Applied statistics: part I

(*Krakow* 12/05/2020).

Contents

1 A primer in Bayesian inference

Different paradigms for statistical inference:

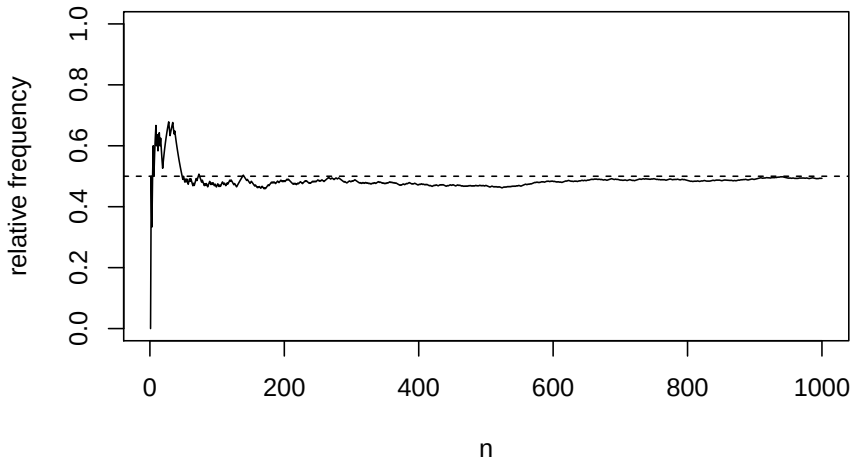
- classical or frequentist,
- Bayesian,
- fiducial,
- ...

1.1 On the meanings of probability

"Probability and Proof" from Prof. Dawid's website (`www.statslab.cam.ac.uk/~apd/`). There is agreement on the math for probability but not on the interpretation.
Different meanings:

- statistical
- classical
- empirical or frequentist
- subjective
- logical
- ...

Illustration of empirical probability



As we will see, there are limitations to this definition which serves as a basis for frequentist statistics. We will now view probability as 'logical' or 'subjective'.

R.T. Cox (1946) Probability, frequency and reasonable expectation, American Journal of Physics. vol 14, pp 1-13. Plausibility quantifier for a **rational person** in a **given context**. It should satisfy the three following properties:

1. degrees of plausibility are represented by real numbers (the larger the number the larger the plausibility)
   notations
   - $\bar{A}$: assertion A is false
   - $P(A)$: plausibility of assertion A
   - $P(AB)$: plausibility that assertions A and B are simultaneously true
   - $P(A+B)$: plausibility that either assertion A or B are true
   - $P(A|B)$: plausibility that A is true knowing that B is true
2. reason consistently;
   - if a conclusion can be reached in more than one way, then it should lead to the same degree of plausibility
   - all the available evidence should be used to quantify plausibility: no information can be left out
   - equivalent states of knowledge should give same plausibility assignments
3. qualitative correspondence with common sense

- consider statments A and B and some contextual information H
- assume that an update of H is given by H'
- assume that the plausibilities are affected as follows:
  - $P(A|H') > P(A|H)$
  - $P(B|AH') = P(B|AH)$

  then the requested property is that

$$P(AB|H') > P(AB|H)$$

  more plausability after update of the information

Cox(1946) this gives us back the Kolmogorov axioms from probability theory; informally,

- *Bounds:*

$$0 \leq P(A|H) \leq 1$$

  where A is an event, $P(A|H) = 0$ if a is impossible and $P(A|H) = 1$ if a is certain in the context H.

- *Addition rule:* If A and B are mutually exclusive (i.e. one at most can occur)

$$P(A \cup B|H) = P(A|H) + P(B|H).$$

- *Multiplication rule:* For any events A and B,

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

  We say that A and B are independent if $P(A \text{ and } B|H) = P(A|H)P(B|H)$ or equivalently $P(A|B, H) = P(A|H)$.

About subjectivity and context:

- All probabilities are conditional on context H
- They are **Your probabilities** for an event, not a property of the event
- Probabilities are therefore subjective and can be given for unique events, e.g. the probability of aliens openly visiting earth in the next 10 years
- They express **Your relationship** to the event - different stakeholders will have different information and different probabilities

Is probability Chance or Ignorance?
We can think of two broad types (at least) of uncertainty

- Aleatory: essentially unpredictable
- Epistemic: due to lack of knowledge

From subjectivist point of view, no need to worry about distinction between them, they are just uncertainties, e.g

- probability that it will rain tomorrow ?

  we often say that the probability of rain tommorow is 3/4. Is tomorrow weather random or deterministic ? meteorologic models are complex dynamic deterministic. But uncertainies w.r.t initial conidtions and chaos phenomena make these previsions uncertain and random...

- what is the height of Nigara Falls ?

