Bayesian Inference: The Basics

Statistical Models

- Most statistical analysis consists of fitting a model to the data
- The model summarises the data
  - should show the main features of the data
- Model has a number of parameters
- Want to estimate the parameters
- Definitions:
  - Data: X Parameters: θ

What is the probability that my bus will be late in the morning?

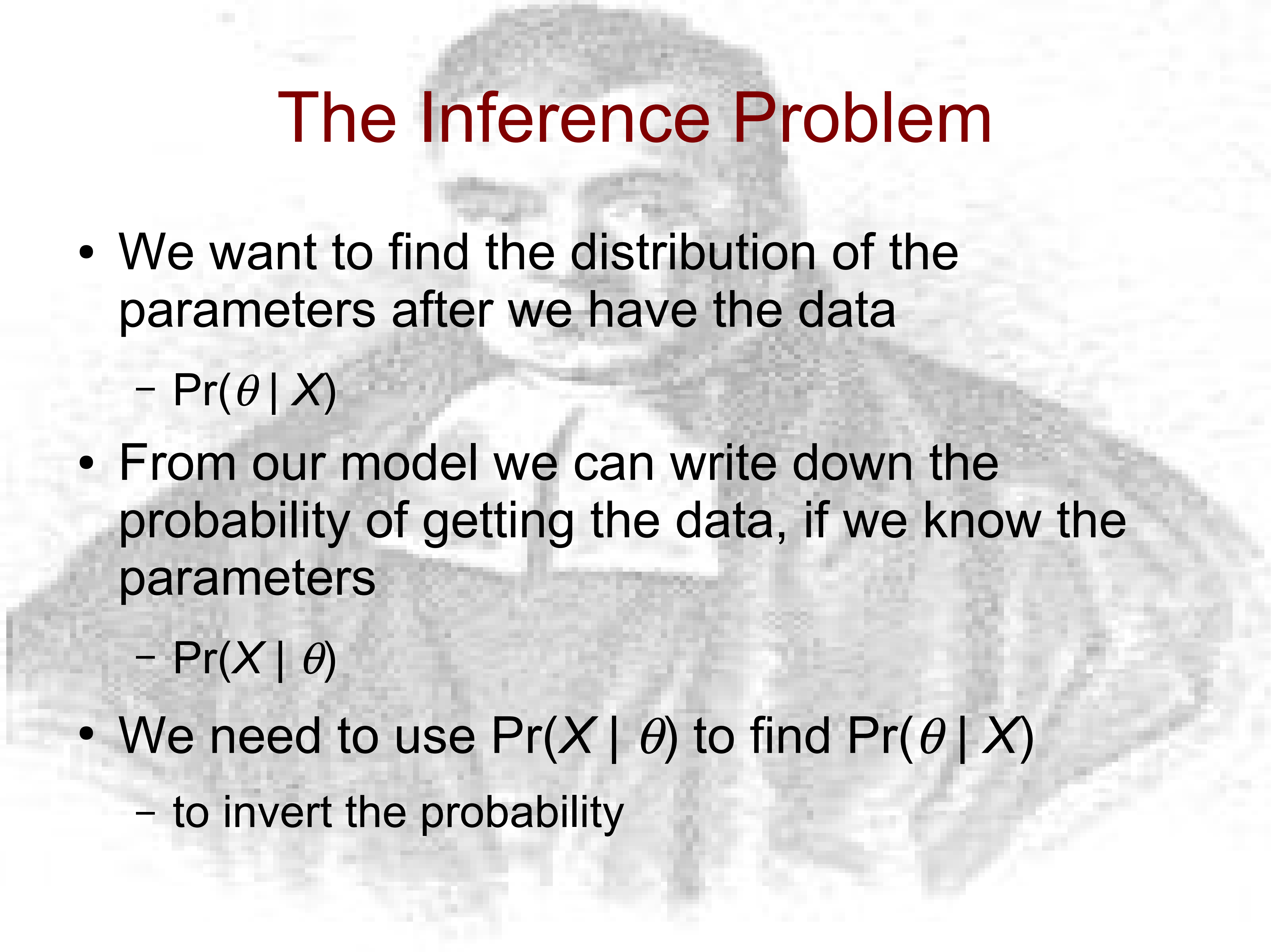
- Let p be the probability that the bus is late
- Model assumptions:
  - the probability is constant
  - events are independent
- Observe N mornings, bus is late n times, then:

$$Pr(N=n|p)=\frac{N!}{N!(N-n)!}p^{n}(1-p)^{N-n}$$

- Binomial Distribution

Bayesian Inference

- We can collect data about the processes that are influenced by the parameters
- Use this data and the models to infer the possible values of the parameters
- Summarise our beliefs about the possible values as probability distributions
- Adding data changes these distributions
  - inference is a learning process



The Inference Problem

- We want to find the distribution of the parameters after we have the data
  - $\Pr(\theta \mid X)$
- From our model we can write down the probability of getting the data, if we know the parameters
  - $\Pr(X \mid \theta)$
- We need to use $\Pr(X \mid \theta)$ to find $\Pr(\theta \mid X)$
  - to invert the probability

Enter Bayes' Theorem

- We know from probability theory that:

