Week 5: Probability

Goals this week

We are taking a turn this wee[illegible]d inference. You will likely find this week and the next (and the problem set) more difficult th[illegible]bout probability is essential for understanding how the majority of data science and machine le[illegible]work. Even if you don't go on to fit probabilistic models directly on your own real-life data, I gua[illegible]s you do use will be rooted in probability calculus (from clustering to regression to neural network[illegible]tion). By learning the underlying fundamentals, you will (1) better remember a method's functio[illegible]cally, (2) understand a method's underlying assumptions and whether those assumptions are met in your data analysis.

The probability of boys and girls

In the 1770's, the French mathematician Pierre Laplace started working on big data. He became interested in the curious biological observation that the ratio of boys to girls at birth isn't exactly 50:50. Village records showed an apparent bias towards more boy births, but the numbers were small and vulnerable to statistical fluctuation. Laplace set out to use some new ideas about probability theory to put the question of sex ratio at birth to a mathematically rigorous test, and he needed big data sets to do it.

Laplace turned to the extensive census records in London and Paris. The Paris census for the years 1745-1770 recorded the birth of 251527 boys and 241945 girls, a 51:49 (1.04) ratio. The London census for 1664-1757 recorded the birth of 737629 boys and 698958 girls, again a 51:49 ratio.

Laplace's work is one of the origins of probability theory. Today, his laborious manual calculations are easy to reproduce in a few lines of Python, and his problem makes a compact example for us to illustrate some key ideas of probabilistic inference

The binomial distribution

Let's call the probability of having a boy p. The probability of having a girl is $1-p$. The probability of having b boys in N total births is given by a binomial distribution:

$$P(b \mid p, N) = \binom{N}{b} p^b (1-p)^{N-b}$$

A couple of things to explain, in case you haven't seen them before:

- $P(b \mid p, N)$ is "the probability of b, given p and N": a **conditional probability**. The vertical line $\mid$ means "given". That is: if I told you p and N, what's the probability of observing data b?
- $\binom{N}{b}$ is conventional shorthand for the binomial coefficient: $\frac{N!}{b!(N-b)!}$.

Suppose $p=0.5$, and $N=493472$ in the Paris data (251527 boys + 241945 girls). The probability of getting 251527 boys is:

$$P(b \mid p, N) = \frac{493472!}{251527!\,241945!} 0.5^{251527} 0.5^{241945}$$

Your calculator isn't likely to be able to deal with that, but Python can. For example, you can use the `pmf` (probability mass function) of `scipy.stats.binom`:

```
import scipy.stats as stats
p = 0.5
b = 251527
N = 493472
Prob = stats.binom.pmf(b, N, p)
print(Prob)
```

which gives $4.5 \cdot 10^{-44}$.

Probabilities sum to one, so You can verify this in Python easily, because there are only $N+1$ possible values for b, from 0

Maximum likelihood estimate of p

Laplace's goal isn't to calculate the probability of the observed data, it's to infer what p is, given the observed census data. One way to approach this is ask, what is the best p that explains the data – what is the value of p that maximizes $P(b \mid p, N)$?

It's easily shown (by taking the derivative of $P(b \mid p, N)$ with respect to p) that this optimal p is just the frequency of boys, $\hat{p} = \frac{251527}{493472} = 0.51$. The hat on $\hat{p}$ denotes an estimated parameter that's been fitted to data.

With $\hat{p} = 0.51$, we get $P(b \mid \hat{p}, N) = 0.001$. So even with the best $\hat{p}$, it's improbable that we would have observed exactly b boys, simply because there are many other b that could have happened. The probability of the data is not the probability of p; $P(b \mid p, N)$ is not $P(p \mid b, N)$. When I deal you a five card poker hand, it's laughably unlikely that I would have dealt you exactly those five cards if I were dealing fairly, but that doesn't mean you should reach for your revolver.

It seems like there ought to be some relationship, though. Our optimal $\hat{p} = 0.51$ does seem like a much better explanation of the observed data b than $p = 0.5$ is.

