Lecture 2: Exact Bayesian inference

Ben Lambert[1]
ben.lambert@some.ox.ac.uk

[1]Somerville College
University of Oxford

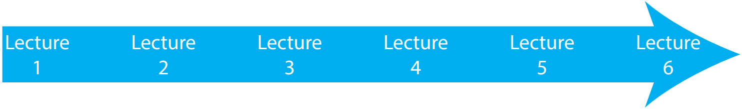
November 2, 2016

Lecture outcomes

By the end of this lecture you should:

1. Understand the elements of Bayes' rule and the intuition behind how it works.
2. Grasp what is meant by a posterior predictive distribution. And how it can be used to:
   - Forecast.
   - Critically assess a model.
3. Appreciate why **exact** Bayesian inference is hard.
4. Know what conjugate priors are and how they can be used to simplify (basic) analyses.

Our progress in the overall course



Lecture 1

Lecture 2

Lecture 3

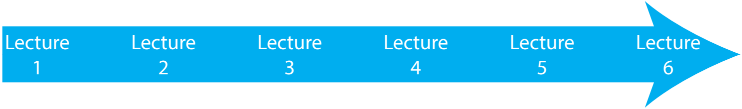
Lecture 4

Lecture 5

Lecture 6

Our progress in the overall course

"I know what inference is."



Lecture 1

Lecture 2

Lecture 3

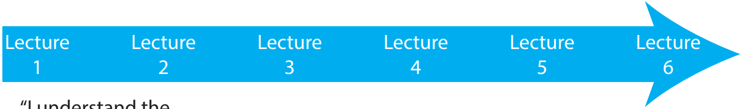
Lecture 4

Lecture 5

Lecture 6

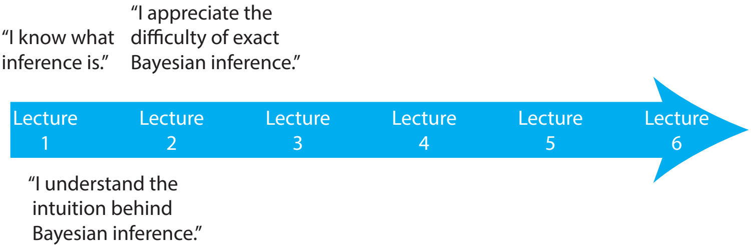
Our progress in the overall course

“I know what inference is.”



“I understand the intuition behind Bayesian inference.”

Our progress in the overall course



"I know what inference is."

"I appreciate the difficulty of exact Bayesian inference."

Lecture 1

Lecture 2

Lecture 3

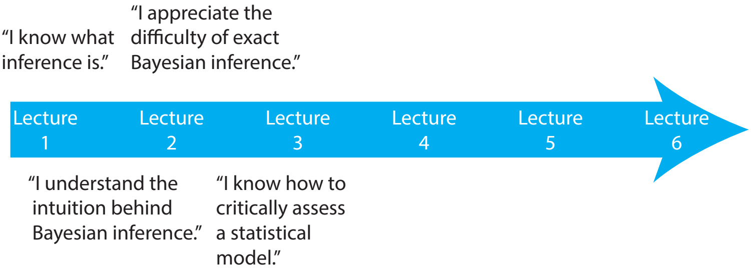
Lecture 4

Lecture 5

Lecture 6

"I understand the intuition behind Bayesian inference."

Our progress in the overall course



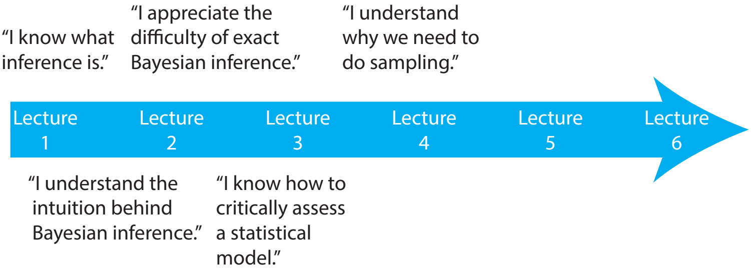
“I know what inference is.”

“I understand the intuition behind Bayesian inference.”

“I appreciate the difficulty of exact Bayesian inference.”

“I know how to critically assess a statistical model.”

Our progress in the overall course



"I know what inference is."

"I appreciate the difficulty of exact Bayesian inference."

"I understand why we need to do sampling."

Lecture 1

Lecture 2

Lecture 3

Lecture 4

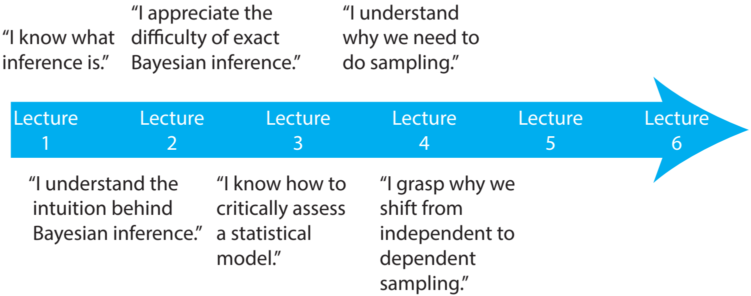
Lecture 5

Lecture 6

"I understand the intuition behind Bayesian inference."

"I know how to critically assess a statistical model."

Our progress in the overall course



Lecture 1

"I know what inference is."

"I understand the intuition behind Bayesian inference."

Lecture 2

"I appreciate the difficulty of exact Bayesian inference."

Lecture 3

"I know how to critically assess a statistical model."

"I understand why we need to do sampling."

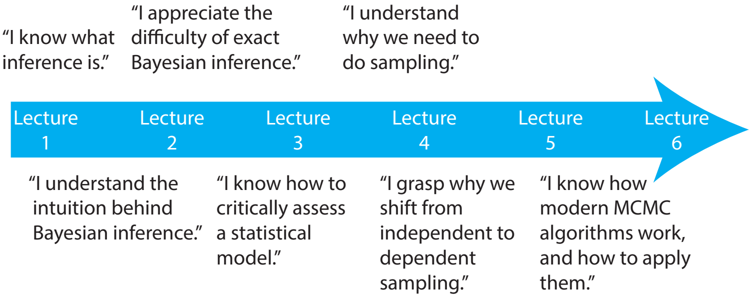
Lecture 4

"I grasp why we shift from independent to dependent sampling."

Lecture 5

Lecture 6

Our progress in the overall course



"I know what inference is."

"I appreciate the difficulty of exact Bayesian inference."

"I understand why we need to do sampling."

Lecture 1 | Lecture 2 | Lecture 3 | Lecture 4 | Lecture 5 | Lecture 6

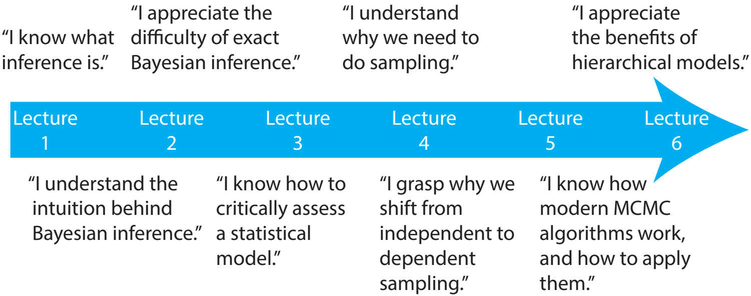
"I understand the intuition behind Bayesian inference."

"I know how to critically assess a statistical model."

"I grasp why we shift from independent to dependent sampling."

"I know how modern MCMC algorithms work, and how to apply them."

Our progress in the overall course



Lecture 1

Lecture 2

Lecture 3

Lecture 4

Lecture 5

Lecture 6

“I know what inference is.”

“I understand the intuition behind Bayesian inference.”

“I appreciate the difficulty of exact Bayesian inference.”

“I know how to critically assess a statistical model.”

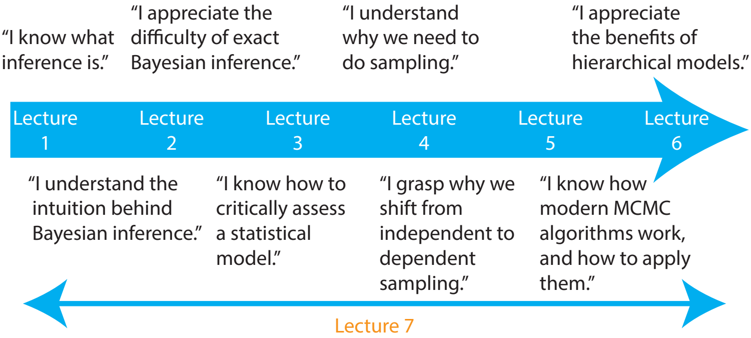
“I understand why we need to do sampling.”

“I grasp why we shift from independent to dependent sampling.”

“I know how modern MCMC algorithms work, and how to apply them.”

“I appreciate the benefits of hierarchical models.”

Our progress in the overall course



“I know what inference is.”

“I appreciate the difficulty of exact Bayesian inference.”

“I understand why we need to do sampling.”

“I appreciate the benefits of hierarchical models.”

Lecture 1

Lecture 2

Lecture 3

Lecture 4

Lecture 5

Lecture 6

“I understand the intuition behind Bayesian inference.”

“I know how to critically assess a statistical model.”

“I grasp why we shift from independent to dependent sampling.”

“I know how modern MCMC algorithms work, and how to apply them.”

Lecture 7

Outline

1. Previous lecture recap

2. Elements of Bayes' rule for inference

3. Why is exact Bayesian statistics hard?

4. Making life simpler: Conjugate priors

1. Previous lecture recap

2. Elements of Bayes’ rule for inference

3. Why is exact Bayesian statistics hard?

4. Making life simpler: Conjugate priors

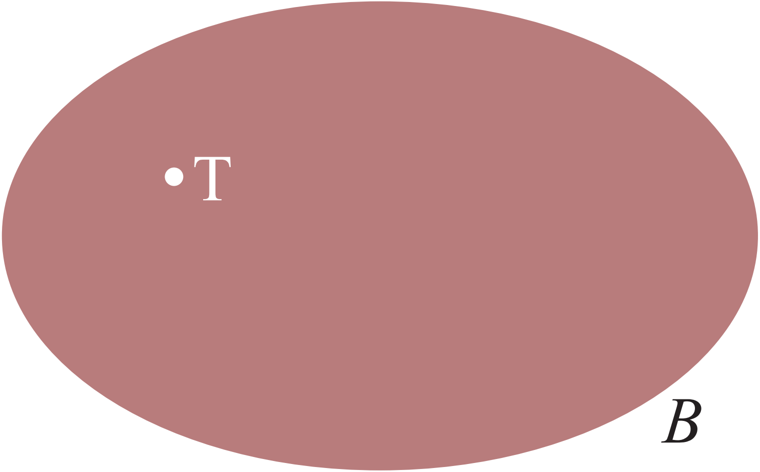
Example problem: paternal discrepancy

- **Paternal discrepancy** is the term given to a child who has a biological father different to their supposed biological father.
- **Question:** how common is it?
- **Answer:** a recent meta-analysis of studies of "paternal discrepancy" (PD) found a rate of $\sim 10\%$[1].
- Suppose we have data for a random sample of 10 children's presence/absence of PD.

