Chapter 1: Foundations

Contents

1 How random is the flip of a coin? 2

1.1 "It's a toss-up" 2

1.2 Probability and Statistics are two sides of the same coin 2

1.3 The Bayesian approach 3

2 Beta-Bernoulli model 4

2.1 Bernoulli distribution 4

2.2 Beta distribution 5

2.3 The posterior 6

3 The cast of characters 8

3.1 Marginal likelihood and posterior predictive 8

3.2 Example: Beta-Bernoulli 8

4 Decision theory 9

4.1 The basics of Bayesian decision theory 10

4.2 Real-world decision problems 11

4.3 Example: Resource allocation for disease prevention/treatment 12

4.4 Frequentist risk and Integrated risk 13

5 Exercises 15

Jeffrey W. Miller (2015). *Lecture Notes on Bayesian Statistics.* Duke University, Durham, NC.

1 How random is the flip of a coin?

1.1 “It’s a toss-up”

- It is so generally assumed that a coin toss comes up heads half of the time, that it has even become a standard metaphor for two events with equal probability.

- But think about it—is it really 50-50? Suppose we always flip a coin starting with heads up. Could the outcome actually be biased toward either heads or tails?

- Assume the coin is physically symmetric. Since an “in-flight” coin is a relatively simple physical system, the outcome should be essentially determined by the initial conditions at the beginning of it’s trajectory. So, any randomness comes from the flipper, not the flipping.

- Experiment: Flip a coin $n = 10$ times, starting with heads up each time. (flip, flip, flip, . . .) Do we know anything more now than when we started? We got some data, so we should know more now. But probably we need more data! How much more? How can we quantify our uncertainty about the answer?

- (And now comes the surprise: Diaconis et al. (2007) argue that, in fact, the outcome is slightly biased due to precession, and will come up the same way it started about 51% of the time! This is based on physical, rather than statistical, evidence.)

1.2 Probability and Statistics are two sides of the same coin

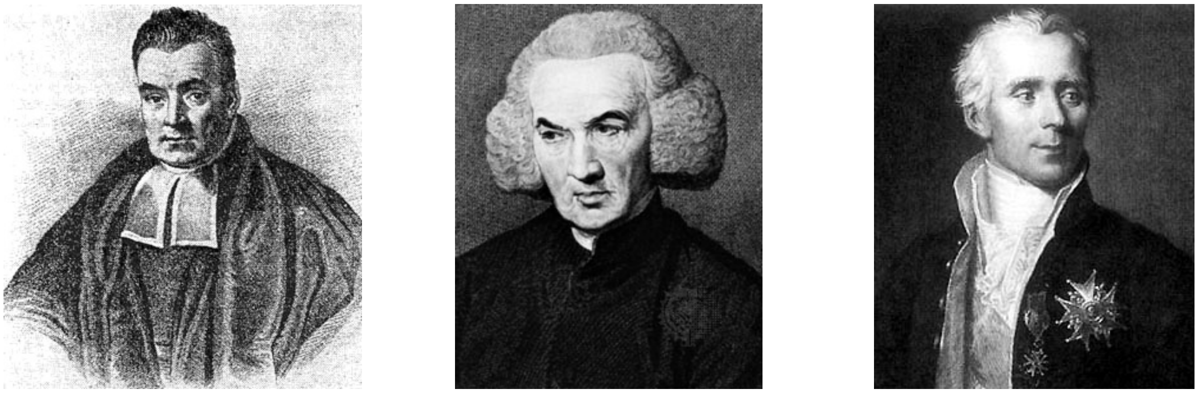
- Let $X_1, \ldots, X_n$ be the outcomes of n coin flips, and suppose they are i.i.d. (independent and identically distributed), with the probability of heads equal to θ.

- This defines a probabilistic model, for which—if we knew θ—we could prove all kinds of things about the distribution of $X_{1:n} = (X_1, \ldots, X_n)$. This is Probability.

- Statistics, meanwhile, goes the other direction—trying to obtain information about θ from $X_{1:n}$.

$$\begin{array}{ll} \text{Probability:} & \theta \longrightarrow X_{1:n} \\ \text{Statistics:} & \theta \longleftarrow X_{1:n} \end{array}$$

- To see if the outcome is biased, based on the data $X_1, \ldots, X_n$, perhaps the first thing that comes to mind is to simply look at the proportion of heads, and see if it’s close to 1/2. But on reflection, there are some issues with this:
  - How close is “close”?
  - How would we quantify our uncertainty about the correct answer?
  - If n is very small, say 2 or 3, there is a good chance that the flips will all come up the same (all heads or all tails), in which case the proportion of heads would be 1 or 0. But from experience, we know θ is unlikely to be close to 1 or 0. Would it be better to take such prior knowledge into account?



$\mathbb{P}(\text{this} = \text{Bayes} \mid \text{data}) < 1$ Richard Price Pierre-Simon Laplace

Figure 1: Founders of Bayesian statistics.

