smfd4.tex Day 4. 18.5.2012

## IV. BAYESIAN STATISTICS

# 1. CLASSICAL STATISTICS AND ITS LIMITATIONS.

Broadly speaking, statistics splits into two main streams:

(i) classical, or frequentist, and

(ii) Bayesian.

Much of classical statistics is devoted to the following general areas: Estimation of parameters (I), Hypothesis testing (II). Again, this is not exhaustive: the main remaining area is Non-parametric statistics (III).

Estimation of parameters itself splits, into

(ia). Point estimation [e.g., maximum-likelihood estimates],

(ib). Interval estimation [e.g., confidence intervals].

Both these are open to interpretational objections. A point estimate is a single number, which will almost certainly be wrong [i.e., will differ from the value of the parameter it estimates]. How wrong? And what should we do about this?

A confidence interval is more informative, because it includes an error estimate. For instance, its mid-point can be regarded as a point estimate, and half its length as an error estimate - leading to conclusions of the form

$$\theta = 3.76 \pm 0.003$$
 (\*)

- with *confidence* 95% [or 99%, or whatever]. What does this mean? It is not a probability statement:

either  $\theta$  lies between 3.73 and 3.79 [when (\*) is true, so holds with probability 100 %]

or it doesn't [when (\*) is false, so holds with probability 0%].

Problem: We don't know which!

Interpretation. If a large number of statisticians independently replicated the analysis leading to (\*), then about 95 % of them would succeed in producing confidence intervals covering the unknown parameter  $\theta$ . But

(a) We wouldn't know which 95 %,

(b) This is of doubtful relevance anyway. The large number of independent replications will usually never take place in practice. So confidence statements like (\*) lack, in practice, a direct interpretation. [They are 'what

happens to probability statements in classical statistics when we put the numbers in'.]

A further problem is that small changes in our data can lead to abrupt discontinuities in our conclusions. In borderline situations,  $\theta$  'just within' the confidence interval and 'just outside' represent diametrically opposite outcomes, but the data may be very close. Small changes in input *should* only lead to small changes in output, rather than abrupt changes.

Hypothesis testing is open to similar objections. It is usual to have a null hypothesis,  $H_0$ , representing our present theory (the 'default option'), and an alternative hypothesis,  $H_1$ , representing some proposed alternative theory. At the end of the investigation, we have to choose between two alternatives. We may be wrong: we may

reject  $H_0$  when it is true, and choose  $H_1$  [Type I error, probability  $\alpha$ , the significance level], or

reject  $H_1$  when it is true, and choose  $H_0$  [Type II error, probability  $\beta$ ].

We then have a trade-off between  $\alpha$  and  $\beta$ . It is not always clear how to do this sensibly, still less optimally [it is customary to choose  $\alpha = 0.05$  or 0.01, and then try to minimise  $\beta$ , but this is merely conventional]. Again, problems present themselves:

(i) We won't know whether our choice between  $H_0$  and  $H_1$  was correct,

(ii) Small changes in the data can lead to abrupt changes between choosing  $H_0$  and choosing  $H_1$ .

Thus both the main branches of classical parametric statistics lead to abruptly discontinuous conclusions and present interpretational difficulties. One justification for Bayesian statistics is that it avoids these. There are many others: we shall argue for Bayesian statistics below on its merits.

### 2. PRIOR KNOWLEDGE AND HOW TO UPDATE IT.

The difficulties identified above arise because in classical statistics we rely entirely on the data, that is, on the sample we obtained. The mathematics involved in classical statistics amounts to comparing the sample we actually obtained with the large (usually, infinite) class of hypothetical samples we might have obtained but didn't. These include the samples that we would obtain if we repeated our sampling independently – or that other statisticians would obtain if they independently replicated our work. This is where the term 'frequentist' for classical statistics originates: e.g., in 95 % confidence intervals, independently replicated confidence intervals would cover the parameter  $\theta$  with frequency 0.95.

