

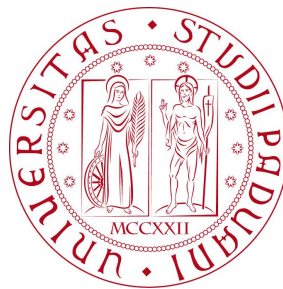
# Statistical Models and Inference - Part I

---

Alberto Garfagnini

Università di Padova

AA 2019/2020 - Stat Lect. 6



## Data Modeling

---

- we perform **experiments** and make **observations** to **learn about a phenomenon**
- to interpret data, we have to model them

### Inference

- make general statements about a phenomenon **through a model**, using **noisy and incomplete data**
  - must **describe** both the **Phenomenon** (i.e. Model) and the **Measurement Process**
- ▷ Key to Data Modeling: use data together with generative model (theory) and measurement model (experimental practice) to derive consistent probabilistic inferences

- given some data,  $D$ , we want to perform three actions:
- ▷ **parameter estimation**:  
for a specific Model  $M$ , with parameters  $\theta$ , infer the values of model parameters, i.e.  $P(\theta | D, M)$ , the **parameter posterior pdf**
- ▷ **model comparison**:  
given a set of models  $\{M_j\}$ , find out which one is best supported by data. This means finding  $P(M_j | D)$ , the **model posterior probability**
- ▷ **prediction**:  
given a model  $M$ , inferred from the data, **predict new data at some new location** (in the parameter space or time)

## Bayesian Model Comparison

---

- we start by looking at model comparison for the simple case of models with no parameters
- ▷ using our data  $D$ , we look for  $P(M | D)$
- since  $M \cdot \bar{M} = 0$  and  $M + \bar{M} = \Omega$ , we can write

$$\begin{aligned} P(D) &= P(DM) + P(D\bar{M}) \\ &= P(D | M) P(M) + P(D | \bar{M}) P(\bar{M}) \end{aligned}$$

- our quantity of interest,  $P(M | D)$ , is related to Bayes' theorem by

$$\begin{aligned} P(M | D) &= \frac{P(D | M) P(M)}{P(D)} = \frac{P(D | M) P(M)}{P(D | M) P(M) + P(D | \bar{M}) P(\bar{M})} \\ &= \frac{1}{1 + \frac{P(D | \bar{M}) P(\bar{M})}{P(D | M) P(M)}} = \frac{1}{1 + \frac{1}{R}} \end{aligned}$$

- with  $R = \frac{P(D | M) P(M)}{P(D | \bar{M}) P(\bar{M})}$  the **posterior odd ratio** of the models

# Bayesian Model Comparison

---

- it is easy to demonstrate that

$$\frac{P(M | D)}{P(\bar{M} | D)} = R = \frac{P(D | M) P(M)}{P(D | \bar{M}) P(\bar{M})}$$

- in order to determine  $P(M | D)$ , we need three quantities:

▷  $P(D | M)$  : the probability of measuring  $D$  when  $M$  is true

▷  $P(D | \bar{M})$  : the probability of measuring  $D$  when  $M$  is not true (i.e. false)

▷  $P(M)$  : the probability that  $M$  is true, independently of the data (and, of course,  $P(\bar{M}) = 1 - P(M) \Rightarrow P(M)$  tells us how probable the model is

- but, shouldn't we have information to tell us that  $M$  is more likely than  $\bar{M}$ , we could set

$$P(M) = P(\bar{M})$$

- and  $R$  becomes the Bayes factor

$$BF = \frac{P(D | M)}{P(D | \bar{M})}$$

- i.e. the ratio of the probability of the data under each model

# Bayesian Model Comparison

---

- should we have more models,  $\{M_j\}$ , with  $\sum P(M_j) = 1$ , the probability of data becomes

$$P(D) = \sum_j P(D | M_j) P(M_j)$$

- and the posterior probability of model # 1,  $M_1$ , becomes

$$P(M_1 | D) = \frac{P(D | M_1) P(M_1)}{P(D)}$$

- if we do not have a complete set of models, we cannot compute the posterior probabilities, but we can still compute the odds ratio or Bayes factor between any two models

$$BF = \frac{P(D | M_1)}{P(D | M_2)} \quad \text{and} \quad R = \frac{P(D | M_1) P(M_1)}{P(D | M_2) P(M_2)}$$

# Example

---

## Problem

- a test for a disease is 90% reliable
- the probability of testing positive, in absence of the disease, is 0.07
- we know that among people aged 40 to 50 with no symptoms 8 in 1000 have the disease

Q: if a person in his/her 40 tests positive, what is the probability that he/she has the disease ?

## Background information

- we build the following propositions:
  - $D$ : a person is tested positive
  - $M$ : a person has the disease
- and probabilities
  - $P(D | M) = 0.9$
  - $P(D | \bar{M}) = 0.07$
  - $P(M) = 0.008$

## Example - analytical solution

---

- we build

$$R = \frac{P(D | M) P(M)}{P(D | \bar{M}) P(\bar{M})} = \frac{9 \cdot 10^{-1} \times 8 \cdot 10^{-3}}{7 \cdot 10^{-2} \times (1 - 8 \cdot 10^{-3})} = 0.1035$$

- therefore

$$P(M | D) = \frac{1}{1 + 1/R} = 0.094$$

- even though a positive test result is quite probable (assuming the person has the disease), it is very unlikely that he/she has the disease
- what is decisive in the computation of  $P(M | D)$  is the ratio between

$$P(D | M) = P(D | M) P(M) = 7.2 \cdot 10^{-3}$$

(positive result, assuming the disease is present)

- and

$$P(D | \bar{M}) = P(D | \bar{M}) P(\bar{M}) = 7 \cdot 10^{-2}$$

(positive result, assuming the disease is absent)