1.3 The Bayesian approach

- Thomas Bayes (1701?–1761) was an ordained minister who was also a talented mathematician and a Fellow of the Royal Society. Bayes came up with an ingenious solution to this problem, but died before publishing it. Fortunately, his friend Richard Price carried his work further and published it in 1764. Apparently independently, Laplace rediscovered essentially the same idea in 1774, and developed it much further. (See Figure 1.)

- The idea is to assume a ***prior*** probability distribution for θ—that is, a distribution representing the plausibility of each possible value of θ before the data is observed. Then, to make inferences about θ, one simply considers the conditional distribution of θ given the observed data. This is referred to as the ***posterior*** distribution, since it represents the plausibility of each possible value of θ after seeing the data.

- Mathematically, this is expressed via ***Bayes' theorem***,

  $$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta), \tag{1.1}$$

  where x is the observed data (for example, $x = x_{1:n}$). In words, we say "the posterior is proportional to the likelihood times the prior". Bayes' theorem is essentially just the definition of conditional probability

  $$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

  extended to conditional densities. (From the modern perspective, Bayes' theorem is a trivial consequence of the definition of a conditional density—however, when Bayes wrote his paper, the idea of a conditional probability density did not yet exist!)

- More generally, the Bayesian approach—in a nutshell—is to assume a prior distribution on any unknowns, and then just follow the rules of probability to answer any questions of interest. This provides a coherent framework for making inferences about unknown parameters θ as well as any future data or missing data, and for making rational decisions based on such inferences.

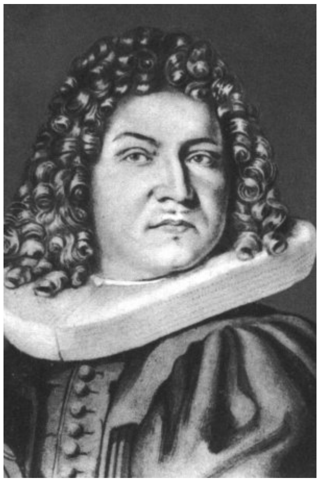


Figure 2: Jacob Bernoulli (not in a bad mood, everyone is just annoying).

2 Beta-Bernoulli model

We now formally explore a Bayesian approach to the coin flipping problem.

2.1 Bernoulli distribution

- The Bernoulli distribution models binary outcomes, i.e., taking two possible values. The convention is to use 0 and 1.
- It is named for Jacob Bernoulli (1655–1705), who is known for various inconsequential trivialities such as developing the foundations of probability theory (including the law of large numbers), combinatorics, differential equations, and the calculus of variations. Oh, and he discovered the constant e.
- The Bernoulli distribution shows up everywhere, due to the ubiquity of binary outcomes (for instance, in computer vision, neuroscience, demographics and polling, public health and epidemiology, etc. etc. etc.).
- We write $X \sim \text{Bernoulli}(\theta)$ to mean that

  $$\mathbb{P}(X = x \mid \theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

  and is 0 otherwise. In other words, the p.m.f. (probability mass function) is

  $$p(x|\theta) = \mathbb{P}(X = x \mid \theta) = \theta^x (1-\theta)^{1-x} \mathbb{1}(x \in \{0,1\}).$$

- The mean (or expectation) is $\mathbb{E}X = \sum_{x \in \{0,1\}} x p(x|\theta) = \theta$.
- Notation: $\mathbb{P}$ denotes "the probability of", and $\mathbb{E}$ denotes the expectation. The ***indicator function***, $\mathbb{1}(E)$, equals 1 when E is true and is 0 otherwise. The symbol $\in$ means "belongs to the set".

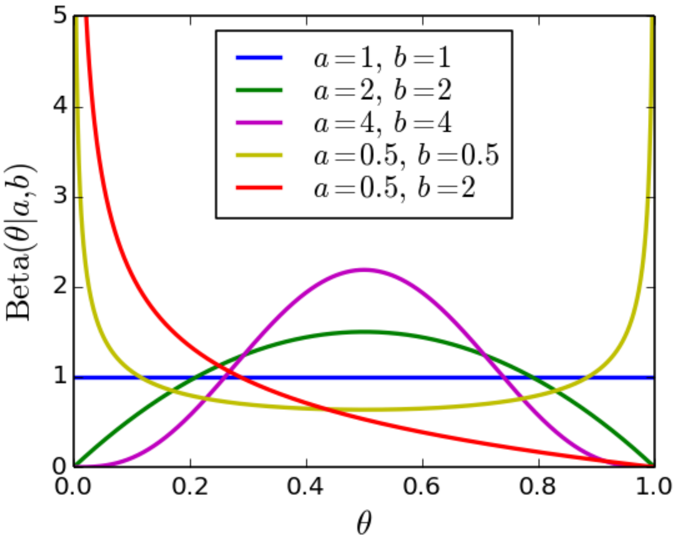


Figure 3: Some Beta p.d.f.s.

