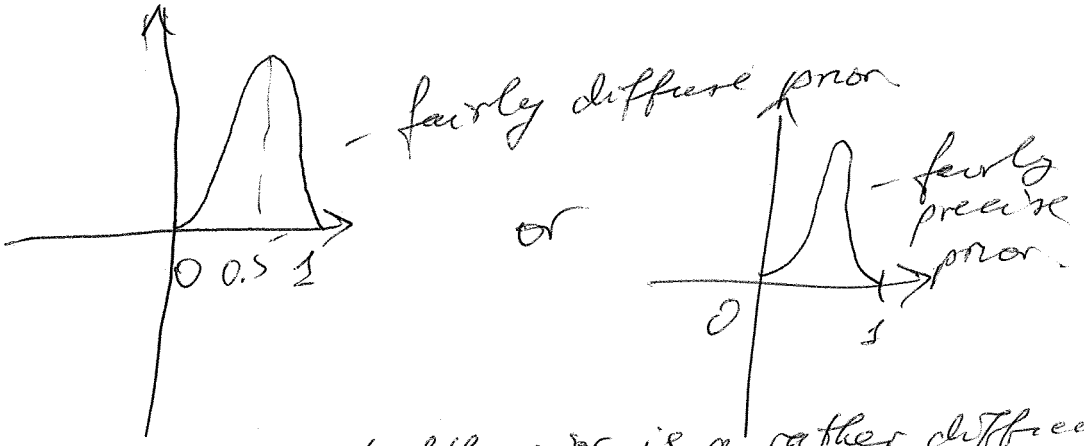
7.1 Prior and Posterior Distributions

Model $\{f_\theta : \theta \in \Omega\}$ for the data $s \in S$ and the prior probability measure Π for θ.

Ex If $\Omega = [0,1]$, θ is the prob. of success in a coin toss, we may have a prior like



Assignment of the prior is a rather difficult problem. Thus, the full set of ingredients in the Bayesian approach are: 1) a distribution for θ - prior Π

2) a set of conditional distributions for the data given θ, i.e. $\{f_\theta | \theta \in \Omega\}$. $\Rightarrow$ the joint probability distr. for (s, θ) is $\pi(\theta) f_\theta(s)$. If the prior distribution is absolutely continuous, the marginal distr. for s is $m(s) = \int \pi(\theta) f_\theta(s) d\theta$ - this is the prior predictive distribution of Ω the data.

Next, posterior distr. of θ is described by

$$\pi(\theta|s) = \frac{\pi(\theta) f_\theta(s)}{m(s)},$$ this is an axiom based on the conditional probability principle, not a theorem!!

$m(s)$ is the inverse normalizing constant for the posterior density since $\pi(\theta|s) \propto \pi(\theta) f_\theta(s)$. In many examples, $m(s)$ need not be computed because we recognize the functional form as a function of θ.

Ex. 7.1.1 Bernoulli model

Let $(X_1, \dots, X_n) \sim \text{Ber}(\theta)$, $\theta \in [0,1]$;

Let $\pi = \text{Beta}(\alpha, \beta)$; then, the posterior of θ is $\sim$ to the likelihood $\prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{n\bar{x}} (1-\theta)^{n(1-\bar{x})}$

multiplied by the prior $B^{-1}(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$\Rightarrow$ The product is proportional to

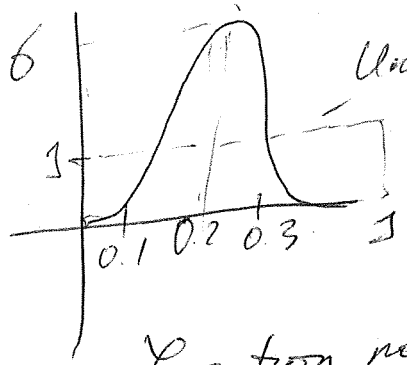
$$\theta^{n\bar{x}+\alpha-1}(1-\theta)^{n(1-\bar{x})+\beta-1}$$

- $\text{Beta}(n\bar{x}+\alpha, n(1-\bar{x})+\beta)$, no need to compute

$m(x_1,\dots,x_n)$. Let us say, we observe $n\bar{x}=10$ in a sample of $n=40$, and $\alpha=\beta=1 \Rightarrow$ uniform prior on θ. Then, the posterior of θ is given by $\text{Beta}(11,31)$.

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

from here comes Beta(α,β) distribution



Ex. 7.1.2 Location normal model

$(x_1,\dots,x_n)\sim N(\mu,\sigma_0^2)$, σ_0^2 is known

$$L(\mu/x_1,\dots,x_n)=\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2\right)$$

Let the prior μ be $N(\mu_0,\tau_0^2)$ for some specified μ_0 and τ_0^2. The posterior density of the μ is, then,

$$e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2}=e^{-\frac{\mu_0^2}{2\tau_0^2}-\frac{n\bar{x}^2}{2\sigma_0^2}}\times$$

$$e^{-\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)\left[\mu^2-2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)\mu\right]}=$$

$$\times e^{-\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)\left(\mu-\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)\right)^2}\times$$

$$=e^{\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)^2-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{n\bar{x}^2}{\sigma_0^2}\right)}\times$$

$\times e^{\dots}$

- this is $\propto N\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right),\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\right)$

the posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\bar{x}$,

weights $\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1} \frac{1}{\tau_0^2}$ and $\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1} \frac{n}{\sigma_0^2}$

Also, the posterior variance is smaller than the variance of the sample mean. The more diffuse the prior is, (<=> the larger τ_0^2 is) — the less influence the prior has. E.g. if $n = 20$, $\sigma_0^2 = 1$, $\tau_0^2 = 1$, the ratio of the posterior variance to the sample mean variance is $\frac{20}{21} \approx 0.95$

Ex 7.13 Multinomial model.

