Introduction to Statistics
Class 10, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the three overlapping "phases" of statistical practice.
2. Know what is meant by the term *statistic*.

2 Introduction to statistics

Statistics deals with data. Generally speaking, the goal of statistics is to make inferences based on data. We can divide this the process into three phases: collecting data, describing data and analyzing data. This fits into the paradigm of the scientific method. We make hypotheses about what's true, collect data in experiments, describe the results, and then infer from the results the strength of the evidence concerning our hypotheses.

2.1 Experimental design

The design of an experiment is crucial to making sure the collected data is useful. The adage 'garbage in, garbage out' applies here. A poorly designed experiment will produce poor quality data, from which it may be impossible to draw useful, valid inferences. To quote R.A. Fisher one of the founders of modern statistics,

> To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

2.2 Descriptive statistics

Raw data often takes the form of a massive list, array, or database of labels and numbers. To make sense of the data, we can calculate summary statistics like the mean, median, and interquartile range. We can also visualize the data using graphical devices like histograms, scatterplots, and the empirical cdf. These methods are useful for both communicating and exploring the data to gain insight into its structure, such as whether it might follow a familiar probability distribution.

2.3 Inferential statistics

Ultimately we want to draw inferences about the world. Often this takes the form of specifying a statistical model for the random process by which the data arises. For example, suppose the data takes the form of a series of measurements whose error we believe follows a normal distribution. (Note this is always an approximation since we know the error must

have some bound while a normal distribution has range $(-\infty, \infty)$.) We might then use the data to provide evidence for or against this hypothesis. Our focus in 18.05 will be on how to use data to draw inferences about model parameters. For example, assuming gestational length follows a $N(\mu, \sigma)$ distribution, we'll use the data of the gestational lengths of, say, 500 pregnancies to draw inferences about the values of the parameters μ and σ. Similarly, we may model the result of a two-candidate election by a Bernoulli(p) distribution, and use poll data to draw inferences about the value of p.

We can rarely make definitive statements about such parameters because the data itself comes from a random process (such as choosing who to poll). Rather, our statistical evidence will always involve probability statements. Unfortunately, the media and public at large are wont to misunderstand the probabilistic meaning of statistical statements. In fact, researchers themselves often commit the same errors. In this course, we will emphasize the meaning of statistical statements alongside the methods which produce them.

Example 1. To study the effectiveness of new treatment for cancer, patients are recruited and then divided into an experimental group and a control group. The experimental group is given the new treatment and the control group receives the current standard of care. Data collected from the patients might include demographic information, medical history, initial state of cancer, progression of the cancer over time, treatment cost, and the effect of the treatment on tumor size, remission rates, longevity, and quality of life. The data will be used to make inferences about the effectiveness of the new treatment compared to the current standard of care.

Notice that this study will go through all three phases described above. The experimental design must specify the size of the study, who will be eligible to join, how the experimental and control groups will be chosen, how the treatments will be administered, whether or not the subjects or doctors know who is getting which treatment, and precisely what data will be collected, among other things. Once the data is collected it must be described and analyzed to determine whether it supports the hypothesis that the new treatment is more (or less) effective than the current one(s), and by how much. These statistical conclusions will be framed as precise statements involving probabilities.

As noted above, misinterpreting the exact meaning of statistical statements is a common source of error which has led to tragedy on more than one occasion.

Example 2. In 1999 in Great Britain, Sally Clark was convicted of murdering her two children after each child died weeks after birth (the first in 1996, the second in 1998). Her conviction was largely based on a faulty use of statistics to rule out sudden infant death syndrome. Though her conviction was overturned in 2003, she developed serious psychiatric problems during and after her imprisonment and died of alcohol poisoning in 2007. See http://en.wikipedia.org/wiki/Sally_Clark

This TED talk discusses the Sally Clark case and other instances of poor statistical intuition: http://www.youtube.com/watch?v=kLmzxmRcUTo

2.4 What is *a* statistic?

We give a simple definition whose meaning is best elucidated by examples.

Definition. A statistic is anything that can be computed from the collected data.

Example 3. Consider the data of 1000 rolls of a die. All of the following are statistics: the average of the 1000 rolls; the number of times a 6 was rolled; the sum of the squares of the rolls minus the number of even rolls. It's hard to imagine how we would use the last example, but it is a statistic. On the other hand, the probability of rolling a 6 is *not* a statistic, whether or not the die is truly fair. Rather this probability is a property of the die (and the way we roll it) which we can estimate using the data. Such an estimate is given by the statistic 'proportion of the rolls that were 6'.

Example 4. Suppose we treat a group of cancer patients with a new procedure and collect data on how long they survive post-treatment. From the data we can compute the average survival time of patients in the group. We might employ this statistic as an estimate of the average survival time for future cancer patients following the new procedure. The actual survival is *not* a statistic.

Example 5. Suppose we ask 1000 residents whether or not they support the proposal to legalize marijuana in Massachusetts. The proportion of the 1000 who support the proposal is a statistic. The proportion of all Massachusetts residents who support the proposal is *not* a statistic since we have not queried every single one (note the word "collected" in the definition). Rather, we hope to draw a statistical conclusion about the state-wide proportion based on the data of our random sample.

The following are two general types of statistics we will use in 18.05.

1. Point statistics: a single value computed from data, such as the sample average $\overline{x}_n$ or the sample standard deviation s_n.
2. Interval statistics: an interval $[a, b]$ computed from the data. This is really just a pair of point statistics, and will often be presented in the form $\overline{x} \pm s$.

3 Review of Bayes' theorem

We cannot stress strongly enough how important Bayes' theorem is to our view of inferential statistics. Recall that Bayes' theorem allows us to 'invert' conditional probabilities. That is, if H and D are events, then Bayes' theorem says

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

In scientific experiments we start with a hypothesis and collect data to test the hypothesis. We will often let H represent the event 'our hypothesis is true' and let D be the collected data. In these words Bayes' theorem says

$$P(\text{hypothesis is true} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis is true}) \cdot P(\text{hypothesis is true})}{P(\text{data})}$$

The left-hand term is the probability our hypothesis is true given the data we collected. This is precisely what we'd like to know. When all the probabilities on the right are known exactly, we can compute the probability on the left exactly. This will be our focus next week. Unfortunately, in practice we rarely know the exact values of all the terms on the

right. Statisticians have developed a number of ways to cope with this lack of knowledge and still make useful inferences. We will be exploring these methods for the rest of the course.

Example 6. Screening for a disease redux
Suppose a screening test for a disease has a 1% false positive rate and a 1% false negative rate. Suppose also that the rate of the disease in the population is 0.002. Finally suppose a randomly selected person tests positive. In the language of hypothesis and data we have:
Hypothesis: $H =$ 'the person has the disease'
Data: $D =$ 'the test was positive.'
What we want to know: $P(H|D) = P(\text{the person has the disease} \mid \text{a positive test})$

In this example all the probabilities on the right are known so we can use Bayes' theorem to compute what we want to know.

$$\begin{aligned} P(\text{hypothesis} \mid \text{data}) &= P(\text{the person has the disease} \mid \text{a positive test}) \\ &= P(H|D) \\ &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{.99 \cdot .002}{.99 \cdot .002 + .01 \cdot .998} \\ &= 0.166 \end{aligned}$$

Before the test we would have said the probability the person had the disease was 0.002. After the test we see the probability is 0.166. That is, the positive test provides some evidence that the person has the disease.

Maximum Likelihood Estimates
Class 10, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to define the likelihood function for a parametric model given data.
2. Be able to compute the maximum likelihood estimate of unknown parameter(s).

2 Introduction

Suppose we know we have data consisting of values $x_1, \ldots, x_n$ drawn from an exponential distribution. The question remains: which exponential distribution?!

We have casually referred to *the* exponential distribution or *the* binomial distribution or *the* normal distribution. In fact the exponential distribution $\exp(\lambda)$ is not a single distribution but rather a one-parameter family of distributions. Each value of λ defines a different distribution in the family, with pdf $f_\lambda(x) = \lambda e^{-\lambda x}$ on $[0, \infty)$. Similarly, a binomial distribution $\text{bin}(n, p)$ is determined by the two parameters n and p, and a normal distribution $N(\mu, \sigma^2)$ is determined by the two parameters μ and σ^2 (or equivalently, μ and σ). Parameterized families of distributions are often called parametric distributions or parametric models.

We are often faced with the situation of having random data which we know (or believe) is drawn from a parametric model, whose parameters we do not know. For example, in an election between two candidates, polling data constitutes draws from a Bernoulli(p) distribution with unknown parameter p. In this case we would like to use the data to estimate the value of the parameter p, as the latter predicts the result of the election. Similarly, assuming gestational length follows a normal distribution, we would like to use the data of the gestational lengths from a random sample of pregnancies to draw inferences about the values of the parameters μ and σ^2.

Our focus so far has been on computing the probability of data arising from a parametric model with known parameters. Statistical inference flips this on its head: we will estimate the probability of parameters given a parametric model and observed data drawn from it. In the coming weeks we will see how parameter values are naturally viewed as hypotheses, so we are in fact estimating the probability of various hypotheses given the data.

3 Maximum Likelihood Estimates

There are many methods for estimating unknown parameters from data. We will first consider the maximum likelihood estimate (MLE), which answers the question:

> For which parameter value does the observed data have the biggest probability?

The MLE is an example of a point estimate because it gives a single value for the unknown parameter (later our estimates will involve intervals and probabilities). Two advantages of

the MLE are that it is often easy to compute and that it agrees with our intuition in simple examples. We will explain the MLE through a series of examples.

Example 1. A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

Before actually solving the problem, let's establish some notation and terms.

We can think of counting the number of heads in 100 tosses as an experiment. For a given value of p, the probability of getting 55 heads in this experiment is the binomial probability

$$P(55 \text{ heads}) = \binom{100}{55} p^{55}(1-p)^{45}.$$

The probability of getting 55 heads depends on the value of p, so let's include p in by using the notation of conditional probability:

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

You should read $P(55 \text{ heads} \,|\, p)$ as:

'the probability of 55 heads given p,'

or more precisely as

'the probability of 55 heads given that the probability of heads on a single toss is p.'

Here are some standard terms we will use as we do statistics.

- Experiment: Flip the coin 100 times and count the number of heads.
- Data: The data is the result of the experiment. In this case it is '55 heads'.
- Parameter(s) of interest: We are interested in the value of the unknown parameter p.
- Likelihood, or likelihood function: this is $P(\text{data} \,|\, p)$. Note it is a function of both the data and the parameter p. In this case the likelihood is

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

Notes: **1.** The likelihood $P(\text{data} \,|\, p)$ changes as the parameter of interest p changes.

2. Look carefully at the definition. One typical source of confusion is to mistake the likelihood $P(\text{data} \,|\, p)$ for $P(p \,|\, \text{data})$. We know from our earlier work with Bayes' theorem that $P(\text{data} \,|\, p)$ and $P(p \,|\, \text{data})$ are usually very different.

Definition: Given data the maximum likelihood estimate (MLE) for the parameter p is the value of p that maximizes the likelihood $P(\text{data} \,|\, p)$. That is, the MLE is the value of p for which the data is most likely.

answer: For the problem at hand, we saw above that the likelihood

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

We'll use the notation $\hat{p}$ for the MLE. We use calculus to find it by taking the derivative of the likelihood function and setting it to 0.

$$\frac{d}{dp}P(\text{data } |p) = \binom{100}{55}(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0.$$

Solving this for p we get

$$\begin{aligned} &55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44} \\ &55(1-p) = 45p \\ &55 = 100p \\ &\text{the MLE is } \hat{p} = .55 \end{aligned}$$

Note: **1.** The MLE for p turned out to be exactly the fraction of heads we saw in our data.

2. The MLE is computed from the data. That is, it is a statistic.

3. Officially you should check that the critical point is indeed a maximum. You can do this with the second derivative test.

3.1 Log likelihood

If is often easier to work with the natural log of the likelihood function. For short this is simply called the log likelihood. Since $\ln(x)$ is an increasing function, the maxima of the likelihood and log likelihood coincide.

Example 2. Redo the previous example using log likelihood.

answer: We had the likelihood $P(55 \text{ heads } | p) = \binom{100}{55}p^{55}(1-p)^{45}$. Therefore the log likelihood is

$$\ln(P(55 \text{ heads } | p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p).$$

Maximizing likelihood is the same as maximizing log likelihood. We check that calculus gives us the same answer as before:

$$\begin{aligned} \frac{d}{dp}(\text{log likelihood}) &= \frac{d}{dp}\left[\ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p)\right] \\ &= \frac{55}{p} - \frac{45}{1-p} = 0 \\ &\Rightarrow 55(1-p) = 45p \\ &\Rightarrow \hat{p} = .55 \end{aligned}$$

3.2 Maximum likelihood for continuous distributions

For continuous distributions, we use the probability density function to define the likelihood. We show this in a few examples. In the next section we explain how this is analogous to what we did in the discrete case.

Example 3. Light bulbs
Suppose that the lifetime of *Badger* brand light bulbs is modeled by an exponential distribution with (unknown) parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ?

answer: We need to be careful with our notation. With five different values it is best to use subscripts. Let X_j be the lifetime of the i^{th} bulb and let x_i be the value X_i takes. Then each X_i has pdf $f_{X_i}(x_i) = \lambda e^{-\lambda x_i}$. We assume the lifetimes of the bulbs are independent, so the joint pdf is the product of the individual densities:

$$f(x_1, x_2, x_3, x_4, x_5 \,|\, \lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5}) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}.$$

Note that we write this as a conditional density, since it depends on λ. Viewing the data as fixed and λ as variable, this density is the likelihood function. Our data had values

$$x_1 = 2,\ x_2 = 3,\ x_3 = 1,\ x_4 = 3,\ x_5 = 4.$$

So the likelihood and log likelihood functions with this data are

$$f(2, 3, 1, 3, 4 \,|\, \lambda) = \lambda^5 e^{-13\lambda}, \quad \ln(f(2, 3, 1, 3, 4 \,|\, \lambda) = 5 \ln(\lambda) - 13\lambda$$

Finally we use calculus to find the MLE:

$$\frac{d}{d\lambda}(\text{log likelihood}) = \frac{5}{\lambda} - 13 = 0 \ \Rightarrow\ \boxed{\hat{\lambda} = \frac{5}{13}}.$$

Note: **1.** In this example we used an uppercase letter for a random variable and the corresponding lowercase letter for the value it takes. This will be our usual practice.

2. The MLE for λ turned out to be the reciprocal of the sample mean $\bar{x}$, so $X \sim \exp(\hat{\lambda})$ satisfies $E(X) = \bar{x}$.

The following example illustrates how we can use the method of maximum likelihood to estimate multiple parameters at once.

Example 4. Normal distributions
Suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a $N(\mu, \sigma^2)$ distribution, where μ and σ are unknown. Find the maximum likelihood estimate for the pair (μ, σ^2).

answer: Let's be precise and phrase this in terms of random variables and densities. Let uppercase $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables, and let lowercase x_i be the value X_i takes. The density for each X_i is

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Since the X_i are independent their joint pdf is the product of the individual pdf's:

$$f(x_1, \ldots, x_n \,|\, \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

For the fixed data $x_1, \ldots, x_n$, the likelihood and log likelihood are

$$f(x_1, \ldots, x_n|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}, \quad \ln(f(x_1, \ldots, x_n|\mu, \sigma)) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Since $\ln(f(x_1,\dots,x_n|\mu,\sigma))$ is a function of the two variables μ, σ we use partial derivatives to find the MLE. The easy value to find is $\hat{\mu}$:

$$\frac{\partial f(x_1,\dots,x_n|\mu,\sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0 \;\Rightarrow\; \sum_{i=1}^{n} x_i = n\mu \;\Rightarrow\; \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

To find $\hat{\sigma}$ we differentiate and solve for σ:

$$\frac{\partial f(x_1,\dots,x_n|\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0 \;\Rightarrow\; \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}.$$

We already know $\hat{\mu} = \overline{x}$, so we use that as the value for μ in the formula for $\hat{\sigma}$. We get the maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \overline{x} && = \text{the mean of the data} \\ \hat{\sigma}^2 &= \sum_{i=1}^{n} \frac{1}{n}(x_i-\hat{\mu})^2 = \sum_{i=1}^{n} \frac{1}{n}(x_i-\overline{x})^2 && = \text{the variance of the data.} \end{aligned}$$

Example 5. Uniform distributions
Suppose our data $x_1,\dots x_n$ are independently drawn from a uniform distribution $U(a,b)$. Find the MLE estimate for a and b.

answer: This example is different from the previous ones in that we won't use calculus to find the MLE. The density for $U(a,b)$ is $\frac{1}{b-a}$ on $[a,b]$. Therefore our likelihood function is

$$f(x_1,\dots,x_n \,|\, a,b) = \begin{cases} \left(\frac{1}{b-a}\right)^n & \text{if all } x_i \text{ are in the interval } [a,b] \\ 0 & \text{otherwise.} \end{cases}$$

This is maximized by making $b-a$ as small as possible. The only restriction is that the interval $[a,b]$ must include all the data. Thus the MLE for the pair (a,b) is

$$\hat{a} = \min(x_1,\dots,x_n) \qquad \hat{b} = \max(x_1,\dots,x_n).$$

Example 6. Capture/recapture method

The capture/recapture method is a way to estimate the size of a population in the wild. The method assumes that each animal in the population is equally likely to be captured by a trap.

Suppose 10 animals are captured, tagged and released. A few months later, 20 animals are captured, examined, and released. 4 of these 20 are found to be tagged. Estimate the size of the wild population using the MLE for the probability that a wild animal is tagged.

answer: Our unknown parameter n is the number of animals in the wild. Our data is that 4 out of 20 recaptured animals were tagged (and that there are 10 tagged animals). The likelihood function is

$$P(\text{data} \,|\, n \text{ animals}) = \frac{\binom{n-10}{16}\binom{10}{4}}{\binom{n}{20}}$$

(The numerator is the number of ways to choose 16 animals from among the $n-10$ untagged ones times the number of was to choose 4 out of the 10 tagged animals. The denominator

is the number of ways to choose 20 animals from the entire population of n.) We can use R to compute that the likelihood function is maximized when $n = 50$. This should make some sense. It says our best estimate is that the fraction of all animals that are tagged is 10/50 which equals the fraction of recaptured animals which are tagged.

