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 is somewhat linked to a Central Limit Theorem. There are theorems in statistics that say 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 additive combinations of lots of other things.

What other distributions naturally describe sums of things?

Even if the theoretical basis of a distribution 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 can result in nonsensical results.

Distributions are fundamental in parametric 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 parametric 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 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 a solution in simple mathematical expression terms (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 guess for the expected number then ask your lecturer to show you how to simulate better approximations.

Bayes class 2

Sean van der Merwe

# The Bayesian model flow

In this class we are going to look at how Bayesian models naturally flow in a mathematical sense.

# The properties of the data

Before we collect data we already know a lot about it, often enough to build a useful model.

Suppose you plan to observe the time that people stand in line at the supermarket. Perhaps a customer complained but the manager says that the customer is lying, and head office is paying you to sort it out statistically. What can you say about the observations right now?

# Distribution

What are some distributions that might work for this problem?

There are much better options, but we are going to use a very simple example for today, specifically a [This might be appropriate if each customer has 5 people in front of them with independent serving times.]

We need the density function, list some places where you might find it, then write it out.

# Likelihood

To keep things simple, we will assume (incorrectly) that observations are independent and identically distributed.

Likelihoods usually work better on the log scale, so the procedure is:

1. Write down the **log density** given a single observation ; 2. **Sum** over

Constants are used to capture the terms that are not of interest, in this case terms that contain no parameters.

# Priors

The prior sums up everything we know about the **parameters** *before* we collect any data.

There are many kinds (categories) of priors that are used in Bayes models, can you name a few?

For this simple problem we will use an easy prior:

# Hierarchy

In Bayes we usually write models in hierarchical notation, meaning in layers:

# Posterior

The posterior distribution is proportional to the prior times the likelihood. Alternatively, the log posterior is equal to the log prior plus the log likelihood plus a constant.

Write down the log posterior for this problem now:

Which leads to . After class check that you can get to this result yourself.

# Predictions

To make good predictions for new data we must consider as much uncertainty as is practically feasible. This includes at least the uncertainty in the observations and the uncertainty in the parameters of the model. However, when making predictions *we don’t really care about the values of the parameters*, only the information and uncertainty they carry, so we integrate over all possible values of the parameters to obtain a distribution for a new observation given all the old observations.

In this case,

Note that in almost all real problems you will find that the integration cannot be performed explicitly. Luckily this integral can be easily approximated via simulation techniques. *For example, ask your lecturer to show the approximate distribution a random future customer and of the maximum wait time of 10 random future customers.*

**The beauty of the simulation approach is that you only need to simulate once and then you can answer as many questions as you want.** Suppose after presenting your results someone asks you, “What is the probability of waiting more than 12 minutes?” This question can be answered easily using the same simulations as before by counting the proportion that meet the condition.

# Improvements

As the course progresses, we will learn how to use more flexible models with more realistic assumptions, and thus make better inferences and predictions. However, the principles do not change and should be learned right now.

# Fun final thought

It might be ironic that the frequentist definition of probability is manipulated 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 (often referring to an infinite replication of an experiment) and doesn’t hold in practice.

Bayes class 3

Sean van der Merwe

# Simulation

By now you should be familiar with the basic concepts of stochastic simulation via pseudo random number generators. If not, read up on the fascinating history of both pseudo and physical random number generation.

What random number generation algorithm is used by statistical programs like R, SAS, etc.? What is the random number generation algorithm used in Excel called and what are its drawbacks?

Can you write down the theorem that helps us simulate from a general univariate distribution, given the pseudo uniform numbers generated by the above algorithms?

# The art of the simulation study

In this class we are going to look at how statistical modelling procedures are compared objectively, given a specific problem that these procedures are meant to address.

# The steps of the simulation study

## What to simulate?

The first step of a simulation study is to decide on what will be simulated. Usually this revolves around comparing a new approach to one or more existing approaches for a problem. It is almost always the same team that developed the new approach that end up doing the simulation study, usually with the goal of showing that their new approach has favourable properties. Where do the existing approaches come from?

