Data Visualisation

University of the Witwatersrand

2024

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.

Lesson Plan

- Revision problem

Lesson Plan

- Revision problem
- Gamma Distribution

Lesson Plan

- Revision problem
- Gamma Distribution
- Beta Distribution

Lesson Plan

- Revision problem
- Gamma Distribution
- Beta Distribution
- A Bayesian introduction to the revision problem

Lesson Plan

- Revision problem
- Gamma Distribution
- Beta Distribution
- A Bayesian introduction to the revision problem
- Preview on Conjugate Priors

Lesson Plan

- Revision problem
- Gamma Distribution
- Beta Distribution
- A Bayesian introduction to the revision problem
- Preview on Conjugate Priors
- Riddle

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.
- $H_0 : p = \frac{1}{2}$

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.
- $H_0 : p = \frac{1}{2}$
- $H_1 : p \neq \frac{1}{2}$

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.
- $H_0 : p = \frac{1}{2}$
- $H_1 : p \neq \frac{1}{2}$
- $\alpha = 0.05$ (arbitrary choice).

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.
- $H_0 : p = \frac{1}{2}$
- $H_1 : p \neq \frac{1}{2}$
- $\alpha = 0.05$ (arbitrary choice).
- P-value $= \binom{5}{0}(\frac{1}{2})^5 + \binom{5}{5}(\frac{1}{2})^5 = \frac{1}{16} = 0.0625$

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.
- $H_0 : p = \frac{1}{2}$
- $H_1 : p \neq \frac{1}{2}$
- $\alpha = 0.05$ (arbitrary choice).
- P-value $= \binom{5}{0}\left(\frac{1}{2}\right)^5 + \binom{5}{5}\left(\frac{1}{2}\right)^5 = \frac{1}{16} = 0.0625$
- We fail to reject the null hypothesis.

The Gamma Distribution

- The Gamma distribution is a generalisation of several distributions. Such as the chi-squared, the Erlang Distribution and our old friend the exponential

The Gamma Distribution

- The Gamma distribution is a generalisation of several distributions. Such as the chi-squared, the Erlang Distribution and our old friend the exponential
- The exponential distribution is often thought of as the waiting time for an event to occur.

The Gamma Distribution

- The Gamma distribution is a generalisation of several distributions. Such as the chi-squared, the Erlang Distribution and our old friend the exponential
- The exponential distribution is often thought of as the waiting time for an event to occur.
- The Gamma distribution is the waiting time for an event to occur k times.

The Gamma Distribution

- The Gamma distribution is a generalisation of several distributions. Such as the chi-squared, the Erlang Distribution and our old friend the exponential
- The exponential distribution is often thought of as the waiting time for an event to occur.
- The Gamma distribution is the waiting time for an event to occur k times.
- This gives the Gamma two parameters. The old λ that our exponential had (often called a scale parameter) and the k (often called a shape parameter).

The Gamma Distribution

- The pdf of the Gamma Distribution is $\frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$ for $x > 0$

The Gamma Distribution

- The pdf of the Gamma Distribution is $\frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$ for $x > 0$
- Notice that when $k = 1$ this is the exponential distribution. $\Gamma(1) = 1$ and the $x^{k-1} = x^0 = 1$ and then we're left with $\lambda e^{-\lambda x}$.

The Gamma Distribution

- The pdf of the Gamma Distribution is $\frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$ for $x > 0$
- Notice that when $k = 1$ this is the exponential distribution. $\Gamma(1) = 1$ and the $x^{k-1} = x^0 = 1$ and then we're left with $\lambda e^{-\lambda x}$.
- To prove that the Gamma is the distribution of a sum of independent exponential random variables we'll use induction on k. We handled the base case above!

The Gamma Distribution

- Assume true for some k and we'll try to prove true for $k+1$. Here we assume that that's the pdf of k exponentials.

The Gamma Distribution

- Assume true for some k and we'll try to prove true for $k+1$. Here we assume that that's the pdf of k exponentials.
-

$$
\begin{aligned}
f_{k+1}(t) &= \int_0^\infty f_1(t-x) f_k(x) dx \\
&= \int_0^t [\lambda e^{-\lambda(t-x)}][\frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}] dx \\
&= \frac{\lambda^{k+1} e^{-\lambda t}}{\Gamma(k)} \int_0^t [x^{k-1}] dx \\
&= \frac{\lambda^{k+1} e^{-\lambda t}}{\Gamma(k)} \frac{t^k}{k} \\
&= \frac{\lambda^{k+1} t^k e^{-\lambda t}}{\Gamma(k+1)}
\end{aligned}
$$

The Beta Distribution

- The beta distribution is another two-parameter family of distributions.