Example 7. Hardy-Weinberg. Suppose that a particular gene occurs as one of two alleles (A and a), where allele A has frequency θ in the population. That is, a random copy of the gene is A with probability θ and a with probability $1 - \theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:

genotype	AA	Aa	aa
probability	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

Suppose we test a random sample of people and find that k_1 are AA, k_2 are Aa, and k_3 are aa. Find the MLE of θ.

answer: The likelihood function is given by

$$P(k_1, k_2, k_3 \,|\, \theta) = \binom{k_1 + k_2 + k_3}{k_1} \binom{k_2 + k_3}{k_2} \binom{k_3}{k_3} \theta^{2k_1} (2\theta(1-\theta))^{k_2} (1-\theta)^{2k_3}.$$

So the log likelihood is given by

$$\text{constant} + 2k_1 \ln(\theta) + k_2 \ln(\theta) + k_2 \ln(1-\theta) + 2k_3 \ln(1-\theta)$$

We set the derivative equal to zero:

$$\frac{2k_1 + k_2}{\theta} - \frac{k_2 + 2k_3}{1-\theta} = 0$$

Solving for θ, we find the MLE is

$$\hat{\theta} = \frac{2k_1 + k_2}{2k_1 + 2k_2 + 2k_3},$$

which is simply the fraction of A alleles among all the genes in the sampled population.

4 Why we use the density to find the MLE for continuous distributions

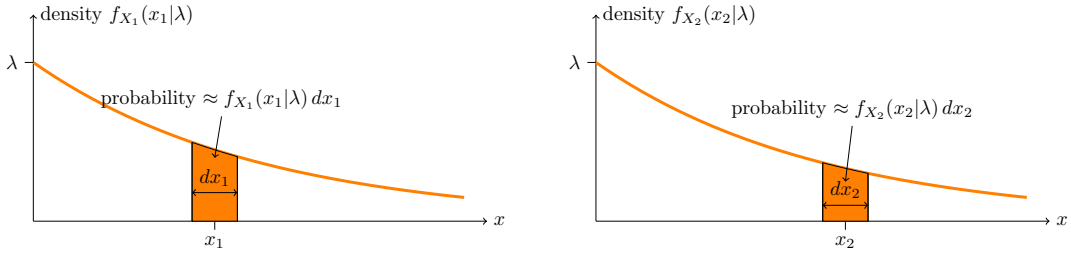
The idea for the maximum likelihood estimate is to find the value of the parameter(s) for which the data has the highest probability. In this section we 'll see that we're doing this is really what we are doing with the densities. We will do this by considering a smaller version of the light bulb example.

Example 8. Suppose we have two light bulbs whose lifetimes follow an exponential(λ) distribution. Suppose also that we independently measure their lifetimes and get data $x_1 = 2$ years and $x_2 = 3$ years. Find the value of λ that maximizes the probability of this data.

answer: The main paradox to deal with is that for a continuous distribution the probability of a single value, say $x_1 = 2$, is zero. We resolve this paradox by remembering that a single

measurement really means a range of values, e.g. in this example we might check the light bulb once a day. So the data $x_1 = 2$ years really means x_1 is somewhere in a range of 1 day around 2 years.

If the range is small we call it dx_1. The probability that X_1 is in the range is approximated by $f_{X_1}(x_1|\lambda)\,dx_1$. This is illustrated in the figure below. The data value x_2 is treated in exactly the same way.



The usual relationship between density and probability for small ranges.

Since the data is collected independently the joint probability is the product of the individual probabilities. Stated carefully

$$P(X_1 \text{ in range, } X_2 \text{ in range}|\lambda) \approx f_{X_1}(x_1|\lambda)\,dx_1 \cdot f_{X_2}(x_2|\lambda)\,dx_2$$

Finally, using the values $x_1 = 2$ and $x_2 = 3$ and the formula for an exponential pdf we have

$$P(X_1 \text{ in range, } X_2 \text{ in range}|\lambda) \approx \lambda e^{-2\lambda}\,dx_1 \cdot \lambda e^{-3\lambda}\,dx_2 = \lambda^2 e^{-5\lambda}\,dx_1\,dx_2.$$

Now that we have a genuine probability we can look for the value of λ that maximizes it. Looking at the formula above we see that the factor $dx_1\,dx_2$ will play no role in finding the maximum. So for the MLE we drop it and simply call the density the likelihood:

$$\text{likelihood } = f(x_1, x_2|\lambda) = \lambda^2 e^{-5\lambda}.$$

The value of λ that maximizes this is found just like in the example above. It is $\hat{\lambda} = 2/5$.

5 Appendix: Properties of the MLE

For the interested reader, we note several nice features of the MLE. These are quite technical and will not be on any exams.

The MLE behaves well under transformations. That is, if $\hat{p}$ is the MLE for p and g is a one-to-one function, then $g(\hat{p})$ is the MLE for $g(p)$. For example, if $\hat{\sigma}$ is the MLE for the standard deviation σ then $(\hat{\sigma})^2$ is the MLE for the variance σ^2.

Furthermore, the MLE is asymptotically unbiased and has asymptotically minimal variance. To explain these notions, note that the MLE is itself a random variable since the data is random and the MLE is computed from the data. Let $x_1, x_2, \ldots$ be an infinite sequence of samples from a distribution with parameter p. Let $\hat{p}_n$ be the MLE for p based on the data $x_1, \ldots, x_n$.

Asymptotically unbiased means that as the amount of data grows, the mean of the MLE converges to p. In symbols: $E(\hat{p}_n) \to p$ as $n \to \infty$. Of course, we would like the MLE to be

close to p with high probability, not just on average, so the smaller the variance of the MLE the better. Asymptotically minimal variance means that as the amount of data grows, the MLE has the minimal variance among all unbiased estimators of p. In symbols: for any unbiased estimator $\tilde{p}_n$ and $\epsilon > 0$ we have that $\mathrm{Var}(\tilde{p}_n) + \epsilon > \mathrm{Var}(\hat{p}_n)$ as $n \to \infty$.

Bayesian Updating with Discrete Priors
Class 11, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to apply Bayes' theorem to compute probabilities.
2. Be able to define the and to identify the roles of prior probability, likelihood (Bayes term), posterior probability, data and hypothesis in the application of Bayes' Theorem.
3. Be able to use a Bayesian update table to compute posterior probabilities.

2 Review of Bayes' theorem

Recall that Bayes' theorem allows us to 'invert' conditional probabilities. If $\mathcal{H}$ and $\mathcal{D}$ are events, then:

$$P(\mathcal{H} \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

Our view is that Bayes' theorem forms the foundation for inferential statistics. We will begin to justify this view today.

2.1 The base rate fallacy

When we first learned Bayes' theorem we worked an example about screening tests showing that $P(\mathcal{D}|\mathcal{H})$ can be very different from $P(\mathcal{H}|\mathcal{D})$. In the appendix we work a similar example. If you are not comfortable with Bayes' theorem you should read the example in the appendix now.

3 Terminology and Bayes' theorem in tabular form

We now use a coin tossing problem to introduce terminology and a tabular format for Bayes' theorem. This will provide a simple, uncluttered example that shows our main points.

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

answer: Let A, B, and C be the event that the chosen coin was type A, type B, and type C. Let $\mathcal{D}$ be the event that the toss is heads. The problem asks us to find

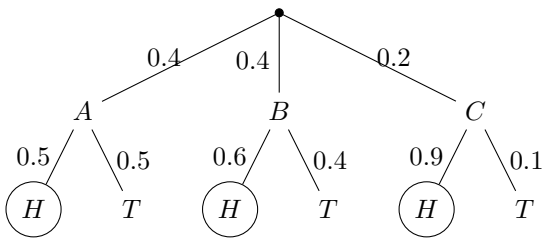
$$P(A|\mathcal{D}), \quad P(B|\mathcal{D}), \quad P(C|\mathcal{D}).$$

Before applying Bayes' theorem, let's introduce some terminology.

- Experiment: pick a coin from the drawer at random, flip it, and record the result.
- Data: the result of our experiment. In this case the event $\mathcal{D}$ = 'heads'. We think of $\mathcal{D}$ as data that provides evidence for or against each hypothesis.
- Hypotheses: we are testing three hypotheses: the coin is type A, B or C.
- Prior probability: the probability of each hypothesis prior to tossing the coin (collecting data). Since the drawer has 2 coins of type A, 2 of type B and 1 or type C we have
$$P(A) = 0.4, \qquad P(B) = 0.4, \qquad P(C) = 0.2.$$
- Likelihood: (This is the same likelihood we used for the MLE.) The likelihood function is $P(\mathcal{D}|\mathcal{H})$, i.e., the probability of the data assuming that the hypothesis is true. Most often we will consider the data as fixed and let the hypothesis vary. For example, $P(\mathcal{D}|A)$ = probability of heads if the coin is type A. In our case the likelihoods are
$$P(\mathcal{D}|A) = 0.5, \qquad P(\mathcal{D}|B) = 0.6, \qquad P(\mathcal{D}|C) = 0.9.$$
The name likelihood is so well established in the literature that we have to teach it to you. However in colloquial language likelihood and probability are synonyms. This leads to the likelihood function often being confused with the probabity of a hypothesis. Because of this we'd prefer to use the name Bayes' term. However since we are stuck with 'likelihood' we will try to use it very carefully and in a way that minimizes any confusion.
- Posterior probability: the probability (posterior to) of each hypothesis given the data from tossing the coin.
$$P(A|\mathcal{D}), \qquad P(B|\mathcal{D}), \qquad P(C|\mathcal{D}).$$
These posterior probabilities are what the problem asks us to find.

We now use Bayes' theorem to compute each of the posterior probabilities. We are going to write this out in complete detail so we can pick out each of the parts (Remember that the data $\mathcal{D}$ is that the toss was heads.)

First we organize the probabilities into a tree:



Probability tree for choosing and tossing a coin.

Bayes' theorem says, e.g. $P(A|\mathcal{D}) = \dfrac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})}$. The denominator $P(\mathcal{D})$ is computed using the law of total probability:

$$P(\mathcal{D}) = P(\mathcal{D}|A)P(A) + P(\mathcal{D}|B)P(B) + P(\mathcal{D}|C)P(C) = 0.5 \cdot 0.4 + 0.6 \cdot 0.4 + 0.9 \cdot 0.2 = 0.62.$$

Now each of the three posterior probabilities can be computed:

$$P(A|\mathcal{D}) = \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})} = \frac{0.5 \cdot 0.4}{0.62} = \frac{0.2}{0.62}$$
$$P(B|\mathcal{D}) = \frac{P(\mathcal{D}|B)P(B)}{P(\mathcal{D})} = \frac{0.6 \cdot 0.4}{0.62} = \frac{0.24}{0.62}$$
$$P(C|\mathcal{D}) = \frac{P(\mathcal{D}|C)P(C)}{P(\mathcal{D})} = \frac{0.9 \cdot 0.2}{0.62} = \frac{0.18}{0.62}$$

Notice that the total probability $P(\mathcal{D})$ is the same in each of the denominators and that it is the sum of the three numerators. We can organize all of this very neatly in a Bayesian update table:

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
A	0.4	0.5	0.2	0.3226
B	0.4	0.6	0.24	0.3871
C	0.2	0.9	0.18	0.2903
total	1		0.62	1

The Bayes numerator is the product of the prior and the likelihood. We see in each of the Bayes' formula computations above that the posterior probability is obtained by dividing the Bayes numerator by $P(\mathcal{D}) = 0.625$. We also see that the law of law of total probability says that $P(\mathcal{D})$ is the sum of the entries in the Bayes numerator column.

Bayesian updating: The process of going from the prior probability $P(\mathcal{H})$ to the posterior $P(\mathcal{H}|\mathcal{D})$ is called Bayesian updating. Bayesian updating uses the data to alter our understanding of the probability of each of the possible hypotheses.

3.1 Important things to notice

1. There are two types of probabilities: Type one is the standard probability of data, e.g. the probability of heads is $p = 0.9$. Type two is the probability of the hypotheses, e.g. the probability the chosen coin is type A, B or C. This second type has prior (before the data) and posterior (after the data) values.

2. The posterior (after the data) probabilities for each hypothesis are in the last column. We see that coin B is now the most probable, though its probability has decreased from a prior probability of 0.4 to a posterior probability of 0.39. Meanwhile, the probability of type C has increased from 0.2 to 0.29.

3. The Bayes numerator column determines the posterior probability column. To compute the latter, we simply rescaled the Bayes numerator so that it sums to 1.

4. If all we care about is finding the most likely hypothesis, the Bayes numerator works as well as the normalized posterior.

5. The likelihood column does not sum to 1. The likelihood function is *not* a probability function.

6. The posterior probability represents the outcome of a 'tug-of-war' between the likelihood and the prior. When calculating the posterior, a large prior may be deflated by a small likelihood, and a small prior may be inflated by a large likelihood.

7. The maximum likelihood estimate (MLE) for Example 1 is hypothesis C, with a likelihood $P(\mathcal{D}|C) = 0.9$. The MLE is useful, but you can see in this example that it is not the entire story, since type B has the greatest posterior probability.

Terminology in hand, we can express Bayes' theorem in various ways:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

With the data fixed, the denominator $P(\mathcal{D})$ just serves to normalize the total posterior probability to 1. So we can also express Bayes' theorem as a statement about the proportionality of two functions of $\mathcal{H}$ (i.e, of the last two columns of the table).

$$P(\text{hypothesis}|\text{data}) \propto P(\text{data}|\text{hypothesis})P(\text{hypothesis})$$

This leads to the most elegant form of Bayes' theorem in the context of Bayesian updating:

$$\boxed{\text{posterior} \propto \text{likelihood} \times \text{prior}}$$

3.2 Prior and posterior probability mass functions

Earlier in the course we saw that it is convenient to use random variables and probability mass functions. To do this we had to assign values to events (head is 1 and tails is 0). We will do the same thing in the context of Bayesian updating.

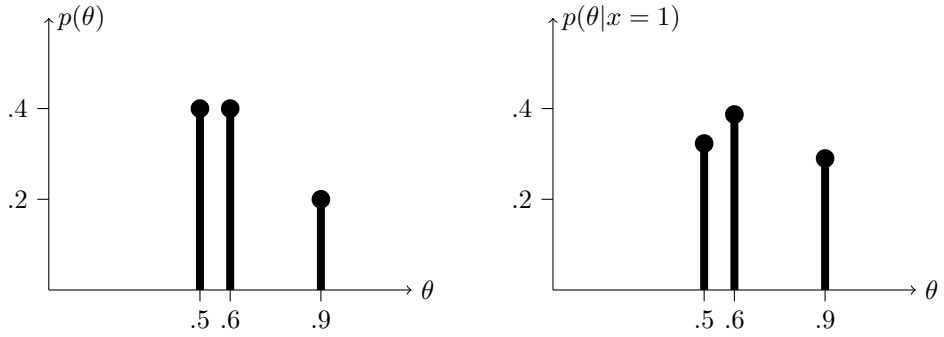
Our standard notations will be:

- θ is the value of the hypothesis.
- $p(\theta)$ is the prior probability mass function of the hypothesis.
- $p(\theta|\mathcal{D})$ is the posterior probability mass function of the hypothesis given the data.
- $p(\mathcal{D}|\theta)$ is the likelihood function. (This is not a pmf!)

In Example 1 we can represent the three hypotheses A, B, and C by $\theta = 0.5, 0.6, 0.9$. For the data we'll let $x = 1$ mean heads and $x = 0$ mean tails. Then the prior and posterior probabilities in the table define the prior and posterior probability mass functions.

Hypothesis	θ	prior pmf $p(\theta)$	poster pmf $p(\theta\|x=1)$
A	0.5	$P(A) = p(0.5) = 0.4$	$P(A\|\mathcal{D}) = p(0.5\|x=1) = 0.3226$
B	0.6	$P(B) = p(0.6) = 0.4$	$P(B\|\mathcal{D}) = p(0.6\|x=1) = 0.3871$
C	0.9	$P(C) = p(0.9) = 0.2$	$P(C\|\mathcal{D}) = p(0.9\|x=1) = 0.2903$

Here are plots of the prior and posterior pmf's from the example.



Prior pmf $p(\theta)$ and posterior pmf $p(\theta|x=1)$ for Example 1

If the data was different then the likelihood column in the Bayesian update table would be different. We can plan for different data by building the entire likelihood table ahead of time. In the coin example there are two possibilities for the data: the toss is heads or the toss is tails. So the full likelihood table has two likelihood columns:

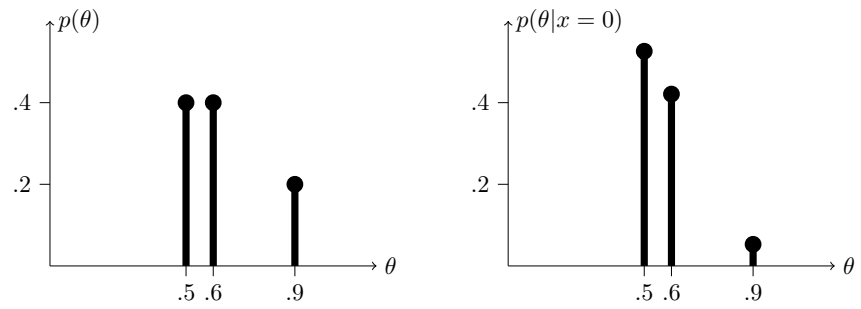
hypothesis	likelihood $p(x\|\theta)$	
θ	$p(x=0\|\theta)$	$p(x=1\|\theta)$
0.5	0.5	0.5
0.6	0.4	0.6
0.9	0.1	0.9

Example 2. Using the notation $p(\theta)$, etc., redo Example 1 assuming the flip was tails.

answer: Since the data has changed, the likelihood column in the Bayesian update table is now for $x=0$. That is, we must take the $p(x=0|\theta)$ column from the likelihood table.

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$p(\theta)$	$p(x=0\,\|\,\theta)$	$p(x=0\,\|\,\theta)p(\theta)$	$p(\theta\,\|\,x=0)$
0.5	0.4	0.5	0.2	0.5263
0.6	0.4	0.4	0.16	0.4211
0.9	0.2	0.1	0.02	0.0526
total	1		0.38	1

Now the probability of type A has increased from 0.4 to 0.5263, while the probability of type C has decreased from 0.2 to only 0.0526. Here are the corresponding plots:



Prior pmf $p(\theta)$ and posterior pmf $p(\theta|x=0)$

3.3 Food for thought. Thinking

Suppose that in Example 1 you didn't know how many coins of each type were in the drawer. You picked one at random and got heads. How would you go about deciding which hypothesis (coin type) if any was most supported by the data?

4 Updating again and again

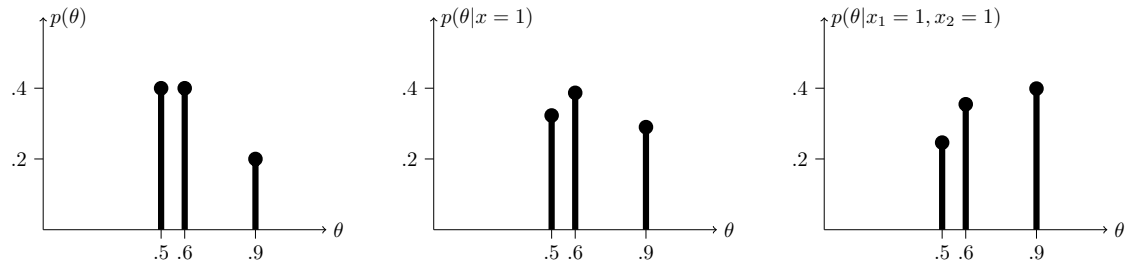
In life we are continually updating our beliefs with each new experience of the world. In Bayesian inference, after updating the prior to the posterior, we can take more data and update again! For the second update, the posterior from the first data becomes the prior for the second data.

