Probabilities & Data
Week 9: Conditional & Bayesian

DR NICK FEWSTER-YOUNG

Topics ☺

- Probabilistic models usually include multiple uncertain numerical quantities. In this chapter we describe how to specify random variables to represent such quantities and their interactions.
- Bayesian Statistics
- We look at updating distributions, that is given prior knowledge or an assumption. We will update our knowledge based off a piece of sample information and create a Posterior Distribution using Bayesian Methods.
- We will look at the Normal Distribution process only!
- There will be simulations in R during the workshop.

Continuous Bivariate Variables

- As in the case of univariate (single) continuous random variables, we characterize the behaviour of several continuous random variables defined on the same probability space through the probability that they belong to unions of intervals.
- You can think of them as being defined on grids in 2 dimensions and lattices for more than 3 variables.
- The concepts are the same, the density function just becomes measuring the volume and its under 1.

Joint Cumulative Distribution

Definition: (Joint cumulative distribution function). Let $X, Y : S \to R$ random variables. The joint cumulative distribution function of X and Y is defined as

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

Conditional Probability

Definition: (Conditional Probability Function). If $F_{X,Y}(x,y)$ is differentiable, then the conditional probability function of Y given X is defined as

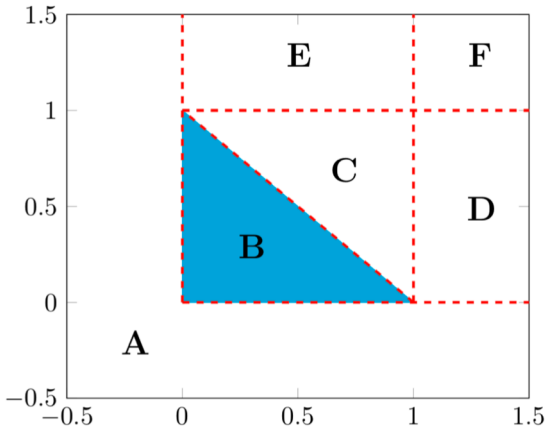
$$f_{Y|X}(y|x) \coloneqq \frac{f_{X,Y}(x,y)}{f_X(x)}, \qquad given \qquad f_X(x) > 0.$$

Definition: (Conditional Cumulative probability Function). If $F_{X,Y}(x,y)$ is differentiable, then the conditional probability function of Y given X is defined as

$$F_{Y|X}(y\,|x) \coloneqq \int_{\infty}^{y} f_{Y|X}(u|x)\,du\,, \qquad if \qquad f_X(x) > 0.$$

Example

Example (Triangle lake). A biologist is tracking an otter that lives in a lake. She decides to model the location of the otter probabilistically. The lake happens to be triangular as shown in Figure 3.1, so that we can represent it by the set



Thus the Lake $\coloneqq \{ x, y : x \geq 0, y \geq 0, x + y \leq 1 \}$.

The biologist has no idea where the otter is, so she models where it is by (x,y), which is uniformly distributed over the lake. Thus,

$$f_{X,Y}(x, y) \coloneqq 2 \quad for\ (x, y)\ in\ Lake.$$

Example (Otters)

The biologist is interested in the probability that the otter is south of x. This information is encoded in the cumulative distribution function below of the random vector, we just need to take the limit when y → ∞ to marginalize over y.

$$F_X\ (x) = \begin{cases} 0, & x < 0 \\ 2x - x^2, & 0 \leq x \leq 1 \\ 1, & x \geq 1 \end{cases}$$

and if we marginalise this then is (reverse integration = differentiation)

$$f_X(x) \coloneqq \begin{cases} 0, & x < 0 \\ 2 - 2x, & 0 \leq x \leq 1 \\ 0, & x \geq 1 \end{cases}$$

Bayesian Continuous Variables

- Recall when we looked at the discrete scenario that the theory was based around Bayes' theorem. Again we have a similar result stated as follows:
- **Theorem:** (Chain Rule / Bayes' continuous version)

$$f_Y(y)f_{X|Y}(x|y) = f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x).$$

Bayesian Process

1. We start with prior information, that is probabilities about a hypothesis or events we know!!
2. We compute the likelihoods, that is the process which is happens by given new sample information.
3. Combining the likelihoods /process and the prior information, we compute the predicative distribution.
4. Finally, we can compute the posterior (updated) distribution given the process and the new sample information.

