Statistics I Lecture Notes

Notes by Prof. Matthias Troffaes

Typeset in LaTeX by Tom Stoneham

Contents

0 Course Introduction

This course is delivered by Matthias Troffaes (matthias.troffaes@durham.ac.uk). Office hours are Mondays, from 08:40 - 10:40, in CM304.

Tutorials will occur once every two weeks, starting in week 12.

Problems classes occur once every two weeks in the Friday lecture spot, starting in week 14.

Homework will be set every Friday, and is to be handed in to your tutor by next Friday at 17:00.

We're working from a new course. Problem sheets and solutions are available on DUO, as are the lecture notes Matthias is working from. Not all past exam questions are going to be directly relevant to the course: it's mostly the more advanced questions from the problem sheets we should use for practice.

We will follow *Probability & Statistics*, 2012, 4ed, by DeGroot and Schervish, for most lectures, referred to in these notes as [DS12]. This is available as an e-book on DUO. For the first few lectures, we will follow *Applied Statistics & Probability for Engineers*, 2003, 3ed, by Montgomery & Runger, referred to as [MR03]. While there is not an e-book freely available, there are a number of copies available from the library. Furthermore, the library have actually scanned the first two chapters of [MR03], and these are available to read via DUO, under "Library Resources".

n.b: Only one person can "borrow" the DS12 ebook at a time, so email the library staff if you can't access it.

Throughout these notes, we will highlight certain information as follows:

Definition: Introduction

Definitions are in dark blue boxes, with the word being defined highlighted in bold.

Related Questions

References to related questions on the problem sheets are in light blue.

Warning

Any warnings/caveats related to a section's content are placed in red.

Further Reading

Further reading (from [DS12], [MR03]) will be placed in green.

1 Simple Frequentist Estimation and Prediction

1.1 Statistical Modelling and Inference

Further Reading

[MR03, section 7.1]
[DS12, section 7.1]

Consider the following practical scientific questions:

- What will be the sea level rise in the next 50 years?
- What's the effectiveness of a new cancer treatment?
- What's the biological impact of introducing a non-native species to an environment?
- How much energy will a new wind farm produce, if built in a certain location?
- What is the distribution of dark matter in the universe?

What do these situations have in common?

Definition: Uncertainty, Data, and Models

Each situation involves:

- **Uncertainty**: There's no exact answer, due to a lack of knowledge, and due to randomness.
- **Data**: Empirical observations, expert knowledge
- **Model**: Some idea of how the world behaves, how data are correlated. May be thought of as a specific way of expressing the objective part of expert knowledge.

We will use probability theory to tackle these types of question.

Probability theory involves a number of concepts we will make use of.

Definition: Probability Theory Concepts

The **possibility space**, notated Ω, is the set of all possible outcomes. This can be huge, for example, the set of all possible distributions of dark matter; or smaller, like if we were to represent sea level rise by a single number.
We don't normally specify possibility spaces directly, but instead focus on **random variables**. A random variable is a *function* from $\Omega \mapsto \mathbb{R}$ (or $\mathbb{R}^k$). This can be observed. Random variables will always be denoted by capital letters: $X, Y, \Theta, ...$
An actual observed value of a random variable will be denoted with a lower-case letter: $x, y, \theta, ...$
A **statistical model** consists of an identification of:

- Relevant random variables (both observable and hypothetically observable) including the data.

  For example: the expansion coefficient of water, actual rise, global temperature.

- Parameters (both known and unknown). We may learn about them, but not observe them directly.

  For example: The likelihood of recovery following the use of a certain treatment.

- A joint probability distribution, through probability mass functions (pmfs) and probability density functions (pdfs), on *all* random variables, and possibly on all unknown paramters.

Warning

Random variables *only* correspond to observable (or *hypothetically* observable) quantities.
We may treat parameters as random variables in a statistical model.

Related Questions

Exercises 1 and 2 on the problem sheet give you textual descriptions of some scenarios, and ask you to identify the statistical model.

Definition: Statistical inference

A **statistical inference** is a procedure which produces a **probabilistic statement** about any part of a statistical model.
A probabilistic statement is just the probability of an event, a mean, a variance, etc; i.e: Anything involving a mathematical statement of probability.

1.1.1 Example: Smartphone Batteries

Consider a smartphone battery production line. Every 50th battery is destructively tested for "quality". Quality, here, is just some proxy for the battery lifetime. We're given the following data for the quality of the last 10 tested batteries:

Battery	1	2	3	4	5	6	7	8	9	10
Quality	90	86	82	77	94	90	87	90	86	86

A battery is deemed faulty if quality is less than 80. Identify the relevant statistical model.

We first need to identify the relevant random variables:

- The quality of the tested batteries (the *data*, known): $X_1, ..., X_n$
- The quality of the untested batteries (hypothetically observable, unknown): $X_{n+1}, X_{n+2}, ...$

$n = 10$ is the sample size.

We now need to assign some joint distribution. We might conjecture the following model, for example:

The X_i are identically distributed (come from the same probability distribution), according to some probability density function $f(\cdot|\theta)$, for example, the normal distribution, with $\mathbb{E}(X_i|\theta) = \theta, \mathbb{V}\text{ar}(X_i|\theta) = 5^2$. This is just an assumption which seems reasonable when we look at the data. θ is an unknown parameter representing the mean of X_i. Θ is a random variable representing θ.

Here, $\theta \in \mathbb{R}$, but in general, you can have $\theta \in \mathbb{R}^k$ (i.e: multiple parameters).

Warning

We must assume that X_i are independent *conditional on* Θ.
Why can't we assume unconditional independence? Let's assume for now that all variables are discrete, for simplicity.
Say that I observe X_1 to learn about X_2. So we're trying to find:

$$\mathbb{P}(X_2 = x_2|X_1 = x_1) = \frac{\mathbb{P}(X_2 = x_2, X_1 = x_1)}{\mathbb{P}(X_1 = x_1)}$$

If we assume the variables are unconditionally independent, we then have:

$$\begin{aligned}\mathbb{P}(X_2 = x_2|X_1 = x_1) &= \frac{\mathbb{P}(X_2 = x_2)\mathbb{P}(X_1 = x_1)}{\mathbb{P}(X_1 = x_1)} \\ &= \mathbb{P}(X_2 = x_2)\end{aligned}$$

So we've learned absolutely nothing about X_2!
If, on the other hand, we assume conditional independence on Θ, we have:

$$\begin{aligned}\mathbb{P}(X_2 = x_2|X_1 = x_1) &= \sum_\theta \mathbb{P}(X_2 = x_2|X_1 = x_1, \Theta = \theta) \cdot \mathbb{P}(\Theta = \theta|X_1 = x_1) \\ &= \sum_\theta \mathbb{P}(X_2 = x_2|\Theta = \theta) \cdot \mathbb{P}(\Theta = \theta|X_1 = x_1) \\ &= \sum_\theta \mathbb{P}(X_2 = x_2|\Theta = \theta) \cdot \frac{\mathbb{P}(X_1 = x_1|\Theta = \theta)\mathbb{P}(\Theta = \theta)}{\mathbb{P}(X_1 = x_1} \\ &\neq \mathbb{P}(X_2 = x_2) \quad \textit{(generally)}\end{aligned}$$

So we can learn about X_2 given the value of X_1! This is actually the *only* way to enable learning, by DeFinetti's representation theorem.

1.1.2 Estimation and Prediction

Typically, we perform statistical inference about either unknown parameters, such as Θ, or about future observations, such as X_{n+1}.

