Joint probability, conditional probability, and multiple comparisons

February 27, 2017
Announcements / Action Items

• Schedule update: final R lab moved to Week 10

• Optional Survey 5 coming soon
Last time

- We introduced the $X^2$ statistic for comparing observed and expected frequencies for *many* proportions at once

- If we have a **single** categorical variable, we can use a $X^2$ test for goodness-of-fit to compare the proportions of responses in each category to a hypothesized model

- If we have **multiple** categorical variables, we can use a $X^2$ test for independence to ask whether the two variables are associated (not independent) or not associated (independent)
This time

- How do probabilities change with multiple comparisons?
- How do we use conditional probabilities to ask and answer the right question?
- How can we calculate conditional probabilities using Bayes’ rule?
- Why do we need to treat non-independent data differently?
Everyday statistics

We are all consumers of statistics

“What I, in my complacency, chose to ignore is just how much of the persuasion now aimed at the average citizen comes in the form of numbers, specifically numbers that tell us about the future, about how likely something is to happen (or not happen) based on how much it happened (or didn’t) in the past. These numbers sing to us the siren song of cause and effect, humanity’s favorite tune. Why do we like it so much? Because knowing what causes events and conditions is the first step toward controlling them, and we human beings are all about controlling our environments. That’s how we ended up ruling this planet, and it’s how some of us hope to save it.”

http://www.slate.com/articles/life/classes/2015/08/take_a_statistics_and_probability_class_in_college_to_improve_critical_thinking.html
Reminder: probability

- $p(\text{not } A) = 1 - p(A)$

- $p(A \text{ or } B) = p(A) + p(B)$ if $A$ and $B$ are mutually exclusive

- $p(A \text{ and } B) = p(A) \times p(B)$ if $A$ and $B$ are independent

- $p(A \mid B) = \frac{p(A \text{ and } B)}{p(B)}$
statistics is about math

statistics is about people
This time

• How do probabilities change with multiple comparisons?

• How do we use conditional probabilities to ask and answer the right question?

• How can we calculate conditional probabilities using Bayes’ rule?

• Why do we need to treat non-independent data differently?
the good: we used a single test to infer that there is an association between city and washing

the bad: but we still don’t know which pairs of cities have different proportions of washers from each other

why didn’t we test for a difference in proportions in ATL vs. CHI, ATL vs. NY, ATL vs. SF, CHI vs. NY, CHI vs. SF, and NY vs. SF?

we need to return to some ideas of probability

<table>
<thead>
<tr>
<th>Observed</th>
<th>Atlanta</th>
<th>Chicago</th>
<th>NY</th>
<th>SF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washed</td>
<td>1175</td>
<td>1329</td>
<td>1169</td>
<td>1521</td>
<td>5194</td>
</tr>
<tr>
<td>Did not wash</td>
<td>413</td>
<td>180</td>
<td>334</td>
<td>215</td>
<td>1142</td>
</tr>
<tr>
<td>Total</td>
<td>1588</td>
<td>1509</td>
<td>1503</td>
<td>1736</td>
<td>6336</td>
</tr>
</tbody>
</table>
Imagine a woman named Linda, who is 31 years old, single, outgoing, and very bright. In college she majored in philosophy. While a student she was deeply concerned with discrimination and social justice and participated in a number of demonstrations for environmental protection. Rank each statement about Linda in terms of the highest probability (1) to the lowest probability (7). (Tversky & Kahnman)

Linda sells insurance
Linda is a registered democrat
Linda sells insurance and is a registered democrat
Linda is a bank teller
Linda is active in women’s rights movements
Linda is a bank teller and is active in women’s rights movements
Linda is a teacher
the probability that two things will *both* happen is always less than (or equal to) the individual probability that one of those things will happen

independent: \( p(A \text{ and } B) = p(A) \times (B) \)
non-independent: \( P(A \text{ and } B) = p(A) \times P(B \mid A) \)
the probability that two things both will not happen is always less than (or equal to) the individual probability that one of those things will not happen

the probability of many specific events can be low, but the probability that at least one happens can be high

independent: \( p(\text{not A and not B}) = p(\text{not A}) \times (\text{not B}) \)
non-independent: \( P(\text{not A and not B}) = p(\text{not A}) \times P(\text{not B} \mid \text{not A}) \)
Powerball