Example

Example. Light bulbs are exponentially distributed.

Lifetime of each $Bulb \sim Exponential\ (\lambda)$

Test 5 bulbs and find lifetimes $x_1, \dots, x_5$.

I. Find the likelihood function.

Let $X_i \sim \exp(\lambda)$ the lifetime of the i bulb. The Likelihood = joint pdf (assuming independence):

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

Example. (recall)

- Bernoulli process with an unknown probability of success p.
- Can hypothesise that p takes any value in [0, 1].
- Imagine a ‘bent coin’ with probability p of heads.
- We could start with the prior information that the probability is p or a guess. Then we simulation or run the experiment as a Bernoulli process and if proportion of heads is not equal to p then the posterior distribution will give a new distribution for the prior guesses.

-- We have done the Discrete Case.

Example (Cont.)

Let $C_{0.25}$ stand for the hypothesis (event) that the chosen coin has probability 0.25 of heads. We want to compute $P(C_{0.25}|data)$

Method 1: Using Bayes' formula and the law of total probability:

$$
\begin{aligned}
P(C_{.25}|\text{data}) &= \frac{P(\text{data}|C_{.25})P(C_{.25})}{P(\text{data})} \\
&= \frac{P(\text{data}|C_{.25})P(C_{.25})}{P(\text{data}|C_{.25})P(C_{.25}) + P(\text{data}|C_{.5})P(C_{.5}) + P(\text{data}|C_{.75})P(C_{.75})} \\
&= \frac{(0.75)^2(1/4)}{(0.75)^2(1/4) + (0.5)^2(1/2) + (0.25)^2(1/4)} \\
&= 0.5
\end{aligned}
$$

Example: A Bayesian table

Method 2: Using a Bayesian update table:

hypotheses $\mathcal{H}$	prior $P(\mathcal{H})$	likelihood $P(\text{data}\|\mathcal{H})$	Bayes numerator $P(\text{data}\|\mathcal{H})P(\mathcal{H})$	posterior $P(\mathcal{H}\|\text{data})$
$C_{0.25}$	$1/4$	$(0.75)^2$	0.141	0.500
$C_{0.5}$	$1/2$	$(0.5)^2$	0.125	0.444
$C_{0.75}$	$1/4$	$(0.25)^2$	0.016	0.056
Total	1		$P(\text{data}) = 0.281$	1

Example 2 &3

2. Waiting time $X \sim \exp(\lambda)$ with unknown λ or Lifetime of a lightbulb. Can hypothesize that λ takes any value greater than 0.

3. Have Normal random variable with unknown μ and σ. Can hypothesisze that (μ, σ) is anywhere in $(-\infty, \infty) \times [0, \infty)$.

- These are both continuous distributions and we can now investigate how these change….hint its similar to the first one.

Normal Bayesian

- In this course for the continuous case, we will only discuss the Normal case. That is where the prior distribution and process is Normally Distributed.
- Thus, we will be talking about means and standard deviations a lot.
- This is a very attractive mathematical distribution due to its very useful properties in statistics.
- Recall that if our sample is large enough (n > 30) then the distribution type is irrelevant if we are dealing with means or averages. ☺ (Central Limit Theorem)
- We will continually be transforming the raw data to the standardize Normal distribution by

$$z = \frac{x - \mu}{\sigma}$$

Bayesian Statistics (Updating)

- Suppose we are dealing with Normal Distributions or means of sample size greater than 25 in this course.
- We will look at two applications (examples) to motivate and explain Bayesian Statistics for Normal Distributions.

