MATH 1051H-A: Lecture #06

Considering Categorical Data

Contingency Tables

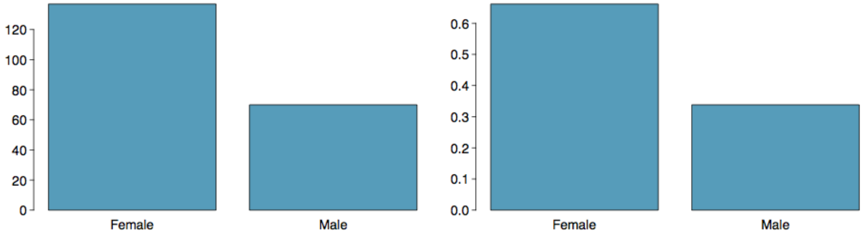
A table that summarizes data for two categorical variables is called a **contingency table**.

The contingency table below shows the distribution of students’ genders and whether or not they are looking for a spouse while in college.

	looking for spouse		
	No	**Yes**	**Total**
gender			
Female	86	51	137
Male	52	18	70
Total			
	138	69	207

Bar Plots

A **bar plot** is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a **relative frequency bar plot**.



How are bar plots different than histograms? Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

Choosing the Appropriate Proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

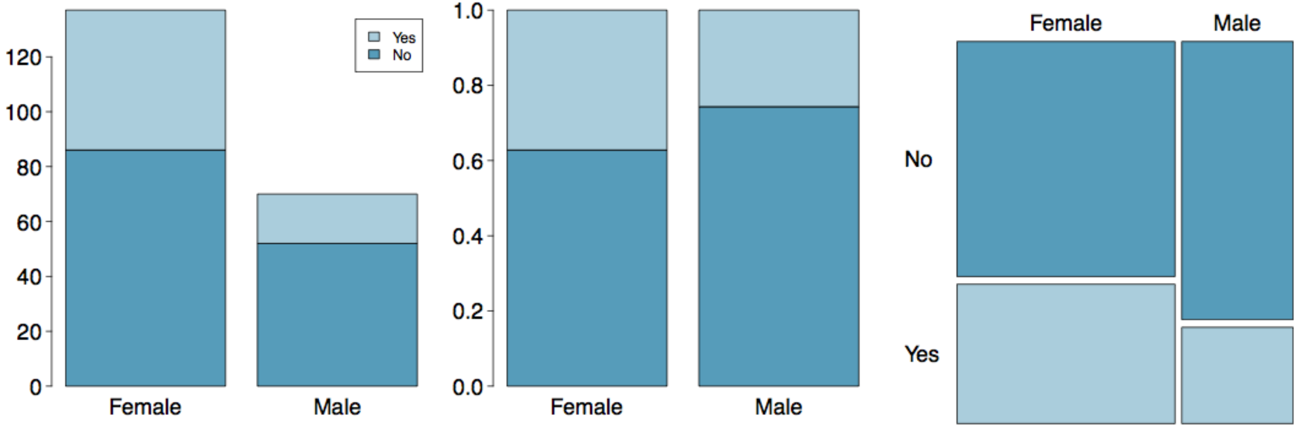
	looking for spouse		
	No	Yes	Total
gender			
Female	86	51	137
Male	52	18	70
Total			
	138	69	207

To answer this question we examine the row proportions:

- % Females looking for a spouse: 51 / 137 ~ 0.37
- % Males looking for a spouse: 18 / 70 ~ 0.26