Definition: Estimation, Prediction

Estimation specifically refers to statistical inference about unknown parameters, which **prediction** refers to statistical inference about future observations.

1.2 Point Estimation

Further Reading

[MR03, section 7.2]

Let's return to the battery example. Consider the sample mean:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

This would be a good choice to estimate θ, because:

$$\begin{aligned}\mathbb{E}(\bar{X}_n|\theta) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i|\theta) \\ &= \frac{1}{n}n\theta \\ &= \theta\end{aligned}$$

We can also use the sample mean to estimate the variance:

$$\begin{aligned}\mathbb{V}\text{ar}(\bar{X}_n|\theta) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\text{ar}(X_i|\theta)\\ &= \frac{1}{n^2}n5^2\\ &= \frac{5^2}{n}\end{aligned}$$

Note that $\mathbb{V}\text{ar}(\bar{X}_n|\theta) \to 0$ as $n \to \infty$. So, we might use $\bar{X}_n$ as an approximation for Θ.

Definition: Statistic, Estimator, Point Estimate

A **statistic** is a real-valued function of the data, for example, $\bar{X}_n$.
An **estimator**, $\hat{T}$, is a statistic which is meant to approximate some real-valued function, $t(\Theta)$, of the parameters. Remember that the parameters are random variables, so a function of a parameter is also a random variable. Usually, $t(\Theta)$ is just the identity function. We write $\hat{\Theta}$ for an estimator of Θ.
A **point estimate**, $\hat{t}$ for $t(\Theta)$ is just a specific realisation of an estimator $\hat{T}fort(\Theta)$ after observing the data (the actual value of $\hat{T}$). We write $\hat{\theta}$ for a point-estimate of Θ.

In our battery example, $\hat{\Theta} = \bar{X}_{10}$ is an estimator for Θ. $\hat{\theta} = \frac{1}{10}(90 + 86 + ...) = 86.8$ is a point estimate of Θ.

1.2.1 Properties of Estimators

Definition: Bias, Errors

For any estimator $\hat{T}$ of $t(\Theta)$, we define:

- **Bias** $:= \mathbb{E}(\hat{T}|\theta) - t(\theta)$

  An estimator with zero bias $\forall\,\theta$ is called **unbiased**.

  We saw that, for the sample mean, the conditional expectation is equal to the value we want to estimate, so we would say this estimator is unbiased.

- **Standard error** $:= \sqrt{\mathbb{V}\text{ar}(\hat{T}|\theta)}$.

  We can think of this as how much an estimator will vary, its conditional standard deviation.

  We want for an estimator to have a low standard error.

- **Mean square error** $:= \mathbb{E}((\hat{T} - t(\theta)^2|\theta)$

  This is what we (usually) *really* want to minimise for a good estimator.

Normally, we want estimators with a low mean square error.

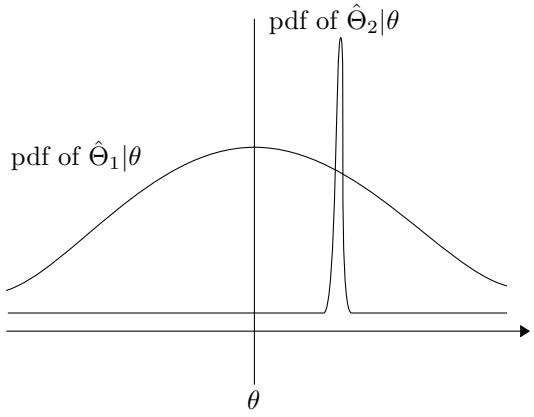
1.2.2 Theorem: Relation of Mean Square Error, Standard Error, and Bias

$$\text{mean square error} = (\text{standard error})^2 + (\text{bias})^2$$
$$\mathbb{E}((\hat{T} - t(\theta))^2|\theta) = \mathbb{V}\text{ar}(\hat{T}|\theta) + (\mathbb{E}(\hat{T}|\theta) - t(\theta))^2$$

Related Questions

Exercise 3 involves proving this relation. The solution is on DUO if you really want to be sure about the proof.

This theorem is important because, since we're trying to minimise mean square error, we might actually prefer a biased estimator, provided its standard error is lower.



In this case, for example, we might choose to use the clearly biased estimator $\hat{\Theta}_2$, since its variance being so low may lead to a lower mean square error than $\hat{\Theta}_1$.

1.2.3 Theorem: Minimum Variance Unbiased Estimator

Consider some sequence of random variables $X_1, X_2, ...$, with $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ where the X_i are independent, identically distributed (IID) conditional on Θ and $\sigma > 0$ is some known constant. Then, $\bar{X}_n$ is the minimum variance unbiased estimator of Θ (in essence: the sample mean is the best estimator we can construct of Θ).

Related Questions

The proof of this theorem is not given, but a related (simpler) proof is required for exercises 3 to 7.

1.2.4 Bias and Error in the Battery Example

In the battery example, we showed:

$$\mathbb{E}(\bar{X}_n|\theta) = \theta$$
$$\mathbb{V}\text{ar}(\bar{X}_n|\theta)\frac{5^2}{n}$$

So $\bar{X}_n$ is an unbiased estimator of Θ.

$\bar{X}_n$ has standard error $\frac{5}{\sqrt{n}}$, so the mean square error is $\frac{5^2}{n}$.

1.3 Interval Estimation

Further Reading

[MR03, section 8.1, 8.2.1]

1.3.1 Central Limit Theorem (CLT)

This is a key result from probability theory.

Consider an infinite sequence of random variables, $X_1, X_2, ...$ We will assume they are IID conditional on some random variable Θ.

We will define the following functions:

$$\mu(\theta) := \mathbb{E}(X_i|\theta)$$
$$\sigma^2(\theta) := \mathbb{V}\text{ar}(X_i|\theta)$$

Note that neither μ nor θ depend on i due to the IID assumption.

We will further define the following random variables:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Z_n := \frac{\bar{X}_n - \mu(\Theta)}{\sigma(\Theta)/\sqrt{n}} \quad \textit{(the standardised sample mean)}$$

Then, $\forall\, z \in \mathbb{R}$, and all possible values of θ, we have:

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z|\theta) = \Phi(z)$$

where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$.

The practical implication of this theorem is that, for large n, we have the approximate result:

$$\bar{X}_n|\theta \sim \mathcal{N}\left(\mu(\theta), \frac{\sigma^2(\theta)}{n}\right)$$

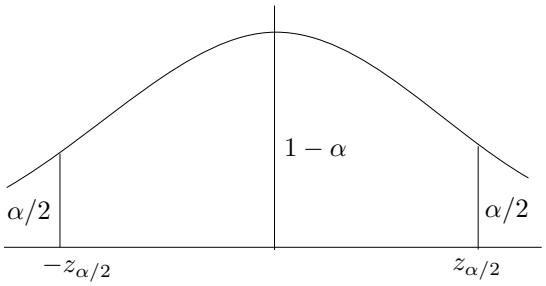
This approximation improves with a larger value of n, or as the distribution of $X_i|\theta$ is closer to the normal.

1.3.2 Application of CLT to Battery Example

In our battery example, $\mu(\theta) = \theta, \sigma^2(\theta) = 5^2$. So, $\bar{X}_{10}|\theta \sim \mathcal{N}\left(\theta, \frac{5^2}{10}\right)$ approximately.

Can we exploit the CLT to make stronger probabilistic statements about $\bar{X}_{10}$ And about Θ?