Aim: infer the prevalence of PD in the population (θ).

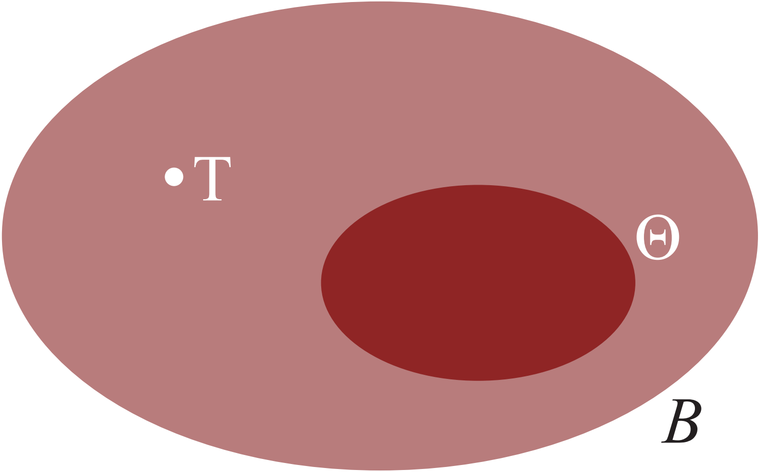
The whole inference process

Define observables (here the sample of 10 children's PD): The Big World.



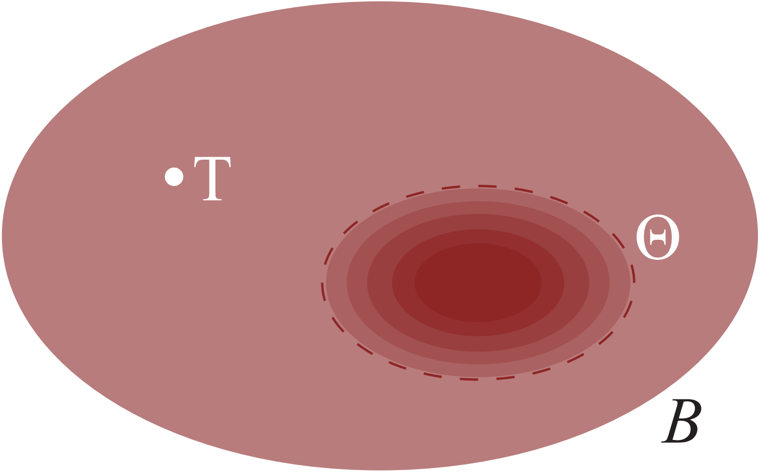
The whole inference process

Specify a likelihood, for example $X \sim \text{binomial}(10, \theta)$ to define the Small World: $\theta \in \Theta$.



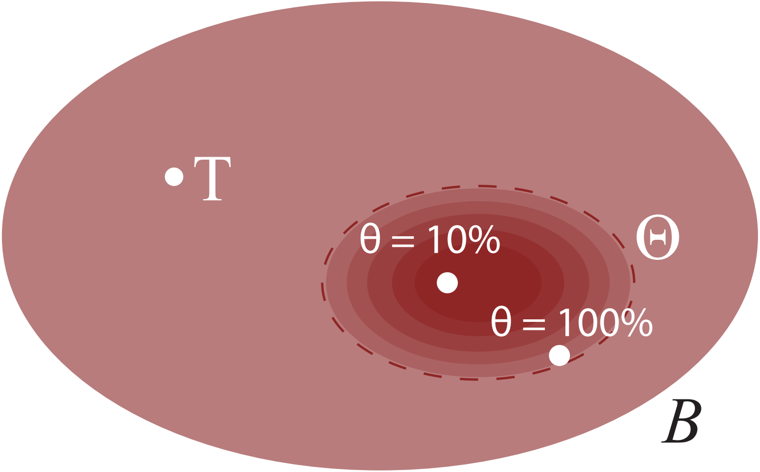
The whole inference process

Specify a prior.



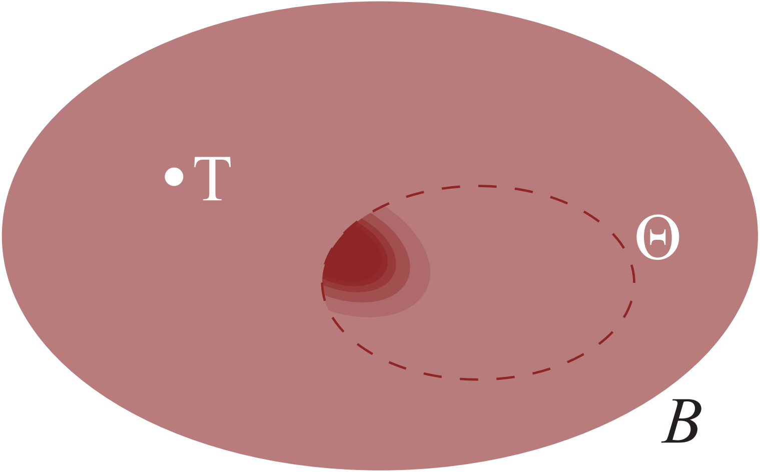
The whole inference process

Specify a prior.



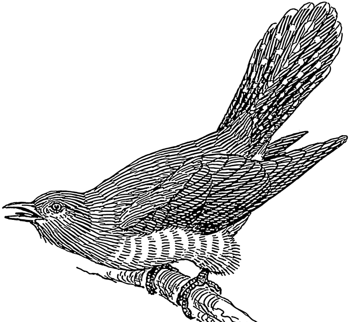
The whole inference process

Input the data.



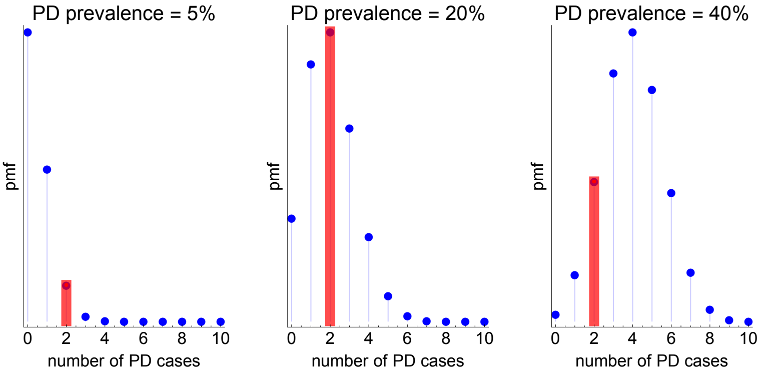
Inference: going from effect back to cause

- We find that 2 out of a sample of 10 children have a true biological father different to their supposed biological father.
    - (Not too far from figure of 10% given by researchers in a recent meta-analysis of studies of "paternal discrepancy" (PD).)
- The question is what was the "cause" of this "effect"?



Potential cuckolding realities

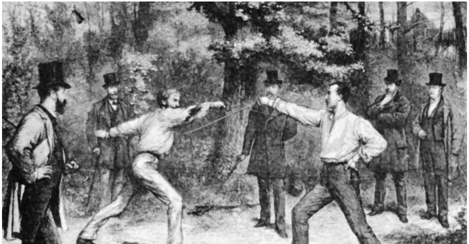
Effect: 2/10 children with Paternal discrepancy (PD).
Potential causes:



Two rival ways of going from effect to cause

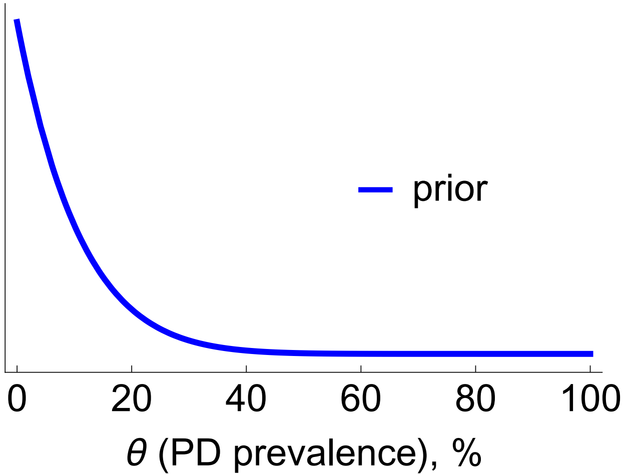
- **Frequentists:**
  - If $Pr(\text{effect}|\text{cause})$ is small (typically $< 5\%$) then reject cause $\implies Pr(\text{cause}|\text{effect}) = 0$.
  - Otherwise, $Pr(\text{cause}|\text{effect}) =?$
- **Bayesians:** use Bayes' rule to directly calculate $Pr(\text{cause}|\text{effect})$:

$$Pr(\text{cause}|\text{effect}) = \frac{Pr(\text{effect}|\text{cause}) \times Pr(cause)}{Pr(effect)} \tag{1}$$



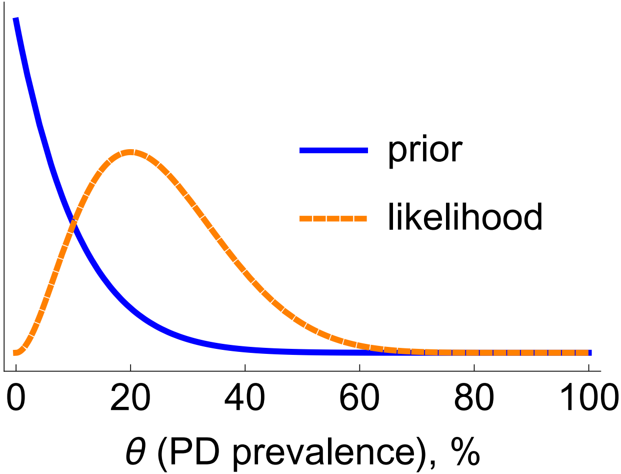
Bayesian cause from effect

Give a prior weighting to all causes (within Small World):



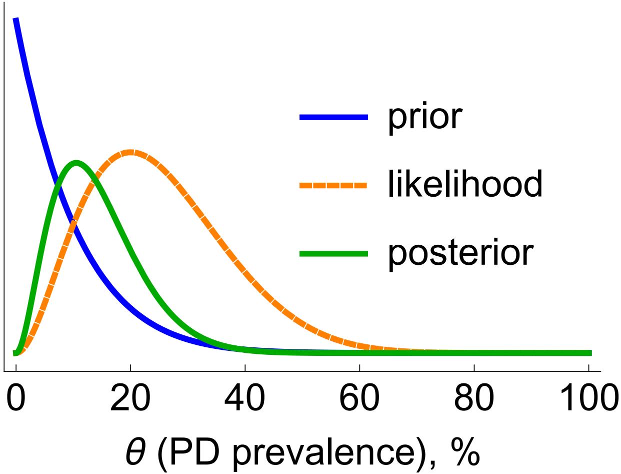
Bayesian cause from effect

Collect the data (2/10 PD) and input to likelihood:

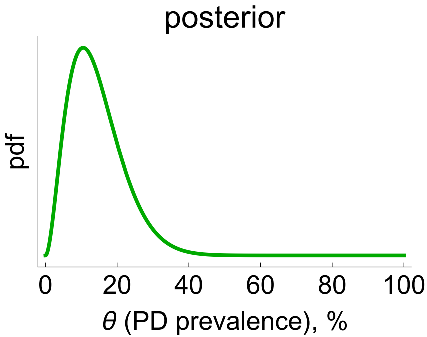


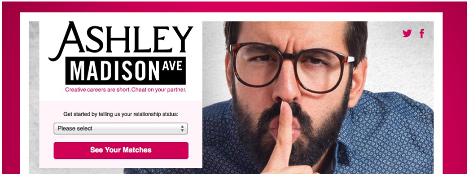
Bayesian cause from effect

Use Bayes' rule to calculate posterior:



Posterior: a weighting across all causes





Get started by telling us your relationship status:

Please select

1. Previous lecture recap
2. Elements of Bayes' rule for inference
3. Why is exact Bayesian statistics hard?
4. Making life simpler: Conjugate priors

Likelihood summary

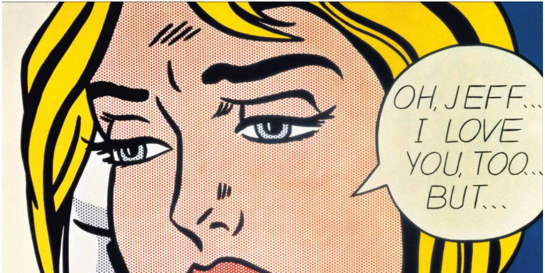
$$p(\theta|X) = \frac{\boxed{\color{blue}p(X|\theta)} \times p(\theta)}{p(X)} \tag{2}$$

- In our example θ is the rate of PD.
- Here X is the data.
- $p(X|\theta)$ represents the *likelihood*.
- Remember *not* a probability distribution because θ varies.
- Most important choice $\implies$ Small World boundaries.
- Encapsulates many **subjective** judgements about analysis.

Paternal discrepancy revisited: choosing a likelihood

Suppose we now have two samples: 2/10 and 0/10 cases of paternal discrepancy (PD). Start by listing assumptions about the world:

- One event of paternal discrepancy is **independent** of all others.
- The underlying rate of paternal discrepancy is the same for the wider population from which the samples were drawn (**identically distributed**).



Paternal discrepancy revisited: choosing a likelihood

These assumptions can of course be challenged:

- **Independence:** violated if for example there are multiple instances of PD within one family.
- **Identically distributed:** more subtle since what is meant by the "population", but could be violated if there are important differences between the two samples.

Paternal discrepancy revisited: calculating the likelihood

We have the following conditions:

- **Independent** and **identically distributed** observations.
- Discrete observable - the number of PD cases.
- Fixed sample size for each sample.

$\Longrightarrow$ **Binomial** distribution.

Paternal discrepancy revisited: calculating the likelihood

For our sample of 2/10 with PD we can find the likelihood using the **equivalence relation** (where θ is the underlying prevalence of PD in the population):

"The **likelihood** for 2 PD cases as a function of θ is equivalent to the **probability** of 2 PD cases given θ"

$\Longrightarrow$

$$\begin{aligned} L(\theta|X=2) &= Pr(X=2|\theta) \\ &= \binom{10}{2} \times \underbrace{\theta^2}_{\text{2 positives}} \times \underbrace{(1-\theta)^{10-2}}_{\text{8 negatives}} \end{aligned}$$

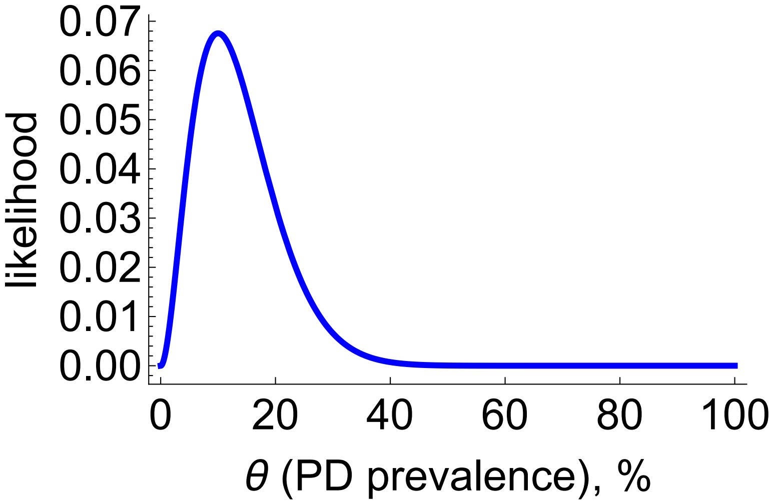
Paternal discrepancy revisited: calculating the likelihood

Since assume samples are (conditionally) independent:

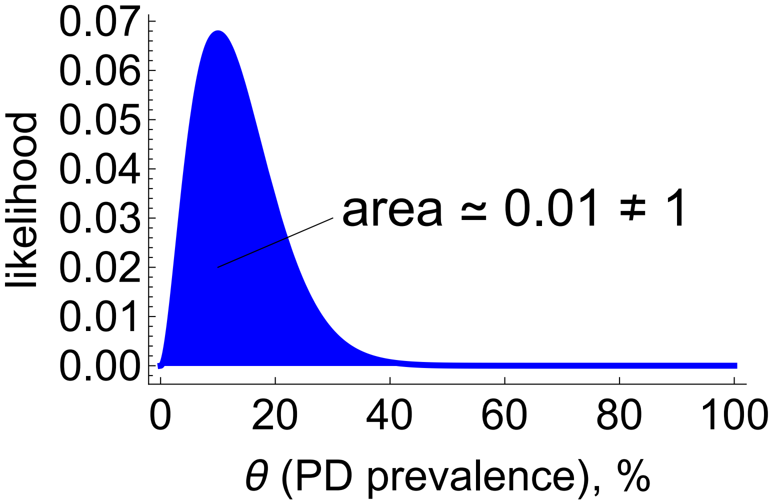
$$\begin{aligned} L(\theta|X_1 = 2, X_2 = 0) &= Pr(X_1 = 2, X_2 = 0|\theta) \\ &= Pr(X_1 = 2|\theta) \times Pr(X_2 = 0|\theta) \\ &= \binom{10}{2}\theta^2(1-\theta)^8 \times \binom{10}{0}\theta^0(1-\theta)^{10} \\ &= 45 \times \theta^2(1-\theta)^{18} \end{aligned}$$

Resultant likelihood the same as that from a Binomial$(20, \theta)$ distribution (due to independence).

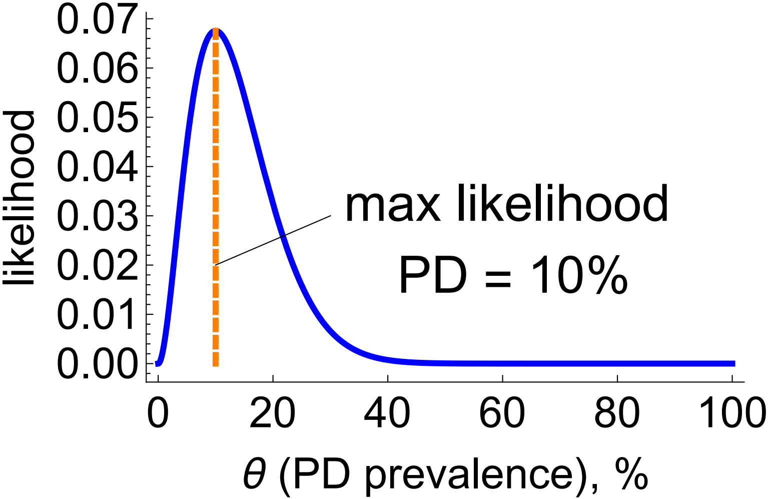
Paternal discrepancy revisited: graphing the likelihood



Paternal discrepancy revisited: graphing the likelihood



Paternal discrepancy revisited: graphing the likelihood



max likelihood
PD = 10%

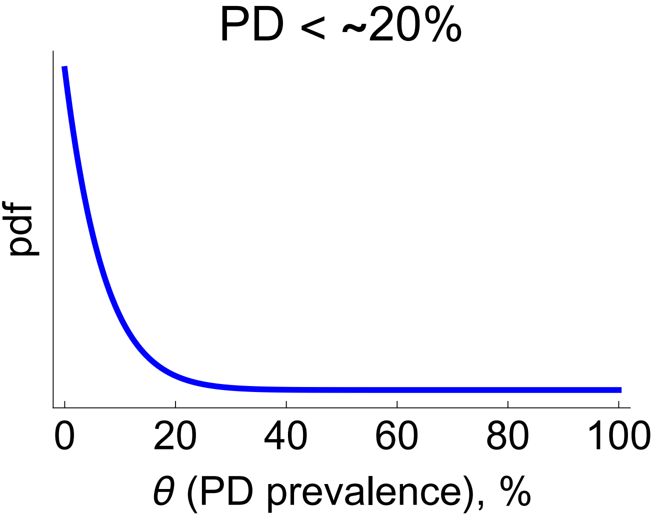
Priors summary

$$p(\theta|X) = \frac{p(X|\theta) \times \enclose{circle}{p(\theta)}}{p(X)} \tag{3}$$

- $p(\theta)$ represents the *prior*.
- A valid probability distribution.
- Determines which areas of the Small World we believe are most likely to contain the true data generating process.
- Similar to the likelihood; it is also subjective.

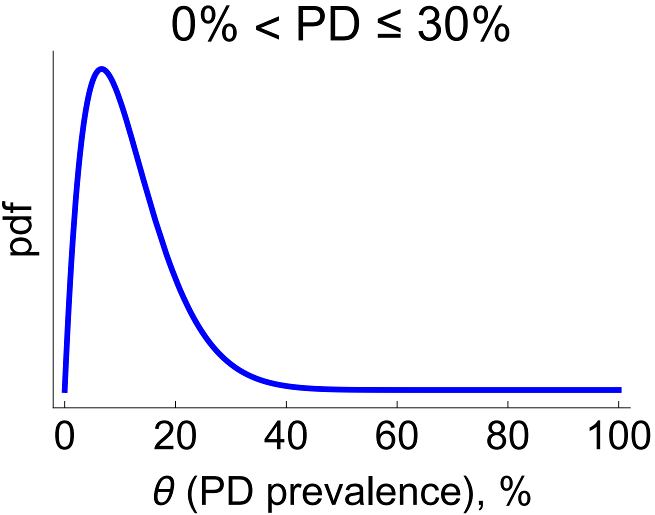
Arriving at a choice of prior

What prevalences of PD do we believe are most likely?



