SFB-1491 Graduate School: Astrostatistics

Section 2: Probability, Bayes, & MCMC

Dr Angus H Wright

2023-02-02

Section 2: Introduction

Section 2: Probability & Decision Making (Weeks 3-5)

Topics include:

- Decision theory
- Fundamentals of probability
- Statistical distributions and their origins

Scenario One: The Drunken Coin Toss

One friend announces to the group that he spent lockdown teaching himself how to predict the future.

He says that he can predict everyone's future, and that he can see the outcome of any event before it happens.

The Test: the group challenges him to correctly predict the outcome of a fair coin toss 10 times in a row.

Scenario Two: The Wine Critic

A second friend says that he spent the lockdown "learning how to be a wine critic".

It's possible that this is true, but it's also possible that he spent the year drinking cheap wine…

The Test: the group challenges him to correctly identifying whether an unlabelled glass of wine comes from a vineyard in France or Spain 10 times in a row.

Scenario Three: The Classical Pianist

A third friend says that she spent the lockdown re-learning how to play classical piano.

She says that she had learned to play classical piano as a child, and the long time in lockdown gave her an opportunity to renew her passion for playing.

The Test: the group challenges her correctly identify whether a piece of classical music is by Beethoven or Mozart 10 times in a row.

Sceptic or Believer?

Each of these scenarios present the person with a challenge.

Without focussing on the statistics (just yet), come up with a figure that demonstrates how likely **you think** the person is to succeed their challenge.

What probability do you think that the person has of succeeding their challenge? Remember that the challenges all state that they must guess correctly *all* 10 *times*. If they get *any* of the 10 guesses wrong, then they fail their challenge.

What does Angus think?

- I don't believe that the friend can see the future. At all. $P(\text{success}) \approx 0$.
- I think it's unlikely that the friend studied wine all lockdown, but I'm not completely closed off to the possibility… $P(\text{success}) \approx 0.5$.
- I think that getting back into piano sounds like something totally reasonable, especially if you had learned it originally as a child, so there's a good chance she will win the challenge. $P(\text{success}) \approx 0.9$

What happens in the challenges?

Scenario One: The Drunken Coin Toss

First up is your good friend the drunken fortune teller.

You take a coin from your own pocket, which you know is fair.

You toss the coin 10 times, and each time your friend guesses the outcome while the coin is in the air.

- **He guesses correctly all 10 times**

What happens in the challenges?

Scenario Two: The Wine Critic

The barman prepares 10 glasses containing wine from either France or Spain.

He numbers the glasses 1 to 10, and gives them to you to administer the challenge.

You give the glasses of wine one-by-one to your friend, and record whether he thinks the wine comes from France or Spain. You take his responses back to the bartender.

- **He has guessed correctly all 10 times**

What happens in the challenges?

Scenario Three: The Classical Pianist

The group goes through their music libraries and selects 10 songs that are written by either Mozart or Beethoven.

They make the test as difficult as possible by purposely avoiding any songs that are widely popular.

They play the 10 songs one at a time for your friend.

- **She guesses correctly all 10 times**

What do you think?

In all three challenges, our friends are victorious!

How do these observations align with what you have in your graphs? Or more importantly:

- **How does the outcomes of the challenges change your opinions?**

- 0 means that they're lying, and that they cannot:
  - see the future
  - tell the difference between wines
  - tell the difference between the classical pieces

- 1 means that they're telling the truth, and that they *can*:
  - see the future
  - tell the difference between wines
  - tell the difference between the classical pieces

Reminder of Set Notation

- A set X is defined as a collection of items, grouped by curly brackets:

$$X = \{x_1, x_2, \dots, x_n\}$$

- The size (or cardinality) of a set is given by $|X|$.
- A set with cardinality of 0 is the empty set, which has the symbol $\emptyset$.
- The data that exist outside of a particular set, called the "compliment" is given as:

$$X^c = \Omega \notin X$$

Events and Outcomes

Given an experiment, there can be a set of possible results.

These results are known as **outcomes**.

Every run of an experiment will produce one, *and only one*, outcome.

Sample Space

When we toss a coin, there are two possible outcomes: Heads (H) or Tails (T).

This is therefore the total set of available outcomes, known as the **sample space** (Ω).

$$\Omega = \{H, T\}$$

What about if we were to toss our coin twice?

Now: consider a game which only ends when you throw a head *and then* a tail. What is the sample space of outcomes of this game?

Probability derives from Outcomes

The available outcomes of our single coin-toss experiment are:

$$\Omega = \{H, T\}$$

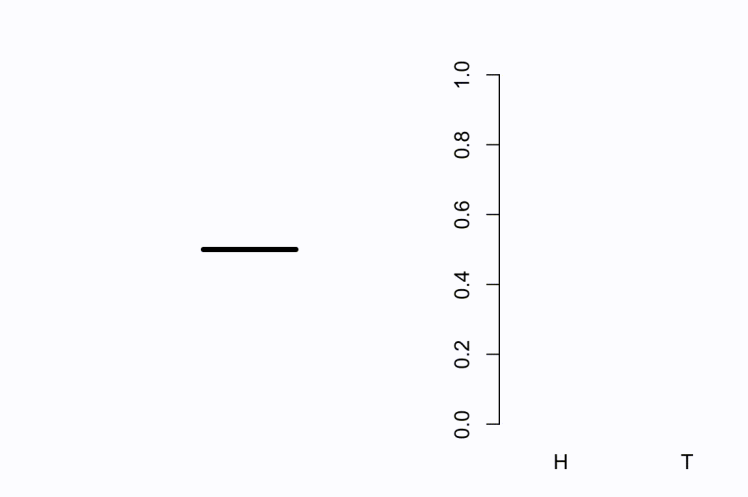
We were to perform this experiment N times (i.e. run N **trials**), and record the number of occurrences of each outcome A.