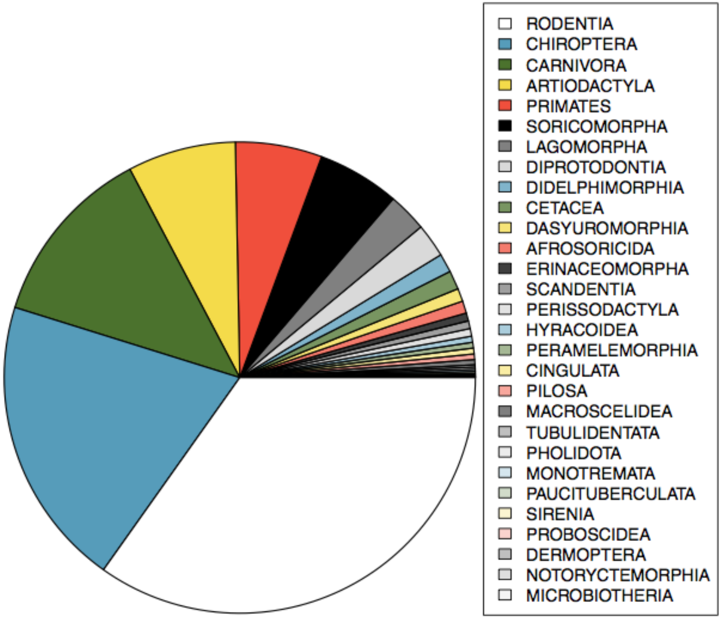
Segmented Bar and Mosaic Plots

What are the differences between the three visualizations shown below?



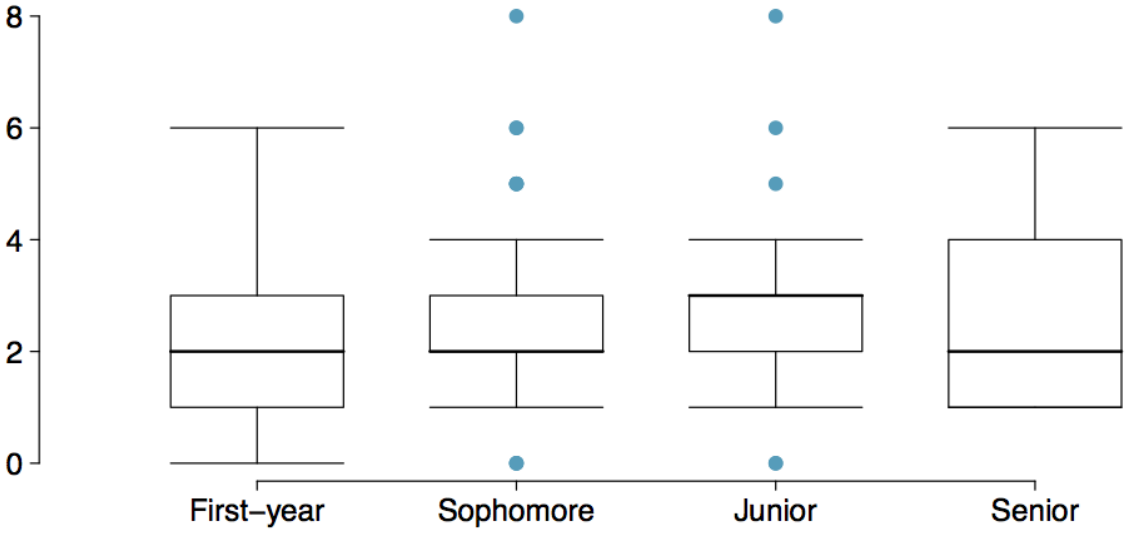
Pie Charts

Can you tell which order encompasses the lowest percentage of mammal species?



Comparing Numerical Data Across Groups

Does there appear to be a relationship between class year and the number of clubs students are in?



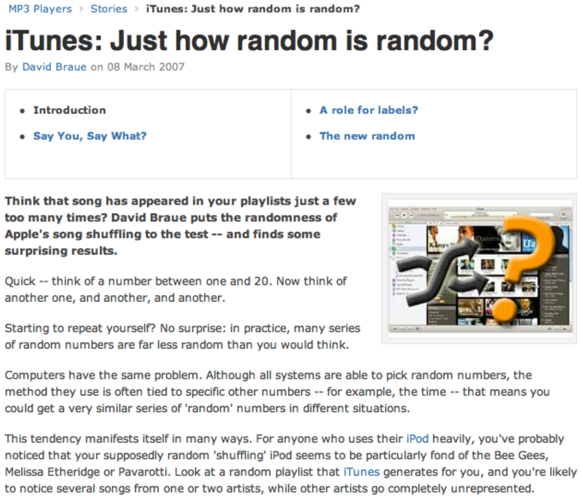
Defining Probability

Note

We will definitely not finish these notes today. This will spill into Friday, and possibly next week. We *are* done Chapter 2 of the 4th edition of the textbook for now, though.