- If $X_1,\ldots,X_n \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(\theta)$ then for $x_1,\ldots,x_n \in \{0,1\}$,

$$\begin{aligned} p(x_{1:n}|\theta) &= \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n \mid \theta) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta) \\ &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}. \end{aligned} \tag{2.1}$$

- Viewed as a function of θ, $p(x_{1:n}|\theta)$ is called the ***likelihood function***. It is sometimes denoted $L(\theta; x_{1:n})$ to emphasize this. Viewed as a distribution on $x_{1:n}$, we will refer to this as the ***generator*** or ***generating distribution*** (sometimes it is referred to as the “sampling distribution”, but this becomes ambiguous when one is also sampling from the posterior).

2.2 Beta distribution

- Bayes used a uniform prior on θ, which is a special case of the beta distribution.
- Given $a, b > 0$, we write $\boldsymbol{\theta} \sim \mathrm{Beta}(a,b)$ to mean $\boldsymbol{\theta}$ has p.d.f. (probability density function)

$$p(\theta) = \mathrm{Beta}(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}(0<\theta<1), \tag{2.2}$$

A posteriori

i.e., $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ on the interval from 0 to 1. Here, $B(a,b)$ is Euler's beta function.

- The mean is $\mathbb{E}\,\boldsymbol{\theta} = \int \theta\, p(\theta) d\theta = a/(a+b)$.

Notation

- $f(x) \propto g(x)$ ("f is proportional to g") means there is a constant c such that $f(x) = cg(x)$ for all x. For functions of multiple variables, say x and y, we write $\underset{x}{\propto}$ to indicate proportionality with respect to x only. This simple device is surprisingly useful for deriving posterior distributions.

- Usually, we use capital letters to denote random variables (e.g., X) and lowercase for particular values (e.g., x). However, in the case of theta, we will use bold font to denote the random variable $\boldsymbol{\theta}$, and unbold for particular values θ.

- We will usually use p for all p.d.f.s and p.m.f.s, following the usual convention that the symbol used (e.g., the θ in the expression $p(\theta)$) indicates which random variable we are talking about.

2.3 The posterior

- Using Bayes' theorem (Equation 1.1), and plugging in the likelihood (Equation 2.1) and the prior (Equation 2.2), the posterior is

$$\begin{aligned} p(\theta|x_{1:n}) &\propto p(x_{1:n}|\theta)p(\theta) \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}(0<\theta<1) \\ &\propto \theta^{a+\sum x_i-1}(1-\theta)^{b+n-\sum x_i-1}\mathbb{1}(0<\theta<1) \\ &\propto \text{Beta}\left(\theta \mid a+\sum x_i,\, b+n-\sum x_i\right). \end{aligned} \tag{2.3}$$

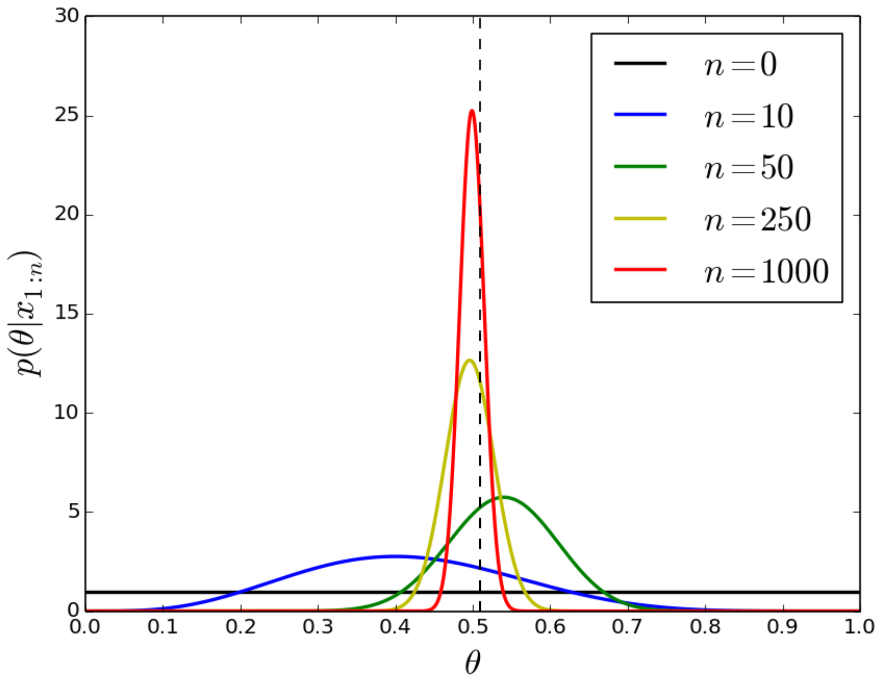


Figure 4: Posterior densities. The dotted line shows the true value of theta.