We can then observe the relative frequency of each outcome:

$$f_A = \frac{\#(A)}{N}$$

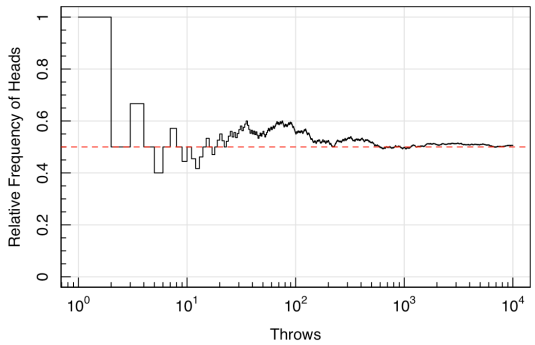
Coin Toss Simulation

We can perform this experiment using a simulated coin toss:



Probability derives from Outcomes

If we plot this as a running "Relative Frequency of Heads":



As the N becomes large, $P(A)$ tends towards 50%.

$$P(A) = \lim_{N \to \infty} \frac{\#(A)}{N}.$$

Our First Probability Laws

This gives us some information about probability already.

> - Because probability derives from relative frequency of observations:
>
> $$0 \leq P(A) \leq 1$$

When we look at the probability of all possible outcomes (i.e. $A_i \in \Omega$):

> - Ω is the sample space of all outcomes, so the sum of all relative frequencies must be 1: $\sum_{A_i \in \Omega} P(A_i) = 1$.

Events

An **event** is defined as a set of outcomes.

Take our example of the prime number roll on a dice:

$$\mathcal{E} = \{2, 3, 5\}$$

The probability of observing event $\mathcal{E}$ is the sum of the probabilities of observing each of the independent outcomes within the event:

$$P(\mathcal{E}) = \sum_{A_i \in \mathcal{E}} P(A_i)$$

If the event contains all possible outcomes (that is, the event space is the sample space: $\mathcal{E} = \Omega$), then we recover our earlier summation

$$P(\Omega) = 1.$$

Given the above two properties, observing any event that is not A, which is denoted as the compliment of A:

$$P(A^c) = 1 - P(A)$$

Events

Given the complementarity rule, if the event space contains no outcomes (that is, it is the empty set: $\mathcal{E} = \Omega^c = \emptyset$), then the probability of the event is 0:

$$\begin{aligned} P(\emptyset) &= 1 - P(\Omega) \\ &= 0. \end{aligned}$$

So, what is the probability of observing a prime number when we roll a fair die:

$$\begin{aligned} P(\mathcal{E} \in \{2,3,5\}) &= \sum_{A_i \in \mathcal{E}} P(A_i) \\ &= P(2) + P(3) + P(5) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ \therefore P(\mathcal{E}) &= 0.5 \end{aligned}$$

Rolling two Dice

Lets complicate the sample space by looking at the outcomes of rolling two dice at the same time, and summing together the results.

Each die has the outcomes $A = \{1, 2, 3, 4, 5, 6\}$. The sample space of the two-dice roll is the set of all possible **ordered combinations** or **permutations** of two draws from these values.

We can construct this sample space by hand:

$$\begin{aligned} \Omega = \{&1+1, 1+2, 1+3, 1+4, 1+5, 1+6, \\ &2+1, 2+2, 2+3, 2+4, 2+5, 2+6, \\ &\vdots \\ &6+1, 6+2, 6+3, 6+4, 6+5, 6+6\} \end{aligned}$$

Rolling two Dice

With two fair die, the probability of observing each of these outcomes is equal:

$$\begin{aligned} P(A_i) &= \frac{1}{|\Omega|} \\ &= \frac{1}{36} \end{aligned}$$

However, we wanted to calculate the *sum* of the dice. The summation doesn't distinguish between $1+4$ or $4+1$, it only matters that we have the *event* $\mathcal{E}_i = 5$.

What is the probability, then, of all distinct *events* in our two-dice roll?

Joint Probability

The **joint probability** of two outcomes is the probability of observing both outcomes at the same time. With our two dice, the joint probability of any two numbers was:

$$P(A \cap B) = P(A) \times P(B)$$

- Observations are **independent** if and only if

$$P(A \cap B) = P(A) \times P(B)$$

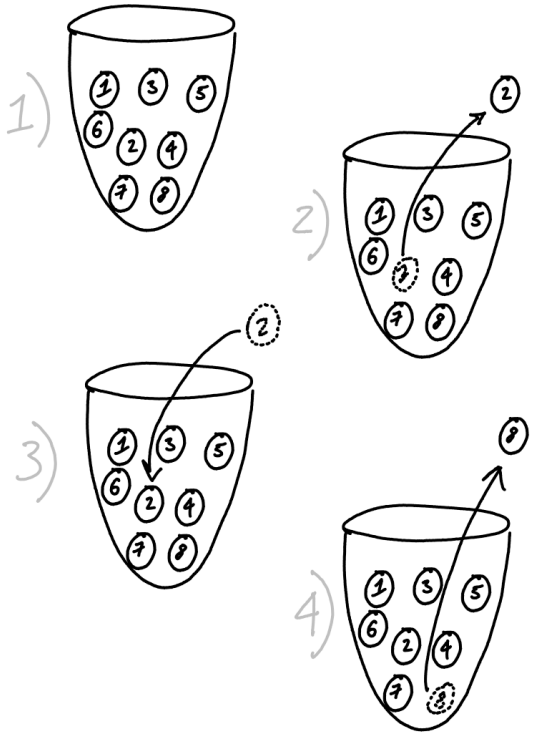
.

Independence and Non-independence

Independent events are extremely important in statistics, especially in the context of random variables (which we will discuss later in this section).

However non-independent events are also extremely important.

Independence and Non-independence



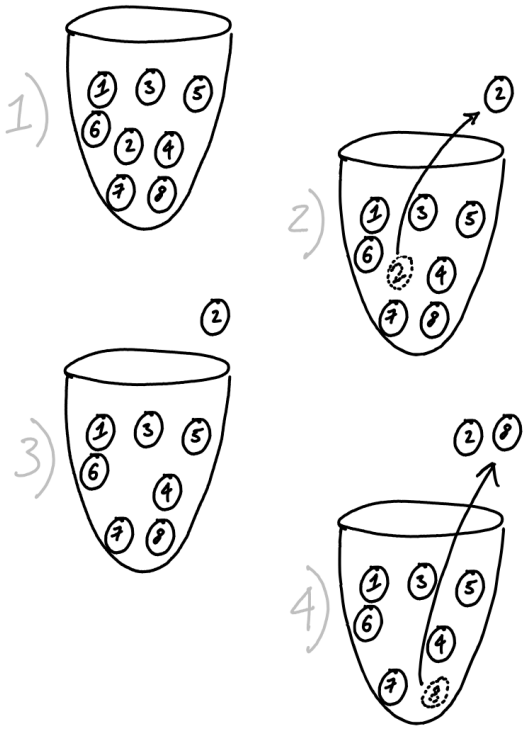
Independence and Non-independence

Here our draws are independent:

$$
\begin{aligned}
P(2 \cap 8) &= P(2) \times P(8) \\
&= \frac{1}{8} \times \frac{1}{8} \\
&= \frac{1}{64}
\end{aligned}
$$

However, how does this change if we choose not to replace the first ball that we draw?