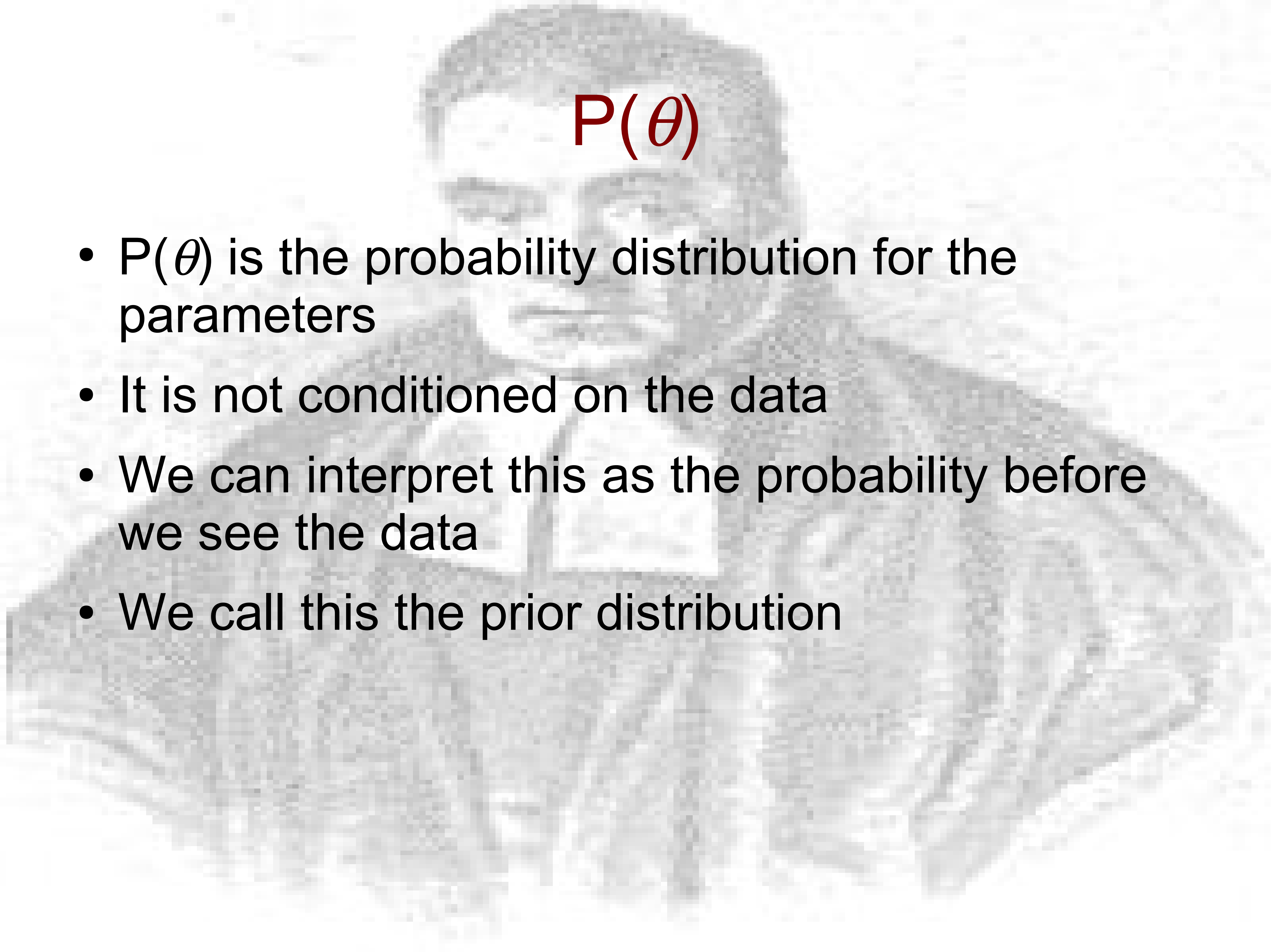
$$Pr(A|B) = \frac{Pr(B|A)\,Pr(A)}{Pr(B)}$$

- This is Bayes' theorem
- It allows us to invert the probabilities:

$$Pr(\theta|X) = \frac{Pr(X|\theta)\,Pr(\theta)}{Pr(X)}$$

- But what are P(θ) and P(X)?

P(θ)



- P(θ) is the probability distribution for the parameters
- It is not conditioned on the data
- We can interpret this as the probability before we see the data
- We call this the prior distribution

Prior Distributions

- Before we see any data, we have some idea about what values the parameters might take
- This may be "somewhere between minus or plus infinity"
- At other times, may be tighter
  - e.g. there are very few people 3m tall
- Our subjective uncertainty about the parameters before we see the data

Priors for late buses

- The parameter, *p*, is limited to be between 0 and 1
- We could assume total ignorance, and use a uniform distribution as a prior:
  - P(p)=1
- Or we could use another distribution, for example the Beta distribution:

$$P(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

The Beta Distribution

- The important bit:

$$P(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

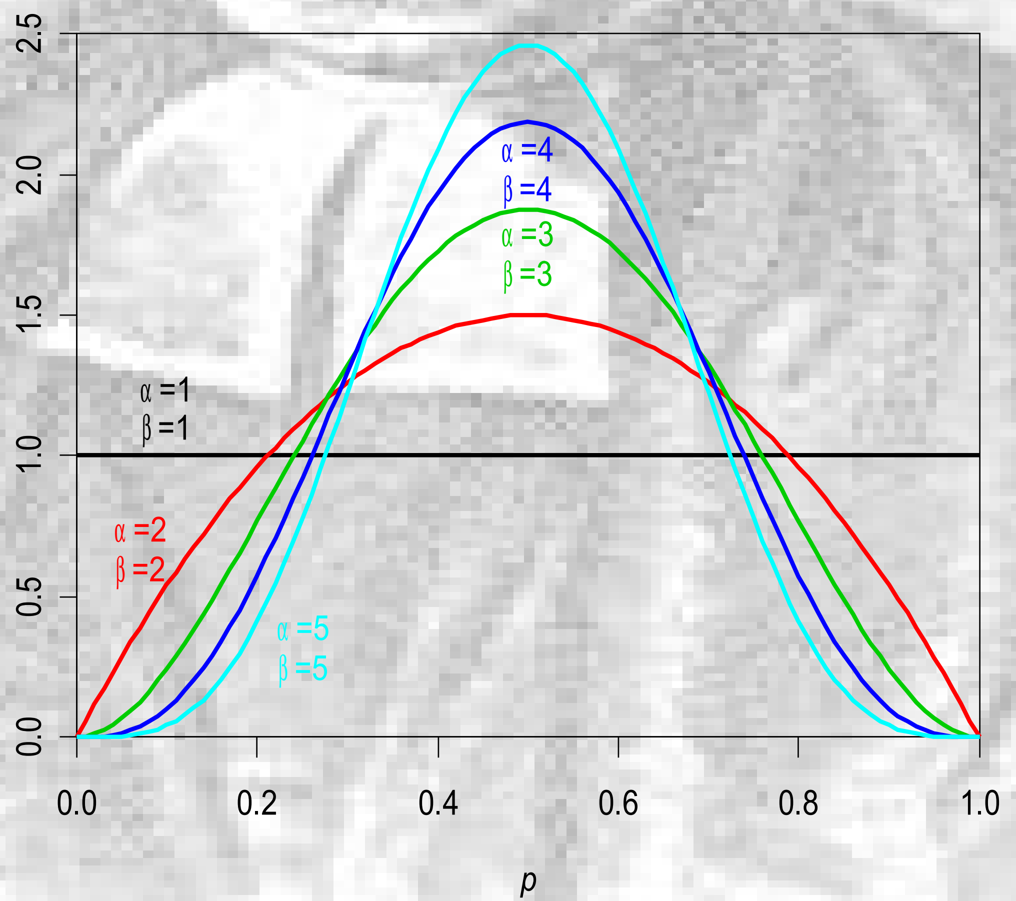
- Symmetric
  - replace p by 1-q, and swap a and ß

$$E(p)=\frac{\alpha}{\alpha+\beta} \qquad Var(p)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Write as Beta(α,β)