Example 3. Suppose you have picked a coin as in Example 1. You flip it once and get heads. Then you flip the same coin and get heads again. What is the probability that the coin was type A? Type B? Type C?

answer: As we update several times the table gets big, so we use a smaller font to fit it in:

hypothesis	prior	likelihood 1	Bayes numerator 1	likelihood 2	Bayes numerator 2	posterior 2
θ	$p(\theta)$	$p(x_1=1\|\theta)$	$p(x_1=1\|\theta)p(\theta)$	$p(x_2=1\|\theta)$	$p(x_2=1\|\theta)p(x_1=1\|\theta)p(\theta)$	$p(\theta\|x_1=1,x_2=1)$
0.5	0.4	0.5	0.2	0.5	0.1	0.2463
0.6	0.4	0.6	0.24	0.6	0.144	0.3547
0.9	0.2	0.9	0.18	0.9	0.162	0.3990
total	1				0.406	1

Note that the second Bayes numerator is computed by multiplying the first Bayes numerator and the second likelihood; since we are only interested in the final posterior, there is no need to normalize until the last step. As shown in the last column and plot, after two heads the type C hypothesis has finally taken the lead!



prior $p(\theta)$, first posterior $p(\theta|x_1 = 1)$, second posterior $p(\theta|x_1 = 1, x_2 = 1)$

5 Appendix: the base rate fallacy

Example 4. A screening test for a disease is both sensitive and specific. By that we mean it is usually positive when testing a person with the disease and usually negative when testing someone without the disease. Let's assume the true positive rate is 99% and the false positive rate is 2%. Suppose the prevalence of the disease in the general population is 0.5%. If a random person tests positive, what is the probability that they have the disease?

answer: As a review we first do the computation using trees. Next we will redo the computation using tables.

Let's use notation established above for hypotheses and data: let $\mathcal{H}_+$ be the hypothesis (event) that the person has the disease and let $\mathcal{H}_-$ be the hypothesis they do not. Likewise, let $\mathcal{T}_+$ and $\mathcal{T}_-$ represent the data of a positive and negative screening test respectively. We are asked to compute $P(\mathcal{H}_+|\mathcal{T}_+)$.

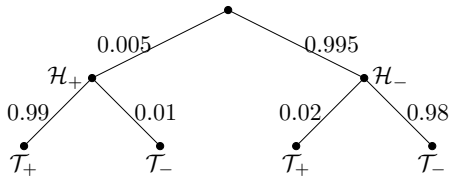
We are given

$$P(\mathcal{T}_+|\mathcal{H}_+) = 0.99, \quad P(\mathcal{T}_+|\mathcal{H}_-) = 0.02, \quad P(\mathcal{H}_+) = 0.005.$$

From these we can compute the false negative and true negative rates:

$$P(\mathcal{T}_-|\mathcal{H}_+) = 0.01, \quad P(\mathcal{T}_-|\mathcal{H}_-) = 0.98$$

All of these probabilities can be displayed quite nicely in a tree.



Bayes' theorem yields

$$P(\mathcal{H}_+|\mathcal{T}_+) = \frac{P(\mathcal{T}_+|\mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{T}_+)} = \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.02 \cdot 0.995} = 0.19920 \approx 20\%$$

Now we redo this calculation using a Bayesian update table:

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{T}_+\|\mathcal{H})$	$P(\mathcal{T}_+\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{T}_+)$
$\mathcal{H}_+$	0.005	0.99	0.00495	0.19920
$\mathcal{H}_-$	0.995	0.02	0.01990	0.80080
total	1	NO SUM	0.02485	1

The table shows that the posterior probability $P(\mathcal{H}_+|\mathcal{T}_+)$ that a person with a positive test has the disease is about 20%. This is far less than the sensitivity of the test (99%) but much higher than the prevalence of the disease in the general population (0.5%).

Bayesian Updating: Probabilistic Prediction
Class 12, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to use the law of total probability to compute prior and posterior predictive probabilities.

2 Introduction

In the previous class we looked at updating the probability of hypotheses based on data. We can also use the data to update the probability of each possible outcome of a future experiment. In this class we will look at how this is done.

2.1 Probabilistic prediciton; words of estimative probability (WEP)

There are many ways to word predictions:

- Prediction: "It will rain tomorrow."
- Prediction using words of estimative probability (WEP): "It is likely to rain tomorrow."
- Probabilistic prediction: "Tomorrow it will rain with probability 60% (and not rain with probability 40%)."

Each type of wording is appropriate at different times.

In this class we are going to focus on probabilistic prediction and precise quantitative statements. You can see http://en.wikipedia.org/wiki/Words_of_Estimative_Probability for an interesting discussion about the appropriate use of words of estimative probability. The article also contains a list of *weasel words* such as 'might', 'cannot rule out', 'it's conceivable' that should be avoided as almost certain to cause confusion.

There are many places where we want to make a probabilistic prediction. Examples are

- Medical treatment outcomes
- Weather forecasting
- Climate change
- Sports betting
- Elections
- ...

These are all situations where there is uncertainty about the outcome and we would like as precise a description of what could happen as possible.

3 Predictive Probabilities

Probabilistic prediction simply means assigning a probability to each possible outcomes of an experiment.

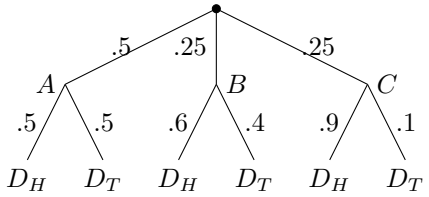
Recall the coin example from the previous class notes: there are three types of coins which are indistinguishable apart from their probability of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins have probability 0.6 of heads
- Type C coins have probability 0.9 of heads

You have a drawer containing 4 coins: 2 of type A, 1 of type B, and 1 of type C. You reach into the drawer and pick a coin at random. We let A stand for the event 'the chosen coin is of type A'. Likewise for B and C.

3.1 Prior predictive probabilities

Before taking data we can compute the probability that our chosen coin will land heads (or tails) if flipped. Let D_H be the event it lands heads and let D_T the event it lands tails. We can use the law of total probability to determine the probabilities of these events. Either by drawing a tree or directly proceeding to the algebra, we get:



$$\begin{aligned} P(D_H) &= P(D_H|A)P(A) + P(D_H|B)P(B) + P(D_H|C)P(C) \\ &= 0.5 \cdot 0.5 + 0.6 \cdot 0.25 + 0.9 \cdot 0.25 = 0.625 \\ P(D_T) &= P(D_T|A)P(A) + P(D_T|B)P(B) + P(D_T|C)P(C) \\ &= 0.5 \cdot 0.5 + 0.4 \cdot 0.25 + 0.1 \cdot 0.25 = 0.375 \end{aligned}$$

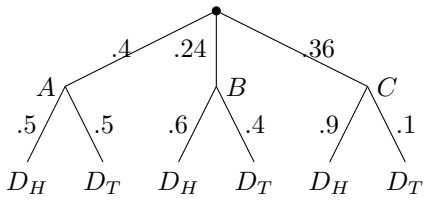
Definition: These probabilities give a (probabilistic) prediction of what will happen if the coin is tossed. Because they are computed before we collect any data they are called prior predictive probabilities.

3.2 Posterior predictive probabilities

Suppose we flip the coin once and it lands heads. We now have data D, which we can use to update the prior probabilities of our hypotheses to posterior probabilities. Last class we learned to use a Bayes table to facilitate this computation:

hypothesis	prior	likelihood	Bayes numerator	posterior
H	$P(H)$	$P(D\|H)$	$P(D\|H)P(H)$	$P(H\|D)$
A	0.5	0.5	0.25	0.4
B	0.25	0.6	0.15	0.24
C	0.25	0.9	0.225	0.36
total	1		0.625	1

Having flipped the coin once and gotten heads, we can compute the probability that our chosen coin will land heads (or tails) if flipped a second time. We proceed just as before, but using the posterior probabilities $P(A|D)$, $P(B|D)$, $P(C|D)$ in place of the prior probabilities $P(A)$, $P(B)$, $P(C)$.



$$\begin{aligned} P(D_H|D) &= P(D_H|A)P(A|D) + P(D_H|B)P(B|D) + P(D_H|C)P(C|D) \\ &= 0.5 \cdot 0.4 + 0.6 \cdot 0.24 + 0.9 \cdot 0.36 = 0.668 \\ P(D_T|D) &= P(D_T|A)P(A|D) + P(D_T|B)P(B|D) + P(D_T|C)P(C|D) \\ &= 0.5 \cdot 0.4 + 0.4 \cdot 0.24 + 0.1 \cdot 0.36 = 0.332 \end{aligned}$$

Definition: These probabilities give a (probabilistic) prediction of what will happen if the coin is tossed again. Because they are computed after collecting data and updating the prior to the posterior, they are called posterior predictive probabilities.

Note that heads on the first toss increases the probability of heads on the second toss.

3.3 Review

Here's a succinct description of the preceding sections that may be helpful:

Each hypothesis gives a different probability of heads, so the total probability of heads is a weighted average. For the prior predictive probability of heads, the weights are given by the prior probabilities of the hypotheses. For the posterior predictive probability of heads, the weights are given by the posterior probabilities of the hypotheses.

Remember: Prior and posterior probabilities are for hypotheses. Prior predictive and posterior predictive probabilities are for data. To keep this straight, remember that the latter predict future data.

Bayesian Updating: Odds
Class 12, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to convert between odds and probability.
2. Be able to update prior odds to posterior odds using Bayes factors.
3. Understand how Bayes factors measure the extent to which data provides evidence for or against a hypothesis.

2 Odds

When comparing two events, it common to phrase probability statements in terms of odds.

Definition The odds of event E versus event E' are the ratio of their probabilities $P(E)/P(E')$. If unspecified, the second event is assumed to be the complement E^c. So the odds of E are:

$$O(E) = \frac{P(E)}{P(E^c)}.$$

For example, $O(\text{rain}) = 2$ means that the probability of rain is twice the probability of no rain (2/3 versus 1/3). We might say 'the odds of rain are 2 to 1.'

Example. For a fair coin, $O(\text{heads}) = \dfrac{1/2}{1/2} = 1$. We might say the odds of heads are 1 to 1 or fifty-fifty.

Example. For a standard die, the odds of rolling a 4 are $\dfrac{1/6}{5/6} = \dfrac{1}{5}$. We might say the odds are '1 to 5 for' or '5 to 1 against' rolling a 4.

Example. The probability of a pair in a five card poker hand is 0.42257. So the odds of a pair are 0.42257/(1-0.42257) = 0.73181.

We can go back and forth between probability and odds as follows.

Conversion formulas: if $P(E) = p$ then $O(E) = \dfrac{p}{1-p}$. If $O(E) = q$ then $P(E) = \dfrac{q}{1+q}$.

Notes:
1. The second formula simply solves $q = p/(1-p)$ for p.
2. Probabilities are between 0 and 1, while odds are between 0 to ∞.
3. The property $P(E^c) = 1 - P(E)$ becomes $O(E^c) = 1/O(E)$.

Example. Let F be the event that a five card poker hand is a full house. Then $P(F) = 0.00145214$ so $O(F) = 0.0014521/(1 - 0.0014521) = 0.0014542$.

The odds not having a full house are $O(F^c) = (1 - 0.0014521)/0.0014521 = 687 = 1/O(F)$.

4. If $P(E)$ or $O(E)$ is small then $O(E) \approx P(E)$. This follows from the conversion formulas.

Example. In the poker example where $F =$ 'full house' we saw that $P(F)$ and $O(F)$ differ only in the fourth significant digit.

3 Updating odds

3.1 Introduction

In Bayesian updating, we used the likelihood of data to update prior probabilities of hypotheses to posterior probabilities. In the language of odds, we will update prior odds to posterior odds. One of our key points will be that the data can provide evidence supporting or negating a hypothesis depending on whether its posterior odds are greater or less than its prior odds.

3.2 Example: Marfan syndrome

Marfan syndrome is a genetic disease of connective tissue that occurs in 1 of every 15000 people. The main ocular features of Marfan syndrome include bilateral ectopia lentis (lens dislocation), myopia and retinal detachment. About 70% of people with Marfan syndrome have a least one of these ocular features; only 7% of people without Marfan syndrome do. (We don't guarantee the accuracy of these numbers, but they will work perfectly well for our example.)

If a person has at least one of these ocular features, what are the odds that they have Marfan syndrome?

answer: This is a standard Bayesian updating problem. Our hypotheses are:
M = 'the person has Marfan syndrome'
M^c = 'the person does not have Marfan syndrome'
The data is:
$F =$ 'the person has at least one ocular feature'.

We are given the prior probability of M and the likelihoods of F given M or M^c:

$$P(M) = 1/15000, \qquad P(F|M) = 0.7, \qquad P(F|M^c) = 0.07.$$

As before, we can compute the posterior probabilities using a table:

hypothesis	prior	likelihood	Bayes numerator	posterior
H	$P(H)$	$P(F\|H)$	$P(F\|H)P(H)$	$P(H\|F)$
M	0.000067	0.7	0.0000467	0.00066
M^c	0.999933	0.07	0.069995	0.99933
total	1		0.07004	1

First we find the prior odds:

$$O(M) = \frac{P(M)}{P(M^c)} = \frac{1/15000}{14999/15000} = \frac{1}{14999} \approx 0.000067.$$

The posterior odds are given by the ratio of the posterior probabilities or the Bayes numerators, since the normalizing factor will be the same in both numerator and denominator.

$$O(M|F) = \frac{P(M|F)}{P(M^c|F)} = \frac{P(F|M)P(M)}{P(F|M^c)P(M^c)} = 0.000667.$$

The posterior odds are a factor of 10 larger than the prior odds. In that sense, having an ocular feature is strong evidence in favor of the hypothesis M. However, because the prior odds are so small, it is still highly unlikely the person has Marfan syndrome.

4 Bayes factors and strength of evidence

The factor of 10 in the previous example is called a Bayes factor. The exact definition is the following.

Definition: For a hypothesis H and data D, the Bayes factor is the ratio of the likelihoods:

$$\text{Bayes factor} = \frac{P(D|H)}{P(D|H^c)}.$$

Let's see exactly where the Bayes factor arises in updating odds. We have

$$\begin{aligned} O(H|D) &= \frac{P(H|D)}{P(H^c|D)} \\ &= \frac{P(D|H)\,P(H)}{P(D|H^c)P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot \frac{P(H)}{P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot O(H) \end{aligned}$$

$$\text{posterior odds} = \textbf{Bayes factor} \times \text{ prior odds}$$

From this formula, we see that the Bayes' factor (BF) tells us whether the data provides evidence for or against the hypothesis.

- If $BF > 1$ then the posterior odds are greater than the prior odds. So the data provides evidence for the hypothesis.
- If $BF < 1$ then the posterior odds are less than the prior odds. So the data provides evidence against the hypothesis.
- If $BF = 1$ then the prior and posterior odds are equal. So the data provides no evidence either way.

The following example is taken from the textbook *Information Theory, Inference, and Learning Algorithms* by David J. C. Mackay, who has this to say regarding trial evidence.

> In my view, a jury's task should generally be to multiply together carefully evaluated likelihood ratios from each independent piece of admissible evidence with an equally carefully reasoned prior probability. This view is shared by many statisticians but learned British appeal judges recently disagreed and actually overturned the verdict of a trial because the jurors *had* been taught to use Bayes' theorem to handle complicated DNA evidence.

Example 1. Two people have left traces of their own blood at the scene of a crime. A suspect , Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and type 'AB' (a rare type, with frequency 1%). Does this data (type 'O' and 'AB' blood were found at the scene) give evidence in favor of the proposition that Oliver was one of the two people present at the scene of the crime?"

answer: There are two hypotheses:
S = 'Oliver and another unknown person were at the scene of the crime
S^c = 'two unknown people were at the scene of the crime'

The data is:
D = 'type 'O' and 'AB' blood were found'

The Bayes factor for Oliver's presence is $BF_{\text{Oliver}} = \dfrac{P(D|S)}{P(D|S^c)}$. We compute the numerator and denominator of this separately.

The data says that both type O and type AB blood were found. If Oliver was at the scene then 'type O' blood would be there. So $P(D|S)$ is the probability that the other person had type AB blood. We are told this is .01, so $P(D|S) = 0.01$.

If Oliver was not at the scene then there were two random people one with type O and one with type AB blood. The probability of this is $2 \cdot 0.6 \cdot 0.01$. The factor of 2 is because there are two ways this can happen –the first person is type O and the second is type AB or vice versa.*

Thus the Bayes factor for Oliver's presence is

$$BF_{\text{Oliver}} = \frac{P(D|S)}{P(D|S^c)} = \frac{0.01}{2 \cdot 0.6 \cdot 0.01} = 0.83.$$

Since $BF_{\text{Oliver}} < 1$, the data provides (weak) evidence against Oliver being at the scene.

*We have assumed the blood types of the two people are independent. This is not precisely true, but for a large population it is close enough. The exact probability is $\dfrac{2 \cdot N_O \cdot N_{AB}}{N \cdot (N-1)}$ where N_O is the number of people with type O blood, N_{AB} the number with type AB blood and N the size of the population. We have $\dfrac{N_O}{N} = 0.6$. For large N we have $N \approx N - 1$, so $\dfrac{N_{AB}}{N-1} \approx 0.01$. This shows the probability is approximately $2 \cdot 0.6 \cdot 0.01$ as claimed.

Example 2. Another suspect Alberto is found to have type 'AB' blood. Do the same data give evidence in favor of the proposition that Alberto was one of the two people present at the crime?

answer: Reusing the above notation with Alberto in place of Oliver we have:

$$BF_{\text{Alberto}} = \frac{P(D|S)}{P(D|S^c)} = \frac{0.6}{2 \cdot 0.6 \cdot 0.01} = 50.$$

Since $BF_{\text{Alberto}} \gg 1$, the data provides strong evidence in favor of Alberto being at the scene.

Notes:
1. In both examples, we have only computed the Bayes factor, not the posterior odds. To compute the latter, we would need to know the prior odds that Oliver (or Alberto) was at the scene based on other evidence.