- So, the posterior has the same form (a Beta distribution) as the prior! When this happens, we say that the prior is ***conjugate*** (more on this later).
- Since the posterior has such a nice form, it is easy to work with—e.g., for computing certain integrals with respect to the posterior, sampling from the posterior, and computing the posterior p.d.f. and its derivatives.

Example

- Suppose we choose $a = 1$ and $b = 1$, so that the prior is uniform. As a simulation to see how the posterior behaves, let's generate data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_0)$ with $\theta_0 = 0.51$.
- Figure 4 shows the posterior p.d.f. for increasing amounts of data. (Note that this will be different each time the experiment is run, because the samples will be different.)

3 The cast of characters

Here's a list of the mathematical objects we will most frequently encounter. So far, we've seen the likelihood, prior, and posterior. In the rest of this chapter, we will get acquainted with the rest of them. Here, we denote the observed data by x, noting that this may consist of many data points, e.g., $x = x_{1:n} = (x_1, \ldots, x_n)$.

generator / likelihood	$p(x\|\theta)$
prior	$p(\theta)$
posterior	$p(\theta\|x)$
marginal likelihood	$p(x)$
posterior predictive	$p(x_{n+1}\|x_{1:n})$
loss function	$\ell(s, a)$
posterior expected loss	$\rho(a, x)$
risk / frequentist risk	$R(\theta, \delta)$
integrated risk	$r(\delta)$

3.1 Marginal likelihood and posterior predictive

The ***marginal likelihood*** is

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

i.e., it is the marginal p.d.f./p.m.f. of the observed data, obtained by integrating θ out of the joint density $p(x, \theta) = p(x|\theta)p(\theta)$. When θ is a vector, this will be a multi-dimensional integral.

When the data is a sequence $x = (x_1, \ldots, x_n)$, the ***posterior predictive distribution*** is the distribution of X_{n+1} given $X_{1:n} = x_{1:n}$. When $X_1, \ldots, X_n, X_{n+1}$ are independent given $\boldsymbol{\theta} = \theta$, the posterior predictive p.d.f./p.m.f. is given by

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}, \theta|x_{1:n})\,d\theta \\ &= \int p(x_{n+1}|\theta, x_{1:n})p(\theta|x_{1:n})\,d\theta \\ &= \int p(x_{n+1}|\theta)p(\theta|x_{1:n})\,d\theta. \end{aligned}$$

3.2 Example: Beta-Bernoulli

If $\boldsymbol{\theta} \sim \text{Beta}(a, b)$ and $X_1, \ldots, X_n \mid \boldsymbol{\theta} = \theta$ are i.i.d. Bernoulli(θ) (as in Section 2), then the marginal likelihood is

$$\begin{aligned} p(x_{1:n}) &= \int p(x_{1:n}|\theta)p(\theta)\,d\theta \\ &= \int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta \\ &= \frac{B\big(a + \sum x_i,\, b + n - \sum x_i\big)}{B(a,b)}, \end{aligned}$$

Blaise Pascal

Abraham Wald

Figure 5: Historical figures in decision theory.

by the integral definition of the Beta function. Letting $a_n = a + \sum x_i$ and $b_n = b + n - \sum x_i$ for brevity, and using the fact (from Equation 2.3) that $p(\theta|x_{1:n}) = \text{Beta}(\theta|a_n, b_n)$,

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1 \mid x_{1:n}) &= \int \mathbb{P}(X_{n+1} = 1 \mid \theta) p(\theta|x_{1:n}) d\theta \\ &= \int \theta \ \text{Beta}(\theta|a_n, b_n) = \frac{a_n}{a_n + b_n},\end{aligned}$$

hence, the posterior predictive p.m.f. is

$$p(x_{n+1}|x_{1:n}) = \frac{a_n^{x_{n+1}} b_n^{1-x_{n+1}}}{a_n + b_n} \mathbb{1}(x_{n+1} \in \{0, 1\}).$$

4 Decision theory

In decision theory, we start with the end in mind—how are we actually going to use our inferences and what consequences will this have? The basic goal is to minimize loss (or equivalently, to maximize utility/gain). While there are multiple ways of making this precise, here we consider the standard Bayesian approach, which is to minimize posterior expected loss.