The other aspect of classical statistics crucial for our purposes is that it ignores everything before sampling. This is often unreasonable. For instance, we may know a good deal about the situation under study, based on prior experience. Such situations are typical in, e.g., industrial quality control: suppose we are employed by a rope manufacturer, and are testing the breaking strain of ropes in a current batch. We may have to hand large amounts of data obtained from tests on previous batches from the same production line. In hypothesis testing, such prior knowledge is tacitly assumed, because we need it to be able to formulate  $H_0$  and  $H_1$  sensibly. But we may not be willing to enter the 'accept or reject' framework of hypothesis testing [which some statisticians believe is inappropriate and damaging]: how then can we use prior knowledge? In the estimation framework also, we may know a lot about  $\theta$  before sampling [as in the rope example above]: indeed, if we do not have some prior knowledge of the situation to be studied, we would in practice not have enough prior interest in it to be willing to invest the time, trouble and money to study it statistically.

Bayesian statistics addresses these aspects by providing a framework in which

1. The statistician knows something before sampling: he has some *prior* knowledge.

2. He then draws a sample, and analyses the *data* to extract some relevant information.

3. He then updates his prior information with his data (or sample) information, to obtain posterior information

(prior: before (sampling); posterior: after (sampling)).

This verbal description of the Bayesian approach is attractive, because it resembles how we learn. Life involves (indeed, largely consists of) a constant, ongoing process of acquiring new information and using it to update our previous ('prior') information/beliefs/attitudes/policies.

To implement the Bayesian approach, we need some mathematics. The formulae below derive from the work of the English clergyman

Thomas BAYES (1702-1761): An essay towards solving a problem in the doctrine of chances (1763, posth.).

Recall that if A, B are events of positive probability,

$$P(A) > 0, \qquad P(B) > 0,$$

the conditional probability of A given (or knowing) B is

$$P(A|B) := P(A \cap B)/P(B).$$

Symmetrically,

$$P(B|A) := P(B \cap A)/P(A) = P(A \cap B)/P(A).$$

Combining,

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A),$$

or

P(B|A) = P(A|B)P(B)/P(A) (BAYES' FORMULA, or BAYES' THEOREM).

Interpretation.

1. Think of A as a 'cause', B as an 'effect'. We naturally first think of P(effect B | cause A). We can use Bayes' formula to get from this to P(cause A | effect B) (think of B as an effect we can see, A as an effect we can't see).

2. Suppose we are interested in event B. We begin with an initial, prior probability P(B) for its occurrence. This represents how probable we initially consider B to be [this depends on us: we will have to estimate P(B)!]. Suppose we then observe that event A occurs. This gives us new information, which affects how probable we should now consider B to be, after observing A [or, to use the technical term, a posteriori]. Bayes' theorem tells us how to do this updating: we multiply by the ratio P(A|B)/P(A):

$$P(B|A) = P(B).P(A|B)/P(A) :$$

posterior probability of B = prior probability of  $B \times$  updating ratio. We first observe some extreme cases.

Independence. If A, B are independent,  $P(A \cap B) = P(A).P(B)$ , so

$$P(B|A) = P(A \cap B) / P(A) = P(A) \cdot P(B) / P(A) = P(B),$$

and similarly P(A|B) = P(A): updating ratio = 1, posterior probability = prior probability – conditioning on something independent has no effect. *Inclusion*.

1.  $A \subset B$ : here,  $P(A \cap B) = P(A)$ ,  $P(A|B) = P(A \cap B)/P(B) = P(A)/P(B)$ ;

updating ratio P(A|B)/P(A) = 1/P(B), posterior probability = 1. 2.  $B \subset A$ : here,  $P(A \cap B) = P(B)$ ,  $P(A|B) = P(A \cap B)/P(B) = P(B)/P(B) = 1$ ;

updating ratio P(A|B)/P(A) = 1/P(A), posterior probability = P(B)/P(A). Partitions. The event B partitions the sample space  $\Omega$  (the space of all possible outcomes) into two disjoint events  $B, B^c$  whose union is  $\Omega$ . Then A is the disjoint union of  $A \cap B$  and  $A \cap B^c$ , so

$$P(A) = P(A \cap B) + P(A \cap B^{c}) = P(A|B)P(B) + P(A|B^{c})P(B^{c}),$$

by definition of conditional probability. Similarly, if  $B_1, B_2, \dots, B_n$  form a partition (are disjoint events with union  $\Omega$ ), A is the disjoint union of events  $A \cap B_1, \dots, A \cap B_n$ . So by the additivity property of probability,

$$P(A) = \sum_{r=1}^{n} P(A \cap B_r) = \sum_{r=1}^{n} P(A|B_r) P(B_r) \quad (\text{FORMULA OF TOTAL PROBABILITY}).$$

using the definition of conditional probability again.