$P(b \mid p, N)$ is called the **likelihood** of p, signifying our intuition that $P(b \mid p, N)$ seems like it should be a relative measure of how well a given p explains our observed data b. We call $\hat{p}$ the **maximum likelihood** estimate of p.

The London and Paris data have different maximum likelihood values of p: 0.5135 versus 0.5097. Besides asking whether the birth sex ratio is 50:50, we might even ask, is the ratio the same in London as it is in Paris?

The probability of p

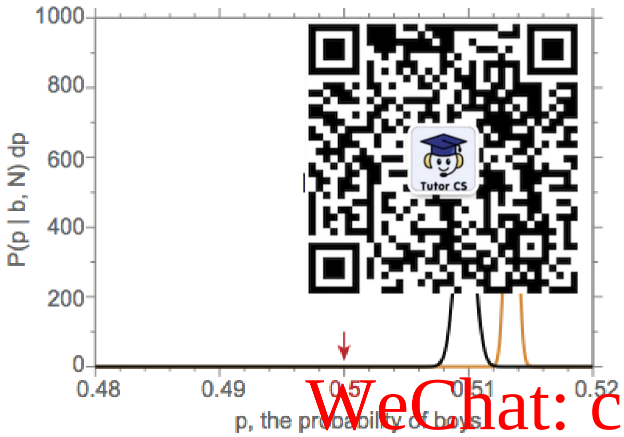
Laplace made an intuitive leap. He reasoned that one value p_1 is more probable than another p_2 by the ratio of these probabilities:

$$\frac{P(p_1 \mid b, N)}{P(p_2 \mid b, N)} = \frac{P(b \mid p_1, N)}{P(b \mid p_2, N)}$$

if we had no other reason to favor one value for p over the other (we'll see later that Laplace implicitly assumed a uniform "prior" for p). So $p = 0.51$ is $\frac{0.001}{4.5e{-44}} \sim 10^{40}$-fold more probable than $p = 0.5$, given the Paris data. This relationship also implies that we can obtain a probability distribution for p by normalizing over the sum over all possible p in the likelihood, which requires an integral, not just a simple sum, since p is continuous:

$$P(p \mid b, N) = \frac{P(b \mid p, N)}{\int_0^1 P(b \mid p, N) dp}$$

This is a **probability density**, because p is continuous. Strictly speaking, the probability of any specific value of p is zero, because there are an infinite number of values of p, and $\int_0^1 P(p \mid b, N)\, dp$ has to be 1.



It's possible (and indeed, it frequently happens) that a probability density function like $P(p \mid b, N)$ can be greater than 1.0 over a small range of p, so don't be confused if you see that; it's the integral $\int_0^1 P(p \mid b, N)\, dp = 1$ that counts. For example, you can see that there are large values of $P(p \mid b, N)\, dp$ in the figure above, where I've plotted Laplace's probability densities for the Paris and London data.

In the figure, it seems clear that $p = 0.5$ is not supported by either the Paris or London data. We also see that the uncertainty around p^{london} does not overlap with the uncertainty around p^{paris}. It appears that the birth sex ratio in Paris and London is different.

We're just eyeballing though, when we say that it seems that the two distributions for p don't overlap 0.5, nor do they overlap each other. Can we be more quantitative?

The cumulative probability of p

Because the probability at any given continuous value of p is actually zero, it's hard to frame a question like "is $p = 0.5$?". Instead, Laplace now framed a question with a probability he could calculate: **what is the probability that p <= 0.5?** If that probability is tiny, then we have strong evidence that $p > 0.5$.

A cumulative probability distribution $F(x)$ is the probability that a variable takes on a value less than or equal to x. For a continuous real-valued variable x with a probability density function $P(x)$, $F(x) = \int_{-\infty}^{x} P(x)$. For a continuous probability p constrained to the range $0..1$, $F(p) = \int_0^p P(p)$ and $p \leq 1$.

