MS = BAYESIAN STATISTICS

Bayesian reasoning provides rules of logic for updating prior beliefs in the light of new evidence. It is named after Thomas Bayes, an XVIII^{e}-century presbytarian minister who did maths on the side. His work "An Essay towards solving a problem in the doctrine of chances", published posthumously in 1763, contains theorems of conditional probability, that set the grounds for what is known now as Bayes's theorem. The Bayesian framework is now widely used across many fields, ranging from machine learning to epidemics, and medical studies.

For two events A and B, Bayes's theorem states that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \propto \mathbb{P}(B|A)\,\mathbb{P}(A)$$

"posterior" belief = updated belief after new evidence is collected

new evidence

"prior" probability or base rate

Update prior belief after having observed new evidence

- Ex = $A = \{$ presence of antibodies $\}$
  $B = \{$ positive test result $\}$
  $\mathbb{P}(B|A) \neq 1$ since medical tests are imperfect.

Bayes's theorem is a simple but powerful statement about proportions. It provides a way to integrate what

you previously thought with what you have learned and reach a conclusion that incorporates them both, with appropriate weights.

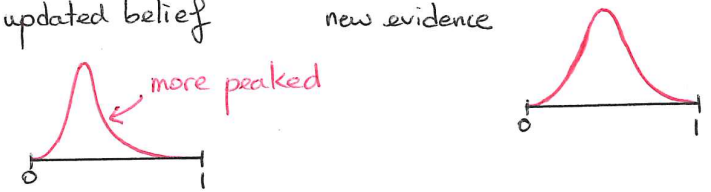
In the formula $\mathbb{P}(A|B) \propto \mathbb{P}(B|A)\,\mathbb{P}(A)$, the prior puts a mass on some probability, and assigns a zero probability to other values. It may be more realistic to spread this belief around this probability, and work instead with the "conditional density" version of Bayes's formula: $f(p|x) \propto f(x|p)\,f(p)$

posterior

updated belief

new evidence

prior density



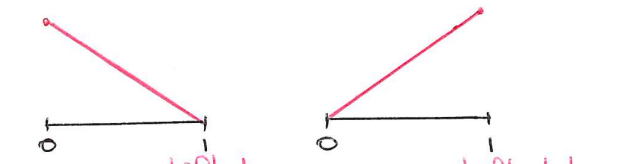
The last written formula describes a situation where

- $X|p \sim B(p)$ i.e. $f(x|p) = p^{x}(1-p)^{1-x}$ $x \in \{0,1\}$.
- $p \sim f(p)$ some prior probability distribution. When nothing is known about p beforehand, it is common to take a uniform density over $[0,1]$,
  $p \sim \mathcal{U}(0,1) = B(1,1)$
  ↑ Beta distribution

The posterior is $f(p|x) = p^{x}(1-p)^{1-x} = B(1+x, 2-x)$

Again, a Beta distribution. We say that Beta is the conjugate prior of the binomial distribution.

In particular, $x=0$ $\sim B(1,2)$ $x=1$ $\sim B(2,1)$



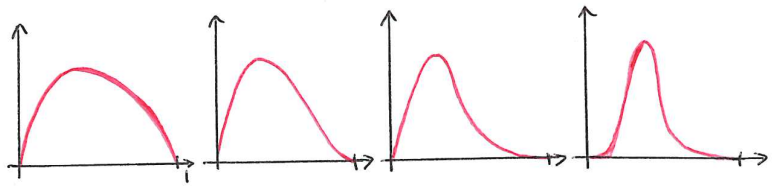
← mass shifted to the left (p more likely to be smaller after observing $x=0$)

→ mass shifted to the right (p more likely to be larger having observed $x=1$)

- The $B(\alpha,\beta)$ distribution has density $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$, with mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

As α increases, the distribution is shifted to the left
β —"— , —"— to the right.
& more peaked as $(\alpha+\beta)$ grows.

Ex:



(α,β) = $(2,3)$ $(2,10)$ $(20,100)$ $(200,1000)$

(and uniform with $(\alpha,\beta)=(1,1)$)

- Suppose now that we collect n observations $\mathcal{L}_n=\{X_1,\dots,X_n\}$, $X_1,\dots,X_n$ iid $B(p)$: $\mathbb{P}(X_i=1)=p$.

→ Consider the conjugate beta prior on $p \sim B(\alpha,\beta)$.

→ The likelihood is

$$L(X_1,\dots,X_n \mid p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

$$= p^{S_n}(1-p)^{n-S_n},$$

where $S_n = \sum_{i=1}^{n} X_i$.

↖ The MLE is $\hat{p} = \frac{S_n}{n}$ = proportion of successes.

→ The posterior is proportional to the product of the likelihood with the prior

$$p \mid \mathcal{L}_n \sim p^{S_n}(1-p)^{n-S_n}\, p^{\alpha-1}(1-p)^{\beta-1}$$

$$= p^{(S_n+\alpha)-1}(1-p)^{(n+\beta-S_n)-1}$$

$$\sim B(\alpha+S_n,\ n+\beta-S_n).$$

We retrieve again the conjugate property of the beta distribution.

→ The posterior mean is

$$\mathbb{E}(p \mid \mathcal{L}_n) = \frac{\alpha+S_n}{\alpha+\beta+n} = \left(\frac{n}{\alpha+\beta+n}\right)\frac{S_n}{n} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta}$$

$\to 1$ as $n\to\infty$ (MLE) ; $\to 0$ as $n\to\infty$ (prior mean)

= weighted linear combination of the prior mean & the MLE

It is a general feature of the posterior distribution : centered at a point that represents a compromise between prior info & the data. Also, the data get more weight as n gets large → $\mathbb{E}(p \mid \mathcal{L}_n) \approx \frac{S_n}{n}$ = MLE.

A uniform prior yields a posterior mean equal to $\frac{S_n+1}{n+2}$. Comparing with the MLE $\frac{S_n}{n}$, this accounts to augmenting the set of observations $\mathcal{L}_n = \{X_1, \dots, X_n\}$ with two additional artificial values: one extra success and one extra failure.

↑ ≡ regularization in small sample cases, preventing overfitting.

Ex: $\mathcal{L}_n = \{X_1 = X_2 = X_3 = X_4 = X_5 = 1\}$, $n=5$, the MLE is equal to 1, while the posterior mean is equal to $6/7$ (prior mean takes into consideration that $p<1$ has strictly positive probability)

→ The POSTERIOR PREDICTIVE DISTRIBUTION of a new observation can be calculated by integration of the posterior distribution:

$$\mathbb{P}(\tilde{X}=1 \mid \mathcal{L}_n) = \int_0^1 \underbrace{\mathbb{P}(\tilde{X}=1 \mid p, \mathcal{L}_n)}_{=p} \underbrace{f(p \mid \mathcal{L}_n)}_{B(\alpha+S_n, n+\beta-S_n)} dp$$

$\tilde{X}$: prediction for a new observation; $\mathcal{L}_n$: posterior

= mean of the $B(\alpha+S_n, n+\beta-S_n)$ distribution.

$$= \frac{\alpha+S_n}{\alpha+\beta+n}.$$

↑

In a Machine Learning context, $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Having observed a new input point x, the posterior predictive

distribution of Y given $X=x$ and $\mathcal{L}_n$ is

$$f(y \mid x, \mathcal{L}_n) = \int_\Theta f(y, \underbrace{\theta}_{\theta = \text{model parameter(s)}} \mid \mathcal{L}_n, x) \, d\underbrace{\theta}_{\uparrow}$$

$$= \int_\Theta \underbrace{f(y \mid \theta, \cancel{\mathcal{L}_n}, x)}_{\text{model distribution}} \underbrace{f(\theta \mid \mathcal{L}_n, x)}_{\text{posterior distribution}} d\theta \, .$$

→ The posterior variance is

$$\text{var}(p \mid \mathcal{L}_n) = \frac{(\alpha+S_n)(n+\beta-S_n)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

$$= \frac{\mathbb{E}(p \mid \mathcal{L}_n)\left(1-\mathbb{E}(p \mid \mathcal{L}_n)\right)}{\alpha+\beta+n-1}$$

↑

As $n \to \infty$, $\mathbb{E}(p \mid \mathcal{L}_n) \approx \frac{S_n}{n} = \text{MLE}$ (see p. 4),

and $\text{var}(p \mid \mathcal{L}_n) \approx \frac{1}{n}\frac{S_n}{n}\left(1-\frac{S_n}{n}\right) = \frac{\hat{p}(1-\hat{p})}{n}$.

Summarizing, as n gets large, $\mathbb{E}(p \mid \mathcal{L}_n) \approx \hat{p}$

& $\text{var}(p \mid \mathcal{L}_n) \approx n^{-1}\hat{p}(1-\hat{p})$

↑ Familiar expressions in non-Bayesian statistics.

In fact, under some regularity conditions, a CLT in Bayesian context holds:

$$\frac{p - \mathbb{E}(p \mid \mathcal{L}_n)}{\sqrt{\text{var}(p \mid \mathcal{L}_n)}} \mid \mathcal{L}_n \xrightarrow{d} \mathcal{N}(0,1), \text{ as } n \to \infty.$$

and we recover the usual "frequentist" CLT

$$\frac{n^{1/2}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow{d} \mathcal{N}(0,1).$$

↖ Combined with Slutsky theorem, replacing p by its consistent estimator $\hat{p}$.

⇒ Under some regularity conditions, the posterior is usually approximated using a normal distribution.

Remark : $\operatorname{var} p = \mathbb{E}\operatorname{var}(p|\mathcal{L}_n) + \operatorname{var}\mathbb{E}(p|\mathcal{L}_n)$
$\geqslant \mathbb{E}\operatorname{var}(p|\mathcal{L}_n)$

prior variance ↑ posterior variance

The posterior variance is on average smaller than the prior variance ⇒ observations reduce the uncertainty about p. This property holds more generally, and is not specific about binomial proportions. ∎

- **Application** : Comparing two proportions.

Suppose that we collect two independent learning samples :

$$X_{n_0} := \{X_{1,0}, \dots, X_{n_0}\}, \quad X_{1,0}, \dots, X_{n_0} \sim \mathcal{B}(p_0)$$
$$Y_{n_1} := \{Y_{1,1}, \dots, Y_{n_1}\}, \quad Y_{1,1}, \dots, Y_{n_1} \sim \mathcal{B}(p_1),$$

and that we wish to infer which one of p_0 or p_1 is the largest. In clinical trials, X_{n_0} might represent a control group, who was administered a placebo, while individuals in Y_{n_1} receive a new treatment. The X_i / Y_i indicate

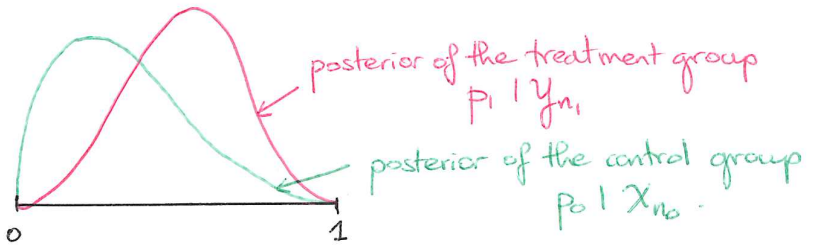
whether an individual recovered after some time, and the goal is to assess the new treatment's effectiveness.

Prior $p_0 \sim \mathcal{B}(\alpha_0, \beta_0)$, $p_1 \sim \mathcal{B}(\alpha_1, \beta_1)$

⇓

Posterior $p_0 | X_{n_0} \sim \mathcal{B}(\alpha_0 + S_{n_0}, n_0 + \beta_0 - S_{n_0})$, $S_{n_0} = \sum_{i=1}^{n_0} X_i$

$p_1 | Y_{n_1} \sim \mathcal{B}(\alpha_1 + S'_{n_1}, n_1 + \beta_1 - S'_{n_1})$, $S'_{n_1} = \sum_{i=1}^{n_1} Y_i$



In particular, we may be interested in the posterior distribution of $p_0 - p_1 \mid X_{n_0} \cup Y_{n_1}$. This can easily be done numerically using the following procedure :

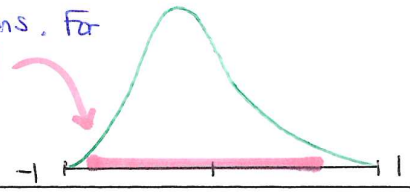
(i) Draw $p_0 | X_{n_0} \sim \mathcal{B}(\alpha_0 + S_{n_0}, n_0 + \beta_0 - S_{n_0})$
$p_1 | Y_{n_1} \sim \mathcal{B}(\alpha_1 + S'_{n_1}, n_1 + \beta_1 - S'_{n_1})$

(ii) Compute $\delta := p_0 - p_1$

(iii) Repeat (i)+(ii) as many times as desired

(iv) Draw a histogram of the values of δ

A summary of this posterior distribution may be used to draw conclusions. For example, is the 95% highest posterior density region strictly contained in some small interval



Remarks (i) To compute the exact analytical expression of the posterior distribution of δ, we proceed as follows.

Let $p_0 \mid X_{n_0} \sim B(\alpha_1, \beta_1)$
$p_1 \mid Y_{n_1} \sim B(\alpha_2, \beta_2)$,

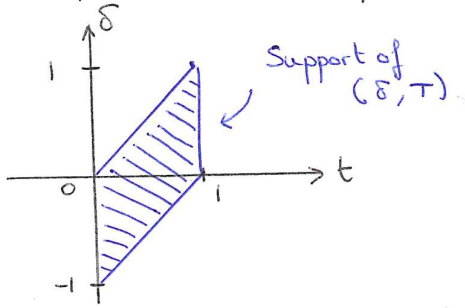
where $\alpha_1 = \alpha_0 + S_{n_0}$, $\beta_1 = n_0 + \beta_0 - S_{n_0}$
$\alpha_2 = \alpha_1 + S'_{n_1}$, $\beta_2 = n_1 + \beta_1 - S'_{n_1}$.

Consider the change of variable $\begin{cases} \delta = p_0 - p_1 \\ T = p_0 \end{cases}$

The Jacobian of this transformation is equal to 1, so that the joint density of (δ, T) is

$$f_{\delta,T}(d,t) = f_{p_0,p_1}(t, t-d)$$
$$= B(t \mid \alpha_1, \beta_1)\, B(t-d \mid \alpha_2, \beta_2),$$

where $0 < t < 1$
$t-1 < d < t$



It "remain" to integrate out the variable T to get the posterior density of δ_0 given $X_{n_0} \cup Y_{n_1}$.

(ii) Compare the approach with the frequentist paradigm, where we usually test for $H_0 : p_0 = p_1$ (versus some alternative that can be $H_1 : p_0 < p_1$),

or construct a confidence interval for $\hat{\delta} = \hat{p}_0 - \hat{p}_1$, based on large sample approximations, see p. 24 in MS: HYPOTHESIS TESTING.

(iii) Large-sample correspondence.

A multivariate normal distribution holds for the asymptotic density of the posterior distribution of the bivariate vector $\begin{pmatrix} p_0 \\ p_1 \end{pmatrix}$, given $X_{n_0} \cup Y_{n_1}$, as $n_0, n_1 \to \infty$, meaning that this holds true as well for the difference $\delta = p_0 - p_1$. In view of the results page 6 and 7, we see that the posterior distribution for δ is asymptotically the same as the repeated sampling distribution derived from frequentist arguments $\Rightarrow$ A 95% central posterior interval for δ will cover the true value 95% of the time under repeated sampling, for any value of $-1 < \delta < 1$.

↑ As we shall see later, this holds true as well in a wide range of interesting cases. ■

↓ So far, we have derived a full Bayesian inference for a binomial proportion. The general procedure can be summarized as follows:

(a) Specify a joint probability on the parameter θ & observations (prior + model given θ)
(b) Compute the posterior distribution of θ given x_n.
(c) Derive the posterior predictive distribution of a new observation.

The Chapter is organised as follows.

- Section I = Inference for the Normal Distribution, (I.1, I.2)
  Proper and Improper Priors (I.3) (I.4)
- Section II = Jeffreys Prior
- Section III = Bayesian vs Frequentists =
  - ↘ Bernstein-von-Mises Thm (III.1)
  - ↘ Hypothesis Testing (III.2)
  - ↘ Sequential Testing (III.3)
- Section IV = Empirical Bayes
  - ↘ The James-Stein Estimator (IV.1)
  - ↘ Bayesian FDR (IV.3)
- Section V = Hierarchical Models
  & Application to Multiple Testing (V.1)
- References.

I. BAYESIAN INFERENCE FOR THE NORMAL DISTRIBUTION

I.1. unknown mean, known variance.

Let $\mathcal{L}_n = \{X_1, \dots, X_n\}$, iid, $X_i \sim \mathcal{N}(\mu, \sigma^2)$

μ: unknown, with $\mu \sim \mathcal{N}(\mu \mid \mu_0, \tau_0^2)$; σ^2: known

The conjugate prior for μ, when σ^2 is known.

$\mu \sim \mathcal{N}(\mu \mid \mu_0, \tau_0^2)$: prior distribution

General definition of conjugacy:

$p(\theta \mid \mathcal{L}_n) \in \mathcal{P} \quad \forall \; p(\cdot \mid \theta) \in \mathcal{F}$

- θ: parameter of interest
- $\mathcal{P}$: a family of distribution
- $p(\cdot \mid \theta) \in \mathcal{F}$ = model distrib
- $p(\cdot) \in \mathcal{P}$ = prior

(+)
- analytical results usually available
- simplifies computations
- building blocks for more complicated models

- The posterior distribution of $\mu \mid \mathcal{L}_n$ is

$$f(\mu \mid \mathcal{L}_n) \propto f(\mathcal{L}_n \mid \mu) f(\mu)$$

$$\propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \frac{1}{\tau_0} \exp\left\{-\frac{1}{\tau_0^2}(\mu-\mu_0)^2\right\}$$

$$= \frac{1}{\sigma^n} \frac{1}{\tau_0} \exp\left\{-\frac{1}{2}\left(\underbrace{\frac{1}{\sigma^2}\sum(x_i-\mu)^2 + \frac{1}{\tau_0^2}(\mu-\mu_0)^2}\right)\right\}$$

write this term $\frac{1}{\tau_n^2}(\mu-\mu_n)^2 + C$ (C: cst indpt of μ)

& identify μ_n and τ_n^2.

expand & complete the squares:

$$\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\mu^2 - 2\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau_0^2}\right)\mu + \dots$$

$$= \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\mu^2 - 2\mu\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau_0^2}\right)\Big/\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right) + \dots\right]$$

$$= \frac{1}{\tau_n^2} \qquad = \mu_n$$

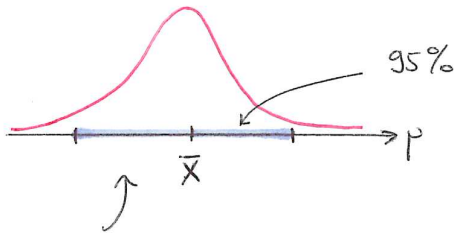
$$\Rightarrow f(\mu \mid \mathcal{L}_n) \sim \mathcal{N}(\mu \mid \mu_n, \tau_n^2), \text{ with}$$

$$\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{x}/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \quad , \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

= weighted average of prior mean & sample mean

= prior precision + data precision.

x Remark: When n is large, $\mu_n \approx \bar{x}$ and $\tau_n^2 \approx \frac{\sigma^2}{n}$, and we get the approximation $\mu \mid \mathcal{L}_n \sim \mathcal{N}(\mu \mid \bar{x}, \frac{\sigma^2}{n})$



95% highest posterior density region is given by $[\bar{x} - z n^{-1/2}\sigma, \bar{x} + z n^{-1/2}\sigma]$, where z = corresponding quantile of the $\mathcal{N}$ distrib.

