1 Probability

Probability

Review

- Counting
- Urns (sampling)
- Area

Words

Disjoint (mutually exclusive)

Complement of an event (set)

Independence

Conditional probability

Marginal probability

Joint probability

Conditional probability

Addition rule: We can add probabilities if the events are disjoint: $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

General addition rule: Think of Venn diagrams.

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Two events are independent of knowing the outcome of one gives no useful information about the outcome of the other and vice versa.

Multiplication rule for independent processes: $\Pr(A \cap B) = \Pr(A) \Pr(B)$

Conditional probability means prior is a smaller space, so just normalise to the smaller space.

Said differently, **conditional probability** $\Pr(A \mid B)$ is the probability of some event B conditioned on some event A.

The **marginal probability** is when we don't condition. It's just a single variable, but implicitly there are other variables.

Joint probability is taking multiple variables into account together.

So general multiplication rule: $\Pr(A \cap B) = \Pr(A \mid B) \Pr(B)$.

Draw addition and multiplication rules somewhere here.

Conditional probability

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

First show plot, then show equation, then highlight B, then highlight intersection, then highlight none.

Then show that if $B = \cup_i B_i$, then the sum is 1 (because sum was $\Pr(B)$).

Show that conditioning on an independent variable has no effect. Intuition, then picture, then formula with trivial cancellation.

$$P(A_1 \mid B) = \frac{\Pr(B \mid A_1)\Pr(A_1)}{\Pr(B \mid A_1)\Pr(A_1) + \cdots + \Pr(B \mid A_k)\Pr(A_k)}$$

Example

In any year, about .35% of the population is diagnosed with dread disease D.

The test for D isn't perfect:

- For people with D, 11% receive a (false) negative test result.
- For people without D, 7% receive a (false) positive test result.

We test person λ at random. The test says positive. What is the probability that λ has D?

What we know:

- $\Pr(D) = .0035$.
- $\Pr(- \mid D) = .011$ (and so $\Pr(+ \mid D) = .089$).
- $\Pr(+ \mid \neg D) = .07$ (and so $\Pr(- \mid \neg D) = .093$).

$$\begin{aligned}\Pr(D \mid +) &= \frac{\Pr(D \cap +)}{\Pr(+)} \\ &= \Pr(+ \mid D)\Pr(D) \\ &= .89 \times .0035 = .00312\end{aligned}$$

Note that $\Pr(+) = \Pr(+ \mid D) + \Pr(+ \mid \neg D)$.

$$\Pr(D \mid -) = \frac{\Pr(D \cap -)}{\Pr(-)} = .00038$$

$$\Pr(\neg D \mid +) = \frac{\Pr(\neg D \cap +)}{\Pr(+)} = ..06976$$

$$\Pr(\neg D \mid -) = \frac{\Pr(\neg D \cap -)}{\Pr(-)} = ..92675$$

Probability Density Function

```
import scipy.stats as ss x = np.linspace(ss.norm.ppf(.01),
ss.norm.ppf(.99), 100) plt.plot(x, ss.norm.pdf(x))
plt.show()
```

The pdf describes the distribution.

Cumulative Density Function

```
import scipy.stats as ss x = np.linspace(ss.norm.ppf(.01),
ss.norm.ppf(.99), 100) plt.plot(x, ss.norm.cdf(x))
plt.show()
```

The cdf also describes the distribution. It's the integral of the pdf.

Counting

Example: two children, what is $P(\exists \text{boy}) = P(\exists b)$?

Draw graphic of bb, bg, gb, gg with prob 1/4.

Draw graphic of bb, bg, gb, gg with prob 1/4 with circled if $\exists$ b.

What about $P(\exists b \mid \exists g)$?

Same graphic, fade bb, circles, show 2/3.

Events have probabilities summing to 1, cover, and are disjoint.

Counting: Count all possibilities (which are equally likely), then probability is the fraction of the collection.

Random processes:

A random variable (roughly) is a variable that has some probability of taking on different sets of events ($X : \Omega \to E$). That use of "probability" is circular. A random process is the time evolution of a random variable.

The probability is the proportion of times the outcome occurs if we observe the random process infinitely long.

It's sometimes useful to model things as random that aren't: traffic, human behaviour, ...

Law of Large Numbers: The larger the sample, the closer to the actual probability. Still deviations, but less.

Example: Brownian motion. Vacuum on one side of the room.

Pólya urn model

r red balls, b black balls in an urn.