Independence and Non-independence



Conditional Probability

Given a sample space Ω of outcomes and a collection of events, the probability of B conditioned on A is the probability that B occurs given that A has definitely occurred

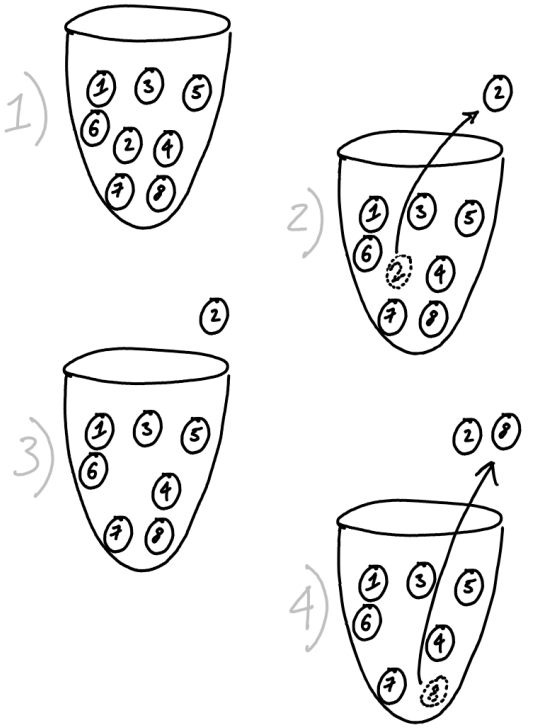
$$P(B|A)$$

With our urn example, for our second draw, what now want to know is the probability of observing an 8 *given that* we just observed a 2.

Said differently, the second draw computes the probability of observing an 8 *conditioned upon* our prior observation of a 2.

Conditional Probability

In this example we can compute the conditional probability logically:



$$
\begin{aligned}
P(2 \cap 8) &= P(2) \times P(8|2) \\
&= \frac{1}{8} \times \frac{1}{7} \\
&= \frac{1}{56}.
\end{aligned}
$$

Conditional Probability II

Suppose we draw two balls from our urn, with replacement. We want to calculate the probability of drawing two balls with a combined value greater than or equal to 10.

The “win” event space is therefore:

$$\begin{aligned}\mathcal{E} = \{&8+2, 8+3, 8+4, 8+5, 8+6, 8+7, 8+8,\\ &7+3, 7+4, 7+5, 7+6, 7+7, 7+8,\\ &6+4, 6+5, 6+6, 6+7, 6+8,\\ &5+5, 5+6, 5+7, 5+8,\\ &4+6, 4+7, 4+8,\\ &3+7, 3+8,\\ &2+8\}.\end{aligned}$$

Conditional Probability II

There are 64 possible ways of drawing 2 balls from a bag of 8 with replacement, which means that we have a $7/16$ chance of winning this game.

However, suppose now that we **know** that our first draw is an 8. How does this information influence our chance of winning?

If we first observe an 8, there are 7 subsequent draws which will earn us a victory:

$$\mathcal{E}|8 = \{8+2, 8+3, 8+4, 8+5, 8+6, 8+7, 8+8\}$$

Therefore the probability of winning given our first draw is an 8 jumps to $P(\mathcal{E}|8) = 7/8$.

Conditional Probability II

What about if we **know** that our first draw is a 2?

$$\mathcal{E}|2 = \{2+8\}$$

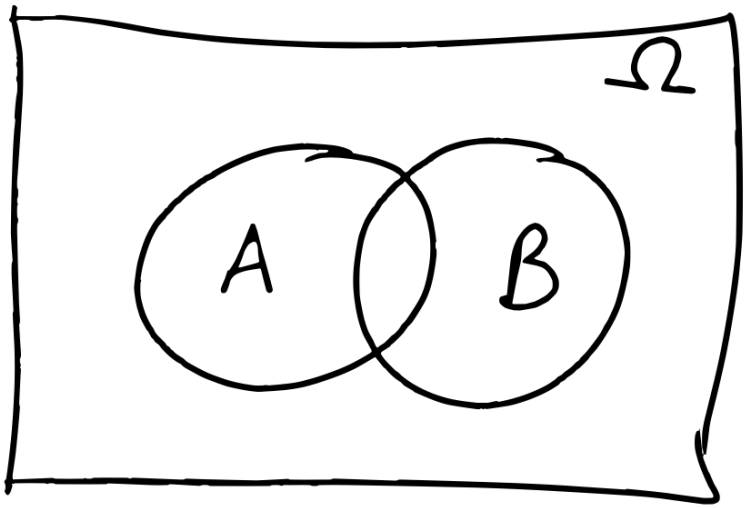
And so our probability of winning is a lowly $P(\mathcal{E}|2) = 1/8$.

So **event probabilities** can be wildly influenced by different conditionalisation.

Computing Conditional Probability

We now want to derive an expression for the conditional probability $P(B|A)$.

We can start with our venn diagram again:



Our conditionalisation means that we know our probability must reside within A. We want to know $P(B|A)$: the probability that it lies within *both* A and B. This is the *intersection* $B \cap A$.

Computing Conditional Probability

If the outcome lies in A, then it must fall within either $A \cap B$ or $A \cap B^c$. Therefore:

$$P(B|A) + P(B^c|A) = 1$$

Additionally, we can use our link between probability and relative frequency to our advantage.

If some outcome $C \cap A$ is k times more likely than $B \cap A$, then $P(C \cap A) = kP(B \cap A)$.