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.

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: Frequentist

- **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.
  - relies on a “multiverse” concept for most problems, “if”

Probability: Bayesian

- **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?

1. exactly 3 heads in 10 coin flips
2. exactly 3 heads in 100 coin flips
3. exactly 3 heads in 1000 coin flips

Practice

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

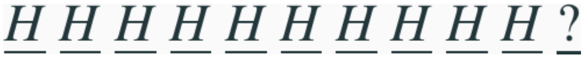
1. exactly 3 heads in 10 coin flips
2. exactly 3 heads in 100 coin flips
3. *exactly 3 heads in 1000 coin flips*

Law of large numbers

The **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.)

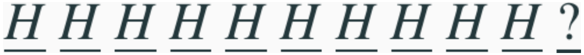
When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?



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

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?



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

Law of large numbers (cont.)

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

* 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 the **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.

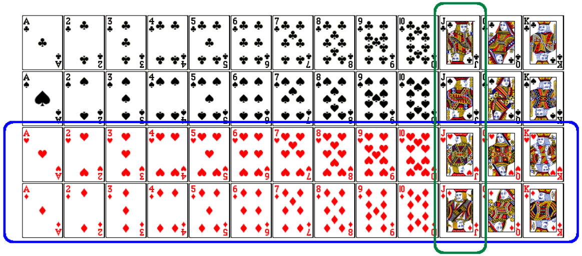
- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

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

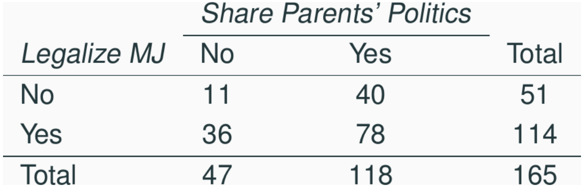
What is the probability of drawing a jack or a red card from a well shuffled full deck?



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

Practice

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

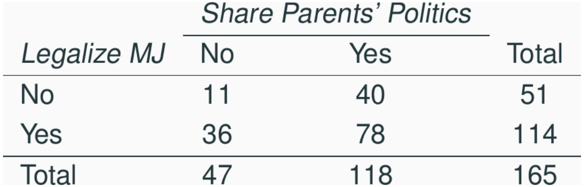


	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes	36	78	114
Total	47	118	165

1. $\frac{40+36-78}{165}$
2. $\frac{114+118-78}{165}$
3. $\frac{78}{165}$
4. $\frac{78}{188}$

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

1. $\frac{40+36-78}{165}$
2. $\frac{114+118-78}{165}$
3. $\frac{78}{165}$
4. $\frac{78}{188}$

Recap

General addition rule $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$

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

Probability distributions

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

- The probability distribution for the gender of one child:

Event	Male	Female
Probability	0.5	0.5

Probability distributions

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

- Rules for probability distributions:
  - The events listed must be disjoint
  - Each probability must be between 0 and 1
  - 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 genders of two children:

Event	MM	FF	MF	FM
Probability	0.25	0.25	0.25	0.25

Practice

In a survey, 52% of respondents said they like pizza. What is the probability that a randomly selected respondent from this sample is a **pizza hater**.

1. 0.48
2. more than 0.48
3. less than 0.48
4. cannot calculate using only the information given

Practice

In a survey, 52% of respondents said they like pizza. What is the probability that a randomly selected respondent from this sample is a **pizza hater**.

1. 0.48
2. more than 0.48
3. less than 0.48
4. *cannot calculate using only the information given*

More: If the only two states people can be in are **pizza likers** and **pizza haters**, then the first option (1) is possible. However, it is also possible there are people who are ambivalent about pizza, or love pizza (more than like?), or who have never eaten pizza! The only answer we can eliminate is (b), because it is impossible by the rules of probability.

Sample space and complements

The **sample space** is the collection of all possible outcomes of a trial (or experiment).

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

Sample space and complements

The **sample space** is the collection of all possible outcomes of a trial (or experiment).

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

Sample space and complements

The **sample space** is the collection of all possible outcomes of a trial (or experiment).

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

Sample space and complements

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

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

Independence

Two processes are said to be **independent** of one another 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 said to be **independent** of one another 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 North Carolina 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?