  From a practical point of view it is also natural to give probability law to non random parameters. For example, it is natural to give a probability of 0,9 that the height of nigara falls is in beween 40 and 70m. This height is of course not random, we are just caracterizing our uncertainty about this true value.

Subjectivity is of often seen as a major disadvantage, however one should not confuse subjectivity with the **difficulty to translate beliefs about an assertion A into the number $P(A|H)$**.

Inversion of probabilities is described by the **Bayes formula.** It provides a formal mechanism for learning from experience. Let us consider two events A, E, such that $P[E \neq 0]$

$$P[A|E] = \frac{P[E|A]P[A]}{P[E|A]P[A] + P[E|A^c]P[A^c]} = \frac{P[E|A]P[A]}{P[E]}$$

where A^c denotes the complementary event.
Now let us look at the following ratio

$$\frac{P[A|E]}{P[B|E]} = \frac{P[E|A]P[A]}{P[E]} \frac{P[E]}{P[E|B]P[B]}$$

If events A and B are equiprobable, then

$$\frac{P[A|E]}{P[B|E]} = \frac{P[E|A]}{P[E|B]}$$

for 2 equiprobable causes, the ratio of the probabilities of the causes conditionally to the effect is the same than the ratio of the probability of the effect conditionally to the causes.

1.2 Bayesian statistical model

The specification of a Bayesian model starts with a statistical model $\mathcal{P} = \left\{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\right\}$ $d \geq 1$, fixed. The beliefs on θ are represented by a random variable $\bar{\Theta} \sim \Pi$, where Π is a **prior** probability distribution endowing the parameter space Θ.

The sample $(X_1, \ldots, X_n)$ is viewed as realizations of the conditional law $X|\theta \sim P_\theta$ given that $\bar{\Theta} = \theta$. The conditional distribution of $\bar{\Theta}|X_1, \ldots, X_n$ is the **posterior** distribution. The model is said to be dominated, when the distributions P_θ and Π admit densities $p(\cdot|\theta)$ and $\pi(\cdot)$ respectively. From now on, for simplicity of notations, all variables will be written in small letters, the context will

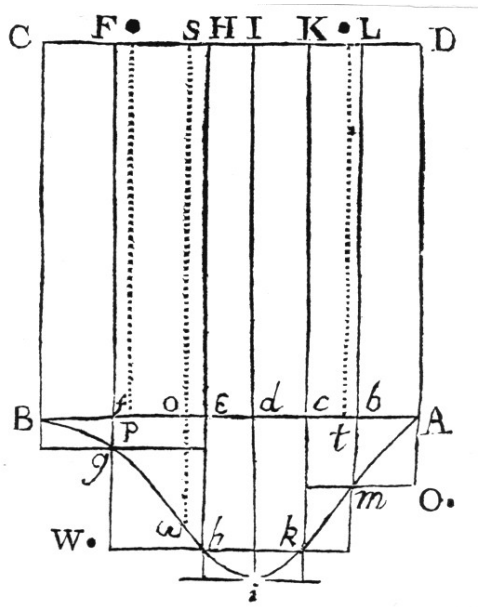
allow to understand whether it concerns random variable or not). Using the Bayes theorem, the posterior is written as follows:

$$\pi(\theta|x_1,\ldots,x_n) = \frac{\prod_{i=1}^n p(x_i|\theta)\pi(\theta)}{\int \prod_{i=1}^n p(x_i|\theta)\pi(\theta)d\theta}$$

Starting from the effects (observations) one try to infer on the causes (the parameters $\theta \in \Theta$) of the Data Generating Process. This is the inversion w.r.t probabilistic approach; Likelihood = inverted density! $l(\theta; x) = p(x|\theta)$.

1.2.1 Historical foundations

Bayes Billard table



(Bayes, T. (1763) An Essay towards solving a problem in the Doctrine of Chances, Philosophical transactions of the Royal Society, 53, 370-418). A billard ball is launched on a line of length 1. The location $\theta \in (0,1)$ where the ball stop is modelled by a uniform law. A second ball is launched n times from identical manner and y is the number of times that this ball is on the right of the first one. Knowing y, what can we say on θ?

$$\begin{array}{rl}
\text{Likelihood} & \left\{ \begin{array}{l} y|\theta \sim Bin(n,\theta) \\ p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}. \end{array} \right. \\
\text{Prior} & \left\{ \begin{array}{l} \theta \sim U[0,1] \\ \pi(\theta) = 1_{\{\theta\in(0,1)\}}. \end{array} \right. \\
\text{Posterior} & \left\{ \begin{array}{l} \theta|y \sim Be(y+1, n-y+1) \\ \pi(\theta|y) \propto \theta^y(1-\theta)^{n-y}. \end{array} \right.
\end{array}$$

```
> theta = 0.17
> # n observations
> n = 60
> # generate example data
> set.seed(123)
> x = rbinom(n = n, prob = theta, size = 1)
```