- Sampling with replacement and duplication (rich get richer)
- Variations: balls into boxes
  - Sampling with replacement
  - Sampling without replacement

Want to know sequence of colours selected, evolution of colour distribution in urn.

Sampling

- We don't know r, b, or even r/b. Estimate.

Area

- Continuous case.
- Events are regions.
- Same constraints: sum of areas is unity. Cover and disjoint.
- Zero probability events!

Discrete: throwing darts at a checker board.

- What's the probability of hitting a white square?
- P(white square in bottom left quadrant)

Continuous: throwing darts at the same board but no colour borders.

- P(region with colour of reddish hue)
- P(region with reddish hue and within 2 cm of blue region)

Distributions

Soon. . .

2 Statistics

Statistics

What is statistics?

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Sadly, sometimes people forget 1. Statistics is about making 2–4 efficient, rigorous, and meaningful.

OpenIntro Statistics, 2nd edition, D. Diez, C. Barr, M. Çetinkaya-Rundel, 2013.

Another view:

Statistics is about comparing observed reality to a model.

Sometimes we can even measure how far off from the model we are.

What is data science?

(Exercise: Is this the same question as the last slide?)

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

What the public thinks.

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data

5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

Where we spend most of our time.

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

The easiest part to forget.

http://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-tha

Good statistical models are often relatively simple.

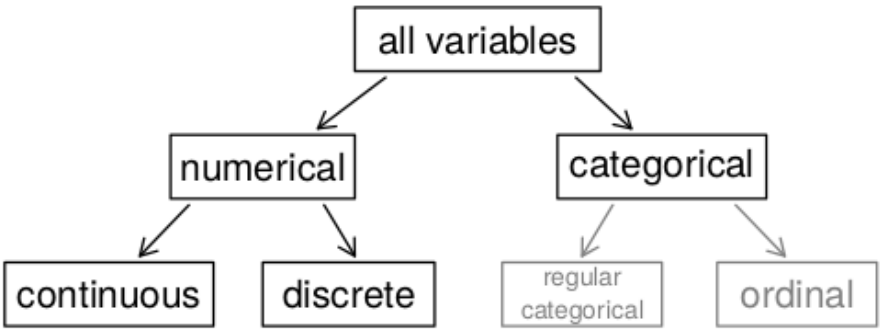
Significance

- Flip a coin N times. Get 49% heads. Is the coin fair?
- Congress/parliament has N members of whom m are male. Do we discriminate against women?
- Do we discriminate against women more or less than we discriminate against immigrants?
- Email classification: spam or not?
- Careful about what we conclude. Here we can't conclude anything about cause or source.

We'll also have to come back to it when we understand about p values and the like.

Words

A **case** or observational unit is an observation or a single sample. E.g., an email.
A **variable** is a characteristic of an observational unit. E.g., number of lines, mean line length, number of words.



Categorical is sometimes called nominal. Categorical variables with logical ordering (empty, half, full) or called ordinal.

Anecdote

Some properties of anecdote:

- is data
- haphazardly collected
- is generally not representative
- sometimes result of selective retention
- does not accumulate to be representative
- might be true (by chance)
- is ok to use as hypothesis, but be clear that hypothesis is anecdote

Study Types

- Observational
- Experimental

What can go wrong?

- Forgetting that association $\neq$ causation
- Not random
- Confounding variables

Example: Sunscreen use associated with skin cancer. But sun exposure is a confounding variable.

Observational studies: prospective (identify individuals, collect information) and retrospective. (Can combine the two.)

Observational studies

Can't generally conclude causation.

Discuss.

Experiment

We do stuff. Maybe randomised experiment.

If properly designed, can conclude causation.

Experiments to generate hypotheses vs experiments to demonstrate hypotheses. (Don't use the same data!)

Example: Phase 0 (10 people) and phase 1 (20-100 people) clinical trials are not randomised.

Phase 0: Pharmacodynamics and pharmacokinetics, particularly oral bioavailability and half-life of the drug. Very small, subtherapeutic doses.

Phase 1: Testing of drug on healthy volunteers for dose-ranging. Often subtherapeutic, but with ascending doses. Determine whether a drug is safe to check for efficacy in phase 2.

- Controlling
- Randomisation
- Replication
- Blocking (optional, advanced)

Control: hold other variables constant. E.g., drinking pill with full glass of water even if we don't care.

Random: Cancel out effects we can't control.

Replication: Enough participants.

Blocking: Subsections if we think some variable influences response.

