



# **Bayesian Inference**

The Basics

# Statistical Analyses

- Our purpose: to fit statistical models to data
  - for whatever reason!
- The model will have a number of parameters
- We want to estimate the parameters
  - and then to check that the model fits

# The Model

- We assume that the data is produced by some mechanism
  - the “data generation mechanism”
- This mechanism involves some random component
  - here “random” is roughly the same as “unknown”
- We then build a mathematical model of the data generation mechanism
  - an approximation to the real mechanism

# The Model

- Definitions
  - $Y$ : Data
  - $X$ : Covariates
  - $\theta$ : Parameters of the model
- These all can be scalars, vectors or matrices
- The model is a way of producing data
- Tells us the probability of getting the data, with the parameters (and covariates)

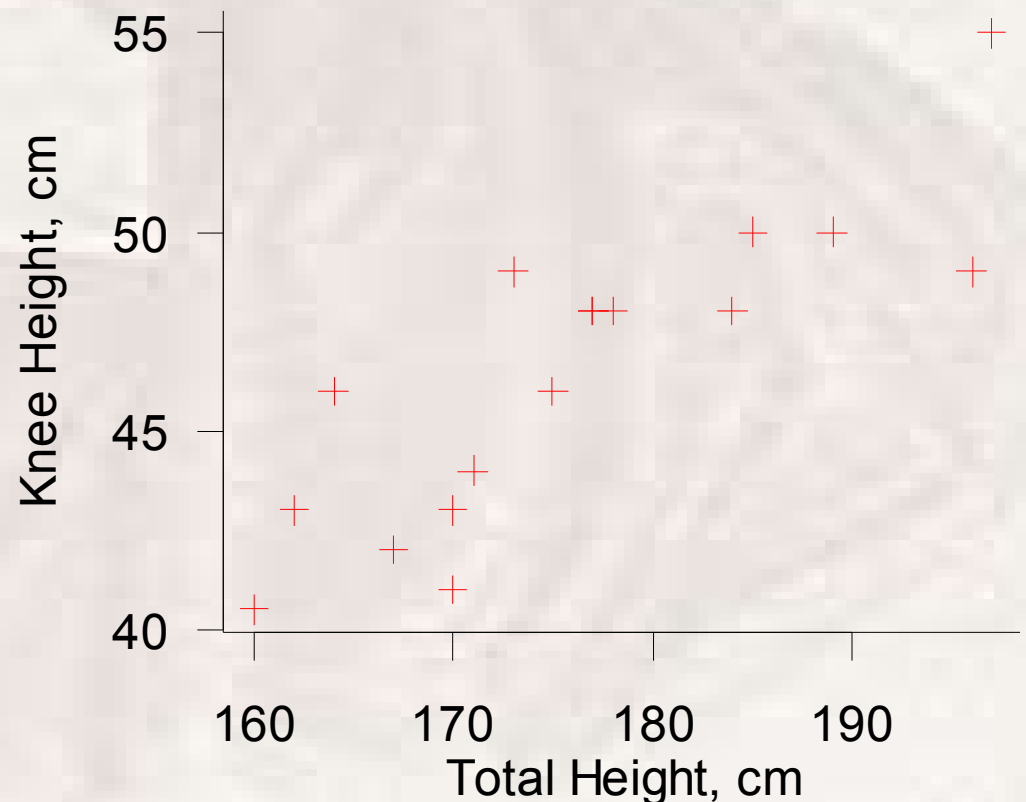
# An Example: Knees

- We want to look at the relationship between knee height ( $x$ ) and total height ( $y$ )
- Can use regression

- The model:

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta x_i$$



# Likelihood

- Models give the probability of obtaining the data, given some parameters:

$$P(X | \theta)$$

- This is called the *likelihood*
- We want to use this to learn about the parameters,  $\theta$

