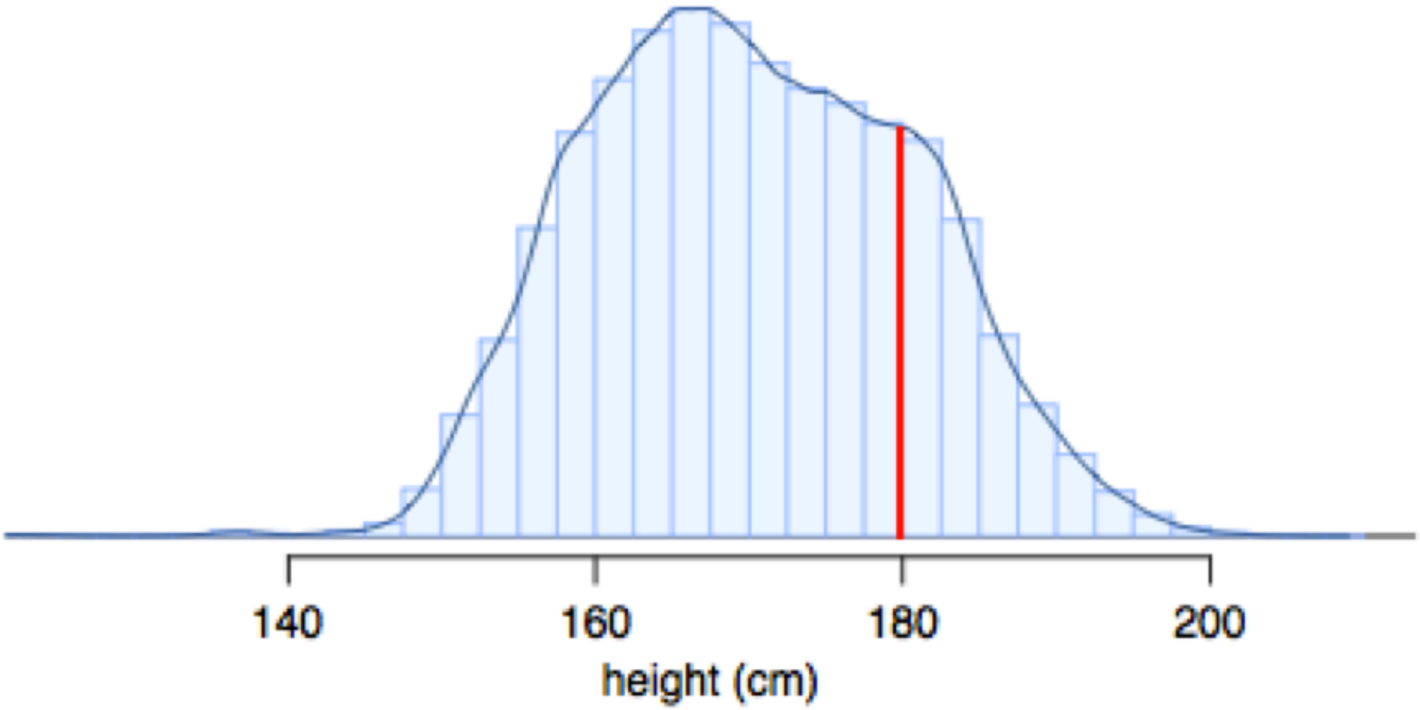
Probabilities from continuous distributions

Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm 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 person being exactly 180 cm (or any exact value) is defined as 0.



Let’s discuss!

Health coverage, frequencies

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health	No	459	727	854	385	99	2,524
Coverage	Yes	4,198	6,245	4,821	1,634	578	17,476
	Total	4,657	6,972	5,675	2,019	677	20,000

(a) If we draw one individual at random, what is the probability that the respondent has an excellent health and doesn't have health coverage?

(b) If we draw one individual at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

Thursday is R Session!

- Don't forget to bring laptop with you.
  - Continue from R Session2 - Graphical Summary for Categorical Variable
  - Probability Simulation
- Office hour from **7pm** via Zoom.
  - **This week, make sure to participate at least one office hour!!**
  - 5 out of 50 participation pts