Such expressions are often used for the denominator in Bayes' formula:

$$P(B_r|A) = P(B_r)P(A|B_r)/P(A) = P(B_r)P(A|B_r)/\Sigma_k P(B_k)P(A|B_k)$$

### 3. PRIOR AND POSTERIOR DENSITIES.

Suppose now we are studying a parameter  $\theta$ . Suppose we have data x [x may be a single number, i.e. a scalar, or a vector  $\mathbf{x} = (x_1, \dots, x_n)$  of numbers. Following O'Hagan [O'H] (for references, see Day 0), we shall simply write x in both scalar and vector cases.] Recall that x is an observed value of a random variable, X say. In the *density case*, this random variable has a (probability) *density* (function), f(x) say, a non-negative function that integrates to 1:

$$f(x) \ge 0, \qquad \int f(x)dx = 1$$

(here and below, integrals with limits unspecified are over everything). Interpretation.  $P(X \in A) = \int_A f(x) dx$  for all subsets A of the real line **R** [actually, we need to restrict to suitable – 'measurable' – sets A, but it suffices for our purposes to consider intervals or half-lines. For instance, if  $A = (-\infty, x]$ ,

$$F(x) := P(X \in (-\infty, x]) = P(X \le x) = \int_{-\infty}^{x} f(y) dy \quad \forall x \in \mathbf{R};$$

as x varies, F(x) gives the (probability) distribution (function) of X.] In brief: the density f(x) expresses, or describes, the *uncertainty* in the data x.

The distinctive feature of Bayesian statistics is that it treats *parameters*  $\theta$  in the same way as *data* x. Our initial (prior) uncertainty about  $\theta$  should also be described by a density  $f(\theta)$ :

$$f(\theta) \ge 0, \qquad \int_{-\infty}^{\infty} f(\theta) d\theta = 1,$$
$$P(\theta \in A) = \int_{A} f(\theta) d\theta \qquad \forall A \subset \mathbf{R},$$

where the probability on the left is a *prior probability*. The analogue for densities of Bayes' formula

$$P(B|A) = P(B)P(A|B)/P(A)$$

now becomes

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x).$$
(\*)

The density on the left is the *posterior density* of  $\theta$  given the data x; it describes our uncertainty about  $\theta$  knowing x.

Now densities integrate to 1:

$$\int f(\theta|x)d\theta = 1,$$

so  $\int [f(\theta)f(x|\theta)/f(x)]d\theta = 1$ :

$$\int f(\theta)f(x|\theta)d\theta = f(x).$$

Combining,

$$f(\theta|x) = f(\theta)f(x|\theta) / \int f(\theta)f(x|\theta)d\theta$$

In the discrete case,  $\theta$  and/or x may take discrete values  $\theta_1, \theta_2, \dots, x_1, x_2, \dots$ only, with probabilities  $f(\theta_1), f(\theta_2), \dots, f(x_1), f(x_2), \dots$  The above formulae still apply, but with *integrals replaced by sums*:

$$P(X \in A) = \sum_{x \in A} f(x), \qquad P(\theta \in B) = \sum_{\theta \in B} f(\theta),$$
$$f(x) = \sum_{\theta} f(\theta f(x|\theta),$$

$$f(\theta|x) = f(\theta)f(x|\theta) / \Sigma_{\theta}f(\theta)f(x|\theta).$$

In the formula  $f(\theta|x) = f(\theta)f(x|\theta)/f(x)$ , it is  $\theta$ , the parameter under study, which is the main focus of interest. Consequently, the denominator f(x) – whose role is simply to ensure that the posterior density  $f(\theta|x)$  integrates to 1 (i.e., really is a density) – can be omitted (or understood from context). This replaces the *equation* above by a *proportionality statement*:

$$f(\theta|x) \propto f(\theta) f(x|\theta)$$

(here  $\propto$ , read as 'is proportional to', relates to the variability in  $\theta$ , which is where the action is). Now  $f(x|\theta)$  can be viewed in two ways:

(i) for fixed  $\theta$  as a function of x. It is then the density of x when  $\theta$  is the true parameter value,

(ii) for fixed/known/given data values x as a function of  $\theta$ . It is then called the *likelihood* of  $\theta$ . Likelihood was systematically studied in the 1920s by the great English statistician and geneticist R. A. (Sir Ronald) FISHER (1890-1962); it is crucially important in classical statistics also (e.g., in maximum likelihood estimation (Ch. I), or likelihood ratio tests (Ch. II)).