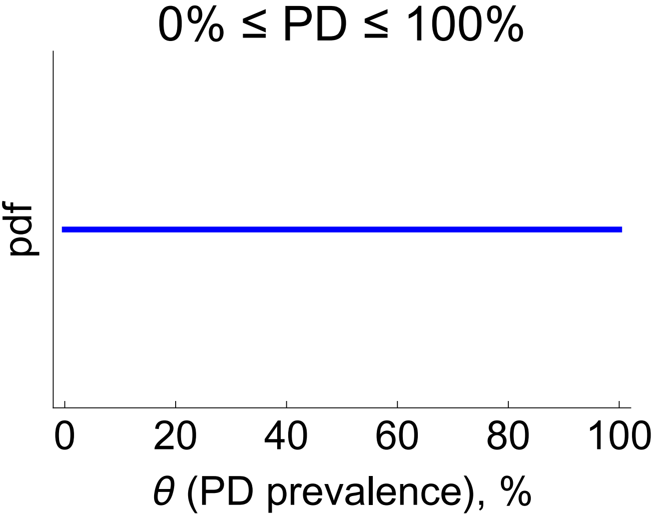
Arriving at a choice of prior

What prevalences of PD do we believe are most likely?



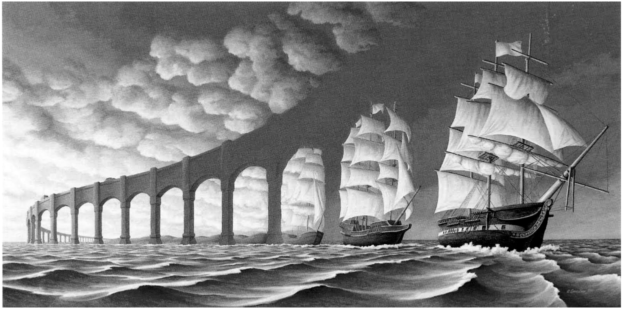
Arriving at a choice of prior

What rates of PD do we believe are most likely?



No "objective" rule for priors

- Embody subjective assumptions about state of the world.
- Essentially measure $Pr(\text{cause}|\text{pre-data knowledge})$.
  - Since knowledge differs between subjects $\implies$ different priors.
- Can be informed by pre-experimental data (for example, previous studies or from a collection of previous studies).



Why do we need priors at all?

Can’t we just allow the data to “speak for itself”?

$$\textit{initial belief} \xrightarrow{\textit{Bayes' rule}} \textit{new beliefs} \tag{4}$$

No. Bayes’ rule only provides us with a way to *update* beliefs. For example,

$$\text{“There are WMDs in Iraq.”} \xrightarrow{\text{Go to Iraq + BR}} \text{“There are NO WMDs in Iraq.”}$$

Can't we just use a unity prior?

Set a uniform prior of the form: $p(\theta) = 1$, then Bayes' rule becomes,

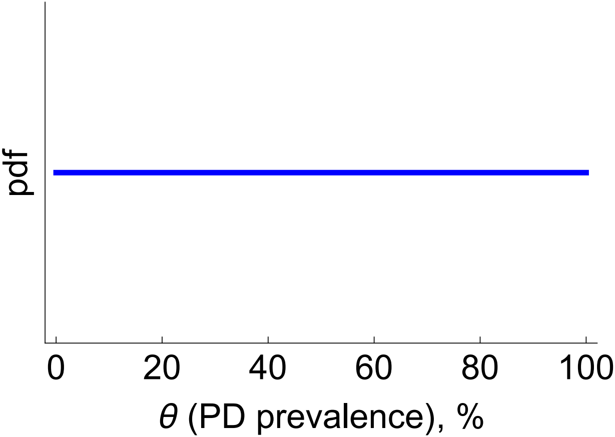
$$p(\theta|X) = \frac{p(X|\theta)}{p(X)} \tag{5}$$

Three arguments against this:

- *Mathematical pedantry*: in general $p(\theta) = 1$, is not a valid probability distribution $\implies$ cannot guarantee that posterior behaves as a valid probability distribution.
- *Reduced variance*: whilst our beliefs may induce a bias, this is more than compensated for by a reduction in variance.
- *Avoid nonsensical answers*: $p(\theta) = 1$ does not normally reflect common sense.

“Uninformative” priors

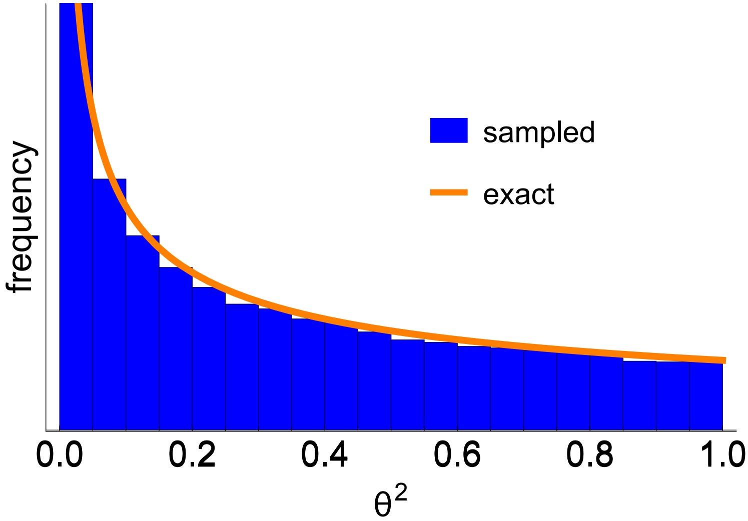
θ represents the probability that one child chosen at random has paternal discrepancy.



“Uninformative” priors

- **Question 1:** What is the probability that 2/2 children have PD? **Answer 1:** $\theta \times \theta = \theta^2$
- **Question 2:** Assuming a uniform prior for θ, what does $p(\theta^2)$ look like?
- **Answer 2:** use sampling! First sample $\theta \sim$ uniform$(0, 1)$, then look at distribution of θ^2.

“Uninformative” priors



The illusion of uninformative priors

- Even priors that appear uninformative in one frame of reference may not in another!
- A more useful concept is that of a "weakly informative" prior.
    - Give most weight to "reasonable" parameter values whilst still allowing considerable freedom.
    - Can considerably help with sampling (lecture 5 and 6: by Gelman's "Folk Theorem").

Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{\enclose{circle}{p(X)}} \tag{6}$$

- $p(X)$ represents the *denominator*.
- Two different interpretations:
  - Before we collect X it is the **prior predictive distribution**.
  - When we have data $X = 2$ it is simply a number (that normalises the posterior).
- Calculated from the numerator.
- Source of some of the difficulty of **exact** Bayesian inference (return to this later).

The denominator before we get data: the prior predictive distribution

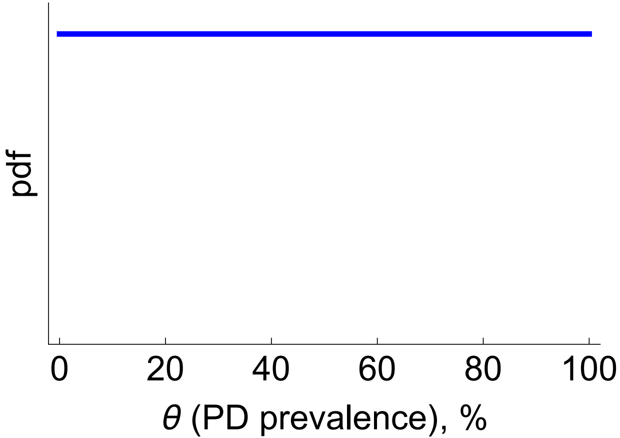
Assume we choose a Binomial likelihood, $p(X|\theta)$, and uniform prior, $p(\theta) = 1$. These form a joint distribution:

$$p(X,\theta) = \overbrace{p(X|\theta)}^{\textit{sampling distribution}} \times \overbrace{p(\theta)}^{\textit{prior}} \tag{7}$$

- $p(X)$ is the probability of obtaining data X conditional on our choice of likelihood and prior.
- Get the marginal distribution $p(X)$ from the joint distribution $p(X,\theta)$ by integration but far easier to use **sampling**.
- But how do we do this?

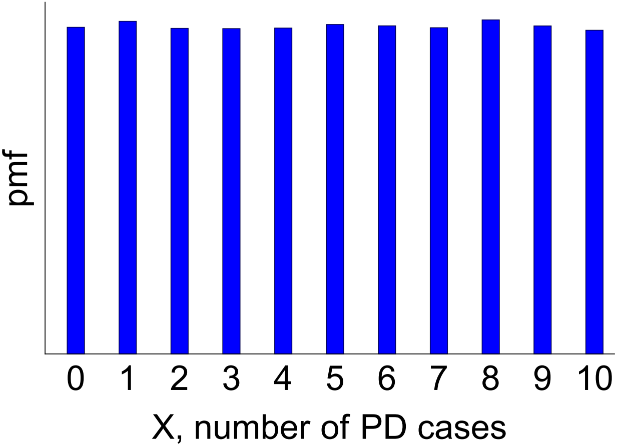
The prior predictive distribution by sampling

First sample $\theta_i \sim p(\theta)$; i.e. the prior.



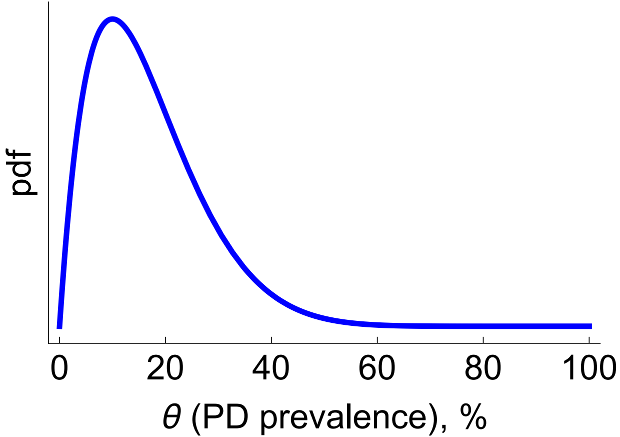
The prior predictive distribution by sampling

Then sample $X_i \sim p(X|\theta_i)$; i.e. the sampling distribution (different from likelihood since θ_i is fixed).



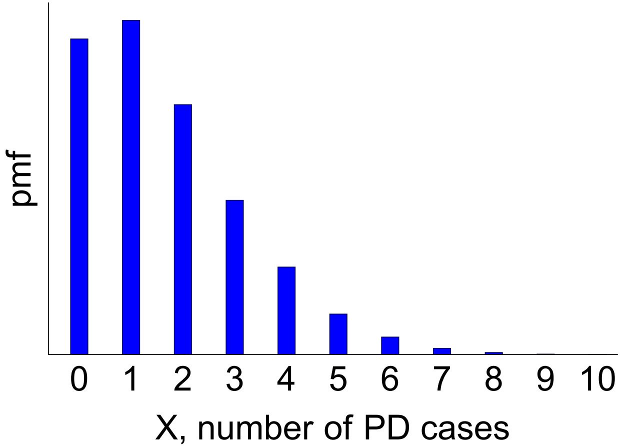
The prior predictive distribution by sampling

For a more informative prior.

