Bayes class 1

Sean van der Merwe

Helping people talk less nonsense is one of the key roles of a statistician.

# Who are you?

What would you like to be called and how would you like to be contacted?

# Models

What is a model (in general)?

Models are how our brains work. The world is complicated, and we perceive but a tiny fraction of it at once, yet we need to interact with it a great deal as we live our lives.

You would not be able to live if you didn’t constantly classify, discriminate (in a good way), and especially predict based on the limited inputs from your senses.

# Good models

Better models mean less stress for everyone (less difference between expectation and reality).

Write down a few properties of a good model. Ask the people near you what they think.

The most important property of a good model is that it works for new data that the model hasn’t seen. New data is not the same as existing data and brings additional uncertainty with it.

# Bayes

**Your brain is Bayesian**: it takes what it perceives and compares it to what it knows, then updates what it knows based on the new information by changing the connections. If you are very certain about what you know then it takes a lot to change your mind; while if you are aware of how little you actually know then it is easier to consider and remember new information. <https://www.nature.com/articles/s41562-021-01247-w>

How would you define a probability?

Knowledge of the past is only useful if it helps us to better approach our future. Prediction can be seen as more important than history, but prediction should flow from history in a sensible way, including both the uncertainty in our view of the past and the uncertainty inherent in moving from the past to the future. Bayes analysis provides a workflow for doing just this.

# Statistical Modelling

In statistical modelling the goal is to make information and knowledge shared more accurate and reliable, allowing for better decision making. Statistical modelling is thus incredibly important for everyone, and becoming more so in the 4th industrial revolution. All humans perform statistical modelling along with general modelling all the time, but all of us have room for improvement.

In order to improve we must expand our statistical modelling toolboxes. The goal of the assessments in this course is to motivate you to gather more tools by testing you on the use of specific tools/techniques.

# Distributions

One of the basic but incredibly important tools for a statistician is statistical distributions. Distributions help us to describe patterns of variation in numbers (or numeric summaries of non-numeric information). Many distributions have a strong theoretical basis for both form and function, and understanding this basis helps us to use them appropriately.

For example, the Normal distribution flows from the Central Limit Theorem (among other places). The CLT says that as your sample size grows the average of a measure will behave more and more like a Normal distribution, regardless of the distribution of the actual measurements (given that certain assumptions hold). Thus, the Normal distribution is often a natural choice for things that are sums of lots of other things.

What other distributions naturally describe sums of things?

Even if the theoretical basis is not of interest, the properties are important because they help us to select an appropriate choice for modelling. If your measure of interest can, and is likely to, take on negative values then using a strictly positive distribution to describe that measure is illogical and will result in nonsensical results.

Distributions are fundamental in formal statistical modelling. In the simplest case a distribution is a model in itself, with observations assumed independent and identically distributed according to a distribution with some (usually unknown) parameter values.

In general, a statistical model begins by specifying that the observations of interest follow a specific distribution where the unknown parameter values depend on known information (directly or indirectly).

**Thus, knowing the basic distributions deeply, without having to even think about it, is essential to being a good statistical modeller.**

Consider the set of axes given. Add a rough line showing the density function of each of these continuous densities, and remember to label each line.

• Standard Normal density

• Standard t density with about 5 degrees of freedom

• Standard Exponential density

• Gamma density with shape parameter 2 and scale parameter 1

• Beta(1,1) density (standard Uniform)

• Beta(2,2) density



# Class challenge

In your first year of statistics you probably encountered some variation of the coupon collection problem. The standard form is that there are $m$ variations of an object, you obtain objects at random one at a time with equal probability, so how many objects would you expect to have to obtain in order to get a complete collection? This form of the problem has an analytical solution (look it up for yourself).

Consider now the more general problem, which I will call the loot box problem: Every time you buy/win a loot box you obtain an item at random (or multiple items, but that doesn’t change the problem meaningfully). The probabilities of the items are far from equal and you are usually most interested in specific rare items.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Item | A | B | C | D | E | F | G | H | I |
| Prob | 0.3 | 0.2 | 0.15 | 0.1 | 0.1 | 0.05 | 0.04 | 0.03 | 0.02 |

Suppose that, in order to complete a basic set, you require 2 of Item E, 1 of Item F, and 1 of Item G.

What is the expected number of loot boxes you need to open in order to complete your set? What is the number of loot boxes you need to open in order to have a 95% or higher probability of completing your set?

The most efficient way to answer these questions is simulation. We can simulate our scenario a large number of times and use the frequentist definition of probability by counting successes over trials. With fixed probabilities (as required by law in many countries, although difficult to test) this is surprisingly easy.

First write down your initial guesses below (you may use a calculator, but maximum 90 seconds) then ask your lecturer to show you how to simulate better approximations.

# Fun final thought

It might be ironic that the frequentist definition of probability is used to great effect to calculate Bayesian probabilities flowing from Bayesian models because Bayesian models are generally fit via simulation and simulation fits the assumption of ‘a large number of trials’. This is in contrast to the common frequentist use of the definition, where the ‘large number of trials’ is hypothetical and doesn’t hold in practice.