## Details, sample sizes, and parameter values

Once it is established which approaches will be compared then there are often still tons of small details to decide upon. These might include sample sizes, model complexity, and especially parameter values. Can you think of a good way to choose parameter values for a simulation study?

## Simulating samples

This is where the coding starts. For each scenario, a number of samples are simulated given the chosen parameter values and stored.

## Pretending to forget

Before any modelling takes place, it is critical that the chosen properties and parameter values are put aside as if they are completely unknown. **Each sample must be viewed as if it was the only sample ever collected and you have no idea how it came to be.** Why is this step important?

## Applying the models

For a single given sample, each model is applied, and the results recorded. At this stage the results can only be checked for sanity – did the models run and do the results make any sense. Can you give an example of a sanity check?

## Oh yes, I just remembered the parameter values

Once the models are done fitting, we can compare the fits to the true parameter values that we chose at the start. We should at least try to be objective here.

The results should include fit statistics of many kinds. Can you give some examples of fit statistics that are used in simulation studies? Remember to ask those around you for more examples.

## Comparing approaches

The final step is to summarise the statistics across many samples, per approach. This enables us to compare approaches, which was the goal all along.

# Examples

When called upon, please explain a simulation study that you read about to the class. This could be a formal study in an academic journal, or perhaps something illustrative on a blog post. You may use up to 3 minutes and up to 1 slide.

Bayes class 4

Sean van der Merwe

# Prior distributions

In this class we are going to look at choosing prior distributions.

# The properties of the model

Models must be connected to the data generating process in order to be useful. However, once a model is decided upon, the parameters of the model are properties of the model, not the data. Parameters are ideally connected to some properties of the population, but not necessarily. Think of something in the real world that is sampled. It could be very small or very big, it could be something tangible or merely an idea/opinion, it could be a habit or a rare event. Pick a partner in the class and describe this thing to them, and how it is sampled. Now write down the thing they thought of:

For the sample they gave you, what is the population?

What are the properties of the population that you might be interested in AND possibly be able to estimate from the sample?

What prior information do you have about those properties and how might you capture that prior information in the form of distributions?

# Types of prior distributions

For each of the following distributions that are used as priors, say what type of prior you think it is and why [there are at least two correct answers for each, see if you can get all of them]:

# Example time

At this point the class will create a sample of data and try to model it in the Bayesian framework.

Bayes class 5

Sean van der Merwe

# Multivariate simulation

In this class we are going to look at issues that come up when doing multivariate simulation.

At all times we must keep in mind the goal of stochastic simulation:

To draw a representative sample from the distribution of interest

Just like with real experimental samples, the more complex the distribution we are drawing from the more samples are required in order to represent the full picture of the object of interest.

# Exploring the mountain

Use the following code in RStudio to follow along more easily with the explanations to come:

plotly::plot\_ly(z=~volcano) |> plotly::add\_surface()

In Bayes, the most common target of simulation is the posterior distribution of the parameters of a model. Given a decent sample size, this distribution will resemble a mountain (formally it approaches a multivariate normal as ). However, for a finite sample size, the mountain will only look smooth from a distance and have all sorts of interesting features up close.

The goal is to explore the mountain and understand all its quirks, including all the rubble at the bottom.

# Univariate simulation first

We have already discussed univariate simulation, here is a recap of some key points:

1. Computers simulate pseudo random uniform(0,1) numbers, as many as we need.
2. These numbers have the full appearance of independence – i.i.d.
3. Given uniform numbers we can get to any univariate distribution.
	1. For all distributions which have a closed form inverse CDF, we simply plug the uniform values into that inverse CDF.
	2. For other standard distributions there are tricks and approaches that produce neat and accurate simulations quickly, usually using simple transformations.
	3. For non-standard distributions there are various useful algorithms, such as sampling-importance resampling, acceptance-rejection, and discretisation.
		1. Discretisation is the simplest approach and always produces simulations quickly, but can be very inaccurate in some cases.

However, multivariate distributions don’t have an inverse CDF in the traditional sense, which means that we need entirely new approaches.