2. Note that if 50% of the population had type O blood instead of 60%, then the Oliver's Bayes factor would be 1 (neither for nor against). More generally, the break-even point for blood type evidence is when the proportion of the suspect's blood type in the general population equals the proportion of the suspect's blood type among those who left blood at the scene.

4.1 Updating again and again

Suppose we collect data in two stages, first D_1, then D_2. We have seen in our dice and coin examples that the final posterior can be computed all at once or in two stages where we first update the prior using the likelihoods for D_1 and then update the resulting posterior using the likelihoods for D_2. The latter approach works whenever likelihoods multiply:

$$P(D_1, D_2|H) = P(D_1|H)P(D_2|H).$$

Since likelihoods are conditioned on hypotheses, we say that D_1 and D_2 are conditionally independent if the above equation holds for every hypothesis H.

Example. There are five dice in a drawer, with 4, 6, 8, 12, and 20 sides (these are the hypotheses). I pick a die at random and roll it twice. The first roll gives 7. The second roll gives 11. Are these results conditionally independent? Are they independent?

answer: These results are conditionally independent. For example, for the hypothesis of the 8-sided die we have:

$$\begin{aligned} P(\text{7 on roll 1} \,|\, \text{8-sided die}) &= 1/8 \\ P(\text{11 on roll 2} \,|\, \text{8-sided die}) &= 0 \\ P(\text{7 on roll 1, 11 on roll 2} \,|\, \text{8-sided die}) &= 0 \end{aligned}$$

For the hypothesis of the 20-sided die we have:

$$\begin{aligned} P(\text{7 on roll 1} \,|\, \text{20-sided die}) &= 1/20 \\ P(\text{11 on roll 2} \,|\, \text{20-sided die}) &= 1/20 \\ P(\text{7 on roll 1, 11 on roll 2} \,|\, \text{20-sided die}) &= (1/20)^2 \end{aligned}$$

However, the results of the rolls are *not* independent. That is:

$$P(\text{7 on roll 1, 11 on roll 2}) \neq P(\text{7 on roll 1})P(\text{11 on roll 2}).$$

Intuitively, this is because a 7 on the roll 1 allows us to rule out the 4- and 6-sided dice, making an 11 on roll 2 more likely. Let's check this intuition by computing both sides precisely. On the righthand side we have:

$$P(\text{7 on roll 1}) = \frac{1}{5}\cdot\frac{1}{8} + \frac{1}{5}\cdot\frac{1}{12} + \frac{1}{5}\cdot\frac{1}{20} = \frac{31}{600}$$
$$P(\text{11 on roll 2}) = \frac{1}{5}\cdot\frac{1}{12} + \frac{1}{5}\cdot\frac{1}{20} = \frac{2}{75}$$

On the lefthand side we have:

$$\begin{aligned} P(\text{7 on roll 1, 11 on roll 2}) &= P(\text{11 on roll 2} \,|\, \text{7 on roll 1})P(\text{7 on roll 1}) \\ &= \left(\frac{30}{93}\cdot\frac{1}{12} + \frac{6}{31}\cdot\frac{1}{20}\right)\cdot\frac{31}{600} \\ &= \frac{17}{465}\cdot\frac{31}{600} = \frac{17}{9000} \end{aligned}$$

Here $\frac{30}{93}$ and $\frac{6}{31}$ are the posterior probabilities of the 12- and 20-sided dice given a 7 on roll 1. We conclude that, without conditioning on hypotheses, the rolls are not independent.

Returning the to general setup, if D_1 and D_2 are conditionally independent for H and H^c then it makes sense to consider each Bayes factor independently:

$$BF_i = \frac{P(D_i|H)}{P(D_i|H^c)}.$$

The prior odds of H are $O(H)$. The posterior odds after D_1 are

$$O(H|D_1) = BF_1 \cdot O(H).$$

And the posterior odds after D_1 and D_2 are

$$\begin{aligned} O(H|D_1, D_2) &= BF_2 \cdot O(H|D_1) \\ &= BF_2 \cdot BF_1 \cdot O(H) \end{aligned}$$

We have the beautifully simple notion that updating with new data just amounts to multiplying the current posterior odds by the Bayes factor of the new data.

Example 3. Other symptoms of Marfan Syndrome

Recall from the earlier example that the Bayes factor for a least one ocular feature (F) is

$$BF_F = \frac{P(F|M)}{P(F|M^c)} = \frac{0.7}{0.07} = 10.$$

The wrist sign (W) is the ability to wrap one hand around your other wrist to cover your pinky nail with your thumb. Assume 10% of the population have the wrist sign, while 90% of people with Marfan's have it. Therefore the Bayes factor for the wrist sign is

$$BF_W = \frac{P(W|M)}{P(W|M^c)} = \frac{0.9}{0.1} = 9.$$

We will assume that F and W are conditionally independent symptoms. That is, among people with Marfan syndrome, ocular features and the wrist sign are independent, and among people without Marfan syndrome, ocular features and the wrist sign are independent. Given this assumption, the posterior odds of Marfan syndrome for someone with both an ocular feature and the wrist sign are

$$O(M|F,W) = BF_W \cdot BF_F \cdot O(M) = 9 \cdot 10 \cdot \frac{1}{14999} \approx \frac{6}{1000}.$$

We can convert the posterior odds back to probability, but since the odds are so small the result is nearly the same:

$$P(M|F,W) \approx \frac{6}{1000+6} \approx 0.596\%.$$

So ocular features and the wrist sign are both strong evidence in favor of the hypothesis M, and taken together they are very strong evidence. Again, because the prior odds are so small, it is still unlikely that the person has Marfan syndrome, but at this point it might be worth undergoing further testing given potentially fatal consequences of the disease (such as aortic aneurysm or dissection).

Note also that if a person has exactly one of the two symptoms, then the product of the Bayes factors is near 1 (either 9/10 or 10/9). So the two pieces of data essentially cancel each other out with regard to the evidence they provide for Marfan's syndrome.

5 Log odds

In practice, people often find it convenient to work with the natural log of the odds in place of odds. Naturally enough these are called the log odds. The Bayesian update formula

$$O(H|D_1,D_2) = BF_2 \cdot BF_1 \cdot O(H)$$

becomes

$$\ln(O(H|D_1,D_2)) = \ln(BF_2) + \ln(BF_1) + \ln(O(H)).$$

We can interpret the above formula for the posterior log odds as the sum of the prior log odds and all the evidence $\ln(BF_i)$ provided by the data. Note that by taking logs, evidence in favor ($BF_i > 1$) is positive and evidence against ($BF_i < 1$) is negative.

To avoid lengthier computations, we will work with odds rather than log odds in this course. Log odds are nice because sums are often more intuitive then products. Log odds also play a central role in logistic regression, an important statistical model related to linear regression.

Bayesian Updating with Continuous Priors
Class 13, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand a parameterized family of distributions as representing a continuous range of hypotheses for the observed data.
2. Be able to state Bayes' theorem and the law of total probability for continous densities.
3. Be able to apply Bayes' theorem to update a prior probability density function to a posterior pdf given data and a likelihood function.
4. Be able to interpret and compute posterior predictive probabilities.

2 Introduction

Up to now we have only done Bayesian updating when we had a finite number of hypothesis, e.g. our dice example had five hypotheses (4, 6, 8, 12 or 20 sides). Now we will study Bayesian updating when there is a continuous range of hypotheses. The Bayesian update process will be essentially the same as in the discrete case. As usual when moving from discrete to continuous we will need to replace the probability mass function by a probability density function, and sums by integrals.

The first few sections of this note are devoted to working with pdfs. In particular we will cover the law of total probability and Bayes' theorem. We encourage you to focus on how these are essentially identical to the discrete versions. After that, we will apply Bayes' theorem and the law of total probability to Bayesian updating.

3 Examples with continuous ranges of hypotheses

Here are three standard examples with continuous ranges of hypotheses.

Example 1. Suppose you have a system that can succeed or fail with probability p. Then we can hypothesize that p is anywhere in the range $[0, 1]$. That is, we have a continuous range of hypotheses. We will often model this example with a 'bent' coin with unknown probability p of heads.

Example 2. The lifetime of a certain isotope is modeled by an exponential distribution $\exp(\lambda)$. In principal, the mean lifetime $1/\lambda$ can be any real number in $(0, \infty)$.

Example 3. We are not restricted to a single parameter. In principle, the parameters μ and σ of a normal distribution can be any real numbers in $(-\infty, \infty)$ and $(0, \infty)$, respectively. If we model gestational length for single births by a normal distribution, then from millions of data points we know that μ is about 40 weeks and σ is about one week.

In all of these examples we modeled the random process giving rise to the data by a distribution with parameters –called a parametrized distribution. Every possible choice of the parameter(s) is a hypothesis, e.g. we can hypothesize that the probability of succcess in Example 1 is $p = 0.7313$. We have a continuous set of hypotheses because we could take any value between 0 and 1.

4 Notational conventions

4.1 Parametrized models

As in the examples above our hypotheses often take the form a certain parameter has value θ. We will often use the letter θ to stand for an arbitrary hypothesis. This will leave symbols like p, f, and x to take there usual meanings as pmf, pdf, and data. Also, rather than saying 'the hypothesis that the parameter of interest has value θ' we will simply say the hypothesis θ.

4.2 Big and little letters

We have two parallel notations for outcomes and probability:

1. (Big letters) Event A, probability function $P(A)$.
2. (Little letters) Value x, pmf $p(x)$ or pdf $f(x)$.

These notations are related by $P(X = x) = p(x)$, where x is a value the discrete random variable X and '$X = x$' is the corresponding event.

We carry these notations over to the probabilities used in Bayesian updating.

1. (Big letters) From hypotheses $\mathcal{H}$ and data $\mathcal{D}$ we compute several associated probabilities

$$P(\mathcal{H}),\ P(\mathcal{D}),\ P(\mathcal{H}|\mathcal{D}),\ P(\mathcal{D}|\mathcal{H}).$$

In the coin example we might have $\mathcal{H}$ = 'the chosen coin has probability 0.6 of heads', $\mathcal{D}$ = 'the flip was heads', and $P(\mathcal{D}|\mathcal{H}) = 0.6$

2. (Small letters) Hypothesis values θ and data values x both have probabilities or probability densities:

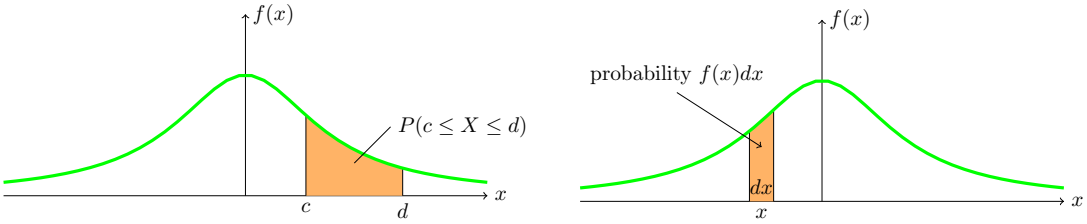
$$\begin{array}{cccc} p(\theta) & p(x) & p(\theta|x) & p(x|\theta) \\ f(\theta) & f(x) & f(\theta|x) & f(x|\theta) \end{array}$$

In the coin example we might have $\theta = 0.6$ and $x = 1$, so $p(x|\theta) = 0.6$. We might also write $p(x = 1|\theta = 0.6)$ to emphasize the values of x and θ, but we will never just write $p(1|0.6)$ because it is unclear which value is x and which is θ.

Although we will still use both types of notation, from now on we will mostly use the small letter notation involving pmfs and pdfs. Hypotheses will usually be parameters represented by Greek letters $(\theta, \lambda, \mu, \sigma, \dots)$ while data values will usually be represented by English letters $(x, x_i, y, \dots)$.

5 Quick review of pdf and probability

Suppose X is a random variable with pdf $f(x)$. Recall $f(x)$ is a density; its units are probability/(units of x).



The probability that the value of X is in $[c, d]$ is given by

$$\int_c^d f(x)\,dx.$$

The probability that X is in an infinitesimal range dx around x is $f(x)\,dx$. In fact, the integral formula is just the 'sum' of these infinitesimal probabilities. We can visualize these probabilities by viewing the integral as area under the graph of $f(x)$.

In order to manipulate probabilities instead of densities in what follows, we will make frequent use of the notion that $f(x)\,dx$ is the probability that X is in an infinitesimal range around x of width dx. Please make sure that you fully understand this notion.

6 Continuous priors, discrete likelihoods

In the Bayesian framework we have probabilities of hypotheses –called prior and posterior probabilities– and probabilities of data given a hypothesis –called likelihoods. In earlier classes both the hypotheses and the data had discrete ranges of values. We saw in the introduction that we might have a continuous range of hypotheses. The same is true for the data, but for today we will assume that our data can only take a discrete set of values. In this case, the likelihood of data x given hypothesis θ is written using a pmf: $p(x|\theta)$.

We will use the following coin example to explain these notions. We will carry this example through in each of the succeeding sections.

Example 4. Suppose we have a bent coin with unknown probability θ of heads. The value of of θ is random and could be anywhere between 0 and 1. For this and the examples that follow we'll suppose that the value of θ follows a distribution with continuous prior probability density $f(\theta) = 2\theta$. We have a discrete likelihood because tossing a coin has only two outcomes, $x = 1$ for heads and $x = 0$ for tails.

$$p(x = 1|\theta) = \theta, \qquad p(x = 0|\theta) = 1 - \theta.$$

Think: This can be tricky to wrap your mind around. We have a coin with an unknown probability θ of heads. The value of the parameter θ is itself random and has a prior pdf $f(\theta)$. It may help to see that the discrete examples we did in previous classes are similar. For example, we had a coin that might have probability of heads 0.5, 0.6, or 0.9. So,

we called our hypotheses $H_{0.5}$, $H_{0.6}$, $H_{0.9}$ and these had prior probabilities $P(H_{0.5})$ etc. In other words, we had a coin with an unknown probability of heads, we had hypotheses about that probability and each of these hypotheses had a prior probability.

7 The law of total probability

The law of total probability for continuous probability distributions is essentially the same as for discrete distributions. We replace the prior pmf by a prior pdf and the sum by an integral. We start by reviewing the law for the discrete case.

Recall that for a discrete set of hypotheses $\mathcal{H}_1, \mathcal{H}_2, \ldots \mathcal{H}_n$ the law of total probability says

$$P(\mathcal{D}) = \sum_{i=1}^{n} P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i). \tag{1}$$

This is the total prior probability of $\mathcal{D}$ because we used the prior probabilities $P(\mathcal{H}_i)$

In the little letter notation with $\theta_1, \theta_2, \ldots, \theta_n$ for hypotheses and x for data the law of total probability is written

$$p(x) = \sum_{i=1}^{n} p(x|\theta_i)p(\theta_i). \tag{2}$$

We also called this the prior predictive probability of the outcome x to distinguish it from the prior probability of the hypothesis θ.

Likewise, there is a law of total probability for continuous pdfs. We state it as a theorem using little letter notation.

Theorem. Law of total probability. Suppose we have a continuous parameter θ in the range $[a, b]$, and discrete random data x. Assume θ is itself random with density $f(\theta)$ and that x and θ have likelihood $p(x|\theta)$. In this case, the total probability of x is given by the formula.

$$p(x) = \int_a^b p(x|\theta)f(\theta)\,d\theta \tag{3}$$

Proof. Our proof will be by analogy to the discrete version: The probability term $p(x|\theta)f(\theta)\,d\theta$ is perfectly analogous to the term $p(x|\theta_i)p(\theta_i)$ in Equation 2 (or the term $P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i)$ in Equation 1). Continuing the analogy: the sum in Equation 2 becomes the integral in Equation 3

As in the discrete case, when we think of θ as a hypothesis explaining the probability of the data we call $p(x)$ the prior predictive probability for x.

Example 5. (Law of total probability.) Continuing with Example 4. We have a bent coin with probability θ of heads. The value of θ is random with prior pdf $f(\theta) = 2\theta$ on $[0, 1]$.

Suppose I flip the coin once. What is the total probability of heads?

answer: In Example 4 we noted that the likelihoods are $p(x = 1|\theta) = \theta$ and $p(x = 0|\theta) = 1 - \theta$. So the total probability of $x = 1$ is

$$p(x = 1) = \int_0^1 p(x = 1|\theta)\, f(\theta)\, d\theta = \int_0^1 \theta \cdot 2\theta\, d\theta = \int_0^1 2\theta^2 d\theta = \frac{2}{3}.$$

Since the prior is weighted towards higher probabilities of heads, so is the total probability.

8 Bayes' theorem for continuous probability densities

The statement of Bayes' theorem for continuous pdfs is essentially identical to the statement for pmfs. We state it including $d\theta$ so we have genuine probabilities:

Theorem. Bayes' Theorem. Use the same assumptions as in the law of total probability, i.e. θ is a continuous parameter with pdf $f(\theta)$ and range $[a, b]$; x is random discrete data; together they have likelihood $p(x|\theta)$. With these assumptions:

$$f(\theta|x)\,d\theta = \frac{p(x|\theta)f(\theta)\,d\theta}{p(x)} = \frac{p(x|\theta)f(\theta)\,d\theta}{\int_a^b p(x|\theta)f(\theta)\,d\theta}. \tag{4}$$

Proof. Since this is a statement about probabilities it is just the usual statement of Bayes' theorem. This is important enough to warrant spelling it out in words: Let Θ be the random variable that produces the value θ. Consider the events

$$H = \text{'}\Theta \text{ is in an interval of width } d\theta \text{ around the value } \theta\text{'}$$

and

$$D = \text{'the value of the data is } x\text{'}.$$

Then $P(H) = f(\theta)\,d\theta$, $P(D) = p(x)$, and $P(D|H) = p(x|\theta)$. Now our usual form of Bayes' theorem becomes

$$f(\theta|x)\,d\theta = P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{p(x|\theta)f(\theta)\,d\theta}{p(x)}$$

Looking at the first and last terms in this equation we see the new form of Bayes' theorem.

Finally, we firmly believe that is is more conducive to careful thinking about probability to keep the factor of $d\theta$ in the statement of Bayes' theorem. But because it appears in the numerator on both sides of Equation 4 many people drop the $d\theta$ and write Bayes' theorem in terms of densities as

$$f(\theta|x) = \frac{p(x|\theta)f(\theta)}{p(x)} = \frac{p(x|\theta)f(\theta)}{\int_a^b p(x|\theta)f(\theta)\,d\theta}.$$

9 Bayesian updating with continuous priors

Now that we have Bayes' theorem and the law of total probability we can finally get to Bayesian updating. Before continuing with Example 4, we point out two features of the Bayesian updating table that appears in the next example:

1. The table for continuous priors is very simple: since we cannot have a row for each of an infinite number of hypotheses we'll have just one row which uses a variable to stand for all hypotheses θ.