The Beta Distribution

- The beta distribution is another two-parameter family of distributions.
- The classical derivation is from the order statistics of uniform $[0, 1]$ distributions.

The Beta Distribution

- The beta distribution is another two-parameter family of distributions.
- The classical derivation is from the order statistics of uniform $[0, 1]$ distributions.
- By order statistics I mean you generate n i.i.d. random variables and choose the k^{th} largest/smallest.

The Beta Distribution

- The beta distribution is another two-parameter family of distributions.
- The classical derivation is from the order statistics of uniform $[0, 1]$ distributions.
- By order statistics I mean you generate n i.i.d. random variables and choose the k^{th} largest/smallest.
- Instead of a n and k usually we use an α and β.

The Beta Distribution

- The beta distribution is another two-parameter family of distributions.
- The classical derivation is from the order statistics of uniform $[0, 1]$ distributions.
- By order statistics I mean you generate n i.i.d. random variables and choose the k^{th} largest/smallest.
- Instead of a n and k usually we use an α and β.
- The pdf of a beta distribution is $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ on $[0, 1]$.

The Beta Distribution

- This is a little bit like a binomial distribution. Those Gammas look suspiciously like a binomial coefficient and $x^{\alpha-1}$ and $(1-x)^{\beta-1}$ feel a lot like p^k and $(1-p)^{n-k}$.

The Beta Distribution

- This is a little bit like a binomial distribution. Those Gammas look suspiciously like a binomial coefficient and $x^{\alpha-1}$ and $(1-x)^{\beta-1}$ feel a lot like p^k and $(1-p)^{n-k}$.
- So in some sense this is a continuous analog of a binomial distribution.

The Beta Distribution

- This is a little bit like a binomial distribution. Those Gammas look suspiciously like a binomial coefficient and $x^{\alpha-1}$ and $(1-x)^{\beta-1}$ feel a lot like p^k and $(1-p)^{n-k}$.
- So in some sense this is a continuous analog of a binomial distribution.
- Let's talk about the k^{th} order statistic of n.

The Beta Distribution

- This is a little bit like a binomial distribution. Those Gammas look suspiciously like a binomial coefficient and $x^{\alpha-1}$ and $(1-x)^{\beta-1}$ feel a lot like p^k and $(1-p)^{n-k}$.
- So in some sense this is a continuous analog of a binomial distribution.
- Let's talk about the k^{th} order statistic of n.
- So we have X_1, X_2,... X_n are i.i.d and we let $Y_1 \leq Y_2 \leq Y_3 \leq .. \leq Y_n$ be those same values but in order.

The Beta Distribution

- This is a little bit like a binomial distribution. Those Gammas look suspiciously like a binomial coefficient and $x^{\alpha-1}$ and $(1-x)^{\beta-1}$ feel a lot like p^k and $(1-p)^{n-k}$.
- So in some sense this is a continuous analog of a binomial distribution.
- Let's talk about the k^{th} order statistic of n.
- So we have X_1, X_2,... X_n are i.i.d and we let $Y_1 \leq Y_2 \leq Y_3 \leq .. \leq Y_n$ be those same values but in order.
- So Y_1 is the minimum of the X_i and Y_n is the maximum of the X_i.

The Beta Distribution

- This is a little bit like a binomial distribution. Those Gammas look suspiciously like a binomial coefficient and $x^{\alpha-1}$ and $(1-x)^{\beta-1}$ feel a lot like p^k and $(1-p)^{n-k}$.
- So in some sense this is a continuous analog of a binomial distribution.
- Let's talk about the k^{th} order statistic of n.
- So we have X_1, X_2,... X_n are i.i.d and we let $Y_1 \leq Y_2 \leq Y_3 \leq .. \leq Y_n$ be those same values but in order.
- So Y_1 is the minimum of the X_i and Y_n is the maximum of the X_i.
- For now we'll only think about continuous distributions. This avoids any chance of ties, which get very case-ish very quickly.