## Why do multivariate distributions not have inverse CDFs?

# Chains

With multivariate distributions in general there are no reliable approaches to produce i.i.d. samples, but this doesn’t matter. We don’t need i.i.d. samples, we need representative samples. A painter painting a mountain doesn’t generally use independent pencil dots, they use brush strokes, and yet they still capture the essence of the mountain in their painting. Thus, attention shifts to approaches of simulation that might not be independent but still able to correctly cover the target distribution.

Modern simulation approaches simulate chains with varying levels of autocorrelation, where each value simulated depends on 1 or more previous values, directly or indirectly.

Why does extreme autocorrelation pose a challenge when trying to draw a representative sample?

In extreme cases of autocorrelation we can discard every 2nd simulated vector and still retain 99% of the information. This process is called thinning and is most often required when simulating complex models via Gibbs sampling (as used by JAGS and BUGS). Stan does **not** require thinning.

## Burn in, warm up, and initial values

Simulating via chains requires that each chain be started at a valid point in the distribution. These points can sometimes be guessed by the software, but not always. Starting at a point too far away from the center of the target distribution will result in a complete failure to converge to the target distribution since you will be in a region where the distribution is almost flat. In short, you can’t explore a mountain if you start so far away that you can’t even see the mountain. Sometimes simpler parameter estimation approaches, like the method of moments or OLS regression, can help generate reasonable starting values.

Regardless, the simulation approach needs time to explore the distribution outline before simulating for real, and we don’t want the exploration to mess with the output, so we throw it away afterwards. When exploring a mountain, it is a good idea to first size it up and prepare properly, but the preparation time is not truly part of the journey.

## Convergence

Remember that in Bayes we are not interested in convergence in the mathematical sense (unless you just want the posterior mode). Instead, we are trying to get a picture of the whole distribution. Thus, we want to see our simulations converge in distribution.

We must check for two things: within-chain convergence and between-chain convergence.

Given a trace plot showing how the simulations move around, how would you assess within-chain convergence?

And what about between-chain convergence?

## Other checks

If you go a step further and simulate from the posterior predictive distribution then you can do many more checks, can you suggest any?

## Example

Ask the lecturer to show an example of a joint distribution, and how simulations approximate it.

Use the back of this page to make notes.

Bayes class 6

Sean van der Merwe

# Heteroscedasticity

There are many types of heteroscedasticity. In most undergraduate statistics programmes, students are taught to watch out for some of these, but seldom taught how to work with them. This is a serious shortcoming as understanding uncertainty is at the heart of what makes a good statistician. Furthermore, predictions from an unstable system are bound to lead to trouble.

## Systematic heteroscedasticity

The first case is where we have or something similar, where the change in variance is directly related to a change in something observed. This can often be addressed with a transformation.

What is usually a good transformation when ?

What is usually a good transformation when ?

These transformations are special cases of what general transformation?

## Specific heteroscedasticity

The opposite case is when each observation has different variance with no discernible pattern. This case can also present as excess kurtosis or extreme values in one or more directions.

One approach used to address such issues is to adapt the residual distribution, usually replacing a non-robust but powerful distribution (*e.g.* normal) with a more robust alternative (*e.g.* Student-t). The result is that outlying observations become less influential, which is useful when modelling central tendencies.

Another approach is switching from a parametric model to a non-parametric model. This is somewhat similar to the above in that we abandon the usual assumptions (*e.g.* normal) and use a more free specification for the residuals.

What are the downsides to changing the assumed residual distribution?

## Class based heteroscedasticity

This is where groups/classes/treatments/levels of observations have naturally different variation.

The standard approach to addressing this is to model the groups/classes/treatments/levels separately.

What does a separate regression for each group imply about the regression lines?

What is the problem with doing a separate regression for each group?

How do we get one regression line with different variances for each group? Give multiple approaches:

Simulate a sample of shoe sizes for boys and girls aged 9 to 18. You should have 5 for each age and gender, so 100 sizes in total in your sample. Both the sizes and their variances should depend on the age and gender. [Hint: a Poisson distribution with would probably work (barely).]