So Laplace framed his question as:

$$P(p \leq 0.5 \mid b, N) = \frac{\int_0^{0.5} P(b \mid p, N) dp}{\int_0^1 P(b \mid p, N) dp}$$

Then Laplace spent a bazillion pages working out those integrals by hand, obtaining an estimated log probability of -42.0615089 (i.e., a probability of $8.7 \cdot 10^{-43}$): decisive evidence that the probability p of having a boy must be greater than 0.5.

These days we can replace Laplace's virtuosic calculations and approximations with one call to Python:

```
import scipy.special as special
p = 0.5
b = 251527
N = 493472

answer = special.betainc(b+1, N-b+1, p)
print (answer)
```


which gives $1.1 \cdot 10^{-42}$. Lap

Beta integrals

I probably won't have time to e board in lecture, but I weant to quickly explain what this `scipy.special.betainc` funct ed a Beta integral.

The complete Beta integral $B(a,b)$ is:

$$B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}\,dp = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

A gamma function $\Gamma(x)$ is a generalization of the factorial from integers to real numbers. For integer a,

$$\Gamma(a) = (a-1)!$$

and conversely

$$a! = \Gamma(a+1)$$

The incomplete Beta integral $B(x;a,b)$ is:

$$B(x;a,b) = \int_0^x p^{a-1}(1-p)^{b-1}\,dp$$

and has no clean analytical expression, but statistics packages typically give you a computational method for calculating it – hence the SciPy scipy.special.betainc function.

If you wrote out Laplace's problem in terms of binomial probability distributions for $P(b \mid p, N)$, you'd see the binomial coefficient cancel out (it's a constant with respect to p), leaving:

$$P(p \leq 0.5 \mid b, N) = \frac{\int_0^{0.5} P(b \mid p, N)dp}{\int_0^1 P(b \mid p, N)dp} = \frac{B(0.5; b+1, N-b+1)}{B(b+1, N-b+1)}$$

Alas, reading documentation in Python is usually essential, and it turns out that the scipy.special.betainc function doesn't just calculate the incomplete beta function; it sneakily calculates a regularized incomplete beta integral $\frac{B(x;a,b)}{B(a,b)}$ by default, which is why all we needed to do was call:

```
answer = special.betainc(b+1, N-b+1, p)
```

Summary

Laplace treated the unknown parameter p like it was something he could infer, and express an uncertain probability distribution over it. He obtained that distribution by inverse probability: by using $P(b|p)$, the probability of the data if the parameter were known and given, to calculate $P(p|b)$, the probability of the unknown parameter given the data.

Laplace's reasoning was clear, and proved to be influential. Soon he realized that the Reverend Thomas Bayes, in England, had derived a very similar approach to inverse probability just a few years earlier. We'll learn about Bayes' 1763 paper in a bit. But for now, let's leave Laplace and Bayes, and lay out some basic terminology of probabilities and probabilistic inference.

A minicourse in probability calculus

1. Random variabl[illegible]

A random variable is someth[illegible]lue. The value might be discrete (like "boy" or "girl", or a roll of a die 1..6) or it might be real-value[illegible]rawn from a Gaussian distribution). We'll denote random variables or events with capital letters, [illegible]es or outcomes with small letters, like x.

When we say $P(X)$ (the pro[illegible]le X), we are envisioning a set of values $P(X=x)$, the probability that we could get each possible outcome x.

Probabilities sum to one. If X has discrete outcomes x, $\sum_x P(X=x) = 1$. If X has continuous outcomes x, $\int_{-\infty}^{\infty} P(X=x) = 1$.

For example, suppose we have a fair die, and a loaded die. With the fair die, the probability of each outcome 1..6 is $\frac{1}{6}$. With the loaded die, let's suppose that the probability of rolling a six is $\frac{1}{2}$, and the probability of rolling anything else (1..5) is 0.1. We have a bag with fair dice and loaded dice in it. We pick a die out of the bag randomly and roll it. What's the probability of rolling 1, 2, 3, 4, 5 or 6? We have two random variables in this example: let's call D the outcome of whether we chose a fair or a loaded die, and R the outcome of our roll. D takes on values f or l (fair or loaded); R takes on values 1..6.