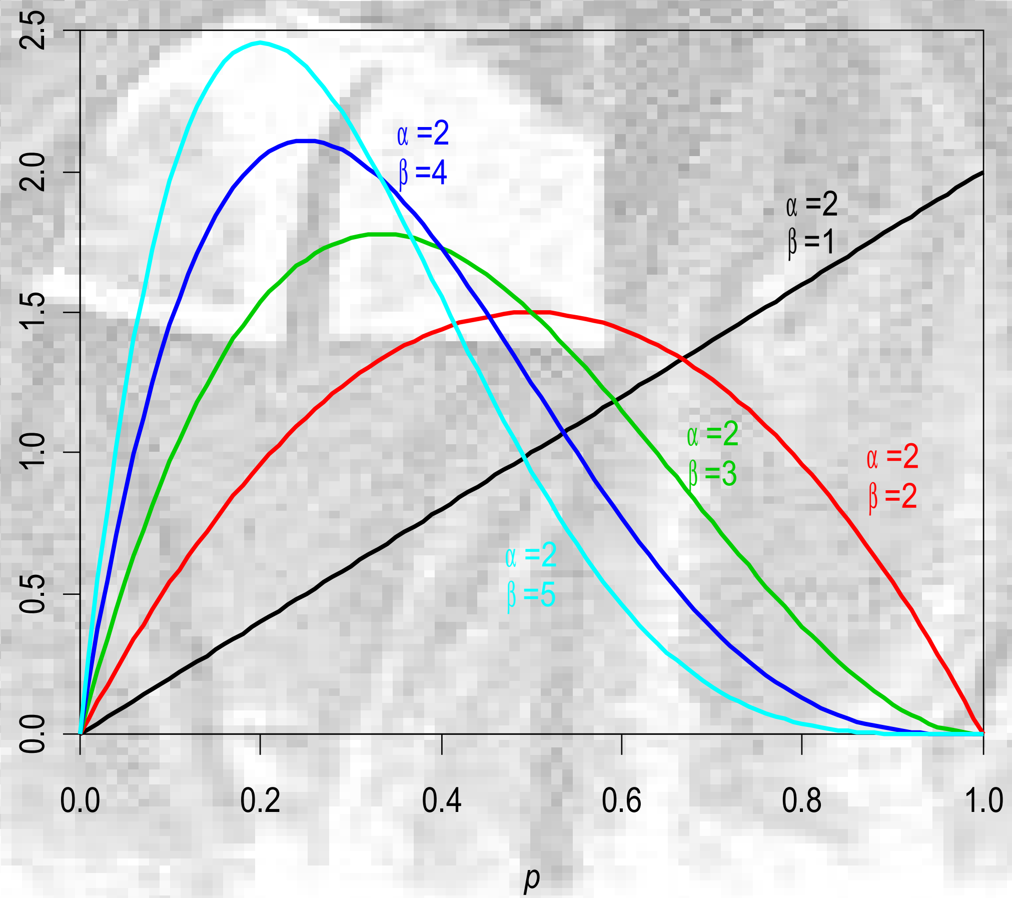
The Shape of the Beta, I

- When increasing α and β, variance gets lower



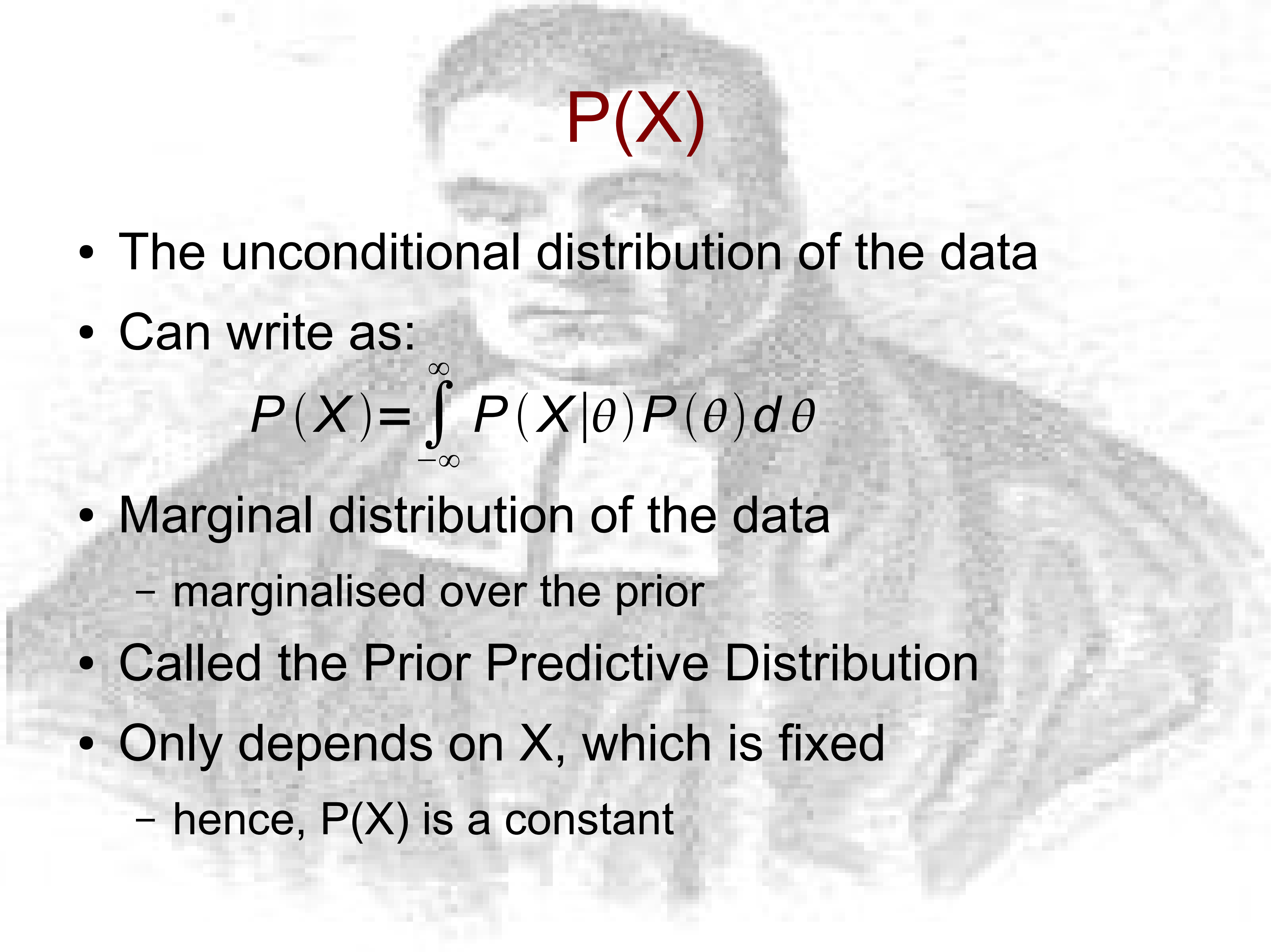
The Shape of the Beta, II

- When ß increases, distribution shifts down
  - similarly when a increases, distribution shifts up



Using the Beta as a Prior

- We can tune the distribution to reflect our prior knowledge
- If $\alpha = \beta = 1$, we have a uniform distribution
- As we increase α and β, the variance decreases
  - use this if we have more knowledge
- if we were talking about coin tossing, we might use $\alpha = \beta =$ something high



P(X)

- The unconditional distribution of the data
- Can write as:

$$P(X) = \int_{-\infty}^{\infty} P(X|\theta)P(\theta)\,d\theta$$

- Marginal distribution of the data
  - marginalised over the prior
- Called the Prior Predictive Distribution
- Only depends on X, which is fixed
  - hence, P(X) is a constant

Bayesian Inference

- As P(X) is a constant, all we need to estimate P(θ|X) are P(θ) and P(X|θ)
- Bayes' rule becomes:

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

- P(θ |X) is called the Posterior Distribution
- Product of the prior and the likelihood
- In practice, the constant of proportionality can usually be ignored

Late Buses

- Firstly, a Uniform Prior:

$$P(p)=1$$

- The likelihood – on N days, bus is late n times:

$$Pr(N=n|p)=\frac{N!}{N!(N-n)!}p^n(1-p)^{N-n}$$

- The Prior Predictive distribution:

$$Pr(N=n)=\int_0^1 \frac{N!}{N!(N-n)!}p^n(1-p)^{N-n}\,dp=1$$

Late Buses: The Posterior

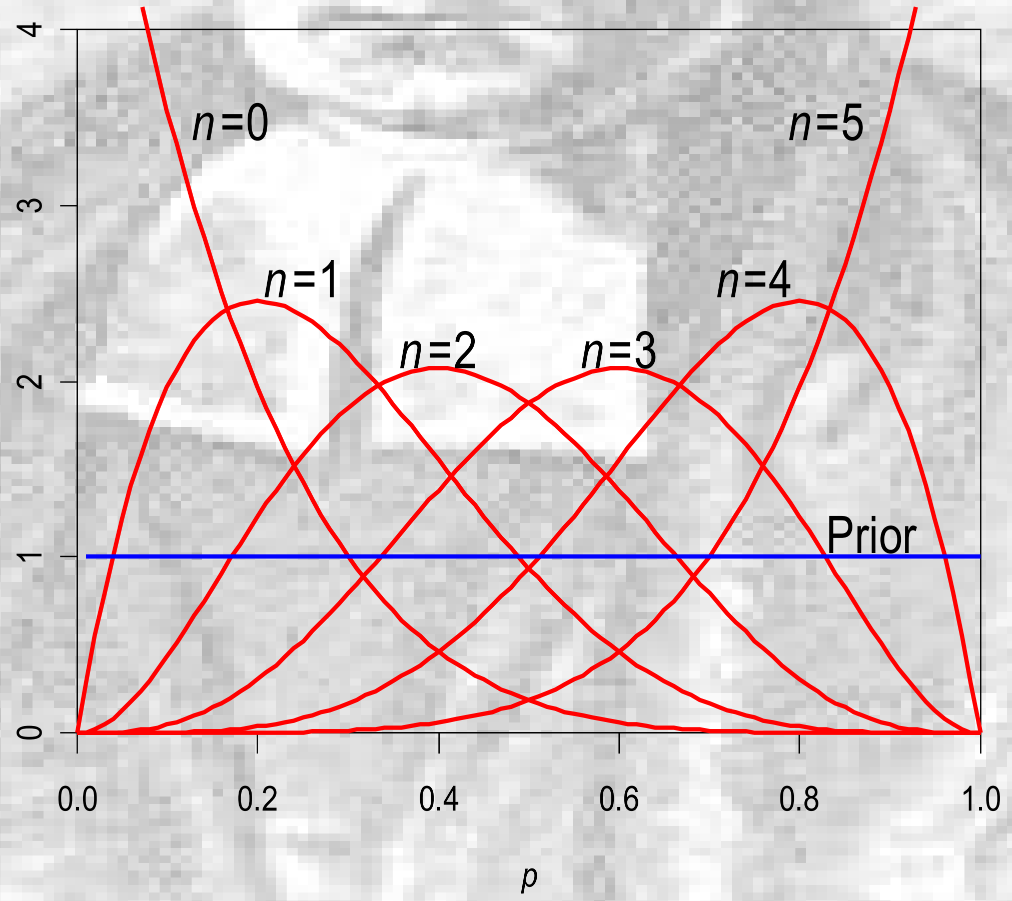
- The posterior:

$$Pr(p|n) = \frac{P(p)P(n|p)}{p(n)}$$
$$= 1\frac{N!}{N!(N-n)!}p^n(1-p)^{N-n}$$
$$\propto p^n(1-p)^{N-n}$$

- This is a Beta distribution!
  - P(p|n) = Beta(n+1, N-n+1)

The Posterior

- For N=5 (one week), possible posteriors:



Late Buses: Beta Prior

- The beta prior:

$$P(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

- The posterior:

$$\begin{aligned} P(p|n) &\propto p^{\alpha-1}(1-p)^{\beta-1} p^{n}(1-p)^{N-n} \\ &= p^{\alpha+n-1}(1-p)^{\beta+N-n-1} \end{aligned}$$

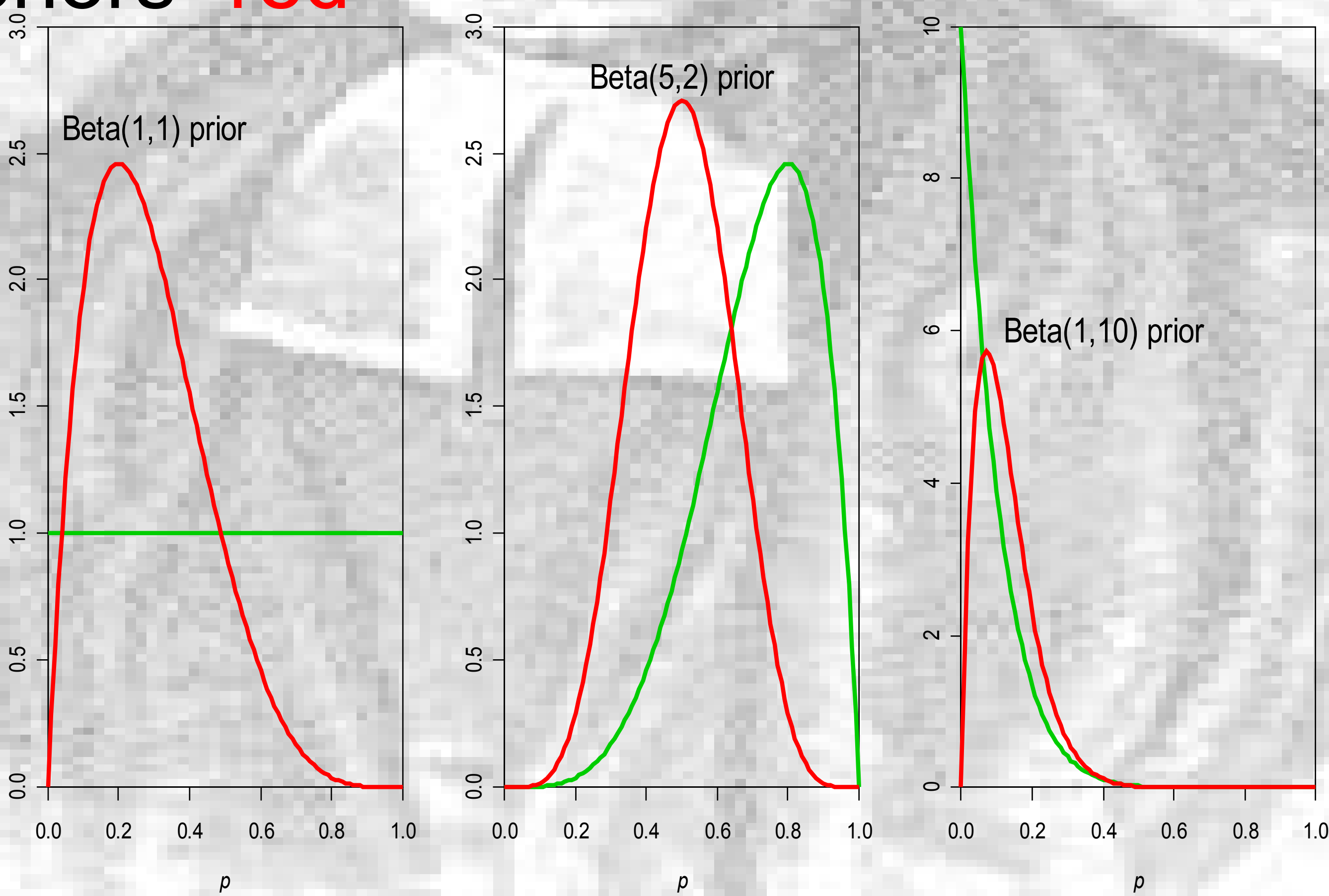
- Also a Beta distribution!
  - P(p|n) = Beta($n+\alpha$, N-$n+\beta$)