The Beta Distribution

- Let's consider the distribution of Y_i when $n = 5$ and the X_i distributed uniformly.

The Beta Distribution

- Let's consider the distribution of Y_i when $n = 5$ and the X_i distributed uniformly.
- We'll start with Y_5.

The Beta Distribution

- Let's consider the distribution of Y_i when $n = 5$ and the X_i distributed uniformly.
- We'll start with Y_5.
- $\mathbb{P}(Y_5 \leq t) = \prod_{i=1}^{5} \mathbb{P}(X_i < t) = p^5$

The Beta Distribution

- Let's consider the distribution of Y_i when $n = 5$ and the X_i distributed uniformly.
- We'll start with Y_5.
- $\mathbb{P}(Y_5 \leq t) = \prod_{i=1}^{5} \mathbb{P}(X_i < t) = p^5$
- This makes the density (pdf) the derivative of p^5 which is $5p^4$. Of course on $[0, 1]$.

The Beta Distribution

- Let's consider the distribution of Y_i when $n = 5$ and the X_i distributed uniformly.
- We'll start with Y_5.
- $\mathbb{P}(Y_5 \leq t) = \prod_{i=1}^{5} \mathbb{P}(X_i < t) = p^5$
- This makes the density (pdf) the derivative of p^5 which is $5p^4$. Of course on $[0, 1]$.
- Let's consider Y_3.

The Beta Distribution

- Let's consider the distribution of Y_i when $n = 5$ and the X_i distributed uniformly.
- We'll start with Y_5.
- $\mathbb{P}(Y_5 \leq t) = \prod_{i=1}^{5} \mathbb{P}(X_i < t) = p^5$
- This makes the density (pdf) the derivative of p^5 which is $5p^4$. Of course on $[0, 1]$.
- Let's consider Y_3.
- $\mathbb{P}(Y_3 \in (t, t + \epsilon))$. Well we need one of our X_i in that $(t, t + \epsilon)$ two below t and two above t. There are $5 \times \binom{4}{2} = 30$ ways to do this. Each has a probability of $F(t)^2(1 - F(t))^2(F(t + \epsilon) - F(t))$.

The Beta Distribution

- Taking a limit as ϵ tends to zero $(F(t+\epsilon) - F(t)) = f(t)$.

The Beta Distribution

- Taking a limit as ϵ tends to zero $(F(t+\epsilon) - F(t)) = f(t)$.
- $30F(t)^2(1-F(t))^2(F(t+\epsilon) - F(t)) = 30t^2(1-t)^2$

The Beta Distribution

- Taking a limit as ϵ tends to zero $(F(t+\epsilon) - F(t)) = f(t)$.
- $30F(t)^2(1 - F(t))^2(F(t+\epsilon) - F(t)) = 30t^2(1-t)^2$
- More generally let's think of Y_i for general n and i the X_i still distributed uniformly.

The Beta Distribution

- Taking a limit as ϵ tends to zero $(F(t+\epsilon) - F(t)) = f(t)$.
- $30F(t)^2(1 - F(t))^2(F(t+\epsilon) - F(t)) = 30t^2(1-t)^2$
- More generally let's think of Y_i for general n and i the X_i still distributed uniformly.
- Well now we can do the same calculation but we'll have $i - 1$ observations below t and $n - i$ above $t + \epsilon$ and one in that very narrow $(t, t + \epsilon)$ interval.

The Beta Distribution

- Taking a limit as ϵ tends to zero $(F(t+\epsilon) - F(t)) = f(t)$.
- $30F(t)^2(1-F(t))^2(F(t+\epsilon) - F(t)) = 30t^2(1-t)^2$
- More generally let's think of Y_i for general n and i the X_i still distributed uniformly.
- Well now we can do the same calculation but we'll have $i-1$ observations below t and $n-i$ above $t+\epsilon$ and one in that very narrow $(t, t+\epsilon)$ interval.
- $n \times \binom{n-1}{i-1}$ ways to place the points.