The formula above now reads, in words:

posterior density is proportional to prior density times likelihood. This is the essence of Bayesian statistics. It shows how Bayes' theorem (of which this formula is a version) may be used to update the prior information on  $\theta$  before sampling by using the information in the data x – which is contained in the likelihood factor  $f(x|\theta)$  by which one multiplies – to give the posterior information on  $\theta$  after sampling. Thus posterior information combines two sources: prior information and data/sample/likelihood information.

#### 4. EXAMPLES.

Example 1. Bernoulli trials with Beta prior ([O'H], Ex. 1.4, p.5). Here  $\theta$  represents the probability of a head on tossing a biased coin. On the basis of prior information,  $\theta$  is assumed to have a prior density proportional to  $\theta^{p-1}(1-\theta)^{q-1}$  ( $0 \le \theta \le 1$ ) for p, q > 0:

$$f(\theta) \propto \theta^{p-1} (1-\theta)^{q-1} \qquad (0 \le \theta \le 1).$$

Writing

$$B(p,q) := \int_0^1 \theta^{p-1} (1-\theta)^{q-1} d\theta$$

(the *Beta function*),

$$f(\theta) = \theta^{p-1} (1-\theta)^{q-1} / B(p,q).$$

We quote the *Eulerian integral* for the Beta function: for

$$\Gamma(p) := \int_0^\infty e^{-x} x^{p-1} dx \quad (p > 0), \quad B(p,q) = \Gamma(p) \Gamma(q) / \Gamma(p+q) \quad (p,q > 0).$$

Note that, as p, q vary, the shape of  $f(\theta)$  varies – e.g, the graph is u-shaped if 0 < p, q < 1, n-shaped if p, q > 1. Here p, q are called *hyperparameters* - they are parameters describing the parameter  $\theta$ .

Suppose now we toss the biased coin n times (independently), observing x heads. Then x is our data. It has a discrete distribution, the binomial  $B(n, \theta)$ , described by

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad (x=0,1,\cdots,n).$$

We apply Bayes' theorem to update our prior information on  $\theta$  – our prior values of p, q – by our data x. Now

$$f(x) = \int f(\theta) f(x|\theta) d\theta = \int \frac{\theta^{p-1} (1-\theta)^{q-1}}{B(p,q)} \cdot \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta$$
$$\binom{n}{x} \cdot \frac{1}{B(p,q)} \cdot \int_0^1 \theta^{p+x-1} (1-\theta)^{q+n-x-1} d\theta = \binom{n}{x} \cdot \frac{B(p+x,q+n-x)}{B(p,q)}$$

So Bayes' theorem gives

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x) = \binom{n}{x} \cdot \frac{1}{B(p,q)} \cdot \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p,q)} \cdot \frac{B(p+x,q+n-x)}{B(p,q)} \cdot \frac{B($$

or

=

$$f(\theta|x) = \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p+x,q+n-x)}.$$

The posterior density of  $\theta$  is thus another Beta density, B(p+x, q+n-x). Summarising:

• prior B(p,q) is updated by data x heads in n tosses to posterior B(p + x, q + n - x).

Graphs. To graph the three functions of  $\theta$  – prior, likelihood and posterior –

first find their maxima.

*Likelihood:*  $f(x|\theta)$  has a maximum where  $\log f(x|\theta)$  has a maximum, i.e. where

 $x \log \theta + (n - x) \log(1 - \theta)$  has a maximum, i.e. where

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$$
:  $x - x\theta = n\theta - x\theta$ :  $\theta = x/n$ .