A famous early example of decision-theoretic reasoning is Pascal's Wager, in which Blaise Pascal (1623–1662) suggested the following argument for believing in God: If God exists then one will reap either an infinite gain or infinite loss (eternity in heaven or hell), depending on whether one believes or not—meanwhile, if he does not exist, the gain or loss is finite. Thus, he reasoned, no matter how small the probability that God exists, the rational decision is to believe. Pascal's loss function can be represented by the following matrix, in which 1 indicates existence, 0 indicates non-existence, and α, β are finite values:

$$\begin{array}{cc|c|c|}
 & & \multicolumn{2}{c}{\text{Belief}} \\
 & & \multicolumn{1}{c}{0} & \multicolumn{1}{c}{1} \\
\cline{3-4}
\text{Truth} & 0 & \alpha & \beta \\
\cline{3-4}
 & 1 & \infty & -\infty \\
\cline{3-4}
\end{array}$$

In statistics, loss functions were used in a limited way during the 1700s and 1800s (most notably Laplace's absolute error and Gauss' quadratic error), but the real developments would have to wait until the 1900s.

The father of statistical decision theory was Abraham Wald (1902–1950). Wald was born in Austria-Hungary, and moved to the United States after the annexation of Austria into Germany in 1938. In 1939, he published a groundbreaking paper establishing the foundations of modern statistical decision theory. Wald also developed sequential analysis, made significant contributions to econometrics and geometry, and provided an important statistical analysis of aircraft vulnerability during World War II.

4.1 The basics of Bayesian decision theory

- The general setup is that there is some unknown state S (a.k.a. the state of nature), we receive an observation x, we take an action a, and we incur a real-valued loss $\ell(S, a)$.

S	state (unknown)
x	observation (known)
a	action
$\ell(s, a)$	loss

- In the Bayesian approach, S is a random variable, the distribution of x depends on S, and the optimal decision is to choose an action a that minimizes the ***posterior expected loss***,
$$\rho(a, x) = \mathbb{E}(\ell(S, a)|x).$$
In other words, $\rho(a, x) = \sum_s \ell(s, a)p(s|x)$ if S is a discrete random variable, while if S is continuous then the sum is replaced by an integral.

- A ***decision procedure*** δ is a systematic way of choosing actions a based on observations x. Typically, this is a deterministic function $a = \delta(x)$ (but sometimes introducing some randomness into a can be useful).

- A ***Bayes procedure*** is a decision procedure that chooses an a minimizing the posterior expected loss $\rho(a, x)$, for each x.

- Note: Sometimes the loss is restricted to be nonnegative, to avoid certain pathologies.

Example 1: Estimating $\boldsymbol{\theta}$, with quadratic loss

- Setup:
  - State: $S = \boldsymbol{\theta}$
  - Observation: $x = x_{1:n}$
  - Action: $a = \hat{\theta}$
  - Loss: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (quadratic loss, a.k.a. square loss)
- Using quadratic loss here works out nicely, since the optimal decision is simply to estimate $\boldsymbol{\theta}$ by the posterior mean—in other words, to choose

$$\hat{\theta} = \delta(x_{1:n}) = \mathbb{E}(\boldsymbol{\theta} | x_{1:n}).$$

- To see why, note that $\ell(\theta, \hat{\theta}) = \theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2$, and thus

$$\rho(\hat{\theta}, x_{1:n}) = \mathbb{E}(\ell(\boldsymbol{\theta}, \hat{\theta}) | x_{1:n}) = \mathbb{E}(\boldsymbol{\theta}^2 | x_{1:n}) - 2\hat{\theta}\mathbb{E}(\boldsymbol{\theta} | x_{1:n}) + \hat{\theta}^2,$$

  which is convex as a function of $\hat{\theta}$. Setting the derivative with respect to $\hat{\theta}$ equal to 0, and solving, we find that the minimum occurs at $\hat{\theta} = \mathbb{E}(\boldsymbol{\theta} | x_{1:n})$.

Example 2: Predicting the next outcome, X_{n+1}, with $0 - 1$ loss

- Assume X_{n+1} is a discrete random variable.
- Setup:
  - State: $S = X_{n+1}$
  - Observation: $x = x_{1:n}$
  - Action: $a = \hat{x}_{n+1}$
  - Loss: $\ell(s, a) = \mathbb{1}(s \neq a)$ (this is called the $0 - 1$ loss)
- Using $0 - 1$ loss here works out nicely, since it turns out that the optimal decision is simply to predict the most probable value according to the posterior predictive distribution, i.e.,

$$\hat{x}_{n+1} = \delta(x_{1:n}) = \arg\max_{x_{n+1}} p(x_{n+1} | x_{1:n}).$$

4.2 Real-world decision problems

Medical decision-making

At what age should you get early screening for cancer (such as prostate or breast cancer)? There have been recent recommendations to delay screening until later ages due to a high number of false positives, which lead to unnecessary biopsies and considerable physical discomfort and mental distress, in addition to medical costs.

Public health policy

The CDC estimates that 5–20% of the US population gets influenza annually, and thousands die. Each year, in order to produce the right kinds of flu shots in sufficient quantities, researchers and vaccine manufacturers have to predict the prevalence of different strains of the virus at least 6 months in advance of flu season.