Now run an ordinary regression and look at the diagnostic plots. Can you describe all the ways in which the resulting patterns differ from what is usually expected?

Now fit a model that accommodates these issues to some extent. Are the coefficients closer to the true values you selected?

**In all cases remember that it is the conditional variance that matters, not the raw variance!**

Bayes class 7

Sean van der Merwe

# Model comparison

In a constructed experiment it may be feasible to have a single model that perfectly captures all relevant aspects of the data generating process and answers the research question. However, this is extremely rare, and even in the best-case scenario people may still need to be convinced that there isn’t a better way. In almost all real experiments the process of answering the research question involves comparing multiple feasible models.

More importantly, most data that is analysed is not from a constructed experiment. Exploring the data generating process properly requires looking at a problem from multiple angles using multiple models.

Summarise the reasons for comparing models in your own words:

## Parsimony as a substitute for generalisation

What does generalisation mean to you?

How does applying the principle of parsimony help with generalisation?

What approaches are available to help us apply the principle of parsimony objectively?

## Information criteria

List some useful information criteria that can be used with Bayes models and give the formula for at least one that is designed to work with Bayes models.

How are Bayes information criteria different to the kind used in frequentist models?

How does one interpret the value of an information criterion?

What are the assumptions of information criteria in general?

Now try to apply some of these approaches to a simulated problem and see whether they give the expected results.

Once you’ve learned about mixed effects models, come back to this worksheet and explain why there are special formulas for comparing mixed effects models:

Bayes class 8

Sean van der Merwe

# Frequentist Parameter Estimation

In the frequentist setting parameters have a fixed value, we observe one sample out of a hypothetical infinite set of possible samples, and we use the idea of sampling variation to try to say something related to the fixed value of the parameter.

For example, based on a sample I obtain a 95% confidence interval for the population mean as (1.2, 3.4). What does this interval say?

# Bayesian Parameter Description

In the Bayesian setting it is the sample that is fixed and we try to obtain as much information about the parameters as we reliably can from the sample. This is done by finding posterior distributions for the parameters. Once we have a distribution for a parameter we can create parameter ‘estimates’ in many ways, because there are many ways to describe or estimate the location of a distribution.

Why is ‘estimates’ in inverted commas here?

Consider the general problem of explaining the location or central tendency of a sample and realise that estimating a parameter value from a posterior is actually the same sort of problem, especially when working with simulations. Give a few measures of central tendency you know:

Some measures are ideal under particular loss functions. Connect 3 measures of central tendency of a posterior to their related loss functions:

# Bayesian Credibility Intervals

In the Bayesian setting the interpretation of an interval is much simpler: A 95% credibility interval covers 95% of a posterior distribution.

It does not say “that there is a 95% probability of the parameter value being in the interval” because in Bayes there is no “the parameter value”.

Credibility intervals can be created in many ways. The simplest is the ‘symmetric’ interval, which is symmetric in probability. If we want an interval covering proportion of the posterior then we estimate the intervals using quantiles (think ) and (think ). A 95% interval would then have 2.5% of the probability outside the left limit and 2.5% outside the right limit, leaving 95% in the inside of the interval.

An alternative is the Highest Density Interval (HDI) / Highest Posterior Density (HPD) interval / shortest interval. A shorter interval with the same coverage carries more information, so in a sense we can say that the best 95% interval is the shortest 95% interval. In theory this can be tough to find, but with a simulated sample it is straightforward: consider every viable 95% interval and use the shortest one.

# Objective Prior Derivation

In the absence of informative or expert prior information, it is often desired to appear as objective as possible in the choice of prior distributions. One might choose a prior based on a theoretically desirable property, rather than just convenience.

The first set of properties relate to parameter estimation. A prior might be preferred if it offered superior abilities to come close to parameter values and desired coverage in a simulated scenario that is similar to the real data scenario. For example, a 95% interval should ideally cover a selected value 95% of the time.

The second set of properties relate to information theory. An objective prior should minimise the information provided by the prior, and maximise the information provided by the sample. In the case of i.i.d. distributed observations we can formally derive such priors, can you give two examples with their algorithms?