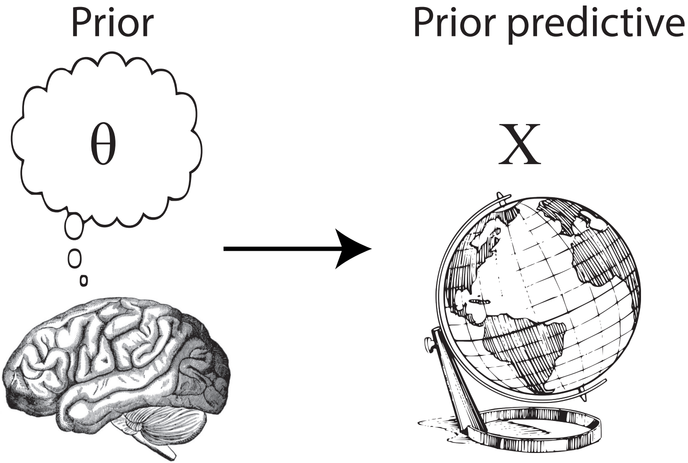


The prior predictive distribution by sampling

Prior predictive similarly peaked.



The prior predictive distribution: from "conceptual" to "observable" pre-data world



Prior

θ

Prior predictive

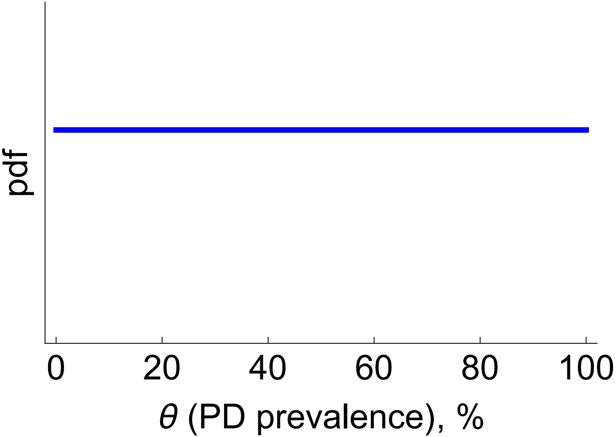
X

The denominator as a normalising factor

- When we obtain data, $X = 2$, the denominator collapses to a probability $Pr(X = 2)$.
- This number normalises the numerator ensuring that the posterior is a valid density.
- (This is part of the difficulty in exact Bayesian inference.)

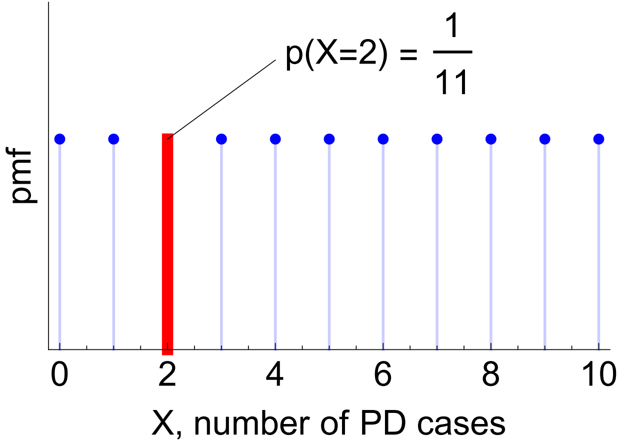
The denominator as a normalising factor

Consider flat prior again.



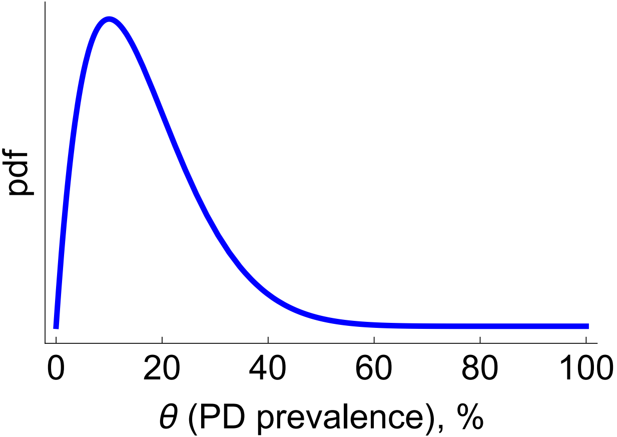
The denominator as a normalising factor

Calculate probability from the prior predictive distribution.



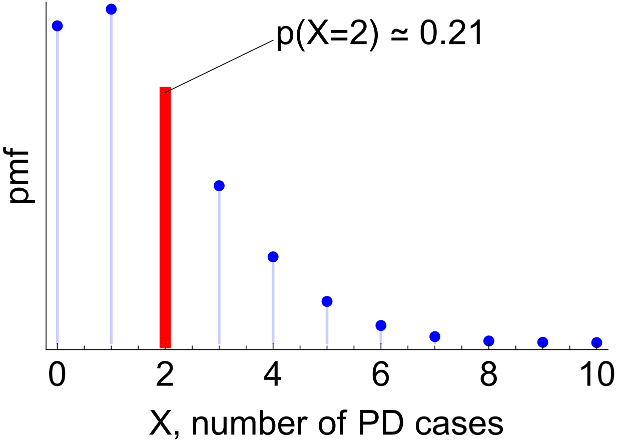
The denominator as a normalising factor

Consider the informative prior again.



The denominator as a normalising factor

Calculate probability from the prior predictive distribution.



Posteriors summary

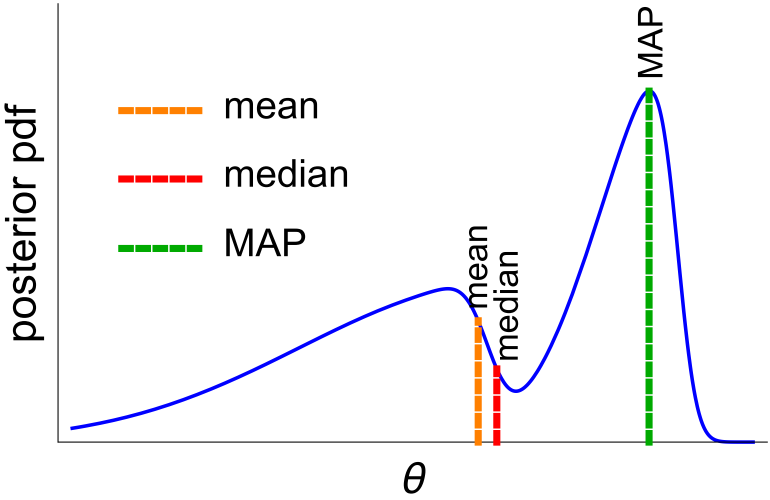
$$\enclose{circle}{p(\theta|X)} = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{8}$$

- $p(\theta|X)$ represents the *posterior*.
- A valid probability distribution.
- Starting point for all further analysis in Bayesian inference.

Posterior point estimates

- Mathematical models and policy makers often require point estimates of parameters.
- In Bayesian inference there are choices for estimates:
  - Posterior mean.
  - Posterior median.
  - Maximum *a posteriori* (MAP); also known as the *mode*.
- (Statistical decision theory: under different loss functions each can be “optimal”.)
- However, generally prefer posterior mean or median over MAP.
  - MAP ignores the measure by focusing solely on density.
  - (Linked) MAP can lie a long way from probability mass.

Posterior point estimates

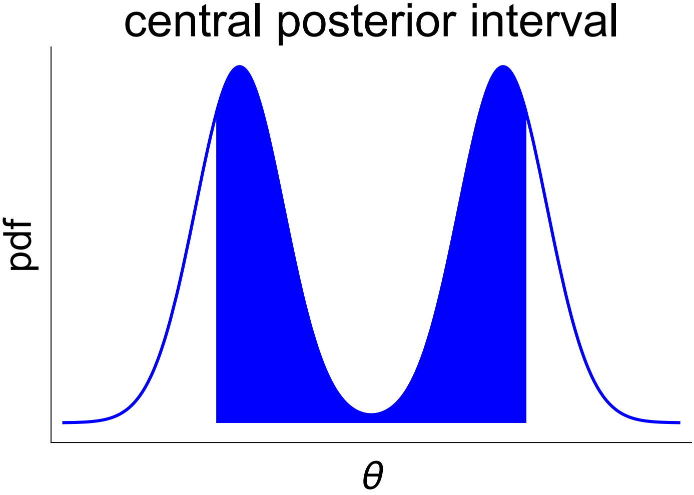


Posterior summaries

- Prefer estimates incorporating uncertainty over point estimates.
- In Bayesian statistics: central posterior interval (CPI) versus highest density interval (HDI).
- (The “optimal” choice can again be determined via decision theory.)

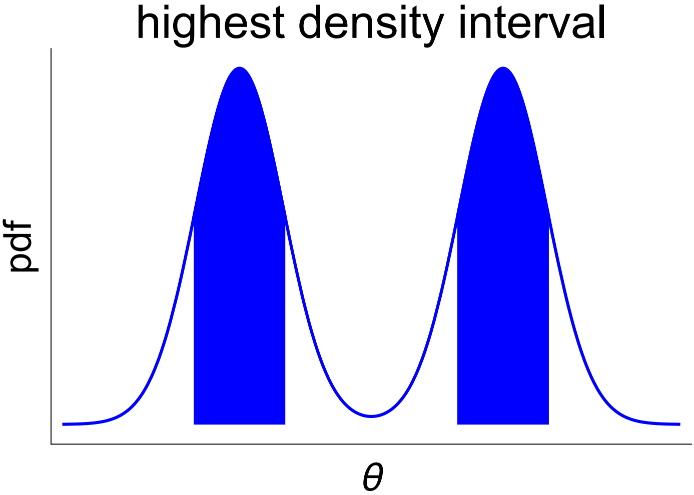
Posterior summaries

If contiguity of regions is important $\Longrightarrow$



Posterior summaries

If avoiding nonsensical (low density) regions is important $\Longrightarrow$



Intuition behind Bayesian analyses

Bayes' rule:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{9}$$

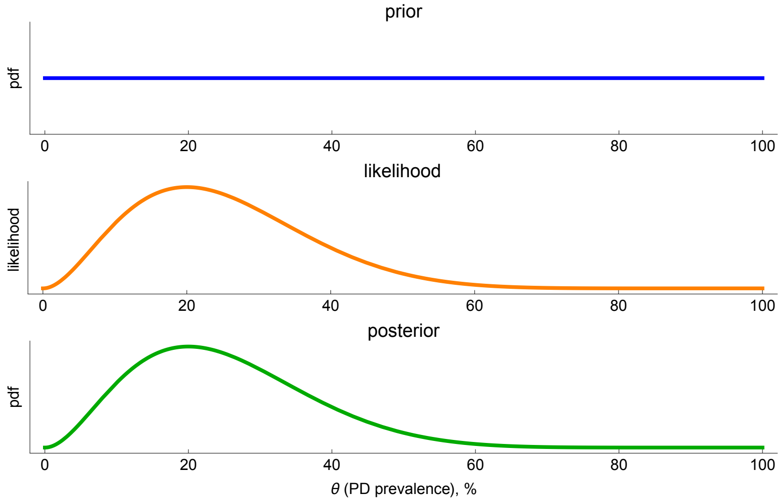
Tells us that:

$$p(\theta|X) \propto p(X|\theta) \times p(\theta) \tag{10}$$

Because $p(X)$ is independent of θ
$\implies$ the posterior is a essentially a weighted (geometric) mean of the prior and likelihood.