Opinions on gun ownership and race ethnicity are most likely

1. complementary
2. mutually exclusive
3. independent
4. dependent
5. disjoint

Practice

Between January 9-12, 2013, SurveyUSA interviewed a random sample of 500 North Carolina 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?

Opinions on gun ownership and race ethnicity are most likely

1. complementary
2. mutually exclusive
3. independent
4. *dependent*
5. disjoint

Checking for independence:

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

The Gun Ownership Question

- P(protects citizens) = 0.58
- P(randomly selected NC resident says gun ownership protects citizens, given that the resident is white) = 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.

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.

More on dependence

We saw that P(protects citizens $|$ White) = 0.67 and P(protects citizens $|$ Hispanic) = 0.64. Under which condition would you be more convinced of a real difference between the proportions of Whites and Hispanics who think gun widespread gun ownership protects citizens? $n = 500$ or $n = 50,000$

Product rule for independent events: $P(A\ and\ B) = P(A) \times P(B)$

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

Example

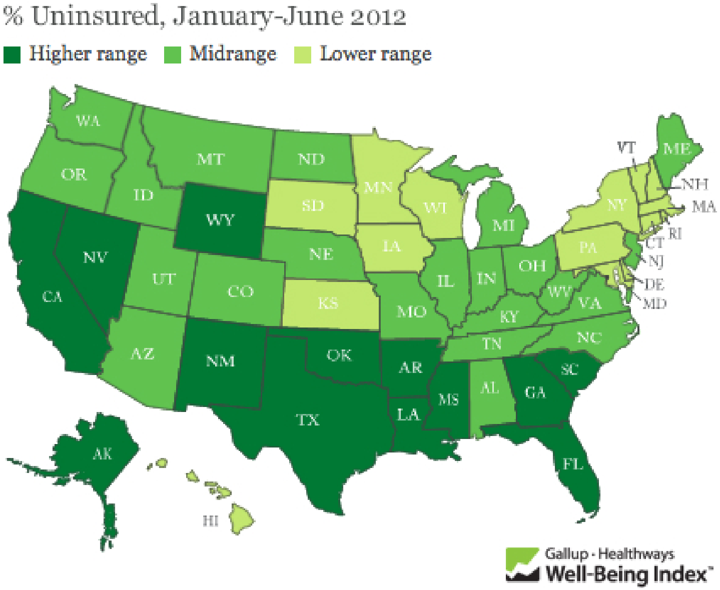
You toss a coin twice, what is the probability of getting two tails in a row?

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

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?

- 25.5^2
- 0.255^2
- 0.255×2
- $(1 - 0.255)^2$



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

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

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?

- $1 - 0.2 \times 3$
- $1 - 0.2^3$
- 0.8^3
- $1 - 0.8 \times 3$
- $1 - 0.8^3$

Practice

$$\begin{aligned} P(\text{at least 1 from veg}) &= 1 - P(\text{none veg}) \\ &= 1 - (1 - 0.2)^3 \\ &= 1 - 0.8^3 \\ &= 1 - 0.512 = 0.488 \end{aligned}$$

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?

- $1 - 0.2 \times 3$
- $1 - 0.2^3$
- 0.8^3
- $1 - 0.8 \times 3$
- $1 - 0.8^3$

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

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**

$$P(\text{relapsed}) = \frac{48}{72} \approx 0.67$$

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

$$P(\text{relapsed and desipramine}) = \frac{10}{72} \approx 0.14$$

Conditional probability

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

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

	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

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

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

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

Conditional probability (cont.)

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

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

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

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

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

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

soc	ial science non	-social science tot	al
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.