Reminder: one classical frequentist solution Use normal approximation and build a 95 percent confidence interval for the proportion:

$$\hat{\theta} \pm z_{0.975} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

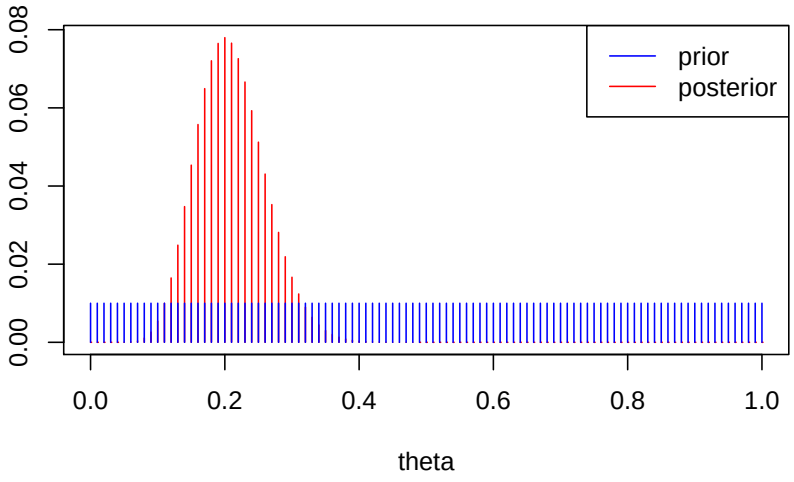
```
theta_hat,  0.2

95% confidence interval,  0.0988 0.3012
```

Remark: The normal approximation deteriorates when the unknown is close to 0 or 1.

1.2.2 Bayesian approach for the Binomial model

using a discrete uniform prior on θ over the set values $\{i/100; 0 \leq i \leq 100\}$.

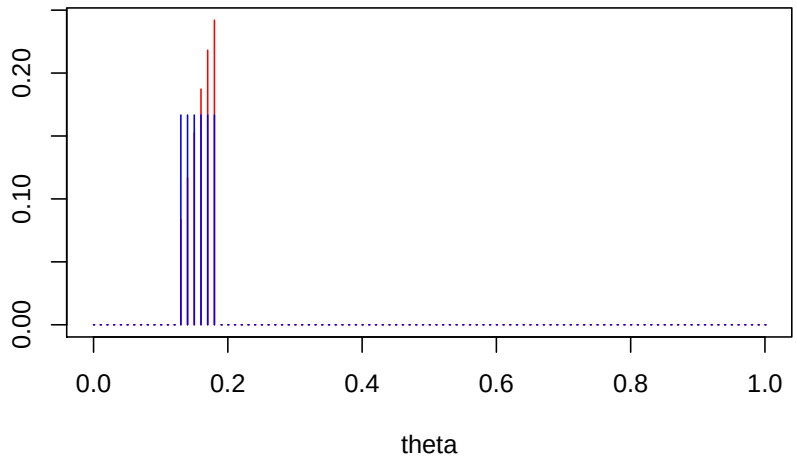
```
prior probability of theta in [0.14;0.19] 0.06

posterior probability of theta in [0.14;0.19] 0.3492532
```



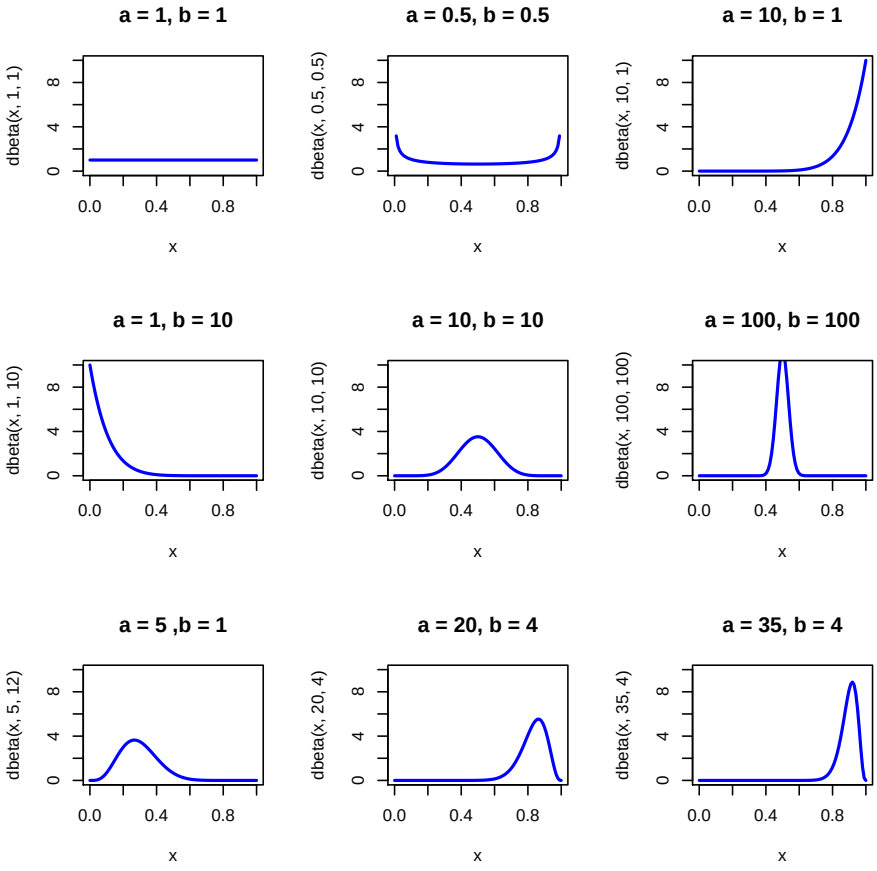
using a truncated discrete uniform prior on θ over the set values $\{i/100; 0 \leq i \leq 100\}$.

```
prior probability of theta in [0.14;0.19] 0.8333333
```

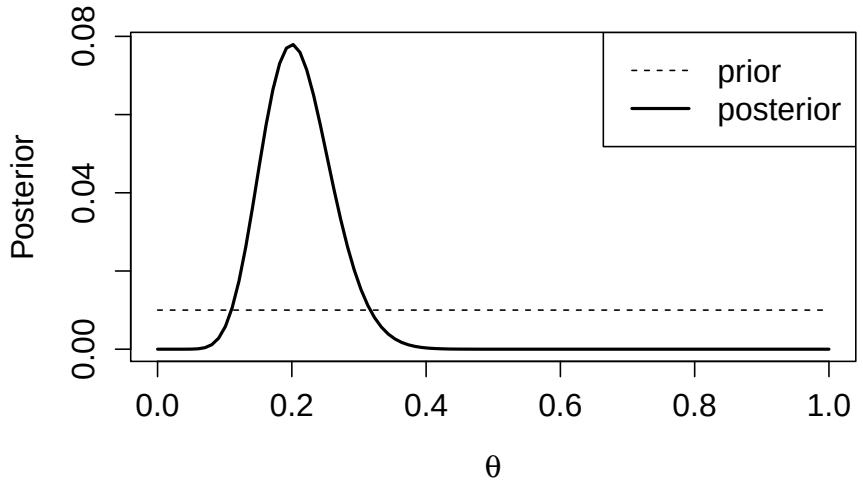
```
posterior probability of theta in [0.14;0.19] 0.9164723
```



using a continuous beta prior Shapes of the beta distribution




```
> #set the prior distribution parameters
> a  = b = 1
> # posterior parameters
> a_post = y + a
> b_post = n - y + b
> # Plot the prior, likelihood and posterior
> L = 100
> theta <- seq(0,1,length=L)
> theta_prior <- dbeta(theta,a, b)/L
> theta_post <- dbeta(theta,a_post, b_post)/L
> cex <- 1.25
> plot(theta,theta_post,type="l",lty=1,lwd=2,
+      cex.lab=cex,cex.axis=cex,
+      xlab=expression(theta),ylab="Posterior" )
> lines(theta,theta_prior,lty=2)
> legend("topright",c("prior", "posterior"), lty=c(2,1),lwd=c(1,2), cex=cex)
>
```



1.2.3 Summarizing the information from the posterior

Definition 1.1. *Consider a parametric Bayesian statistical model for the observation x, which consists in a sampling family $\{P_\theta; \theta \in \Theta\}$ and a prior distrubtion Π on $\theta \in \Theta$. We define*

- *the posterior mean*

$$\bar{\theta} = \mathbb{E}_{\theta|x}(\theta) = \int \theta d\Pi(\theta|x)$$

- *the posterior mode, denoted as $\hat{\theta}_m$: i.e., the point(s) where the posterior density is maximum,*

$$\bar{\theta}_m = \arg\max_{\theta \in \Theta} \pi(\theta|x).$$

- *the posterior variance*

$$\bar{v} = \mathbb{V}ar(\theta|x) = \int (\theta - \bar{\theta})^2 d\Pi(\theta|x).$$

Definition 1.2. *Let $\Theta \subset \mathbb{R}$, and $F_{\theta|x}$ be the cdf of the posterior distribution $\Pi(\theta|x)$. Suppose $F_{\theta|x}$ as an inverse $F^{-1}_{\theta|x}$. We define the posterior quantiles as*

$$q_{\theta|x}(t) = F^{-1}_{\theta|x}(t).$$

Definition 1.3 (credible region)**.** *A region C of Θ is said to be a $(1-\alpha)$ credible region for $\Pi(\cdot|x)$ if and only if*

$$\Pi(\theta \in C|x) \geq 1 - \alpha.$$

Remarks:

- there exists an infinite number of $(1-\alpha)$ credible regions. In single parameter case, we seek for
  - the shortest interval,
  - an equal tailed interval, i.e., $[l, u]$ such that $\Pi(\theta \leq u|x) = \Pi(\theta \geq l|x) = \epsilon/2$,
  - an Highest Posterior Denisty (HPD) interval.

Applied to our previous example

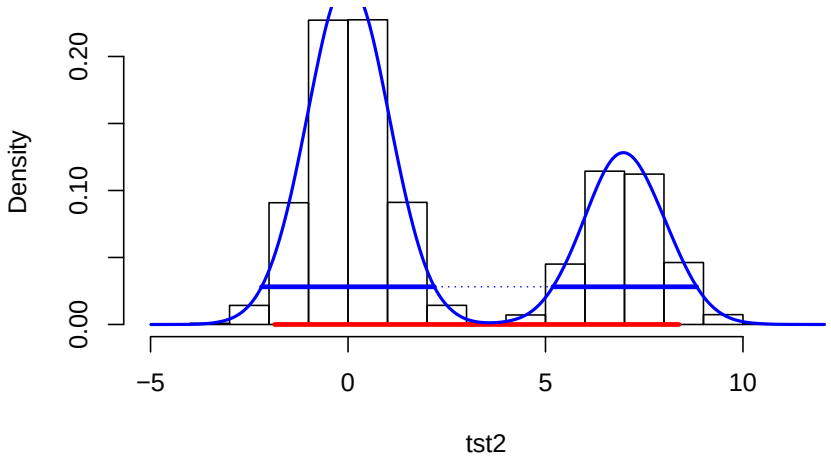
```
> # Summarize the posterior in a table:
> A    <- y+a
> B    <- n-y+b
> Mean <- A/(A+B)
> Var  <- A*B/((A+B)*(A+B)*(A+B+1))
> SD   <- sqrt(Var)
> Q05  <- qbeta(0.05,A,B)
> Q95  <- qbeta(0.95,A,B)
> P0.1  <- pbeta(0.1,A,B, lower.tail = FALSE)
> output <- cbind(Mean,SD,Q05,Q95,P0.1)
> output <- round(output,4)
> output

       Mean     SD    Q05    Q95   P0.1
[1,] 0.2097 0.0513 0.1309 0.2992 0.9934

> cat("95 % credible interval", c(qbeta(0.025, a_post, b_post), qbeta(0.975, a_post, b_post)))

95 % credible interval 0.1186424 0.3184212
```

Definition 1.4 (HPD region). *C_α^π is an HPD region if and only if $C_\alpha^\pi = \{\theta; \pi(\theta|x) \geq h_\alpha\}$, where h_α is some value such that* $\sup_h \int_{\{\theta:\pi(\theta|x)>h\}} \pi(\theta|x)d\theta \geq 1-\alpha$.

Here we illustrate a very appealing feature of HPI intervals considering a bimodal distribution

```
    lower     upper
-1.855720  8.377764
attr(,"credMass")
[1] 0.95

    lower     upper
-1.936889  8.443473
attr(,"credMass")
[1] 0.95
attr(,"height")
[1] 0.02812475

[1] 0.02812475
```

```
         begin      end
[1,] -2.187219 2.185216
[2,]  5.189179 8.827312
attr(,"credMass")
[1] 0.95
attr(,"height")
[1] 0.02812475
```



'blue horizontal lines' correspond to the HPD interval; 'red horizontal line' to an equal tail credible interval.

1.2.4 Monte carlo approximation to posterior characteristics

We are typically interested in quantities of the form

$$\mathbb{E}_{\theta|x}(g(\theta)) = \int_{\Theta} g(\theta)\pi(\theta|x)d\theta. \tag{1}$$

Since $\pi(\theta|x)$ is a density, we can use a sample $\theta_1, \ldots, \theta_M$ generated from the posterior density $\pi(\theta|x)$ and approximate (1) by an empirical average

$$\bar{g}_M := \frac{1}{M}\sum_{m=1}^{M} g(\theta_m).$$

By the SLLN, $\bar{g}_M$ converges a.s. to $\mathbb{E}_{\theta|x}(g(\theta))$.
The rate of convergence of $\bar{g}_M$ can be assess by the variance

$$\mathbb{V}ar(\bar{g}_M) = \frac{1}{M}\int_{\Theta}(g(\theta) - \bar{g}_M)^2\pi(\theta|x)d\theta$$

which is estimated from the sample $\theta_1, \ldots, \theta_M$ by

$$\hat{\mathbb{V}ar}(\bar{g}_M) = \frac{1}{M^2} \sum_{m=1}^{M} (g(\theta_m) - \bar{g}_M)^2.$$

```
> M=10000
> theta_post <- rbeta(n = M, a_post, b_post)
> cat("mean", mean(theta_post))

mean 0.2098775

> cat("95 % credible interval", c(round(quantile(theta_post, probs = 0.025),4),
+                                 round(quantile(theta_post, probs = 0.975), 4)))

95 % credible interval 0.1185 0.3167

> cat("The posterior probability that theta is smaller than 0.1 is", sum(theta_post<0.1)/M)

The posterior probability that theta is smaller than 0.1 is 0.0061
```

1.2.5 Convergence of the posterior distribution

In a pure Bayesian approach everything is random, there is no true parameter. Bayesian inference stops with exploiting the posterior. Nevertheless, we can consider an hybrid Bayesian framework under which we will study the properties of the posterior under the law P_{θ_0} from which the observations have been generated (alike in the frequentist approach). We are then interested in the **frequentist properties of the posterior** among which the consistency as defined hereafter for any $\delta > 0$:

$$\Pi(\{\theta : \|\theta - \theta_0\| > \delta\} \, | x_1, \ldots, x_n) \to 0, \; P_{\theta_0}, \text{ as } n \to \infty.$$

Hereafter we give a more general result which basically states that bayesian procedures (for frequentist inference) give the same results as if we were using the asymptotic distribution of maximum likelihood estimators. Before giving this result, we give the following definition and lemma.

Definition 1.5. *Let P, Q be two probability measures with $dP = pd\mu$ and $dQ = qd\mu$. The L_1-distance between P and Q is $\|P - Q\|_1 = \int |p - q| d\mu$.*

Lemma 1.1. *Let P, Q be two probability measures on $\mathcal{X}$ with $\sigma-$algebra $\mathcal{A}$. Then,*

$$\|P - Q\|_1 = 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

Theorem 1.1 (Bernstein von-Mises)**.** *Let $\mathcal{P} = \left\{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\right\}$ be a Bayesian (regular) statistical model. Let x be our sample generated from P_{θ_0}, $\theta_0 \in \Theta$. Suppose that the prior distribution on Θ satisfies*

- Π *has a density π w.r.t the Lebesgue measure on $\mathbb{R}^d$.*

- $\pi(\theta_0) > 0$ *and* $\pi(.)$ *is continuous at* θ_0

Assume that the Fisher information matrix $I(\theta_0)$ *is invertible at* θ_0*. Let* $\hat{\theta}^{MLE}$ *be the MLE in this model. Then,*

$$\left\| \Pi(\cdot|x) - \mathcal{N}\left(\hat{\theta}^{MLE}, \frac{I(\theta_0)^{-1}}{n}\right)(\cdot) \right\|_1 \overset{P_{\theta_0}}{\to} 0, \ n \to \infty.$$

In the Bayesian setting, there is no need to estimate the Fisher information matrix which is generally in practice unknown.

Theorem 1.2. *Consider a statistical model for which the parameter space* $\Theta \subset \mathbb{R}$ *and that we obtain a posterior* $\Pi(\cdot|x)$ *from the prior* Π *and data* $x = (x_1, \ldots, x_n)$*. Let* $\alpha \in (0,1)$*,* z_α *be the* α*-quantile of a standard normal. Suppose that the BvM theorem applies, then, for* $l_n(x), u_n(x)$ *defined as*

$$\Pi\left((-\infty, l_n)|x\right) = \alpha/2$$
$$\Pi\left((u_n, \infty)|x\right) = \alpha/2.$$

Then,

$$[l_n, u_n] = \left[\hat{\theta}^{MV} - \frac{z_{1-\alpha/2}}{\sqrt{nI(\theta_0)}}\left(1 + o_P(1)\right), \hat{\theta}^{MV} + \frac{z_{1-\alpha/2}}{\sqrt{nI(\theta_0)}}\left(1 + o_P(1)\right)\right],$$

where $o_P(1)$ is an arbitrary quantity going to 0 as $n \to \infty$ under P_{θ_0}.

1.3 Conclusions

- principles of bayesian inference: given a statistical model, all its parameters are random, then equipped with probability distributions.
- inference is done conditionally on the observed data: **inversion principle**
- Main critics:
  - **from frequentists:** prior specification !
  - **from machine learning perspective:** need to build a full statistical model on the data and therefore it is a limitation for large and complex data.
- Bayesian society `https://bayesian.org/`
- A key for the future of deep learning `http://bayesiandeeplearning.org/`.
- A timeline for bayesian statistics history [insert picture]

1.4 Exercice

1.4.1 Exercise 1: Comparing two proportions

The results of clinical trial to test wether an intra-uterin device (IUD) emitting progesteron can avoid emdometriosis in woman suffering of breast cancer under Tamoxifen Treatment (TF). At the end of the study, 5 of the 56 women in the IUD group had a fibrome while in the control group, 13 over 53 had one.