Bayesian Statistics

Suppose that a retailer is interested in the distribution of weekly sales at a particular store. He is willing to assume that the random variable for weekly sales, $\tilde{x}$, is Normally distributed with an unknown mean and a known variance $\sigma^2 = 90000$. It should be stressed that this is a subjective assumption.

In assuming an independent Normal data-generating model, he is assuming that there is no trend or seasonal trend to worry about.

We start by illustrating the use of discrete prior distributions in conjunction with a continuous data process. Suppose that the retailer decides to consider only 5 potential values of the sample mean, 1100, 1150, 1200, 1250 and 1300.

Bayesian Statistics Example

Based on the experience with similar stores, he assesses the following prior probabilities:

$$P(\tilde{\mu} = 1100) = 0.15$$

$$P(\tilde{\mu} = 1150) = 0.2$$

$$P(\tilde{\mu} = 1200) = 0.3$$

$$P(\tilde{\mu} = 1250) = 0.2$$

$$P(\tilde{\mu} = 1100) = 0.15$$

Perhaps he assessed the prior probabilities as we did previously in regards to selling cars in relation to odds or rates.

The retailer decides that he would like to obtain more information about the sales of the store, so he takes a sample from the past sales records of the store, assuming that weekly sales from different weeks can be regarded as independent.

Suppose he takes a sample of size 60, and finds the average weekly sales to be $\bar{x} = 1240$. Prior to seeing this sample, the conditional distribution of $\bar{x},\ given$ the mean μ was Normally distributed with mean μ and variance

$$\frac{\sigma^2}{n} = 1500.$$

This is known as the prior sampling distribution of means.

So far, we have the prior information and the process for which governs our new information.

Likelihoods

We can compute the likelihoods understanding the process is Normally distributed. This does require us to convert the raw scores to standard scores (first year statistics)

$$f(1240 \mid \mu = 1100, \sigma/\sqrt{n} = 38.73) = f\left(\frac{1240-1100}{38.73} \mid 0,1\right)$$
$$= f(3.61 \mid 0,1) = 0.0006$$
$$f(1240 \mid \mu = 1150\ \sigma/\sqrt{n} = 38.73) = f\left(\frac{1240-1150}{38.73} \mid 0,1\right)$$
$$= f(2.32 \mid 0,1) = 0.0270$$
$$f(1240 \mid \mu = 1200, \sigma/\sqrt{n} = 38.73) = f\left(\frac{1240-1200}{38.73} \mid 0,1\right)$$
$$= f(1.03 \mid 0,1) = 0.2347$$

Continued

$$f(1240 \mid \mu = 1250, \sigma/\sqrt{n} = 38.73) = f\left(\frac{1240 - 1250}{38.73} \mid 0,1\right)$$

$$= f(-0.26 \mid 0,1) = 0.3857$$

$$f(1240 \mid \mu = 1300, \sigma/\sqrt{n} = 38.73) = f\left(\frac{1240 - 1300}{38.73} \mid 0,1\right)$$

$$= f(-1.55 \mid 0,1) = 0.1200$$

Posterior Distribution

Note that since $\frac{\sigma}{\sqrt{n}} = 38.73$ is the same for all values of $\tilde{\mu}$!!

Remember we are interested in $\tilde{\mu}$! By an application of Bayes' theorem here, we can update the prior information with the new posterior information!

$$P(\mu = 1100 \mid \bar{x} = 1240) = \frac{f(\bar{x}=1240|\mu=1100,\sigma/\sqrt{n})\, P(\mu=1100)}{P(\bar{x}=1240)}$$

$$P(\mu = 1150 \mid \bar{x} = 1240) = \frac{f(\bar{x}=1240|\mu=1150,\sigma/\sqrt{n})\, P(\mu=1150)}{P(\bar{x}=1240)}$$

$$P(\mu = 1200 \mid \bar{x} = 1240) = \frac{f(\bar{x}=1240|\mu=1200,\sigma/\sqrt{n})\, P(\mu=1200)}{P(\bar{x}=1240)}$$

$$P(\mu = 1250 \mid \bar{x} = 1240) = \frac{f(\bar{x}=1240|\mu=1250,\sigma/\sqrt{n})\, P(\mu=1250)}{P(\bar{x}=1240)}$$

Posterior Distribution (cont.)

$$P(\mu = 1300 \mid \bar{x} = 1240) = \frac{f(\bar{x}=1240|\mu=1300,\sigma/\sqrt{n})\, P(\mu=1300)}{P(\bar{x}=1240)}.$$

We still need to calculate the $P(\bar{x} = 1240)$, this is the predicative probability for the introduced sample information.