Government regulations

Per watt produced, there are approximately 4000 times more deaths due to coal power generation than due to nuclear power—not counting the environmental costs. Nonetheless, regulation of nuclear power is very stringent, perhaps due to a misperception of the risk.

Personal financial decisions

Should you buy life insurance? To make this decision, you would need to think about your probability of dying early and the financial impact it would have on your family, weighed against the cost of the policy.

A word of caution

Use good judgment! A formal decision analysis is almost always oversimplified, and it's a bad idea to adhere strictly to such a procedure. Decision-theoretic analysis can help to understand and explore a decision problem, but after all the analysis, decisions should be made based on your best judgment.

4.3 Example: Resource allocation for disease prevention/treatment

- Suppose public health officials in a small city need to decide how much resources to devote toward prevention and treatment of a certain disease, but the fraction θ of infected individuals in the city is unknown.
- Suppose they allocate enough resources to accomodate a fraction c of the population. If c is too large, there will be wasted resources, while if it is too small, preventable cases may occur and some individuals may go untreated. After deliberation, they tentatively adopt the following loss function:

$$\ell(\theta, c) = \begin{cases} |\theta - c| & \text{if } c \geq \theta \\ 10|\theta - c| & \text{if } c < \theta. \end{cases}$$

- By considering data from other similar cities, they determine a prior $p(\theta)$. For simplicity, suppose $\boldsymbol{\theta} \sim \text{Beta}(a, b)$ (i.e., $p(\theta) = \text{Beta}(\theta|a, b)$), with $a = 0.05$ and $b = 1$.
- They conduct a survey assessing the disease status of $n = 30$ individuals, $x_1, \ldots, x_n$. This is modeled as $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, which is reasonable if the individuals are uniformly sampled and the population is large. Suppose all but one are disease-free, i.e., $\sum_{i=1}^n x_i = 1$.

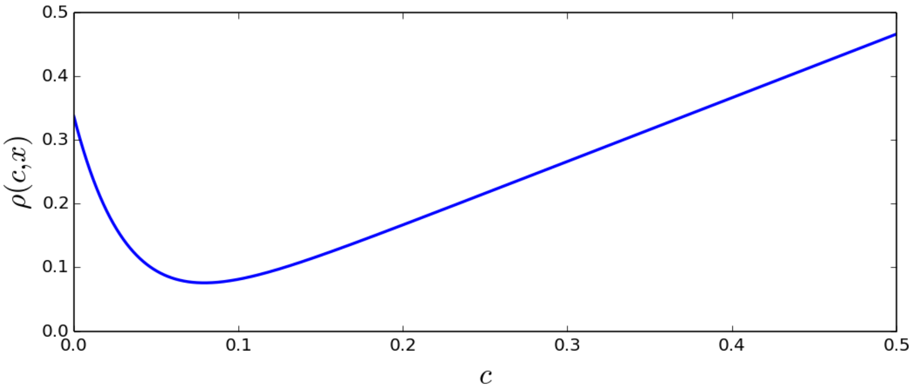


Figure 6: Posterior expected loss for the disease prevalence example.

The Bayes procedure

- The Bayes procedure is to minimize the posterior expected loss

$$\rho(c,x) = \mathbb{E}(\ell(\boldsymbol{\theta},c)|x) = \int \ell(\theta,c)p(\theta|x)d\theta$$

  where $x = x_{1:n}$. We know $p(\theta|x)$ from Equation 2.3, so we can numerically compute this integral for each c.

- Figure 6 shows $\rho(c,x)$ for our example. To visualize why it looks like this, think about the shape of $\ell(\theta,c)$ as a function of c, for some fixed θ—then imagine how it changes as θ goes from 0 to 1, and think about taking a weighted average of these functions, with weights determined by $p(\theta|x)$.

- The minimum occurs at $c \approx 0.08$, so under the assumptions above, this is the optimal amount of resources to allocate. Note that this makes more sense than naively choosing $c = \bar{x} = 1/30 \approx 0.03$, which does not account for uncertainty in θ and the large loss that would result from possible under-resourcing.

4.4 Frequentist risk and Integrated risk

- Consider a decision problem in which $S = \boldsymbol{\theta}$.

- The ***risk*** (or ***frequentist risk***) associated with a decision procedure δ is

$$R(\theta,\delta) = \mathbb{E}\big(\ell(\boldsymbol{\theta},\delta(X)) \mid \boldsymbol{\theta} = \theta\big),$$

  where X has distribution $p(x|\boldsymbol{\theta})$. In other words,

$$R(\theta,\delta) = \int \ell(\theta,\delta(x))\, p(x|\theta)\, dx$$

  if X is continuous, while the integral is replaced with a sum if X is discrete.

$$\begin{tikzcd}
& \begin{array}{c}\text{Loss}\\ L=\ell(\boldsymbol{\theta},\delta(X))\end{array} \arrow[dl] \arrow[dr] & \\
\begin{array}{c}\text{Post. exp. loss}\\ \mathbb{E}(L \mid X=x)\end{array} \arrow[dr] & & \begin{array}{c}\text{Frequentist risk}\\ \mathbb{E}(L \mid \boldsymbol{\theta}=\theta)\end{array} \arrow[dl] \\
& \begin{array}{c}\text{Integrated risk}\\ \mathbb{E}(L)\end{array} &
\end{tikzcd}$$

Figure 7: Visualizing the relationships between different decision-theoretic objects.