But this must be true regardless of whether A is observed first or not (the order of observation doesn't change the relative positions of items in our venn diagram). So $P(C|A) = kP(B|A)$. This means:

$$P(B|A) \propto P(B \cap A)$$

To determine the coefficient of proportionality (c) we can use the above expressions and find:

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

The intersection and the conditional probability are therefore very closely related.

The intersection probability has range $0 \leq P(B \cap A) \leq P(A)$, while the conditional probability has the range $0 \leq P(B|A) \leq 1$.

Computing Conditional Probability

This can be a guide as to how to think about the intersection (i.e. $P(B \cap A)$) and the conditional probability.

The former provides a probability in the *absence* of any additional information/observations.

The conditional probability, however, provides probability based on the *knowledge* that we have already made some observation.

Conditional Probability & Independence

Lastly, there is one additional (very important!) observation we can make. Given that the intersection of two probabilities is unchanged under ordering: $P(B \cap A) = P(A \cap B)$, this means that:

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

This turns out to be an extremely valuable relationship known as **Bayes Rule**.

Modelling conditionality & independence: Criminal Trial

You are on the jury of a criminal trial.

A criminal was identified by matching a sample of DNA (from a crimescene) to a database of many thousands of people.

The probability of **incorrectly** matching DNA to a random person is 0.01%, and such incorrect/chance matches are independent.

There are $20\,000$ people in the DNA database. The police find a match, and take the matching person to trial.

Modelling conditionality & independence: Criminal Trial

The prosecution stands before the jury and says they have *damning* evidence.

The probability of the DNA match being wrong is 0.01%, and so there is a 99.99% chance that this person is guilty.

What conditionality is important?

The prosecutor has quoted the following conditional probability:

$$P(\text{match}|\text{guilty}) = 0.9999$$

or equivalently

$$P(\text{match}|\text{innocent}) = 0.0001$$

Is that what we want to know?

The Prosecutors Fallacy

This trial is a ultimately a question of innocence and guilt.

So what we care about is:

$$P(\text{innocent} \mid \text{match})$$

A prosecutor at court presents evidence (i.e. the match) $\mathcal{E}$. They argue that the defendant is guilty because the probability of finding the evidence given innocence $P(\mathcal{E}|\mathcal{I})$ is small. But this is totally irrelevant. The real question is what is the probability that the defendant is innocent given the evidence: $P(\mathcal{I}|\mathcal{E})$.

The Prosecutors Fallacy

Using Bayes Rule:

$$P(\mathcal{I}|\mathcal{E}) = \frac{P(\mathcal{E}|\mathcal{I})P(\mathcal{I})}{P(\mathcal{E})}$$

So the assumption that $P(\mathcal{I}|\mathcal{E}) \approx P(\mathcal{E}|\mathcal{I})$ is only true if the probability of innocence $P(\mathcal{I})$ is equal to the probability of seeing the evidence $P(\mathcal{E})$.

The probability of seeing the evidence here is important, because it includes *all* the ways of producing a positive match:

$$\begin{aligned} P(\mathcal{E}) &= P(\mathcal{E}|\mathcal{I})P(\mathcal{I}) + P(\mathcal{E}|\mathcal{I}^c)P(\mathcal{I}^c) \\ &= 0.0001 \times \frac{N-1}{N} + 0.9999 \times \frac{1}{N} \\ &\approx 0.00015 \end{aligned}$$

So:

$$\begin{aligned} P(\mathcal{I}|\mathcal{E}) &= \frac{0.0001 \times \frac{N-1}{N}}{0.00014999} \\ &\approx 66.67\% \end{aligned}$$

Anomaly detection

The last discussion that we will have in this section on conditional probability is regarding the difficulty of anomaly detection: that is, why it's difficult to reliably detect rare events.

There are many cases in Astronomy and Physics where anomaly detection is interesting/desirable. Discovering new and rare phenomena is an obvious example, be they exotic particles in a detector or exotic transients in the universe.

Anomaly detection

When discussing accuracy of detection it is worth understanding the different types of result:

```
##          Compare         Reality.Positive        Reality.Negative
## 1  Measured-True            True Positive False Positive (Type I)
## 2 Measured-False False Negative (Type II)           True Negative
```

The “Types” are included because these names are sometimes used for specific types of failures.

The difficulty in anomaly detection arises because, as an event becomes rare, the accuracy of tests required to minimise false positives (Type 1) becomes prohibitively large.

Let’s consider two examples: detecting a common event, and detecting a rare event, with an experiment of fixed accuracy.

A common event

A decay process occurs in nature with probability 0.4. You have an experiment that detects this emission with a probability of 0.6, and produces a false positive with probability 0.1. What is the conditional probability that you witness a true decay *and* the experiment produces a positive detection?

```
##              Compare True.Decay No.Decay
## 1 Positive Detection        0.6      0.1
## 2 Negative Detection        0.4      0.9
```

Let d be the event that a true decay occurs, and $\mathcal{P}$ be the event that the experiment produces a positive detection.

$$
\begin{aligned}
P(d|\mathcal{P}) &= \frac{P(\mathcal{P}|d)P(d)}{P(\mathcal{P})} \\
&= \frac{P(\mathcal{P}|d)P(d)}{P(\mathcal{P}|d)P(d) + P(\mathcal{P}|d^c)P(d^c)} \\
&= \frac{0.6 \times 0.4}{0.6 \times 0.4 + 0.1 \times 0.6} \\
&= 0.8
\end{aligned}
$$

So your experiment only has to be approximately accurate to produce reliable detections when the event it common.

A rare event

Let’s now consider a similar scenario, except the probability of our decay occurring in nature is very small: $P(d) = 0.001$. The experiment, though, has improved to 99.9% accuracy, and only 1% false-positive rate.

```
##              Compare True.Decay No.Decay
## 1 Positive Detection      0.999     0.01
## 2 Negative Detection      0.001     0.99
```

Again we compute our probability that we actually detected the event:

$$
\begin{aligned}
P(d|\mathcal{P}) &= \frac{P(\mathcal{P}|d)P(d)}{P(\mathcal{P})} \\
&= \frac{P(\mathcal{P}|d)P(d)}{P(\mathcal{P}|d)P(d) + P(\mathcal{P}|d^c)P(d^c)} \\
&= \frac{0.999 \times 0.001}{0.999 \times 0.001 + 0.01 \times 0.999} \\
&= 0.09
\end{aligned}
$$

So despite our experiment becoming much much more accurate, the probability that we make a true detection is less than 10%. Said differently, 10 out of every 11 detections will be false.