By the Law of Total Probability, this is just the sum of all the numerators above in the calculations of the posterior probabilities.

We can use to calculate this!

Bayesian Table

A table of posterior and prior and likelihood probabilities.

$\tilde{\mu}$	Prior Probability	Likelihood	Prior Prob x Likelihood	Posterior
1100	0.15	0.0006	0.00009	0.001
1150	0.2	0.0270	0.0054	0.032
1200	0.3	0.2347	0.07041	0.412
1250	0.2	0.3857	0.07714	0.450
1300	0.15	0.1200	0.018	0.105

- **Exercise**: In **R**, let's create some probability plots!!

The plots: Conclusions

- We notice that the first plot of the prior distribution is lovely and symmetric!
- The second plot is not so symmetric, the posterior distribution and we can see with the added information of a sample – how it has changed the distribution!
- Very cool!
- We did sort of make an assumption which was very valid based off the sample size which is the prior distribution is technically discrete but we approximated it to be Normal!
- What this means that if we know what the process is and the prior information. Then we can update our prior knowledge! We could continue this process over and over again, each month and this produces the basics for Bayesian Algorithm for MACHINE LEARNING.

Continuous Prior and Continuous Process

- Now suppose everything is Continuous!
- This means we are given the original distribution (Normal) with a mean μ, and standard deviation σ.
- In the case, where a sample is given to us with a sample mean $\bar{x}$ and size n. Then the posterior distribution is

$$Normal\left(\bar{x}, \frac{\sigma}{\sqrt{n}}\right)$$

where $\bar{x}$ acts as *a point estimate*.

- Now, the other common case is we may be given a "*credible interval*" which is a confidence interval where the estimated mean is believed to be contained - especially from simulated data.

Example

Suppose a new store opens with a ticketing system to monitor how many enter the store and allows us to know how many staff are required to serve customers.

For the past month, data was collected for 25 work days and showed a sample mean of 100 customers per day. Suppose that the ticketing system/customers attending has a Normal Distribution with standard deviation of 30 customers per day.

- Based on this observed information presented to us, what is the posterior distribution of the mean of customers per day?

Bayesian Posterior (without prior/known sample mean)

- If the prior distribution is not known and we take the sample mean as the prior then we obtain the following result for the posterior distribution
- This is just the CLT at work where the sample mean is a point estimate to the true mean and is working behind the scenes! But yes, this is Bayesian – interesting!
- **Theorem :** If the process is Normally distributed and the sample mean given by $\bar{x}$ and the standard deviation is given by σ, then the posterior distribution is described by a Normal distribution:

$$Normal\left(\bar{x}, \frac{\sigma}{\sqrt{n}}\right).$$

Thus, $\bar{x}$ becomes a point estimate for the true mean μ.

Note: σ can be replaced by $s = sample\ standard\ deviation$.

Example (cont.)

Therefore, the posterior distribution of the mean of customers per day is given by

$$N\left(100, \frac{30}{\sqrt{25}}\right) = N(100, 6).$$

The probability that μ will be between 90 and 110 customers can be calculated using the Normal Probability table and by standardizing.

$$P(90 \leq \mu \leq 110) = \Phi\left(\frac{110 - 100}{6}\right) - \Phi\left(\frac{90 - 100}{6}\right) = \Phi(1.67) - \Phi(-1.67)$$
$$= 0.9525 - 0.0475$$
$$= 0.9050.$$

The probability the true mean is between 90 and 110 is 90.50% .