- H_0: Independent (or null) hypothesis
- H_A: Alternative hypothesis

Simple random sampling

Uniform and independent.

- Uniform: Each case has the same probability of being included.
- Independent: Knowing that a particular case is included provides no useful information about other cases.

Example: Footballers in Europe, put all the names in a bucket.

What can go wrong:

- Not actually random
- Convenience sample
- Non-response bias

Stratified Sampling

Group like with like, then usually simple random sampling of each stratum.

Divide and conquer.

Example: Footballers in Europe, sample k from each team. (Hypothesis: footballers are paid about the same within a team.)

What will go wrong:

- Harder to analyse than simple random sampling.

Cluster Sampling

Simple random sampling to form clusters, then simple random sampling.

Example: Sample k from each village. Works if villages have high diversity, not so much if villages are substantially different from one another.

Example: Ebola. SRS expensive (census of each village, then maybe have to visit them all). So select n villages and sample k within each. Hope that villages are similar enough that this is valid as a first step.

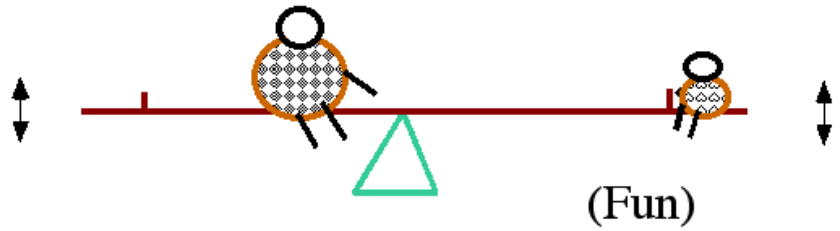
What will go wrong:

- Harder to analyse than simple random sampling.

Mean

- Weighted and unweighted
- Centroid to physicists

Sample mean vs true (population) mean.



$$\mu = E(X) = \sum w_i x_i = \mathbf{w} \cdot \mathbf{x}$$

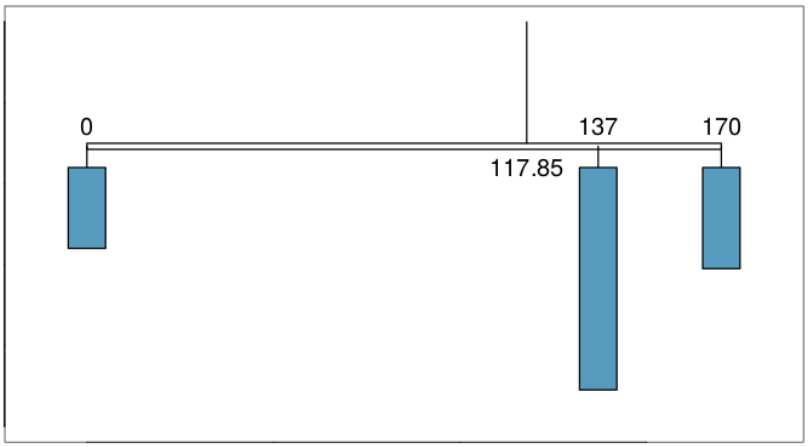
http://telescopes.stardate.org/images/research/teeter-totter/TT4.gif

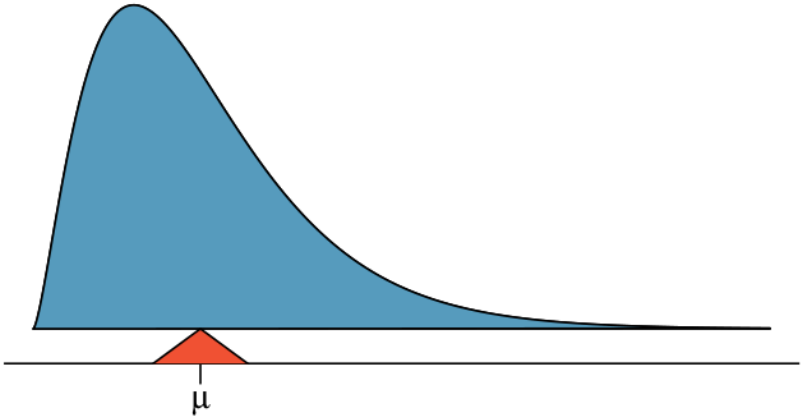
$$\mu = E(X) = \sum \Pr(X = x_i) x_i$$

http://telescopes.stardate.org/images/research/teeter-totter/TT4.gif

$$\mu = E(X) = \int x f(x)\,\mathrm{d}x$$

http://telescopes.stardate.org/images/research/teeter-totter/TT4.gif





Variation

Deviation is distance from mean.

