Statistics

Marco Caserta
marco.caserta@ie.edu

IE University

1. Definition of Probability
2. Conditional 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.

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?

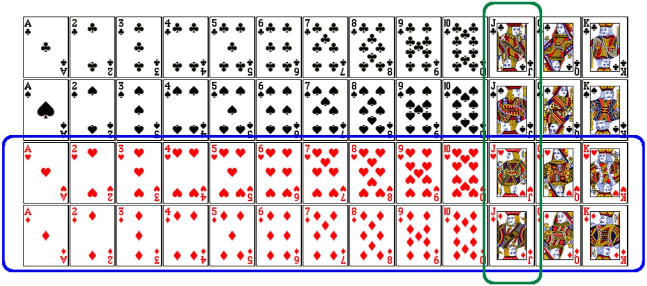
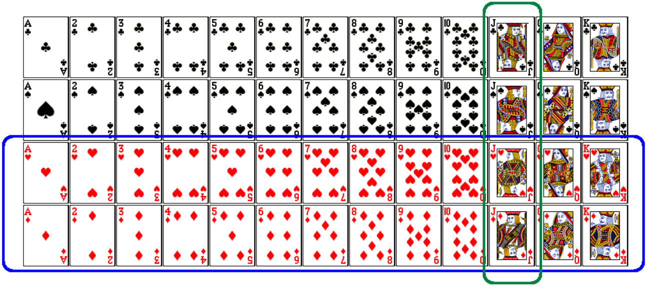


Figure from http://www.milefoot.com/math/discrete/counting/cardfreq.htm.

Union of non-disjoint events

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



$$P(jack\ or\ red) = P(jack) + P(red) - P(jack\ and\ red)$$
$$= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$

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

Recap

General addition rule

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

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 the gender of one kid:

Event	Male	Female
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 the gender of one kid:

Event	Male	Female
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 the gender of one kid:

Event	Male	Female
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 the genders of two kids:

Event	*MM*	*FF*	*MF*	*FM*
Probability	*0.25*	*0.25*	*0.25*	*0.25*

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. 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.

- A couple has one kid, what is the sample space for the gender of this kid? $S = \{M, F\}$
- A couple has two kids, what is the sample space for the gender of these kids?

Sample space and complements

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

- A couple has one kid, what is the sample space for the gender of this kid? $S = \{M, F\}$
- A couple has two kids, what is the sample space for the gender of these kids? $S = \{MM, FF, FM, MF\}$

Sample space and complements

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

- A couple has one kid, what is the sample space for the gender of this kid? $S = \{M, F\}$
- A couple has two kids, what is the sample space for the gender of these kids? $S = \{MM, FF, FM, MF\}$

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

- A couple has one kid. If we know that the kid is not a boy, what is gender of this kid? { ~~M~~, F } $\rightarrow$ Boy and girl are complementary outcomes.
- A couple has two kids, if we know that they are not both girls, what are the possible gender combinations for these kids?

Sample space and complements

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

- A couple has one kid, what is the sample space for the gender of this kid? $S = \{M, F\}$
- A couple has two kids, what is the sample space for the gender of these kids? $S = \{MM, FF, FM, MF\}$

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

- A couple has one kid. If we know that the kid is not a boy, what is gender of this kid? { ~~M~~, F } $\rightarrow$ Boy and girl are complementary outcomes.
- A couple has two kids, if we know that they are not both girls, what are the possible gender combinations for these kids? { *MM, ~~FF~~, FM, MF* }

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

Between January 9-12, 2013, SurveyUSA interviewed a random sample of 500 NC residents asking them whether they think widespread gun ownership protects law abiding citizens from crime, or makes society more dangerous. 58% of all respondents said it protects citizens. 67% of White respondents, 28% of Black respondents, and 64% of Hispanic respondents shared this view. Which of the below is true?

Opinion on gun ownership and race ethnicity are most likely

a. complementary
b. mutually exclusive
c. independent
d. dependent
e. disjoint

http://www.surveyusa.com/client/PollReport.aspx?g=a5f460ef-bba9-484b-8579-1101ea26421b

Practice

Between January 9-12, 2013, SurveyUSA interviewed a random sample of 500 NC residents asking them whether they think widespread gun ownership protects law abiding citizens from crime, or makes society more dangerous. 58% of all respondents said it protects citizens. 67% of White respondents, 28% of Black respondents, and 64% of Hispanic respondents shared this view. Which of the below is true?