- 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,

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

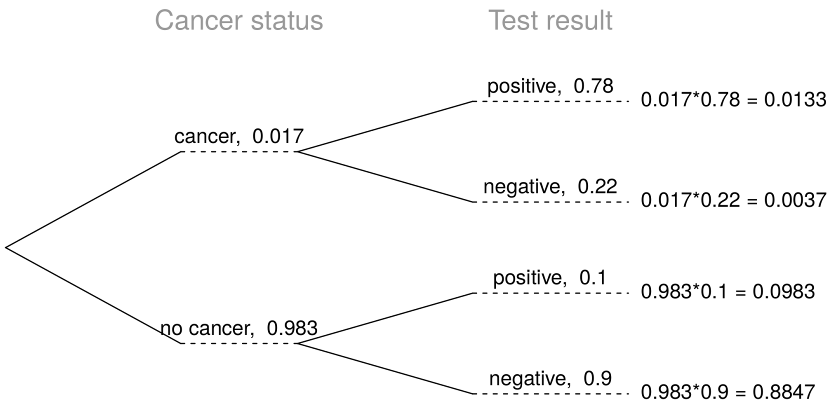
Breast cancer screening

- American Cancer Society estimates that about 1.7% of women have breast cancer.
- Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.
- An article published in 2003 suggests that up to 10% of all mammograms result in false positives for patients who do not have cancer.

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 (ctd.)

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

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?

- 0.017
- 0.12
- 0.0133
- 0.88

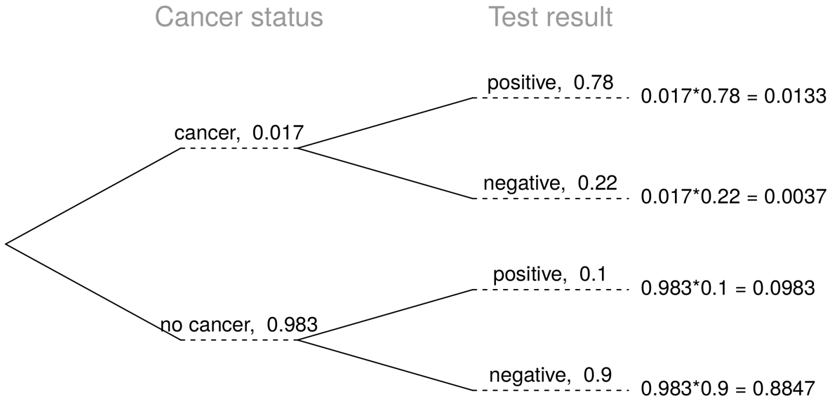
Practice

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?

- 0.017
- *0.12*
- 0.0133
- 0.88

Practice

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



Practice

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

- 0.0936
- 0.088
- 0.48
- 0.52

Practice

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

- 0.0936
- 0.088
- 0.48
- *0.52*

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

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:**

$$
\begin{aligned}
&P(\text{outcome } A_1 \text{ of variable } 1 | \text{outcome B of variable } 2) \\
&\qquad = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}
\end{aligned}
$$

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

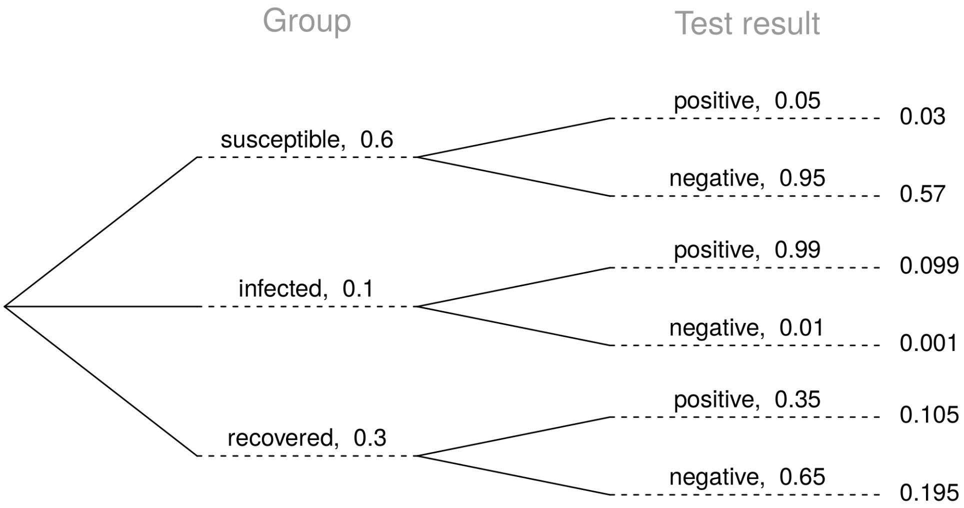
Application activity: Inverting probabilities

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.)



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