What are the odds that I will win Powerball?
1 in 292 million

What are the odds that someone will win Powerball?
1 - p(nobody wins)
1 - p(P1 doesn’t win and P2 doesn’t win and …)
1 - p(P1 doesn’t win) * p(P2 doesn’t win) * …
1 - p(I don’t win)^number of players

with 50000000 independent players, 16% chance that someone will win

* this is a binomial probability problem with 0 successes, 50000000 trials, and π = (1 / 292 million)
Birthdays

If we select 25 random people, what is the probability that two people will share the same birthday?

If we weren’t sure how to work out the math, we could use simulation

for many repetitions

  draw 25 random numbers out of the numbers 1 to 365 (sample with replacement)

  determine whether there is a duplicate

calculate proportion of times we saw a duplicate

see accompanying R code for this lecture
Birthdays

If we select 25 random people, what is the probability that two people will share the same birthday?

$$p(\text{shared birthday})$$

$$= 1 - p(\text{no shared birthday})$$

$$= 1 - p(\text{P1 and P2 don’t share and P3 and P2 don’t share and P3 and P1 don’t share and …})$$

$$= 1 - p(\text{P1 and P2 don’t share}) \cdot p(\text{P3 and P2 don’t share}) \cdot p(\text{P3 and P1 don’t share}) \cdot …$$

$$= 1 - p(\text{random pair doesn’t share})^{\text{total number of pairs in our group}}$$

$$= 1 - p\left(\frac{364}{365}\right)^{\begin{pmatrix}25 \\ 2\end{pmatrix}} = 1 - p\left(\frac{364}{365}\right)^{300} = .56$$
Penny flipping

- If we flip a coin 100 times, the probability of getting 60 or more heads is .03
- If we flip five coins 100 times, the probability that at least one of them will get 60 or more heads is .13
- If we flip a coin 26 times the odds of getting *all* heads are about 1 in 73 million
- If we had 300 million people flip a coin 26 times the odds that *someone* will get all heads is about 98%
Lotteries

- September 9, 1981 the lottery number 8092 was drawn in both Massachusetts and New Hampshire (described as a 1 in 100 million event; 1 in 10000 if we notice that we would have been equally impressed with any number; much less if we notice we would have been impressed with any two lotteries on any two days)

- A woman in New Jersey won the lottery twice, in 1985 and 1986 (1 in 17 trillion, down to 1.5 million if we condition on already having won, adjust for the fact that she played weekly for years, down to about 1 in 1000, adjust for the fact that there are many lottery winners, no big deal)
March madness brackets

- 63 games total, participation estimated to be 10 million to 60 million

- Longest (verified) streak so far is picking the 34 games correctly — also an unverified report of a perfect bracket through first two rounds (48 games)
  - Is this person necessarily a basketball savant?
  - How do you think this person will do next year?

- Warren Buffett offered 1 billion dollar prize for a perfect NCAA bracket — was this foolish, given what we’ve seen?
  - Completely random guessing: 1 in 9.2 quintillion
  - Not random: expert opinions range from 1 in 5 billion to 1 in 135 billion
  - Not independent (people tend to pick similar brackets)
  - Puts odds of someone winning at about 1 in 3.3 million
  - Put another way, would need 90 billion brackets before odds of someone winning was 50-50
Putting the Polling Miss of the 2016 Election in Perspective

It was the biggest polling miss in a presidential election in decades.

Yet in many ways, it wasn't wholly out of the ordinary.

Over all, the national polls missed the result by only a few points: Hillary Clinton is on track to win the popular vote by around 1.5 percentage points, not especially far from her roughly four-point lead in an average of national polls.