Opinion on gun ownership and race ethnicity are most likely

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

http://www.surveyusa.com/client/PollReport.aspx?g=a5f460ef-bba9-484b-8579-1101ea26421b

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(protects citizens) = 0.58*
- *P(protects citizens | White) = 0.67*
- *P(protects citizens | Black) = 0.28*
- *P(protects citizens | Hispanic) = 0.64*

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(protects citizens) = 0.58*
- *P(protects citizens | White) = 0.67*
- *P(protects citizens | Black) = 0.28*
- *P(protects citizens | Hispanic) = 0.64*

$P(protects\ citizens)$ *varies by race/ethnicity, therefore opinion on gun ownership and race ethnicity are most likely* dependent*.*

Product rule for independent events

$$P(A \ and \ B) = P(A) \times P(B)$$

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

Product rule for independent events

$$P(A \; and \; B) = P(A) \times P(B)$$

Or more generally, $P(A_1 \; and \; \cdots \; and \; A_k) = P(A_1) \times \cdots \times P(A_k)$

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

Product rule for independent events

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

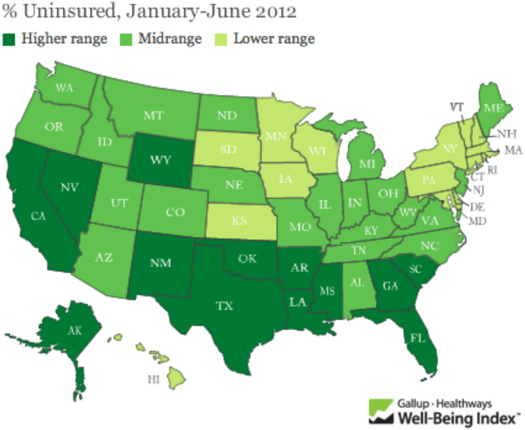
Or more generally, $P(A_1 \text{ and } \cdots \text{ and } A_k) = P(A_1) \times \cdots \times P(A_k)$

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 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

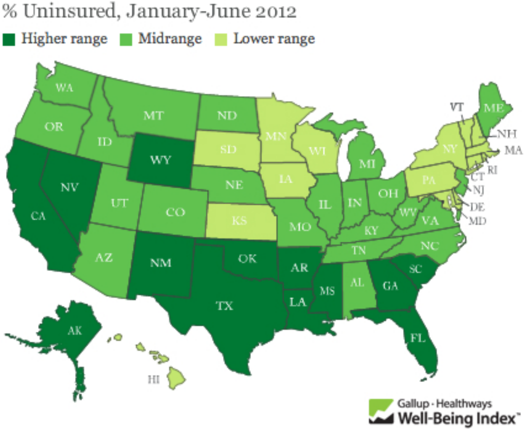
(c) 0.255×2

(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 intro two categories:

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

Putting everything together...

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

$$Prob(at\ least\ 1\ uninsured) \quad = \quad 1 - Prob(none\ uninsured)$$

Putting everything together...

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

$$\begin{aligned} Prob(at\ least\ 1\ uninsured) &= 1 - Prob(none\ uninsured) \\ &= 1 - [(1 - 0.255)^5] \end{aligned}$$

Putting everything together...

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

$$
\begin{aligned}
Prob(at\ least\ 1\ uninsured) &= 1 - Prob(none\ uninsured) \\
&= 1 - [(1 - 0.255)^5] \\
&= 1 - 0.745^5
\end{aligned}
$$

Putting everything together...

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

$$\begin{aligned} Prob(at\ least\ 1\ uninsured) &= 1 - Prob(none\ uninsured) \\ &= 1 - [(1 - 0.255)^5] \\ &= 1 - 0.745^5 \\ &= 1 - 0.23 \end{aligned}$$

Putting everything together...

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

$$\begin{aligned} Prob(at\ least\ 1\ uninsured) &= 1 - Prob(none\ uninsured) \\ &= 1 - [(1 - 0.255)^5] \\ &= 1 - 0.745^5 \\ &= 1 - 0.23 \\ &= 0.77 \end{aligned}$$

At least 1

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

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. $1 - 0.8^3$

$$P(at\ least\ 1\ from\ veg) = 1 - P(none\ veg)$$
$$= 1 - (1 - 0.2)^3$$
$$= 1 - 0.8^3$$
$$= 1 - 0.512 = 0.488$$

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) $= \frac{48}{72} \approx 0.67$

Joint probability

What is the probability that a patient received the 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 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) $= \frac{10}{72} \approx 0.14$

Conditional probability

Conditional probability

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

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

Conditional probability

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

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

Conditional probability

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

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

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

Conditional probability

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) $= \frac{10}{24} \approx 0.42$

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 | desipramine) = $\frac{10}{24} \approx 0.42$