Example.

A common question is regarding probabilities where the customers pour in to the store and it becomes unmanageable by staff.

Suppose that it becomes unmanageable when there is more than 110 customers, thus the probability based off our sample for a **single** day:

$$P(X > 110) = P\left(Z > \frac{110 - \bar{x}}{\sqrt{\sigma^2 + \frac{\sigma^2}{n}}}\right) = \mathrm{P}\left(\mathrm{Z} > \frac{10}{30.60}\right) = 1 - \Phi(0.33) = 1 - 0.6293 = 0.3707$$

Thus, there is a 37.07% chance that they will be unmanageable on a given day.

Note that the additional of $\frac{\sigma^2}{n}$ which takes into account for a single day in the sample.

Credible Intervals

Now suppose that before the store opens for customers, we run a simulation to predict the number of customers per day to coordinate the number of staff needed.

This produces a credible interval or confidence interval based off the simulation alone of say

$$(\bar{x} - z\times\frac{s}{\sqrt{n}}, \bar{x} - z\times\frac{s}{\sqrt{n}}) = (95 - 9, 95 + 9) = (86,104)$$

with a 95% level of confidence (z=1.96).

This is telling us that the mean number of customers is between 86 and 104 on a particular day.

Credible Intervals

- This is just an estimate for the sample mean and works as our new information.
- Therefore, if we know the prior distribution then we can use Bayes theorem to determine the new distribution (posterior distribution) under the effects of the simulation!
- **Theorem.** If the prior distribution of the mean is assumed to be Normal $N(\mu, \frac{\sigma^2}{n})$, and a sample (simulation) is introduced which has a sample mean $\bar{x}$ and standard deviation s, size n_1, then the prior distribution of μ is

$$\mathrm{N}\left(\bar{x}, \frac{s}{\sqrt{n_1}}\right)$$

And the posterior distribution of μ is Normal with

$$\tilde{\mu} = \frac{\mu\left(\frac{s}{\sqrt{n_1}}\right)^2 + \bar{x}\left(\frac{\sigma}{n}\right)^2}{\left(\frac{s}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma}{\sqrt{n}}\right)^2}, \qquad \text{and} \qquad \tilde{\sigma} = \sqrt{\frac{\left(\frac{s}{\sqrt{n_1}}\right)^2\left(\frac{\sigma}{\sqrt{n}}\right)^2}{\left(\frac{s}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma}{\sqrt{n}}\right)^2}}$$

Credible Intervals

This produces a credible interval or confidence interval based off the simulation alone of say

$$(\bar{x} - z\times\frac{s}{\sqrt{n}}, \bar{x} - z\times\frac{s}{\sqrt{n}}) = (95 - 9, 95 + 9) = (86,104)$$

with a 95% level of confidence (z=1.96).

Thus for our example, by calculating s from the expression above, the mean number of customers is distributed $N(95, 3.06)$. Now apply the previous theorem to calculate the posterior distribution of μ

Turn over slide

Posterior Distribution Example

$$\tilde{\mu} = \frac{\mu\left(\frac{s}{\sqrt{n_1}}\right)^2 + \bar{x}\left(\frac{\sigma}{\sqrt{n}}\right)^2}{\left(\frac{s}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma}{\sqrt{n}}\right)^2} = \frac{100 \times 3.06^2 + 95 \times 6^2}{3.06^2 + 6^2} =$$

$$\text{and} \quad \tilde{\sigma} = \sqrt{\frac{\left(\frac{s}{\sqrt{n_1}}\right)^2\left(\frac{\sigma}{\sqrt{n}}\right)^2}{\left(\frac{s}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma}{\sqrt{n}}\right)^2}} = \sqrt{\frac{(3.06)^2 6^2}{3.06^2 + 6^2}} =$$

Finish off!

We can plot these distributions in **R.**

Next week, we will look at applications in Stock Market and Brownian Motion.

Have a good weekend!