Let's assume the conditions of the central limit theorem, with $\mu(\theta) = \theta$ and $\sigma^2(\theta) = \sigma^2$ (i.e: σ^2 is constant). Fix any $\alpha \in [0,1]$ and let $z_{\frac{\alpha}{2}} := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$



By the CLT:

$$\mathbb{P}(|Z_n| \leq z_{\alpha/2}|\theta) = 1 - \alpha$$

By the definition of Z_n:

$$\mathbb{P}\left(\left|\frac{\bar{X}_n - \theta}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}|\theta\right) = 1 - \alpha$$

or equivalently:

$$\mathbb{P}\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\theta\right) \leq 1-\alpha$$

This is a **confidence interval**, which we will define after a caveat.

Warning

Note that, here, the θ on the left is a constant as we're conditional on θ, not a random variable, and X_n is a random variable. So, if we have a value for the sample mean, we can construct the interval.
This does *not* say that the probability that θ lies in this specific interval is $1-\alpha$.
See section 1.3.3 for more information on the issues with thinking about confidence intervals like this.

Definition: Confidence Interval

Assume $X_1, X_2, ..., X_n$ are IID, conditional on θ, and assume that $\mathbb{E}(X_i|\theta) = \theta$, $\mathbb{V}\text{ar}(X_i|\theta) = \sigma^2 > 0$ is known, and constant independent of θ. Then, $\forall\, \alpha \in [0,1]$:

$$\left[\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is a $100 \cdot (1-\alpha)\%$ **confidence interval** on Θ.

Common values for $z_{\alpha/2}$:

$1-\alpha$	0.80	0.90	0.95	0.98	0.99
$z_{\alpha/2}$	1.28	1.64	1.96	2.33	2.58

Now that we've defined a confidence interval, we should return to the battery example. In this example, $n = 10, \bar{x}_{10} = 86.8, \sigma = 5$. For a 95% confidence interval, we need to calculate:

$$\bar{x}_n - 1.96\frac{\sigma}{\sqrt{n}} = 83.7$$
$$\bar{x}_n + 1.96\frac{\sigma}{\sqrt{n}} = 89.9$$

So the 95% confidence interval for Θ is given by $[83.6, 89.9]$.

Related Questions

Problem 11 is highly relevant here, and is the closest so far to an actual exam question we would expect.
Problem 9 also relies on confidence intervals and would be useful to attempt.

1.3.3 What Does a Confidence Interval Mean?

In the battery example, does $[83.6, 89.9]$ really capture the uncertainty of Θ? In particular, does the event $\Theta \in [83.7, 89.9]$ have probability 0.95? *No.*

$$\mathbb{P}\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\theta\right) \leq 1-\alpha$$

The probability in this equation, which we used to define a confidence interval, is *conditional on knowing* $\Theta = \theta$. This isn't really what we want! We want it to be conditional on $\bar{X}_n = \bar{x}_n$.

Instead, we *only* know that $\forall\, \theta \in \mathbb{R}$, the following event:

$$\theta \in \left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

has probability 0.95, conditional on $\Theta = \theta$.

We used the distribution of $\bar{X}_n$ conditional on $\Theta = \theta$. We *really* want the distribution of Θ conditional on $X_1 = x_1, ..., X_n = x_n$.

How can we do that? We'll use Bayes's theorem in a later section. section. However, the computations are much harder, and we must be able to treat Θ as a random variable, which we've managed to avoid while thinking about confidence intervals.

1.4 Interval Prediction

Further Reading

[MR03, section 8.6]

Let's think about the battery example again. What can we say about the quality level of an untested battery, X_{n+1}? In other words, having observed $X_1, ..., X_n$, what can we say about X_{n+1}?

In prediction, we can no longer rely on the central limit theorem as we did in estimation, and must instead make another assumption about the distribution of X_n. So, we will assume normality, again, IID conditional on θ:

$$X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$$

We then have:

$$\begin{aligned}\mathbb{E}(X_{n+1} - \bar{X}_n|\theta) &= \mathbb{E}(X_{n+1}|\theta) - \mathbb{E}(\bar{X}_n|\theta) \\ &= \theta - \theta \\ &= 0\end{aligned}$$

$$\begin{aligned}\mathbb{V}\text{ar}(X_{n+1} - X_n|\theta) &= \mathbb{V}\text{ar}(X_{n+1}|\theta) + \mathbb{V}\text{ar}(\bar{X}_n|\theta) && \textit{(by IID of } X_1, ..., X_{n+1}\textit{)} \\ &= \sigma^2 + \frac{\sigma^2}{n} \\ &= \sigma^2\left(1 + \frac{1}{n}\right)\end{aligned}$$

Note that X_{n+1} is distributed normally, and each X_i is normal, so the sample mean (a sum of normals) will also be normally distributed, and the difference $X_{n+1} - \bar{X}_n$ will also have a normal distribution, as follows:

$$X_{n+1} - \bar{X}_n \sim \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$$

and consequently, via a similar calculation to how we analysed confidence intervals, we have:

$$\mathbb{P}\left(\bar{X}_n - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{X}_n + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}|\theta\right) = 1 - \alpha$$

1.4.1 Interval Prediction in the Battery Example

We have $n = 10, \bar{X}_n = 86.8, \sigma = 5$, so we attain:

$$\bar{x}_n - 1.96\sigma\sqrt{1+\frac{1}{n}} = 76.5$$
$$\bar{x}_n + 1.96\sigma\sqrt{1+\frac{1}{n}} = 97.1$$

Recall that a battery was designated faulty if its quality was less than 80, so the 95% prediction interval includes values less than this threshold. As such, we'd expect to see more failures than we might be comfortable with, and we might wish to redesign the battery production process.

The interval $[76.5, 97.1]$ is called a 95% frequentist prediction interval for X_{n+1}.

Definition: Frequentist Prediction Interval

Assume $X_1, X_2, ..., X_n$ are IID, conditional on θ, and assume that $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$, where σ is independent of θ. Then, $\forall\, \alpha \in [0,1]$:

$$\left[\bar{x}_n - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}, \bar{x}_n + z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}\right]$$

is a $100 \cdot (1-\alpha)\%$ **frequentist prediction interval** for X_{n+1}.

Warning

Similar concerns to those involved in interval estimation apply: We used the distribution of $\bar{X}_n$ and X_{n+1} conditional on $\Theta = \theta$, but we actually want the distribution of X_{n+1} conditional on the data: $X_1 = x_1, ..., X_n = x_n$!

1.5 Parameters as Random Variables

Further Reading

[DS12, section 7.1]

For what we've done so far, we have avoided having to consider Θ as a random variable, because we only used $\mathbb{P}(\cdot|\theta)$, which we could consider as a distribution parametrised by θ. This may be notated (and in textbooks, commonly is notated) as $\mathbb{P}_\theta(\cdot)$, or even just $\mathbb{P}(\cdot)$.

Doing this is appropriate if we only want to make probability statements indexed by θ. However, this is very limiting.

If we want to make more advanced probability statements, can we treat Θ as a random variable? *Yes*, provided that Θ is (hypothetically) observable.

How can we make Θ observable? Consider, for instance, the battery example, in which we observed the quality of batteries. By the strong rule of large numbers, we have:

$$\mathbb{P}\left(\Theta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i\right) = 1$$

This just means that as we take a very large value of n for our sample size, we expect the sample average to converge to Θ.

So in this specific example, we can treat Θ as if it were the observable quantity $\frac{1}{n}\sum_{i=1}^{n} X_i$ for very large n.

This limit construction is possible for a very wide range of practical statistical models (and, in particular, every model that we'll encounter in first year).