2. By including $d\theta$, all the entries in the table are probabilities and all our usual probability rules apply.

Example 6. (Bayesian updating.) Continuing Examples 4 and 5. We have a bent coin with unknown probability θ of heads. The value of θ is random with prior pdf $f(\theta) = 2\theta$. Suppose we flip the coin once and get heads. Compute the posterior pdf for θ.

answer: We make an update table with the usual columns. Since this is our first example the first row is the abstract version of Bayesian updating in general and the second row is Bayesian updating for this particular example.

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$f(\theta)\,d\theta$	$p(x=1\|\theta)$	$p(x=1\|\theta)f(\theta)\,d\theta$	$f(\theta\|x=1)\,d\theta$
θ	$2\theta\,d\theta$	θ	$2\theta^2\,d\theta$	$3\theta^2\,d\theta$
total	$\int_a^b f(\theta)\,d\theta = 1$		$p(x=1) = \int_0^1 2\theta^2\,d\theta = 2/3$	1

Therefore the posterior pdf (after seeing 1 heads) is $\boxed{f(\theta|x) = 3\theta^2}$.

We have a number of comments:

1. Since we used the prior probability $f(\theta)\,d\theta$, the hypothesis should have been:
'the unknown paremeter is in an interval of width $d\theta$ around θ'.
Even for us that is too much to write, so you will have to think it everytime we write that the hypothesis is θ.

2. The posterior pdf for θ is found by removing the $d\theta$ from the posterior probability in the table.

$$f(\theta|x) = 3\theta^2.$$

3. (i) As always $p(x)$ is the total probability. Since we have a continuous distribution instead of a sum we compute an integral.

(ii) Notice that by including $d\theta$ in the table, it is clear what integral we need to compute to find the total probability $p(x)$.

4. The table organizes the continuous version of Bayes' theorem. Namely, the posterior pdf is related to the prior pdf and likelihood function via:

$$f(\theta|x)d\theta = \frac{p(x|\theta)\,f(\theta)d\theta}{\int_a^b p(x|\theta)f(\theta)\,d\theta} = \frac{p(x|\theta)\,f(\theta)}{p(x)}$$

Removing the $d\theta$ in the numerator of both sides we have the statement in terms of densities.

5. Regarding both sides as functions of θ, we can again express Bayes' theorem in the form:

$$f(\theta|x) \propto p(x|\theta) \cdot f(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

9.1 Flat priors

One important prior is called a flat or uniform prior. A flat prior assumes that every hypothesis is equally probable. For example, if θ has range $[0, 1]$ then $f(\theta) = 1$ is a flat prior.

Example 7. (Flat priors.) We have a bent coin with unknown probability θ of heads. Suppose we toss it once and get tails. Assume a flat prior and find the posterior probability for θ.

answer: This is the just Example 6 with a change of prior and likelihood.

hypothesis θ	prior $f(\theta)\,d\theta$	likelihood $p(x=0\|\theta)$	Bayes numerator	posterior $f(\theta\|x=0)\,d\theta$
θ	$1\cdot d\theta$	$1-\theta$	$(1-\theta)\,d\theta$	$2(1-\theta)\,d\theta$
total	$\int_a^b f(\theta)\,d\theta = 1$		$p(x=0) = \int_0^1 (1-\theta)\,d\theta = 1/2$	1

9.2 Using the posterior pdf

Example 8. In the previous example the prior probability was flat. First show that this means that a priori the coin is equally like to be biased towards heads or tails. Then, after observing one heads, what is the (posterior) probability that the coin is biased towards heads?

answer: Since the parameter θ is the probability the coin lands heads, the first part of the problem asks us to show $P(\theta > .5) = 0.5$ and the second part asks for $P(\theta > .5 \,|\, x = 1)$. These are easily computed from the prior and posterior pdfs respectively.

The prior probability that the coin is biased towards heads is

$$P(\theta > .5) = \int_{.5}^1 f(\theta)\,d\theta = \int_{.5}^1 1\cdot d\theta = \theta|_{.5}^1 = \frac{1}{2}.$$

The probability of $1/2$ means the coin is equally likely to be biased toward heads or tails. The posterior probabilitiy that it's biased towards heads is

$$P(\theta > .5|x=1) = \int_{.5}^1 f(\theta|x=1)\,d\theta = \int_{.5}^1 2\theta\,d\theta = \theta^2\big|_{.5}^1 = \frac{3}{4}.$$

We see that observing one heads has increased the probability that the coin is biased towards heads from $1/2$ to $3/4$.

10 Predictive probabilities

Just as in the discrete case we are also interested in using the posterior probabilities of the hypotheses to make predictions for what will happen next.

Example 9. (Prior and posterior prediction.) Continuing Examples 4, 5, 6: we have a coin with unknown probability θ of heads and the value of θ has prior pdf $f(\theta) = 2\theta$. Find the prior predictive probability of heads. Then suppose the first flip was heads and find the posterior predictive probabilities of both heads and tails on the second flip.

answer: For notation let x_1 be the result of the first flip and let x_2 be the result of the second flip. The prior predictive probability is exactly the total probability computed in Examples 5 and 6.

$$p(x_1 = 1) = \int_0^1 p(x_1 = 1|\theta) f(\theta)\,d\theta = \int_0^1 2\theta^2\,d\theta = \frac{2}{3}.$$

The posterior predictive probabilities are the total probabilities computed using the posterior pdf. From Example 6 we know the posterior pdf is $f(\theta|x_1 = 1) = 3\theta^2$. So the posterior predictive probabilities are

$$p(x_2 = 1|x_1 = 1) = \int_0^1 p(x_2 = 1|\theta, x_1 = 1) f(\theta|x_1 = 1)\, d\theta = \int_0^1 \theta \cdot 3\theta^2\, d\theta = 3/4$$

$$p(x_2 = 0|x_1 = 1) = \int_0^1 p(x_2 = 0|\theta, x_1 = 1) f(\theta|x_1 = 1)\, d\theta = \int_0^1 (1-\theta) \cdot 3\theta^2\, d\theta = 1/4$$

(More simply, we could have computed $p(x_2 = 0|x_1 = 1) = 1 - p(x_2 = 1|x_1 = 1) = 1/4$.)

11 From discrete to continuous Bayesian updating

To develop intuition for the transition from discrete to continuous Bayesian updating, we'll walk a familiar road from calculus. Namely we will:

(i) approximate the continuous range of hypotheses by a finite number.

(ii) create the discrete updating table for the finite number of hypotheses.

(iii) consider how the table changes as the number of hypotheses goes to infinity.

In this way, will see the prior and posterior pmf's converge to the prior and posterior pdf's.

Example 10. To keep things concrete, we will work with the 'bent' coin with a flat prior $f(\theta) = 1$ from Example 7. Our goal is to go from discrete to continuous by increasing the number of hypotheses

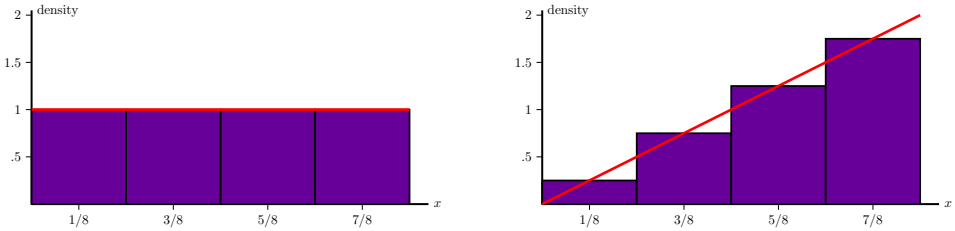
4 hypotheses. We slice $[0, 1]$ into 4 equal intervals: $[0, 1/4]$, $[1/4, 1/2]$, $[1/2, 3/4]$, $[3/4, 1]$. Each slice has width $\Delta\theta = 1/4$. We put our 4 hypotheses θ_i at the centers of the four slices:

$$\theta_1\text{: '}\theta = 1/8\text{'}, \quad \theta_2\text{: '}\theta = 3/8\text{'}, \quad \theta_3\text{: '}\theta = 5/8\text{'}, \quad \theta_4\text{: '}\theta = 7/8\text{'}.$$

The flat prior gives each hypothesis a probability of $1/4 = 1 \cdot \Delta\theta$. We have the table:

hypothesis	prior	likelihood	Bayes num.	posterior
$\theta = 1/8$	1/4	1/8	$(1/4) \times (1/8)$	1/16
$\theta = 3/8$	1/4	3/8	$(1/4) \times (3/8)$	3/16
$\theta = 5/8$	1/4	5/8	$(1/4) \times (5/8)$	5/16
$\theta = 7/8$	1/4	7/8	$(1/4) \times (7/8)$	7/16
Total	1	–	$\sum_{i=1}^{n} \theta_i\, \Delta\theta$	1

Here are the density histograms of the prior and posterior pmf. The prior and posterior pdfs from Example 7 are superimposed on the histograms in red.

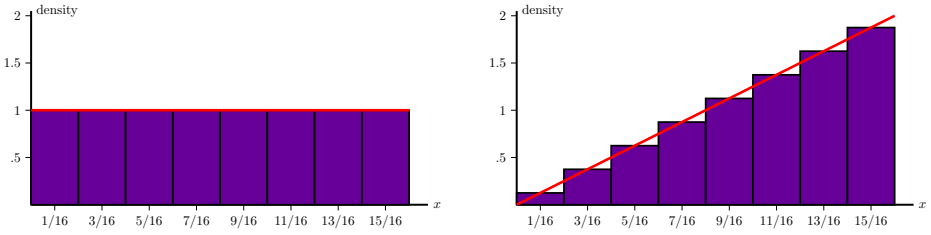


8 hypotheses. Next we slice [0,1] into 8 intervals each of width $\Delta\theta = 1/8$ and use the center of each slice for our 8 hypotheses θ_i.

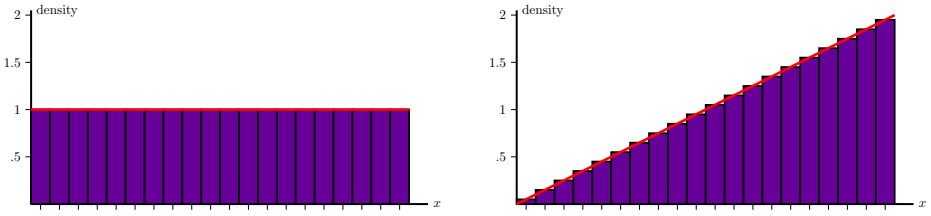
θ_1: '$\theta = 1/16$', θ_2: '$\theta = 3/16$', θ_3: '$\theta = 5/16$', θ_4: '$\theta = 7/16$'
θ_5: '$\theta = 9/16$', θ_6: '$\theta = 11/16$', θ_7: '$\theta = 13/16$', θ_8: '$\theta = 15/16$'

The flat prior gives each hypothesis the probablility $1/8 = 1 \cdot \Delta\theta$. Here are the table and density histograms.

hypothesis	prior	likelihood	Bayes num.	posterior
$\theta = 1/16$	1/8	1/16	$(1/8) \times (1/16)$	1/64
$\theta = 3/16$	1/8	3/16	$(1/8) \times (3/16)$	3/64
$\theta = 5/16$	1/8	5/16	$(1/8) \times (5/16)$	5/64
$\theta = 7/16$	1/8	7/16	$(1/8) \times (7/16)$	7/64
$\theta = 9/16$	1/8	9/16	$(1/8) \times (9/16)$	9/64
$\theta = 11/16$	1/8	11/16	$(1/8) \times (11/16)$	11/64
$\theta = 13/16$	1/8	13/16	$(1/8) \times (13/16)$	13/64
$\theta = 15/16$	1/8	15/16	$(1/8) \times (15/16)$	15/64
Total	1	–	$\sum_{i=1}^{n} \theta_i \, \Delta\theta$	1



20 hypotheses. Finally we slice [0,1] into 20 pieces. This is essentially identical to the previous two cases. Let's skip right to the density histograms.



Looking at the sequence of plots we see how the prior and posterior density histograms converge to the prior and posterior probability density functions.

Notational conventions
Class 13, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to work with the various notations and terms we use to describe probabilities and likelihood.

2 Introduction

We've introduced a number of different notations for probability, hypotheses and data. We collect them here, to have them in one place.

3 Notation and terminology for data and hypotheses

The problem of labeling data and hypotheses is a tricky one. When we started the course we talked about outcomes, e.g. heads or tails. Then when we introduced random variables we gave outcomes numerical values, e.g. 1 for heads and 0 for tails. This allowed us to do things like compute means and variances. We need to do something similar now. Recall our notational conventions:

- Events are labeled with capital letters, e.g. A, B, C.
- A random variable is capital X and takes values small x.
- The connection between values and events: '$X = x$' is the event that X takes the value x.
- The probability of an event is capital $P(A)$.
- A discrete random variable has a probability mass function small $p(x)$ The connection between P and p is that $P(X = x) = p(x)$.
- A continuous random variable has a probability density function $f(x)$ The connection between P and f is that $P(a \le X \le b) = \int_a^b f(x)\,dx$.
- For a continuous random variable X the probability that X is in an infinitesimal interval of width dx round x is $f(x)\,dx$.

In the context of Bayesian updating we have similar conventions.

- We use capital letters, especially $\mathcal{H}$, to indicate a hypothesis, e.g. $\mathcal{H}$ = 'the coin is fair'.

- We use lower case letters, especially θ, to indicate the hypothesized value of a model parameter, e.g. the probability the coin lands heads is $\theta = 0.5$.
- We use upper case letters, especially $\mathcal{D}$, when talking about data as events. For example, $\mathcal{D}$ = 'the sequence of tosses was HTH.
- We use lower case letters, especially x, when talking about data as values. For example, the sequence of data was $x_1, x_2, x_3 = 1, 0, 1$.
- When the set of hypotheses is discrete we can use the probability of individual hypotheses, e.g. $p(\theta)$. When the set is continuous we need to use the probability for an infinitesimal range of hypotheses, e.g. $f(\theta)\,d\theta$.

The following table summarizes this for discret θ and continuous θ. In both cases we are assuming a discrete set of possible outcomes (data) x. Tomorrow we will deal with a continuous set of outcomes.

	hypothesis	prior	likelihood	Bayes numerator	posterior
	$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
Discrete θ:	θ	$p(\theta)$	$p(x\|\theta)$	$p(x\|\theta)p(\theta)$	$p(\theta\|x)$
Continuous θ:	θ	$f(\theta)\,d\theta$	$p(x\|\theta)$	$p(x\|\theta)f(\theta)\,d\theta$	$f(\theta\|x)\,d\theta$

Remember the continuous hypothesis θ is really a shorthand for 'the parameter θ is in an interval of width $d\theta$ around θ'.

Beta Distributions
Class 14, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be familiar with the 2-parameter family of beta distributions and its normalization.
2. Be able to update a beta prior to a beta posterior in the case of a binomial likelihood.

2 Beta distribution

The beta distribution beta(a, b) is a two-parameter distribution with range $[0, 1]$ and pdf

$$f(\theta) = \frac{(a+b-1)!}{(a-1)!(b-1)!}\theta^{a-1}(1-\theta)^{b-1}$$

We have made an applet so you can explore the shape of the Beta distribution as you vary the parameters:

http://mathlets.org/mathlets/beta-distribution/.

As you can see in the applet, the beta distribution may be defined for any real numbers $a > 0$ and $b > 0$. In 18.05 we will stick to integers a and b, but you can get the full story here: http://en.wikipedia.org/wiki/Beta_distribution

In the context of Bayesian updating, a and b are often called hyperparameters to distinguish them from the unknown parameter θ representing our hypotheses. In a sense, a and b are 'one level up' from θ since they parameterize its pdf.

2.1 A simple but important observation!

If a pdf $f(\theta)$ has the form $c\theta^{a-1}(1-\theta)^{b-1}$ then $f(\theta)$ is a beta(a, b) distribution and the normalizing constant must be

$$c = \frac{(a+b-1)!}{(a-1)!\,(b-1)!}.$$

This follows because the constant c must normalize the pdf to have total probability 1. There is only one such constant and it is given in the formula for the beta distribution.

A similar observation holds for normal distributions, exponential distributions, and so on.

2.2 Beta priors and posteriors for binomial random variables

Example 1. Suppose we have a bent coin with unknown probability θ of heads. We toss it 12 times and get 8 heads and 4 tails. Starting with a flat prior, show that the posterior pdf is a beta$(9, 5)$ distribution.

answer: This is nearly identical to examples from the previous class. We'll call the data from all 12 tosses x_1. In the following table we call the leading constant factor in the posterior column c_2. Our simple observation will tell us that it has to be the constant factor from the beta pdf.

The data is 8 heads and 4 tails. Since this comes from a binomial(12, θ) distribution, the likelihood $p(x_1|\theta) = \binom{12}{8}\theta^8(1-\theta)^4$. Thus the Bayesian update table is

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$1 \cdot d\theta$	$\binom{12}{8}\theta^8(1-\theta)^4$	$\binom{12}{8}\theta^8(1-\theta)^4\,d\theta$	$c_2\,\theta^8(1-\theta)^4\,d\theta$
total	1		$T = \binom{12}{8}\int_0^1 \theta^8(1-\theta)^4\,d\theta$	1

Our simple observation above holds with $a = 9$ and $b = 5$. Therefore the posterior pdf

$$f(\theta|x_1) = c_2\theta^8(1-\theta)^4$$

follows a beta$(9, 5)$ distribution and the normalizing constant c_2 must be

$$c_2 = \frac{13!}{8!\,4!}.$$

Note: We explicitly included the binomial coefficient $\binom{12}{8}$ in the likelihood. We could just as easily have given it a name, say c_1 and not bothered making its value explicit.

Example 2. Now suppose we toss the same coin again, getting n heads and m tails. Using the posterior pdf of the previous example as our new prior pdf, show that the new posterior pdf is that of a beta$(9 + n, 5 + m)$ distribution.

answer: It's all in the table. We'll call the data of these $n + m$ additional tosses x_2. This time we won't make the binomial coefficient explicit. Instead we'll just call it c_3. Whenever we need a new label we will simply use c with a new subscript.

hyp.	prior	likelihood	Bayes posterior	numerator
θ	$c_2\theta^8(1-\theta)^4\,d\theta$	$c_3\,\theta^n(1-\theta)^m$	$c_2c_3\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$	$c_4\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$
total	1		$T = \int_0^1 c_2c_3\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$	1

Again our simple observation holds and therefore the posterior pdf

$$f(\theta|x_1, x_2) = c_4\theta^{n+8}(1-\theta)^{m+4}$$

follows a beta$(n + 9, m + 5)$ distribution.

Note: Flat beta. The beta$(1, 1)$ distribution is the same as the uniform distribution on $[0, 1]$, which we have also called the flat prior on θ. This follows by plugging $a = 1$ and $b = 1$ into the definition of the beta distribution, giving $f(\theta) = 1$.