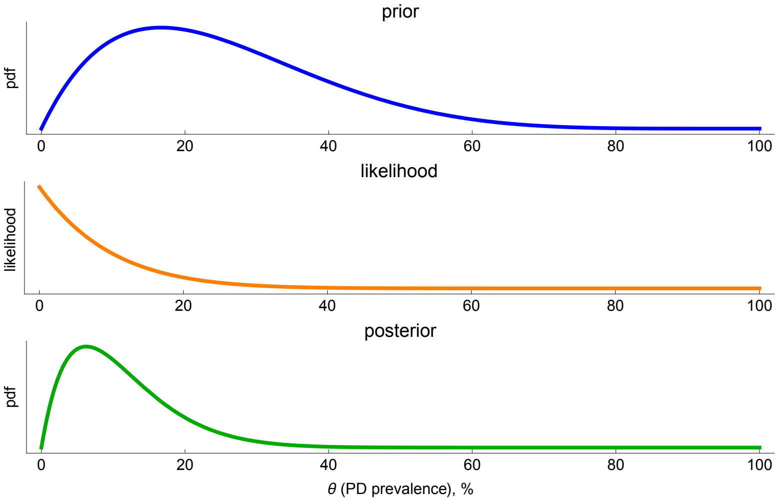
Intuition behind Bayesian analyses: PD rate again

Consider single sample of 10 children; 2 of which have PD.



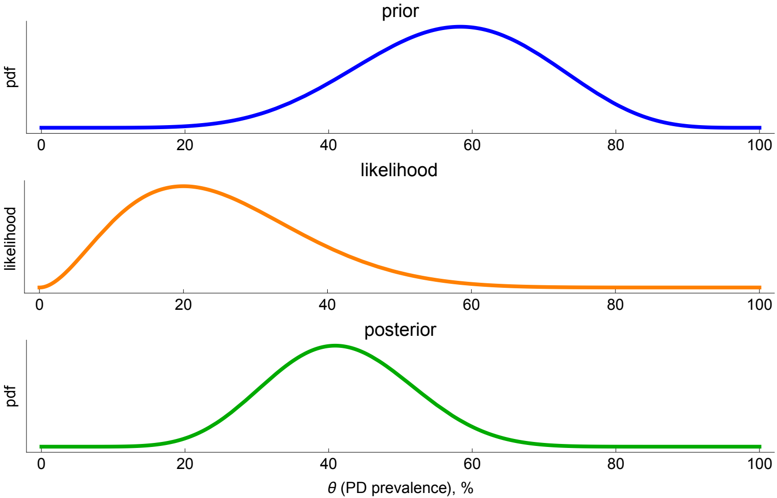
Intuition behind Bayesian analyses: PD rate again

Now holding prior constant and varying proportion with PD.

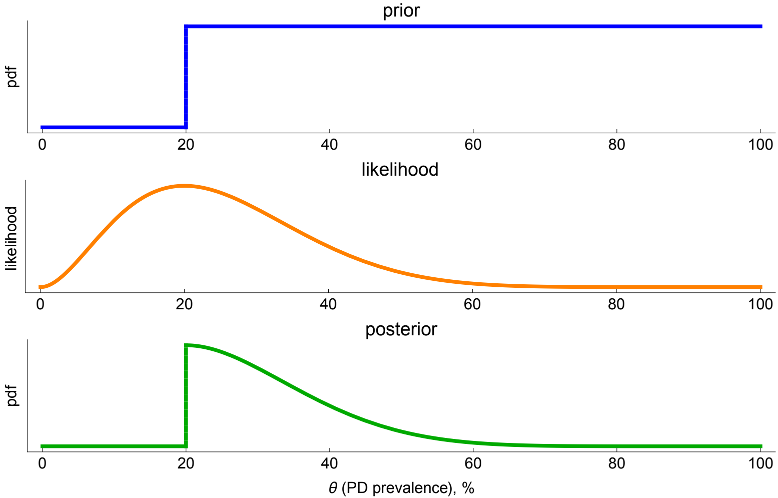


Intuition behind Bayesian analyses: PD rate again

Constant prior and proportion with PD (20%); sample size↑.



An exception: zero priors (avoid these)



Intuition behind Bayesian analyses: summary

- The posterior is a weighted average of the prior and likelihood (data).
- Changes in position of prior or likelihood are reflected in posterior.
- The weighting towards the likelihood increases as more data is collected $\implies$ models with a lot of data are less dependent on priors.
- Exception to this is “zero” priors.

Forecasting

- Consider a new data sample $\tilde{X}$.
- Want to find $p(\tilde{X}|X)$; the probability of the new data sample given our current data X.
- We call $p(\tilde{X}|X)$ the **posterior predictive distribution**, and can be used:
  - To forecast.
  - To check model.
- Similar logic as that of the prior predictive distribution, but using the posterior instead.

Prior vs posterior predictive distributions

To obtain **prior** predictive distribution $p(X)$ we sampled from the joint distribution:

$$p(X, \theta) = \overbrace{p(X|\theta)}^{\textit{sampling distribution}} \times \overbrace{p(\theta)}^{\textit{prior}} \tag{11}$$

We did this stepwise:

1. Sample $\theta_i \sim p(\theta)$; i.e. from the prior.
2. Sample $X_i \sim p(X|\theta_i)$; i.e. from the sampling distribution.

Prior vs posterior predictive distributions

To obtain **posterior** predictive distribution $p(\tilde{X}|X)$ we sample from the joint distribution:

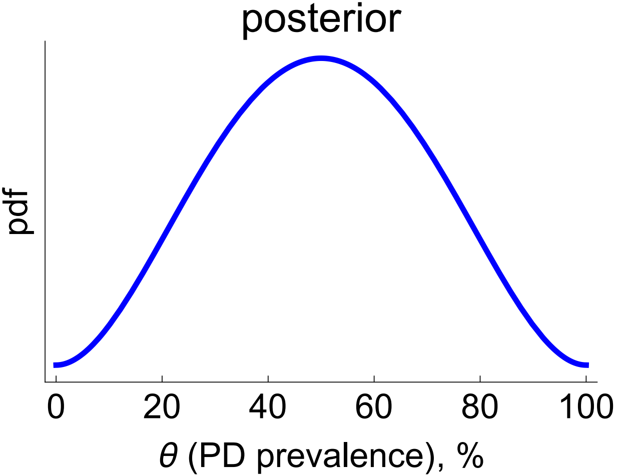
$$
\begin{aligned}
p(\tilde{X},\theta|X) &= p(\tilde{X}|\theta,X) \times p(\theta|X) \\
&= \overbrace{p(\tilde{X}|\theta,\cancel{X})}^{\textit{independent}} \times p(\theta|X) \\
&= \overbrace{p(\tilde{X}|\theta)}^{\textit{sampling distribution}} \times \overbrace{p(\theta|X)}^{\textit{posterior}}
\end{aligned}
$$

Again do this stepwise:

1. Sample $\theta_i \sim p(\theta|X)$; i.e. from the posterior.
2. Sample $\tilde{X}_i \sim p(\tilde{X}|\theta_i)$; i.e. from the sampling distribution.

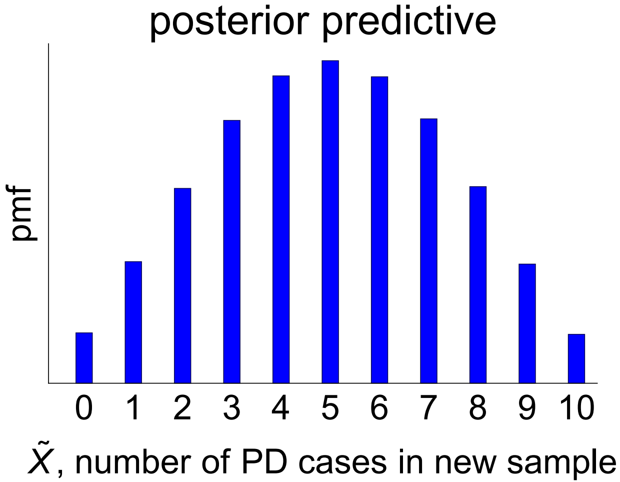
Posterior predictive distribution

1. Sample θ_i from posterior.



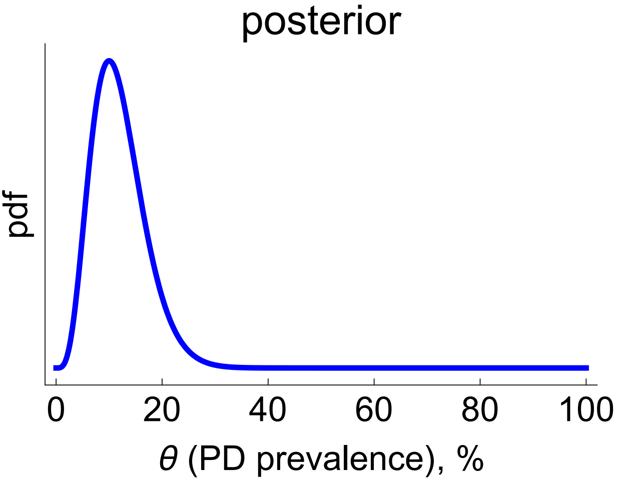
Posterior predictive distribution

2. Sample $\tilde{X}_i$ from sampling distribution $\Longrightarrow$



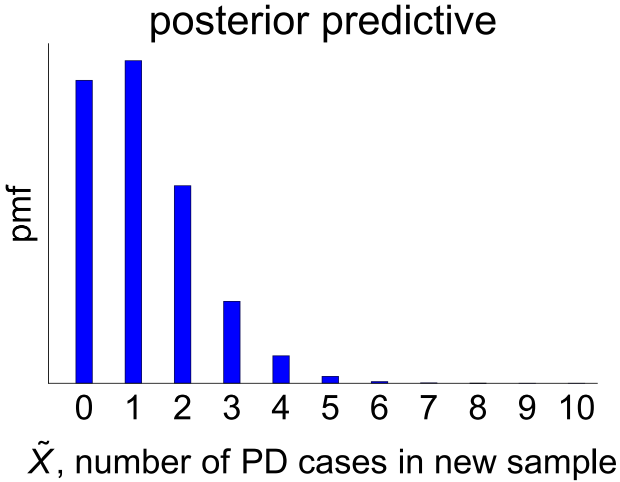
Posterior predictive distribution

A more concentrated posterior...



Posterior predictive distribution

…yields a narrower posterior predictive range.