2. Conditional probability

$P(X \mid Y)$ is a conditional probability distribution: the probability that X takes on some value, given a value of Y.

To put numbers to a discrete conditional probability distribution $P(X \mid Y)$, envision a table with a row for each variable Y, and a column for each variable X. Each row sums to one: $\sum_X P(X \mid Y) = 1$.

In our example, I told you $P(R \mid D)$: the probability of rolling the possible outcomes 1..6, when you know whether the die is fair or loaded.

roll $R =$	1	2	3	4	5	6
D = fair:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
D = loaded:	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{2}$

3. Joint probability

$P(X, Y)$ is a joint probability: the probability that X takes on some value and Y takes on some value.

Again, envision a table with a row (or column) for each variable Y, and a column (or row) for each variable X – but now the whole table sums to one, $\sum_{XY} P(X, Y) = 1$.

For instance, we might want to know the probability that we chose a loaded die and we rolled a six. You don't know the joint distribution yet in our example, because I haven't given you enough information.

4. Relationship between conditional and joint probability

The joint probability that X and Y both happen is the probability that Y happens, then X happens given Y:

$$P(X,Y) = P(X \mid Y)P(Y)$$

Also, conversely, because we're not talking about causality (with a direction), only about statistical dependency:

$$P(X,Y) = P(Y \mid X)P(X)$$

so:

$$P(Y) = P(Y \mid X)P(X)$$

So for our example, let's sup of choosing a fair die from the bag is $\frac{9}{10}$, and the probability of choosing a loaded one is $\frac{1}{10}$. can calculate the joint probability distribution $P(R,D)$ as $P(R \mid D)P(D)$.

roll $R =$	1	2	3	4	5	6
D = fair:	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$
D = loaded:	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{20}$

If you have additional random variables in play, they stay where they are on the left or right side of the $\mid$. For example, if the joint probability of X, Y was conditional on Z:

$$P(X,Y \mid Z) = P(Y \mid X,Z)P(X \mid Z) = P(X \mid Y,Z)P(Y \mid Z)$$

Or, if I start from the joint probability $P(X,Y,Z)$:

$$P(X,Y,Z) = P(Y,Z \mid X)P(X) = P(X,Z \mid Y)P(Y)$$

5. Marginalization

If we have a joint distribution like $P(X,Y)$, we can "get rid" of one of the variables X or Y by summing it out:

$$P(X) = \sum_Y P(X,Y)$$

It's called marginalization because imagine a 2-D table with rows for Y's values and columns for X's values. Each entry in the table is $P(X,Y)$ for two specific values x, y. If you sum across the columns to the right margin, your row sums give you $P(Y)$. If you sum down the rows to the bottom margin of the table, your column sums give you $P(X)$. When we obtain a distribution $P(X)$ by marginalizing $P(X,Y)$, we say $P(X)$ is the **marginal distribution** of X.

In our example, we can marginalize our joint probability matrix:

roll $R =$	1	2	3	4	5	6	$P(D)$
D = fair:	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$	$\frac{9}{60}$	**0.9**
D = loaded:	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{20}$	**0.1**
$P(R)$:	**0.16**	**0.16**	**0.16**	**0.16**	**0.16**	**0.20**	

Now we have the marginal distribution $P(R)$. This is the probability that you're going to observe a specific roll of 1..6, if you don't know what kind of die you pulled out of the bag. You've marginalized over your uncertainty of an unknown variable Y. Because sometimes you pull a loaded die out of the bag, the probability that you're going to roll a six is slightly higher than $\frac{1}{6}$.