Classical Inference

During the course so far, we have frequently described the concept of extracting information about a “population” Ω from a “sample” of observations drawn from that population.

This process is called statistical inference, and until this point we have been interested in “**Classical Inference**”.

A Collider Problem

Let’s take an example. A scientist wants to know (or rather estimate) the production rate of a particular exotic particle across a range of interactions.

It’s not possible to run all possible interaction combinations, so they instead run a sample of n interactions as their test.

If θ is true fraction of interactions that lead to a particle emission, then each interaction in the sample n, independent of all others, will be produce a particle with probability θ.

A Collider Problem

Let X be the random variable corresponding to the number of particles of interest produced by the the sample of n interactions.

The scientist will therefore use $X = x$ to draw some inference about the true underlying population parameter θ.

Such an inference could be of the form of:

- a **point estimate**: $\hat{\theta} = 0.1$
- a **confidence interval**: $\theta \in [0.08, 0.12]$ at 95% confidence
- a **hypothesis test**: reject the hypothesis that $\theta < 0.07$ at the 5% significance level
- a **prediction**: predict that 15% of future tests will produce the particle
- a **decision**: that this field of study isn’t worth persuing

A Collider Problem

Generally speaking, such inference will be made by specifying some probability model:

$$p(x|\theta)$$

which determines the probability of observing $X = x$ given a particular value of θ.

In our collider example, we have two possible outcomes per test (particle is produced or not produced). So the appropriate probability model is the binomial function:

$$X|\theta \sim \text{Bin}(n, \theta)$$

A Collider Problem

One method of estimating the value of θ is to maximise the likelihood of $X = x$ with respect to θ.

In the simplest terms we are finding the value of θ that is *most likely* to produce the observed value of $X = x$. In this case, we have the **maximum likelihood** estimate of θ.

A Collider Problem

Let's say we observe $x = 10$ given $n = 10$. Every interaction produces our particle.

Given the binomial function:

$$\begin{aligned}
\text{Bin}(n = 10, \theta = 0.0) &= 0.00\% \\
\text{Bin}(n = 10, \theta = 0.1) &= 0.00\% \\
\text{Bin}(n = 10, \theta = 0.2) &= 0.00\% \\
\text{Bin}(n = 10, \theta = 0.3) &= 0.00\% \\
\text{Bin}(n = 10, \theta = 0.4) &= 0.01\% \\
\text{Bin}(n = 10, \theta = 0.5) &= 0.10\% \\
\text{Bin}(n = 10, \theta = 0.6) &= 0.60\% \\
\text{Bin}(n = 10, \theta = 0.7) &= 2.82\% \\
\text{Bin}(n = 10, \theta = 0.8) &= 10.74\% \\
\text{Bin}(n = 10, \theta = 0.9) &= 34.87\% \\
\text{Bin}(n = 10, \theta = 1.0) &= 100.00\%
\end{aligned}$$

So the maximum likelihood is $\theta = 1$, and that is our best estimate of the value of θ.

In this case, the parameter θ is being treated as a **constant**. This is the cornerstone of classical statistical theory.

Bayesian Inference

The fundamental difference between Classical and Bayesian statistics is that:

- in Bayesian statistical inference, θ is treated as a **random** quantity.

This means that inference can be made by analysing probabilities alone.

Bayes Rule

Bayesian Inference is based on the concept of the **posterior probability distribution**:

$$p(\theta|x)$$

We obtain the posterior probability distribution via **Bayes Rule**, which we have already seen during our discussion of conditional probability:

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

where the denominator in this equation:

$$p(x) = \int p(\theta)p(x|\theta)\mathrm{d}\theta$$

is the probability of observing the data x, is independent of θ for a fixed x, and can be considered a constant.

Most Importantly: The Prior

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

$p(\theta)$ is the **prior probability distribution**, which represents the beliefs that we have about the possible values of θ *prior* to observing any information (i.e. from the data).

When trying to estimate θ, we almost always have some prior knowledge, or belief, about the value of θ before taking any data into consideration.

Our Binomial example

$$X|\theta \sim \text{Bin}(n = 10, p = \theta)$$

with $n = 10, x = 10$.

We saw already that classical statistics says that best estimate of θ here in 1.

The **maximum likelihood** of $p(x|\theta)$ in the classical inference sense is $\theta = 1$.

Do you think that this is reasonable?

Back to the start…

We had three scenarios that we considered. They were:

- **The Drunken Coin Toss**: A drunk friend correctly predicts the outcome of 10 tosses of a fair coin.
- **The Wine Critic**: Another friend correctly picks the origin of the wine 10 times in a row.
- **The Classical Pianist**: Another friend correctly identifies a classical composer 10 times in a row.

The classical interpretation

Each of these scenarios presents the same binomial experiment, with the same outcome:

$X|\theta \sim \text{Bin}(10, \theta)$ and $x = 10$.

Based on these data alone, we would be forced to draw the same conclusion in each case; $\theta = 1$.

However our **prior beliefs** are likely to be somewhat at odds with this result. In classical inference this is of no consequence, however in *Bayesian* inference this can have a significant influence over our conclusions…

Prior Belief

The values that you wrote down prior to seeing the data represents your **prior** on θ in each of the three scenarios.

Look back on your priors now. How do they compare with mine?

- Perhaps you're much more open to the possibility that your friend can actually see into the future?
- Or perhaps you have no idea about wine, and so had no idea if it was hard or difficult to tell the difference between bottles from France and Spain?

Importantly, **there are no wrong answers**. The graphs you've drawn are **your** prior belief.

Opinions about the results

After we observed the data, I asked you to write down a probability that you thought each person was telling the truth.