2 Prior and Posterior Distributions

2.1 Prior Distributions

Further Reading

[DS12, section 7.2]

Definition: Prior PDF/PMF

The **prior pdf or pmf** of a *parameter* Θ is the pdf/pmf of Θ in advance of observing any data. This can be used to quantify uncertainty in advance of observing actual values. Mathematically, this is the *unconditional* marginal pdf/pmf of Θ.
We don't have prior pdf/pmfs for all random variables, only parameters.

2.1.1 Examples

Let's go back to the battery example we looked at throughout section 1, which we modelled as follows:

$$X_i|\theta \sim \mathcal{N}(\theta, 5^2)$$

Prior to seeing the data, the manager supposes that the parameter Θ is normally distributed, with mean 90 and standard deviation 10. So we have:

$$\Theta \sim \mathcal{N}(90, 10^2), \implies f(\theta) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\theta - 90}{10}\right)^2\right)$$

This is the prior pdf for Θ. *It is unconditional on any* X_i.

Warning

The point of a prior pdf/pmf is that it is entirely unconditional on any X_i. You *cannot* use any data parameters observed in a sample to determine the prior distribution.

Let's move on to another example. Consider the outcome X of a coin toss. $X|\theta \sim \text{Bin}(\theta, 1)$. Let $X = 1$ represent heads, and $X = 0$ tails. We know that either:

- The coin is fair: $\theta = \frac{1}{2}$,
- Or the coin has two heads: $\theta = 1$

Before we do any experimentation, we'll make a judgement about whether we think the coin is fair or not. Let's suppose we know the coin is fair with 80% probability. We can express this information using a probability mass function:

$$p(\theta) = \begin{cases} 0.8, & \theta = \frac{1}{2} \\ 0.2, & \theta = 1 \end{cases}$$

This is a prior pmf on Θ.

We now have two examples of prior pdf/pmfs, and the distribution of data. Upon observing data, can we update our potential pdf/pmfs to quantify the uncertainty in Θ?

2.2 Posterior Distribution and Likelihood

Definition: Posterior Distribution

The **posterior distribution** of a *parameter* Θ is the pdf or pmf of Θ after observing the data (i.e: the *conditional* pdf/pmf of Θ) given $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ (or any other observed random variables comprising the data, although in this course the observed random vars will always be in this form).

Usually, the posterior distribution is calculated via Bayes's theorem.

2.2.1 Posterior Distribution for the Coin Toss Example

Consider again the coin toss problem, and say that we observe heads: $X = 1$. This provides some evidence that it might be more likely that there are two heads.

Then, by Bayes's theorem, we have:

$$\begin{aligned}\mathbb{P}\left(\Theta = \frac{1}{2}|X = 1\right) &= \frac{\mathbb{P}\left(X = 1|\Theta = \frac{1}{2}\right)\mathbb{P}\left(\Theta = \frac{1}{2}\right)}{\mathbb{P}\left(X = 1|\Theta = \frac{1}{2}\right)\mathbb{P}\left(\Theta = \frac{1}{2}\right) + \mathbb{P}(X = 1|\Theta = 1)\mathbb{P}(\Theta = 1)} \\ &= \frac{\frac{1}{2} \cdot 0.8}{\frac{1}{2} \cdot 0.8 + 1 \cdot 0.2} \\ &= \frac{2}{3}\end{aligned}$$

We could also evaluate the probability is 1 by performing the same calculation to attain $\mathbb{P}(\theta = 1|X = 1) = \frac{1}{3}$. So the probability that the coin is biased has increased, which makes intuitive sense.

It's also worth checking for yourself that $\mathbb{P}\left(\Theta = \frac{1}{2}|x = 0\right) = 1$.

2.2.2 Theorem: Posterior Distribution in General

Assume $X_1, ..., X_n$ are IID, conditional on Θ, with pdf $f(x|\theta)$ and Θ has a prior pdf $f(\theta)$. Then, Θ has posterior pdf given by:

$$f(\theta|x_1, ..., x_n) = \frac{f(\theta)\prod_{i=1}^n f(x_i|\theta)}{\int f(\theta')\prod_{i=1}^n f(x_i|\theta')d\theta'}$$

θ' is just a dummy variable over which we're integrating here.

If we instead have a probability mass function for the data, the formula is exactly the same, we just replace $f(x_i|\theta)$ with $p(x_i|\theta)$.

If the prior distribution is a pmf rather than a pdf, replace $f(\theta)$ with $p(\theta)$, replace $f(\theta|x_1, ..., x_n)$ with $p(\theta|x_1, ..., x_n)$, and note that the integral $\int ...d\theta'$ becomes a sum $\sum_{\theta'}$.

Proof Given a model as above with IID observations and a prior pdf, we can write down the full joint density for the model by the definition of conditional density:

$$\begin{aligned}f(x_1, ..., x_n, \theta) &= f(x_1, ..., x_n|\theta) \cdot f(\theta) \\ &= f(\theta) \cdot \prod_{i=1}^n f(x_i|\theta) \qquad \textit{(by IID of } X_i \textit{ cond on } \Theta\textit{)}\end{aligned}$$

Also, we have that the integral of this product over θ will give us the joint distribution of the X_is, by the partition theorem:

$$f(x_1, ..., x_n) = \int f(x_1, ..., x_n, \theta')d\theta'$$
$$= \int f(\theta') \cdot \prod_{i=1}^{n} f(x_i|\theta')d\theta'$$

Finally, by the definition of conditional density, we have that:

$$f(\theta|x_1, ..., x_n) = \frac{f(x_1, ..., x_n, \theta)}{f(x_1, ..., x_n)}$$
$$= \frac{f(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int f(\theta')\prod_{i=1}^{n} f(x_i|\theta')d\theta'} \qquad \square$$

The proof is very similar if we use pmfs instead.

Tackling the Integral The tricky part of using this theorem to calculate posterior distributions tends to be evaluating the integral in the denominator. It is useful to note, however, that it only depends on the data $x_1, ..., x_n$, not the parameter θ. It only acts to normalise the numerator (to make sure the integral of the density equals 1).

We can exploit this fact to write the following:

$$f(\theta|x_1, ..., x_n) \overset{(\theta)}{\propto} f(\theta) \cdot \prod_{i=1}^{n} f(x_i|\theta)$$

Warning

The symbol $\overset{(\theta)}{\propto}$ means "is equal to, up to a factor independent of θ". So the constant of proportionality (the *normalisation constant*) can depend on anything except θ.

Definition: Likeliness Function, Likelihood

We define the **likeliness function** or **likelihood** to be the conditional pdf/pmf of the data given the parameter:

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where the x_i are IID are conditional on Θ.
So, the posterior is proportional to the prior distribution, multiplied by the likelihood.

2.2.3 Assorted Lemmata

For working with normal priors and likelihoods, the following results are very useful:

Lemma 2.2.3(a) Let's say we have the following:

$$f(y|z) \overset{(y)}{\propto} \exp -\frac{1}{2}\left(\frac{y - \mu(z)}{\sigma(z)}\right)^2$$

This is equivalent to saying:

$$Y|z \sim \mathcal{N}(\mu(z), \sigma^2(z))$$

The proof of this is trivial given the definition of the normal distribution.