- find the posterior distribution of the Proportion of Fibrome (PF) in each group using a uniform prior.
- compute 95% credibility interval for the PF for each group.

  In a previous clinical study based on 20 patients and 20 controls, respectively 8 and 12 had a fibrome.
- You do not trust fully this experiment but you are still willing to use this information as prior information in the previous models. Hence you decide to give it a 10 unit sample size equivalence for both treatment and control group. Find the corresponding posterior distribution of TF in each group.
- find the distribution of the difference in proportions of TF.
- compute the posterior probability that the proportion of TF is smaller in the tamoxifen group.
- answer to the previous quesions using normal approximations to the posterior.

Solution: Let n_1 be the number of patients in the group 1.
n_2 be the number of patients in the group 2.
By assumption $y_1 \perp y_2|\theta_1, \theta_2$.

$$\begin{aligned} y_1|\theta_1 &\sim Bin(n_1, \theta_1) \\ y_2|\theta_2 &\sim Bin(n_2, \theta_2). \end{aligned}$$

we are interested in $\delta = \theta_2 - \theta_1 = g(\theta_1, \theta_2)$
Set up an independent prior

$$\begin{aligned} \pi(\theta_1, \theta_2) &= \pi(\theta_1)\pi(\theta_2) \\ \theta_1 &\sim Be(\alpha_1, \beta_1) \\ \theta_2 &\sim Be(\alpha_2, \beta_2). \end{aligned}$$

The **joint** posterior is

$$\begin{aligned} \pi(\theta_1, \theta_2|y_1, y_2) &\propto \theta_1^{y_1}(1-\theta_1)^{n_1-y_1}\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}\theta_2^{y_2}(1-\theta_2)^{n_2-y_2}\theta_2^{\alpha_2-1}(1-\theta_2)^{\beta_2-1} \\ &= \theta_1^{y_1+\alpha_1-1}(1-\theta_1)^{n_1-y_1+\beta_1-1}\theta_2^{y_2+\alpha_2-1}(1-\theta_2)^{n_2-y_2+\beta_2-1}. \end{aligned}$$

We remark directly that the posterior can be factorized as follows

$$\pi(\theta_1, \theta_2|y_1, y_2) = \pi(\theta_1|y_1)\pi(\theta_2|y_2)$$

where

$$\theta_1|y_1 \sim Be(\alpha_1^*, \beta_1^*)$$
$$\theta_2|y_2 \sim Be(\alpha_2^*, \beta_2^*).$$

We are interested in evaluating

$$\int_A \pi(\theta_1, \theta_2|y_1, y_2) d\theta_1 d\theta_2,$$

where $A = \{(\theta_1, \theta_2) : \theta_2 - \theta_1 \leq c\}$.
We can use the monte carlo principle to evaluate this integral:
At the m-th iteration $(1 \leq m \leq M)$

- 1. generate

$$\theta_1^{(m)} \text{ from } \theta_1|y_1 \sim Be(\alpha_1^\star, \beta_1^\star)$$
$$\theta_2^{(m)} \text{ from } \theta_2|y_2 \sim Be(\alpha_2^\star, \beta_2^\star)$$

- 2. compute $\delta^{(m)} = \theta_2^{(m)} - \theta_1^{(m)}$
  $\left\{\delta^{(m)}, 1 \leq m \leq M\right\}$ is a sample from $\pi(\delta|y_1, y_2)$.

- 3. return to step 1.

Normal approximation to the posterior
There are 2 types of possible approximation: moment matching and modal approximation

- **moment matching**

  $$\theta|x \dot{\sim} \mathcal{N}(\mathbb{E}(\theta|x), \mathbb{V}ar(\theta|x)),$$

  where $\dot{\sim}$ denote a distribution approximation.

  It allows to build $(1-\alpha)$-level credible intervals

  $$\mathbb{E}(\theta|x) \pm z_{1-\alpha/2}\sqrt{\mathbb{V}ar(\theta|x)}.$$

- **modal approximation**

  Let f be a nonnegative unimodal function with mode $\tilde{\theta}$. A quadratic Taylor expansion of $\log f(\theta)$ at $\tilde{\theta}$ gives

  $$\log f(\theta) \approx \log(f(\tilde{\theta})) - \frac{1}{2}(\theta - \tilde{\theta})' Q (\theta - \tilde{\theta}),$$

  where $Q_{ij} = -\left[\frac{\partial^2}{\partial\theta_i\theta_j} \log f(\theta)\right]_{\theta=\tilde{\theta}}$

  Exponentiate each side

  $$f(\theta) \approx f(\tilde{\theta}) e^{-\frac{1}{2}(\theta-\tilde{\theta})' Q (\theta-\tilde{\theta})}$$