For my results:

- my opinion of the fortune teller is unchanged $P(\text{truth}|\text{success}) \approx 0.$;
- my opinion about the wine-drinker has flipped: I now believe that there's a high chance that he's telling the truth. $P(\text{truth}|\text{success}) = 0.8$;
- for the classical pianist, the observations of the data have further hardened my belief that she is telling the truth: $P(\text{truth}|\text{success}) \approx 1.$

Bayesian Inference and the Prior

What we've just drawn on the last slide (and on your own pages) is essentially your prior belief on the outcome of the challenge, and the resulting **posterior probability distribution** after observing the data.

This is the important part:

- Every one of you was presented with the same data.

- However you all will have different priors. This means that *you will all have different posterior distributions too*.

The **essential basis** of Bayesian inference is that experiments are not abstract devices. Invariably we have some knowledge about the process being investigated before we observe any data, and Bayesian statistics provides us with a mechanism for drawing inference from this combination of prior knowledge and data.

Characteristics of Bayesian Statistics

There are four fundamental aspects of the the Bayesian approach to statistical inference.

- Prior Information
- Subjective Probability
- Self-consistency
- no "ad-hoc"-ery

Characteristics of Bayesian Statistics

Prior Information

All problems are unique to having their own context.

That context derives prior information, and it is the formulation and exploitation of this prior knowledge that sets Bayesian statistics apart from Classical statistics.

Subjective Probability

Bayesian Statistics formalises explicitly the notion that all probabilities are subjective; based on an individuals prior knowledge and the knowledge at hand.

Characteristics of Bayesian Statistics

Self-consistency

By treating θ as a random variable, it emerges that the whole development of Bayesian inference stems from probability theory only.

This means that all statistical inference issues can be addressed as probability statements about θ, which we can derive directly from the posterior distribution.

No "adhockery"

Bayesian inference side-steps the tendency (in classical statistics and inference) to invent ad-hoc criteria for judging and comparing estimators (point estimates, confidence intervals, etc).

This is done by relying on the posterior distribution itself to express (in straightforward probabilistic terms) the entire inference about an unknown θ.

Review of Bayes Theorem (I)

In its basic form, Bayes Theorem is a simple result of conditional probabilities. Indeed, this is how we first came to discover it a few of lectures ago.

If A and B are two events with $P(A) > 0$, then:

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

The use of Bayes Theorem in probability applications is to reverse the conditioning of events. That is, to show how the probability of $B|A$ is related to $A|B$.

Review of Bayes Theorem (I)

A slight extension of Bayes Theorem is obtained by conditioning events $C_1, \dots, C_k$, which partition the sample space Ω so that $C_i \cap C_j = 0$ if $i \neq j$, and $C_1 \cup C_2 \cup \cdots \cup C_k = \Omega$.

In this circumstance the computation of Bayes Theorem becomes:

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{\sum_{j=0}^{k} P(A|C_j)P(C_j)} \quad \forall\ i = 1, \dots, k$$

This formulation is useful because it allows us to consider simple experiments in order to explore the details of Bayes Theorem.

Return of the Urn

Consider an urn that contains six balls of unknown colours.

Three balls are drawn at random *without* replacement, and all are found to be black.

What is the probability that there are **no** black balls left in the urn?

Return of the Urn

Let A be the event that 3 black balls are drawn from the urn, and C_i be the event that there are truly i black balls in the urn.

By Bayes Theorem:

$$P(C_3|A) = \frac{P(A|C_3)P(C_3)}{\sum_{j=0}^{6} P(A|C_j)P(C_j)}$$

The probability $P(A|C_3)$ is simple to calculate:

$$P(A|C_3) = \frac{3}{6} \times \frac{2}{5} \times \frac{1}{4}$$

However the crucial issue is this:

What values do we assign to $P(C_0), \dots, P(C_6)$.

Remember that these are our prior beliefs about there being i black balls in the bag *prior* to seeing any data.

Return of the Urn

Without any additional information, we might logically assume that all outcomes are equally likely:

$$P(C_i) = \frac{1}{k}$$

where $k = 7$ (because there are 7 possible outcomes; $C_0, C_1, \dots, C_6$).

Using this prior:

$$\begin{aligned} P(C_3|A) &= \frac{P(A|C_3)P(C_3)}{\sum_{j=0}^{6} P(A|C_j)P(C_j)} \\ &= \frac{\left(\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4}\right) \times \frac{1}{7}}{\frac{1}{7} \times \left[0 + 0 + 0 + \left(\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4}\right) + \left(\frac{4}{6} \times \frac{3}{5} \times \frac{2}{4}\right) + \left(\frac{5}{6} \times \frac{4}{5} \times \frac{3}{4}\right) + 1\right]} \\ &= \frac{1}{35} \end{aligned}$$

So the data have updated our prior belief from $P(C_3) = \frac{1}{7}$ to $P(C_3|A) = \frac{1}{35}$. Put in words, the event that there is only 3 black balls in the Urn is much less likely having seen the data than we believed it to be previously.

Review of Bayes Theorem (II)

Stated in terms of random variables (with probability densities generally denoted by p) Bayes Theorem then takes the form:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(\theta)p(x|\theta)\mathrm{d}\theta}.$$

As per normal, when x is a continuous random variable p will represent the *probability density function (pdf)* of x, whereas when x is a discrete random variable p will refer to the *probability mass function (pmf)* of x.

Similarly, θ can be continuous or discrete, but in the discrete case the integral in the denominator becomes the summation that we've already encountered:

$$\sum_j p(\theta_j)p(x|\theta_j).$$

Note that the denominator of Bayes Theorem marginalises over θ (and so is only a function of x). Therefore for fixed data, Bayes Theorem can be rewritten as the proportionality:

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

That is, the posterior probability is proportional to the prior probability times the likelihood $p(x|\theta)$.

Bayesian Updating

There are 4 key steps in the Bayesian approach:

1. Specification of the likelihood model $p(x|\theta)$;
2. Determination of the prior $p(\theta)$;
3. Calculation of the posterior distribution $p(\theta|x)$; and
4. Draw inference from the posterior distribution.

Multi-parameter Models

For our purposes in the natural sciences, the examples that we've been using up until this point are not particularly useful.

We have been looking largely at examples that analyse a single variable, such as the binomial coin-toss, a Gaussian distribution with known variance, etc. However in practice all problems that we will encounter will involve more than one variable.