P(relapse | lithium) = $\frac{18}{24} \approx 0.75$
P(relapse | placebo) = $\frac{20}{24} \approx 0.83$

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

$P(\text{desipramine} \mid \text{relapse}) = \frac{10}{48} \approx 0.21$

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) $= \frac{10}{48} \approx 0.21$

P(lithium | relapse) $= \frac{18}{48} \approx 0.375$
P(placebo | relapse) $= \frac{20}{48} \approx 0.42$

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

$$P(A \ and \ B) = P(A|B) \times P(B)$$

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

$$P(A \; and \; B) = P(A|B) \times P(B)$$

  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 she is 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 she is 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 she is 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, $P(A\ and\ B) = P(A) \times P(B)$. Then,

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

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

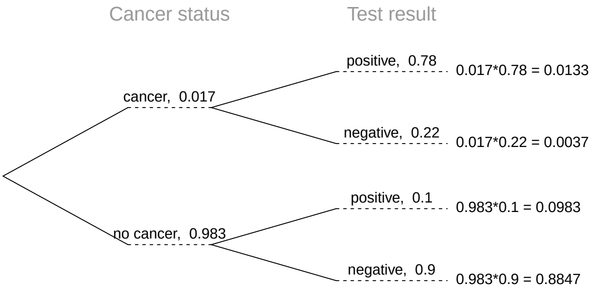
These percentages are approximate, and very difficult to estimate.

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?

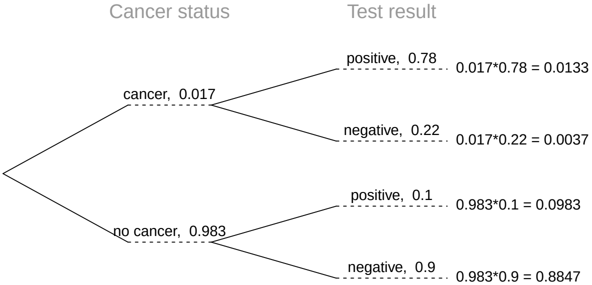
Inverting probabilities

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

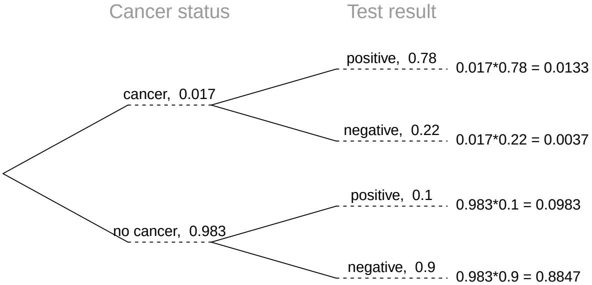
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



$P(C|+)$

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?

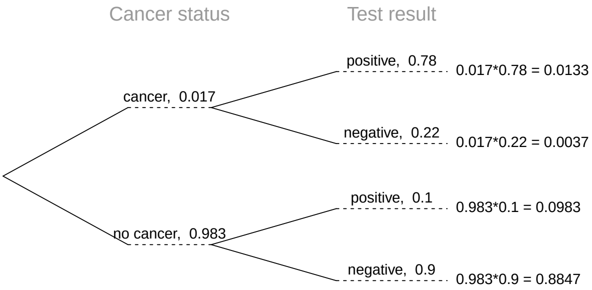


$P(C|+)$

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

Inverting probabilities

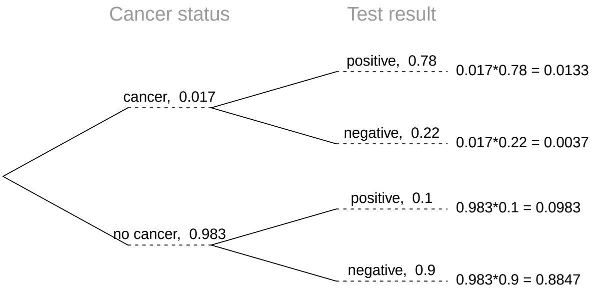
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



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

Inverting probabilities

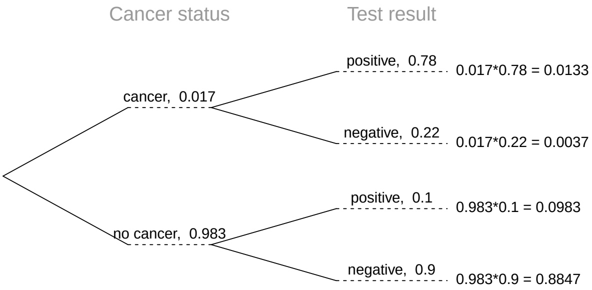
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



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

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?