Variance is mean square of deviations

Standard deviation is square root of variance

Sample standard deviation (s) and sample variance (s^2), divide by $n-1$.

Population standard deviation (σ) and population variance (σ^2), divide by n.

Population statistics may not be feasible or possible to compute.

$$s^2 = \frac{(\overline{x} - x_1)^2 + \cdots (\overline{x} - x_n)^2}{n-1}$$

$$\sigma^2 = \frac{(\overline{x} - x_1)^2 + \cdots (\overline{x} - x_n)^2}{n}$$

$$\text{Var}(X) = \sigma^2 = (\overline{x} - x_1)^2 \Pr(X = x_1) + \cdots (\overline{x} - x_n)^2 \Pr(X = x_n)$$

Linear combinations of random variables

Assuming X and Y are independent:

$$\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y)$$

$$\mathbf{Var}(aX + bY) = a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y)$$

Games

Multiple choice exam, 10 questions, all have four choices of which precisely one is correct. Guess at random.

- Probability of getting 100%? $\left(\frac{1}{4}\right)^{10}$
- Probability of getting 0%? $\left(\frac{3}{4}\right)^{10}$
- Expected score? $\mathbf{E}(X) = 10\left(\frac{1}{4}\right) = \frac{5}{2}$
- Standard deviation $\mathbf{Var}(X) = 10\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{30}{16} = \frac{15}{8}$

We'll come back to: if student gets 9 of 10 correct, do we really believe that they guessed at random?

Experimental design

A telephone survey company dials digits at random (some of them aren't valid telephone numbers) rather than using a phone directory. A statistics student wants to determine if social network usage among his peers (students at his school) affects academic performance. Possibilities:

- He messages his class snapchat group.
- He posts paper fliers at his university.
- He gets a list of students from the registrar. (What about opt-in/opt-out policies? What if he doesn't know how many have opted out/failed to opt in?
- Other?

Due to a flu epidemic, 10% of students in a class miss the (final) exam. The professor decides to offer a make-up exam but to require 11% of the students who took the original exam to take the make-up exam as well in order to calibrate the new exam and make results comparable.

Discuss. Explain flaws. Discuss what researcher should have done differently.

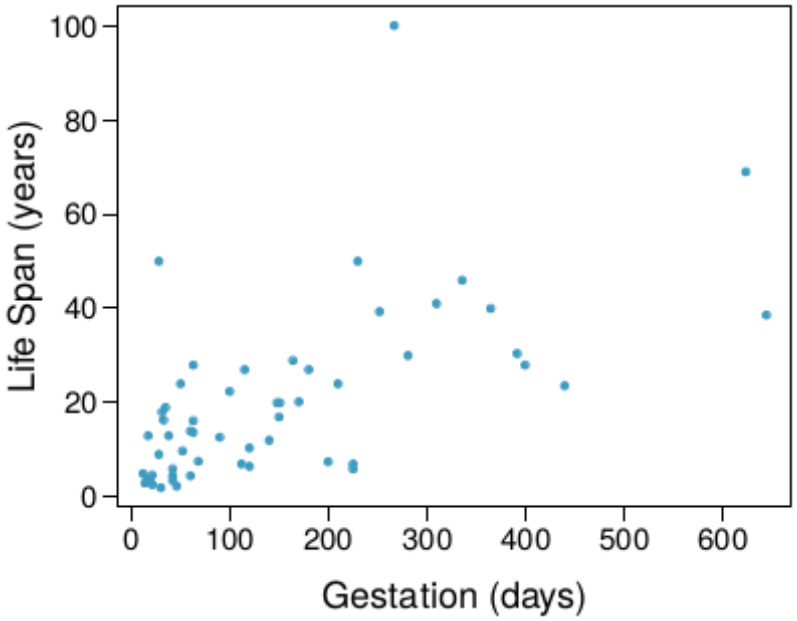
Reasoning about data

Students at an elementary school are given a questionnaire that they are required to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude thata great majority of the parents have no difficulty spending time with their kids after school. A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later, however, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers. An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems. The news reports that 27 year old Andreas Lubitz, co-pilot of crashed flight 4U 9525, had researched suicide before killing himself.

Explain flaws. Discuss what researcher should have done differently.

Example

Mammals



Discuss. Are the variables independent?

Questions?

`purple.com/talk-feedback`

3 Break

Break