Bayes class 9

Sean van der Merwe

# Fixed effects

Fixed effects have a closed set of category levels.

It is assumed that the entire population of relevant category levels are represented in the data.

Future data will only present as one of the studied category levels.

For example, a business might mark customers in their data base as Gender = “Male/Female/Other” knowing that a future customer will be easily placed in one of these categories, as easily as existing and past customers. They might also classify customers as Risk = “LowRisk/HighRisk” knowing that all customers (past, present, and future) can be classified into these categories deterministically if desired.

Revision: what would the design matrix look like for a regression on the interaction of these factors? Give column headings and the values for a customer that is Female and High Risk.

# Random effects

Random effects occur when categorical effects are a random sample from a population of possible effects. Random effects are incorporated into a model when future data may include categories not previously sampled.

For example, a business might manufacture their products from a standard base, say cheap wood, that they source from various suppliers. They might choose to model the suppliers as random effects if they are open to using new suppliers in the near future. Even if they don’t, the quality of the supplied wood may vary randomly from month to month without warning or signs, so the business might model “Month” as a random effect in their input model.

How would the model differ if they could measure the quality of the wood prior to accepting delivery?

# Mixed effects

Mixed effects models occur when the design matrix includes both fixed and random effects. Usually you would have a treatment variable that is a fixed effect, a subject variable that is a random effect, and confounding variables (possibly continuous) that may interact with either or both sets of effects.

For example, in medical trials the medicine/treatment is a fixed effect because the statistician is only interested in a closed set of medicines being studied and future people will also know exactly which medicine they are taking. Confounding variables such as blood pressure, gender, or BMI might be included in the model as they could interact with the medicine and change its effect. However, even after accounting for the confounding variables, each person still reacts differently to the medicine due to random aspects of their bodies that are not measured. Critically, it is the effect of the medicine on random future people (subjects) that we care about, not the set already observed.

Consider a different example: Suppose you are studying child behaviour and how it is affected by a specific set of teaching approaches. What would be possible fixed effects and random effects in this case? What would be the treatment and what would be the subject?

Fixed effects:

Random effects:

# Uncertain/Debatable effects

Sometimes an effect can be either a fixed or a random effect based on problem context.

For example, a business might initially consider the town/city of origin of a customer as a fixed effect when they operate in a closed area of the country, but then consider origin a random effect when they want to expand to new parts of the country. They could possibly revert to setting it as a fixed effect again when they are done expanding.

The key issue is that random effects can help to account for random variation in groups due to factors that we cannot measure.

# Including a nuisance effect

Consider again the examples discussed so far. In each case it might be convenient to simply ignore the random effect variables. Including ‘subject’ in the model makes it more complicated, and this can seem unjustified when we don’t actually care about any of the specific ‘subjects’.

When is it important to account for such nuisance factors in an experiment and when can we just ignore them as part of the random variation captured in the residuals?

Hint: How many measurements are being taken per subject? Is it the same for each subject?

Bayes class 10

Sean van der Merwe

# Censoring

Clinical trials and other experiments involving sick patients/animals/subjects tend to have multiple types of censoring.

In difficult studies such as cancer trials the best treatment is the one with the longest time to death. What types of censoring would you expect in such cases?

In happier experiments the best treatment is the one with the shortest time to recovery. What types of censoring would you expect in such cases?

In lab experiments we often have difficulty taking exact measurements. What are some types of censoring that might occur as a result?

In all cases of censoring there is a true value that is unobserved. How is a censored value different from a missing value?

How is a censored value different to a truncated value?

## Working with censored values in the dependent variable

When a value is observed precisely, we ask, “How likely was this exact value given the model?”

When a value is censored (or observed approximately) we ask, “How likely was it for the value to fall in the interval where it ended up, given the model?” Thus, instead of a density value we use a probability.