This is where another aspect of Bayesian statistics is much more straight-forward than classical statistics. For highly complex multi-parameter models, **no new methods are required**.

Multi-parameter Models

We now have a vector $\vec{\theta} = \{\theta_1, \ldots, \theta_k\}$ of unknown parameters which we wish to make inference about.

We specify a multivariate model prior distribution $p(\vec{\theta})$ for $\vec{\theta}$, and combine the likelihood $p(x|\vec{\theta})$ via Bayes Theorem to obtain:

$$p(\vec{\theta}|x) = \frac{p(x|\vec{\theta})p(\vec{\theta})}{p(x)}$$

exactly as before.

We often want to draw conclusions about one or more parameters at the same time.

These **marginal distributions** can be obtained in a straight-forward manner using probability calculations on the joint distributions.

Inference of Multi-Parameter Models

For example, the marginal distribution of θ_1 is obtained by integrating out all of the other components of $\vec{\theta}$.

$$p(\theta_1|x) = \int_{\theta_2} \ldots \int_{\theta_k} p(\vec{\theta}|x)\mathrm{d}\theta_2 \ldots \mathrm{d}\theta_k$$

Equivalently, we can use **simulation** to draw samples from the joint distribution and then look at the parameters of interest (i.e. ignore the values of the other parameters).

Inference about multi-parameter models creates the following complications:

1. **Prior Specification**: priors are now multivariate distributions. This means that we need to express dependencies between parameters as well, which is often complicated.
2. **Computation**: we now have even more complicated integrals to evaluate, which creates the necessity for complex numerical techniques.
3. **Interpretation**: the structure of the posterior distribution may be highly complex, which causes difficulties in interpretation.

Posterior Simulation, and Markov Chain Monte Carlo

The arguably most popular/useful method of posterior interpretation is **posterior simulation**.

More than any other technique, **Markov Chain Monte Carlo** has been responsible for the current resurgence of Bayesian Statistics in the natural sciences.

This is because **MCMC** allows us to estimate a vast array of Bayesian models with ease.

The idea of MCMC was first introduced as a method for the efficient simulation of energy levels of atoms in crystalline lattices. It was subsequently adapted for broader use within statistics.

The concept of MCMC is as follows:

Suppose we have some arbitrary "**target distribution**" π:

$$\pi(x), \quad x \in \Omega \in \mathbb{R}^p.$$

If π is sufficiently complex that we are unable to sample from it directly, then an indirect method for sampling from it is to construct a **Markov Chain** within a state space Ω, whose "stationary distribution" is $\pi(x)$.

If we run the chain for long enough, simulated values from the chain can be treated as samples from the target distribution, and used as a basis for summarising the important features of π.

There is **a lot** of jargon above, but don't fret. We will make this clear in the following slides.

The Markov Chain

What is a **markov chain**?

A sequence of random variables X_t is defined as a Markov Chain if it follows the conditional probability:

$$P(X_{t+1} = x_{t+1} | X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_0 = x_0) = P(X_{t+1} = y | X_t = x_t)$$

That is, it is a sequence of numbers that, given some current state, is independent of the past.

The probability of transition from state $x = x_t$ to $y = x_{t+1}$ is given by some **transition probability**.

Visualising the Markov Chain

You may already be familiar with the concept of the Markov chain, even if you don't know it by that name specifically.

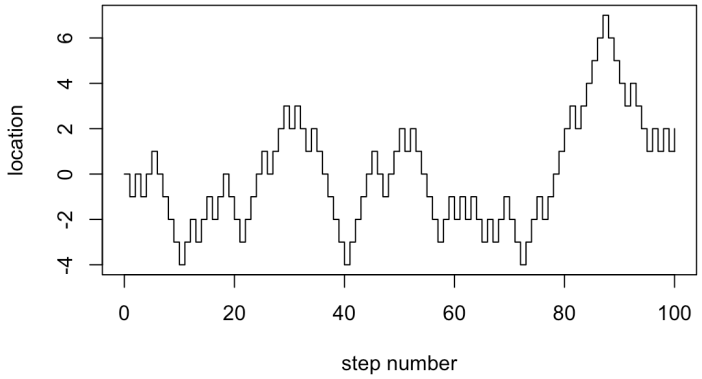
This is because there is a special type of Markov Chain that is somewhat well known: the **random walk**.

A random walk is a Markov chain whereby the transition probability is uniform for the points immediately adjacent to x:

$$P(x, y) = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \end{pmatrix}$$

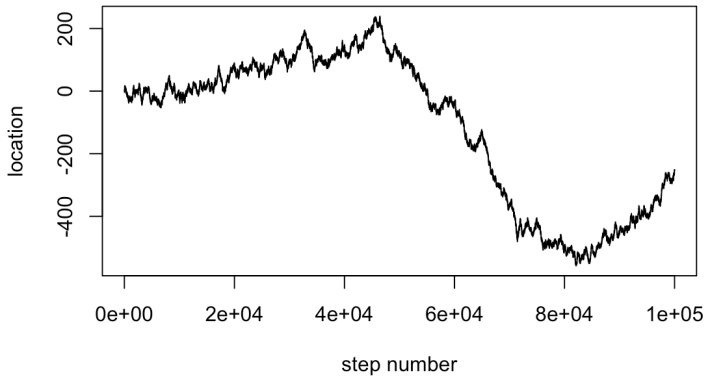
Visualising the Markov Chain

A "random walk" is therefore just a sequence of steps:



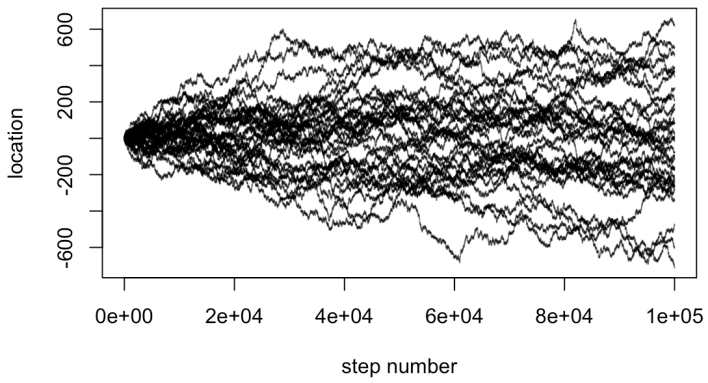
Visualising the Markov Chain

Simplifying this code into a different (faster) implementation:



Visualising the Markov Chain

We can also run many of the Markov Chains from the same starting point:



Using this random walk, we can now explore the transition probability we discussed earlier:

$$\begin{aligned} P^n(x,y) &= P(X_n = y | X_0 = x) \\ &= \sum_{x_1} \dots \sum_{x_{n-1}} P(x,x_1)P(x_1,x_2)\dots P(x_{n-1},y) \end{aligned}$$

Analytically, this probability requires us to sum over all possible values of $x_1, x_2, \dots, x_{n-1}$.