Summary: If the probability of heads is θ, the number of heads in $n+m$ tosses follows a binomial$(n+m, \theta)$ distribution. We have seen that if the prior on θ is a beta distribution then so is the posterior; only the parameters a, b of the beta distribution change! We summarize precisely how they change in a table. We assume the data is n heads in $n+m$ tosses.

hypothesis	data	prior	likelihood	posterior
θ	$x = n$	beta(a, b)	binomial$(n+m, \theta)$	beta$(a+n, b+m)$
θ	$x = n$	$c_1\theta^{a-1}(1-\theta)^{b-1}\,d\theta$	$c_2\theta^n(1-\theta)^m$	$c_3\theta^{a+n-1}(1-\theta)^{b+m-1}\,d\theta$

2.3 Conjugate priors

In the literature you'll see that the beta distribution is called a conjugate prior for the binomial distribution. This means that if the likelihood function is binomial, then a beta prior gives a beta posterior. In fact, the beta distribution is a conjugate prior for the Bernoulli and geometric distributions as well.

We will soon see another important example: the normal distribution is its own conjugate prior. In particular, if the likelihood function is normal with known variance, then a normal prior gives a normal posterior.

Conjugate priors are useful because they reduce Bayesian updating to modifying the parameters of the prior distribution (so-called hyperparameters) rather than computing integrals. We saw this for the beta distribution in the last table. For many more examples see:
http://en.wikipedia.org/wiki/Conjugate_prior_distribution

Continuous Data with Continuous Priors
Class 14, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to construct a Bayesian update table for continuous hypotheses and continuous data.
2. Be able to recognize the pdf of a normal distribution and determine its mean and variance.

2 Introduction

We are now ready to do Bayesian updating when both the hypotheses and the data take continuous values. The pattern is the same as what we've done before, so let's first review the previous two cases.

3 Previous cases

1. Discrete hypotheses, discrete data

Notation

- Hypotheses $\mathcal{H}$
- Data x
- Prior $P(\mathcal{H})$
- Likelihood $p(x \,|\, \mathcal{H})$
- Posterior $P(\mathcal{H} \,|\, x)$.

Example 1. Suppose we have data x and three possible explanations (hypotheses) for the data that we'll call A, B, C. Suppose also that the data can take two possible values, -1 and 1.

In order to use the data to help estimate the probabilities of the different hypotheses we need a prior pmf and a likelihood table. Assume the prior and likelihoods are given in the following table. (For this example we are only concerned with the formal process of of Bayesian updating. So we just made up the prior and likelihoods.)

hypothesis $\mathcal{H}$	prior $P(\mathcal{H})$
A	0.1
B	0.3
C	0.6

Prior probabilities

hypothesis	likelihood $p(x \mid \mathcal{H})$	
$\mathcal{H}$	$x = -1$	$x = 1$
A	0.2	0.8
B	0.5	0.5
C	0.7	0.3

Likelihoods

Naturally, each entry in the likelihood table is a likelihood $p(x \mid \mathcal{H})$. For instance the 0.2 row A and column $x = -1$ is the likelihood $p(x = -1 \mid A)$.

Question: Suppose we run one trial and obtain the data $x_1 = 1$. Use this to find the posterior probabilities for the hypotheses.

answer: The data picks out one column from the likelihood table which we then use in our Bayesian update table.

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$p(x = 1 \mid \mathcal{H})$	$p(x \mid \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H} \mid x) = \dfrac{p(x \mid \mathcal{H})P(\mathcal{H})}{p(x)}$
A	0.1	0.8	0.08	0.195
B	0.3	0.5	0.15	0.366
C	0.6	0.3	0.18	0.439
total	1		$p(x) = 0.41$	1

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $p(x)$ is the sum of the probabilities in the Bayes numerator column; and we divide by $p(x)$ to normalize the Bayes numerator.

2. Continuous hypotheses, discrete data

Now suppose that we have data x that can take a discrete set of values and a continuous parameter θ that determines the distribution the data is drawn from.

Notation

- Hypotheses θ
- Data x
- Prior $f(\theta)\, d\theta$
- Likelihood $p(x \mid \theta)$
- Posterior $f(\theta \mid x)\, d\theta$.

Note: Here we multiplied by $d\theta$ to express the prior and posterior as probabilities. As densities, we have the prior pdf $f(\theta)$ and the posterior pdf $f(\theta \mid x)$.

Example 2. Assume that $x \sim \text{Binomial}(5, \theta)$. So θ is in the range $[0, 1]$ and the data x can take six possible values, 0, 1, ..., 5.

Since there is a continuous range of values we use a pdf to describe the prior on θ. Let's suppose the prior is $f(\theta) = 2\theta$. We can still make a likelihood table, though it only has one row representing an arbitrary hypothesis θ.

hypothesis	likelihood $p(x \mid \theta)$					
	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
θ	$\binom{5}{0}(1-\theta)^5$	$\binom{5}{1}\theta(1-\theta)^4$	$\binom{5}{2}\theta^2(1-\theta)^3$	$\binom{5}{3}\theta^3(1-\theta)^2$	$\binom{5}{4}\theta^4(1-\theta)$	$\binom{5}{5}\theta^5$

Likelihoods

Question: Suppose we run one trial and obtain the data $x_1 = 2$. Use this to find the posterior pdf for the parameter (hypotheses) θ.

answer: As before, the data picks out one column from the likelihood table which we can use in our Bayesian update table. Since we want to work with probabilities we write $f(\theta)d\,\theta$ and $f(\theta \mid x_1)\,d\theta$ for the pdf's.

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$f(\theta)\,d\theta$	$p(x = 2 \mid \theta)$	$p(x \mid \theta)f(\theta)\,d\theta$	$f(\theta \mid x)\,d\theta = \dfrac{p(x \mid \theta)f(\theta)\,d\theta}{p(x)}$
θ	$2\theta\,d\theta$	$\binom{5}{2}\theta^2(1-\theta)^3$	$2\binom{5}{2}\theta^3(1-\theta)^3\,d\theta$	$f(\theta \mid x)\,d\theta = \dfrac{3!\,3!}{7!}\theta^3(1-\theta)^3\,d\theta$
total	1	$p(x) = \int_0^1 2\binom{5}{2}\theta^2(1-\theta)^3\,d\theta = 2\binom{5}{2}\frac{3!\,3!}{7!}$		1

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $p(x)$ is the integral of the probabilities in the Bayes numerator column; and we divide by $p(x)$ to normalize the Bayes numerator.

4 Continuous hypotheses and continuous data

When both data and hypotheses are continuous, the only change to the previous example is that the likelihood function uses a pdf $f(x \mid \theta)$ instead of a pmf $p(x \mid \theta)$. The general shape of the Bayesian update table is the same.

Notation

- Hypotheses θ
- Data x
- Prior $f(\theta)d\theta$

- Likelihood $f(x \mid \theta)\, dx$
- Posterior $f(\theta \mid x)\, d\theta$.

Simplifying the notation. In the previous cases we included $d\theta$ so that we were working with probabilities instead of densities. When both data and hypotheses are continuous we will need both $d\theta$ and dx. This makes things conceptually simpler, but notationally cumbersome. To simplify the notation we will allow ourselves to dx in our tables. This is fine because the data x is a fixed. We keep the $d\theta$ because the hypothesis θ is allowed to vary.

For comparison, we first show the general table in simplified notation followed immediately afterward by the table showing the infinitesimals.

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$f(\theta)\, d\theta$	$f(x \mid \theta)$	$f(x \mid \theta) f(\theta)\, d\theta$	$f(\theta \mid x) = \dfrac{f(x \mid \theta) f(\theta)\, d\theta}{f(x)}$
total	1		$f(x) = \int f(x \mid \theta) f(\theta)\, d\theta$	1

Bayesian update table without dx

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$f(\theta)\, d\theta$	$f(x \mid \theta)\, dx$	$f(x \mid \theta) f(\theta)\, d\theta\, dx$	$f(\theta \mid x)\, d\theta = \dfrac{f(x \mid \theta) f(\theta)\, d\theta\, dx}{f(x)\, dx} = \dfrac{f(x \mid \theta) f(\theta)\, d\theta}{f(x)}$
total	1		$f(x)\, dx = \left(\int f(x \mid \theta) f(\theta)\, d\theta\right)\, dx$	1

Bayesian update table with $d\theta$ and dx

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $f(x)\, dx$ is the integral of the probabilities in the Bayes numerator column; we divide by $f(x)\, dx$ to normalize the Bayes numerator.

5 Normal hypothesis, normal data

A standard example of continuous hypotheses and continuous data assumes that both the data and prior follow normal distributions. The following example assumes that the variance of the data is known.

Example 3. Suppose we have data $x = 5$ which wass drawn from a normal distribution

with unknown mean θ and standard deviation 1.

$$x \sim \mathrm{N}(\theta, 1)$$

Suppose further that our prior distribution for θ is $\theta \sim \mathrm{N}(2, 1)$.

Let x represent an arbitrary data value.

(a) Make a Bayesian table with prior, likelihood, and Bayes numerator.

(b) Show that the posterior distribution for θ is normal as well.

(c) Find the mean and variance of the posterior distribution.

answer: As we did with the tables above, a good compromise on the notation is to include $d\theta$ but not dx. The reason for this is that the total probability is computed by integrating over θ and the $d\theta$ reminds of us that.

Our prior pdf is

$$f(\theta) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(\theta-2)^2/2}.$$

The likelihood function is

$$f(x = 5 \,|\, \theta) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(5-\theta)^2/2}.$$

We know we are going to multiply the prior and the likelihood, so we carry out that algebra first. In the very last step we simplify the constant factor into one constant we call c_1.

$$\begin{aligned}
\text{prior} \cdot \text{likelihood} &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(5-\theta)^2/2} \\
&= \frac{1}{2\pi} \mathrm{e}^{-(2\theta^2 - 14\theta + 29)/2} \\
&= \frac{1}{2\pi} \mathrm{e}^{-(\theta^2 - 7\theta + 29/2)} \quad \text{(complete the square)} \\
&= \frac{1}{2\pi} \mathrm{e}^{-((\theta-7/2)^2 + 9/4)} \\
&= \frac{\mathrm{e}^{-9/4}}{2\pi} \mathrm{e}^{-(\theta-7/2)^2)} \\
&= c_1 \mathrm{e}^{-(\theta-7/2)^2}
\end{aligned}$$

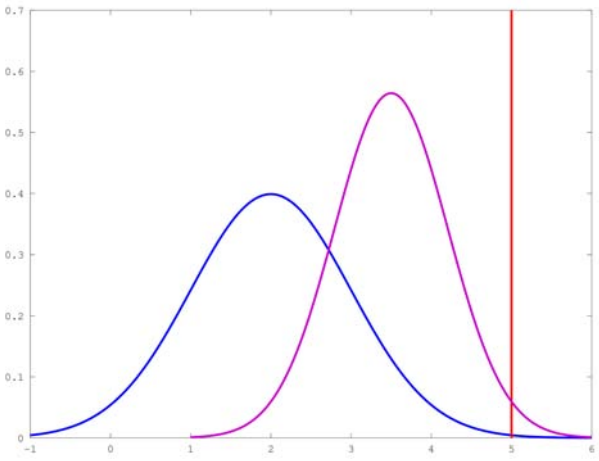
In the last step we replaced the complicated constant factor by the simpler expression c_1.

hypothesis	prior	likelihood	Bayes numerator	posterior $f(\theta \,|\, x = 5)\, d\theta$
θ	$f(\theta)\, d\theta$	$f(x = 5 \,|\, \theta)$	$f(x = 5 \,|\, \theta) f(\theta)\, d\theta$	$\dfrac{f(x = 5 \,|\, \theta) f(\theta)\, d\theta}{f(x = 5)}$
θ	$\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(\theta-2)^2/2}\, d\theta$	$\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(5-\theta)^2/2}$	$c_1 \mathrm{e}^{-(\theta-7/2)^2}$	$c_2 \mathrm{e}^{-(\theta-7/2)^2}$
total	1		$f(x = 5) = \int f(x = 5 \,|\, \theta) f(\theta)\, d\theta$	1

We can see by the form of the posterior pdf that it is a normal distribution. Because the exponential for a normal distribution is $e^{-(\theta-\mu)^2/2\sigma^2}$ we have mean $\mu = 7/2$ and $2\sigma^2 = 1$, so variance $\sigma^2 = 1/2$.

We don't need to bother computing the total probability; it is just used for normalization and we already know the normalization constant $\frac{1}{\sigma\sqrt{2\pi}}$ for a normal distribution.

Here is the graph of the prior and the posterior pdf's for this example. Note how the data 'pulls' the prior towards the data.



prior = blue; posterior = purple; data = red

Now we'll repeat the previous example for general x. When reading this if you mentally substitute 5 for x you will understand the algebra.

Example 4. Suppose our data x is drawn from a normal distribution with unknown mean θ and standard deviation 1.

$$x \sim \mathrm{N}(\theta, 1)$$

answer: As before, we show the algebra used to simplify the Bayes numerator: The prior pdf and likelihood function are

$$f(\theta) = \frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2} \qquad f(x \mid \theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}.$$

The Bayes numerator is the product of the prior and the likelihood:

$$\begin{aligned}
\text{prior} \cdot \text{likelihood} &= \frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} \\
&= \frac{1}{2\pi} e^{-(2\theta^2-(4+2x)\theta+4+x^2)/2} \\
&= \frac{1}{2\pi} e^{-(\theta^2-(2+x)\theta+(4+x^2)/2)} \quad \text{(complete the square)} \\
&= \frac{1}{2\pi} e^{-((\theta-(1+x/2))^2-(1+x/2)^2+(4+x^2)/2)} \\
&= c_1 e^{-(\theta-(1+x/2))^2}
\end{aligned}$$

Just as in the previous example, in the last step we replaced all the constants, including the exponentials that just involve x, by by the simple constant c_1.

Now the Bayesian update table becomes

hypothesis	prior	likelihood	Bayes numerator	posterior $f(\theta \mid x)\,d\theta$
θ	$f(\theta)\,d\theta$	$f(x \mid \theta)$	$f(x \mid \theta)f(\theta)\,d\theta$	$\dfrac{f(x \mid \theta)f(\theta)\,d\theta}{f(x)}$
θ	$\frac{1}{\sqrt{2\pi}}e^{-(\theta-2)^2/2}\,d\theta$	$\frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2}$	$c_1e^{-(\theta-(1+x/2))^2}$	$c_2e^{-(\theta-(1+x/2))^2}$
total	1		$f(x) = \int f(x \mid \theta)f(\theta)\,d\theta$	1

As in the previous example we can see by the form of the posterior that it must be a normal distribution with mean $1 + x/2$ and variance $1/2$. (Compare this with the case $x = 5$ in the previous example.)

6 Predictive probabilities

Since the data x is continuous it has prior and posterior predictive pdfs. The prior predictive pdf is the total probability density computed at the bottom of the Bayes numerator column:

$$f(x) = \int f(x|\theta)f(\theta)\,d\theta,$$

where the integral is computed over the entire range of θ.

The posterior predictive pdf has the same form as the prior predictive pdf, except it use the posterior probabilities for θ:

$$f(x_2|x_1) = \int f(x_2|\theta, x_1)f(\theta|x_1)\,d\theta,$$

As usual, we usually assume x_1 and x_2 are conditionally independent. That is,

$$f(x_2|\theta, x_1) = f(x_2|\theta).$$

In this case the formula for the posterior predictive pdf is a little simpler:

$$f(x_2|x_1) = \int f(x_2|\theta)f(\theta|x_1)\,d\theta,$$

Conjugate priors: Beta and normal
Class 15, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand the benefits of conjugate priors.
2. Be able to update a beta prior given a Bernoulli, binomial, or geometric likelihood.
3. Understand and be able to use the formula for updating a normal prior given a normal likelihood with known variance.

2 Introduction and definition

In this reading, we will elaborate on the notion of a conjugate prior for a likelihood function. With a conjugate prior the posterior is of the same type, e.g. for binomial likelihood the beta prior becomes a beta posterior. Conjugate priors are useful because they reduce Bayesian updating to modifying the parameters of the prior distribution (so-called hyperparameters) rather than computing integrals.

Our focus in 18.05 will be on two important examples of conjugate priors: beta and normal. For a far more comprehensive list, see the tables herein:

http://en.wikipedia.org/wiki/Conjugate_prior_distribution

We now give a definition of conjugate prior. It is best understood through the examples in the subsequent sections.

Definition. Suppose we have data with likelihood function $f(x|\theta)$ depending on a hypothesized parameter. Also suppose the prior distribution for θ is one of a family of parametrized distributions. If the posterior distribution for θ is in this family then we say the the prior is a conjugate prior for the likelihood.

3 Beta distribution

In this section, we will show that the beta distribution is a conjugate prior for binomial, Bernoulli, and geometric likelihoods.

3.1 Binomial likelihood

We saw last time that the beta distribution is a conjugate prior for the binomial distribution. This means that if the likelihood function is binomial and the prior distribution is beta then the posterior is also beta.

More specifically, suppose that the likelihood follows a binomial(N, θ) distribution where N is known and θ is the (unknown) parameter of interest. We also have that the data x from one trial is an integer between 0 and N. Then for a beta prior we have the following table:

hypothesis	data	prior	likelihood	posterior
θ	x	beta(a, b)	binomial(N, θ)	beta$(a + x, b + N - x)$
θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$c_2\theta^x(1-\theta)^{N-x}$	$c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

The table is simplified by writing the normalizing coefficient as c_1, c_2 and c_3 respectively. If needed, we can recover the values of the c_1 and c_2 by recalling (or looking up) the normalizations of the beta and binomial distributions.

$$c_1 = \frac{(a+b-1)!}{(a-1)!\,(b-1)!} \qquad c_2 = \binom{N}{x} = \frac{N!}{x!\,(N-x)!} \qquad c_3 = \frac{(a+b+N-1)!}{(a+x-1)!\,(b+N-x-1)!}$$

3.2 Bernoulli likelihood

The beta distribution is a conjugate prior for the Bernoulli distribution. This is actually a special case of the binomial distribution, since Bernoulli(θ) is the same as binomial(1, θ). We do it separately because it is slightly simpler and of special importance. In the table below, we show the updates corresponding to success ($x = 1$) and failure ($x = 0$) on separate rows.

hypothesis	data	prior	likelihood	posterior
θ	x	beta(a, b)	Bernoulli(θ)	beta$(a + 1, b)$ or beta$(a, b + 1)$
θ	$x = 1$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	θ	$c_3\theta^a(1-\theta)^{b-1}$
θ	$x = 0$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$1-\theta$	$c_3\theta^{a-1}(1-\theta)^b$

The constants c_1 and c_3 have the same formulas as in the previous (binomial likelihood case) with $N = 1$.