State Polling Errors in 2016 Were the Largest in Decades

<table>
<thead>
<tr>
<th>Year</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>3.4 pts.</td>
</tr>
<tr>
<td>1992</td>
<td>3.4 pts.</td>
</tr>
<tr>
<td>1996</td>
<td>2.3 pts.</td>
</tr>
<tr>
<td>2000</td>
<td>1.8 pts.</td>
</tr>
<tr>
<td>2004</td>
<td>1.7 pts.</td>
</tr>
<tr>
<td>2008</td>
<td>1.7 pts.</td>
</tr>
<tr>
<td>2012</td>
<td>2.3 pts.</td>
</tr>
<tr>
<td>2016</td>
<td>3.9 pts.</td>
</tr>
</tbody>
</table>

In 2016, state polls understated Republican support by large margins in many small states. In the five states in bold, Hillary Clinton lost the statewide vote despite leading in the polls.

Arrows show simple average of state polls conducted in the three weeks before Election Day in states with at least three polls, compared with the projected final election result.

In 2012, the state and national polls were off by a similar amount, but in the other direction.
Planned comparisons

The odds of a random DNA match are 1 in 1000000

You’re on the jury, and you hear that DNA at the crime scene is a match for the defendant

How convincing is this evidence if:

• The police found a suspect based on other evidence, tested the suspect’s DNA, and found that it was a match (a “planned comparison”)

• The police found DNA at the crime scene, ran it through a database of millions of stored DNA samples, found it matched someone in the database, and identified this person as a suspect (an “exploratory analysis”)
through a lens of multiple comparisons, people or events that seemed special begin to seem ordinary

the relevant probability question is not what is the probability that this specific event happened but rather what is the probability that an event like this would happen eventually
Class data

the good, people tended to believe that the probability of someone winning the lottery was greater than the probability of a specific person winning the lottery, and that the probability of at least one coin having 60+ heads was more than the probability of a single coin having 60+ heads

the bad (but typical), the impact of these multiple comparisons was under-estimated

\[ p(\text{single coin 60+ heads}), \text{median} = .40 \]
\[ p(\text{sharing birthdays}), \text{median} = .09 \]
\[ p(\text{duplicate lottery numbers}), \text{median} = .0004 \]
Class data

- \( p(\text{a duplicate number is selected from 500 draws from 2.4 million numbers}) = 0.05 \)
- \( p(\text{at least 2 people out of 25 people share a birthday}) = 0.56 \)
- \( p(\text{flipping a coin 100 times and getting > 60 heads}) = 0.028 \)

Many comparisons \( \rightarrow \) increases probability of at least one occurrence.

Central limit theorem (large \( n \) \( \rightarrow \) decreased standard error).
The test-wise alpha level is the risk of a Type I error for an individual hypothesis test ($\alpha$).

When an experiment involves several different hypothesis tests, the family-wise alpha level is the total probability of a Type I error in at least one test.

Imagine we did six pairwise tests of proportions (ATL-CHI, ATL-NY, ATL-SF, CHI-NY, CHI-SF, NY-SF) and they were all independent (they’re not) and that in the population there were no differences in proportions ($H_0$ is true):

- Family-wise alpha level
  
  $$
  = p(\text{at least one Type I Error})
  
  = 1 - p(\text{no Type I Errors})
  
  = 1 - (0.95)^6
  
  = 0.26
  $$

by calculating a single statistic (such as with $\chi^2$), we’ve kept our family-wise alpha at 0.05.
Follow-up tests

But we do want to know which cities are significantly different from each other

Typical workflow

Perform single test to control family-wise $\alpha$

If $p > \alpha$, stop

If $p < \alpha$, perform follow-up tests to zoom in on specific differences

pairwise tests for planned comparisons of interest (in this case, hypothesis tests or confidence intervals for differences in proportions)

all six possible pairwise tests (we now feel more justified doing this …) (in this case, hypothesis tests or confidence intervals for differences in proportions)

also are procedures to perform pairwise comparisons while still controlling family-wise $\alpha$
What if we want to do many different tests that can’t be combined under a single test?