Lemma 2.2.3(b) Let's say we have a sum of the following form:

$$\sum_{i=1}^{n} a_i(\theta - b_i)$$

We then have the following:

$$\sum_{i=1}^{n} a_i(\theta - b_i)^2 = \left(\sum_{i=1}^{n} a_i\right) \cdot \left(\theta - \frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i}\right)^2 + ... \qquad \textit{(all other terms are independent of } \theta\textit{)}$$

Lemma 2.2.3(c) If we just have two terms in the previous sum:

$$a_1(\theta - b_1)^2 + a_2(\theta - b_2)^2 = (a_1 + a_2) \cdot \left(\theta - \frac{a_1 b_1 + a_2 b_2}{a_1 + a_2}\right)^2 + \left(\frac{1}{a_1} + \frac{1}{a_2}\right)^{-1} \cdot (b_1 - b_2)^2$$

2.2.4 Posterior Distribution for the Battery Example

We'll assume that the distribution of quality is normal, IID conditional on θ:

$$X_i|\theta \sim \mathcal{N}(\theta, 5^2)$$

What is the likelihood (up to $\overset{(\theta)}{\propto}$)?

$$\begin{aligned}
f(x_1, ..., x_n|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\
&= \prod_{i=1}^{n} \frac{1}{5\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x_i - \theta}{5}\right)^2 \\
&\overset{(\theta)}{\propto} \exp -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{5^2}(\theta - x_i)^2\right) \\
&= \exp -\frac{1}{2}\left(\frac{n}{5^2}\left(\theta - \frac{\sum_{i=1}^{n}\frac{x_i}{5^2}}{n/5^2}\right)^2 + F\right) \qquad \textit{(by lemma 2.2.3(b), } F \textit{ independent of } \theta\textit{)} \\
&\overset{(\theta)}{\propto} \exp -\frac{n}{2 \cdot 5^2}(\theta - \bar{x}_n)^2 \\
&= \exp -\frac{10}{2 \cdot 5^2}(\theta - 86.8)^2
\end{aligned}$$

So, up to a factor independent of θ, our likelihood only depends on the data through $\bar{x}_n$. Consequently, so will the posterior! Therefore, for this model, $\bar{X}_n$ is called a **sufficient statistic**: We don't need to know the full data to make inferences about Θ, we only need to know the sample mean $\bar{X}_n$!

In order to find a posterior distribution, we need to suppose a prior distribution for Θ, which is based on expert knowledge. Let's say an expert gives the following prior distribution:

$$\Theta \sim \mathcal{N}(90, 10^2)$$

What is the posterior distribution for Θ?

$$\begin{aligned}
f(\theta|x_1,...,x_n) &\overset{(\theta)}{\propto} f(\theta)f(x_1,...,x_n|\theta) \\
&\overset{(\theta)}{\propto} \exp -\frac{1}{2}\left(\frac{\theta-90}{10}\right)^2 \cdot \exp -\frac{n}{2\cdot 5^2}(\theta-\bar{x}_n)^2 \\
&= \exp -\frac{1}{2}\left[\left(\frac{\theta-90}{10}\right)^2 + \frac{n}{5^2}(\theta-\bar{x}_n)^2\right] \\
&= \exp -\frac{1}{2}\left(\frac{1}{10^2}+\frac{n}{5^2}\right)\left(\theta - \frac{\frac{90}{10^2}+\frac{n\bar{x}_n}{5^2}}{\frac{1}{10^2}+\frac{n}{5^2}}\right)^2 \\
&= \exp -\frac{1}{2}\left(\frac{1}{1.56}\right)^2\left(\theta - \frac{\frac{90}{10^2}+\frac{n\bar{x}_n}{5^2}}{\frac{1}{10^2}+\frac{n}{5^2}}\right)^2 \\
&= \exp -\frac{1}{2}\left(\frac{1}{1.56}\right)^2(\theta-86.9)^2
\end{aligned}$$

$$\implies \Theta|x_1,...,x_n \sim \mathcal{N}(86.9, 1.56^2)$$

We can now construct genuinely useful probability intervals for Θ conditional on the data!

$$\mathbb{P}(86.9 - 1.96\cdot 1.56 \leq \Theta \leq 86.9 + 1.96\cdot 1.56|x_1,...,x_n) = 0.95$$

$[83.8, 90.0]$ is called a 95% credible interval on θ.

Warning

We should note that the 95% credible interval for θ, $[83.8, 90.0]$ is very similar to the 95% confidence interval, $[83.7, 89.9]$, obtained earlier from the same data. This is not a coincidence as we will see later, but we *must* keep in mind the different interpretations of the two intervals.

Definition: Credible Interval

An interval $[l, u]$ is called a $100\cdot(1-\alpha)\%$ **credible interval** for Θ given $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ when the posterior probability of $\Theta \in [l, u]$ is $1-\alpha$.

$$\mathbb{P}(l \leq \Theta \leq u|X_1 = x_1, ..., X_n = x_n) = 1-\alpha$$

$X_1, ..., X_n$ could be any other variable that comprise the data.

To bring into sharper relief the difference between a confidence interval and a credible interval, let's consider what happens when we have a more informative prior, say:

$$\Theta \sim \mathcal{N}(90, 0.1^2)$$

We have the same calculations as before, leaving us with:

$$\Theta|x_1,...,x_n \sim \mathcal{N}(89.99, 0.0998^2)$$

This is almost unchanged from the prior, due to the extremely low variance of the prior reflecting strong prior knowledge in the value of $\Theta \approx 90$.

2.2.5 Theorem: General Case for Normal Sampling

Consider a general situation as before, with σ, σ_0 and μ_0 known constants, X_i IID conditional on θ, distributed as follows:

$$X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$$

and prior distribution:

$$\Theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

Then:

$$\Theta|x_1, ..., x_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

where:

$$\mu_n := \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}_n}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$\sigma_n^2 := \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

These formulae should probably just be memorised.

The proof is similar to the battery example given earlier.

So if $\frac{1}{\sigma^2} \ll \frac{n}{\sigma^2}$, then the data will drive the posterior, and the credible interval will be more similar to the confidence interval.

Whereas, if $\frac{1}{\sigma^2} \gg \frac{n}{\sigma^2}$, then the prior will drive the posterior.

If $\frac{1}{\sigma^2} \approx \frac{n}{\sigma^2}$, then both the data and the prior will influence the posterior.

Warning

Thinking about the prior/posterior in this way really makes it clear why σ_0, μ_0 (the std. dev. and mean for the prior) should not be based on the data! They must reflect the uncertainty about θ as if you had not seen the data.

2.3 Sequential Updating and Prediction

Let's look at the posterior after a single observation, $X_1 = x_1$:

$$f(\theta|x_1) \overset{(\theta)}{\propto} f(\theta)f(x_1|\theta)$$

Now after the second observation, $X_2 = x_2$:

$$\begin{aligned} f(\theta|x_1, x_2) &\overset{(\theta)}{\propto} f(\theta)f(x_1|\theta)f(x_2|\theta) \\ &\overset{(\theta)}{\propto} f(\theta|x_1)f(x_2|\theta) \end{aligned}$$

So, we can treat the posterior after our first observation as our prior for the posterior following the second! Similarly, after the third observation, $X_3 = x_3$:

$$\begin{aligned} f(\theta|x_1, x_2, x_3) &\overset{(\theta)}{\propto} f(\theta)f(x_1|\theta)f(x_2|\theta)f(x_3|\theta) \\ &\overset{(\theta)}{\propto} f(\theta|x_1, x_2)f(x_3|\theta) \end{aligned}$$

2.3.1 The Sequential Updating Theorem

$$f(\theta|x_1,...,x_{n+1}) \overset{(\theta)}{\propto} f(\theta|x_1,...,x_n)f(x_{n+1}|\theta)$$

We can treat the posterior after n observations as the prior for updating with observation $n+1$.