Beta Distribution Priors

- Not restricted to a uniform distribution
  - e.g. Assume we observe 1 late bus in a week
- Look at different priors:
  - uniform – Beta(1,1)
  - British prior – Beta(2,5)
  - Finnish prior – Beta(10,1)

Informative Priors

- Prior – green
- Posteriors -red



Adding More Data

- Observe data, X_1, get a posterior distribution:

$$Pr(\theta|X_1) \propto Pr(X_1|\theta)\,Pr(\theta)$$

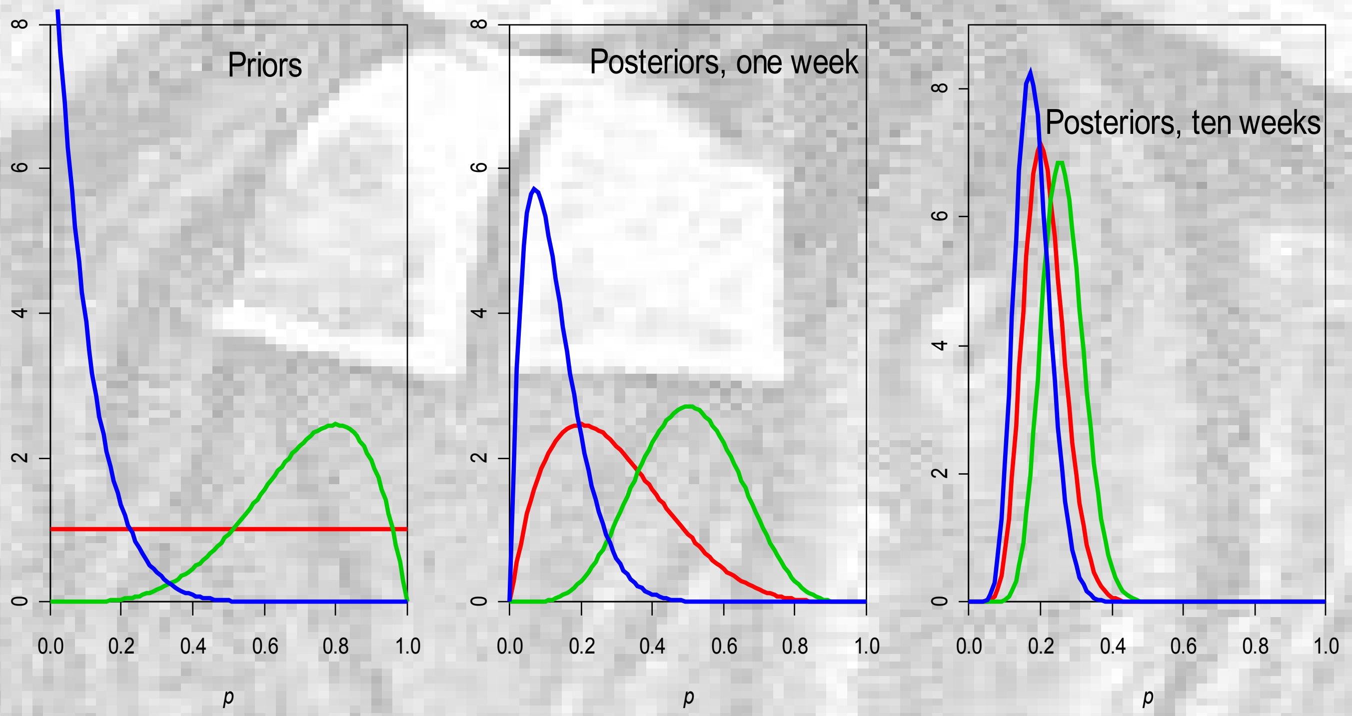
- What if we later observe more data, X_2?
- If this is independent of the first data set, then P(X_1 and X_2|θ)=P(X_1|θ).P(X_2|θ). Hence

$$\begin{aligned} Pr(\theta|X_1, X_2) &\propto Pr(\theta)\,Pr(X_1|\theta)\,Pr(X_2|\theta) \\ &= Pr(\theta|X_1)\,Pr(X_2|\theta) \end{aligned}$$

- i.e. we use the first posterior as the prior for the second posterior

Adding More Data

- After 10 weeks, observe 10 late buses
  - out of 50



- As evidence accumulates, our beliefs converge

Learning

- The Bayesian approach is often talked about as a learning process
- As we get more data, we add it to our store of information by multiplying it by our current posterior distribution.
- It has been argued that this can form the basis of a philosophy of science
  - Philosophers of science are not know for their success at explaining science