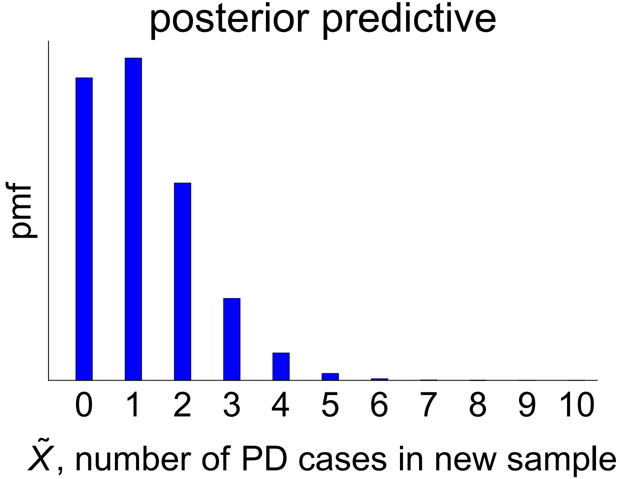
Posterior predictive distribution: uses

Why should we estimate this distribution?

- **Forecasts:**
  - A valid probability distribution.
  - $\implies$ no extra work to obtain predictive intervals.
- **Check model's suitability:**
  - Use posterior predictive distribution to obtain "simulated" data.
  - If model fits data $\implies$ should "look" like real data.
  - Exhaustive and creative way of checking **any** aspect of a model (come back to this next lecture).

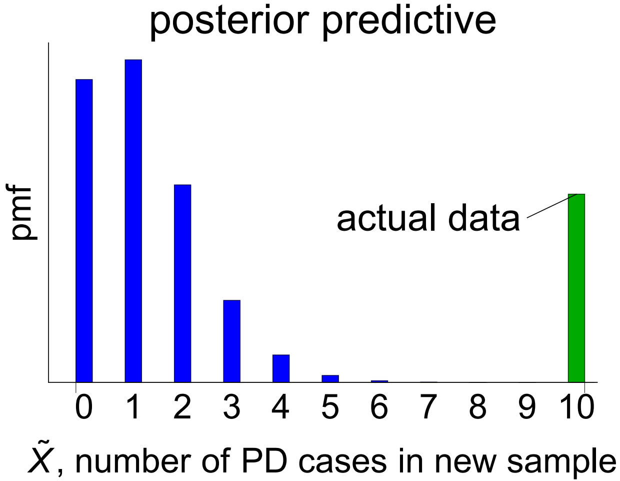
Quick introduction to posterior predictive checks

Start with posterior predictive distribution.

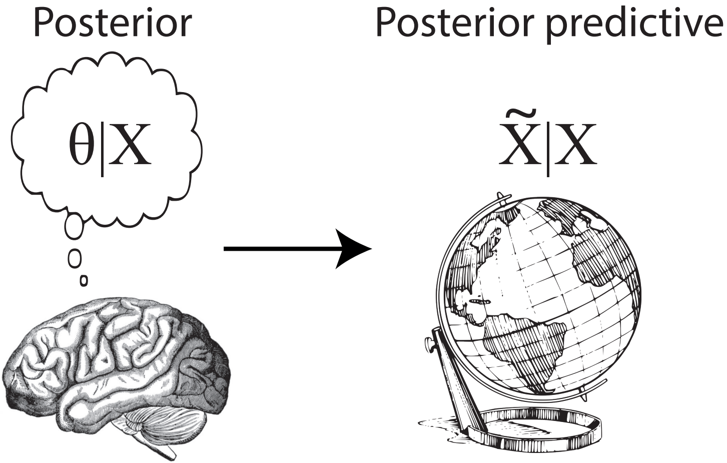


Quick introduction to posterior predictive checks

Compare with actual data! (More next time.)



The posterior predictive distribution: from “conceptual” to “observable” post-data world



Posterior

$\theta|X$

Posterior predictive

$\tilde{X}|X$

Bayes' rule for inference: summary

Bayes' rule for inference requires that we specify:

- **Likelihood:** determines the probability model framework we consider (Small World boundaries).
- **Prior:** subjective measure of our belief in certain parameter values (informed by past data/experience).

$\Longrightarrow$ **Posterior** is a weighted average of these two; in general the more data the higher the weighting towards the likelihood.

Bayes' rule for inference: summary

- The likelihood and prior determine the **prior predictive density** $\implies$ useful for analysing the implications of prior "conceptions" in the real world.
- The likelihood and posterior determine the **posterior predictive density**. Useful for:
  - **Forecasting:** automatically gives predictive intervals for parameters.
  - **Checking a model:** compare simulated data with actual. If different then the model is deficient in some way (more next time).

1. Previous lecture recap

2. Elements of Bayes' rule for inference

3. Why is exact Bayesian statistics hard?

4. Making life simpler: Conjugate priors

The denominator revisited

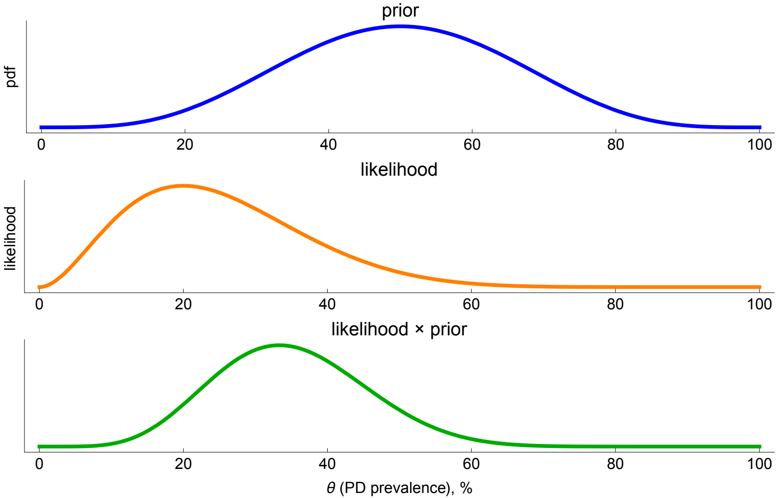
$$p(\theta|X=2) = \frac{p(X=2|\theta) \times p(\theta)}{\enclose{circle}{\color{blue}{p(X=2)}}} \tag{12}$$

Where we suppose we have data $X = 2$ out of a sample of 10 in our PD example. We obtain the denominator by averaging out all θ dependence. This is equivalent to integrating across all θ:

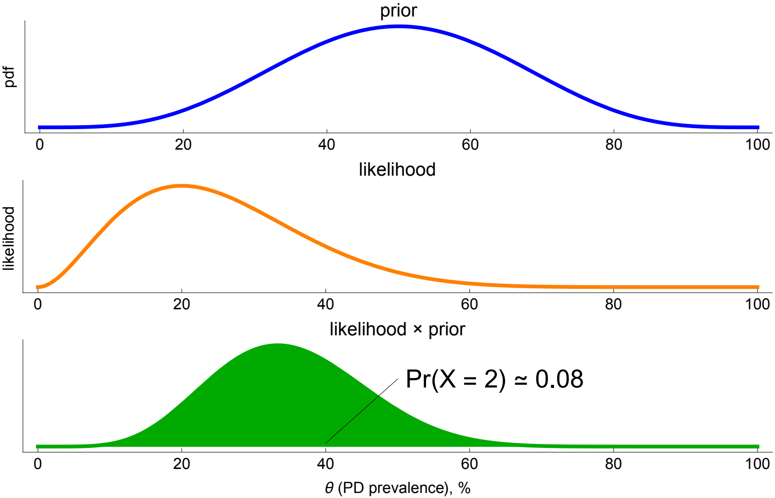
$$p(X=2) = \int_{0}^{1} p(X=2|\theta) \times p(\theta)\mathrm{d}\theta \tag{13}$$

(We approximately determined this using sampling previously.)

The denominator as an area



The denominator as an area

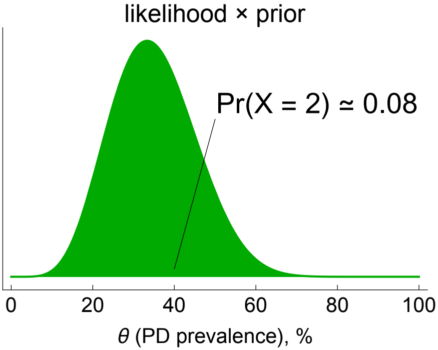


Calculating the denominator in 1 dimension

For our PD example there is a single parameter $\theta \implies$

$$p(X=2) = \int_0^1 p(X=2|\theta) \times p(\theta)\mathrm{d}\theta \tag{14}$$

This is equivalent to working out an **area** under a curve.



Calculating the denominator in 2 dimensions

If we considered a different model where there were two parameters $\theta_1 \in (0,1),\ \theta_2 \in (0,1) \implies$:

$$p(X=2) = \int_0^1 \int_0^1 p(X=2|\theta_1,\theta_2) \times p(\theta_1,\theta_2)\mathrm{d}\theta_1\mathrm{d}\theta_2 \tag{15}$$

This is equivalent to working out a **volume** contained within a surface.

Calculating the denominator in d dimensions

If we considered a different model where there were d parameters $(\theta_1, ..., \theta_d)$ all defined to lie between 0 and 1 $\implies$:

$$p(X=2) = \int_0^1 ... \int_0^1 p(X=2|\theta_1, ..., \theta_d) \times p(\theta_1, ..., \theta_d) \mathrm{d}\theta_1 ... \mathrm{d}\theta_d \tag{16}$$

This is equivalent to working out a $(d+1)$-dimensional **volume** contained within a d-dimensional (hyper-surface)!

The difficult denominator

- Calculating the denominator possible for $d <\sim 20$ using computers.
- Numerical quadrature and many other approximate schemes struggle for larger d.
- Many models have **thousands** of parameters.

Arrrghhh!

Other difficult integrals

Assume we can calculate posterior:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{17}$$

Typically we want summary measures of posterior, for example, the mean of θ_1:

$$\begin{aligned}\mathrm{E}(\theta_1|X) &= \int\limits_{\Theta_1} \theta_1 \left[\int\limits_{\Theta_2} ... \int\limits_{\Theta_d} p(\theta_1, \theta_2, ..., \theta_d|X) \mathrm{d}\theta_d ... \mathrm{d}\theta_2 \right] \mathrm{d}\theta_1 \\ &= \int\limits_{\Theta_1} \theta_1 \; p(\theta_1|X) \mathrm{d}\theta_1\end{aligned}$$

As difficult as denominator!

Problems with exact Bayesian inference, and solutions

Problems:

- In general we cannot calculate the denominator of Bayes' rule $\implies$ no analytic posterior.
- To use posterior for most further analyses we need to do more integrals! $\implies$ Even if we can calculate posterior we still run into problems.

Solutions:

- *Conjugate priors*: use simple-fitting distributions where we can follow basic rules to find posteriors (no calculation necessary!)
- *Sampling*: understand a distribution by sampling from it, rather than exact calculation (next time).

1. Previous lecture recap

2. Elements of Bayes' rule for inference

3. Why is exact Bayesian statistics hard?

4. Making life simpler: Conjugate priors

What are conjugate priors?

Judicious choice of prior and likelihood can make posterior calculation trivial.

- Choose a likelihood L.
- Choose a prior $\theta \sim f \in F$, where:
  - F is a family of distributions.
  - f is a member of that **same** family.
- If posterior, $\theta|X \sim f' \in F \implies$ conjugate!
- In other words both the **prior** and **posterior** are members of the same distribution!