The constant of proportionality has a special meaning, encapsulated in the following theorem.

2.3.2 Theorem: Predictive Distribution from Sequential Updates

$$f(\theta|x_1,...,x_{n+1}) = \frac{f(\theta|x_1,...,x_n)f(x_{n+1}|\theta)}{f(x_{n+1}|x_1,...,x_n)}$$

So the constant of proportionality is the predictive distribution! We will use this to predict the next observation given the previous observations.

Proof

$$\begin{aligned} f(\theta|x_1,...,x_{n+1})f(x_{n+1}|x_1,...,x_n) &= f(\theta, x_{n+1}|x_1,...,x_n) \\ &= f(x_{n+1}|\theta, x_1,...,x_n)f(\theta|x_1,...,x_n) \\ &= f(x_{n+1}|\theta)f(\theta|x_1,...,x_n) \qquad \textit{(by IID on } \theta\textit{)} \quad \square \end{aligned}$$

Definition: Prior Predictive PDF

We define the **prior predictive pdf** to be $f(x_{n+1})$, and the **posterior predictive pdf** to be $f(x_{n+1}|x_1,...,x_n)$.
We use the prior predictive pdf to predict X_{n+1} before having seen the data, and the posterior predictive to predict X_{n+1} after having seen the data.

2.3.3 Prediction in the Battery Example

Consider again the battery example with X_i IID conditional on Θ distributed according to:

$$X_i|\theta \sim \mathcal{N}(\theta, 5^2)$$

with the prior distribution for Θ given by:

$$\Theta \sim \mathcal{N}(90, 10^2)$$

Let's find $f(x_{n+1}|x_1,...,x_n)$ for our specific data. One way to do so is to use theorem 2.3.2, and if need be, evaluate up to $\overset{(x_{n+1})}{\propto}$:

$$f(x_{n+1}|x_1,...,x_n) = \frac{f(\theta|x_1,...,x_n)f(x_{n+1}|\theta)}{f(\theta|x_1,...,x_{n+1})}$$

Note that, since the left hand side is not a function of θ, you should either see that factors of θ on the right hand side cancel out, or just fix a value of θ where $f(\theta|x_1,...,x_{n+1}) > 0$.

$$f(x_{n+1}|x_1,...,x_n) \overset{(x_{n+1})}{\propto} \frac{f(x_{n+1}|\theta)}{f(\theta|x_1,...,x_{n+1})}$$

Alternatively, we could also evaluate, up to $\overset{(x_{n+1})}{\propto}$:

$$f(x_{n+1}|x_1,...,x_n) = \int f(x_{n+1}|\theta)f(\theta|x_1,...,x_n)d\theta$$

In our case, with $\sigma = 5, \mu_n = 86.9, \sigma_n = 1.56$, we have:

$$\begin{aligned}
f(x_{n+1}|x_1,...,x_n) &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x_{n+1}-\theta}{\sigma}\right)^2 \cdot \frac{1}{\sigma_n\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{\theta-\mu_n}{\sigma_n}\right)^2 d\theta \\
&\overset{(x_{n+1})}{\propto} \int_{-\infty}^{\infty} \exp -\frac{1}{2}\left(\left(\frac{x_{n+1}-\theta}{\sigma}\right)^2 + \left(\frac{\theta-\mu_n}{\sigma_n}\right)^2\right) d\theta \\
&= \int_{-\infty}^{\infty} \exp -\frac{1}{2}\left[\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}\right)\left(\theta - \frac{\frac{x_{n+1}}{\sigma^2}+\frac{\mu_n}{\sigma_n^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}}\right)^2 + (\sigma^2+\sigma_n^2)^{-1}(x_{n+1}-\mu_n)^2\right] d\theta \\
&= \exp\left(-\frac{1}{2}\frac{(x_{n+1}-\mu_n)^2}{\sigma^2+\sigma_n^2}\right) \cdot \int_{-\infty}^{\infty} \exp -\frac{1}{2}\left(\frac{\theta-\mu_{n+1}}{\sigma n+1}\right)^2 \\
&\overset{(x_{n+1})}{\propto} \exp -\frac{1}{2}\frac{1}{\sigma^2+\sigma_n^2}(x_{n+1}-\mu_n)^2
\end{aligned}$$

$$\implies X_{n+1}|x_1,...,x_n \sim \mathcal{N}(\mu_n, \sigma^2+\sigma_n^2)$$

Substituting our specific data, we have $\mu_n = 86.9$, $\sqrt{\sigma^2+\sigma_n^2} = 5.24$. So:

$$\mathbb{P}(86.9 - 1.96 \cdot 5.24 \leq X_{n+1} \leq 86.9 + 1.96 \cdot 5.24 | x_1,...,x_n) = 0.95$$

and the 95% posterior prediction interval for X_{n+1} is:

$$[76.6, 97.2]$$

This is very similar to the 95% frequentist prediction interval $[76.5, 97.1]$ for the same data obtained earlier.

If, instead, we had a prior pdf for Θ such that:

$$\Theta \sim \mathcal{N}(90, 0.1^2)$$

we would have $\mu_n = 89.99, \sigma_n = 0.0998$, as seen earlier, so the 95% posterior prediction interval is:

$$[89.99 - 1.96\sqrt{5^2+0.0998^2}, 89.99 + 1.96\sqrt{5^2+0.0998^2}] = [80.188, 99.792]$$

Due to the very low variance in the prior pdf, we would expect this prediction interval to be almost identical to the 95% prior prediction interval. This would be:

$$[90 - 1.96\sqrt{5^2+0.1}, 90 + 1.96\sqrt{5^2+0.1}] = [80.198, 99.802]$$

Related Questions

These questions are probably appropriate for the whole of section 2.
[DS12, section 7.2, exercises 1, 2, 3, 4, 5, 6, 7, 10, 11]

Aside: Summary of Content Covered Thus Far

Frequentist Method

- Needs a likelihood, given by:

$$\prod_{i=1}^{n} f(x_i|\theta)$$

- We constructed confidence intervals, conditioned on the parameter θ:

$$\mathbb{P}(L \leq \theta \leq U|\theta) = 1 - \alpha$$

  where L, U are functions of the data, e.g: $\bar{X}_n \pm z_{\alpha/2}, ...$

- We don't have a sense of having a distribution for a predictive quantity based on the data in the frequentist approach.

- We constructed frequentist prediction intervals:

$$\mathbb{P}(L \leq X_{n+1} \leq U|\theta) = 1 - \alpha$$

Bayesian Method

- Needs a posterior $\propto$ prior $\times$ likelihood:

$$f_{\alpha_0}(\theta|x_1, ..., x_n) \overset{(\theta)}{\propto} f_{\alpha_0} \prod_{i=1}^{n} f(x_i|\theta)$$

  Here, α_0 is a hyper-parameter, like μ_0, σ_0: It's a parameter that determines the distribution of other parameters.

- We constructed credible intervals, conditioned on the data:

$$\mathbb{P}(l \leq \Theta \leq u|x_1, ..., x_n) = 1 - \alpha$$

  where l, u are functions of the data and the hyper-parameters (so $\alpha_0, x_1, ..., x_n$).