The Beta Distribution

- Each of these comes up with probability $F(t)^{i-1}(1 - F(t))^{n-i}(F(t+\epsilon) - F(t))$, for a uniform this is $\frac{n!}{(i-1)!(n-i)!} t^{i-1}(1-t)^{n-i}$

The Beta Distribution

- Each of these comes up with probability $F(t)^{i-1}(1-F(t))^{n-i}(F(t+\epsilon)-F(t))$, for a uniform this is $\frac{n!}{(i-1)!(n-i)!}t^{i-1}(1-t)^{n-i}$
- Which is $\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)}t^{i-1}(1-t)^{n-i}$. Which is a Beta distribution with parameters $i-1$ and $n-i$

The Beta Distribution

- Each of these comes up with probability $F(t)^{i-1}(1-F(t))^{n-i}(F(t+\epsilon)-F(t))$, for a uniform this is $\frac{n!}{(i-1)!(n-i)!}t^{i-1}(1-t)^{n-i}$
- Which is $\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)}t^{i-1}(1-t)^{n-i}$. Which is a Beta distribution with parameters $i-1$ and $n-i$
- These are the number of values below you and above you.

Moments of a Beta distribution

- $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$

Moments of a Beta distribution

- $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$
- $\mathbb{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Moments of a Beta distribution

- $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$
- $\mathbb{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Intuitively if you think of the Beta as a re-scaled and normalized Binomial. Then the expectation is p instead of np because.

Moments of a Beta distribution

- $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$
- $\mathbb{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Intuitively if you think of the Beta as a re-scaled and normalized Binomial. Then the expectation is p instead of np because.
- The variance of a binomial is $np(q-p)$ so a re-scaled version "should" have variance $\frac{pq}{n}$. The Beta's variance can be thought of as roughly $p = \frac{\alpha}{\alpha+\beta}$ times $1-p = \frac{\beta}{\alpha+\beta}$ times $\frac{1}{n} = \frac{1}{\alpha+\beta+1}$

A Bayesian Approach to the Revision Problem

- In the above hypothesis test we used a frequentist approach.

A Bayesian Approach to the Revision Problem

- In the above hypothesis test we used a frequentist approach.
- That's because it's pretty natural to think that coins could well be fair. They're usually made to be.

A Bayesian Approach to the Revision Problem

- In the above hypothesis test we used a frequentist approach.
- That's because it's pretty natural to think that coins could well be fair. They're usually made to be.
- On the other hand we could live in a world where they're not usually fair. Or be dealing with some other application (say classifying emails) where we don't have any good reason to pre-suppose that $p = \frac{1}{2}$.

A Bayesian Approach to the Revision Problem

- In the above hypothesis test we used a frequentist approach.
- That's because it's pretty natural to think that coins could well be fair. They're usually made to be.
- On the other hand we could live in a world where they're not usually fair. Or be dealing with some other application (say classifying emails) where we don't have any good reason to pre-suppose that $p = \frac{1}{2}$.
- Let's look a Bayesian approach.

A Bayesian Approach to the Revision Problem

- Let's imagine that while we don't know if the coin is fair but we can reasonably assume that it comes up heads with some fixed but unknown probability p.

A Bayesian Approach to the Revision Problem

- Let's imagine that while we don't know if the coin is fair but we can reasonably assume that it comes up heads with some fixed but unknown probability p.
- For this example we'll assume that p is uniform on $[0, 1]$.

A Bayesian Approach to the Revision Problem

- Let's imagine that while we don't know if the coin is fair but we can reasonably assume that it comes up heads with some fixed but unknown probability p.
- For this example we'll assume that p is uniform on $[0, 1]$.
- This is pretty arbitrary but no more arbitrary than assuming that $p = \frac{1}{2}$

A Bayesian Approach to the Revision Problem

- Let's imagine that while we don't know if the coin is fair but we can reasonably assume that it comes up heads with some fixed but unknown probability p.
- For this example we'll assume that p is uniform on $[0, 1]$.
- This is pretty arbitrary but no more arbitrary than assuming that $p = \frac{1}{2}$
- Then the data comes in! After it's in we won't know p but we'll have a better idea than before we got the data.

A Bayesian Approach to the Revision Problem

- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$

A Bayesian Approach to the Revision Problem

- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)

A Bayesian Approach to the Revision Problem

- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^5$ because if we know that the probability of heads is p then the probability of us getting five heads is p^5.

A Bayesian Approach to the Revision Problem

- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^5$ because if we know that the probability of heads is p then the probability of us getting five heads is p^5.
- $\mathbb{P}(X) = \int_0^1 p^5 dp = \frac{1}{6}$

A Bayesian Approach to the Revision Problem

- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^5$ because if we know that the probability of heads is p then the probability of us getting five heads is p^5.
- $\mathbb{P}(X) = \int_0^1 p^5 dp = \frac{1}{6}$
- We want $\mathbb{P}(\theta|X)$. Plugging things in and juggling the algebra gives $\mathbb{P}(\theta|X) = 6p^5$ (for p in $[0, 1]$).