Compare this interval with the 95% confidence interval for μ → the two coincide when n is large.

- The posterior predictive distribution is

$$f(x \mid \mathcal{L}_n) = \int \mathcal{N}(x \mid \mu, \sigma^2)\, \mathcal{N}(\mu \mid \mu_n, \tau_n^2)\, d\mu$$

x: new sample point; $\mathcal{N}(x \mid \mu, \sigma^2)$: model distribution; $\mathcal{N}(\mu \mid \mu_n, \tau_n^2)$: posterior distribution (see top of p. 6)

convolution of two normal distribution is still normal.

$$\sim \mathcal{N}(x \mid m_n, v_n^2), \quad \begin{cases} \text{(a) } \mu \mid \mathcal{L}_n \sim \mathcal{N}(\mu_n, \tau_n^2) \\ \text{(b) } X \mid \mu, \mathcal{L}_n \sim \mathcal{N}(\mu, \sigma^2) \end{cases}$$

with

$$m_n := \mathbb{E}(X \mid \mathcal{L}_n) = \mathbb{E}_\mu \mathbb{E}(X \mid \mu, \mathcal{L}_n) = \mathbb{E}(\mu \mid \mathcal{L}_n) = \mu_n$$

$$v_n^2 := \mathrm{var}(X \mid \mathcal{L}_n) = \mathbb{E}\underbrace{\mathrm{var}(X \mid \mathcal{L}_n)}_{=\sigma^2} + \underbrace{\mathrm{var}\,\mathbb{E}(X \mid \mathcal{L}_n)}_{=\mathrm{var}\,\mu_n = \tau_n^2}$$

$$\Rightarrow \boxed{f(x \mid \mathcal{L}_n) \sim \mathcal{N}(x \mid \mu_n, \sigma^2 + \tau_n^2)}$$

(posterior predictive distribution of a new point)

I.2. know mean, unknown variance.

This case is rarely useful in practice, but important as a building block for more complicated models.

$$\mathcal{L}_n = \{X_1, \dots, X_n\}, \text{ iid}, \quad X_i \sim \mathcal{N}(\mu, \sigma^2)$$

μ: known; σ^2: unknown, with $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \tau_0^2)$ (scaled inverse χ^2 distrib with ν_0 d.o.f.)

Toolbox

(a) $X \sim \chi^2(\nu_0)$ = Chi-square distribution with ν_0 degrees of freedom, $\nu_0 > 0$

X has density

$$f(x) = \frac{1}{\Gamma(\nu_0/2)} 2^{-\nu_0/2} x^{\nu_0/2-1} e^{-x/2}, \quad x>0$$

& $\mathbb{E}X = \nu_0$; $\text{var}\, X = 2\nu_0$.

(b) If $X \sim \chi^2(\nu_0)$, then $\frac{1}{X} \sim \text{Inv-}\chi^2(\nu_0)$, $\nu_0 > 0$

A variable $\theta \sim \text{Inv-}\chi^2(\nu_0)$ has density

$$f(\theta) = \frac{2^{-\nu_0/2}}{\Gamma(\nu_0/2)} \theta^{-(\nu_0/2+1)} e^{-1/2\theta}, \quad \theta>0$$

& $\mathbb{E}\theta = \frac{1}{\nu_0-2}$ (for $\nu_0>2$) & $\text{var}\,\theta = \frac{2}{(\nu_0-2)^2(\nu_0-4)}$ (for $\nu_0>4$)

(c) $\theta \sim \text{Inv-}\chi^2(\nu_0, \tau_0^2)$ has density

$$f(\theta) = \frac{(\tau_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \theta^{-(\nu_0/2+1)} e^{-\nu_0\tau_0^2/2\theta}, \quad \theta>0$$

$$\mathbb{E}\theta = \frac{\nu_0}{\nu_0-2}\tau_0^2 \quad \& \quad \text{var}\,\theta = \frac{2\nu_0^2\tau_0^4}{(\nu_0-2)^2(\nu_0-4)}$$

If $\theta \sim \text{Inv-}\chi^2(\nu_0, \tau_0^2)$, then

$$\frac{\theta}{\nu_0\tau_0^2} \sim \text{Inv-}\chi^2(\nu_0)$$

- The posterior distribution of σ^2 given $\mathcal{L}_n$ is

$$f(\sigma^2|\mathcal{L}_n) \propto f(\mathcal{L}_n|\sigma^2) f(\sigma^2)$$

$$\propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\} \times \left(\frac{1}{\sigma^2}\right)^{\nu_0/2+1} \exp\left\{-\frac{\nu_0\tau_0^2}{2\sigma^2}\right\}$$

$$= \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+n}{2}+1} \exp\left\{-\frac{\nu_0\tau_0^2 + \sum(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$\sim \text{Inv-}\chi^2\left(\nu_0+n, \frac{\nu_0\tau_0^2 + \sum(x_i-\mu)^2}{\nu_0+n}\right)$$

ν_0+n: prior d.o.f. + data d.o.f.

$$= \left(\frac{\nu_0}{\nu_0+n}\right)\tau_0^2 + \left(\frac{n}{\nu_0+n}\right)\frac{1}{n}\sum(x_i-\mu)^2$$

τ_0^2: prior scale; $\frac{1}{n}\sum(x_i-\mu)^2$: data empirical variance.

↳ The prior on σ^2 behaves as ν_0 observations, with an averaged squared deviation equal to τ_0^2.

I.3. Proper and Improper prior distributions

Recall the two examples from sections I.1 and I.2:

$X_1, \dots, X_n$ iid $\sim \mathcal{N}(\mu, \sigma^2)$

- μ unknown, σ^2 known

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2) \Rightarrow \mu|\mathcal{L}_n \sim \mathcal{N}(\mu_n, \tau_n^2)$$

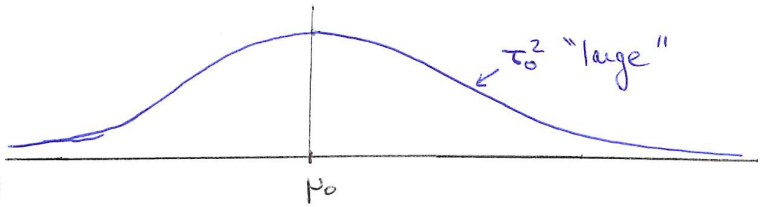
$$\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{x}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}$$

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

When $\frac{n}{\sigma^2} \gg \frac{1}{\tau_0^2}$, then $\mu_n \simeq \bar{x}$, $\tau_n^2 \simeq \frac{n}{\sigma^2}$

($\frac{n}{\sigma^2}$: data precision; $\frac{1}{\tau_0^2}$: prior precision; as if $\tau_0^2 = +\infty$)

This situation arises when the prior is "flat" $\Rightarrow$ little prior knowledge about the most likely values of μ are incorporated into the model.



The limiting case $\tau_0^2 = +\infty$ is not a proper density: it does not integrate to 1. However, provided one data point is observed, the posterior $\mathcal{N}(\bar{x}, n/\sigma^2)$ is a proper distribution.

"flat" constant prior on μ: $f(\mu) \sim 1$

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known

"the likelihood dominates the prior"

The joint distribution is not properly defined, but, provided one data point is observed, the posterior $\mu \mid \mathcal{L}_n$ is properly defined.

- μ known, σ^2 unknown

$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$

$\Rightarrow \sigma^2 \mid \mathcal{L}_n \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0 \sigma_0^2 + \sum (x_i - \mu)^2}{\nu_0 + n}\right)$

ν_0 = prior d.o.f $\equiv$ carries as much information as ν_0 observations

If $n \gg \nu_0$, then $\sigma^2 \mid \mathcal{L}_n \approx \text{Inv-}\chi^2\left(n, n^{-1}\sum (x_i - \mu)^2\right)$

(n: data d.o.f; ν_0: prior d.o.f.)

The limiting case $\nu_0 = 0$ corresponds to the case where the prior contains as much information as 0 observation $\Rightarrow$ uninformative. This limiting case is also obtained by defining an improper prior on σ^2, given by $f(\sigma^2) \sim \frac{1}{\sigma^2}$.

($\sim 1/\sigma^2$: uninformative prior on σ^2 (improper); $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ (μ known)) $\Rightarrow$ joint distribution is not defined, but, posterior $\sigma^2 \mid \mathcal{L}_n$ is proper, provided at least one data point is observed

($\frac{1}{\sigma^2}$ does not integrate to 1)

I.4. unknown mean and variance

Consider the case where $X_1, \dots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We proceed as before:

(a) Compute the posterior distribution $f(\mu, \sigma^2 \mid \mathcal{L}_n)$

(b) Derive the posterior predictive distribution $f(x \mid \mathcal{L}_n)$ of a new observation X.

→ Consider first a non-informative prior on (μ, σ^2):

$f(\mu, \sigma^2) \sim \frac{1}{\sigma^2}$. (we assume a more general conjugate prior later)

→ Posterior is

$$f(\mu, \sigma^2 \mid \mathcal{L}_n) \sim \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right\}$$

$$f(\mu,\sigma^2\mid \mathcal{L}_n) \sim \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}\Big[\underbrace{\sum_{i=1}^n (x_i-\bar{x})^2}_{=:(n-1)s^2} + n(\bar{x}-\mu)^2\Big]\right\} \quad (19)$$

$$= \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}\Big[(n-1)s^2 + n(\bar{x}-\mu)^2\Big]\right\}$$

$\uparrow$ $\qquad$ $\uparrow$ A

We decompose this expression further into the product

$$f(\mu,\sigma^2\mid \mathcal{L}_n) = f(\mu\mid\sigma^2,\mathcal{L}_n)\, f(\sigma^2\mid \mathcal{L}_n)$$

This term (σ^2 fixed) was considered in section I.1

$$f(\mu\mid\sigma^2,\mathcal{L}_n) = \mathcal{N}\left(\bar{x},\frac{\sigma^2}{n}\right)$$

Make use of

$$f(\sigma^2\mid\mathcal{L}_n) = \int f(\mu,\sigma^2\mid\mathcal{L}_n)\,d\mu$$

$$\sim \frac{1}{\sigma^{n+2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \underbrace{\int e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} d\mu}_{=\left(2\pi\frac{\sigma^2}{n}\right)^{1/2}}$$

$$\Downarrow$$

$$f(\sigma^2\mid\mathcal{L}_n) \sim \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}$$

$$= \text{Inv-}\chi^2(n-1,s^2)$$

We get

$$\boxed{\begin{aligned} f(\mu,\sigma^2\mid\mathcal{L}_n) &= \mathcal{N}\left(\mu\mid\bar{x},\frac{\sigma^2}{n}\right)\text{Inv-}\chi^2(\sigma^2\mid n-1,s^2) \\ &= f(\mu\mid\sigma^2,\mathcal{L}_n)\; f(\sigma^2\mid\mathcal{L}_n) \end{aligned}}$$

Next, we derive the posterior predictive distribution of a new observation

$$f(x\mid\mathcal{L}_n) = \int f(x,\mu,\sigma^2\mid\mathcal{L}_n)\,d\mu\,d\sigma^2$$

$$f(x\mid\mathcal{L}_n) = \int \underbrace{f(x\mid\mu,\sigma^2,\mathcal{L}_n)}_{\text{model distribution } \mathcal{N}(x\mid\mu,\sigma^2)}\; \underbrace{f(\mu\mid\sigma^2,\mathcal{L}_n)\, f(\sigma^2\mid\mathcal{L}_n)}_{\text{derived on page 19 (posterior distribution)}} \quad (20)$$

$$d\mu\,d\sigma^2$$

$$= \int \mathcal{N}(x\mid\mu,\sigma^2)\,\mathcal{N}\left(\mu\mid\bar{x},\frac{\sigma^2}{n}\right)\text{Inv-}\chi^2(\sigma^2\mid n-1,s^2)\,d\mu\,d\sigma^2$$

$$= \int \left[\mathcal{N}(x\mid\mu,\sigma^2)\,\mathcal{N}\left(\mu\mid\bar{x},\frac{\sigma^2}{n}\right)d\mu\right]\text{Inv-}\chi^2\,d\sigma^2$$

see top of p.14 ↓

$$= \int \mathcal{N}\left(x\mid\bar{x},\left(1+\frac{1}{n}\right)\sigma^2\right)\text{Inv-}\chi^2(\sigma^2\mid n-1,s^2)\,d\sigma^2$$

Note that $\sigma^2 \sim \text{Inv-}\chi^2(\sigma^2\mid n-1,s^2)$ (bottom of p. 15)

$$\Leftrightarrow \frac{\sigma^2}{(n-1)s^2} \sim \text{Inv-}\chi^2(\sigma^2\mid n-1) \quad \text{(p. 15, (b))}$$

$$\Leftrightarrow \frac{(n-1)s^2}{\sigma^2} \sim \chi^2\left(\frac{1}{\sigma^2}\mid n-1\right) \quad \text{(with } \lambda := \tfrac{1}{\sigma^2}\text{)}$$

$$\Leftrightarrow (n-1)s^2\lambda \sim \gamma\left(\lambda\mid\tfrac{n-1}{2},\tfrac{1}{2}\right) \quad \text{(scaling (use MGF))}$$

$$\Leftrightarrow \lambda \sim \gamma\left(\lambda\mid\tfrac{n-1}{2},\tfrac{(n-1)s^2}{2}\right) \quad \text{(} \chi^2 \text{ \& } \gamma \text{ distributions)}$$

$$\Rightarrow f(x\mid\mathcal{L}_n) = \int \mathcal{N}\left(x\mid\bar{x},\left(1+\tfrac{1}{n}\right)\lambda^{-1}\right)\gamma\left(\lambda\mid\tfrac{n-1}{2},\tfrac{(n-1)s^2}{2}\right)d\lambda$$

$$= t\left(n-1,\ \bar{x},\ s^2\left(1+\tfrac{1}{n}\right)\right),$$

d.o.f. $\uparrow$ $\qquad$ location $\uparrow$ $\qquad$ scale is $s\left(1+\frac{1}{n}\right)^{1/2}$ $\nwarrow$

see p.11/12 in SL: BAYESIAN LINEAR MODELS for details.

$$\left[X\sim t(\nu,\mu,\sigma^2) \text{ has density } \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sigma\sqrt{\nu\pi}}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}\right]$$

• Summary : To draw samples $X \sim$ posterior predictive distribution, having observed $\mathcal{L}_n$, we can proceed in two equivalent ways :

(1) Draw $\sigma^2 \mid \mathcal{L}_n \sim \text{Inv-}\chi^2(\sigma^2 \mid n-1, s^2)$

(2) Draw $\mu \mid \sigma^2, \mathcal{L}_n \sim \mathcal{N}(\mu \mid \bar{x}, \frac{\sigma^2}{n})$

(3) Draw $X \mid \mu, \sigma^2, \mathcal{L}_n \sim \mathcal{N}(x \mid \mu, \sigma^2)$

$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

or directly using a scaled & shifted t distribution

(i) Draw $X \mid \mathcal{L}_n \sim t(n-1, \bar{x}, s^2(1+\frac{1}{n}))$

↑ $x_1, \dots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ + non-informative prior on (μ, σ^2)

Remarks (i) The marginal posterior distribution of μ given $\mathcal{L}_n$ can be analytically computed :

$$f(\mu \mid \mathcal{L}_n) \sim A^{-n/2} \underbrace{\int_0^{+\infty} u^{\frac{n-2}{2}} e^{-u} du}_{\text{unnormalized Gamma integral}}$$

A defined on top of page 19 & $u := A/2\sigma^2$

$$\sim \left[(n-1)s^2 + n(\mu - \bar{x})^2 \right]^{-n/2}$$

$$\sim \left[1 + \frac{(\mu - \bar{x})^2}{(n-1)(s^2/n)} \right]^{-\frac{n}{2}}$$

$$\Rightarrow f(\mu \mid \mathcal{L}_n) = t(n-1, \bar{x}, \frac{s^2}{n})$$

$$\Leftrightarrow \frac{\mu - \bar{x}}{s/\sqrt{n}} \mid \mathcal{L}_n \sim t_{n-1}$$

frequentist & Bayesian intervals are equal

& compare with the frequentist result : $\frac{\bar{x} - \mu}{s/\sqrt{n}} \mid \mu, \sigma^2 \sim t_{n-1}$

compare also with the result on p. 13, with μ unknown & σ^2 known

(ii) We may assume more generally a conjugate prior for (μ, σ^2) (the non-informative case can be recovered here as well as a limit).

Motivated by the factorization of the posterior on p. 19, $f(\mu, \sigma^2 \mid \mathcal{L}_n) = f(\mu \mid \sigma^2, \mathcal{L}_n) f(\sigma^2 \mid \mathcal{L}_n)$

($\mathcal{N}$, Inv-χ^2)

a similar factorization must hold for the prior $f(\mu, \sigma^2) = f(\mu \mid \sigma^2) f(\sigma^2)$

($\mathcal{N}$, Inv-χ^2) the conjugate prior.

It is convenient to parametrize the conjugate prior as follows =

$$\mu \mid \sigma^2 \sim \mathcal{N}(\mu \mid \mu_0, \frac{\sigma^2}{\kappa_0})$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\sigma^2 \mid \nu_0, \sigma_0^2)$$

Then it can be shown after calculations that

$$f(\mu, \sigma^2 \mid \mathcal{L}_n) = f(\mu \mid \sigma^2, \mathcal{L}_n) f(\sigma^2 \mid \mathcal{L}_n)$$

$$= \mathcal{N}(\mu \mid \mu_n, \frac{\sigma^2}{\kappa_n}) \text{Inv-}\chi^2(\sigma^2 \mid \nu_n, \sigma_n^2)$$

where

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n} \mu_0 + \frac{n}{\kappa_0 + n} \bar{x}$$

interpretation ?

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{x} - \mu_0)^2$$

The marginal posterior distributions can be computed as well,

$$f(\mu \mid \mathcal{L}_n) = t\left(\nu_n, \mu_n, \frac{\sigma_n^2}{\kappa_n}\right) \quad \text{(i)}$$

$$f(\sigma^2 \mid \mathcal{L}_n) = \text{Inv-}\chi^2(\nu_n, \sigma_n^2) \quad \text{(ii)}$$

We get from (i) that $\dfrac{\mu - \mu_n}{\sigma_n / \sqrt{\kappa_n}} \sim t_{\nu_n}$ (compare with the expression p.21)

Summary = inference for μ.

	μ unknown, σ^2 known		μ unknown, σ^2 unknown	
	non-informative prior	conjugate prior	non-informative prior	conjugate prior
BAYESIAN	$f(\mu) \sim 1$	$\sim \mathcal{N}(\mu \mid \mu_0, \tau_0^2)$ ↖ p.12	$f(\mu, \sigma^2) \sim \frac{1}{\sigma^2}$ ↖ p.18	$\sim \mathcal{N}(\mu \mid \mu_0, \frac{\sigma^2}{\kappa_0}) \times \text{Inv}\chi^2(\sigma^2 \mid \nu_0, \sigma_0^2)$ ↖ p.22
	marginal posterior $\mu \mid \mathcal{L}_n \sim \mathcal{N}(\mu \mid \bar{x}, \frac{\sigma^2}{n})$	↙ p.13 $\mu \mid \mathcal{L}_n \sim \mathcal{N}(\mu \mid \mu_n, \tau_n^2)$; n large: $\mu_n \approx \bar{x}$, $\tau_n^2 \approx \frac{\sigma^2}{n}$	↙ p.20 $\mu \mid \mathcal{L}_n \sim t(n-1, \bar{x}, \frac{s^2}{n})$	↙ p.23 $\mu \mid \mathcal{L}_n \sim t(\nu_n, \mu_n, \frac{\sigma_n^2}{\kappa_n})$; n large: $\mu_n \approx \bar{x}$, $\sigma_n^2 \approx s^2$, $\kappa_n \approx n$
	Credible bounds = Confidence bands	↑ Same when n is large	Credible bounds = Confidence bands	↑ Same when n is large
FREQUENTIST	$\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$; $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$		$\bar{x} \sim t(n-1, \mu, \frac{s^2}{n})$; $\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$	

II. JEFFREYS PRIOR

We often want to consider a prior that carries as little information as possible. When the parameter space is discrete and contains exactly m distinct values, then we just put a mass $1/m$ at each parameter value. When the parameter space is continuous however, there are several ways of achieving this. We discussed in Section I.3 uniform priors for the normal distribution, taken as limits of conjugate priors:

$\mathcal{N}(\mu, \tau^2)$, μ & τ^2 unknown $\rightarrow (\mu, \sigma^2) \sim \frac{1}{\sigma^2}$

Similarly, a uniform prior for a binomial proportion p is $\mathcal{U}(0,1)$.

↖ flat over the parameter space $[0,1]$, which is compact $\Rightarrow$ proper prior

Although a uniform distribution carries no information about p itself, $\sqrt{p}$ has a higher density near 1 than 0.

↳ ignorance about p leads to knowledge about $\sqrt{p}$.

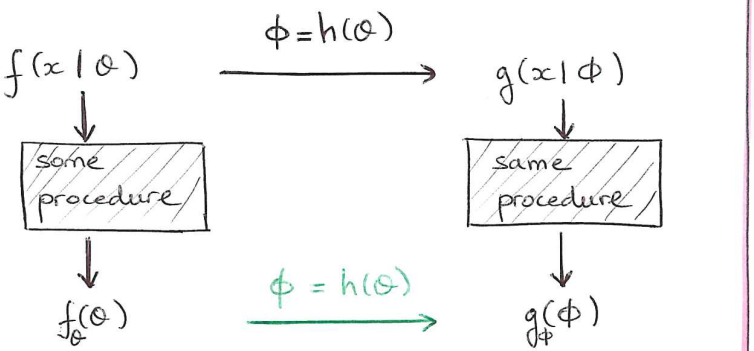
Likewise, under the assumption that $p \sim \mathcal{U}[0,1]$, the transformation $\phi := \log\left(\frac{p}{1-p}\right)$ (log odds) has density $f_\phi(\phi) = \dfrac{e^\phi}{(1+e^\phi)^2}$ ← not a flat prior over $\mathbb{R}$: some values of ϕ are more likely than others.

This was considered a major drawback in the early days

of Bayesian statistics. The uniform prior cannot be accepted as a "universal" non-informative prior, since it carries information under monotonic transformations of the "canonical" parameter.

Let θ denote the original parameter, and consider $\phi := h(\theta)$, for some 1-1 monotone mapping h. The goal is to find a procedure / principle that produces a prior $f_\theta(\theta)$ for any input model density $f(x|\theta)$ parametrized by θ.

(*)



Flat priors (proper & improper) fail to satisfy (*).

Indeed, taking ~~some procedure~~ $\equiv$ ~~flat prior~~, and applying it to the binomial distribution parametrized by $\theta = p$ and $\phi = \log\left(\frac{p}{1-p}\right)$, we see that we should get $f_\theta(\theta) = \mathcal{U}[0,1]$ and $g(\phi) \sim 1$,

while we saw on page 24 that transforming $f_\theta(\theta)$ yields the prior $\frac{e^\phi}{(1+e^\phi)^2}$. A valid universal procedure must therefore produce the same prior $g_\phi(\phi)$ whether (a) we first transform $\phi = h(\theta)$ & apply the procedure or (b) apply the procedure to $f(x|\theta)$ and then consider the change of variable $\phi = h(\theta)$.

- Note that for $\theta \sim f_\theta(\theta)$ and the change of variable $\phi = h(\theta)$ $[\theta = h^{-1}(\phi)]$, the density of ϕ is given by

$$g_\phi(\phi) = f_\theta(\theta)\left|\frac{d\theta}{d\phi}\right| = f_\theta(h^{-1}(\phi))\left|\frac{d\,h^{-1}(\phi)}{d\phi}\right|.$$

- Likewise, for $\phi \sim g_\phi(\phi)$ and the change of variable $\theta = h^{-1}(\phi)$ $[\phi = h(\theta)]$, the density of θ is given by

$$f_\theta(\theta) = g_\phi(\phi)\left|\frac{d\phi}{d\theta}\right| = g_\phi(h(\theta))\left|\frac{d\,h(\theta)}{d\theta}\right|$$

Harold Jeffreys (1946) proposed $f_\theta(\theta) \sim \sqrt{I_\theta(\theta)}$ as the default "objective" choice of prior distribution.

Fisher information

Recall that Fisher information is given by

$$I_\theta(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right]$$

under regularity conditions

$$= -\mathbb{E}\left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta}\right]$$

(cf MS: MAXIMUM LIKELIHOOD ESTIMATION).

We show that Jeffreys prior is invariant under smooth monotone transformations of the input parameter.

We have $\boxed{I_\theta(\theta) = |h'(\theta)|^2 \, I_\phi(h(\theta))}$ —— (**)

Indeed, $I_\theta(\theta) = \mathbb{E}\left(\frac{\partial}{\partial \theta} f(X|\theta)\right)^2$

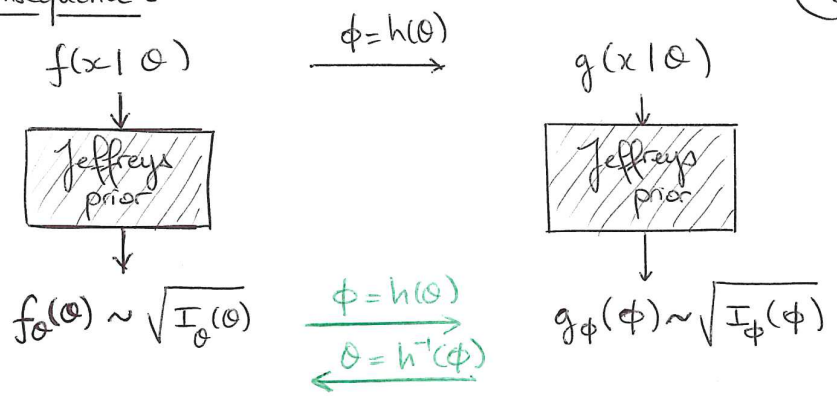
$= \mathbb{E}\left(\frac{\partial}{\partial \theta} g(X|h(\theta))\right)^2$

$= \mathbb{E}\left(\frac{\partial}{\partial \phi} g(X|h(\theta)) \frac{\partial \phi}{\partial \theta}\right)^2$

$= \left(\frac{\partial \phi}{\partial \theta}\right)^2 \mathbb{E}\left(\frac{\partial}{\partial \phi} g(X|\phi)\right)$

$= |h'(\theta)|^2 \, I_\phi(\phi)$ ∎

× Consequence =



Indeed, starting with $g_\phi(\phi) \sim \sqrt{I_\phi(\phi)}$ and considering the change of variable $\theta = h^{-1}(\phi)$ yields the density

$$g_\phi(\phi) \left|\frac{d\phi}{d\theta}\right| \sim \sqrt{I_\phi(\phi)} \left|\frac{d\phi}{d\theta}\right| = \sqrt{I_\theta(\theta)} \sim f_\theta(\theta),$$

as required.

(**) p. 27

× Remarks (i) If $\int \sqrt{I_\theta(\theta)}\, d\theta = c_\theta < +\infty$, Jeffreys prior is proper and given by $f_\theta(\theta) = \frac{1}{c_\theta}\sqrt{I_\theta(\theta)}$.

If the integral is infinite, the prior is improper. This is not an issue as long as the posterior distribution is a proper density.

(ii) Binomial distribution.

• $\theta = p \longrightarrow f(x|\theta) = \theta^x (1-\theta)^{1-x} \quad x \in \{0,1\}$

- $\phi = h(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$

$\hookrightarrow g(x|\phi) = f(x|h^{-1}(\phi))$

$= \left(\frac{e^{\phi}}{1+e^{\phi}}\right)^{x}\left(\frac{1}{1+e^{\phi}}\right)^{1-x}$

$= \frac{e^{\phi x}}{(1+e^{\phi})^{2}}$.

$h'(u) = \frac{1}{u(1-u)}$

$h^{-1}(u) = \frac{e^{u}}{1+e^{u}}$

- Then $\log f(x|\theta) = x\log\theta + (1-x)\log(1-\theta)$

& $\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$

$\Rightarrow I_{\theta}(\theta) = \frac{1}{\theta(1-\theta)}$ & Jeffreys prior is $\sim \frac{1}{\sqrt{\theta(1-\theta)}}$

$\equiv B(1/2, 1/2)$.

- Also, $\log g(x|\phi) = \phi x - \log(1+e^{\phi})$

& $\frac{\partial^2}{\partial\phi^2}\log g(x|\phi) = -\frac{e^{\phi}}{(1+e^{\phi})^{2}}$

$\Rightarrow I_{\phi}(\phi) = \frac{e^{\phi}}{(1+e^{\phi})^{2}}$ $\Rightarrow$ Jeffreys prior is $\sim \frac{e^{\phi/2}}{1+e^{\phi}}$.

- Summarizing, $f_{\theta}(\theta) \sim \frac{1}{\sqrt{\theta(1-\theta)}}$ and $g_{\phi}(\phi) \sim \frac{e^{\phi/2}}{1+e^{\phi}}$.

$\theta = h^{-1}(\phi) = \frac{e^{\phi}}{1+e^{\phi}}$

The change of variable $\theta = h^{-1}(\phi)$ yields the density $g_{\phi}(h(\theta))\left|\frac{dh(\theta)}{d\theta}\right| \sim \frac{e^{h(\theta)/2}}{1+e^{h(\theta)}} \times \frac{1}{\theta(1-\theta)}$

$= \frac{1}{\sqrt{\theta(1-\theta)}}$, as required ∎

(iii) $N(\mu, \sigma^2)$, μ unknown, σ^2 known.

Then $I_{\mu}(\mu) = \frac{1}{\sigma^2}$ so Jeffreys prior is ~ 1 (improper)

(iv) Multivariate case: $f_{\theta}(\theta) \sim \sqrt{\det(I_{\theta}(\theta))}$

Ex: $N(\mu, \sigma^2)$, μ and σ^2 unknown

Then one can show that $I_{\theta}(\theta) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}$

$\theta = (\mu, \sigma^2)$

So that $\det(I_{\theta}(\theta)) = \frac{1}{2\sigma^6}$

Jeffreys prior is $f(\mu, \sigma^2) \sim \frac{1}{\sigma^3}$

differs from the non-informative prior $\frac{1}{\sigma^2}$ encountered before.

(v) Bernardo (1979) showed that Jeffreys prior maximizes the (average) KL divergence between the prior and the posterior: the one that maximizes the contribution of the data and minimizes the contribution of the prior.

III - BAYESIAN vs FREQUENTIST

III.1. Bernstein - Von Mises theorem

- We establish in this section some connections between posterior-based inference and the frequentist approach.
- Assume that $\mathcal{X}_n = \{X_1, \dots, X_n\}$ are independent observations sampled from a common fixed distribution $\mathbb{P}_0$, with density f_0.

↖ A common assumption in frequentist statistics. The posterior distribution will be interpreted under this framework.

- The data are modeled using the parametric family $f(x|\theta)$, $\theta \in \Theta$ and prior $f(\theta)$. The posterior distribution is denoted $f(\theta|\mathcal{X}_n)$. Put

$$\Pi(B|\mathcal{X}_n) := \int_B f(\theta|\mathcal{X}_n)\, d\theta$$

↖ Posterior Probability of B

↖ Borel set

(*) Assume that the parametric family $f(\cdot|\theta)$ includes the "true" density f_0; i.e.
$\exists \theta_0 \in \Theta$ s.t. $f_0(x) = f(x|\theta_0)$

↖ We relax this assumption later, and consider the misspecified case where there is no such θ_0.

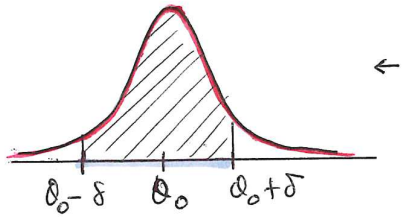
Under (*) and some regularity conditions (given more explicitely below for a more general result), CONSISTENCY follows:

As the number of observations tends to $+\infty$, the posterior distribution collapses to a point mass at θ_0.

$\forall \delta > 0 \quad \Pi(\{\theta : \|\theta - \theta_0\| > \delta\} | \mathcal{X}_n) \to 0$

as $n \to +\infty$, in probability (under $\mathbb{P}_0$).

↖ No that consistency holds in probability under the true $\mathbb{P}_0$, hence the need to introduce it. In fact, it is otherwise possible for most interesting problems to construct a sequence $x_1, x_2, \dots$ that violates the convergence to 0.



← We can put as much mass as possible in any neighborhood of θ_0, by taking n large enough. Under some regularity conditions, we can establish that the posterior distribution approaches normality as $n \to \infty$. This result is known as the Bernstein - Von Mises theorem.

[Reg Conditions]

(1) $X_1, \dots, X_n$ iid $\sim \mathbb{P}_0$. (otherwise the data may not bring enough information about the parameter of interest)

(2) The prior puts a positive mass around θ_0. (otherwise the likelihood has no chance to dominate the prior)

(3) Smoothness (twice continuously differentiable) & identifiability of the log-likelihood. (otherwise there is not a single point θ_0 the posterior can converge to

(4) θ_0 is an interior point of Θ. (otherwise the normal distribution may not be appropriate)

[BERNSTEIN - VON-MISES THM]

Under the regularity conditions stated above,

$$\sup_B \left| \Pi(B \mid \mathcal{L}_n) - \mathcal{N}\left(\hat{\theta}_n, [n I_1(\theta_0)]^{-1}\right) \right| \longrightarrow 0$$

(TV distance)

in probability (under $\mathbb{P}_0$), as $n \to \infty$.

($\hat{\theta}_n$: MLE; $I_1(\theta_0)$: Fisher Information of 1 observation)

↖ A very powerful result. It shows that for any measurable set B, the full posterior distribution is close to the normal distribution. It has huge implications when interpreting credible intervals constructed from the posterior density. Indeed, under regularity conditions, the MLE $\hat{\theta}_n$ is consistent ($\hat{\theta}_n \to \theta_0$ in probability) and asymptotically normal $[n I_1(\theta_0)]^{1/2} (\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1)$

↖ Same covariance matrix as in the B-VM theorem.

"Frequentist" confidence intervals constructed from the asymptotic normal distribution of the MLE take the form

$$C_n = \hat{\theta}_n \pm z_{1-\alpha/2}\, n^{-1/2}\, I_1(\theta_0)^{-1/2}$$

← "Wald" interval

↑ Has approximate coverage $(1-\alpha)$ i.e. $\mathbb{P}_0(\theta_0 \in C_n) \approx 1-\alpha$

↖ Usually estimated by replacing θ_0 with $\hat{\theta}_n$.

Let B_n be such that $\Pi(B_n \mid \mathcal{L}_n) = 1-\alpha$. The Bernstein-Von-Mises theorem ensures that $B_n = C_n + o_{\mathbb{P}_0}(1)$.

⇒ Under regularity assumptions, Bayesian credible sets are valid confidence sets in the frequentist sense: central sets of posterior probability $(1-\alpha)$ cover the parameter at confidence level $(1-\alpha)$.

× Ex: $X_1, \dots, X_n \sim \mathcal{B}(\theta_0)$.

Then $I_1(\theta_0) = \dfrac{1}{\theta_0(1-\theta_0)}$ (see page 29) ⇒ $\hat{\theta}_n = \overline{X}$

Frequentist interval is $\overline{X} \pm z_{1-\frac{\alpha}{2}}\, n^{-1/2} [\theta_0(1-\theta_0)]^{-1/2}$

(replace with $\overline{X}$)

With a conjugate Beta prior $\mathcal{B}(\alpha, \beta)$, the posterior distrib. is $\mathcal{B}(\alpha + S_n, n + \beta - S_n)$ (see page 4), with mean $\approx S_n/n$ and variance $\approx n^{-1}\overline{X}(1-\overline{X})$ when n is large. A central interval of posterior probability $(1-\alpha)$ therefore is of the same form as the frequentist interval. ∎

• Sketch of proof = We provide here some heuristics explaining why the posterior distribution converges to the normal distribution as the sample size increases. A Taylor series expansion of the log posterior $\log f(\theta|\mathcal{L}_n)$, centered at the posterior mode yields:

$$\log f(\theta|\mathcal{L}_n) = \log f(\hat\theta|\mathcal{L}_n) + \frac{1}{2}(\theta-\hat\theta)^t \left.\frac{\partial^2 \log f(\theta|\mathcal{L}_n)}{\partial\theta^2}\right|_{\theta=\hat\theta} (\theta-\hat\theta) + \dots$$

$$f(\theta|\mathcal{L}_n) \propto f(\theta) f(\mathcal{L}_n|\theta)$$

$$= \left.\frac{\partial^2 \log f(\theta)}{\partial\theta^2}\right|_{\theta=\hat\theta} + \sum_{i=1}^{n} \left.\frac{\partial^2 \log f(\mathcal{L}_n|\theta)}{\partial\theta^2}\right|_{\theta=\hat\theta}.$$

↑ constant as a function of θ

↑ dominating term since the sum of n quantities

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(\mathcal{L}_n|\theta)}{\partial\theta^2} \quad \text{for some } \theta\in\Theta$$ = empirical Fisher information

$$\xrightarrow{\text{SLLN}} \mathbb{E}_{P_0}\left[\frac{\partial^2 \log f(X|\theta)}{\partial\theta^2}\right] = -I_1(\theta)$$

$$\Rightarrow \frac{1}{n}\log f(\theta|\mathcal{L}_n) \approx \text{constant} - \frac{1}{2}(\theta-\hat\theta)^t [I_1(\hat\theta)]^{-1}(\theta-\hat\theta) + \underbrace{\text{something that goes to 0.}}_{\text{quadratic in } (\theta-\hat\theta)}$$

We are being sloppy. Not a rigorous proof.

$$\Rightarrow f(\theta|\mathcal{L}_n) \approx \mathcal{N}(\theta|\,\hat\theta, [n I_1(\hat\theta)]^{-1}), \text{ as required.}$$

Under regularity conditions, the posterior mode $\hat\theta$ is consistent for θ_0, and so is the MLE $\hat\theta_n$. This ensures that $f(\theta|\mathcal{L}_n) \approx \mathcal{N}(\theta|\,\hat\theta_n, [n I_1(\theta_0)]^{-1})$, as stated in the Bernstein-Von-Mises theorem.

× Counterexamples =

(a) Let $X_1,\dots,X_n \sim \mathcal{N}(0,1)$ (true distribution)

Data are modeled with $\mathcal{N}(\theta,1)$, with $\theta\geq 0$ constrained to be positive & non-informative prior on θ. The true parameter thus lies on the edge of the parameter space $\Theta = [0,+\infty)$.

The posterior distribution for θ is normal (p. 23), centered at $\bar X$, and truncated to be positive. As $n\to\infty$, it converges to half of a normal distribution centered around 0 and truncated to be positive.

(b) No convergence to a single point. (= underidentified model)

Consider the mixture distribution

$$f(x|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\lambda) \leftarrow \theta=(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\lambda)\in\mathbb{R}^5$$

$$= \lambda \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right) + (1-\lambda)\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right).$$

↑ The likelihood remains the same if we interchange (μ_1,μ_2), (σ_1^2,σ_2^2) and $(\lambda, 1-\lambda)$ so the posterior has at least two modes and cannot converge to a single point. A way around this is to impose constraints on the parameters.

x Bernstein-Von-Mises theorem under misspecification.

Kleijn & van der Vaart (2012)

The case where the true generating distribution $\mathbb{P}_0$ is not a member of the parametric family $f(\cdot|\theta)$ was considered by Kleijn and van der Vaart, extending the original Bernstein-Von-Mises theorem. In the authors' words, « this misspecified version of the Bernstein-Von-Mises theorem has important consequence for the interpretation of Bayesian credible sets. In the misspecified situation the posterior distribution of a parameter shrinks to the point within the model at minimum Kullback-Leibler divergence to the true distribution, a consistency property that it shares with the maximum likelihood estimator. Consequently one can consider both the Bayesian procedure and the MLE as estimates of this minimum KL point. A confidence region for this minimum KL point can be built around the MLE based on its asymptotic normal distribution, involving the sandwich covariance. One might also hope that a Bayesian credible set automatically yields a valid confidence set for the minimum KL point. However, the misspecified Bernstein-Von-Mises theorem shows the latter to be false. »

Let $\theta^* = \underset{\theta}{\operatorname{argmin}}\, KL(\mathbb{P}_0 \| f(\cdot|\theta)) = \underset{\theta}{\operatorname{argmin}} \int \mathbb{P}_0(x) \log \frac{\mathbb{P}_0(x)}{f(x|\theta)}\, dx$

Then, under some regularity conditions,

$$\sup_B \left| \Pi(B|\mathcal{L}_n) - \mathcal{N}(\hat{\theta}_n, [n I_1(\theta^*)]^{-1})(B) \right| \to 0$$

in probability (w.r.t. $\mathbb{P}_0$), as $n \to \infty$, where

$$I_1(\theta) = - \mathbb{E}_{\mathbb{P}_0} \left\{ \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \right\}.$$

However, the MLE $\hat{\theta}_n$ satisfies (van der Vaart (1998)):

$$n^{1/2}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma(\theta^*))$$

with $\Sigma(\theta) = I_1^{-1}(\theta)\, W_1(\theta)\, I_1^{-1}(\theta)$ [aka the sandwich cov.]

$$W_1(\theta) = \mathbb{E}_{\mathbb{P}_0} \left\{ \left(\frac{\partial \log f(X|\theta)}{\partial \theta} \right)^2 \right\}.$$

⇒ The asymptotic covariance matrix of the posterior distribution and of the MLE differ in the misspecified case.

— Bayesian credible sets are not valid confidence sets when the model is misspecified —

- Example 1 (example 2.1 in Kleijn & van der Vaart (2012))

Let $X_1, \dots, X_n$ iid $\sim \mathbb{P}_0 = \mathcal{N}(0, \sigma^2)$ = true distribution

Data are modeled using $f(\cdot|\theta) = \mathcal{N}(\cdot|\theta, 1)$, where θ is the parameter of interest.

Note that $KL(\varphi_0 \| f(\cdot|\theta)) =$

$$\text{constant} - \frac{1}{2}\int \varphi_0(x)(x-\theta)^2\,dx,$$

which is minimized at $\theta^* = \mathbb{E}_{P_0} X = 0$.

In addition, $\log f(x|\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}(x-\theta)^2$

$$\frac{\partial \log f(x|\theta)}{\partial \theta} = x - \theta$$

$$\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} = -1$$

We get that $\cdot\ I_1(\theta^*) = 1$

$\cdot\ W_1(\theta^*) = \mathbb{E}_{P_0} X^2 = \sigma^2$,

so that $\Sigma(\theta^*) = I_1^{-1}(\theta^*)\, W_1(\theta^*)\, I_1^{-1}(\theta^*) = \sigma^2$.

and the MLE satisfies $\boxed{n^{1/2}\,\hat{\theta}_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)}$.

$\Rightarrow$ The $(1-\alpha)$ confidence set around the mean has radius $z_{1-\frac{\alpha}{2}}\,\sigma\, n^{-1/2}$

On the other hand, according to the misspecified Bernstein-Von-Mises theorem, the posterior distribution yields an $(1-\alpha)$ Bayesian credible set with radius $z_{1-\frac{\alpha}{2}}\, n^{-1/2}$, since $f(\cdot\,|\,\mathcal{L}_n) \approx \mathcal{N}(\hat{\theta}_n, 1)$. The Bayesian credible set thus has a frequentist coverage that can be arbitrarily close to 0 or 1 (depending on $\sigma^2 < 1$ or $\sigma^2 > 1$). ∎

• Example 2 : taken from White (1982), Maximum Likelihood Estimation of Misspecified Models, Econometrica, vol 50, n° 1, pp 1-25.

↘ Let $X_1, \dots, X_n$ iid $\sim P_0$ with density φ_0 (true distribution)

Put $\mu_0 = \mathbb{E}_{P_0} X$ $\qquad \sqrt{\beta} = \dfrac{\mathbb{E}_{P_0}(X-\mu_0)^3}{\sigma_0^3}$ (skewness)

$\sigma_0^2 = \mathrm{Var}\, X$ $\qquad \gamma = \dfrac{1}{\sigma_0^4}\,\mathbb{E}_{P_0}(X-\mu_0)^4$ (kurtosis)

↘ Data are modeled using the normal distribution $f(\cdot|\theta) = \mathcal{N}(\cdot\,|\,\mu, \sigma^2)$ and the goal is to estimate μ and σ^2. ↖ $\theta = (\mu, \sigma^2)$

↘ It is straightforward to see that

$$\theta^* = (\mu^*, (\sigma^*)^2) = \underset{(\mu,\sigma^2)}{\operatorname{argmin}}\ KL(\varphi_0 \| f(\cdot|\theta)) = (\mu_0, \sigma_0^2)$$

Indeed,

$$KL(\varphi_0 \| f(\cdot|\theta)) = \int \varphi_0(x) \log \frac{\varphi_0(x)}{f(x|\theta)}\,dx$$

$$= \text{constant} - \log\sigma - \frac{1}{2\sigma^2}\underbrace{\int \varphi_0(x)(x-\mu)^2\,dx}_{=}$$

$$\mathbb{E}_{P_0} X^2 - 2\mu\,\mathbb{E}_{P_0} X + \mu^2$$

minimized at $\mu^* = \mu_0\ (= \mathbb{E}_{P_0} X)$.

At $\mu_* = \mu_0$, $KL(\mathbb{P}_0 \| f(\cdot|\theta)) = cst - \log\sigma - \frac{\sigma_0^2}{2\sigma^2}$,

which is minimized at $\sigma^* = \sigma_0$.

With $\log f(X|\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2$

- $\frac{\partial}{\partial\mu}\log f(X|\theta) = \frac{1}{\sigma^2}(X-\mu)$
- $\frac{\partial}{\partial\sigma^2}\log f(X|\theta) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X-\mu)^2$
- $\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f(X|\theta) = -\frac{X-\mu}{\sigma^4}$
- $\frac{\partial^2}{\partial\mu^2}\log f(X|\theta) = -\frac{1}{\sigma^2}$
- $\frac{\partial^2}{\partial(\sigma^2)^2}\log f(X|\theta) = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(X-\mu)^2$,

we get

$$\rightarrow \quad I_1(\theta^*) = \begin{bmatrix} \frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{1}{2\sigma_0^4} \end{bmatrix}$$

$$\rightarrow \quad W_1(\theta^*) = \begin{bmatrix} 1/\sigma_0^2 & \sqrt{\beta}/2\sigma_0^3 \\ \sqrt{\beta}/2\sigma_0^3 & (\gamma-1)/4\sigma_0^4 \end{bmatrix}$$

$$\rightarrow \quad \Sigma(\theta^*) := I_1^{-1}(\theta^*) W_1(\theta^*) I_1^{-1}(\theta^*) = \begin{bmatrix} \sigma_0^2 & \sqrt{\beta}\sigma_0^3 \\ \sqrt{\beta}\sigma_0^3 & (\gamma-1)\sigma_0^4 \end{bmatrix}$$

Unless $\beta=0$ & $\gamma=3$ (ie. $\mathbb{P}_0 = \mathcal{N}$), $\Sigma(\theta^*) \neq I_1^{-1}(\theta^*)$ indeed.

III.2. Hypothesis testing

- Fisher vs Neyman vs Jeffreys

Suppose that $X \sim f(x|\theta)$, and we wish to test for $\theta = \theta_0$. Fisher, Neyman and Jeffreys would all disagree on how to proceed

Fisher's significance testing =

$H_0 : \theta = \theta_0$

Select a statistic $T(X)$ so that large values of T reflect evidence against H_0

Compute the p-value for the observed data x

$p = \mathbb{P}_{H_0}(T(X) \geq T(x))$

and reject H_0 if p is small, e.g. $p \leq 0.05$

Neyman-Pearson hypothesis testing

Test $H_0 : \theta = \theta_0$ vs an alternative $H_1 : \theta = \theta_1$.

Reject H_0 if $T > c$, for a pre-selected critical value c

Compute $\alpha = \mathbb{P}_{H_0}(\text{rejecting } H_0)$

$\beta = \mathbb{P}_{H_1}(\text{"accepting" } H_0)$

Jeffrey's approach

Define Bayes Factor

$B(x) = \frac{f(x|H_0)}{f(x|H_1)}$

Reject H_0 if $B(x) < 1$, otherwise "accept" it.

Repat the posterior proba

$P(H_0|x) = \frac{B(x)}{1+B(x)}$

Fisher's p-values are often poorly interpreted. A p-value less than 0.05 indicates that the chances are less than 5% of having obtained the observed response or any more extreme response if H_0 is true.

p-values * are not * an error probability for rejecting H_0

p-values * are not * the posterior probability that H_0 is true

— p-values are usually misinterpreted —

We present next the derivation of Bayes Factor in some simple cases, and highlight severe discrepancies between the Bayesian and frequentist approaches to testing that arise from it (the Jeffrey-Lindley & Bartlett paradoxes). We then show that transformations of Fisher's p-value lead to its interpretation as odds or posterior probability of the hypothesis (Sellke, Bayarri & Berger (2001)). We conclude this section discussing multiple and sequential testing.

• Bayes Factor

Let $X|\theta \sim f(x|\theta) \rightarrow$ sample $\mathcal{L}_n = \{X_1, \dots, X_n\}$

We test $H_0 : \theta \in \Theta_0$

vs $H_1 : \theta \in \Theta_1$

with prior distributions

↘ prior probabilities $\mathbb{P}(H_0)$, $\mathbb{P}(H_1)$ of the hypothesis

↘ prior densities $f_0(\theta)$ and $f_1(\theta)$ on Θ_0 and Θ_1.

(can reduce to a point mass)

The marginal likelihoods are $f(\mathcal{L}_n|H_i) = \int f(\mathcal{L}_n|\theta, H_i) f_i(\theta) d\theta$

The posterior probabilities of the hypotheses are:

$$\mathbb{P}(H_0|\mathcal{L}_n) = \frac{f(\mathcal{L}_n|H_0)\,\mathbb{P}(H_0)}{f(\mathcal{L}_n|H_0)\,\mathbb{P}(H_0) + f(\mathcal{L}_n|H_1)\,\mathbb{P}(H_1)}$$

$$= 1 - \mathbb{P}(H_1|\mathcal{L}_n)$$

Then

$$\frac{\mathbb{P}(H_0|\mathcal{L}_n)}{\mathbb{P}(H_1|\mathcal{L}_n)} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \times \frac{f(\mathcal{L}_n|H_0)}{f(\mathcal{L}_n|H_1)}$$

posterior odds = prior odds × Bayes Factor B_{01}

$$\Rightarrow \mathbb{P}(H_0|\mathcal{L}_n) = \left(1 + \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)}\,\frac{1}{B_{01}}\right)^{-1}$$

Conclusions are drawn based (1) on the posterior odds (H_0 is rejected if $\mathbb{P}(H_0|\mathcal{L}_n) < \mathbb{P}(H_1|\mathcal{L}_n)$) or (2) on Bayes Factor B_{01}, with the default $\mathbb{P}(H_0) = \mathbb{P}(H_1)$.

Jeffreys (1961) suggested the scale:

B_{01}	Strength of evidence
1 to 3	Barely worth mentioning
3 to 10	Substantial
10 to 30	Strong
30 to 100	Very Strong
>100	Decisive

We derive next the expression of Bayes Factor in the simplest case of testing for the mean of a normal population.

• Testing for the mean of a normal population.

↘ Case 1: $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known.

Test for $\begin{cases} H_0 : \mu = 0 \quad (\text{w.p. } 1) \\ H_1 : \mu \sim \mathcal{N}(0, \frac{\sigma^2}{m}) = f_1(\mu) \end{cases}$

m controls the width of the prior $\equiv$ sample size

We compute $B_{01} = \dfrac{f(\mathcal{L}_n \mid H_0)}{f(\mathcal{L}_n \mid H_1)}$

- $f(\mathcal{L}_n \mid H_0) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\dfrac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right\}$

- $f(\mathcal{L}_n \mid H_1) = \int f(\mathcal{L}_n \mid \mu, H_1)\, f_1(\mu)\, d\mu$

$$= \int \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \times \frac{1}{(2\pi \frac{\sigma^2}{m})^{1/2}} \exp\left\{ -\frac{m}{2\sigma^2} \mu^2 \right\} d\mu$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(2\pi\frac{\sigma^2}{m})^{1/2}} \int \exp\left\{ -\frac{1}{2\sigma^2} \underbrace{\left[\sum (x_i - \mu)^2 + m\mu^2 \right]}_{\begin{array}{c} \| \\ (n+m)\mu^2 - 2\mu n \bar{x} + n\bar{x^2} \\ \| \\ (n+m)(\mu - \mu_n)^2 - \frac{(n\bar{x})^2}{n+m} + n\bar{x^2} \end{array}} \right\} d\mu$$

where $\mu_n = \dfrac{n}{n+m} \bar{x}$.

⇓ The integral is

$$\left(2\pi \frac{\sigma^2}{n+m}\right)^{1/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[-\frac{(n\bar{x})^2}{n+m} + n\bar{x^2} \right] \right\}$$

and

$$f(\mathcal{L}_n \mid H_1) = \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(2\pi\frac{\sigma^2}{m})^{1/2}} \left(2\pi\frac{\sigma^2}{n+m}\right)^{1/2} \exp\left\{ -\frac{1}{2\sigma^2} [\dots] \right\}$$

After simplifications,

$$B_{01} = \frac{f(\mathcal{L}_n \mid H_0)}{f(\mathcal{L}_n \mid H_1)} = \left(\frac{n+m}{m}\right)^{1/2} \exp\left(-\frac{1}{2} \frac{n}{n+m} z^2 \right), \quad z = \frac{\bar{x}}{\sigma/\sqrt{n}}$$

= function of the prior sample size, the data sample size, and the frequentist z-score z.

Our intuition is that when the z-score z is large; i.e. when $\bar{x}$ is away from the tested value 0, H_1 should be favored over H_0 and B_{01} should be small.

↖ in frequentist terms, a large z-score would yield a small p-value & we would reject H_0.

However =

↘ Jeffreys-Lindley paradox = for large n, $B_{01} \sim \sqrt{n} \exp\left(-\frac{1}{2} z^2\right)$, so that a classical frequentist test can strongly reject the null (which happens with a large z), while a Bayesian analysis can strongly support the null (if $B_{01} \sim \sqrt{n} \exp\left(-\frac{1}{2} z^2\right)$ is large).

↘ Bartlett paradox = Recall that under H_1, $\mu \sim \mathcal{N}(0, \frac{\sigma^2}{m})$. As $\frac{\sigma^2}{m} \to \infty \Leftrightarrow m \to 0$ for σ^2 fixed, a non-informative flat prior for μ is used. However, $B_{01} \to +\infty$ as $m \to 0$, irrespectively how large the z-score is. A Bayesian analysis systematically picks H_0 over H_1.

↘ Case 2: $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 unknown (nuisance)

Test for $\begin{cases} H_0: \mu = 0 \text{ wp1} & \sigma \sim \frac{1}{\sigma} \text{ (Jeffrey's suggestion)} \\ H_1: \mu|\sigma \sim \mathcal{N}(0, \frac{\sigma^2}{m}), & \sigma \sim \frac{1}{\sigma} \end{cases}$

Same prior as before, conditionally on σ

A common improper distribution is used for the common nuisance parameter σ

We compute $B_{01} = \dfrac{f(\mathcal{L}_n | H_0)}{f(\mathcal{L}_n | H_1)}$

- $f(\mathcal{L}_n | H_0) = \int_0^\infty f(\mathcal{L}_n | \mu=0, \sigma, H_0) f(\sigma | H_0)\, d\sigma$

  $\sim (1/\sigma)^n \exp\left(-\frac{1}{2\sigma^2}\sum X_i^2\right)$ (= $\exp(\cdot)$) $\quad$ ($= 1/\sigma$)

  ↳ we omit the term $(1/2\pi)^{n/2}$ which will cancel out with the denominator in B_{01}.

$$= \int_0^\infty \left(\frac{1}{\sigma}\right)^{n+1} \exp\left(-\frac{1}{2\sigma^2} T\right) d\sigma \qquad T = n(s^2 + \bar{x}^2)$$

$$s^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$$

$x = \sigma^2$ ↓

$$= \int_0^\infty \left(\frac{1}{x}\right)^{\frac{n+1}{2}} \exp\left(-\frac{T}{2x}\right) \frac{dx}{2\sqrt{x}}$$

$$= \frac{1}{2}\int_0^\infty \left(\frac{1}{x}\right)^{\frac{n}{2}+1} \exp\left(-\frac{T}{2x}\right) dx$$

inverse gamma distribution ↓

$$= \frac{1}{2}\Gamma\left(\frac{n}{2}\right)\left(\frac{T}{2}\right)^{-n/2}$$

$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}$

- $f(\mathcal{L}_n | H_1) = \iint \left(\frac{1}{\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\underbrace{\sum (x_i - \mu)^2}_{= n(s^2 + (\bar{x}-\mu)^2)}\right) \mathcal{N}(\mu | 0, \frac{\sigma^2}{m}) \frac{1}{\sigma}\, d\mu\, d\sigma$

$$= \int \left\{ \int \exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right) \mathcal{N}(\mu | 0, \frac{\sigma^2}{m})\, d\mu \right\} \left(\frac{1}{\sigma}\right)^{n+m} \times \exp\left(-\frac{ns^2}{2\sigma^2}\right) d\sigma$$

$\left(\frac{m}{n+m}\right)^{1/2} \exp\left(-\frac{nm\bar{x}^2}{2(n+m)\sigma^2}\right)$

Let $T_2 := \frac{nm}{n+m}\bar{x}^2 + ns^2$ ↓

$$= \left(\frac{m}{n+m}\right)^{1/2} \int_0^\infty \left(\frac{1}{\sigma}\right)^{n+1} \exp\left(-\frac{T_2}{2\sigma^2}\right) d\sigma$$

$$= \frac{1}{2}\left(\frac{m}{n+m}\right)^{1/2} \Gamma\left(\frac{n}{2}\right)\left(\frac{T_2}{2}\right)^{-n/2}$$

It follows that

$$B_{01} = \frac{f(\mathcal{L}_n | H_0)}{f(\mathcal{L}_n | H_1)} = \left(\frac{n+m}{m}\right)^{1/2} \left(\frac{n + \frac{mt^2}{n+m}}{n + t^2}\right)^{n/2}, \quad t = \frac{\bar{x}}{s/\sqrt{n}}$$

↳ t plays the same role as z before, replacing σ^2 with its estimate s^2.
t is the frequentist t-statistic. A large value of t yields a small p-value, and the rejection of the null hypothesis.

- Jeffreys-Lindley paradox shows up again: as $n \to \infty$, $B_{01} \sim \sqrt{n} \Rightarrow$ a Bayesian analysis favor H_0 over H_1, irrespectively how large t may be.
- Bartlett paradox: A flat prior on μ is inadequate here as well, since as $m \to 0$, $B_{01} \to \infty$ and Bayes factor picks H_0 over H_1 all the time.

In addition to these paradoxes, note that as $t^2 \to \infty$, Bayes Factor B_{01} tends to a constant value (n, m fixed), which is also counterintuitive. As $t^2 \to \infty$, $\bar{X}$ differs more and more from 0 and one would expect B_{01} to converge to 0.

One possible solution is to use a Cauchy prior on $\mu | \sigma$ instead of the conjugate Gaussian prior, as suggested by Jeffreys. While the numerator in B_{01} is still available in close form, the denominator must be computed numerically. Jeffreys gave a (bad) approximation to B_{01} in this case, $B_{01} \approx \sqrt{\frac{\pi n}{2}} \exp\left(-\frac{t^2}{2}\right)$, which is sufficient to show that $B_{01} \to 0$ as $t^2 \to \infty$, while Lindley's paradox remains since $B_{01} \to \infty$ as $n \to \infty$.

× Remark: Jeffreys recommend to use improper priors for common parameters to H_0 and H_1 (such as σ^2). Kass & Vaidyanathan (1992) showed that Bayes Factors are ≈ insensitive to the choice of a common prior under weaker assumptions than orthogonality (i.e. diagonal Fisher info matrix). Jeffreys advocate the use of proper priors for non-common parameters between H_0 and H_1. The prior should be centered & symmetric about 0, has scale σ, and no moments (like Cauchy)

× Remark = To further support Jeffreys recommendation not to use vague or improper priors for common parameters, consider again

the case $X_1, \dots, X_n \sim \mathcal{N}(\mu, \tau^2)$, τ^2 known, where we test for $H_0 : \mu = 0$ (w.p. 1).

- vague prior for H_1 : $\mu \sim \mathcal{U}(-c, c)$

$$B_{01} = \frac{f(\mathcal{L}_n | H_0)}{\int_{-c}^{c} f(\mathcal{L}_n | \mu, H_1) \frac{1}{2c} d\mu} \approx \frac{2c f(\mathcal{L}_n | H_0)}{\int_{-\infty}^{+\infty} f(\mathcal{L}_n | \mu, H_1) d\mu} = 2c f(\mathcal{L}_n | H_0)$$

c large enough

The value of B_{01} depends heavily on the choice of c.

- improper prior for H_1 : $\mu \sim 1$ or $\mu \sim 2$ yields Bayes Factors $f(\mathcal{L}_n | H_0)$ and $2 f(\mathcal{L}_n | H_0)$, respectively.
  ⇒ Which normalization to use ?

Back to Jeffreys-Lindley's paradox.

↳ denotes a situation where a frequentist test would strongly reject the null, while B_{01} would favor heavily the null over the alternative.

This situation is likely to occur when

(i) the sample size n is large

(ii) H_0 is relatively precise (e.g. $\Theta = 0$ w.p.1), while the alternative is relatively diffuse (vague, improper, but not only).

× How to make sense of this paradox ?

In large sample situations, the Bayesian analysis requires more evidence to detect an effect (reject H_0);

while the frequentist approach tends to reject too easily the null

↘ One way around this is to adjust the level α of the test as n grows, instead of keeping it constant.

Indeed, as n grows, $\bar{x}$ concentrates around its mean. This allows us to move from confidence intervals with fixed coverage $1-\alpha$ =

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}}\sigma n^{-1/2},\ \bar{x} + z_{1-\frac{\alpha}{2}}\sigma n^{-1/2}\right] \rightarrow \text{coverage } 1-\alpha$$

to intervals of the form =

$$\left[\bar{x} - z_n\sigma n^{-1/2},\ \bar{x} + z_n\sigma n^{-1/2}\right].$$

↳ coverage $1-h_n \longrightarrow 1$ as $n \to +\infty$, provided $h_n \to 0$, and $z_n = \Phi^{-1}(1-h_n)$ such that $z_n n^{-1/2} \to 0$ (otherwise the confidence interval blows up)

Take e.g. $h_n \sim n^{-p}$, $p > 1$, see Naaman (2016)

The last written interval has coverage $1-h_n > 1-\alpha$

$\Rightarrow$ it includes the value 0 more often than the classic interval, and therefore the null is "accepted" more often.

Since $\bar{x}$ concentrate around its mean as n grows, the alternative prior under H_1 in the Bayesian setting consider a wide range of alternatives that are meaningless to describe the data.

B_{01} tends to pick up H_0 even if it poorly describes the data. This is even more problematic with diffuse priors.

↳ The prior is treated very differently in Bayesian posterior inference & Bayesian testing:

$f(\mu|\mathcal{L}_n) \propto f(\mathcal{L}_n|\mu) f(\mu)$ vs $\int f(\mathcal{L}_n|\mu) f(\mu)\, d\mu$

"update belief about the prior" — "average over all possible values"

× Summary = We considered so far Bayesian tests where

→ A point null is used $H_0: \theta = 0$ w.p.1

→ An alternative whose prior puts positive mass on $\Theta_0 = \{0\}$ $H_1: \theta \sim \mathcal{N}(0, \frac{\sigma^2}{m})$ or Cauchy.

↳ As noted by Johnson & Rossell (2010), this is a problematic assumption since such alternative priors "do not incorporate any notion of a minimally significant separation between the null and the alternative hypotheses". The authors show that

"For a true null hypothesis, the Bayes factor in favor of the alternative hypothesis decreases only at rate $O_p(n^{-1/2})$"

while

"For a true alternative hypothesis, the Bayes factor in favor of the null hypothesis decreases expanentially fast"

recall the expression of B_{01} in p.46

When H_0 is true, evidence for the null accumulates slowly, while when the alternative is true, evidence accumulates exp. fast.

⇒ We can either fix the null, or the alternative & reconcile the Bayesian approach with the frequentist one.

- Johnson & Rossell (2010) consider a class of "non-local alternative priors", that put zero prior mass for all values $\theta \in \Theta_0$, an a strictly positive mass on Θ_1. "Although this class of prior densities does not provide exponential convergence of Bayes factors in favor of true null hypothesis, it offers a substantial improvement in convergence rates over local alternative priors."

- Casella & Berger (1987) consider non-point nulls $H_0: \theta \leq 0$ vs $H_1: \theta > 0$ and show closer agreement to the frequentist approach. They prove that for many classes of prior distributions, the infimum of the Bayesian posterior probability of H_0 is either equal to or bounded above by the p-value.

- Calibration of p-values.

Since frequentist p-values are often misinterpreted, some attempts to transform them so that they are interpretable as odds ratio or posterior probability of the hypothesis have been considered in the litterature. Following Sellke, Bayarri & Berger (2001), under H_0: $p \sim \mathcal{U}(0,1)$ $\quad p$ = p-value

(see e.g. p.26 in MS = HYPOTHESIS TESTING)

Test H_0 versus H_1: $p \sim f(p|\alpha) = B(p|\alpha, \beta=1) = \alpha p^{\alpha-1}$, with some prior $\pi(\alpha)$ on α.

Then $B_{01}(\pi) = \dfrac{f(p|\alpha=1)}{\int_0^1 f(p|\alpha)\pi(\alpha)\,d\alpha}$ ↖ since $\mathcal{U}(0,1) = B(1,1)$

One can show that

$$\underline{B} = \inf_{\pi} B_{01}(\pi) = \begin{cases} -ep\log p & \text{if } p < e^{-1} \\ 1 & \text{o/w} \end{cases}$$

⇒ Calibration $-ep\log p$, where p is the frequentist p-value, can be taken as an objective lower bound for any prior π on α an the odds of H_0 to H_1.

- Comparing two means.

Suppose that $X_1, \dots, X_{n_0} \sim \mathcal{N}(\mu_X, \sigma^2)$ (σ^2 assumed equal)

$Y_1, \dots, Y_{n_1} \sim \mathcal{N}(\mu_Y, \sigma^2)$.

We wish to test for $H_0: \mu_X = \mu_Y$ vs the alternative $\mu_X \neq \mu_Y$.

To proceed, it is convenient to reparametrize μ_X and μ_Y using the overall mean $\mu = \frac{1}{2}(\mu_X + \mu_Y)$, and $\delta = \mu_X - \mu_Y$:

$X \sim \mathcal{N}(\mu + \frac{\delta}{2}, \sigma^2)$

$Y \sim \mathcal{N}(\mu - \frac{\delta}{2}, \sigma^2)$,

and testing for $\delta = 0$ vs $\delta \neq 0$ (H_0) (H_1).

μ, σ^2 = common parameters for H_0 and H_1 → ok to use a non-informative prior $f(\mu, \sigma^2 | H_i) \propto \frac{1}{\sigma^2}$ $i = 0, 1$.

δ = diffuse priors are not recommended.

Consider first a conjugate normal prior $\frac{\delta}{\sigma} \sim \mathcal{N}(0, \sigma_0^2)$.

Then $$B_{01} = \frac{f(\mathcal{L}_n | H_0)}{f(\mathcal{L}_n | H_1)} = \frac{\iint f(\mathcal{L}_n | \delta=0, \mu, \sigma^2) f(\mu, \sigma^2 | H_0)\, d\mu\, d\sigma^2}{\iiint f(\mathcal{L}_n | \delta, \mu, \sigma^2) f(\delta|\sigma^2) f(\mu, \sigma^2 | H_1)\, d\delta\, d\mu\, d\sigma^2}$$

B_{01} can be calculated exactly in this case, and is expressed in terms of the frequentist pooled-variance two-sample t statistic,

$$BF_{01} = \left(\frac{1 + t^2/\upsilon}{1 + t^2/[\upsilon(1+n\tau_0^2)]} \right)^{-\frac{\upsilon+1}{2}} (1+n\tau_0^2)^{1/2},$$

where $t = \dfrac{\bar{X}-\bar{Y}}{S_p\sqrt{n}}$, $n = \left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-1}$

$$S_p = \frac{1}{n_0+n_1-2}\left((n_0-1)S_0^2 + (n_1-1)S_1^2\right)$$

$$\upsilon = n_0+n_1-2,$$

see Gönen, Johnson, Lu & Westfall (2005).

↑ This approach shows that t arises from a Bayesian formulation, but suffers from the same paradoxes as in the single population problem.

- Jeffreys-Lindley: For large n, $B_{01} \propto \sqrt{n}\, e^{-t^2/2}$, so a strong frequentist effect (large value of t) may lead to strong evidence for H_0.
- As $t \to \infty$, BF_{01} tends to a constant, while we would expect $B_{01} \to 0$ to be in agreement with the frequentist approach.

↳ Use a Cauchy prior instead (B_{01} must be computed numerically)

- Testing based on the posterior distribution.

A more traditional approach can be used for testing, based directly on the posterior distribution directly.

Consider the problem of comparing two binomial proportions,

revisiting the situation on p.7/8:

$X \sim B(p_0)$ $\quad p_0 \sim B(\alpha_0, \beta_0)$ $\quad p_0 | X \sim B(\alpha_0 + S_{n_0}, n_0 + \beta_0 - S_{n_0})$

$Y \sim B(p_1)$ $\quad p_1 \sim B(\alpha_1, \beta_1)$ $\quad p_1 | Y \sim B(\alpha_1 + S_{n_1}, n_1 + \beta_1 - S_{n_1})$

From these, the posterior distribution of the difference $\delta := p_0 - p_1$ can easily be computed (numerically).

To decide which one of $H_0: \delta \leq 0$ or $H_1: \delta > 0$ is more appropriate, we may reject H_0 if the 95% posterior interval is fully contained in the positive half space. Other similar criteria may be considered.

This approach is similar in spirit to the frequentist method which rejects H_0 at 95% confidence level if the one-sided confidence interval excludes 0 ($\Leftrightarrow$ p-value less than 0.05). The two approaches coincide when the posterior credible interval has a frequentist interpretation (or approximately in large sample situations via the Bernstein von Mises theorem)

A more Bayesian treatment introduces a loss, characterizing the cost associated with making a mistake. Let

$$\ell_0(p_0, p_1) := \max(p_1 - p_0, 0)$$
$$\ell_1(p_0, p_1) := \max(p_0 - p_1, 0)$$

If H_0 is true, $p_0 \leq p_1$ and $p_1 - p_0$ is the price to pay to draw erroneous conclusions (linear in $p_1 - p_0$). The cost is 0 is the right decision is made.

The expected loss is

$$\int_0^1 \int_0^1 l_i(p_0, p_1)\, f(p_0, p_1 \mid x \cup y)\, dp_0 p_1 \quad , \quad i = 0, 1.$$

↑ can easily be calculated numerically

In practice, we select a threshold $\varepsilon > 0$ and stop the experiment when the expected loss for $i=0$ or $i=1$ falls below ε.

↑ For example, suppose that historical data indicate that $p_0 \approx 0.20$. If we want to detect a relative effect of about 2% improvement from this base rate, leading to $(p_1 - p_0)/p_0 = 2\%$, $p_1 = 0.204$, we set $\varepsilon = 0.004$.

x Remark: Superiority of BF vs posterior-based testing for a particular application should be established (at least numerically via simulations) by controlling for the type I and type II error rates.

III.3. Sequential Testing

In sequential testing, samples are collected until enough evidence for H_0 or H_1 is obtained, and a decision is made. In frequentist statistics, repeated peeks at the data is known to inflate the type I error. Boundaries for optimal stopping must be updated accordingly, see e.g. p. 37-47 in MS: HYPOTHESIS TESTING. We explore here how Bayesian techniques for sequential testing compare, and discuss testing based on
- the posterior distribution
- Bayes Factors.

(i) Posterior distribution for sequential testing.

Consider the simple case where $X_1, \dots, X_n$ are received sequentially & $\mathcal{N}(\mu, \sigma^2)$ distributed, testing for μ ($H_0: \mu = 0$ vs $H_1: \mu \neq 0$), σ^2 unknown. With a non-informative prior $f(\mu, \sigma^2) \sim \frac{1}{\sigma^2}$, the posterior distribution of $\mu \mid \underline{x}_n$ is $t(n-1, \bar{x}, \frac{1}{n} s^2)$, see page 21.

In this particular case, credible bounds & frequentist bounds exactly coincide, and

> 95% credible bound excludes 0
> $\Leftrightarrow$ 95% confidence interval excludes 0
> $\Leftrightarrow$ p-value < 0.05

↑ of $z = \frac{\bar{x}}{s/\sqrt{n}} \sim t(n-1)$

$\Rightarrow$ Repeated peeks at the data will inflate the type I error with this procedure. With an informative prior, credible & confidence intervals do not match exactly, and the equivalence does not hold. However, in large sample situations, Bernstein von Mises ensures that the posterior distribution has a frequentist interpretation, and inflation of the type I error follows. In fact, even in the non-asymptotic case, the type I error rate needs to be controlled, see Shi & Yin (2019). The authors present a procedure for adjusting the bounds, following the work of Pocock & O'Brien in the frequentist case (see MS: HYPOTHESIS TESTING).

What makes this approach prone to type I error inflation is the vanishing effect on the prior, resulting in the posterior being "too frequentist", therefore inheriting issues associated with sequential testing

posterior ∝ likelihood × prior

The prior is treated very differently in Bayes Factors. Its effect is persistent as new data are collected, since the likelihood is integrated with respect to the prior distribution:

$\int$ likelihood × prior ← no feedback loop

Although this is a limitation of BF to some extent (the prior has a huge effect on the value of BF & on the decision), it turns out to be a nice feature for sequential testing, being more robust to the inflation of the type I error.

(iii) Bayes Factors for sequential testing.

"The rules of governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" Edwards, Lindman & Savage (1963)

Suppose that $X_1, \dots, X_n \sim f_0$ under H_0
$\sim f_1$ under H_1
(with appropriate priors on the parameters)

$$BF_{01} = \frac{f(x_n \mid H_0)}{f(x_n \mid H_1)}.$$

Consider the following procedure, described in Schönbrodt, Wagenmakers, Zehetleitner & Perugini (2015)

(i) Select a threshold indicating the decisiveness of evidence
Ex: B_{01} of 10 for H_0 & the reciprocal value of 1/10 for H_1,

(ii) Collect a minimum of n_{min} observations before stopping the experiment; compute BF_{01} after each observation is collected.
(to avoid misleading evidence leading to early termination)

(iii) When BF_{01} crosses one of the boundaries, stop the experiment & report the final BF_{01} & mean and 95 % credible bounds (posterior).

× Remarks: (i) As $n \to \infty$, BF_{01} is consistent (it converges to $+\infty$ if H_0 is true or to 0 if H_1 is true). and the procedure is guaranted to converge

(ii) The false positive rate (α) is not controlled by the experimenter (a major drawback of this approach, compared to the frequentist procedure). The FPR is a function of the priors, the chosen threshold, and the true effect size in the population. It cannot be computed in advance, but simulations can be performed to get an idea of its magnitude. The FNR magnitude can also be estimated via simulations.

(iii) The procedure suffers from the same drawbacks as in the non-sequential case: evidence accumulates faster when a true effect is present than when there is none (recomended to "separate" the null from H_1, see page 52).

(iv) Simulations indicate that sample size requirements are much lower here than in the frequentist case (e.g. when calculated from a usual power study).
⇒ less data are needed to achieve better performance (FP and FN remain small in %, well below 5% for α, and well below 5% for β, even for relatively small true effect sizes) ← Simulations performed by the authors in the context of comparison of two means.

(v) Downwards bias reported in the point estimates. The procedure tends to underestimate the true effect.

(vi) BF_{01} provides evidence * for * H_0, and not solely against H_0, as happens in the frequentist approach.

(vii) Rouder (2014) shows that in simple situations at least, BF_{01} keeps its interpretation whether optimal stopping is performed or not.

IV. EMPIRICAL BAYES

IV.1. The James-Stein Estimator

Let $X_1, \dots, X_n$ denote n iid univariate random variables, with mean θ.

A common justification for taking $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is its desirable properties. It is unbiased for θ ($\mathbb{E}\bar{X} = \theta$), and no other unbiased function of the data, linear or non-linear, can estimate θ more accurately than $\bar{X}$ in terms of variance (= mean square error).

In the 30's, Neyman, Pearson & Wald considered a different approach to this estimation problem & dropped unbiasedness as a criterion. ↳ Statistical Decision Theory

They examine all functions $\hat{\theta} = \theta(X_1, \dots, X_n)$ of the data & estimators are compared through a risk function, such as the square loss: $R(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2$.

x Ex = Let $X_1, \dots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known, and consider the following estimators of θ:

$\hat{\theta}_1 = \bar{X}$, $\hat{\theta}_2 = \frac{1}{2}\bar{X}$, $\hat{\theta}_3 = \text{median}(X_1, \dots, X_n)$

We compare the risk of $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

. $R(\hat{\theta}_1) = \mathbb{E}(\bar{X} - \theta)^2 = \text{var}\,\bar{X} = \frac{\sigma^2}{n}$
($\mathbb{E}\hat{\theta}_1$)

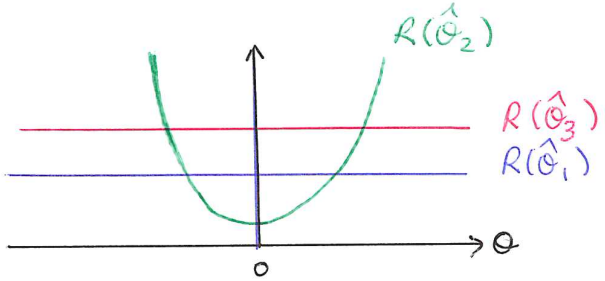
- $R(\hat{\theta}_2) = \mathbb{E}\left(\frac{1}{2}\bar{X} - \theta\right)^2 = \frac{1}{4}\mathbb{E}(\bar{X} - 2\theta)^2$

$$= \frac{1}{4}\left(\mathbb{E}(\bar{X}^2) - 4\theta\mathbb{E}X + 4\theta^2\right)$$

$$= \frac{1}{4n^2}\mathbb{E}\left(\sum_i X_i^2 + \sum_{i\neq j} X_i X_j\right)$$

$$\dots = \frac{1}{4}\theta^2 + \frac{1}{4n}\sigma^2$$

- For large n, it is possible to show that the sample median is $\approx \mathcal{N}\left(\theta, \frac{\pi}{2}\frac{\sigma^2}{n}\right)$, so that $R(\hat{\theta}_3) \simeq \frac{\pi}{2}\frac{\sigma^2}{n}$.



↘ Estimator $\hat{\theta}_1$ has smaller risk than $\hat{\theta}_3$ for any value of θ. We say that $\hat{\theta}_3$ is INADMISSIBLE.

↘ Estimator $\hat{\theta}_2$ has smaller risk than $\hat{\theta}_1$ for small values of θ only. A natural question arises: can we beat the sample mean estimator $\hat{\theta}_1 = \bar{X}$ for any value of θ?
An answer to this question was provided by Blyth, Lehmann and Hodges in 1950: there are no such estimator. In other words, $\hat{\theta}_1$ is ADMISSIBLE.

Stein considered the estimation of several means at one time, where the risk is taken as the sum of the expected values of

of square errors of estimation for all individual means:

$$R(\hat{\theta}) = \mathbb{E}\|\hat{\theta} - \theta\| = \sum_{i=1}^{d}\mathbb{E}(\hat{\theta}_i - \theta_i)^2,$$

where $\theta = (\theta_1, \dots, \theta_d)^t \in \mathbb{R}^d$ is the mean vector to be estimated.

In 1955, Stein showed that the vector of sample means is also admissible if we estimate simultaneously 2 means, but inadmissible if there are 3 means or more.

↑ In fact, Stein showed this in 1955 only if the number of means is very large. He extended the result together with James in 1961 for all $n \geq 3$, and did so in a constructive manner.

- James-Stein estimator

Let $X_1, \dots, X_d$ be d independent variables with

$$X_i | \theta_i \sim \mathcal{N}(\theta_i, \sigma_0^2).$$

Put $X = (X_1, \dots, X_d)^t \in \mathbb{R}^d$

$\theta = (\theta_1, \dots, \theta_d)^t \in \mathbb{R}^d$, so that $X|\theta \sim \mathcal{N}(\theta, \sigma_0^2 I)$

We are looking for estimators $\hat{\theta} = \theta(X)$ of θ.

Consider the case of a single observation X. Each component of θ is therefore estimated using a single noisy measurement. The maximum likelihood approach gives $\hat{\theta}_{ML} = X$. It has risk

$$R(\hat{\theta}_{ML}) = \sum_{i=1}^{d}\mathbb{E}(X_i - \theta_i)^2 = d\sigma_0^2.$$

↑ since the i-th component of $\hat{\theta}_{ML}$ is X_i.

James & Stein (1961) introduced a novel estimator of θ that has a strictly smaller risk than $\hat{\theta}_{ML}$ for any $d \geq 3$. It can be derived in a Bayesian context.

but analysed using frequentist tools.

Consider the model

$$X \mid \theta \sim \mathcal{N}(x \mid \theta, \sigma_0^2 I)$$
$$\theta \sim \mathcal{N}(\theta \mid 0, \alpha I)$$

... general covariance matrix (not necessarily diagonal)

More generally ...

... mean m

... known or unknown

$(\alpha > 0, \ \sigma_0^2 > 0)$

... not necessarily normal

We consider the simplest case here, where both α and σ_0^2 are known.

The posterior distribution is $\theta \mid X \sim \mathcal{N}(\theta \mid \beta X, \beta I)$, where $\beta = \dfrac{\alpha}{\alpha + \sigma_0^2}$.

The posterior mean is then $\hat{\theta}_{post} = \beta X = \dfrac{\alpha}{\alpha + \sigma_0^2} X$

$$\hat{\theta}_{post} = \left(1 - \frac{\sigma_0^2}{\alpha + \sigma_0^2}\right) X$$

All components of X are shrunk towards 0

When α is unknown, it can be estimated from the data. Estimating the prior parameters using the data lies at the heart of Empirical Bayes (EB) methods. We explain the approach more generally in the next subsection. In the context of the James-Stein estimator, note that the marginal distribution of

X is $\mathcal{N}(x \mid 0, (\alpha + \sigma_0^2) I)$ — indpt components

$$\Rightarrow \sum_{i=1}^{d} \frac{X_i^2}{\alpha + \sigma_0^2} \sim \chi_d^2$$

Denoting $\|X\|^2 = \sum_{i=1}^{d} X_i^2$, we see that $\dfrac{\|X\|^2}{\alpha + \sigma_0^2} \sim \chi_d^2$

$$\frac{\alpha + \sigma_0^4}{\|X\|^2} \sim \text{Inv-}\chi_d^2$$

$$\Rightarrow \mathbb{E}\left(\frac{\alpha + \sigma_0^4}{\|X\|^2}\right) = \frac{1}{d-2}$$

$$\mathbb{E}\left(\frac{d-2}{\|X\|^2}\right) = \frac{1}{\alpha + \sigma_0^4}.$$

We can replace the term $\dfrac{1}{\alpha + \sigma_0^2}$ appearing in the expression of the posterior mean with $\dfrac{d-2}{\|X\|^2}$, an unbiased estimator of it. We obtain precisely the definition of the James-Stein estimator of θ:

$$\hat{\theta}_{JS} = \left(1 - \frac{(d-2)\sigma_0^2}{\|X\|^2}\right) X$$

shrinkage factor

- The term $\|X\|^2$ assesses if each component of X is far from 0.
- If X is close to zero, $\|X\|^2$ is small & the shrinkage is large : the JS estimator shrinks each component of X a lot since there is confidence that X is close to 0.
- If X is far from the origin, $\|X\|^2$ is large & the shrinkage is small (since not much confidence in shrinking values towards 0)

Theorem [James & Stein (1961)]

Suppose $X \mid \theta \sim \mathcal{N}(x \mid \theta, \sigma_0^2 I)$, $\sigma_0^2 > 0$ known, $X, \theta \in \mathbb{R}^d$.

For $d \geq 3$, the James-Stein estimator $\hat{\theta}_{JS}$ everywhere dominates the maximum likelihood estimate $\hat{\theta}_{ML}$ under the square loss,

$$\mathbb{E} \| \hat{\theta}_{JS} - \theta \|^2 < \mathbb{E} \| \hat{\theta}_{ML} - \theta \|^2,$$

for every fixed θ.

A frequentist result! (θ fixed) Not a Bayesian one!

$\Rightarrow$ The JS estimator is biased for θ, but has lower MSE than the MLE for any fixed value of θ (the increase in bias is compensated by the large decrease in variance).

This makes the MLE inadmissible for $d \geq 3$.

Remark: The result holds as well for m iid observations $X_1, \dots, X_m$ since if $X_i \mid \theta \sim \mathcal{N}(x \mid \theta, \sigma_0^2 I_d)$, then $\bar{X} = \frac{1}{m} \sum_{i=1}^{m} X_i \sim \mathcal{N}(x \mid \theta, \frac{\sigma_0^2}{m} I_d)$, and the previous analysis holds.

$\Rightarrow$ Repeat the derivation of the JS estimator replacing σ_0^2 with $\frac{\sigma_0^2}{m}$,

$$\hat{\theta}_{JS} = \left(1 - \frac{(d-2)\frac{\sigma_0^2}{m}}{\| \bar{X} \|^2} \right) \bar{X}$$

Each component of $\bar{X} \in \mathbb{R}^d$ is shrank by the same factor. ∎

proof =

$$R(\hat{\theta}_{JS}) = \mathbb{E} \| X - \theta - \frac{(d-2)\sigma_0^2}{\|X\|^2} X \|^2$$

$$= \mathbb{E} \| X - \theta \|^2 - 2(d-2)\sigma_0^2 \sum_{i=1}^{d} \mathbb{E}\left\{ \frac{X_i (X_i - \theta_i)}{\|X\|^2} \right\} + (d-2)^2 \sigma_0^4 \, \mathbb{E}\left\{ \frac{1}{\|X\|^2} \right\}. \quad (*)$$

$$(*) = \int \underbrace{\frac{x_i}{\|x\|^2}}_{u} \underbrace{(x_i - \theta_i) \frac{1}{(2\pi\sigma_0^2)^{d/2}} \exp\left\{ -\frac{1}{2\sigma_0^2} \| x - \theta \|^2 \right\}}_{v'} dx$$

integration by parts

$$= \int \underbrace{\frac{\|x\|^2 - 2x_i}{\|x\|^4}}_{u'} \underbrace{\sigma_0^2 \frac{1}{(2\pi\sigma_0^2)^{d/2}} \exp\left\{ -\text{"}-\text{"} - \right\}}_{v} dx$$

$$= \sigma_0^2 \, \mathbb{E}\left\{ \frac{\|X\|^2 - 2X_i}{\|X\|^4} \right\}$$

$$\sum_{i=1}^{d} \mathbb{E}\left\{ \frac{X_i (X_i - \theta_i)}{\|X\|^2} \right\} = d\sigma_0^2 \, \mathbb{E}\left\{ \frac{1}{\|X\|^2} \right\} - 2\sigma_0^2 \, \mathbb{E}\left\{ \frac{\sum_{i=1}^{d} X_i}{\|X\|^4} \right\}$$

$$= (d-2)\sigma_0^2 \, \mathbb{E}\left\{ \frac{1}{\|X\|^2} \right\}.$$

We get

$$R(\hat{\theta}_{JS}) = d\sigma_0^2 - 2(d-2)^2 \sigma_0^4 \, \mathbb{E}\left\{ \frac{1}{\|X\|^2} \right\} + (d-2)^2 \sigma_0^4 \, \mathbb{E}\left\{ \frac{1}{\|X\|^2} \right\}$$

$$= d\sigma_0^2 - \underbrace{(d-2)^2 \sigma_0^4 \, \mathbb{E}\left\{ \frac{1}{\|X\|^2} \right\}}_{>0 \text{ for } d \geq 3}$$

$$< d\sigma_0^2 = R(\hat{\theta}_{ML}).$$

∎

Remarks (i) The James-Stein estimator derived on page shrinks the components of X towards 0. Other shrinking value can be considered. For example, Efron & Morris (1977) shrink the components of $X \in \mathbb{R}^d$ towards its mean $\bar{X} = \frac{1}{d}\sum_{i=1}^{d} X_i$:

$$\hat{\theta}_{EM} = \bar{X} + \left(1 - \frac{(d-3)\sigma_0^2}{\|X - \bar{X}\|^2}\right)(X - \bar{X})$$

↑ $\hat{\theta}_{EM}$ has strictly smaller risk than $\hat{\theta}_{ML}$ when $d \geq 4$.

It can be derived in a similar manner as before, under the assumption that $X_i | \theta_i \sim \mathcal{N}(\theta_i, \sigma_0^2)$, and $\theta_i \sim \mathcal{N}(m, \alpha)$.

(ii) In the expression of the original JS estimator, the shrinkage factor $\left(1 - \frac{(d-2)\sigma_0^2}{\|X\|^2}\right)$ may be less than 0, and arbitrary large negatively, which is counterintuitive. This can be avoided by taking the positive-part of the shrinkage factor. Baranchick (1964, 1970) showed that the JS-positive-part estimator

$$\hat{\theta}^{+}_{JS} = \left(1 - \frac{(d-2)\sigma_0^2}{\|X\|^2}\right)_{+} X$$

↑ $(u)_+ = \begin{cases} u & \text{if } u \geq 0 \\ 0 & \text{otherwise} \end{cases}$

dominates $\hat{\theta}_{JS}$ for any θ. ↑ smaller MSE

In other words, the JS estimator is inadmissible.

Brown (1971) showed that $\hat{\theta}^{+}_{JS}$ also is inadmissible (but improving on $\hat{\theta}^{+}_{JS}$ is known to be hard)

(iii) James-Stein & Risk Minimization.

Since the JS estimator achieves lower MSE than the MLE, and is of the form cX for $c = \left(1 - \frac{(d-2)\sigma_0^2}{\|X\|^2}\right)$, we may directly consider the class of estimators cX for $c \in \mathbb{R}$, and look for the optimal value of c.

$$\begin{aligned} R(cX) &= \mathbb{E}\|\theta - cX\|^2 \\ &= \|\theta\|^2 + c\,\underbrace{\mathbb{E}\|X\|^2}_{d\sigma_0^2 + \|\theta\|^2} - 2c\sum_{i=1}^{d}\underbrace{\mathbb{E}(X_i\theta_i)}_{\|\theta\|^2} \\ &= \|\theta\|^2(1-c)^2 + c^2 d\sigma_0^2. \end{aligned}$$

The minimum is achieved at $c^* = \dfrac{\|\theta\|^2}{\|\theta\|^2 + d\sigma_0^2}$.

This value cannot be computed in practice since it relies on the unknown θ. However, note that $\mathbb{E}\|X\|^2 = \|\theta\|^2 + d\sigma_0^2$, so that $\|X\|^2$ is an unbiased estimator for $(\|\theta\|^2 + d\sigma_0^2)$. This leads to

$\hat{\theta} = \left(1 - \frac{d\sigma_0^2}{\|X\|^2}\right)X$, which has the same form as $\hat{\theta}_{JS}$, except that the optimal value of $(d-2)$ is replaced with d.

(iv) James-Stein & Ridge Regression.

[Ridge Regression] see SL: RIDGE REGRESSION & LASSO

The linear model is $y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, so that $y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ + Bayesian prior $\beta \sim \mathcal{N}(0, \tau I_d)$.

The posterior is $\beta | \mathcal{L}_n \sim \mathcal{N}(\beta | \mu_n, \Sigma_n)$, with

$$\begin{cases} \mu_n = \frac{1}{\sigma^2} \Sigma_n X^t y \\ \Sigma_n^{-1} = \frac{1}{\tau} I + \frac{1}{\sigma^2} X^t X \end{cases}$$

The posterior mean is thus $\frac{1}{\sigma^2}\left(\frac{1}{\tau} I + \frac{1}{\sigma^2} X^t X\right)^{-1} X^t y$

$$= \left(\frac{\sigma^2}{\tau} I + X^t X \right)^{-1} X^t y$$

$\frac{\sigma^2}{\tau} = \lambda$

$\Rightarrow$ With $\tau = \frac{\sigma^2}{\lambda}$, the Bayesian view of RR is equivalent to the penalized optimization problem

$$\min_{\beta} RSS_2(\lambda), \text{ where } RSS_2(\lambda) = \|y - X\beta\|^2 + \lambda \|\beta\|^2,$$

since the value $\hat{\beta}_\lambda$ minimizing $RSS_2(\lambda)$ is precisely given by $\hat{\beta}_\lambda = (\lambda I + X^t X)^{-1} X^t y$.

Note that when $\lambda = 0$, $\hat{\beta}_0 = \hat{\beta}$ = least squares solution $= (X^t X)^{-1} X^t y$.

One can show that $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^t X)^{-1})$.

[James-Stein] Assume first that the columns of X are orthogonal, i.e. $X^t X = I_d$. Then the Bayesian linear model is equivalent to

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}(\beta, \sigma^2 I_d) \\ \beta &\sim \mathcal{N}(0, \tau I_d) \end{aligned}$$

and the James-Stein estimator of β is

$$\hat{\beta}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{\|\hat{\beta}\|^2}\right) \hat{\beta} \quad \text{vs} \quad \hat{\beta}_\lambda = \frac{1}{1+\lambda} \hat{\beta}$$

Entries of $\hat{\beta}$ are shrank in both cases. However,

the JS procedures enforce a specific value of the shrinkage factor that depends on $\hat{\beta}$ itself, while the RR approach determines the optimal λ via other techniques such as cross-validation.

In the general case, when $\begin{aligned} \hat{\beta} &\sim \mathcal{N}(\beta, \sigma^2 (X^t X)^{-1}) \\ \beta &\sim \mathcal{N}(0, \tau I_d), \end{aligned}$

the JS-like estimator can be taken as

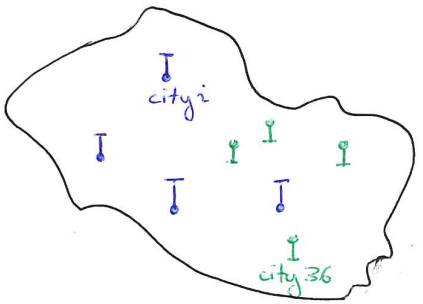
$$\tilde{\beta}_{JS} = \left(1 - \frac{(d-2)\sigma^4}{\hat{\beta}^t (X^t X) \hat{\beta}}\right) \hat{\beta},$$

see for example Bock (1975).

(v) Stein's paradox.

Ex: Toxoplasmosis, Efron & Morris (1977).

- ↳ disease of the blood endemic in Central America
- ↳ study in El Salvador on ~ 5000 people in 36 cities.
- ↳ interested in the incidence rate (w.r.t. the national average)
- ↳ incidence rates are ≈ normally distributed.



$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{36} \end{pmatrix},$$

where θ_i = incidence rate of city i with respect to the national average.

Frequentist estimator $\bar{X}$, whose i-th coordinate corresponds to the sample average for city i

Janes-Stein estimator, which shrinks the components of $\bar{X}$ by some factor.

Which one to choose?

If interested in one city in particular, then consider the MLE i.e. the observed rate for that city

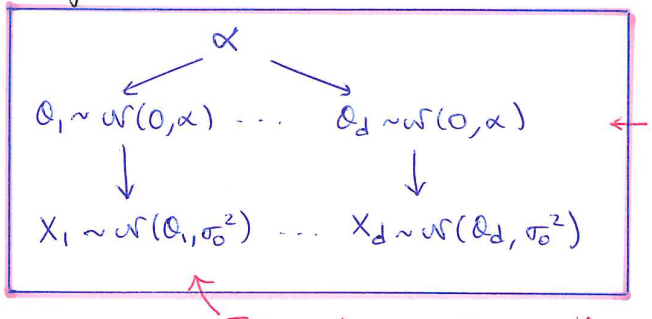
If interested in the precision at the national level, then select the JS estimator. For example, if the Ministry of Health wants to build local hospitals. The hospitals may end up at the wrong place or have the wrong size, but the sum of the mismatches will be smaller than for the observed rates.

There is a flaw in the reasoning. Suppose that you augment the vector of unknowns θ with an extra component θ_{37} corresponding to the incidence rate of a tumor in a group of rats. Including this extra coordinate may reduce the overall risk compared with the MLE. Combining two unrelated problems into one, and reducing the global risk, is at the origin at what is known as STEIN'S PARADOX.

↳ In particular, note that the shrinkage factor in the expression of the JS estimator depends on $\|X\|^2$, i.e. on all the components of $X = (X_1, \dots, X_d)^t$, despite the assumption that the X_i are independent ⇒ as if the incidence rate estimate of a tumor in a group of rats carries information for estimating the incidence rate of toxoplasmosis in El Salvador.

The Empirical Bayes formulation of the JS estimator brings light to the picture. The components of θ are assumed to be $\mathcal{N}(0, \alpha)$ distribution.

⇒ The common variance amongst the θ_i indicate that they are all similar to each other, up to some random fluctuations ⇒ θ_1 can learn from $(\theta_2, \dots, \theta_d)$ and X_1 together.



← sample of the population distrib $\sim \mathcal{N}(0, \alpha)$: the $\theta_1, \dots, \theta_d$ are connected via α.

This picture emphasizes the underlying structure / hierarchy in the approach.

↑ Works well when studying clusters of individuals. In particular, using the data to estimate the prior parameters is inherent to the Empirical Bayes (EB) approach, and is described further in the next section.

III.2. The Empirical Bayes Approach.

Empirical Bayes method are usually not considered "full Bayesian", since the prior parameters are estimated from the data. We illustrate the procedure in the binomial case with Beta prior:

$$X_i \mid p_i \sim Bi(x_i \mid n_i, p_i) \quad \text{iid}$$
$$p_i \mid \alpha, \beta \sim B(p_i \mid \alpha, \beta) \quad \text{iid}$$

EB uses point estimates $\hat{\alpha}, \hat{\beta}$ of the prior distribution, maximizes the MARGINAL LIKELIHOOD:

$$f(\mathcal{X}_n \mid \alpha, \beta) = f(x_1, \dots, x_n \mid \alpha, \beta)$$

Same idea as the frequentist Maximum likelihood estimation, except that maximization occurs on the prior parameters instead of the parameters themselves.

$$= \int f(x_1, \dots, x_n, p_1, \dots, p_n \mid \alpha, \beta)\, dp_1 \dots dp_n$$

(full joint distribution)

$$= \int f(x_1, \dots, x_n \mid p_1, \dots, p_n, \alpha, \beta) \times f(p_1, \dots, p_n \mid \alpha, \beta)\, dp_1 \dots dp_n$$

(independence)

$$= \prod_{i=1}^{n} \int f(x_i \mid p_i) f(p_i \mid \alpha, \beta)\, dp_i.$$

Focusing on a single term

$$\int_0^1 Bi(x \mid n, p)\, B(p \mid \alpha, \beta)\, dp = \text{The Beta-Binomial distribution}$$

$$= \dots = \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}$$

↖ the beta function (not distribution)

↳ Then consider $(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)}{\operatorname{argmax}}\ f(\mathcal{X}_n \mid \alpha, \beta)$

Or, alternatively, use the method of moments

EB violates the Bayesian principle that the prior should be chosen independently of the data. A "more Bayesian" approach, described in Section V, puts a prior distribution on (α, β).

- Robbin's formula, or Non Parametric Empirical Bayes

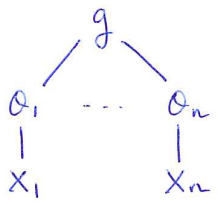
Consider the case of a Poisson distribution

$$\begin{cases} X_i \mid \Theta_i \sim \mathcal{P}(\Theta_i) \\ \Theta_i \sim g \end{cases}$$

where g is left unspecified.

We are interested in the posterior mean $\mathbb{E}(\Theta_i \mid X_i = x_i)$

We drop the subscript i for convenience:



$$\mathbb{E}(\Theta \mid X = x) = \frac{\int_0^\infty \Theta\, p(x \mid \Theta)\, g(\Theta)\, d\Theta}{\int_0^\infty p(x \mid \Theta)\, g(\Theta)\, d\Theta}$$

$=: f(x)$, the marginal likelihood

$$= \frac{\int_0^\infty e^{-\Theta} \frac{\Theta^{x_n}}{x!}\, g(\Theta)\, d\Theta}{\int_0^\infty e^{-\Theta} \frac{\Theta^{x}}{x!}\, g(\Theta)\, d\Theta} = (x+1)\, \frac{\int_0^\infty e^{-\Theta} \frac{\Theta^{x_n}}{(x+1)!}\, g(\Theta)\, d\Theta}{\int_0^\infty e^{-\Theta} \frac{\Theta^{x}}{x!}\, g(\Theta)\, d\Theta}$$

$$= (x+1)\, \frac{f(x+1)}{f(x)}$$

We estimate the marginal likelihood non-parametrically (instead of maximizing the marginal likelihood in a parametric context)

Let $\hat{f}(x+1) = \frac{1}{n}(\#\text{ observations} = x+1) =: \frac{n_{x+1}}{n}$

Then $\mathbb{E}(\Theta \mid X = x) \approx (x+1)\, \frac{n_{x+1}}{n_x}$.

ROBBIN'S FORMULA

IV.3. Bayesian False Discovery Rate

• Context: Multiple testing ; d tests, where each test assumes that H_0 has prior π_0 ; $X \sim f_0 \sim \mathcal{N}(0,1)$ on H_0
H_1 —"— $\pi_1 = 1-\pi_0$; $X \sim f_1$ "away from 0" on H_1

The cumulative distribution functions are denoted F_0 and F_1, respectively.

Then $X \sim f$ = mixture distribution ; $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$.

Let $F(\mathcal{X}) := \int_{\mathcal{X}} f(x)\,dx$ for some measurable set $\mathcal{X}$.

We consider the quantity

$$\phi(\mathcal{X}) := \mathbb{P}(H_0 \mid x \in \mathcal{X}) = \frac{\mathbb{P}(x \in \mathcal{X} \mid H_0)\,\pi_0}{\mathbb{P}(x \in \mathcal{X})} = \frac{F_0(\mathcal{X})\,\pi_0}{F(\mathcal{X})}$$

↖ aka the Bayesian False Discovery Rate ; also usually denoted $\mathrm{Fdr}(\mathcal{X})$.

[if we report $x \in \mathcal{X}$ as non-null, then $\phi(\mathcal{X})$ represents the probability of false-discovery].

× Remark = Typical application where X is a z-score i.e.

$z_i = \frac{\bar{X}}{s/\sqrt{n}}$ for the i-th test $\sim \mathcal{N}(0,1)$ under H_0.

• Empirical Bayes estimate of $\phi(\mathcal{X})$.

Assumptions :

↘ f_0 is known since it is the distribution of a statistic of interest under the null.

↘ f_1 is unlikely to be known.

↘ π_0 is unknown.

↘ F can be easily estimated with $\hat{F}(\mathcal{X}) = \frac{1}{d}\#\{X_i \in \mathcal{X}\}$.

⇒ We get the estimator $\hat{\phi}(\mathcal{X}) = \frac{F_0(\mathcal{X})\,\pi_0}{\hat{F}(\mathcal{X})}$

• How good of an estimator of ϕ is $\hat{\phi}$?

Let

$n_0(\mathcal{X})$ = # of null X_i falling into $\mathcal{X}$ (unobservable)

$n_1(\mathcal{X})$ = # of non-null X_i falling into $\mathcal{X}$ (unobservable)

$n_+(\mathcal{X}) = n_0(\mathcal{X}) + n_1(\mathcal{X})$ (observable)

Then

$$e_0(\mathcal{X}) := \mathbb{E}\,n_0(\mathcal{X}) = \mathbb{E}\sum_{i=1}^{d} \mathbb{1}(X_i \in \mathcal{X} \mid X_i \in H_0)\,\mathbb{1}(X_i \in H_0) = d\,\pi_0\,F_0(\mathcal{X})$$

$$e_1(\mathcal{X}) := \mathbb{E}\,n_1(\mathcal{X}) = d\,\pi_1\,F_1(\mathcal{X})$$

$$e_+(\mathcal{X}) := \mathbb{E}\,n_+(\mathcal{X}) = d\,F(\mathcal{X}).$$

With this notation,

$$\phi(\mathcal{X}) = \frac{F_0(\mathcal{X})\,\pi_0}{F(\mathcal{X})} = \frac{e_0(\mathcal{X})}{e_+(\mathcal{X})}$$

$$\hat{\phi}(\mathcal{X}) = \frac{\mathbb{E}\,n_0(\mathcal{X})}{n_+(\mathcal{X})} = \frac{e_0(\mathcal{X})}{n_+(\mathcal{X})}$$

In addition, we consider

$$\mathrm{FdP}(\mathcal{X}) := \frac{n_0(\mathcal{X})}{n_+(\mathcal{X})} = \text{proportion of } X_i \text{ from the null in } \mathcal{X}$$

Assume that $e_0(x) = \mathbb{E}(n_0(x) \mid n_1(x))$; which holds when $n_0(x) \perp n_1(x)$. Then

$$\boxed{\mathbb{E}(\mathrm{FdP}(x) \mid n_1(x)) \leq \mathbb{E}(\hat{\phi}(x) \mid n_1(x))}$$

"For any value of $n_1(x)$, the EB estimate $\hat{\phi}(x)$ is a conservatively biased estimate of the actual $\mathrm{FdP}(x)$".

- proof: $\mathrm{FdP}(x) = \dfrac{n_0(x)}{n_0(x)+n_1(x)} = \varphi(n_0(x)) - \varphi = \text{concave}$

Jensen inequality $\Rightarrow$

$$\mathbb{E}(\mathrm{FdP}(x) \mid n_1(x)) \leq \frac{e_0(x)}{e_0(x)+n_1(x)}.$$

On the other hand,

$$\hat{\phi}(x) = \frac{e_0(x)}{n_0(x)+n_1(x)} = \tilde{\varphi}(n_0(x)) - \tilde{\varphi} = \text{convex}$$

Applying Jensen again:

$$\mathbb{E}(\hat{\phi}(x) \mid n_1(x)) \geq \frac{e_0(x)}{e_0(x)+n_1(x)}.$$

Combining the two inequalities yields the desired result. ∎

- Mean and Variance of $\hat{\phi}(x)$

Suppose that the $X_1, \dots, X_d$ are independent. Then $n_+(x) \sim \mathrm{Bi}(d, F(x))$.

$$\hat{\phi}(x) = \frac{e_0(x)}{n_+(x)} = \frac{e_0(x)}{e_+(x)} \times \frac{1}{1 + \dfrac{n_+(x)-e_+(x)}{e_+(x)}}$$

$$\hat{\phi}(x) \approx \frac{e_0(x)}{e_+(x)}\left[1 - \frac{n_+(x)-e_+(x)}{e_+(x)} + \left(\frac{n_+(x)-e_+(x)}{e_+(x)}\right)^2 \cdots\right],$$

where $\dfrac{n_+(x)-e_+(x)}{e_+(x)}$ has mean 0 and variance

$$\gamma(x) := \frac{\mathrm{var}\, n_+(x)}{e_+^2(x)}.$$

It follows that $\dfrac{\hat{\phi}(x)}{\phi(x)}$ has approximatively mean $1+\gamma(x)$ and variance $\gamma(x)$.

Under the independence assumption, $\mathrm{var}\, n_+(x) = d\,F(x)(1-F(x))$

$$\Rightarrow \gamma(x) = \frac{d\,F(x)\,(1-F(x))}{d^2 F^2(x)} = \frac{1-F(x)}{dF(x)} = \frac{1-F(x)}{e_+(x)} \approx \frac{1}{e_+(x)}$$

since we are mostly interested in regions x where $F(x)$ is small.

$\Rightarrow$ $\dfrac{\hat{\phi}(x)}{\phi(x)}$ has mean $\approx 1 + \dfrac{1}{e_+(x)}$ & variance $\approx \dfrac{1}{e_+(x)}$

These expressions allow us to evaluate if $\hat{\phi}$ is a biased estimator of ϕ, and to construct approx confidence bounds for ϕ.

The accuracy of $\hat{\phi}(x)$ heavily depends on $e_+(x)$ = the expected number of observations falling into x

* Illustration = $\mathcal{X}$ = tail interval of the form $(-\infty, x)$ or $(x, +\infty)$.

Take e.g. $\mathcal{X} = (3, +\infty)$, and suppose that $n_+(\mathcal{X}) = 50$.

Then $\hat{\phi}(\mathcal{X}) = \dfrac{d\,(\pi_0)\,(1 - \Phi(3))}{n_+(\mathcal{X})} = 0.1$ (say)

take $\pi_0 = 1$ to get an upper bound on $\hat{\phi}(\mathcal{X})$ (conservative approach)

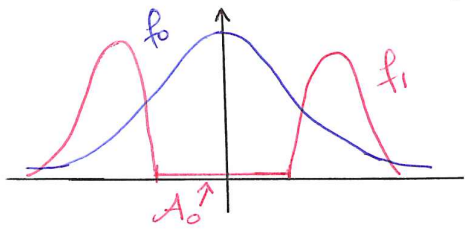
meaning that $\approx 10\%$ of the 50 discoveries are false discoveries.

* Remarks (i) This approach is particularly well suited when testing thousands of hypothesis where many are expected to have zero effect ($\pi_0 \approx 1$).
Ex: effect of a treatment on gene expression.

(ii) Estimation of π_0.

π_0 can be estimated under additional assumptions on f_1.

Suppose that f_1 is zero for some subset A_0 around 0 ("separability" condition)



Then $n_+(A_0) = \#\{X_i \in A_0\}$

and $\mathbb{E}\, n_+(A_0) = \pi_0\, d\, F_0(A)$, which yields

$$\hat{\pi}_0 = \frac{n_+(A_0)}{d \times F_0(A_0)}$$

← You may vary the size of A_0 and check visually the stability of $\hat{\pi}_0$.

(iii) Relation to the JS estimator

The generative process $X \sim f = \pi_0 f_0 + \pi_1 f_1$ can be rewritten

$$(*)\quad \begin{cases} X \mid \mu \sim \mathcal{N}(\mu, 1) \\ \mu \sim \pi_0\, \delta_0(\mu) + (1 - \pi_0)\, g_1(\mu), \end{cases}$$

leading to $f_1(x) = \int \varphi(x - \mu)\, g_1(\mu)\, d\mu$.

In comparison with the JS estimator, which is constructed from the model

$$(**)\quad \begin{cases} X \mid \mu \sim \mathcal{N}(\mu, 1) \\ \mu \sim \mathcal{N}(0, \alpha) \end{cases}$$

"learning from others" (p. 65)

Expression $(*)$ and its comparison with the smooth version $(**)$ highlights that each test is learning from one another to control the FDR.

- JS → performs well even when d is not too large
- Bayesian FDR → the variance of $\hat{\phi}(\mathcal{X})$ is of order $1/e_+(\mathcal{X})$ so in comparison d needs to be quite large to achieve an accurate estimation of $\phi(\mathcal{X})$.

(iv) Link with the Benjamini-Hochberg (BH) procedure

The BH procedure was introduced p. 29-33 in the Chapter MS: HYPOTHESIS TESTING.

Consider d null hypotheses $H_{0,1}, \dots, H_{0,d}$, with z-scores $z_1, \dots, z_d$ and p-values $p_1, \dots, p_d$.

Let $p_{(1)} \leq \dots \leq p_{(d)}$ be the ordered p-values, and $H_{0,(i)}$ be the null hypothesis associated with $p_{(i)}$.

Let k be the largest i for which $p_{(i)} \leq \frac{i\alpha}{d}$.

* BH: reject all $H_{(i)}$ for $i = 1, \dots, k$.
* Guarantee: if the $p_1, \dots, p_d$ are independent, then

$$FDR = \mathbb{E}\left\{ \frac{FP}{FP+TP} \mathbb{1}(FP+TP \geq 1) \right\} \leq \alpha.$$

Suppose that the p_i correspond to the left p-values, so that $p_i = F_0(z_i)$. Then $p_{(i)} = F_0(z_{(i)})$.

On the other hand, ordering $z_{(1)} \leq \dots \leq z_{(d)}$ yields $\hat{F}(z_{(i)}) = \frac{i}{d}$.

Then

$$p_{(i)} \leq \frac{i\alpha}{d} \Leftrightarrow \pi_0 \frac{F_0(z_{(i)})}{\hat{F}(z_{(i)})} \leq \pi_0 \alpha \leq \alpha$$

[BH]

$$\Leftrightarrow \Phi((-\infty, z_{(i)})) \leq \alpha$$

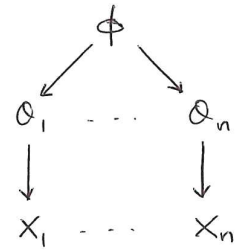
The BH criterion can be equivalently re-expressed in terms of the Bayesian FDR.

(Major) Consequence = For the first time α has the intuitive meaning that a rejected hypothesis $H_{0,i}$ is actually correct (i.e. a Bayesian meaning).

V - HIERARCHICAL MODELS

While Empirical Bayes considers point estimates for the prior's parameters, Hierarchichal Models put a prior on them

• Empirical Bayes



$\rightarrow \Phi$ replaced with a point estimate

$\theta_1, \dots, \theta_n \sim f(\theta|\Phi)$

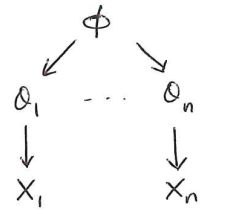
$X_1, \dots, X_n \sim f(x|\theta)$

$\hat{\Phi} = \text{argmax}\, f(\mathcal{L}_n|\Phi)$

$f(\mathcal{L}_n|\Phi) = \int f(\mathcal{L}_n|\theta) f(\theta|\Phi) d\theta$

= marginal likelihood.

• Hierarchical Models



$\Phi \sim f(\Phi)$ - hyperprior

$\theta_1, \dots, \theta_n \sim f(\theta|\Phi)$ - prior

$X_1, \dots, X_n \sim f(x|\theta)$ - likelihood

(i) How to select the hyperprior $f(\Phi)$?

(ii) How to (compute) sample from the joint posterior $(\theta, \Phi)|\mathcal{L}_n$ and the posterior predictive distribution?

A general procedure :
(1) Sample ϕ from $\phi \mid \mathcal{L}_n$
(2) Sample θ from $\theta \mid \phi, \mathcal{L}_n$
(3) Sample a new X from $X \mid \theta, \mathcal{L}_n$

↳ Step (3) is easy

Note that step (1) is feasible when $f(\phi \mid \mathcal{L}_n)$ is known.

$$f(\phi \mid \mathcal{L}_n) = \frac{f(\theta, \phi \mid \mathcal{L}_n)}{f(\theta \mid \phi, \mathcal{L}_n)}$$

⇒ We need to compute $f(\theta \mid \phi, \mathcal{L}_n)$ and $f(\theta, \phi \mid \mathcal{L}_n)$.

this posterior distribution can easily be computed analytically for conjugate models :

Also useful for step 2 ↳ $f(\theta \mid \phi, \mathcal{L}_n) \propto f(\mathcal{L}_n \mid \phi, \theta) f(\theta \mid \phi)$

posterior — likelihood — prior

$$f(\theta, \phi \mid \mathcal{L}_n) \propto f(\mathcal{L}_n \mid \phi, \theta) f(\theta \mid \phi) f(\phi)$$

common terms (need to be computed only once)

We illustrate the procedure for the Beta - Binomial model.

- $X_1, \dots, X_n$; $X_i \mid \theta_i \sim Bi(n_i, \theta_i)$
- $\theta_i \sim B(\alpha, \beta)$
- $(\alpha, \beta) \sim$ some distribution $f(\alpha, \beta)$.

step (2) $f(\theta \mid \mathcal{L}_n, \alpha, \beta) \propto f(\mathcal{L}_n \mid \theta, \alpha, \beta) f(\theta \mid \alpha, \beta)$

Assuming the θ_i are iid generated ↗

$$= \prod_{i=1}^n Bi(x_i \mid n_i, \theta_i) \prod_{i=1}^n B(\theta_i \mid \alpha, \beta)$$

$$= \prod_{i=1}^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1} (1-\theta_i)^{\beta-1} \binom{n_i}{x_i} \theta_i^{x_i} (1-\theta_i)^{n_i - x_i}$$

$$= \prod_{i=1}^n B(\theta_i \mid \alpha + x_i, \beta + n_i - x_i)$$

The joint posterior is

$$f(\theta, \alpha, \beta \mid \mathcal{L}_n) \propto f(\alpha, \beta) \prod_{i=1}^n Bi(x_i \mid n_i, \theta_i) \prod_{i=1}^n B(\theta_i \mid \alpha, \beta)$$

The ratio of these two terms allow us to derive step (1) :

$$f(\alpha, \beta \mid \mathcal{L}_n) \propto f(\alpha, \beta) \prod_{i=1}^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)}{\Gamma(\alpha+\beta+n_i)}$$

After simplifications.

$f(\alpha, \beta \mid \mathcal{L}_n)$ is usually computed numerically & normalized to get a proper distribution.

Remains the choice of the prior $f(\alpha, \beta)$.

Gelman et al argue that the prior $f(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$ yield a proper posterior distribution $f(\alpha, \beta \mid \mathcal{L}_n)$ as long as $0 < x_i < n_i$ for at least one $i = 1, \dots, n$.

V. 1. Application to multiple testing

Hierarchical models prove very useful when performing multiple tests. It complements alternative methods based on the Bayesian FDR which are more adapted to cases where the number of tests is large, and the number of nulls is also believed to be large.

The ideas developed here have been succesfully applied in social sciences application, or psychology. See for example Gelman, Hill and Yajima (2012).

* 1 group

Suppose we are interested in estimating the treatment effect in a single group of population, divided in control and treatment sub-groups.

The variable of interest is denoted Y, and may be modeled using $Y = \mu + \delta \mathbb{1}(\text{Treatment}) + \varepsilon$,

$\varepsilon \sim \mathcal{N}(0, \sigma_Y^2)$

δ represents the treatment effect

× J groups

$Y = \mu_Y + \delta_1 \mathbb{1}(\text{Treatment}) + \varepsilon$ — group 1

$Y = \mu_Y + \delta_2 \mathbb{1}(\text{Treatment}) + \varepsilon$ — group 2

↑ .../... $Y = \mu_Y + \delta_J \mathbb{1}(\text{Treatment}) + \varepsilon$ — group J

The frequentist would test for each $\delta_j = 0$ separately, and probably adjusting the thresholds for rejecting the nulls accordingly (e.g. Bonferroni).

The Bayesian would take a different path, ...

$Y = \mu_Y + \delta_1 \mathbb{1}(\text{Treatment}) + \varepsilon$

$Y = \mu_Y + \delta_2 \mathbb{1}(\text{Treatment}) + \varepsilon$

... / ... $Y = \mu_Y + \delta_J \mathbb{1}(\text{Treatment}) + \varepsilon$

$\delta \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2)$

... assuming the existence of a common underlying distribution for the δ :

$\delta_1, \dots, \delta_J$ indpt $\sim \mathcal{N}(\mu_\delta, \sigma_\delta^2)$.

This approach allows for some heterogeneity accross the groups; while allowing each group to learn from one another.

The procedure is commonly referred to as "partial pooling".

• Effect on the z-scores.

We derive some algebra in a simple case, highlighting the main difference between the frequentist and the Bayesian approach.

In a pairwise set-up, each value Y is itself a difference between paired values (Y_1, Y_2) i.e. $Y = Y_1 - Y_2$; and we are testing for the departure of the mean value from 0. We can write $Y = \delta_i + \varepsilon$ for group i, and we are testing for $\delta_i = 0$.

Then $\begin{cases} \underline{Y} = (Y_1, \dots, Y_d)^t \sim \mathcal{N}(\underline{\delta}, \sigma_Y^2 \underline{\underline{I}}_d) \\ \underline{\delta} = (\delta_1, \dots, \delta_d)^t \sim \mathcal{N}(\underline{\mu}, \sigma_\delta^2 \underline{\underline{I}}_d) \end{cases}$

The posterior distribution of $\delta \mid Y_1, \dots, Y_n$ is normal with mean μ_n and covariance matrix Σ_n, ($\mathcal{L}_n$)

- $\mu_n = \mathbb{E}(\delta \mid \mathcal{L}_n) = \dfrac{\frac{\mu}{\sigma_\delta^2} + \frac{n}{\sigma_Y^2}\bar{Y}}{\frac{1}{\sigma_\delta^2} + \frac{n}{\sigma_Y^2}}$ ← $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$

- $\Sigma_n^{-1} = \left(\frac{1}{\sigma_\delta^2} + \frac{n}{\sigma_Y^2}\right) \underline{\underline{I}}_d \quad \Rightarrow \quad \mathrm{var}(\delta \mid \mathcal{L}_n) = \left(\frac{1}{\sigma_\delta^2} + \frac{n}{\sigma_Y^2}\right)^{-1}$.

In such a set-up, we are also interesting in comparing groups two by two, evaluating the pairwise differences $\mathbb{E}(\delta_j - \delta_k \mid \mathcal{L}_n)$. Straightforward calculations show that

$$\mathbb{E}(\delta_j - \delta_k \mid \mathcal{L}_n) = \frac{n/\sigma_Y^2}{1/\sigma_\delta^2 + n/\sigma_Y^2}(\bar{Y}_j - \bar{Y}_k)$$

$$\mathrm{var}(\delta_j - \delta_k \mid \mathcal{L}_n) = \frac{2\sigma_\delta^2 \frac{\sigma_Y^2}{n}}{\sigma_\delta^2 + \sigma_Y^2/n}.$$

⇓

The posterior z-score is

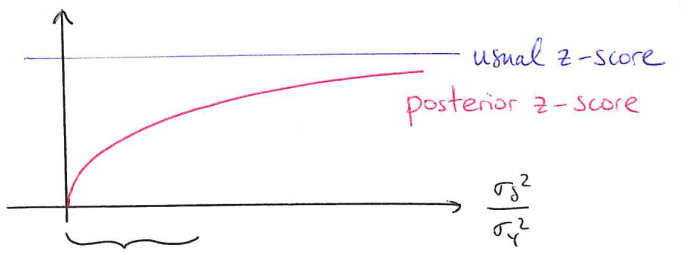
$$\frac{\mathbb{E}(\delta_j - \delta_k \mid \mathcal{L}_n)}{\sqrt{\mathrm{var}(\delta_j - \delta_k \mid \mathcal{L}_n)}} = \frac{\bar{Y}_j - \bar{Y}_k}{\sqrt{2}\sqrt{\sigma_Y^2/n}} \times \frac{1}{\sqrt{\left(1 + \frac{\sigma_Y^2/n}{\sigma_\delta^2}\right)}}$$

The posterior means are pulled together

The usual z-score

shrinkage factor < 1 (correction from partial pooling)

[Plot]

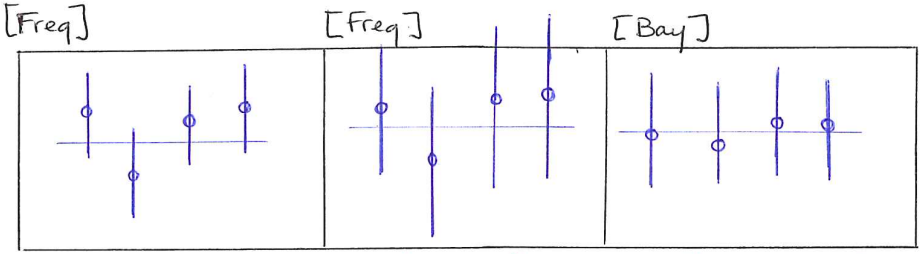


when $\sigma_\delta^2 \ll \sigma_Y^2$, the groups are homogeneous ⇒ large shrinkage.

Conclusion: Frequentist vs Bayesian

- z-scores untouched
- rejection region adjusted

- z-scores shrinked towards 0
- no adjustment of the critical region (or coverage of the credible bands).



↑ Credible intervals are larger

↑ Point Estimates are pulled closer together.

References.

x Baranchick, A. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report No 51, Department of Statistics, Standford University.

x Baranchick, A. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. Ann. Math. Stat., 41, p. 642 - 645.

x Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference. JRSS B, vol 41, no 2, p. 113-147.

x Bock, M. E. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. Ann. of Statistics, vol 3, no 1, pp. 209-218.

x Brown, L. D. (1977). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. Ann Math. Stats, vol 42, pp. 855 - 903.

x Casella G. & Berger R. L. (1987). Reconciling Bayesian and Frequentist Evidence in the One - Sided Testing Problem. JASA, vol 82, no 397, pp. 106 - 111.

x Edwards W., Lindman H. & Savage L. J. (1963). Bayesian Statistical Inference for Psychological Research. Psychological Review, vol 70, no 3, pp. 193 - 242.

x Efron B. (2010). Large Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction. Lecture Notes. Standford University.

x Efron B, & Morris C. (1977). Stein's Paradox in Statistics. Scientific American, vol 236, no 5, pp. 119 - 127.

x Gelman A. et. al (2014). Bayesian Data Analysis, 3rd Edition. CRC - Press.

x Gelman A., Hill J. & Yajima M. (2012). Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness. Vol 5, Issue 2, pp 189 - 211

x Gönen M., Johnson W. O, Lu Y. & Westfall P. (2005). The Bayesian two-sample t-test. The American Statistician. vol 59, pp 252-257.

x James W. & Stein C. (1961). Estimation with quadratic loss, Proc. 4th Berkeley Symp. Math. Stat. Prob, vol 1, pp. 361 - 379.

x Jeffreys H. (1961). Theory of Probability, 3rd Ed. Oxford Classic Texts in the Physical Sciences.

x Johnson V. & Rossell D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. JRSS B. vol 72, Part 2, pp. 143 - 170.

x Kass R. E. & Vaidyanathan (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of two Binomial Proportions. JRSS B, vol 54, no 1, pp. 129 - 144.

x Kleijn B. J. K. and van der Vaart A. W. (2012). The Bernstein-Von-Mises Theorem under misspecification. Electronic J. Stat, vol 6, pp. 354 - 381.

x Rouder J.N (2014). Optional Stopping: No Problem for Bayesians. Psychonomic Bulletin & Review, vol 21, pp. 301-308.

x Schönbrodt et al (2017). Sequential Hypothesis Testing with Bayes Factors = Efficiently Testing Mean Differences, vol 22, no 2, pp. 322-339. in Psychol. Methods.

x Sellke T., Bayarri M.J. & Berger J.O. (2001). Calibration of p-values for Testing Precise Null Hypotheses. The American Statistician, vol 55 pp. 62-71.

x Shi H. & Yin G. (2019). Control of Type I. Error Rates in Bayesian Sequential Design. Bayesian Analysis, vol 14, no 2, pp. 399 - 425.

x Van der Vaart (1998). Asymptotic Statistics. Cambridge University Press.

x White H. (1982). Maximum Likelihood Estimation of Misspecified Models, Econometrica, vol 50, no 1, pp. 1-25.