If I have additional random variables in play, again they stay where they are. Thus:

$$P(X \mid Z) = \sum_Y P(X, Y \mid Z)$$

and:

$$P(X, Z) = \sum_Y P(X, Y, Z).$$

6. Independence

Two random variables X and Y are independent if:

$$P(X, Y) = P(X)P(Y)$$

which necessarily also means:

$$P(X \mid Y) = P(X)$$
$$P(Y \mid X) = P(Y)$$

In our example, the outcome of a die roll R is not independent of the die type D, of course. However, if I chose a die and rolled it N times, we can assume the individual rolls are independent, and the joint probability of those rolls could be factored into a product of their individual probabilities:

$$P(X_1 \dots X_N \mid D) = \prod_{i=1}^{N} P(X_i \mid D)$$

In probability modeling, we will often use independence assumptions to break a big joint probability distribution down into a smaller set of terms, to reduce the number of parameters in our models. The most careful way to invoke an independence assumption is in two steps: first write the joint probability out as a product of conditional probabilities, then specify which conditioning variables are going to be dropped. For example, we can write $P(X, Y, Z)$ as:

$$P(X, Y, Z) = P(X \mid Y, Z)P(Y \mid Z)P(Z)$$

Then state, “and I assume Y is independent of Z, so:”

$$\simeq P(X \mid Y, Z)P(Y)P(Z)$$

It’s possible to have a situation where X is dependent on Y in $P(X \mid Y)$, but when a variable Z is introduced, $P(X \mid Y, Z) = P(X \mid Z)$. In this case we say that X is conditionally independent of Y given Z. For example, Y could cause Z, and Z could cause X; Y‘s effect on X is entirely through Z. This starts to get at ideas from Bayesian networks, a class of methods that give us tools for manipulating conditional dependencies and doing inference in complicated networks.

7. Bayes’ theorem

We’re allowed to apply the above rules repeatedly, algebraically, to manipulate probabilities. Suppose we know $P(X \mid Y)$ but we want to know $P(Y \mid X)$. From the definition of conditional probability we can obtain:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

and from the definition of marginalization we know:

$$P(X) = \sum_Y P(X,Y) = \sum_Y P(X \mid Y)P(Y)$$

Congratulations, you've just derived and proven Bayes' theorem:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{\sum_Y P(X \mid Y)P(Y)}$$

If you assume that $P(Y)$ is a [illegible] or) it cancels out, and you recognize Laplace's inverse probability calculation.

Less trivially than just talking [illegible] om variables, we can use Bayes' theorem when talking more generally about observed da[illegible] I:

$$[illegible] = \frac{P(D \mid H)P(H)}{\sum_H P(D \mid H)P(H)}$$

The probability of our hypothesis, given the observed data, is proportional to the probability of the data given the hypothesis, times the probability of the hypothesis before you saw any data. The denominator, the normalization factor, is $P(D)$: the probability of the data summed over all possible hypotheses.

$P(D \mid H)$, the probability of the data, is usually the easiest bit. This is often called the likelihood. (It's the probability of the data D; it's the likelihood of the model H.)

$P(H)$ is the prior.

$P(D)$ is sometimes called the evidence: the marginal probability of the data, summed over all the possible hypotheses that could've generated it.

$P(H \mid D)$ is called the posterior probability of H.

So Bayes' theorem gives us a principled way to calculate the posterior probability of a hypothesis H, given data D that we've observed.

But: where do we get $P(H)$ from, if it's supposed to be a probability of something before any data have arrived? We may have to make a subjective assumption about it, like saying we assume a uniform prior: assume that all hypotheses H are equiprobable before the data arrive.

How do we enumerate all possible hypotheses H? Sometimes we'll be in a hypothesis test situation of explicitly comparing one hypothesis against another, but in general, there's always more we could come up with.

And what does it mean to talk about the probability of a hypothesis?

Further reading

- Sean Eddy and David J.C. MacKay, Is the Pope the Pope? (https://www.nature.com/articles/382490a0.pdf), Nature, 1996.
- Sean Eddy, What is Bayesian statistics? (http://www.nature.com/nbt/journal/v22/n9/full/nbt0904-1177.html), Nature Biotechnology, 2004.
- David J.C. Mackay, Bayesian Interpolation (https://pdfs.semanticscholar.org/8e68/c54f39e87daf3a8bdc0ee005aece3c652d11.pdf), Neural Computation, 1992.