Terminology

- Uninformative prior
  - Uniform, as wide as possible
  - sometimes called flat priors
  - problem: often difficult to define
- Informative Prior
  - not uniform
  - assume we have some prior knowledge
- Conjugate Prior
  - prior and posterior have same distribution
  - often makes the maths easier

Summarising Posteriors

- Giving the full posterior distribution can be an awkward way of presenting the results of an analysis
  - especially if we have a lot of parameters.
- Often we are only interested in some parameters, or have particular questions to answer

Probabilities of events

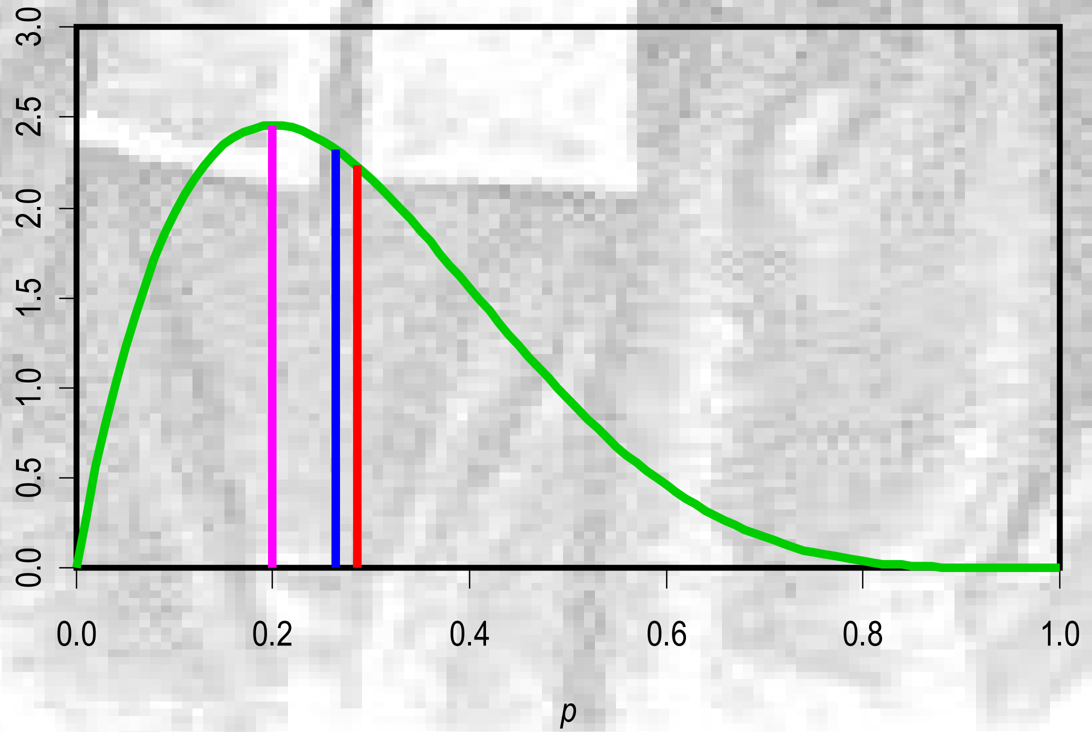
- The posterior distribution is a probability density from which we can calculate probabilities of events
  - e.g. P(θ>1)
- This is a straightforward interpretation of the probability.
- Contrast this with a frequentist p-value:
  - The probability of getting a statistic above a certain value, if the model and the parameter estimates are correct

Summaries

- Rather than give the full posterior for a parameter, we can give summary statistics
  - e.g. the posterior mode
    - equivalent to the ML estimate
    - the most likely value
  - Posterior mean or median
    - average values
    - consensus values
    - may not be very likely!

Late buses

- Uniform prior, observe 1 late bus in a week
  - Posterior: Beta(2, 5)
    - Mode Median Mean



Measures of Spread

- Posterior standard deviation
  - equivalent to standard error - the standard deviation of a statistic
- Bayesians don't make a distinction between parameters and statistics
- Late buses:

$$sd(p) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

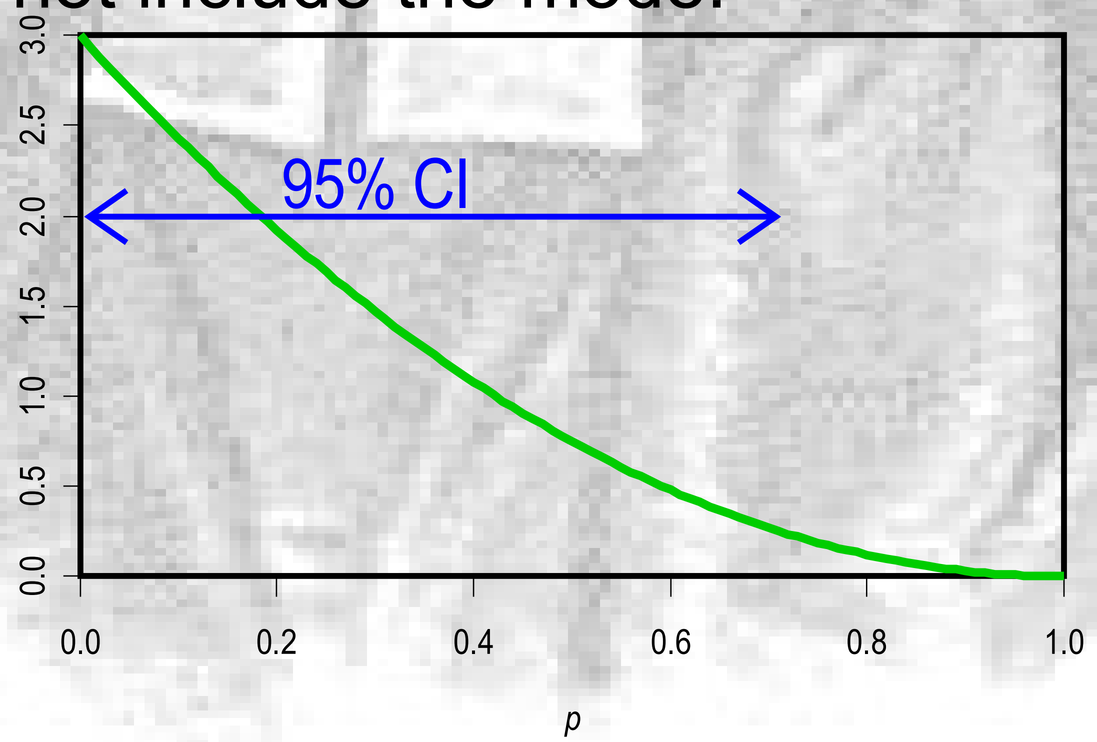
  - One bus late: sd(p) = 0.160
  - 10 out of 50: sd(p) = 0.056