*Prior:* similarly,  $f(\theta)$  has a maximum where log  $f(\theta)$  does, i.e. where

$$\frac{p-1}{\theta} - \frac{q-1}{1-\theta} = 0: \quad p - p\theta - 1 + \theta = q\theta - \theta: \quad \theta = (p-1)/(p+q-2).$$

Posterior: replacing p, q by  $p + x, q + n - x, f(\theta|x)$  has a maximum where

$$\theta = (p + x - 1)/(p + q + n - 2).$$

Now

$$\frac{p+x-1}{p+q+n-2} = \lambda(\frac{x}{n}) + (1-\lambda)(\frac{p-1}{p+q-2}), \quad \text{where} \quad \lambda = \frac{n}{n+p+q-2}:$$

the posterior maximum is a 'weighted average' of the prior and likelihood maxima, with weights  $1 - \lambda = (p + q - 2)/(n + p + q - 2), \lambda = n/(n + p + q - 2)$  as above. When p > 1, q > 1, p + q - 2 > 0 and both weights are positive: this is then a genuine weighted average, a *convex combination* or *mixture*. The posterior combines, or synthesises, the prior and the likelihood: it compromises between them by giving something intermediate.

Note. 1. When p, q > 1, the relative weight n on the likelihood maximum is the sample size, that is, of how much data information we have; the relative weight p+q-2 on the prior maximum is a measure of how concentrated about its maximum the prior density is – that is, of how much prior information we have. Thus both weights have clear interpretations, and the weighting has a clear interpretation as a compromise between them.

2. For an estimator  $\hat{\theta}$  of a parameter  $\theta$ , recall that the variance  $var \hat{\theta}$  measures the degree of uncertainty/spread/scatter of the values of  $\hat{\theta}$  about the mean  $E\hat{\theta} [= \theta \text{ if } \hat{\theta} \text{ is unbiased}]$ . Desirable properties of estimators include: (i) bias  $E\hat{\theta} - \theta$  zero or small,

(ii) variance  $var \hat{\theta}$  minimum or small - that is, its reciprocal  $1/var \hat{\theta}$  maximum or large.

It is often preferable to work with the reciprocal of the variance,  $1/var \hat{\theta}$ ,

rather than the variance itself.

**Definition.** For an estimator  $\hat{\theta}$  of a parameter  $\theta$ , the **precision** of  $\hat{\theta}$  is  $prec \ \hat{\theta} := 1/var \ \hat{\theta}$ .

Example 2. Normal family with normal prior ([O'H], Ex. 1.5 p.7). Suppose x is the sample mean of a sample of n independent readings from a normal distribution  $N(\theta, \sigma^2)$ , with  $\sigma$  known and  $\theta$  the parameter of interest. So x is  $N(\theta, \sigma^2/n)$ :

$$f(x|\theta) = \frac{1}{\sqrt{2\pi} . \sigma/\sqrt{n}} \exp\{-\frac{1}{2}(x-\theta)^2 / \frac{\sigma^2}{n}\}.$$

Suppose that on the basis of past experience [prior knowledge] the prior distribution of  $\theta$  is taken to be  $N(\mu, \tau^2)$ :

$$f(\theta) = \frac{1}{\sqrt{2}\pi\tau} \exp\{-\frac{1}{2}(\theta - \mu)^2/\tau^2\}.$$

Now  $f(x) = \int f(\theta) f(x|\theta) d\theta$ :

$$f(\theta)f(x|\theta) = \frac{1}{2\pi \tau \sigma / \sqrt{n}} \exp\{-\frac{1}{2} \left[\frac{(\theta - \mu)^2}{\tau^2} + \frac{(x - \theta)^2}{\sigma^2 / n}\right]\}.$$

The RHS has the functional form of a bivariate normal distribution. So to evaluate the  $\theta$ -integration, we need to *complete the square* (cf. solving quadratic equations!). First,

$$(x-\theta)^2 = [(x-\mu) - (\theta-\mu)]^2 = (x-\mu)^2 - 2(x-\mu)(\theta-\mu) + (\theta-\mu)^2.$$

We write for convenience

$$c := \frac{1}{\tau^2} + \frac{1}{\sigma^2/n}.$$

Then

$$f(\theta)f(x|\theta) = const. \exp\{-\frac{1}{2}\left[c(\theta-\mu)^2 - \frac{2}{\sigma^2/n}(\theta-\mu)(x-\mu) + \text{function of } x\right]\}$$
$$= const. \exp\{-\frac{1}{2}c\left[(\theta-\mu)^2 - \frac{2(\theta-\mu)(x-\mu)}{c\sigma^2/n} + \text{function of } x\right]\}$$
$$= const. \exp\{-\frac{1}{2}c\left(\theta-\mu - \frac{x-\mu}{c\sigma^2/n}\right)^2 + \text{function of } x\}$$

(this last step is 'completing the square').