Conjugate priors: PD example revisited

Sample 10 children and count number (X) with PD:

- For likelihood (if independent and identically-distributed):

$$X \sim \mathit{Binomial}(10, \theta) \implies p(X|\theta) \propto \theta^{X}(1-\theta)^{10-X} \tag{18}$$

- For prior assume a Beta distribution (a reasonable choice if $\theta \in (0, 1)$):

$$\theta \sim \mathit{Beta}(\alpha, \beta) \implies p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{19}$$

- Numerator of Bayes' rule for inference:

$$p(X|\theta) \times p(\theta) \propto \theta^{X}(1-\theta)^{10-X} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{20}$$

Conjugate priors: PD example revisited

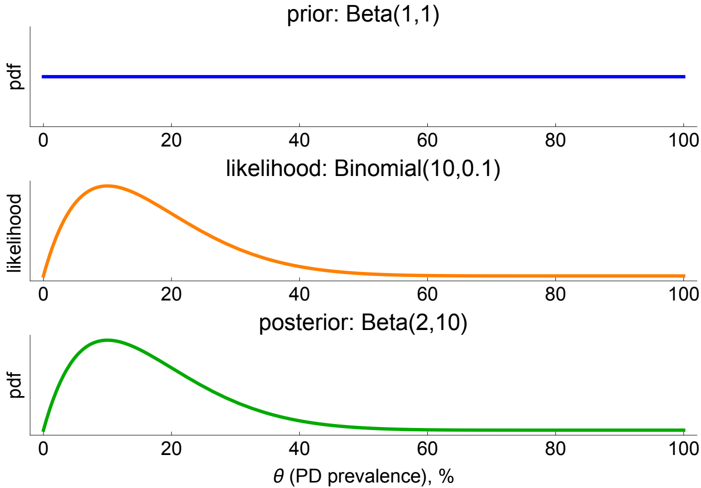
- Numerator of Bayes' rule for inference:

$$
\begin{aligned}
p(X|\theta) \times p(\theta) &\propto \theta^{X}(1-\theta)^{10-X} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{X+\alpha-1}(1-\theta)^{10-X+\beta-1}
\end{aligned}
$$

- This has same θ-dependence as a $\mathit{Beta}(X+\alpha, 10-X+\beta)$ density $\implies$ must be this distribution!
- $\therefore$ a Beta prior is *conjugate* to a Binomial likelihood.

Conjugate priors: PD example revisited

Varying the prior.



Conjugate priors: PD example revisited

Varying the data.

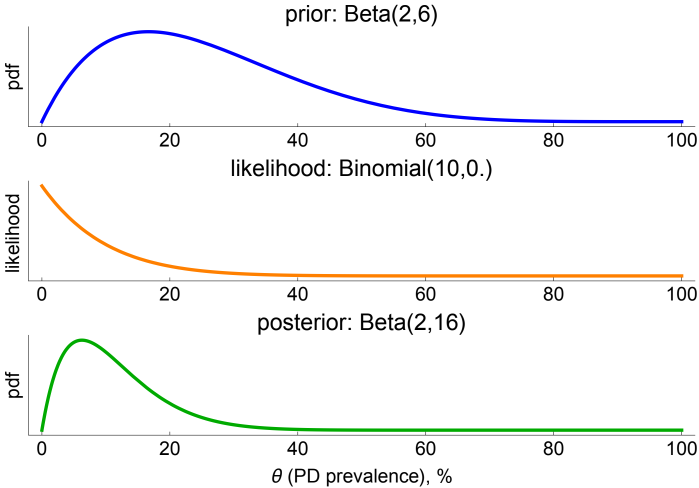


Table of common conjugate pairs of likelihoods and priors

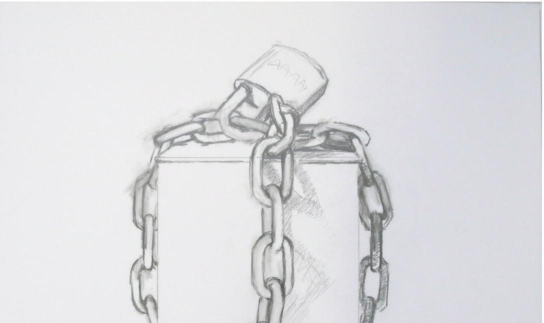
No need to do any integrals! Just lookup rules:

Likelihood	**Prior**	**Posterior**
Bernoulli	Beta(α, β)	Beta$(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i)$
Binomial	Beta(α, β)	Beta$(\alpha + \sum_{i=1}^{n} X_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} X_i)$
Poisson	Gamma(α, β)	Gamma$(\alpha + \sum_{i=1}^{n} X_i, \beta + n)$
Multinomial	Dirichlet$(\boldsymbol{\alpha})$	Dirichlet$(\boldsymbol{\alpha} + \sum_{i=1}^{n} \boldsymbol{X}_i)$
Normal	Normal-inv-Γ	Normal-inv-Γ

Limits of conjugate modelling

Using conjugate priors is limiting because:

- Restricted to univariate (or in some cases bivariate) problems.
  - $\implies$ we could just use numerical quadrature instead.
- Required to use relevant conjugate prior for a given likelihood $\impliedby$ may not be sufficient to capture pre-data beliefs of analyst.



Conjugate priors: summary

- Conjugate priors available for quite a few univariate distributions.
- For parameter dimension $d > 2$ this is generally not possible.
- In many cases conjugate priors are overly restrictive.
- Modelling difficulty $\implies$ model choice $\therefore$ not ideal.

Summary

- Posterior is a weighted average of prior and likelihood, where weight of likelihood determined by amount of data.
- Prior and posterior predictive distributions show implications of the prior and posterior on the observable world.
- Exact Bayes is hard due to difficulty of calculating posterior, and other high dimensional integrals.
- Conjugate priors can make analysis simpler, although are highly restrictive.

Reading list

Directly relevant:

- Chapters 2 (single parameter models) and 3 (multiparameter models) from "Bayesian data analysis", Gelman et al. (2014), 3rd edition.
- Chapters 1 (modelling) and 2 (posteriors etc.) from "Statistical Rethinking", by McElreath (2016).

Reading list

Bits on the side:

- “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures”, Golder and Macy (2011), *Science*.
- “Detecting influenza epidemics using search engine query data”, Ginsberg et al. (2009), *Nature*.
- “The Parable of Google Flu: Traps in Big Data Analysis”, Lazer et al. (2014), *Science*.
- “Measuring paternal discrepancy and its public health consequences”, Bellis et al. (2005), *Journal of epidemiology and community health*.

Problem set details

- Problem set posted on `www.ben-lambert.com/bayesian-lecture-slides` along with any additional files.
- Class upstairs in the IT Suite. If in doubt, come along!

Not sure I understand?

Unconfidence interval:

$$1\% \leq \text{my understanding} \leq 10\% \qquad (21)$$

Uncredible interval:

$$> 50\%$$

Answer to the “Monte Carlo” game

- Imagine the three states of the world; you initially pick:
  - A door with a **car** behind it $\frac{1}{4}$ of the time.
  - A door with **nothing** behind it $\frac{1}{2}$ of the time.
  - A door with a **penalty** behind it $\frac{1}{4}$ of the time.
- The game show host then opens a door revealing that it is empty (you always know that he will do this).

Answer to the "Monte Carlo" game

If you remain with your original choice your probability of each of the outcomes is unchanged:

$$(p(\text{car}), p(\text{nothing}), p(\text{penalty})) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) \tag{22}$$

However, if you decide to change then the expected outcomes *given* that you change depends on your the initial state of the world. If you initially picked:

- A **car**: $(p(\text{car}), p(\text{nothing}), p(\text{penalty})) = (0, \frac{1}{2}, \frac{1}{2})$.
- **Nothing**: $(p(\text{car}), p(\text{nothing}), p(\text{penalty})) = (\frac{1}{2}, 0, \frac{1}{2})$.
- A **penalty**: $(p(\text{car}), p(\text{nothing}), p(\text{penalty})) = (\frac{1}{2}, \frac{1}{2}, 0)$.

We can then calculate the *unconditional* probability of each of the outcomes, by taking the above and weighting them by the probability of each state of the world.

Answer to the "Monte Carlo" game

This gives us the following probabilities:
Remain:

$$(p(\text{car}), p(\text{nothing}), p(\text{penalty})) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) \tag{23}$$

Change:

$$\begin{aligned} p(\text{car}) &= \frac{1}{4} \times 0 + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{3}{8} \\ p(\text{nothing}) &= \frac{1}{4} \times \frac{1}{2} + \frac{1}{2} \times 0 + \frac{1}{4} \times \frac{1}{2} = \frac{2}{8} \\ p(\text{penalty}) &= \frac{1}{4} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times 0 = \frac{3}{8} \end{aligned}$$

Answer to the “Monte Carlo” game

- Both strategies have an expected value of zero.
- However the “changing” strategy places more weight on extreme outcomes (winning a car, or losing the penalty) than the “remain” strategy $\implies$ higher variance!
- Therefore prefer the “remain” strategy because the individual is risk averse.

Bayesian solution to Monte Carlo problem

- Call E_2 the variable which indicates whether the host opens door 2; showing it to be empty.
- I_1 represents the initial choice of the contestant as door 1.
- We want to find $Pr(C_1|E_2, I_1)$ to begin with:

$$Pr(C_1|E_2, I_1) = \frac{Pr(E_2|C_1, I_1)Pr(C_1|I_1)}{Pr(E_2|I_1)} \tag{24}$$

Bayesian solution to Monte Carlo problem

In order to calculate the terms, it is easiest to enumerate all the possible outcomes: (2 times - one for each E) CPEE, CEPE, CEEP, PCEE, PECE, PEEC, EECP, ECEP, ECPE, EEPC, EPEC, EPCE.
We can hence calculate the previous expression:

$$\begin{aligned} Pr(C_1|E_2, I_1) &= \frac{\left(\frac{1}{2} \times \frac{2}{3}\right)\frac{1}{4}}{4 \times \frac{1}{2} \times \frac{1}{12} + 2 \times 1 \times \frac{1}{12}} \\ &= \frac{1}{4} \end{aligned}$$

Bayesian solution to Monte Carlo problem

Hence, the probability of winning a car if you stick with your initial choice is remains at $\frac{1}{4}$. Now we need to calculate $Pr(C_3|E_2, I_1)$ - the probability of a change to door 3 resulting in a car:

$$\begin{aligned} Pr(C_3|E_2, I_1) &= \frac{Pr(E_2|C_3, I_1)Pr(C_3|I_1)}{Pr(E_2|I_1)} \\ &= \frac{\left(\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1 + \frac{1}{3} \times 0\right)\frac{1}{4}}{\frac{1}{3}} \\ &= \frac{3}{8} \end{aligned}$$

Hence by switching, you are more likely to win the car, but the same is true for the penalty $Pr(C_3|E_2, I_1) = \frac{3}{8}$. Hence, if you want to maximise your chances of winning, then you should change. However, if you are risk averse, then you should stick.