95% Confidence Intervals

- An interval where there is a probability of 95% that the parameter is within the interval
  - Bayesian Confidence Interval
  - Credible Interval
- Frequentist CI defined as an interval that the statistic will be in 95% of the time if the ML estimate is correct.
- Problem: asymmetric distributions

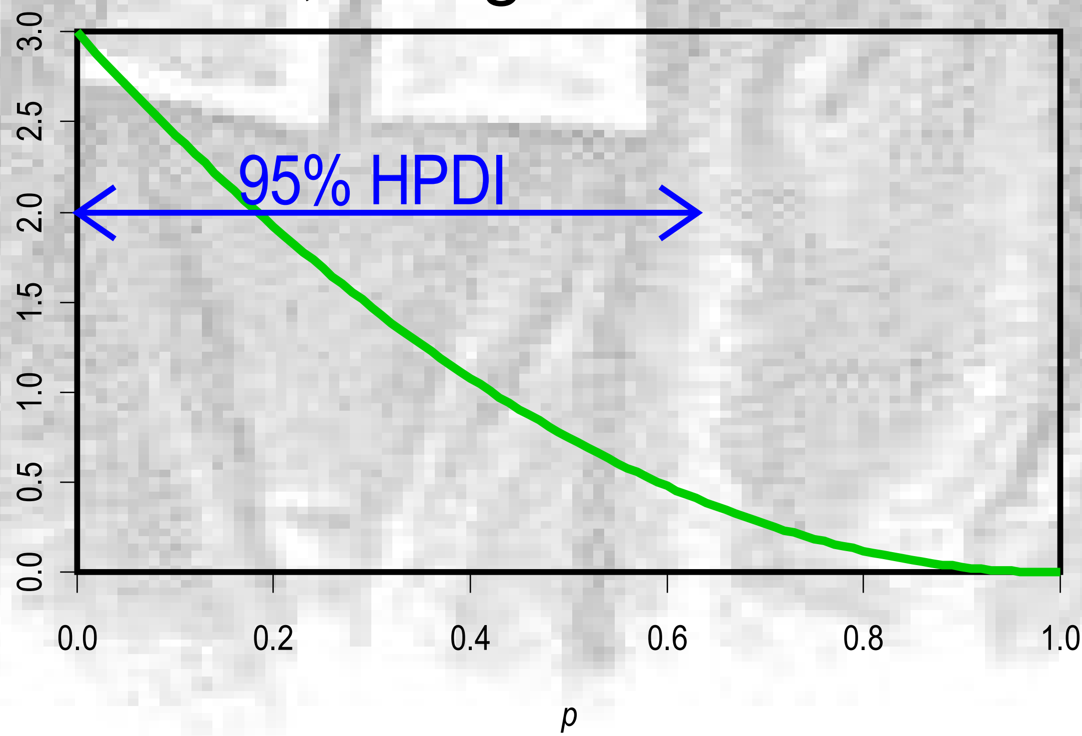
Traditional CIs

- Set at 2.5% and 97.5% limits
- An extreme example: Beta(1,3)
  - mode is 0
  - CI does not include the mode!



Solution

- Highest Posterior Densities
  - shortest 95% interval
  - all points inside interval have higher densities than those outside
  - For bimodal data, can get 2 intervals



Hypothesis Tests

- If we have 2 models, M_1 and M_2, how can we choose between them?
- From Bayes' Rule:
  - $\Pr(M_1|X) = \Pr(X|M_1).\Pr(M_1)$
  - $\Pr(M_2|X) = \Pr(X|M_2).\Pr(M_2)$
- Usually compare these with a Bayes' Factor:

$$B.F. = \frac{Pr(M_1|X)}{Pr(M_2|X)} = \frac{Pr(X|M_1)}{Pr(X|M_2)} \times \frac{Pr(M_1)}{Pr(M_2)}$$

- Change in odds of the models as we get data

Prediction

- One practical use of statistics
- A usual problem: using point estimates ignores the uncertainty in the estimates
- Our predictions are calculated as:

$$P(X_{new}|X) = \int P(X_{new}|\theta) . P(\theta|X)\, d\theta$$

- Posterior Predictive Distribution
- Include the uncertainty in the parameters
- Less precise than predictions from point estimates

How many times will my bus be late this week?

- 5 days. Last week, late once out of 5 times
- Uniform prior
- The posterior predictive distribution is difficult to calculate
- Instead, we use Monte Carlo simulation
- First, we simulate p^{new} from a Beta(2,5) distribution a lot of times
- Then for each p^{new}, we simulate n_{new} from a Binomial distribution with $p=p^{new}$

The Prediction

- Posterior Predictive Standard Deviation 1.4 times larger than for the prediction from the point estimate