Then integrating

$$\int f(\theta)d\theta \approx f(\tilde{\theta}) \int e^{-\frac{1}{2}(\theta-\tilde{\theta})'Q(\theta-\tilde{\theta})}d\theta$$
$$= \frac{f(\tilde{\theta})}{\sqrt{|Q|/2\pi}}.$$

If $f(\theta)$ is a non normalized density, then the modal approximation yields

$$\theta|x \dot{\sim} \mathcal{N}(\tilde{\theta}, Q^{-1}).$$

Back to our case, $f(\theta) = \pi(\theta|y) \propto \theta^{\alpha^\star - 1}(1-\theta)^{\beta^\star - 1}$

Consider our specific example in the binomial model $\theta|x \sim Bin(\alpha, \beta)$

$$Mode(\theta|x) = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

$$\theta|x \dot{\sim} \mathcal{N}\left(\frac{\alpha-1}{\alpha+\beta-2}, \frac{(\alpha-1)(\beta-1)}{(\alpha+\beta-2)^3}\right).$$

```
> y.T.1 <- 5    # number of fibroms in treatment group
> n.T.1 <- 56  # number of patients in treatment group
> y.C.1 <- 13 # number of fibroms in control group
> n.C.1 <- 53 # number of patients in control group
> M <- 10000 # number of monte carlo iterations
> # generate sample of size M from the posterior distributions
> post.T <- rbeta(n = M, y.T.1+1, n.T.1-y.T.1+1)
> post.C <- rbeta(n = M, y.C.1+1, n.C.1-y.C.1+1)
> # 95% credible interval within each group
> quantile(post.T, c(0.025,0.975) )

      2.5%      97.5%
0.03918247 0.19389898

> quantile(post.C, c(0.025,0.975) )

     2.5%     97.5%
0.1484413 0.3793201

> # data from the previous experiment for both control and treatment group
> y.T.2 <- 8
> n.T.2 <- 20
> y.C.2 <- 12
> n.C.2 <- 20
> # a priori informatif with equivalent sample size of 10 units per group
> alpha.T <- y.T.2/2
> beta.T <- (n.T.2 - y.T.2)/2
> alpha.C <- y.C.2/2
```

```
> beta.C <- (n.C.2 - y.C.2)/2
> # parameter of the posterior with this informative prior
> alpha.prime.T <- y.T.1  + alpha.T
> beta.prime.T <- n.T.1 - y.T.1 + beta.T
> alpha.prime.C <-y.C.1  + alpha.C
> beta.prime.C <- n.C.1 - y.C.1 + beta.C
> # utiliser cette information dans l'a priori : ?quivalent ?chantillon de 10 par groupe
> post.T <- rbeta(n = M, alpha.prime.T, beta.prime.T)
> post.C <- rbeta(n = M, alpha.prime.C, beta.prime.C)
> quantile(post.T, c(0.025,0.975) )

      2.5%      97.5%
0.06467356 0.22839754

> quantile(post.C, c(0.025,0.975) )

     2.5%     97.5%
0.1950910 0.4208191

> post.diff <- post.T - post.C
> quantile(post.diff, c(0.025,0.975) )

       2.5%       97.5%
-0.30706649 -0.02744102

> # posterior probability
> sum(post.diff <0)/M

[1] 0.9898

> # normal approximation of the posterior by moment matching
> X = seq(0,1, length = 10000)
> post_mean_T = alpha.prime.T / (alpha.prime.T + beta.prime.T)
> post_var_T = (alpha.prime.T * beta.prime.T) /
+   ( (alpha.prime.T + beta.prime.T)^2 * (alpha.prime.T + beta.prime.T +1))
> post_mean_C = alpha.prime.C / (alpha.prime.C + beta.prime.C)
> post_var_C = (alpha.prime.C * beta.prime.C) /
+   ( (alpha.prime.C + beta.prime.C)^2 * (alpha.prime.C + beta.prime.C +1))
> hist(post.T, breaks = 50, prob = T)
> lines(X, dnorm(X, post_mean_T, sd = sqrt(post_var_T)), col = 2)
> hist(post.C, breaks = 50, prob = T)
> lines(X, dnorm(X, post_mean_C, sd = sqrt(post_var_C)), col = 2)
>
```

Additional remark: from the joint posterior we can easily compute meaningful quantities alike the odd ratio or log odd ratio.

```
> OR = (post.T / (1- post.T)) / (post.C / (1- post.C))
> hist(OR, breaks = 50, prob = T)
> cat("95% CI log odd ratio of the effect of treatment",
+     c(quantile(log(OR), 0.25),quantile(log(OR), 0.975)))

95% CI log odd ratio of the effect of treatment -1.341752 -0.165543
```