Bayes class 11: hypothesis testing and GoF

Sean van der Merwe

# Bayesian Hypothesis Testing

In the classical hypothesis testing framework:

We assume that the world is boring. There is no change, no difference, no impact, all coefficients are zero, all models are the same. We observe something interesting. We ask ourselves, “how likely is it to observe something at least this interesting under the assumption that the world is boring?”

If the answer is large then we don’t conclude anything and continue to assume our gloomy outlook.

If the answer is small then we take it as evidence against the null hypothesis of boringness and conclude that there is at least one interesting thing going on.

Problems:

Evidence against the null hypothesis is not necessarily evidence for the alternative. If the alternative is at least as unlikely to have produced the interesting result as the null then we are bound to make false conclusions.

It does not take practical significance into account at all. Statistical significance is established first and then practical significance is considered as an entirely separate step.

It does not account for any prior knowledge or similar tests already constructed. This can create bias in both directions very easily.

Can you think of a case where prior tests would favour the null and where it would favour the alternative?

Bayesian approaches:

Two popular Bayesian alternatives are:

1. Regions of posterior equivalence. This is where we explicitly bring in a measure of practical significance and calculate a posterior probability of being within or outside the region, or of posterior regions overlapping an interesting amount.
2. Model comparison. This is where we build models that fit each (of possibly many) reasonable hypotheses and see which is best supported by the data.

How would you test equality of means and equality of variances at the same time when your data is grouped?

# Goodness of Fit (GoF)

One of the popular classes of hypothesis tests is those that try to ask whether ‘a model fits the data’. At least that’s what some people would like to ask, but that is not actually a testable or sensible question. In reality, data comes from the true data generating process (complicated) while the model is much simpler and thus cannot truly ever fit the data perfectly.

A better question is, “Assuming that the model is an accurate approximation of the data generating process, how surprising is the discrepancy between the observed data and what would be expected based on model predictions?”

We would like to summarise this discrepancy using a p-value. What do we need in order to calculate a p-value?

## Independent and identically distributed observations

For distributions with no parameters, like the U(0,1), goodness of fit is very easy. There are many options for statistics (AD and KS are the most popular) and the distribution of that statistic is only dependent on sample size.

For distributions that are truly location-scale invariant we can often find statistics which have distributions that are also invariant to the location and scale of the original data. These statistics merely need to account for the sample size and the fact that the parameters must be estimated, not the actual values of the parameters. Examples include most normality tests (like Shapiro-Wilk) and exponentiality tests.

For most distributions though, goodness of fit statistics have distributions that depend on both the sample size and the values of the unknown parameters, meaning that they cannot be established exactly. This implies that we must use a parametric bootstrap procedure to get a p-value. So the procedure is then: 1. Estimate parameters, 2. Get statistic, 3. Simulate many new samples using parameter estimates, 4. For each new sample estimate those parameters, 5. For each new sample get the statistic, 6. Get proportion of new statistics that exceed (more extreme/surprising) than the original statistic (and call the final result a p-value).

How does this change in a Bayesian model fit?

## Regression models

In a regression model, every observation follows a different distribution. Sometimes people say that the residuals are i.i.d., but that only applies to the theoretical residuals, not the observed residuals. Testing the observed residuals in an ordinary regression model for normality actually tests for a lot of things, but normality is surprisingly low on the list of things it tests for. Thus, we need to adjust for this if we want to check goodness of fit.

Previously, people attempted using the parametric bootstrap approach for such problems too, but it is very computationally intensive, and unstable.

Luckily there is a clever new approach that works rather well: DHARMa. The principle is simple:

1. Get the predictive distribution for each observation (or a simulated sample from it).
2. Get the empirical CDF value of each observation according to its own distribution (proportion of simulations less than the observation).
3. Test whether these empirical CDF values come from a U(0,1) distribution (as they should if the model was a perfect fit).

While this approach works for both frequentist (e.g. GLM) models and Bayesian models, it flows naturally from any Bayesian regression model. It is essentially equivalent to the notion that x% prediction intervals should cover x% of the observations, regardless of the value of x.

Further, a U(0,1) qqplot can help identify where the model and the data most disagree, highlighting possible outliers. How might an outlier present itself on a qqplot?