Monte Carlo Simulation

Monte Carlo simulation refers to the generation of realisations (of data/models/steps/distributions) using random draws from a probability distribution.

So: The **MCMC** process involves constructing a Markov Chain that generates Monte Carlo samples for an arbitrary probability distribution.

Building an MCMC Sampler: The Gibbs Sampler

The first MCMC sampler that we are going to look at is the **Gibbs Sampler**.

Consider a pair of variables X, Y, whose joint distribution is denoted by $\pi(x,y)$.

The Gibbs Sampler generates a sample from $\pi(x)$, i.e. the marginal density of π with respect to x, by sampling (in turn) from the conditional distributions $\pi(x|y)$ and $\pi(y|x)$, which frequently known in statistical models of complex data.

This is done by generating a "gibbs sequence" of random variables:

$$Y'_0, X'_0, Y'_1, X'_1, \dots, Y'_N, X'_N$$

The initial value Y'_0 is chosen arbitrarily, and all other values are found iteratively by generating values from:

$$\begin{aligned} X'_t &\sim \pi(x|Y'_t) \\ Y'_{t+1} &\sim \pi(x|X'_t) \end{aligned}$$

This is known as **Gibbs Sampling**.

It turns out that, under reasonably general conditions, the distribution of X'_N converges to $\pi(x)$ as $N \to \infty$.

Said differently, provided N is large enough, the final observation of X'_N is effectively a sample from $\pi(x)$.

The Gibbs Sampler: Algorithm

1. Initialise $\vec{X} = \left(X_0^{(1)}, \dots, X_0^{(d)}\right)$
2. Simulate $X_1^{(1)}$ from the conditional distributions of
$$\pi(X^{(1)} | X_0^{(2)}, \dots, X_0^{(d)})$$
3. Simulate $X_1^{(2)}$ from the conditional distributions of
$$\pi(X^{(2)} | X_1^{(1)}, X_0^{(3)}, \dots, X_0^{(d)})$$
4. …
5. Iterate

Gibbs Demonstration: Bivariate Gaussian

Consider a single observation (y_1, y_2) from a bivariate normally distributed population with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance matrix:

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

With a uniform prior distribution on θ, the posterior distribution is

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} | y \sim N\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Although we can sample from this directly, let's pretend that we cannot.

To apply the Gibbs sampler we must first know the form of the conditional posterior distributions.

From the properties of the multivariate normal distribution these are:

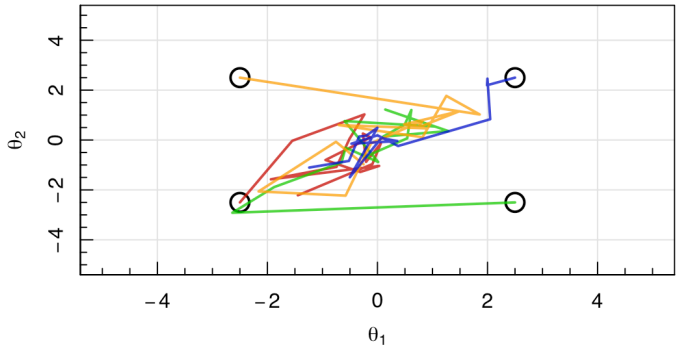
$$\begin{aligned} \theta_1 | \theta_2, y &\sim N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2) \\ \theta_2 | \theta_1, y &\sim N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2) \end{aligned}$$

Let's set $\rho = 0.8$, and $(y_1, y_2) = (0,0)$, and use the initial guess $X_0^1, X_0^2 = (\pm 2.5, \pm 2.5)$ (that is, we're running 4 distinct chains).

```r
#Gibbs Sampler in R
Gibbs<-function(start,n=1e3,rho=0.8) {
  #Initialise the variables
  theta_1<-theta_2<-rep(NA,n)
  #Enter the initial guesses
  theta_1[1]<-start[1]
  theta_2[1]<-start[2]
  #Loop over the next n steps
  for (i in 2:n) {
    #Generate the sample for X
    theta_1[i]<-rnorm(1,rho*theta_2[i-1],sqrt(1-rho^2))
    theta_2[i]<-rnorm(1,rho*theta_1[i],sqrt(1-rho^2))
  }
  return(cbind(theta_1,theta_2))
}
```

Gibbs Demonstration: Bivariate Gaussian

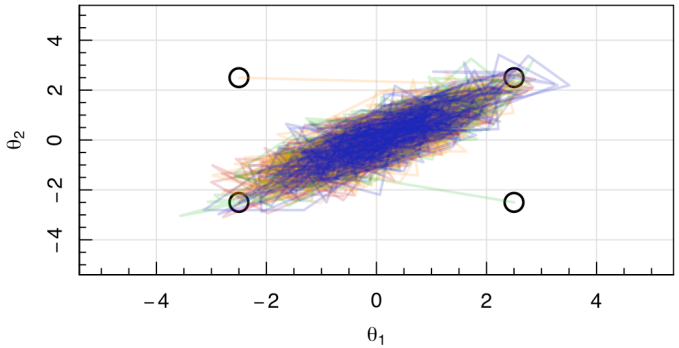
Let's start by plotting the first
15 samples from each of our Gibbs sequences:



You can see our four starting guesses as the black circles, and the subsequent samples of the sequence.

Gibbs Demonstration: Bivariate Gaussian

Let's now crank it up to many samples:



Gibbs Demonstration: Bivariate Gaussian

With so many samples the individual points now become more useful to plot: