# Appendix

Christophe Ambroise

2021-2022

# Snooker Frequentist example I

Let consider a snooker table, and a ball. The ball $A$ is launched perpendicularly to a reference edge and stop at a distance $l$ from this billiard edge.

A second ball $B$ is thrown $n$ times and we denote by $X$ the number of times that $B$ stops at a distance $l'$ of the edge such that $l' > l$.

We try to estimate the proportion $p$ the number of times u $l' > l$ knowing that $X = x$. We assume that $X \sim \mathcal{B}(n, p)$ (binomial distribution).

In this case we have a sample of size 1 and the likelihood of parameter $p$ writes:

$$\ell(p; x) = C_n^x p^x (1-p)^{n-x},$$

and by canceling the first derivative of the log-likelihood, we obtain

$$\hat{p}_{MV} = arg \max_p \ell(p; x) = \frac{x}{n}$$

# Snooker Frequentist example II

If many launches have been performed, the estimate of $p$ will be satisfactory. On the other hand, if only one throw is observed, we find:

- $x = 0$ gives $\hat{p} = 0$;
- $x = 1$, which gives $\hat{p} = 1$.

In both cases the estimate appears intuitively of very poor quality.

# Bayesian Statistics in a nutshell I

The maximum likelihood framework produces point estimates

## Reverend T. Bayes (1701-1761)

Two years after the death of the Reverend T. Bayes (1701-1761), a friend of this one, published his *essay in view of solving the doctrine of chances* (Bayes 1763).

## Parameters are random variables

In this little booklet, which is the source of the inference modern Bayesian statistic,

- Parameters are no longer treated as deterministic quantities but random as are observations.
- The dual role of the parameters $\theta$ and the observations $x$ is described thanks to conditioning by Bayes' theorem:

# Bayes Theorem

For a distribution (called *Prior*) $\pi$ on the parameter $\theta$, and a observation $x$ of density $f(x|\theta)$, the distribution of $\theta$ conditionally on $x$

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

The main innovation of the Bayesian statistical model is the law $\pi$ on the model parameters.

# Prior, Posterior, Cost

Thus, in the context of a Bayesian statistical approach three functions must be specified:

- the law of the observations, the so-called likelihood $f(x|\theta)$;
- the prior distribution on the parameters, $\pi(\theta)$;
- the cost $C$ associated with the decision $\delta$ for parameters $\theta$.

Cost is a numerical measure of the quality of a decision.

# Bayes estimator

We call the Bayes estimator associated with a prior distribution $\pi$ and a cost $C$, any estimator $\delta^\pi$ which, given an observation vector $x$, minimizes the cost a posteriori

$$\rho(\pi, \delta|x) = E^\pi[C(\theta, \delta)|x] = \int_\theta \underbrace{C(\theta, \delta)}_{(\theta-\delta)^2}\pi(\theta|x)d\theta.$$

$$\mathbb{E}_{\theta|x}\left[(\theta-\delta)^2|x\right] = \mathbb{E}\left[\theta^2 - 2\delta\theta + \delta^2\right]$$

$$\hat{\theta} = \arg\min_\delta \rho(\pi,\delta|x) \quad \frac{\partial \rho(\pi,\delta|x)}{\partial\delta} = -2\mathbb{E}[\theta|x] + 2\delta = 0$$

$$\delta = \mathbb{E}_{\theta|x}[\theta]$$

# Conjugate Prior

When the prior and the posterior have the same form, we say that the prior is a conjugate prior for the corresponding likelihood. Conjugate priors are widely used because they simplify computation, and are easy to interpret

The posterior $p(\theta|x)$ summarizes everything we know about the unknown quantities $\theta$.

# Posterior Mean, Median, Mode

We can easily compute a point estimate of an unknown quantity by computing the posterior mean, median or mode.

However, the posterior mode, aka the MAP estimate, is the most popular choice:

- it reduces to an optimization problem, for which efficient algorithms often exist.
- MAP estimation can be interpreted in non-Bayesian terms, by thinking of the log prior as a regularizer (see Lasso e.g.)

Fonctions Gamma & Beta

$$y = (x-1)!$$

La fonction Gamma est "version" continue de

$$f(x) = (x-1)!$$

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x}\, dx$$

si $\alpha \in \mathbb{N}^*$ alors $\Gamma(\alpha) = (\alpha-1)!$

Montrons que $\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)$

$u = e^{-x}$ $\quad u' = e^{-x}$

$v = x^{\alpha}$ $\quad v' = \alpha\, x^{\alpha-1}$

$$\Gamma(\alpha+1) = \int_0^{\infty} x^{\alpha} e^{-x}\, dx$$

$$= \left[uv\right]_0^{\infty} - \int uv'$$

$$= \left[-e^{-x} x^{\alpha}\right]_0^{\infty} + \int \underbrace{+e^{-x}\alpha}\, x^{\alpha-1}\, dx$$

$$= \underbrace{\lim_{x\to\infty} -e^{-x} x^{\alpha}}_{0} + \alpha \underbrace{\int_0^{\infty} e^{-x} x^{\alpha-1}\, dx}_{\Gamma(\alpha)}$$

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

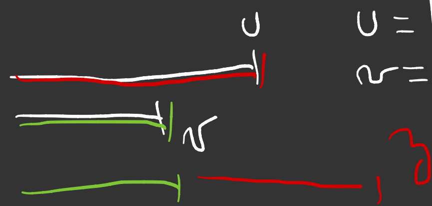
Mentrens que $B(\alpha,\beta)=\dfrac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$$\Gamma(x)\Gamma(y)=\int_{u=0}^{\infty} e^{-u}u^{x-1}\,du\int_{v=0}^{\infty}e^{-v}v^{y-1}\,dv$$

$$=\int_{u=0}^{\infty}\int_{v=0}^{\infty} e^{-u-v}u^{x-1}v^{y-1}\,du\,dv$$

$u=zt \qquad v=z(1-t)$ avec $t\in[0,1]$

$z\in\mathbb{R}^+$




$u=tz$

$v=(1-t)z$

$$F:(z,t)\longrightarrow(u,v)$$
$$(zt \quad z(1-t))$$

$$J_F=\begin{pmatrix}\partial u/\partial t & \partial u/\partial z\\ \partial v/\partial t & \partial v/\partial z\end{pmatrix}=\begin{bmatrix} z & t\\ -z & (1-t)\end{bmatrix}$$

$$|\det(J_F)|=$$
$$|z(1-t)+zt|=z$$

$$\int_{z=0}^{\infty}\int_{t=0}^{1} e^{-z}(zt)^{x-1}(z(1-t))^{y-1} z\,dt\,dz =$$

$$\underbrace{\int_{z=0}^{\infty} e^{-z} z^{x+y-1}\,dz}_{\Gamma(x+y)} \underbrace{\int_0^1 t^{x-1}(1-t)^{y-1}\,dt}_{B(x,y)}$$

$\Gamma(x)\Gamma(y) =$

---

Distribution de probabilité Beta

$\alpha, \beta > 0$ ,

$$p(x \mid \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{[x\in[0,1]]}}{B(\alpha,\beta)}$$

Supposons $X \sim B(\alpha, \beta)$

$$\mathbb{E}[X] = \frac{1}{B(\alpha,\beta)} \int_0^1 x \, x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

$$= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} =$$

$$= \frac{\Gamma(\alpha+1)\,\Gamma(\cancel{\beta})}{\Gamma(\alpha+1+\beta)} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\cancel{\beta})}$$

$$= \frac{\alpha\,\cancel{\Gamma(\alpha)}}{(\alpha+\beta)\,\cancel{\Gamma(\alpha+\beta)}} \times \frac{\cancel{\Gamma(\alpha+\beta)}}{\cancel{\Gamma(\alpha)}}$$

$$= \frac{\alpha}{\alpha+\beta}$$

$$X \sim \mathcal{B}(n,p)$$

$$P(X=x \mid p) = \binom{n}{x} p^{x} (1-p)^{n-x}$$

$$p \sim U[0,1] \qquad f(p) = \mathbb{1}(p \in [0,1])$$

$$P(a < p < b,\ X=x) = \int_a^b \binom{n}{x} p^{x} (1-p)^{n-x}\, dp$$

AVEC $0 \leq a \leq b \leq 1$

$$P(X=x) = \int_0^1 \binom{n}{x} p^{x} (1-p)^{n-x}\, dp = \binom{n}{x} B(x+1, n-x+1)$$

$$P(a < p < b \mid X=x) = \frac{\int_a^b \cancel{\binom{n}{x}} p^{x}(1-p)^{n-x}\, dp}{\cancel{\binom{n}{x}}\, B(x+1, n-x+1)}$$

$$p \mid X=x \ \underset{\text{loi}}{\sim} \mathcal{B}(x+1, n-x+1)$$

$$\mathbb{E}[p \mid X=x] = \frac{x+1}{\cancel{x}+1+n-\cancel{x}+1} = \frac{x+1}{n+2}$$

# Example of Bayesian inference I

Let us take the snooker example again and look for Bayesian approach

- the law on the observations is a binomial, $X \sim \mathcal{B}(\cancel{\backslash,})$; $\mathcal{B}(n,p)$
- the ball can equally probably stop ny distance from the edge. Hence $p \sim U_{[0,1]}$;
- let consider a quadratic cost: $C(p, \delta) = (p - \delta)^2$.

In that case,

$$
\begin{aligned}
\pi(p|X = x) &= \frac{C_n^x p^x (1-p)^{n-x} \mathbb{1}_{\{p \in [0,1]\}}}{\int_0^1 C_n^x p^x (1-p)^{n-x} dp} \\
&= \frac{p^x (1-p)^{n-x} \mathbb{1}_{\;p\;\, in[0,1]\}}}{\int_0^1 p^x (1-p)^{n-x} dp}.
\end{aligned}
$$

# Example of Bayesian inference II

The posterior distribution is therefore a beta distribution, $\mathcal{B}e(x+1, n-x+1)$. It is easy to show that the Bayes estimator associated with a distribution $\pi$ and a quadratic cost is the posterior mean

$$\delta^\pi(x) = E^\pi[p|x] = \int p \cdot \pi(p|x) dp.$$

The expectation of a random variable $X$ following a beta distribution, $\mathcal{B}e(\alpha, \beta)$ is given by

$$E[X] = \frac{\alpha}{\alpha + \beta}.$$

The Bayes estimator associated with the quadratic cost is therefore written

$$\delta^\pi(x) = \frac{x+1}{n+2}.$$

# Example of Bayesian inference III

If many launches have been performed, the estimate of $p$ by this Bayesian procedure will be very close to the estimator of the maximum of likelihood. On the other hand, if only one throw is observed, we find:

- $x = 0$ gives $\hat{p} = \frac{1}{3}$;
- $x = 1$ gives $\hat{p} = \frac{2}{3}$.

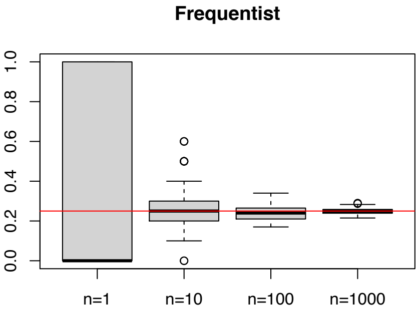
Both of these results seem reasonable.

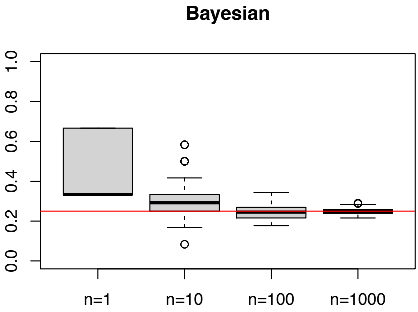
Note that by taking a cost which is equal to 0 if the decision is correct and 1 otherwise (cost 0-1), the Bayes estimator is, in this case, the same as that obtained by the maximum likelihood method.

Note that the Bayes estimators are justified for a size of finite sample, unlike estimators of the maximum of likelihood which only have asymptotic properties.

# R code

Comparing Bayesian and Frequentist estimator for $p = \frac{1}{4}$ and $n \in \{1, 10, 100, 1000\}$.



# Dirichlet distribution $Dir(\boldsymbol{\alpha})$ I

- Peter Gustav Lejeune Dirichlet (13 février 1805 , Düren – · 5 mai 1859 , Göttingen)
- continuous multivariate probability distribution parameterized by a vector $\boldsymbol{\alpha}$ of positive reals.
- used as prior distributions in Bayesian statistics,
- conjugate prior of the categorical distribution and multinomial distribution.

The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, ..., \alpha_K > 0$ has a probability density function

$$f\left(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K\right) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where the normalizing constant is the multivariate beta function.

# Dirichlet distribution $Dir(\alpha)$ II

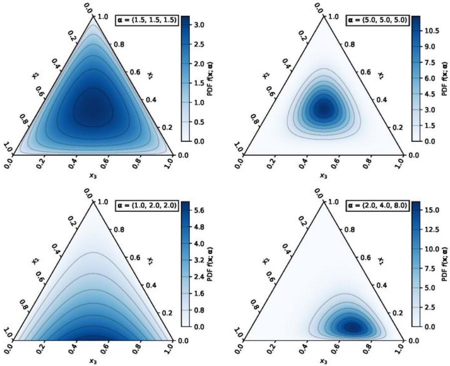



Figure 1: Dirichlet distributions

# Moments

Denoting $\alpha_0 = \sum_{i=1}^{K} \alpha_i$, we have

$$E[X_i] = \frac{\alpha_i}{\alpha_0},$$

$$Var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}.$$

$$Cov[X_i, X_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}.$$