Next we'll generalize!

- We'll still start with a prior distribution where p is uniform in $[0, 1]$.

Next we'll generalize!

- We'll still start with a prior distribution where p is uniform in $[0, 1]$.
- Supposing that we get h heads and t tails.

Next we'll generalize!

- We'll still start with a prior distribution where p is uniform in $[0, 1]$.
- Supposing that we get h heads and t tails.
- That's the preview, we'll also discuss the even more general concept of conjugate priors.

Things to note

- $\mathbb{P}(X)$ is a constant.

Things to note

- $\mathbb{P}(X)$ is a constant.
- It turns out that we can figure out this kind of constant from the fact that $\mathbb{P}(\theta|X)$ is a distribution (so integrates to one).

Things to note

- $\mathbb{P}(X)$ is a constant.
- It turns out that we can figure out this kind of constant from the fact that $\mathbb{P}(\theta|X)$ is a distribution (so integrates to one).
- We can write $\mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$ as

Things to note

- $\mathbb{P}(X)$ is a constant.
- It turns out that we can figure out this kind of constant from the fact that $\mathbb{P}(\theta|X)$ is a distribution (so integrates to one).
- We can write $\mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$ as
- $\mathbb{P}(\theta|X) = \frac{\mathbb{P}(\theta)\mathbb{P}(X|\theta)}{\mathbb{P}(X)}$

Things to note

- $\mathbb{P}(X)$ is a constant.
- It turns out that we can figure out this kind of constant from the fact that $\mathbb{P}(\theta|X)$ is a distribution (so integrates to one).
- We can write $\mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$ as
- $\mathbb{P}(\theta|X) = \frac{\mathbb{P}(\theta)\mathbb{P}(X|\theta)}{\mathbb{P}(X)}$
- Equivalently $\mathbb{P}(\theta|X) \propto \mathbb{P}(\theta)\mathbb{P}(X|\theta)$

Things to note

- $\mathbb{P}(X)$ is a constant.
- It turns out that we can figure out this kind of constant from the fact that $\mathbb{P}(\theta|X)$ is a distribution (so integrates to one).
- We can write $\mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$ as
- $\mathbb{P}(\theta|X) = \frac{\mathbb{P}(\theta)\mathbb{P}(X|\theta)}{\mathbb{P}(X)}$
- Equivalently $\mathbb{P}(\theta|X) \propto \mathbb{P}(\theta)\mathbb{P}(X|\theta)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)

Things to note

- $\mathbb{P}(X)$ is a constant.
- It turns out that we can figure out this kind of constant from the fact that $\mathbb{P}(\theta|X)$ is a distribution (so integrates to one).
- We can write $\mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$ as
- $\mathbb{P}(\theta|X) = \frac{\mathbb{P}(\theta)\mathbb{P}(X|\theta)}{\mathbb{P}(X)}$
- Equivalently $\mathbb{P}(\theta|X) \propto \mathbb{P}(\theta)\mathbb{P}(X|\theta)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- This gives $\mathbb{P}(\theta|X) \propto p^2(1-p)^2$ which means that $\mathbb{P}(\theta|X)$ is a beta distribution!

Things to note

- $\mathbb{P}(X)$ is a constant.
- It turns out that we can figure out this kind of constant from the fact that $\mathbb{P}(\theta|X)$ is a distribution (so integrates to one).
- We can write $\mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$ as
- $\mathbb{P}(\theta|X) = \frac{\mathbb{P}(\theta)\mathbb{P}(X|\theta)}{\mathbb{P}(X)}$
- Equivalently $\mathbb{P}(\theta|X) \propto \mathbb{P}(\theta)\mathbb{P}(X|\theta)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- This gives $\mathbb{P}(\theta|X) \propto p^2(1-p)^2$ which means that $\mathbb{P}(\theta|X)$ is a beta distribution!
- It's worth noting that this works because we know that it's proportional to a beta distribution and that it's a distribution. If it was a rescaled beta it wouldn't integrate to one!