*example, looking for associations between thousands of genes and health outcomes → thousands of t-statistics*

imagine that all null hypotheses were true, there were *no* associations between any genes and any outcomes

what would the distribution of t-statistics look like

Are the genes that are in the tails of this distribution *special*?
Multiple comparisons across tests

Data dredging: “the [questionnaire] we used produced 1,066 variables, and the additional questions we asked sorted survey-takers according to 26 possible characteristics (left- or right-handed, for example). This vast data set allowed us to do 27,716 regressions in just a few hours. (You can see the full results on GitHub.)” (http://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/)
Multiple comparisons *across* tests

See also [http://www.tylervigen.com/spurious-correlations](http://www.tylervigen.com/spurious-correlations)

---

**Divorce rate in Maine**

correlates with

**Per capita consumption of margarine**

---

**Per capita cheese consumption**

correlates with

**Number of people who died by becoming tangled in their bedsheets**

---

Careful, these same analyses start to seem much more convincing when the outcome variables are health outcomes.
Multiple comparisons across tests

- We’ve been treating the outcomes of hypothesis tests agnostically — but in reality it is hard to publish research if you fail to reject the null hypothesis.

- Imagine 100 scientists all test the same null hypothesis, which happens to be true:
  - 95 fail to reject the null hypothesis → they move on, put their result in their “file drawer”
  - 5 do reject the null hypothesis → publish their findings in scientific journals

- The file drawer problem: I look at scientific journals and find 5 reports of an inference that the alternative hypothesis is true, 0 reports of failure to reject the null hypothesis.
similar principles apply to experimenting with ‘multiple tests’ within a single test
Multiple comparison within a test

Unplanned one-tailed tests: if I choose to do a one-tailed test after looking at my data to choose my tail, I have performed two one-tailed tests and selected the preferred one — now my $\alpha = .10$ (Type I Error sums here because of no potential for overlap between the two tails) see R Lab 3 for simulation

Researcher degrees of freedom: small changes that researchers can make in the way they structure their analysis to answer a question — amounts to doing many (non-independent) tests and then selecting the preferred one
Experimenter degrees of freedom

https://fivethirtyeight.com/features/science-isnt-broken/
Hack Your Way To Scientific Glory

You're a social scientist with a hunch: The U.S. economy is affected by whether Republicans or Democrats are in office. Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1. CHOOSE A POLITICAL PARTY
   - Republicans
   - Democrats

2. DEFINE TERMS
   - Which politicians do you want to include?
     - Presidents
     - Governors
     - Senators
     - Representatives
   - How do you want to measure economic performance?
     - Employment
     - Inflation
     - GDP
     - Stock prices
   - Other options
     - Factor in power
     - Weight more powerful positions more heavily
     - Exclude recessions
     - Don't include economic recessions

3. IS THERE A RELATIONSHIP?
   - Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.

4. IS YOUR RESULT SIGNIFICANT?
   - If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.

   Result: Publishable
   - You achieved a p-value of 0.03 and showed that Democrats have a negative effect on the economy. Get ready to be published!

   If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

   Data from The United States Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

   If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

   Data from The United States Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

5. IS YOUR RESULT SIGNIFICANT?
   - If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.

   Result: Unpublishable
   - With a p-value of 0.40, your findings are not statistically significant. Try defining your terms differently.

Data from The United States Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

https://fivethirtyeight.com/features/science-isnt-broken/
Hack Your Way To Scientific Glory

You're a social scientist with a hunch: The U.S. economy is affected by whether Republicans or Democrats are in office. Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1. **Choose a Political Party**
   - Republicans
   - Democrats

2. **Define Terms**
   - Which politicians do you want to include?
     - Presidents
     - Governors
     - Senators
     - Representatives
   - How do you want to measure economic performance?
     - Employment
     - Inflation
     - GDP
     - Stock prices
   - Other options
     - Factor in power
     - Exclude recessions
     - Don't include economic recessions

3. **Is There a Relationship?**
   - Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.

4. **Is Your Result Significant?**
   - If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.

Result: **Publishable**
- You achieved a p-value of less than 0.01 and showed that Democrats have a positive effect on the economy. Get ready to be published!

Alternatively, if you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The Fullyinformed Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

https://fivethirtyeight.com/features/science-isnt-broken/
Experimenter degrees of freedom

**Same Data, Different Conclusions**
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

Referees are three times as likely to give red cards to dark-skinned players.

Statistically significant results showing referees are more likely to give red cards to dark-skinned players.

Twice as likely

Equally likely

Non-significant results

Source: Brian Nosek et al.

https://fivethirtyeight.com/features/science-isnt-broken/
What about exploration?

- Exploring data (including by making multiple comparisons) gives rise to new hypotheses and ideas.

- One option:
  - Clearly mark *planned* vs. *exploratory analyses*.
  - Seek independent *planned* replication of results from *exploratory analyses*. 
This time

• How do probabilities change with multiple comparisons?

• How do we use conditional probabilities to ask and answer the right question?

• How can we calculate conditional probabilities using Bayes’ rule?

• Why do we need to treat non-independent data differently?
Reminder: conditional probability

\( p(A \mid B) \) is the *probability of \( A \) given that \( B \) has occurred*

\( p(\text{wash}) \) is the probability that a random person in a bathroom washes their hands

\( p(\text{wash} \mid \text{SF}) \) is the probability that a random person in a bathroom in San Francisco washes their hands

When used correctly, conditioning gives us more information and lets us make better predictions than if we did not have the additional information (also with means: recall variance explained by prediction based on *group*

We just need to make sure we’re asking the right question
The prosecutor’s fallacy

**OJ Simpson on trial for 1994 murder of wife, Nicole Brown Simpson and friend Ron Goldman**

- Prosecution argument: OJ Simpson had a history of spousal abuse → increased likelihood of murdering his wife

- Defense argument: fewer than 1 in 2500 men who abuse their wives go on to murder them
  
  - \( p(\text{murdered wife} \mid \text{abused wife}) < \frac{1}{2500} \)

- But this is conditioning on the wrong information

  - \( p(\text{murdered wife} \mid \text{abused wife and wife has been murdered}) \) is closer to 8/9
The prosecutor’s fallacy

Sally Clark on trial in 1998 for killing her two children, although she claimed they died accidentally in their sleep

• Prosecution argument: probability of one child dying of SIDS is $1 / 8500 \rightarrow$ probability of two children in the same family dying of SIDS is $(1 / 8500)^2 = 1$ in 73 million (*mistake #1, these are not independent events)

• $p(\text{two children in one family die of SIDS}) = 1$ in 73 million

• But this is not the correct probability

• $p(\text{children were murdered} \mid \text{two children in one family die while sleeping}) = \?, \text{but certainly much greater than } 1 \text{ in 73 million (*remember the multiple comparisons problem)}$
Imagine a test for a disease that is 99% accurate.

100 people have the disease and 100 people do not have the disease, and we give all of them the test.

If the test is positive, what is the probability that the person has the disease?

<table>
<thead>
<tr>
<th></th>
<th>positive test</th>
<th>negative test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>have disease</td>
<td>99</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>don’t have disease</td>
<td>1</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

\[
p(\text{positive test | disease}) = \frac{99}{100} = .99
\]

\[
p(\text{disease | positive test}) = \frac{99}{100} = .99
\]
The base rate

Imagine a test for a disease that is 99% accurate

100 people have the disease and 100000 people do not have the disease, and we give all of them the test

If the test is positive, what is the probability that the person has the disease?

<table>
<thead>
<tr>
<th></th>
<th>positive test</th>
<th>negative test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>have disease</td>
<td>99</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>don’t have disease</td>
<td>1000</td>
<td>99000</td>
<td>100000</td>
</tr>
<tr>
<td>total</td>
<td>1099</td>
<td>99001</td>
<td>100100</td>
</tr>
</tbody>
</table>

\[ p(\text{positive test} \mid \text{disease}) = \frac{99}{100} = .99 \]

\[ p(\text{disease} \mid \text{positive test}) = \frac{99}{1099} = .09 \]

we must consider the base rate of the disease
The base rate and comparisons

*a comparison that accounts for the base rate*

\[ p(\text{get disease} \mid \text{vaccinated}) = \frac{10}{90} = .11 \]

\[ p(\text{get disease} \mid \text{not vaccinated}) = \frac{5}{10} = .50 \]

*a comparison that ignores the base rate*

\[ p(\text{vaccinated} \mid \text{get disease}) = \frac{10}{15} = .67 \]

\[ p(\text{not vaccinated} \mid \text{get disease}) = \frac{5}{15} = .33 \]

<table>
<thead>
<tr>
<th></th>
<th>vaccinated</th>
<th>not vaccinated</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>get disease</td>
<td>10</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>don’t get disease</td>
<td>80</td>
<td>5</td>
<td>85</td>
</tr>
<tr>
<td>total</td>
<td>90</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>
This time

• How do probabilities change with multiple comparisons?

• How do we use conditional probabilities to ask and answer the right question?

• How can we calculate conditional probabilities using Bayes’ rule?

• Why do we need to treat non-independent data differently?
Bayes’ rule

How are \( p(A \mid B) \) and \( p(B \mid A) \) related?

\[
p(B \mid A) = \frac{p(A \text{ and } B)}{p(A)} \quad \rightarrow \quad p(A \text{ and } B) = p(B \mid A)p(A)
\]

\[
p(A \mid B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B \mid A)p(A)}{p(B)}
\]

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>not B</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>p(A and B)</td>
<td>-</td>
<td>p(A)</td>
</tr>
<tr>
<td>not A</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>p(B)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Using Bayes’ rule

\[ p(A|B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)} \]

If a person has the disease, the probability of a positive test is .99
If a person does not have the disease, the probability of a positive test is .01
The probability that a person has the disease is .000999
If the test is positive, what is the probability that the person has the disease?

\[
p(\text{disease} | \text{positive test}) = p(\text{positive test} | \text{disease}) \times p(\text{disease}) \div p(\text{positive test})
\]
\[
= .99 \times .000999 \div p(\text{positive test})
\]
Using Bayes’ rule

\[ p(A|B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)} \]

If a person has the disease, the probability of a positive test is .99
If a person does not have the disease, the probability of a positive test is .01
The probability that a person has the disease is .000999
If the test is positive, what is the probability that the person has the disease?

**What is \( p(\text{positive test}) \)?**

\[
p(\text{positive test})
\]

\[
= p(\text{positive test and disease}) \text{ or (positive test and no disease)}
\]

\[
= p(\text{positive test and disease}) + p(\text{positive test and no disease})
\]

\[
= p(\text{positive test} | \text{disease}) \cdot p(\text{disease})
\]

\[
\quad + p(\text{positive test} | \text{no disease}) \cdot p(\text{no disease})
\]

\[
= .99 \cdot .000999 + .01 \cdot (1 - .000999) = .011
\]

(this general idea is called the law of total probability)
Using Bayes’ rule

\[ p(A|B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)} \]

If a person has the disease, the probability of a positive test is .99

If a person does not have the disease, the probability of a positive test is .01

The probability that a person has the disease is .000999

If the test is positive, what is the probability that the person has the disease?

\[ p(\text{disease} | \text{positive test}) = p(\text{test} | \text{disease}) \times p(\text{disease}) / p(\text{positive test}) \]

\[ = .99 \times .000999 / .011 = .09 \]
Using Bayes’ rule

\[
p(A|B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}
\]

If a person has the disease, the probability that they are vaccinated is .67
If a person does not have the disease, the probability that they are vaccinated is .94
The probability that a person has the disease is .15
If the person is vaccinated, what is the probability that the person has the disease?

\[
p(\text{disease} \mid \text{vaccinated}) = p(\text{vaccinated} \mid \text{disease}) \times p(\text{disease}) / p(\text{vaccinated})
\]
\[
= .67 \times .15 / p(\text{vaccinated})
\]
Using Bayes’ rule

\[ p(A|B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)} \]

If a person has the disease, the probability that they are vaccinated is .67.

If a person does not have the disease, the probability that they are vaccinated is .94.

The probability that a person has the disease is .15.

If the person is vaccinated, what is the probability that the person has the disease?

**What is** \( p(\text{vaccinated}) \)?

\[
p(\text{vaccinated})
= p(\text{disease and vaccinated}) \text{ or } (\text{no disease and vaccinated})
= p(\text{disease and vaccinated}) + p(\text{no disease and vaccinated})
= p(\text{vaccinated | disease}) \times p(\text{disease})
  + p(\text{vaccinated | no disease}) \times p(\text{no disease})
= .67 \times .15 + .94 \times (1 - .15) = .90
\]

(this general idea is (again) called the law of total probability)
Using Bayes’ rule

\[ p(A|B) = \frac{p(A \text{ and } B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)} \]

If a person has the disease, the probability that they are vaccinated is .67

If a person does not have the disease, the probability that they are vaccinated is .94

The probability that a person has the disease is .15

If the person is vaccinated, what is the probability that the person has the disease?

\[
p(\text{disease} \mid \text{vaccinated}) = p(\text{vaccinated} \mid \text{disease}) \cdot p(\text{disease}) / p(\text{vaccinated})
\]

\[= .67 \cdot .15 / .90 = .11 \]
why is this useful?
one reason: if we are given the wrong conditional probabilities for our question, sometimes we can figure them out using Bayes' rule
Taxicabs

A taxi hit a pedestrian one night and fled the scene.

Here's what we know:

(a) There are only two taxi companies in town, 'Blue Cabs' and 'Green Cabs'. On the night in question, 85% of all taxis on the road were green and 15% were blue.

(b) A single witness saw the accident and says that he saw the pedestrian struck by a blue taxi. The witness has undergone an extensive vision test under conditions similar to those on the night in question, and has demonstrated that he can successfully distinguish a blue taxi from a green taxi 80% of the time.

Based on this information, what's the probability that the taxi that struck the pedestrian was blue? (Which taxi company should we investigate first?)

we’ve been given $p(\text{witness identified as blue} \mid \text{blue})$

is this the relevant conditional probability?

$p(\text{blue} \mid \text{witness identified as blue})$

$= p(\text{witness identified as blue} \mid \text{blue}) \times p(\text{blue}) / p(\text{witness identified as blue})$

$= .80 \times .15 / p(\text{witness identified as blue})$
A taxi hit a pedestrian one night and fled the scene. Here's what we know:

(a) There are only two taxi companies in town, 'Blue Cabs' and 'Green Cabs'. On the night in question, 85% of all taxis on the road were green and 15% were blue.

(b) A single witness saw the accident and says that he saw the pedestrian struck by a blue taxi. The witness has undergone an extensive vision test under conditions similar to those on the night in question, and has demonstrated that he can successfully distinguish a blue taxi from a green taxi 80% of the time.

Based on this information, what's the probability that the taxi that struck the pedestrian was blue? (Which taxi company should we investigate first?)

\[
p(\text{witness identified as blue}) = p(\text{witness IDs blue | blue}) \times p(\text{blue}) + p(\text{witness IDs blue | green}) \times p(\text{green})
\]

\[
= .80 \times .15 + .20 \times .84
\]

\[
= .29
\]
A taxi hit a pedestrian one night and fled the scene.

Here's what we know:

(a) There are only two taxi companies in town, 'Blue Cabs' and 'Green Cabs'. On the night in question, 85% of all taxis on the road were green and 15% were blue.

(b) A single witness saw the accident and says that he saw the pedestrian struck by a blue taxi. The witness has undergone an extensive vision test under conditions similar to those on the night in question, and has demonstrated that he can successfully distinguish a blue taxi from a green taxi 80% of the time.

Based on this information, what's the probability that the taxi that struck the pedestrian was blue? (Which taxi company should we investigate first?)

We’ve been given $p(\text{witness identified as blue} \mid \text{blue})$

Is this the relevant conditional probability?

$p(\text{blue} \mid \text{witness identified as blue})$

$= p(\text{witness identified as blue} \mid \text{blue}) \times p(\text{blue}) / p(\text{witness identified as blue})$

$= .80 \times .15 / .29 = .41$  

The most likely situation is that the taxi was green! (we need to consider the base rate)
Taxicabs, class guesses

\[ p(\text{blue} \mid \text{witness identified as blue}) = \frac{p(\text{witness identified as blue} \mid \text{blue}) \times p(\text{blue})}{p(\text{witness identified as blue})} \]

\[ = 0.80 \times 0.15 / 0.29 = 0.41 \]

\[ \text{the most likely situation is that the taxi was green!} \]
\[ \text{(we need to consider the base rate)} \]
This time

• How do probabilities change with multiple comparisons?

• How do we use conditional probabilities to ask and answer the right question?

• How can we calculate conditional probabilities using Bayes’ rule?

• Why do we need to treat non-independent data differently?
Two seats

There is one empty seat on a plane and two passengers who haven’t arrived at the gate. In this situation, passengers tend to show up 75% of the time. What is the probability that both passengers will show up?

If they are independent

→ p(one shows up) * p(other shows up)
→ .75 * .75 = .56

If they are not independent

→ p(one shows up | the other shows up) * p(the other shows up)

if they are traveling together
→ 1 * .75 = .75

if they are busy arguing over which one needs to go on the trip
→ 0 * .75 = 0
From earlier today

What are the odds that I will win Powerball?

1 in 292 million

What are the odds that someone will win Powerball?

1 - p(nobody wins)

1 - p(Person A doesn’t win and Person B doesn’t win and …)

1 - p(Person A doesn’t win) * p(Person B doesn’t win) * ...

1 - p(I don’t win)^number of players

with 50000000 independent players, 16% chance that someone will win

* this is a binomial probability problem with 0 successes, 50000000 trials, and \( \pi = (1 / 292M) \)

What if all players coordinate and choose the same number?

→ p(someone wins) = 1 / 292 million chance someone wins

What if all players coordinate to choose distinct numbers?

→ p(someone wins) = 50 million / 292 million = 17% chance someone wins
From earlier today

*Sally Clark on trial in 1998 for killing her two children, although she claimed they died accidentally in their sleep*

- Prosecution argument: probability of one child dying of SIDS is $1 / 8500 \rightarrow$ probability of two children in the same family dying of SIDS is $(1 / 8500)^2 = 1$ in 73 million (*mistake #1, these are not independent events*)

- If you know one child in a family had a disease, does this change the probability that another child in that family has the disease?
Failure models

July 19, 1989, United Flight 232 crash landed in Sioux City, Iowa

• Pilots lost control of plane after all three hydraulic systems failed

• Odds of all three hydraulic systems failing simultaneously previously calculated as a billion to one

• These calculations assumed that the systems failed independently, i.e., the failure of one system does not impact the probability of the failure of another system

• In this case, fragments from a different failing component punctured all three hydraulic systems simultaneously (not independent events)

• (And there have been multiple cases of all three systems failing simultaneously — suggesting that this was not a one in a billion event)
Independence matters

• Probabilities change dramatically depending on independence of events or observations

• Paired design: observations are *not* independent
  → must take this into account in our analyses

• Also must consider independence across multiple comparisons — some things that seem surprising might not be
Recap

• The ‘surprising-ness’ of an event depends on whether we were interested in that specific event occurring or one of many possible events occurring — implications for multiple comparisons in hypothesis testing

• Choosing the correct conditional probability matters

• We can use Bayes’ rule to find these correct conditional probabilities from partial information

• We need to account for whether our observations are independent
Questions