3.3 Geometric likelihood

Recall that the geometric(θ) distribution describes the probability of x successes before the first failure, where the probability of success on any single independent trial is θ. The corresponding pmf is given by $p(x) = \theta^x(1-\theta)$.

Now suppose that we have a data point x, and our hypothesis θ is that x is drawn from a geometric(θ) distribution. From the table we see that the beta distribution is a conjugate prior for a geometric likelihood as well:

hypothesis	data	prior	likelihood	posterior
θ	x	beta(a, b)	geometric(θ)	beta$(a + x, b + 1)$
θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$\theta^x(1-\theta)$	$c_3\theta^{a+x-1}(1-\theta)^b$

At first it may seem strange that the beta distribution is a conjugate prior for both the binomial and geometric distributions. The key reason is that the binomial and geometric likelihoods are proportional as functions of θ. Let's illustrate this in a concrete example.

Example 1. While traveling through the Mushroom Kingdom, Mario and Luigi find some rather unusual coins. They agree on a prior of $f(\theta) \sim$ beta(5,5) for the probability of heads,

though they disagree on what experiment to run to investigate θ further.

a) Mario decides to flip a coin 5 times. He gets four heads in five flips.

b) Luigi decides to flip a coin until the first tails. He gets four heads before the first tail.

Show that Mario and Luigi will arrive at the same posterior on θ, and calculate this posterior.

answer: We will show that both Mario and Luigi find the posterior pdf for θ is a beta$(9, 6)$ distribution.

Mario's table

hypothesis	data	prior	likelihood	posterior
θ	$x = 4$	beta$(5, 5)$	binomial$(5,\theta)$	???
θ	$x = 4$	$c_1\theta^4(1-\theta)^4$	$\binom{5}{4}\theta^4(1-\theta)$	$c_3\theta^8(1-\theta)^5$

Luigi's table

hypothesis	data	prior	likelihood	posterior
θ	$x = 4$	beta$(5, 5)$	geometric(θ)	???
θ	$x = 4$	$c_1\theta^4(1-\theta)^4$	$\theta^4(1-\theta)$	$c_3\theta^8(1-\theta)^5$

Since both Mario and Luigi's posterior has the form of a beta$(9, 6)$ distribution that's what they both must be. The normalizing factor is the same in both cases because it's determined by requiring the total probability to be 1.

4 Normal begets normal

We now turn to another important example: the normal distribution is its own conjugate prior. In particular, if the likelihood function is normal with known variance, then a normal prior gives a normal posterior. Now both the hypotheses and the data are continuous.

Suppose we have a measurement $x \sim N(\theta, \sigma^2)$ where the variance σ^2 is known. That is, the mean θ is our unknown parameter of interest and we are given that the likelihood comes from a normal distribution with variance σ^2. If we choose a normal prior pdf

$$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$$

then the posterior pdf is also normal: $f(\theta|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ where

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{x}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2} \tag{1}$$

The following form of these formulas is easier to read and shows that μ_{post} is a weighted average between μ_{prior} and the data x.

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{1}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{2}$$

With these formulas in mind, we can express the update via the table:

hypothesis	data	prior	likelihood	posterior
θ	x	$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$	$f(x\|\theta) \sim \mathrm{N}(\theta, \sigma^2)$	$f(\theta\|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$
θ	x	$c_1 \exp\left(\frac{-(\theta-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}\right)$	$c_2 \exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right)$	$c_3 \exp\left(\frac{-(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\right)$

We leave the proof of the general formulas to the problem set. It is an involved algebraic manipulation which is essentially the same as the following numerical example.

Example 2. Suppose we have prior $\theta \sim \mathrm{N}(4, 8)$, and likelihood function likelihood $x \sim \mathrm{N}(\theta, 5)$. Suppose also that we have one measurement $x_1 = 3$. Show the posterior distribution is normal.

answer: We will show this by grinding through the algebra which involves completing the square.

$$\text{prior: } f(\theta) = c_1\, \mathrm{e}^{-(\theta-4)^2/16}; \qquad \text{likelihood: } f(x_1|\theta) = c_2\, \mathrm{e}^{-(x_1-\theta)^2/10} = c_2\, \mathrm{e}^{-(3-\theta)^2/10}$$

We multiply the prior and likelihood to get the posterior:

$$\begin{aligned} f(\theta|x_1) &= c_3\, \mathrm{e}^{-(\theta-4)^2/16}\, \mathrm{e}^{-(3-\theta)^2/10} \\ &= c_3 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right) \end{aligned}$$

We complete the square in the exponent

$$\begin{aligned} -\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} &= -\frac{5(\theta-4)^2 + 8(3-\theta)^2}{80} \\ &= -\frac{13\theta^2 - 88\theta + 152}{80} \\ &= -\frac{\theta^2 - \frac{88}{13}\theta + \frac{152}{13}}{80/13} \\ &= -\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}. \end{aligned}$$

Therefore the posterior is

$$f(\theta|x_1) = c_3\, \mathrm{e}^{-\frac{(\theta-44/13)^2 + 152/13 - (44/13)^2}{80/13}} = c_4\, \mathrm{e}^{-\frac{(\theta-44/13)^2}{80/13}}.$$

This has the form of the pdf for $\mathrm{N}(44/13, 40/13)$. QED

For practice we check this against the formulas (2).

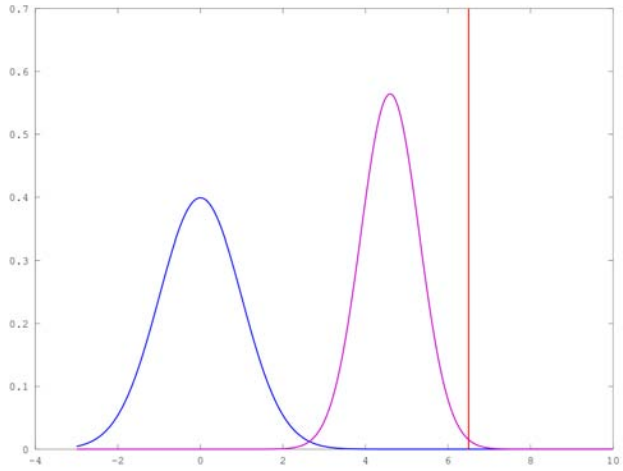
$$\mu_{\text{prior}} = 4, \quad \sigma^2_{\text{prior}} = 8, \quad \sigma^2 = 5 \;\Rightarrow\; a = \frac{1}{8}, \quad b = \frac{1}{5}.$$

Therefore

$$\begin{aligned} \mu_{\text{post}} &= \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{44}{13} = 3.38 \\ \sigma^2_{\text{post}} &= \frac{1}{a+b} = \frac{40}{13} = 3.08. \end{aligned}$$

Example 3. Suppose that we know the data $x \sim \mathrm{N}(\theta, 1)$ and we have prior $\mathrm{N}(0, 1)$. We get one data value $x = 6.5$. Describe the changes to the pdf for θ in updating from the prior to the posterior.

answer: Here is a graph of the prior pdf with the data point marked by a red line.



Prior in blue, posterior in magenta, data in red

The posterior mean will be a weighted average of the prior mean and the data. So the peak of the posterior pdf will be be between the peak of the prior and the read line. A little algebra with the formula shows

$$\sigma^2_{\text{post}} = \frac{1}{1/\sigma^2_{\text{prior}} + 1/\sigma^2} = \sigma^2_{\text{prior}} \cdot \frac{\sigma}{\sigma^2_{\text{prior}} + \sigma^2} < \sigma^2_{\text{prior}}$$

That is the posterior has smaller variance than the prior, i.e. data makes us more certain about where in its range θ lies.

4.1 More than one data point

Example 4. Suppose we have data x_1, x_2, x_3. Use the formulas (1) to update sequentially.

answer: Let's label the prior mean and variance as μ_0 and σ_0^2. The updated means and variances will be μ_i and σ_i^2. In sequence we have

$$\begin{aligned}
&\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}; && \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1}{\sigma^2} \\
&\frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{2}{\sigma^2}; && \frac{\mu_2}{\sigma_2^2} = \frac{\mu_1}{\sigma_1^2} + \frac{x_2}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1+x_2}{\sigma^2} \\
&\frac{1}{\sigma_3^2} = \frac{1}{\sigma_2^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{3}{\sigma^2}; && \frac{\mu_3}{\sigma_3^2} = \frac{\mu_2}{\sigma_2^2} + \frac{x_3}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1+x_2+x_3}{\sigma^2}
\end{aligned}$$

The example generalizes to n data values $x_1, \ldots, x_n$:

Normal-normal update formulas for n data points

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{n\bar{x}}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{n}{\sigma^2}, \qquad \bar{x} = \frac{x_1 + \ldots + x_n}{n}. \tag{3}$$

Again we give the easier to read form, showing μ_{post} is a weighted average of μ_{prior} and the sample average $\bar{x}$:

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{n}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + b\bar{x}}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{4}$$

Interpretation: μ_{post} is a weighted average of μ_{prior} and $\bar{x}$. If the number of data points is large then the weight b is large and $\bar{x}$ will have a strong influence on the posterior. If σ^2_{prior} is small then the weight a is large and μ_{prior} will have a strong influence on the posterior. To summarize:

1. Lots of data has a big influence on the posterior.
2. High certainty (low variance) in the prior has a big influence on the posterior.

The actual posterior is a balance of these two influences.

Choosing priors
Class 15, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Learn that the choice of prior affects the posterior.
2. See that too rigid a prior can make it difficult to learn from the data.
3. See that more data lessons the dependence of the posterior on the prior.
4. Be able to make a reasonable choice of prior, based on prior understanding of the system under consideration.

2 Introduction

Up to now we have always been handed a prior pdf. In this case, statistical inference from data is essentially an application of Bayes' theorem. When the prior is known there is no controversy on how to proceed. The art of statistics starts when the prior is not known with certainty. There are two main schools on how to proceed in this case: Bayesian and frequentist. For now we are following the Bayesian approach. Starting next week we will learn the frequentist approach.

Recall that given data D and a hypothesis H we used Bayes' theorem to write

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}.$$

Bayesian: Bayesians make inferences using the posterior $P(H|D)$, and therefore always need a prior $P(H)$. If a prior is not known with certainty the Bayesian must try to make a reasonable choice. There are many ways to do this and reasonable people might make different choices. In general it is good practice to justify your choices and to explore a range of priors to see if they all point to the same conclusion.

Frequentist: Very briefly, frequentists do not try to create a prior. Instead, they make inferences using the likelihood $P(D|H)$.

We will compare the two approaches in detail once we have more experience with each. For now we simply list two benefits of the Bayesian approach.

1. The posterior probability $P(H|D)$ for the hypothesis given the evidence is usually exactly what we'd like to know. The Bayesian can say something like 'the parameter of interest has probability 0.95 of being between 0.49 and 0.51.'
2. The assumptions that go into choosing the prior can be clearly spelled out.

More good data: It is always the case that more good data allows for stronger conclusions and lessens the influence of the prior. The emphasis should be as much on good data (quality) as on more data (quantity).

3 Example: Dice

Suppose we have a drawer full of dice, each of which has either 4, 6, 8, 12, or 20 sides. This time, we do not know how many of each type are in the drawer. A die is picked at random from the drawer and rolled 5 times. The results in order are 4, 2, 4, 7, and 5.

3.1 Uniform prior

Suppose we have no idea what the distribution of dice in the drawer might be. In this case it's reasonable to use a flat prior. Here is the update table for the posterior probabilities that result from updating after each roll. In order to fit all the columns, we leave out the unnormalized posteriors.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	1/5	1/4	0.370	1/4	0.542	1/4	0.682	0	0.000	0	0.000
H_6	1/5	1/6	0.247	1/6	0.241	1/6	0.202	0	0.000	1/6	0.000
H_8	1/5	1/8	0.185	1/8	0.135	1/8	0.085	1/8	0.818	1/8	0.876
H_{12}	1/5	1/12	0.123	1/12	0.060	1/12	0.025	1/12	0.161	1/12	0.115
H_{20}	1/5	1/20	0.074	1/20	0.022	1/20	0.005	1/20	0.021	1/20	0.009

This should look familiar. Given the data the final posterior is heavily weighted towards hypthesis H_8 that the 8-sided die was picked.

3.2 Other priors

To see how much the above posterior depended on our choice of prior, let's try some other priors. Suppose we have reason to believe that there are ten times as many 20-sided dice in the drawer as there are each of the other types. The table becomes:

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.071	1/4	0.222	1/4	0.453	1/4	0.650	0	0.000	0	0.000
H_6	0.071	1/6	0.148	1/6	0.202	1/6	0.193	0	0.000	1/6	0.000
H_8	0.071	1/8	0.111	1/8	0.113	1/8	0.081	1/8	0.688	1/8	0.810
H_{12}	0.071	1/12	0.074	1/12	0.050	1/12	0.024	1/12	0.136	1/12	0.107
H_{20}	0.714	1/20	0.444	1/20	0.181	1/20	0.052	1/20	0.176	1/20	0.083

Even here the final posterior is heavily weighted to the hypothesis H_8.

What if the 20-sided die is 100 times more likely than each of the others?

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.0096	1/4	0.044	1/4	0.172	1/4	0.443	0	0.000	0	0.000
H_6	0.0096	1/6	0.030	1/6	0.077	1/6	0.131	0	0.000	1/6	0.000
H_8	0.0096	1/8	0.022	1/8	0.043	1/8	0.055	1/8	0.266	1/8	0.464
H_{12}	0.0096	1/12	0.015	1/12	0.019	1/12	0.016	1/12	0.053	1/12	0.061
H_{20}	0.9615	1/20	0.889	1/20	0.689	1/20	0.354	1/20	0.681	1/20	0.475

With such a strong prior belief in the 20-sided die, the final posterior gives a lot of weight to the theory that the data arose from a 20-sided die, even though it extremely unlikely the

20-sided die would produce a maximum of 7 in 5 roles. The posterior now gives roughly even odds that an 8-sided die versus a 20-sided die was picked.

3.3 Rigid priors

Mild cognitive dissonance. Too rigid a prior belief can overwhelm any amount of data. Suppose I've got it in my head that the die has to be 20-sided. So I set my prior to $P(H_{20}) = 1$ with the other 4 hypotheses having probability 0. Look what happens in the update table.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0	1/4	0	1/4	0	1/4	0	0	0	0	0
H_6	0	1/6	0	1/6	0	1/6	0	0	0	1/6	0
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	1/8	0
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	1/12	0
H_{20}	1	1/20	1	1/20	1	1/20	1	1/20	1	1/20	1

No matter what the data, a hypothesis with prior probability 0 will have posterior probability 0. In this case I'll never get away from the hypothesis H_{20}, although I might experience some mild cognitive dissonance.

Severe cognitive dissonance. Rigid priors can also lead to absurdities. Suppose I now have it in my head that the die must be 4-sided. So I set $P(H_4) = 1$ and the other prior probabilities to 0. With the given data on the fourth roll I reach an impasse. A roll of 7 can't possibly come from a 4-sided die. Yet this is the only hypothesis I'll allow. My unnormalized posterior is a column of all zeros which cannot be normalized.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	unnorm. post_4	post_4
H_4	1	1/4	1	1/4	1	1/4	1	0	0	???
H_6	0	1/6	0	1/6	0	1/6	0	0	0	???
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	???
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	???
H_{20}	0	1/20	0	1/20	0	1/20	0	1/20	0	???

I must adjust my belief about what is possible or, more likely, I'll suspect you of accidently or deliberately messing up the data.

4 Example: Malaria

Here is a real example adapted from *Statistics, A Bayesian Perspective* by Donald Berry:

By the 1950's scientists had begun to formulate the hypothesis that carriers of the sickle-cell gene were more resistant to malaria than noncarriers. There was a fair amount of circumstantial evidence for this hypothesis. It also helped explain the persistance of an otherwise deleterious gene in the population. In one experiment scientists injected 30 African volunteers with malaria. Fifteen of the volunteers carried one copy of the sickle-cell gene and the other 15 were noncarriers. Fourteen out of 15 noncarriers developed malaria while only 2

out of 15 carriers did. Does this small sample support the hypothesis that the sickle-cell gene protects against malaria?

Let S represent a carrier of the sickle-cell gene and N represent a non-carrier. Let $D+$ indicate developing malaria and $D-$ indicate not developing malaria. The data can be put in a table.

	$D+$	$D-$	
S	2	13	15
N	14	1	15
	16	14	30

Before analysing the data we should say a few words about the experiment and experimental design. First, it is clearly unethical: to gain some information they infected 16 people with malaria. We also need to worry about bias. How did they choose the test subjects. Is it possible the noncarriers were weaker and thus more susceptible to malaria than the carriers? Berry points out that it is reasonable to assume that an injection is similar to a mosquito bite, but it is not guaranteed. This last point means that if the experiment shows a relation between sickle-cell and protection against injected malaria, we need to consider the hypothesis that the protection from mosquito transmitted malaria is weaker or non-existent. Finally, we will frame our hypothesis as 'sickle-cell protects against malaria', but really all we can hope to say from a study like this is that 'sickle-cell is correlated with protection against malaria'.

Model. For our model let θ_S be the probability that an injected carrier S develops malaria and likewise let θ_N be the probabilitiy that an injected noncarrier N develops malaria. We assume independence between all the experimental subjects. With this model, the likelihood is a function of both θ_S and θ_N:

$$P(\text{data}|\theta_S, \theta_N) = c\,\theta_S^2(1-\theta_S)^{13}\theta_N^{14}(1-\theta_N).$$

As usual we leave the constant factor c as a letter. (It is a product of two binomial coefficients: $c = \binom{15}{2}\binom{15}{14}$.)

Hypotheses. Each hypothesis consists of a pair (θ_N, θ_S). To keep things simple we will only consider a finite number of values for these probabilities. We could easily consider many more values or even a continuous range of hypotheses. Assume θ_S and θ_N are each one of 0, 0.2, 0.4, 0.6, 0.8, 1. This leads to two-dimensional tables.

First is a table of hypotheses. The color coding indicates the following:
1. Light orange squares along the diagonal are where $\theta_S = \theta_N$, i.e. sickle-cell makes no difference one way or the other.
2. Pink and red squares above the diagonal are where $\theta_N > \theta_S$, i.e. sickle-cell provides some protection against malaria.
3. In the red squares $\theta_N - \theta_S \geq 0.6$, i.e. sickle-cell provides a lot of protection.
4. White squares below diagonal are where $\theta_S > \theta_N$, i.e. sickle-cell actually increases the probability of developing malaria.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	(0,1)	(.2,1)	(.4,1)	(.6,1)	(.8,1)	(1,1)
0.8	(0,.8)	(.2,.8)	(.4,.8)	(.6,.8)	(.8,.8)	(1,.8)
0.6	(0,.6)	(.2,.6)	(.4,.6)	(.6,.6)	(.8,.6)	(1,.6)
0.4	(0,.4)	(.2,.4)	(.4,.4)	(.6,.4)	(.8,.4)	(1,.4)
0.2	(0,.2)	(.2,.2)	(.4,.2)	(.6,.2)	(.8,.2)	(1,.2)
0	(0,0)	(.2,0)	(.4,0)	(.6,0)	(.8,0)	(1,0)

Hypotheses on level of protection due to S:
red = strong; pink = some; orange = none; white = negative.