- The ***integrated risk*** associated with δ is

$$r(\delta)=\mathbb{E}(\ell(\boldsymbol{\theta},\delta(X))=\int R(\theta,\delta)\,p(\theta)\,d\theta.$$

- The diagram in Figure 7 (denoting $L=\ell(\boldsymbol{\theta},\delta(X))$ for brevity) helps to visualize the relationships between all of these concepts.

4.4.1 Example: Resource allocation, revisited

- The frequentist risk provides a useful way to compare decision procedures in a prior-free way.
- In addition to the Bayes procedure above, consider two other possibilities: choosing $c=\bar{x}$ (sample mean) or $c=0.1$ (constant).
- Figure 8 shows each procedure as a function of $\sum x_i$, the observed number of diseased cases. For the prior we have chosen, the Bayes procedure always picks c to be a little bigger than $\bar{x}$.
- Figure 9 shows the risk $R(\theta,\delta)$ as a function of θ for each procedure. Smaller risk is better. (Recall that for each θ, the risk is the expected loss, averaging over all possible data sets. The observed data doesn't factor into it at all.)
- The constant procedure is fantastic when θ is near 0.1, but gets very bad very quickly for larger θ. The Bayes procedure is better than the sample mean for nearly all θ's. These curves reflect the usual situation—some procedures will work better for certain θ's and some will work better for others.
- A decision procedure is called ***admissible*** if there is no other procedure that is at least as good for all θ and strictly better for some. That is, δ is admissible if there is no δ' such that

$$R(\theta,\delta')\le R(\theta,\delta)$$

for all θ and $R(\theta,\delta')<R(\theta,\delta)$ for at least one θ.

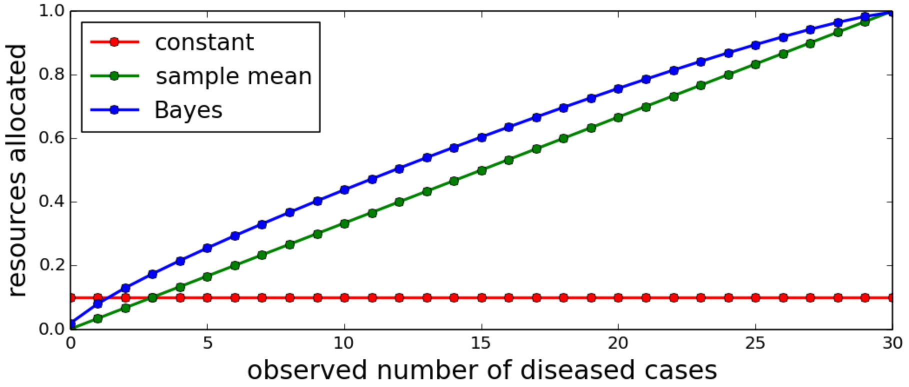


Figure 8: Resources allocated c, as a function of the number of diseased individuals observed, $\sum x_i$, for the three different procedures.

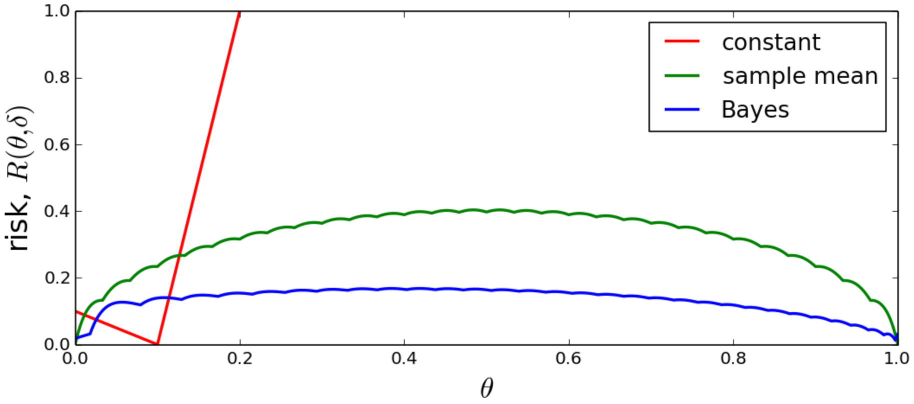


Figure 9: Risk functions for the three different procedures.

- Bayes procedures are admissible under very general conditions.
- Admissibility is nice to have, but it doesn't mean a procedure is necessarily good. Silly procedures can still be admissible—e.g., in this example, the constant procedure $c = 0.1$ is admissible too!