# Example - R solution

```
post <- function(p.d.m, p.d.notm, p.m) {
  p.notm <- 1 - p.m
  odds.ratio <- (p.d.m * p.m) /
                (p.d.notm * p.notm)
  p.m.d <- 1/(1 + 1/odds.ratio)
}

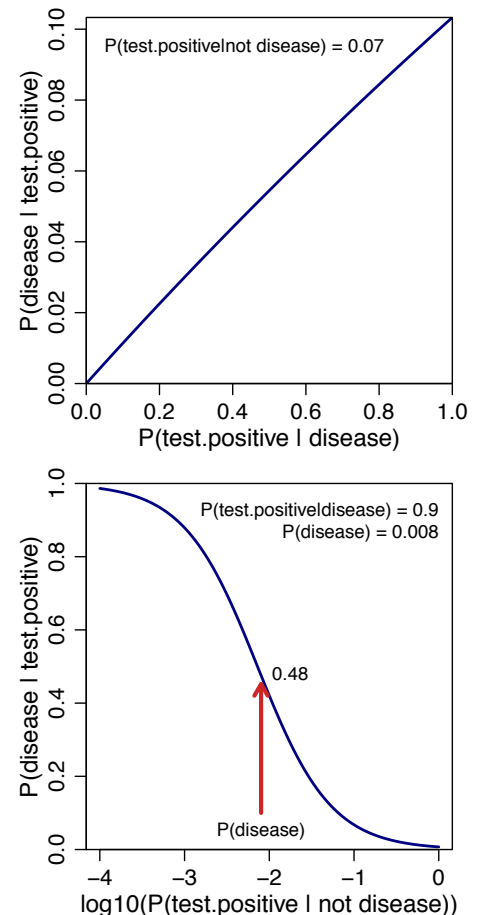
p.d.m <- seq(0, 1, 0.01) # True positive
p.d.notm <- 0.07         # False positive
p.m <- 0.008             # Disease Prior

p.m.d <- post(p.d.m, p.d.notm, p.m)
plot(p.d.m, p.m.d, type='l', lwd=2, col='navy')

p.d.m <- 0.9             # True positive
p.d.notm <- 10^seq(-4,0, 0.02) # False positive
p.m <- 0.008             # Disease Prior

p.m.d <- post(p.d.m, p.d.notm, p.m)
plot(log10(p.d.notm), p.m.d, type='l', col='navy')
```

- only once the false positive rate drops below the base rate ( $P(M)$ ) does the test starts to be useful



## Data Modeling with Parametric Models

- generative model** : theory predicting observable data from model parameters
  - the model just studied did not have any parameter: it was either true or false
- the simplest generative model is a straight line

$$f(x; a, b) = a + b \cdot x$$

- but our measurements will differ from the model due to noise



$$y = f(x; a, b) + \epsilon$$

- and the noise model - we call it the **measurement model** - has also parameters
  - given our set of data  $D = \{y_j\}$  at specified values  $\{x_j\}$ , we want to infer the values of the parameters for the generative model
  - in some cases we want to find the best set of parameters that predicts the data
  - but data are noisy  $\rightarrow$  there is no unique solution
- we look for the probability distributions of the parameters,  $P(\theta | D M)$ , also called **parameter posterior pdf**. Thanks to Bayes' theorem

$$P(\theta | D M) = \frac{P(D | \theta M) P(\theta | M)}{P(D | M)}$$

# The Likelihood

---

- $P(D \mid \theta M)$  is the Likelihood probability
  - it is a key function since it describes both the phenomenon and the data
  - it tells us the probability of getting the data we measured, given some value of the parameters
- $M$  specifies:
  - a generative model  the equation for the straight line  $f(x; a, b)$
  - a measurement model  how the measurement of  $y$  at a given  $x$  differs from  $f(x; a, b)$  due to noise
- the measurement model describes  $\epsilon$  in  $y = f(x; a, b) + \epsilon$ 
  - example: Gaussian distribution with variance  $\sigma^2$ . The Likelihood for any measurement is

$$P(y \mid \theta M) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(y - f(x; a, b))^2}{2\sigma^2}\right)$$

- telling us that the measurement has a Gaussian distribution about the true value
- $\theta = \theta(a, b; \sigma)$  is the union of the generative and measurements models

# The Prior

---

- $P(\theta \mid M)$  is the Prior probability
  - it encapsulates all the information we have, independent of the data
- it is called Prior because is the background information we have before obtaining the Data
- different people may have different information, or different opinion on what prior information is important
- this is not a weakness of inference
- it just reflects reality: we do not only use our immediate measurements to reach scientific conclusions

# The Posterior

---

- $P(\theta | D M)$  is the Posterior probability
  - it is the pdf over the model parameters, given data and background information
- from Bayes' theorem

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- the proportionality is through  $P(D | M)$ , a normalization factor which is independent of  $\theta$ . Therefore:

$$P(\theta | D M) = \frac{1}{Z} P(D | \theta M) P(\theta | M)$$

- with  $Z = P(D | M)$
  - from a conceptual point of view, inference is really that straightforward
  - Bayesian inference is the process of improving our knowledge of the model parameters by using the data
- ▷ we update the Prior using the Likelihood to obtain the Posterior

# The Evidence

---

- $P(D | M)$  is the Evidence
  - is the denominator of Bayes's equation and it gives the probability of observing the Data  $D$ , assuming the model  $M$  to be true, for any values of  $\theta$

$$P(D | M) = \int P(D | \theta M) P(\theta | M) d\theta$$

- evidence plays a key role in model comparison
- as a normalization constant, it is very important if we want to compute certain quantities from the posterior
- sometimes the integral can be calculated analytically, but for many real-world problems, we have to resort to numerical integration → Markov Chain Monte Carlo