- We looked at the posterior predictive distribution:

$$f_{\alpha_0}(x_{n+1}|x_1, ..., x_n) \overset{(x_{n+1})}{\propto} \frac{f(x_{n+1}|\theta)}{f_{\alpha_0}(\theta|x_1, ..., x_{n+1})}$$

- We constructed posterior prediction intervals, also conditioned on the data:

$$\mathbb{P}(l \leq X_{n+1} \leq u|x_1, ..., x_n) = 1 - \alpha$$

- We looked at the prior predictive distribution:

$$f_{\alpha_0}(x_{n+1}) \overset{(x_{n+1})}{\propto} \frac{f(x_{n+1}|\theta)}{f_{\alpha_0}(\theta)}$$

- We constructed prior predictive intervals:

$$\mathbb{P}(l' \leq X_{n+1} \leq u') = 1 - \alpha$$

  where l', u' are functions of α_0 only.

3 Conjugate Distributions

3.1 Sampling from a Normal with Known Variance

We saw that, in the context of the battery example, if $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ IID conditional on Θ, and $\Theta \sim \mathcal{N}(\mu_0, \sigma_0)$, then the posterior distribution for Θ is also normal:

$$\Theta|x_1, ..., x_n \sim \mathcal{N}(\mu_n, \sigma_n^2())$$

with simple formulae for μ_n, σ_n as functions of $\mu_0, \sigma_0, x_1, ..., x_n$ to update from the prior to the posterior.

The property that the posterior is from the same class of distribution as the prior is not unique to this scenario!

3.2 Sampling from a Bernoulli Distribution

3.2.1 Example: Clinical Trial

150 patients are randomly selected, and receive different treatments for depression: Imipramine (treatment 1), lithium carbonate (treatment 2), a combination of the two (treatment 3), or a placebo (treatment 4). The following table shows the number of patients who suffered a relapse within three years:

Treatment	1	2	3	4	Total
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

For now, we'll just focus on treatment 1, and try to apply Bayesian methods in order to make statistical inferences. Let Θ be the proportion of patients from the general population suffering from depression who do not relapse under this treatment.

The first step is to identify the statistical model: We will define the random variable as follows:

$$X_i = \begin{cases} 0, & \text{if patient relapses} \\ 1 & \text{if patient does not relapse} \end{cases}$$

$$X_i|\theta \sim \text{Bernoulli}(\theta)$$

(so $\mathbb{P}(X_i = 1|\theta) = \theta, \mathbb{P}(X_i = 0|\theta) = 1 - \theta$).

We will assume the following prior distribution for Θ:

$$\Theta \sim \text{Beta}(\alpha, \beta), \quad \alpha > 0, \beta > 0$$

Definition: Beta Distribution

Let $\alpha > 0, \beta > 0$. We say that θ follows a **Beta distribution**, $\text{Beta}(\alpha, \beta)$, when Θ has the following pdf:

$$f(\theta) := \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}, & 0 \leq \theta \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

N.b: $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the normalisation constant.
Note that, for the beta distribution, we have the following expressions:

$$\mathbb{E}(\Theta) = \frac{\alpha}{\alpha+\beta}$$
$$\mathbb{V}\text{ar}(\Theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

You don't need to remember these formulae, they'll be given to you if you have to use them in an exam setting.
Γ is the **gamma function**. We don't need to know much about it, other than that for $z > 0$, we have $\Gamma(z) > 0$. Again, if we need to use its properties in an exam, we will be given them. There is not a closed form expression for its entire domain.

Returning to the clinical trial example, let's look at the likelihood. As this is a discrete case, we'll have a pmf rather than a pdf, which by the IID cond on Θ assumption, is given by:

$$\begin{aligned} p(x_1, ..., x_n|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^y(1-\theta)^{n-y}, \quad \left(\textit{where } y := \sum_{i=1}^{n} x_i\right) \end{aligned}$$

So $Y_i = \sum_{n=1}^{n} X_i$ is a sufficient statistic, and since we have the data, we have a sufficient statistic!
Now let's find the posterior distribution:

$$\begin{aligned} f(\theta|x_i, ..., x_n) &\overset{(\theta)}{\propto} f(\theta)p(x_1, ..., x_n|\theta) \\ &\overset{(\theta)}{\propto} \begin{cases} \theta^{\alpha+y-1} \cdot (1-\theta)^{\beta+n-y-1}, & \theta \in [0,1] \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

So the posterior distribution must be a $\text{Beta}(\alpha + y, \beta + x - y)$ distribution.

3.2.2 Theorem

If $x_i|\theta \sim \text{Bernoulli}(\theta)$ and IID conditional on Θ, and $\Theta \sim \text{Beta}(\alpha_0, \beta_0)$ for some constants $\alpha_0, \beta_0 > 0$ then we've shown that $\Theta|x_1, ..., x_n \sim \text{Beta}(\alpha_n, \beta_n)$ where, with $y := \sum_{i=1}^{n} x_i$, we have:

$$\alpha_n := \alpha_0 + y$$
$$\beta + n := \beta_0 + n - y$$

3.2.3 Application of Bayesian Stats to Clinical Trial

In the clinical trial above, if $\Theta \sim \text{Beta}(1, 1)\ (= \text{Unif}(0, 1))$, then for the first treatment, since $n = 40, y = 22$, we have that:

$$\Theta|x_1, ..., x_n \sim \text{Beta}(1 + 22, 1 + 40 - 22) = \text{Beta}(23, 19)$$

For the posterior expectation/variance, we have:

$$\mathbb{E}(\Theta|x_1, ..., x_n) = \frac{1+22}{1+22+1+40-22} = \frac{23}{42} = 0.45$$

$$\mathbb{V}\text{ar}(\Theta|x_1,...,x_n) = \frac{23 \cdot 19}{42^2 \cdot 43} = 0.076^2$$

For comparison with the prior, we had $\mathbb{E}(\Theta) = \frac{1}{2}, \mathbb{V}\text{ar}(\Theta) = 0.29^2$. So the uncertainty has really decreased.

For the credible interval, unfortunately, there's no closed form when using the beta distribution.

3.3 Conjugate Families and Hyper-Parameters

Definition: Conjugate Families

Assume we have data X_i, IID conditional on the parameter Θ with pdf $f(x|\theta)$, or pmf $p(x|\theta)$. Let $f_\alpha(\theta)$ represent a family of densities for Θ, indexed by the *hyper-parameter* $\alpha \in \mathcal{A} \subseteq \mathbb{R}^k$. Then this family is said to be a **conjugate family of priors** for sampling from $f(x|\theta)$ if $\forall\, \alpha_0 \in \mathcal{A}$, and all $n \in \mathbb{N}$, and all possible samples $x_1, ..., x_n$, then $\exists\, \alpha_n \in \mathcal{A}$ such that:

$$f(\theta) = f_{\alpha_0}(\theta) \implies f(\theta|x_1,...,x_n) = f_{\alpha_n}(\theta)$$

i.e: Whenever the prior belongs to the family, then so does the posterior.

The benefit of using a conjugate family is that *updating the distribution is reduced to updating just the hyper-parameters!*

Let's consider, for example, the case using the normal distribution. $\mathcal{N}(\mu_0, \sigma_0^2)$, with hyperparameters $\mu_0 \in \mathbb{R}, \sigma_0 > 0$, is a conjugate family for $X_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$, assuming that σ is known.

$$\mu_n := \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$\sigma_n^2 := \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

Now considering a beta distribution, $\text{Beta}(\alpha_0, \beta_0)$ with hyper-parameters $\alpha_0, \beta_0 > 0$ is a conjugate family for $X_i|\theta \sim \text{Bernoulli}(\theta)$, with:

$$\alpha_n := \alpha_0 + \sum_{i=1}^{n} x_i$$

$$\beta_n := \beta_0 + n - \sum_{i=1}^{n} x_i$$

3.4 Sampling from an Exponential Distribution

Further Reading

[DS12]

3.4.1 Example: Lifetimes of Electrical Components

Consider a company selling electrical components. Each component is assumed to fail with probability θdt in any time interval $[t, t + dt]$ for small values of dt, where θ is an unknown parameter. So the component lifetimes are exponentially distributed, as follows:

$$X_i|\theta \sim \text{Exp}(\theta)$$

in which θ is called the rate parameter. The pdf for this distribution is given by:

$$f(x_i|\theta) = \begin{cases} \theta \exp(-\theta x_i), & x_i \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

We make the usual assumption that the X_i are IID conditional on Θ.