5 Exercises

Gamma-Exponential model

We write $X \sim \text{Exp}(\theta)$ to indicate that X has the Exponential distribution, that is, its p.d.f. is

$$p(x|\theta) = \text{Exp}(x|\theta) = \theta \exp(-\theta x)\mathbb{1}(x > 0).$$

The Exponential distribution has some special properties that make it a good model for certain applications. It has been used to model the time between events (such as neuron spikes, website hits, neutrinos captured in a detector), extreme values such as maximum daily rainfall over a period of one year, or the amount of time until a product fails (lightbulbs are a standard example).

Suppose you have data $x_1, \ldots, x_n$ which you are modeling as i.i.d. observations from an Exponential distribution, and suppose that your prior is $\boldsymbol{\theta} \sim \text{Gamma}(a, b)$, that is,

$$p(\theta) = \text{Gamma}(\theta|a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta)\mathbb{1}(\theta > 0).$$

1. Derive the formula for the posterior density, $p(\theta|x_{1:n})$. Give the form of the posterior in terms of one of the distributions we've considered so far (Bernoulli, Beta, Exponential, or Gamma).

2. Now, suppose you are measuring the number of seconds between lightning strikes during a storm, your prior is Gamma$(0.1, 1.0)$, and your data is

$$(x_1, \ldots, x_8) = (20.9, 69.7, 3.6, 21.8, 21.4, 0.4, 6.7, 10.0).$$

   Using the programming language of your choice, plot the prior and posterior p.d.f.s. (Be sure to make your plots on a scale that allows you to clearly see the important features.)

3. Give a specific example of an application where an Exponential model would be reasonable. Give an example where an Exponential model would NOT be appropriate, and explain why.

Decision theory

4. Show that if ℓ is $0 - 1$ loss and S is a discrete random variable, then the action a that minimizes the posterior expected loss $\rho(a, x_{1:n}) = \mathbb{E}(\ell(S, a)|x_{1:n})$ is the a that maximizes $\mathbb{P}(S = a \mid x_{1:n})$.

5. Consider the Beta-Bernoulli model. Intuitively, how would you predict x_{n+1} based on observations $x_1, \ldots, x_n$? Using your result from Exercise 4, what is the Bayes procedure for making this prediction when ℓ is $0 - 1$ loss?

6. Are there settings of the "hyperparameters" a, b for which the Bayes procedure agrees with your intuitive procedure? Qualitatively (not quantitatively), how do a and b influence the Bayes procedure?

7. What is the posterior mean $\mathbb{E}(\boldsymbol{\theta}|x_{1:n})$, in terms of a, b, and $x_1, \ldots, x_n$? Express this as a convex combination of the sample mean $\bar{x} = \frac{1}{n}\sum x_i$ and the prior mean (that is, write it as $t\bar{x} + (1 - t)\mathbb{E}(\boldsymbol{\theta})$ for some $t \in [0, 1]$).

8. Now, consider the loss function and the prior from the example in Section 4.3. Using the programming language of your choice, reproduce the plot in Figure 6. Do the integrals numerically using a Riemann sum approximation, such as $\int_0^1 f(x)dx \approx \frac{1}{N}\sum_{i=1}^N f((i - \frac{1}{2})/N)$ for a suitably large N.

9. Come up with a scenario in which S is discrete but the $0 - 1$ loss would NOT be appropriate, and give an example of the loss function that would be more suitable.

References and supplements

Probability basics

- Hoff (2009), Sections 2.2–2.6.

- mathematicalmonk videos, Probability Primer (PP) 2.1 – 5.5
  https://www.youtube.com/playlist?list=PL17567A1A3F5DB5E4

Beta-Bernoulli model

- Hoff (2009), beginning of Section 3.1.

- mathematicalmonk videos, Machine Learning (ML) 7.5 and 7.6
  https://www.youtube.com/playlist?list=PLD0F06AA0D2E8FFBA

Coin flipping bias

- Diaconis, P., Holmes, S., & Montgomery, R. (2007). Dynamical bias in the coin toss. SIAM review, 49(2), 211-235.

Decision theory

- *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Robert, C. P. (2001). Springer Texts in Statistics.

- *Statistical Decision Theory and Bayesian Analysis*. Berger, J.O. (1985). Springer.

- mathematicalmonk videos, Machine Learning (ML) 3.1 – 3.4 and 11.1 – 11.3
  https://www.youtube.com/playlist?list=PLD0F06AA0D2E8FFBA

History

- *The history of statistics: The measurement of uncertainty before 1900*. Stigler, S.M. (1986). Harvard University Press.

- Wolfowitz, J. (1952). Abraham Wald, 1902-1950. The Annals of Mathematical Statistics, 1-13.

- Interesting podcast about Wald