Another example

- This time let's do six flips and have four heads

Another example

- This time let's do six flips and have four heads
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$

Another example

- This time let's do six flips and have four heads
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)

Another example

- This time let's do six flips and have four heads
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^4(1-p)^2$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{6}{4}p^4(1-p)^2$.

Another example

- This time let's do six flips and have four heads
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^4(1-p)^2$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{6}{4}p^4(1-p)^2$.
- $\mathbb{P}(X) = \int_0^1 \binom{6}{4}p^4(1-p)^2 dp = \dots$

Another example

- This time let's do six flips and have four heads
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^4(1-p)^2$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{6}{4} p^4(1-p)^2$.
- $\mathbb{P}(X) = \int_0^1 \binom{6}{4} p^4(1-p)^2 dp = \ldots$
- Well it's some constant!

Another example

- This time let's do six flips and have four heads
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^4(1-p)^2$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{6}{4}p^4(1-p)^2$.
- $\mathbb{P}(X) = \int_0^1 \binom{6}{4}p^4(1-p)^2 dp = \ldots$
- Well it's some constant!
- In fact it's a beta-integral and the form is pretty well known but we're going to ignore that. The point here is that it becomes a constant.

Another example

- This time let's do six flips and have four heads
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- In our example. $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^4(1-p)^2$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{6}{4}p^4(1-p)^2$.
- $\mathbb{P}(X) = \int_0^1 \binom{6}{4}p^4(1-p)^2 dp = \ldots$
- Well it's some constant!
- In fact it's a beta-integral and the form is pretty well known but we're going to ignore that. The point here is that it becomes a constant.
- So we have that $\mathbb{P}(\theta|X) \propto \mathbb{P}(\theta)\mathbb{P}(X|\theta) = p^4(1-p)^2$ and is therefore beta

More generally

- Imagine that we have n flips and get k heads.

More generally

- Imagine that we have n flips and get k heads.
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$

More generally

- Imagine that we have n flips and get k heads.
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)

More generally

- Imagine that we have n flips and get k heads.
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^k(1-p)^{n-k}$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{n}{k}p^k(1-p)^{n-k}$.

More generally

- Imagine that we have n flips and get k heads.
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0,1]$)
- $\mathbb{P}(X|\theta) = p^k(1-p)^{n-k}$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{n}{k}p^k(1-p)^{n-k}$.
- $\mathbb{P}(X) = \int_0^1 \binom{n}{k}p^k(1-p)^{n-k}dp = \ldots$

More generally

- Imagine that we have n flips and get k heads.
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^k(1-p)^{n-k}$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{n}{k} p^k (1-p)^{n-k}$.
- $\mathbb{P}(X) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \ldots$
- Again it's a beta integral but it's got to be some constant. some constant!

More generally

- Imagine that we have n flips and get k heads.
- $\mathbb{P}(\theta \cap X) = \mathbb{P}(\theta)\mathbb{P}(X|\theta) = \mathbb{P}(X)\mathbb{P}(\theta|X)$
- Once again $\theta = p$ and $\mathbb{P}(\theta) = 1$ (when $p \in [0, 1]$)
- $\mathbb{P}(X|\theta) = p^k(1-p)^{n-k}$ because if we know that the probability of heads is p then the probability of us getting four heads is $\binom{n}{k} p^k(1-p)^{n-k}$.
- $\mathbb{P}(X) = \int_0^1 \binom{n}{k} p^k(1-p)^{n-k} dp = \ldots$
- Again it's a beta integral but it's got to be some constant. some constant!
- Once again $\mathbb{P}(\theta|X) \propto \mathbb{P}(\theta)\mathbb{P}(X|\theta) = p^k(1-p)^{n-k}$ and is therefore beta

Something other than the uniform

- If we start with a beta-distribution instead of a uniform what happens?

Something other than the uniform

- If we start with a beta-distribution instead of a uniform what happens?
- Well the uniform is a beta (with parameters one and one)

Something other than the uniform

- If we start with a beta-distribution instead of a uniform what happens?
- Well the uniform is a beta (with parameters one and one)
- So having a beta with parameters α and β becomes exactly the same as having a uniform and getting $\alpha - 1$ tails and $\beta - 1$ heads.