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

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

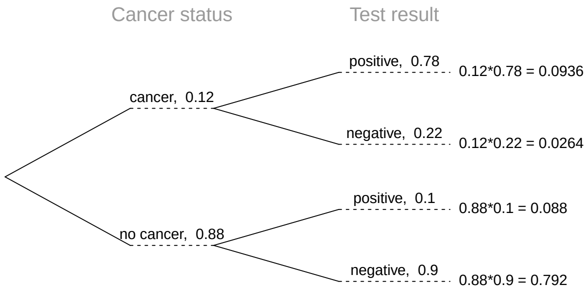
What is the probability that this woman has cancer if this second mammogram also yielded a positive result?

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?

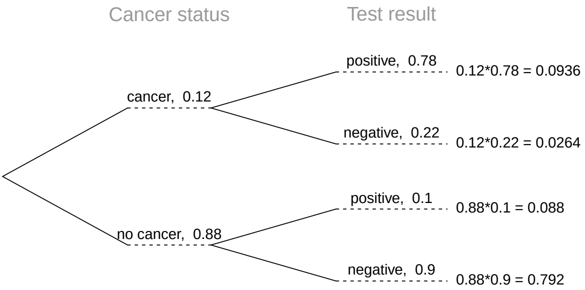
- (a) 0.0936
- (b) 0.088
- (c) 0.48
- (d) *0.52*



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*



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

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:

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

$$= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

where A_1, A_2, $\cdots$, A_k represent all other possible outcomes of variable 1.

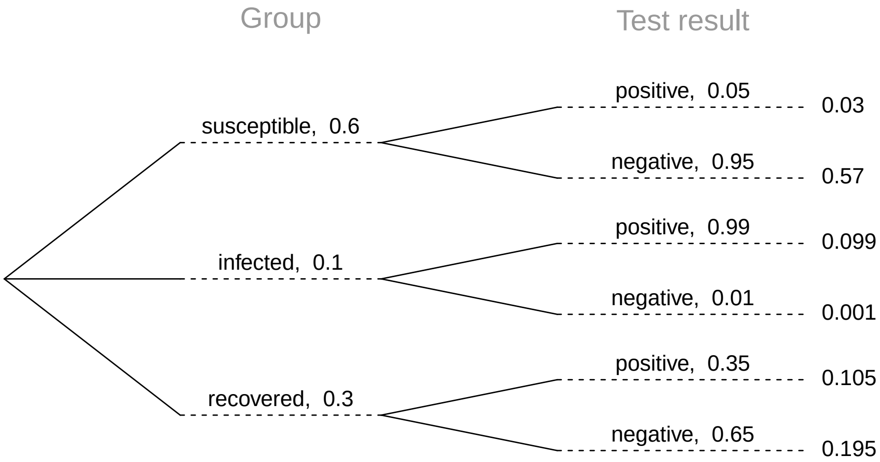
Application activity: Inverting probabilities

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.

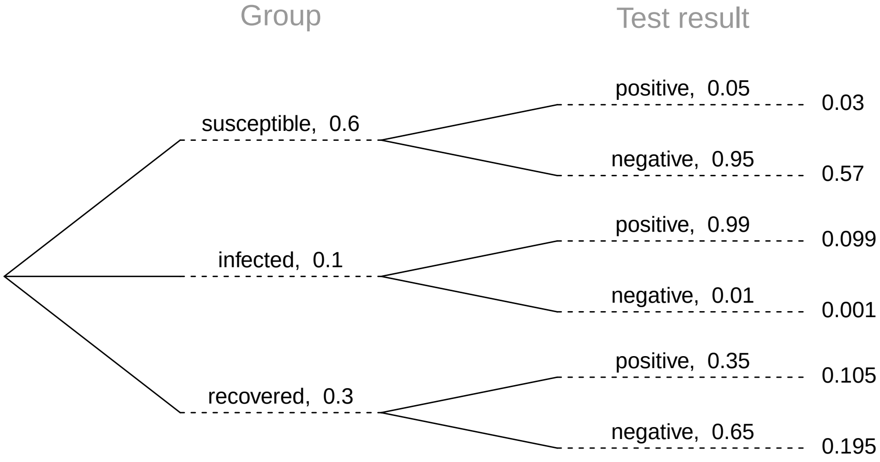
Imagine a population in the midst of an epidemic where 60% of the population is considered susceptible, 10% is infected, and 30% is recovered. The only test for the disease is accurate 95% of the time for susceptible individuals, 99% for infected individuals, but 65% for recovered individuals. (Note: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



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