Next is the table of likelihoods. (Actually we've taken advantage of our indifference to scale and scaled all the likelihoods by $100000/c$ to make the table more presentable.) Notice that, to the precision of the table, many of the likelihoods are 0. The color coding is the same as in the hypothesis table. We've highlighted the biggest likelihoods with a blue border.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	1.93428	0.18381	0.00213	0.00000	0.00000
0.6	0.00000	0.06893	0.00655	0.00008	0.00000	0.00000
0.4	0.00000	0.00035	0.00003	0.00000	0.00000	0.00000
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Likelihoods $p(\text{data}|\theta_S, \theta_N)$ scaled by $100000/c$

4.1 Flat prior

Suppose we have no opinion whatsoever on whether and to what degree sickle-cell protects against malaria. In this case it is reasonable to use a flat prior. Since there are 36 hypotheses each one gets a prior probability of 1/36. This is given in the table below. Remember each square in the table represents one hypothesis. Because it is a probability table we include the marginal pmf.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.8	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(\theta_S)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Flat prior $p(\theta_S, \theta_N)$: every hypothesis (square) has equal probability

To compute the posterior we simply multiply the likelihood table by the prior table and

normalize. Normalization means making sure the entire table sums to 1.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88075	0.08370	0.00097	0.00000	0.00000	0.96542
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00002	0.00000	0.00000	0.00000	0.00018
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91230	0.08670	0.00100	0.00000	0.00000	1.00000

Posterior to flat prior: $p(\theta_S, \theta_N \mid \text{data})$

To decide whether S confers protection against malaria, we compute the posterior probabilities of 'some protection' and of 'strong protection'. These are computed by summing the corresponding squares in the posterior table.

Some protection: $P(\theta_N > \theta_S) = \text{sum of pink and red} = .99995$

Strong protection: $P(\theta_N - \theta_S > .6) = \text{sum of red} = .88075$

Working from the flat prior, it is effectively certain that sickle-cell provides some protection and very probable that it provides strong protection.

4.2 Informed prior

The experiment was not run without prior information. There was a lot of circumstantial evidence that the sickle-cell gene offered some protection against malaria. For example it was reported that a greater percentage of carriers survived to adulthood.

Here's one way to build an informed prior. We'll reserve a reasonable amount of probability for the hypotheses that S gives no protection. Let's say 24% split evenly among the 6 (orange) cells where $\theta_N = \theta_S$. We know we shouldn't set any prior probabilities to 0, so let's spread 6% of the probability evenly among the 15 white cells below the diagonal. That leaves 70% of the probability for the 15 pink and red squares above the diagonal.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	0.04667	0.04667	0.04667	0.04667	0.04667	0.04000	0.27333
0.8	0.04667	0.04667	0.04667	0.04667	0.04000	0.00400	0.23067
0.6	0.04667	0.04667	0.04667	0.04000	0.00400	0.00400	0.18800
0.4	0.04667	0.04667	0.04000	0.00400	0.00400	0.00400	0.14533
0.2	0.04667	0.04000	0.00400	0.00400	0.00400	0.00400	0.10267
0	0.04000	0.00400	0.00400	0.00400	0.00400	0.00400	0.06000
$p(\theta_S)$	0.27333	0.23067	0.18800	0.14533	0.10267	0.06000	1.0

Informed prior $p(\theta_S, \theta_N)$: makes use of prior information that sickle-cell is protective.

We then compute the posterior pmf.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88076	0.08370	0.00097	0.00000	0.00000	0.96543
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00001	0.00000	0.00000	0.00000	0.00017
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91231	0.08669	0.00100	0.00000	0.00000	1.00000

Posterior to informed prior: $p(\theta_S, \theta_N|\text{data})$

We again compute the posterior probabilities of 'some protection' and 'strong protection'.

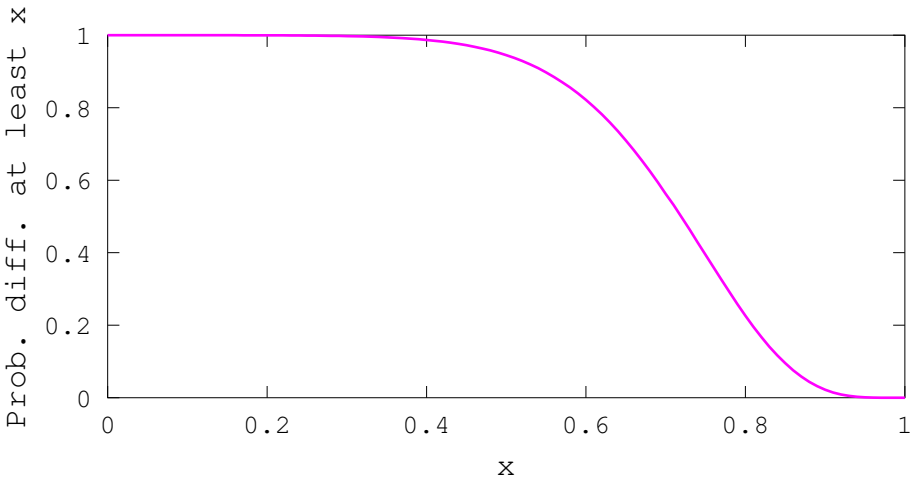
Some protection: $P(\theta_N > \theta_S) =$ sum of pink and red $= .99996$

Strong protection: $P(\theta_N - \theta_S > .6) =$ sum of red $= .88076$

Note that the informed posterior is nearly identical to the flat posterior.

4.3 PDALX

The following plot is based on the flat prior. For each x, it gives the probability that $\theta_N - \theta_S \geq x$. To make it smooth we used many more hypotheses.



Probability the difference $\theta_N - \theta_S$ is at least x (PDALX).

Notice that it is virtually certain that the difference is at least .4.

Probability intervals
Class 16, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to find probability intervals given a pmf or pdf.
2. Understand how probability intervals summarize belief in Bayesian updating.
3. Be able to use subjective probability intervals to construct reasonable priors.
4. Be able to construct subjective probability intervals by systematically estimating quantiles.

2 Probability intervals

Suppose we have a pmf $p(\theta)$ or pdf $f(\theta)$ describing our belief about the value of an unknown parameter of interest θ.

Definition: A p-probability interval for θ is an interval $[a, b]$ with $P(a \le \theta \le b) = p$.

Notes.
1. In the discrete case with pmf $p(\theta)$, this means $\sum_{a \le \theta_i \le b} p(\theta_i) = p$.

2. In the continuous case with pdf $f(\theta)$, this means $\int_a^b f(\theta)\, d\theta = p$.

3. We may say 90%-probability interval to mean 0.9-probability interval. Probability intervals are also called credible intervals to contrast them with confidence intervals, which we'll introduce in the frequentist unit.

Example 1. Between the 0.05 and 0.55 quantiles is a 0.5 probability interval. There are many 50% probability intervals, e.g. the interval from the 0.25 to the 0.75 quantiles.

In particular, notice that the p-probability interval for θ is not unique.

Q-notation. We can phrase probability intervals in terms of **quantiles**. Recall that the s-quantile for θ is the value q_s with $P(\theta \le q_s) = s$. So for $s \le t$, the amount of probability between the s-quantile and the t-quantile is just $t - s$. In these terms, a p-probability interval is any interval $[q_s, q_t]$ with $t - s = p$.

Example 2. We have 0.5 probability intervals $[q_{0.25}, q_{0.75}]$ and $[q_{0.05}, q_{0.55}]$.

Symmetric probability intervals.
The interval $[q_{0.25}, q_{0.75}]$ is symmetric because the amount of probability remaining on either side of the interval is the same, namely 0.25. If the pdf is not too skewed, the symmetric interval is usually a good default choice.

More notes.
1. Different p-probability intervals for θ may have different widths. We can make the width

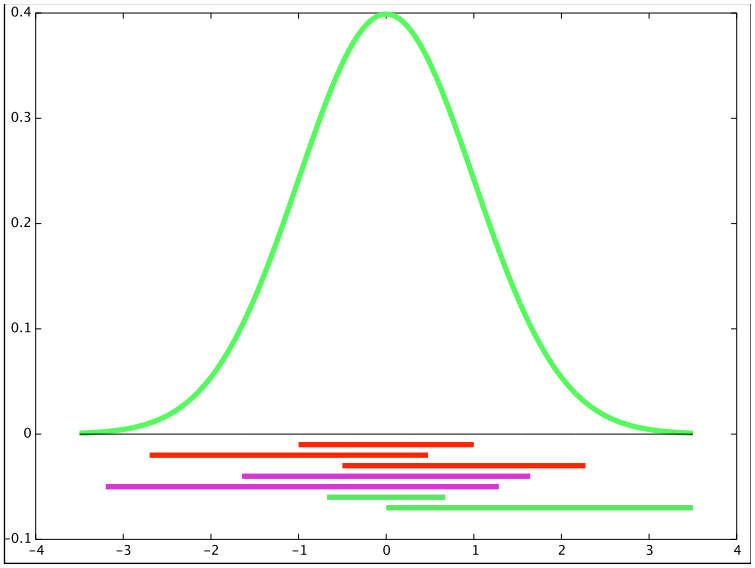
smaller by centering the interval under the highest part of the pdf. Such an interval is usually a good choice since it contains the most likely values. See the examples below for normal and beta distributions.

2. Since the width can vary for fixed p, a larger p does not always mean a larger width. Here's what is true: if a p_1-probability interval is fully contained in a p_2-probability interval, then p_1 is bigger than p_2.

Probability intervals for a normal distribution. The figure shows a number of probability intervals for the standard normal.

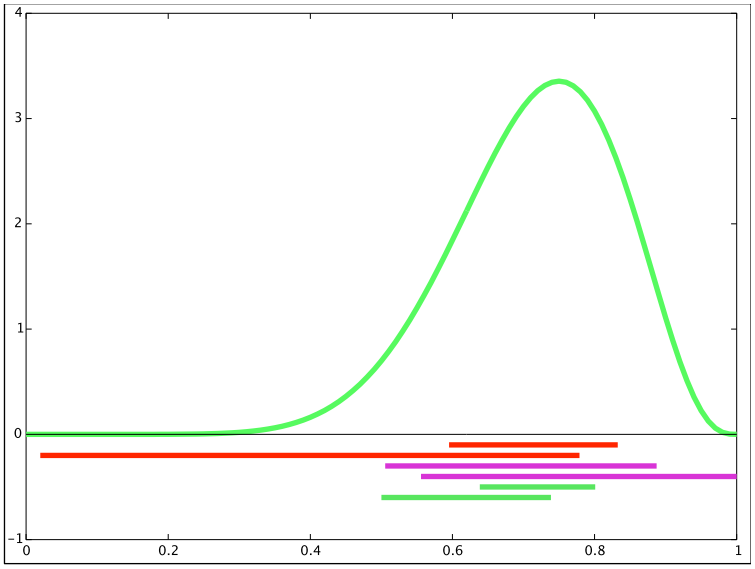
1. All of the red bars span a 0.68-probability interval. Notice that the smallest red bar runs between -1 and 1. This runs from the 16th percential to the 84th percentile so it is a symmetric interval.

2. All the magenta bars span a 0.9-probability interval. They are longer than the red bars because they include more probability. Note again that the shortest magenta bar is symmetric.



red = 0.68, magenta = 0.9, green = 0.5

Probabilitiy intervals for a beta distribution. The following figure shows probability intervals for a beta distribution. Notice how the two red bars have very different lengths yet cover the same probability $p = 0.68$.



red = 0.68, magenta = 0.9, green = 0.5

3 Uses of probability intervals

3.1 Summarizing and communicating your beliefs

Probability intervals are an intuitive and effective way to summarize and communicate your beliefs. It's hard to describe an entire function $f(\theta)$ to a friend in words. If the function isn't from a parameterized family then it's especially hard. Even with a beta distribution, it's easier to interpret "I think θ is between 0.45 and 0.65 with 50% probability" than "I think θ follows a beta(8,6) distribution". An exception to this rule of communication might be the normal distribution, but only if the recipient is also comfortable with standard deviation. Of course, what we gain in clarity we lose in precision, since the function contains more information than the probability interval.

Probability intervals also play well with Bayesian updating. If we update from the prior $f(\theta)$ to the posterior $f(\theta|x)$, then the p-probability interval for the posterior will tend to be shorter than than the p-probability interval for the prior. In this sense, the data has made us more certain. See for example the election example below.

4 Constructing a prior using subjective probability intervals

Probability intervals are also useful when we do not have a pmf or pdf at hand. In this case, subjective probability intervals give us a method for constructing a reasonable prior for θ "from scratch". The thought process is to ask yourself a series of questions, e.g., 'what is my expected value for θ?'; 'my 0.5-probability interval?'; 'my 0.9-probability interval?' Then build a prior that is consistent with these intervals.

4.1 Estimating the intervals directly

Example 3. Building priors
In 2013 there was a special election for a congressional seat in a district in South Carolina. The election pitted Republican Mark Sanford against Democrat Elizabeth Colbert Busch. Let θ be the fraction of the population who favored Busch. Our goal in this example is to build a subjective prior for θ. We'll use the following prior evidence.

- Sanford is a former S. Carolina Congressman and Governor
- He had famously resigned after having an affair in Argentina while he claimed to be hiking the Appalachian trail.
- In 2013 Sanford won the Republican primary over 15 primary opponents.
- In the district in the 2012 presidential election the Republican Romney beat the Democrat Obama 58% to 40%.
- The Colbert bump: Elizabeth Colbert Busch is the sister of well-known comedian Stephen Colbert.

Our strategy will be to use our intuition to construct some probability intervals and then find a beta distribution that approximately matches these intervals. This is subjective so someone else might give a different answer.

Step 1. Use the evidence to construct 0.5 and 0.9 probability intervals for θ.

We'll start by thinking about the 90% interval. The single strongest prior evidence is the 58% to 40% of Romney over Obama. Given the negatives for Sanford we don't expect he'll win much more than 58% of the vote. So we'll put the top of the 0.9 interval at 0.65. With all of Sanford's negatives he could lose big. So we'll put the bottom at 0.3.

$$0.9 \text{ interval:} \quad [0.3,\ 0.65]$$

For the 0.5 interval we'll pull these endpoints in. It really seems unlikely Sanford will get more votes than Romney, so we can leave 0.25 probability that he'll get above 57%. The lower limit seems harder to predict. So we'll leave 0.25 probability that he'll get under 42%.

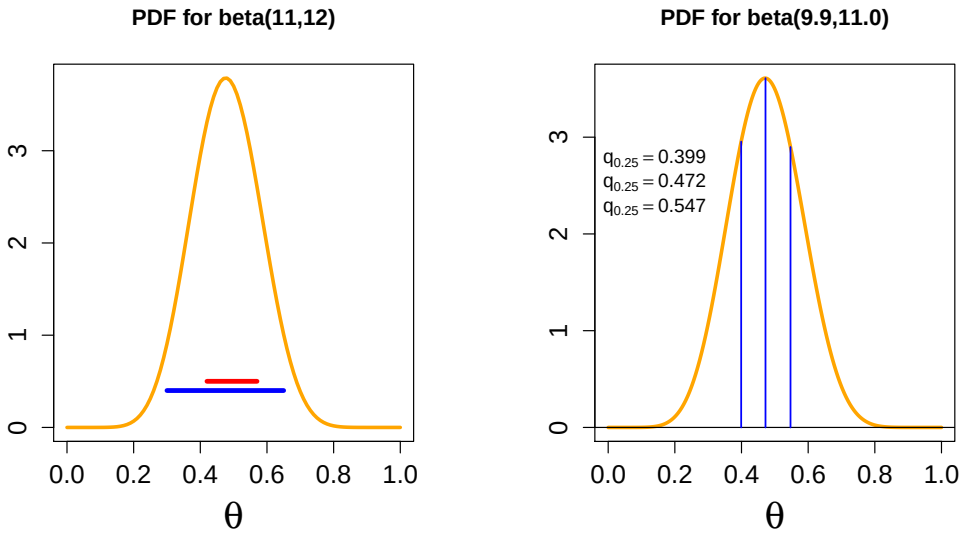
$$0.5 \text{ interval:} \quad [0.42,\ 0.57]$$

Step 2. Use our 0.5 and 0.9 probability intervals to pick a beta distribution that approximats these intervals. We used the R function `pbeta` and a little trial and error to choose beta(11,12). Here is our R code.

```
a = 11
b = 12
pbeta(0.65, a, b) - pbeta(0.3, a, b)
pbeta(0.57, a, b) - pbeta(0.42, a, b)
```

This computed $P([0.3, 0.65]) = 0.91$ and $P([0.42, 0.57]) = 0.52$. So our intervals are actually 0.91 and 0.52-probability intervals. This is pretty close to what we wanted!

At right is a graph of the density of beta(11,12). The red line shows our interval [0.42, 0.57] and the blue line shows our interval [0.3, 0.65].



beta(11,12) found using probability intervals and beta(9.9,11.0) found using quantiles

4.2 Constructing a prior by estimating quantiles

The method in Example 3 gives a good feel for building priors from probability intervals. Here we illustrate a slightly different way of building a prior by estimating quantiles. The basic strategy is to first estimate the median, then divide and conquer to estimate the first and third quantiles. Finally you choose a prior distribution that fits these estimates.

Example 4. Redo the Sanford vs. Colbert-Busch election example using quantiles.

answer: We start by estimating the median. Just as before the single strongest evidence is the 58% to 40% victory of Romney over Obama. However, given Sanford's negatives and Busch's Colbert bump we'll estimate the median at 0.47.

In a district that went 58 to 40 for the Republican Romney it's hard to imagine Sanford's vote going a lot below 40%. So we'll estimate Sanford 25th percentile as 0.40. Likewise, given his negatives it's hard to imagine him going above 58%, so we'll estimate his 75th percentile as 0.55.

We used R to search through values of a and b for the beta distribution that matches these quartiles the best. Since the beta distribution does not require a and b to be integers we looked for the best fit to 1 decimal place. We found beta(9.9, 11.0). Above is a plot of beta(9.9,11.0) with its actual quartiles shown. These match the desired quartiles pretty well.

Historic note. In the election Sanford won 54% of the vote and Busch won 45.2%. (Source: http://elections.huffingtonpost.com/2013/mark-sanford-vs-elizabeth-colbert-busch-sc1