Something other than the uniform

- If we start with a beta-distribution instead of a uniform what happens?
- Well the uniform is a beta (with parameters one and one)
- So having a beta with parameters α and β becomes exactly the same as having a uniform and getting $\alpha - 1$ tails and $\beta - 1$ heads.
- So in this case our posterior is like having $k + \beta - 1$ heads and $n - k + \alpha - 1$ tails. Which we solved for above.

Something other than the uniform

- If we start with a beta-distribution instead of a uniform what happens?
- Well the uniform is a beta (with parameters one and one)
- So having a beta with parameters α and β becomes exactly the same as having a uniform and getting $\alpha - 1$ tails and $\beta - 1$ heads.
- So in this case our posterior is like having $k + \beta - 1$ heads and $n - k + \alpha - 1$ tails. Which we solved for above.
- Cool part of the conjugate priors is that all of this algebra and computation is already done within the maths.

The negative binomial

- This can be seen by again using the same proportionality argument above

The negative binomial

- This can be seen by again using the same proportionality argument above
- Turns out that coin flips have the same information whatever the stopping criteria.

The negative binomial

- Counts the number of tails until the r^{th} head

The negative binomial

- Counts the number of tails until the r^{th} head
- Support is the non-negative integers.

The negative binomial

- Counts the number of tails until the r^{th} head
- Support is the non-negative integers.
- PDF is $\binom{k+r-1}{k} p^k (1-p)^r$ where k is our variable

The negative binomial

- Counts the number of tails until the r^{th} head
- Support is the non-negative integers.
- PDF is $\binom{k+r-1}{k} p^k (1-p)^r$ where k is our variable
- Sum of r independent geometric random variables. So mean and variance are $\frac{r(1-p)}{p}$ and $\frac{r(1-p)}{p^2}$ respectively.

The negative binomial

- Counts the number of tails until the r^{th} head
- Support is the non-negative integers.
- PDF is $\binom{k+r-1}{k} p^k (1-p)^r$ where k is our variable
- Sum of r independent geometric random variables. So mean and variance are $\frac{r(1-p)}{p}$ and $\frac{r(1-p)}{p^2}$ respectively.
- It's conjugate prior is again the beta!!

Other conjugate priors

- We saw Bernoulli, uniform and Beta have a conjugate prior of the Beta.

Other conjugate priors

- We saw Bernoulli, uniform and Beta have a conjugate prior of the Beta.
- The negative binomial also has conjugate prior the Beta

Other conjugate priors

- We saw Bernoulli, uniform and Beta have a conjugate prior of the Beta.
- The negative binomial also has conjugate prior the Beta
- Categorical and multi-nomial distributions have conjugate prior the Drichlet (generalisation of Beta)

Other conjugate priors

- We saw Bernoulli, uniform and Beta have a conjugate prior of the Beta.
- The negative binomial also has conjugate prior the Beta
- Categorical and multi-nomial distributions have conjugate prior the Drichlet (generalisation of Beta)
- The Poisson has conjugate prior the Gamma.

Riddle

- A chicken farmer can make 6kg, 9kg and 20kg bags of chicken. However he can't make anything else, so for example if you asked him for 7kg of chicken he couldn't do it. On the other hand if you asked him for 15kg he would be able to do that by selling you both a 6kg bag and a 9 kg bag.

Riddle

- A chicken farmer can make 6kg, 9kg and 20kg bags of chicken. However he can't make anything else, so for example if you asked him for 7kg of chicken he couldn't do it. On the other hand if you asked him for 15kg he would be able to do that by selling you both a 6kg bag and a 9 kg bag.
- A natural question to ask is which amounts can he make and which amounts can he not make, a sub-question here is to ask what the largest integer amount he cannot make is. It should be noted that he can make as many bags of each weight as he wants.

Riddle

- A chicken farmer can make 6kg, 9kg and 20kg bags of chicken. However he can't make anything else, so for example if you asked him for 7kg of chicken he couldn't do it. On the other hand if you asked him for 15kg he would be able to do that by selling you both a 6kg bag and a 9 kg bag.

Riddle

- A chicken farmer can make 6kg, 9kg and 20kg bags of chicken. However he can't make anything else, so for example if you asked him for 7kg of chicken he couldn't do it. On the other hand if you asked him for 15kg he would be able to do that by selling you both a 6kg bag and a 9 kg bag.
- A natural question to ask is which amounts can he make and which amounts can he not make, a sub-question here is to ask what the largest integer amount he cannot make is. It should be noted that he can make as many bags of each weight as he wants.