A priori, the company judges that $\mathbb{E}(\Theta) = 0.5, \mathbb{V}\text{ar}(\Theta) = 0.5^2$. We then observe: $x_1 = 3, x_1 = 1.5, x_3 = 2.1$. What can we say about future failures?

We need to identify:

1. A family of conjugate priors for this case (i.e: the case of sampling from an exponential distribution)
2. The posterior pdf of Θ
3. The posterior *predictive* pdf of Θ

Definition: Gamma Distribution

We say that $\Theta \sim \text{Gamma}(\alpha, \beta)$ for $\alpha, \beta > 0$ if the pdf is given by:

$$f(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, & \theta \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

This is the family of **gamma distributions**.
For a gamma distribution, we have the following results:

- $\mathbb{E}(\Theta) = \frac{\alpha}{\beta}$
- $\mathbb{V}\text{ar}(\Theta) = \frac{\alpha}{\beta^2}$
- If $\alpha = 1$, then we have $\text{Gamma}(1, \beta) = \text{Exp}(\beta)$

In this example, if $\Theta \sim \text{Gamma}(\alpha_0, \beta_0)$ for some known constants $\alpha_0, \beta_0 > 0$, then for $\theta \geq 0$, the posterior distribution follows:

$$\begin{aligned} f(\theta|x_1, ..., x_n) &\overset{(\theta)}{\propto} f(\theta) \prod_{i=1}^{n} f(x_i|\theta) \\ &\overset{(\theta)}{\propto} \theta^{\alpha_0 - 1} e^{-\beta_0 \theta} \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &= \theta^{\alpha_0 + n - 1} \exp -(\beta_0 + \sum_{i=1}^{n} x_i)\theta \end{aligned}$$

This is a $\text{Gamma}(\alpha_n, \beta_n)$ pdf, with:

$$\alpha_n := \alpha_0 + n$$
$$\beta_n := \beta_0 + \sum_{i=1}^{n} x_i$$

So, the family of Gamma distributions is *conjugate for exponential sampling.*

In this example, we need that $\mathbb{E}(\Theta) = \frac{\alpha_0}{\beta_0} = \frac{1}{2}$, and $\mathbb{V}\text{ar}(\Theta) = \frac{\alpha_0}{\beta_0^2} = \frac{1}{4}$, so we can very easily see that $\alpha_0 = 1, \beta_0 = 2$.

The posterior distribution will be a $\text{Gamma}(\alpha_n, \beta_n)$ distribution, with:

$$\alpha_n = \alpha_0 + n = 1 + 3 = 4$$
$$\beta_n = \beta_0 + \sum_{i=1}^{n} x_i = 8.6$$

Note that, from the posterior distribution, we therefore have $\mathbb{E}(\Theta|x_1, ..., x_n) = \frac{4}{8.6} \approx 0.47 \approx \frac{1}{2}$, so the expectation is still pretty close to the prior distribution. However, when considering the variance, $\mathbb{V}\text{ar}(\Theta|x_1, ..., x_n) = \frac{4}{8.6^2} = 0.054 \ll \frac{1}{4}$, so the variance has greatly decreased.

To find the posterior predictive pdf, we may use integration. For $x_{n+1} \geq 0$, we have:

$$\begin{aligned} f(x_{n+1}|x_1, ..., x_n) &= \int_0^\infty f(x_{n+1}|\theta) \cdot f(\theta|x_1, ..., x_n) d\theta \\ &\overset{(x_{n+1})}{\propto} \int_0^\infty \theta \exp(-\theta x_{n+1}) \theta^{\alpha_n - 1} \exp(-\beta_n \theta) d\theta \\ &= \int_0^\infty \theta^{\alpha_n} \exp(-(\beta_n + x_{n+1})\theta) d\theta \end{aligned}$$

The integrand here is the pdf of a Gamma$(\alpha_n + 1, \beta_n + x_{n+1})$ distribution, up to some normalisation constant, so the integral is the Gamma function over such a constant:

$$\begin{aligned} &= \frac{\Gamma(\alpha_n + 1)}{(\beta_n + x_{n+1})^{\alpha_n + 1}} \\ &\overset{(x_{n+1})}{\propto} (\beta_n + x_{n+1})^{-(\alpha_n + 1)} \end{aligned}$$

An alternative method for finding this result is to consider:

$$\begin{aligned} f(x_{n+1}|x_1, ..., x_n) &= \frac{f(x_{n+1}|\theta) f(\theta|x_1, ..., x_n)}{f(\theta|x_1, ..., x_{n+1})} \\ &= \frac{\theta e^{-x_{n+1}\theta} \beta_n^{\alpha_n} / \Gamma(\alpha_n) \cdot \theta^{\alpha_n - 1} e^{-\beta_n \theta}}{\beta_{n+1}^{\alpha_{n+1}} / \Gamma(\alpha_{n+1}) \cdot \theta^{\alpha_{n+1} - 1} e^{-\beta_{n+1}\theta}} \\ &= \frac{\Gamma(\alpha_n + 1)}{\Gamma(\alpha_n)} \cdot \frac{\beta_n^{\alpha_n}}{(\beta_n + x_{n+1})^{\alpha_n + 1}} \end{aligned}$$

This step is allowed because we have $\alpha_{n+1} = \alpha_n + 1, \beta_{n+1} = \beta_n + x_{n+1}$.

We then use a "magical" result that is left unproven for this course: the first term equals α_n:

$$f(x_{n+1}|x_1, ..., x_n) = \frac{\alpha_n \beta_n^{\alpha_n}}{(\beta_n + x_{n+1})^{\alpha_n + 1}}$$

This is the same result.

Note that the posterior predictive pdf found here is *not* exponential! It follows the **Lomax Distribution**: $X_{n+1}|x_1, ..., x_n \sim \text{Lomax}(\alpha_n, \beta_n)$. For the Lomax distribution, we have:

- $\mathbb{E}(X_{n+1}|x_1, ..., x_n) = \frac{\beta_n}{\alpha_n - 1}$ if $\alpha_n > 1$
- $\mathbb{V}\text{ar}(X_{n+1}|x_1, ..., x_n) = \frac{\beta_n^2 \alpha_n}{(\alpha_n - 1)^2 (\alpha_n - 2)}$ if $\alpha > 2$

In our case, the posterior predictive pdf of X_4 is:

$$f(x_4|x_1, x_2, x_3) \overset{(x_4)}{\propto} (8.6 + x^4)^{-5}$$

with:

$$\mathbb{E}(x_4|x_1, ..., x_3) = \frac{8.6}{3} = 2.87$$
$$\mathbb{V}\text{ar}(x_4|x_1, ..., x_3) = \frac{8.6^2 \cdot 4}{3^2 \cdot 2} = 4.05^2$$

Further Reading

[DS12, section 7.3] contains many more examples

Related Questions

[DS12, section 7.3, exercises 1 - 13, 17, 19, 20]
Extra exercises: [DS12, section 7.3, exercises 14, 15, 16, 23, 24]