# Bayesian Probability

- Bayesians view probability as a statement of uncertainty
  - e.g. there is a 90% chance that this set of slides include a typo
- We can talk about uncertainty in data
  - e.g. the probability that someone with a knee height of 45cm is less than 170cm tall
- Or we can talk about uncertainty in a parameter
  - e.g. the probability that the slope is  $>1$

# **Whose Uncertainty**

## **Is It Anyway?**

- A key point is that Bayesian probabilities are subjective
- My uncertainty may be different to yours
  - e.g. I am convinced Kimi Raikkonen is an interesting person
- We can therefore talk about Bayesian probabilities as statements of belief
  - e.g. I am 90% sure that Kimi Raikkonen is an interesting person

# The Inference Problem

- We observe some data,  $X$ , and want to make inferences about the parameters from the data
  - i.e. find out about  $P(\theta | X)$
- We have a model, which gives us the likelihood
  - $P(X | \theta)$
- We need to use  $P(X | \theta)$  to find  $P(\theta | X)$ 
  - i.e. to invert the probability

# Enter Bayes' Theorem

- We know from probability theory that:

$$P(\theta | X) = \frac{P(\theta)P(X | \theta)}{P(X)}$$

- This is Bayes' Theorem
- Allows us to invert the probability
  - go from  $P(X | \theta)$  to  $P(\theta | X)$
- But what are  $P(\theta)$  and  $P(X)$ ?

# **$P(\theta)$**

- $P(\theta)$  is the probability distribution for the parameters  $\theta$
- 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

# **$P(X)$**

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

$$P(X) = \int P(X | \theta) P(\theta) d\theta$$

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

# Bayesian Inference

- As  $P(X)$  is a constant, all we need to estimate  $P(\theta | X)$  are  $P(\theta)$  and  $P(X | \theta)$

- Bayes' rule becomes:

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

- $P(\theta | X)$  is called the *posterior distribution*
- Product of the prior and the likelihood
- We can ignore the constant of proportionality

# Adding More Data

- Suppose we observe some data,  $X_1$ , and get a posterior distribution:

$$P(\theta | X_1) \propto P(\theta)P(X_1 | \theta)$$

- What if we later observe more data,  $X_2$ ?

- If this is independent of  $X_1$ , then P

$(X_1, X_2 | \theta) = P(X_1 | \theta) P(X_2 | \theta)$ , so that

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

- i.e. the first posterior is used as the prior to get the second posterior

# 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

# Factorisation

- Suppose we have several parameters
  - e.g.  $\theta_1$  and  $\theta_2$

- The posterior is  $P(\theta_1, \theta_2 | X)$

- It is often convenient to factor this as:

$$P(\theta_1, \theta_2 | X) = P(\theta_1 | \theta_2, X)P(\theta_2 | X)$$

- Which comes straight from the definition of conditional probability ( $P(A|B)=P(A,B)/P(B)$ )
- This can be extended:

# Marginalisation

- If we only want to look at some parameters, then we have to remove the others
- We can do this by taking the *marginal distribution*:

$$\begin{aligned}P(\theta_1 | X) &= \int P(\theta_1, \theta_2 | X) d\theta_2 \\ &= \int P(\theta_1 | \theta_2, X) P(\theta_2 | X) d\theta_2\end{aligned}$$

- The uncertainty in  $\theta_2$  is included into the marginal distribution of  $\theta_1$
- This is easy in practice

# 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.
- Therefore reporting marginal distributions is preferable
- The whole distribution is too much information to report
- Want to summarise these distributions

# 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!

# Probabilities of events

- The posterior distribution is a probability density from which we can calculate probabilities of events
  - e.g.  $P(\theta > 1)$
- This is a straightforward interpretation of the probability
- Especially useful if the parameters are indicators for different models

# Prediction

- Predicting from point estimates ignores the uncertainty in the estimates
- Bayesian predictions are calculated from

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

- *Posterior Predictive Distribution*
- Includes the uncertainties in the parameters

# Prior 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