A categorical response x that takes k values, $x \in S = \{1, 2, \dots, k\}$. Say, a bowl contains chips labelled $1, 2, \dots, k$. A proportion θ_i of the chips are labelled i. We randomly draw a chip, observing its label. With unknown θ_i, we have

$\{P(\theta_1, \dots, \theta_k) : (\theta_1, \dots, \theta_k) \in \Omega\}$ where $P(x = i) = \theta_i$.

and $\Omega = \{(\theta_1, \dots, \theta_k) : 0 \le \theta_i \le 1, i = 1, \dots, k, \sum_{i=1}^{k} \theta_i = 1\}$

A sample $(s_1, \dots, s_n)$ is observed; let the ith category count be x_i; $\Rightarrow L(\theta_1, \dots, \theta_k | s_1, \dots, s_n) = \theta_1^{x_1} \theta_2^{x_2} \dots \theta_k^{x_k}$

The prior: $(\theta_1, \dots, \theta_{k-1}) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k)$

with density $\frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}$,

$\alpha_i > 0$ are constants that represent the prior belief about $(\theta_1, \theta_2, \dots, \theta_k)$. The choice $\alpha_1 = \alpha_2 = \dots = \alpha_k = 1$ — uniform distribution.

$\Rightarrow$ The posterior density of $(\theta_1, \dots, \theta_{k-1})$ is

$\propto \theta_1^{x_1 + \alpha_1 - 1} \theta_2^{x_2 + \alpha_2 - 1} \dots \theta_k^{x_k + \alpha_k - 1} \sim \text{Dirichlet}(x_1 + \alpha_1, \dots, x_k + \alpha_k)$

Location - Scale Normal Model.

Now, $(x_1, \dots, x_n) \sim N(\mu, \sigma^2)$: $\quad -\frac{n}{2\sigma^2}(\bar{x}-\mu)^2 - \frac{n-1}{2\sigma^2}s^2$

$\mathcal{L}(\mu, \sigma^2 / x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2}\, e$

1. $\mu/\sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2)$,

2. $\frac{1}{\sigma^2} \sim \text{Gamma}(\alpha_0, \beta_0)$ - inverse Gamma -

From here (no details):

$$\mu(\sigma^2, x_1, \dots, x_n \sim N\left(\mu_x, \left(n+\frac{1}{\tau_0^2}\right)^{-1}\sigma^2\right)$$

and $\frac{1}{\sigma^2} / x_1, \dots, x_n \sim \text{Gamma}\left(\alpha_0 + n/2, \beta_x\right)$

where $\mu_x = \left(n+\frac{1}{\tau_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2} + n\bar{x}\right)$

and $\beta_x = \beta_0 + \frac{n-1}{2}s^2 + \frac{1}{2}\frac{n(\bar{x}-\mu_0)^2}{1+n\tau_0^2}$

As $\tau_0 \to \infty$ (the prior is increasingly diffuse)

the conditional posterior distribution of μ given σ^2

$\xrightarrow{D} N\left(\bar{x}, \frac{\sigma^2}{n}\right)$ because $\mu_x \to \bar{x}$ and $\left(n+\frac{1}{\tau_0^2}\right)^{-1} \to \frac{1}{n}$

Furthermore, as $\tau_0 \to \infty$ and $\beta_0 \to 0$, the marginal posterior of $\frac{1}{\sigma^2} \xrightarrow{D} \text{Gamma}\left(\alpha_0 + \frac{n}{2}, \frac{(n-1)s^2}{2}\right)$ because $\beta_x \to (n-1)s^2/2$.

Note that there's no proper prior distribution as $\tau_0 \to \infty$ and $\beta_0 \to 0$

Corresponds to the diffuse cap

where for $\text{Gamma}(\alpha_0, \beta_0)$ $\quad \text{Var} = \frac{\alpha_0}{\beta_0^2} \xrightarrow{\beta_0 \to 0} \infty$

7.2. Inferences based on the posterior distribution.

7.2.1. Estimation

To estimate $\psi(\theta)$, two most common approaches are: 1) to compute the posterior density of $\psi(\theta)$ and use the posterior mode $\hat{\psi}$. To do so, one needs to maximize $w(\psi|s)$, or equivalently $m(s)\,w(\psi|s)$ (so as not to compute the normalizing constant), as a function of ψ. Moreover, in general, any 1-1 increasing function of $\psi(\cdot|s)$ can be maximized to yield the same result.

2) Another common alternative is to use the posterior mean $\mathbb{E}(\psi(\theta)|s)$. When the posterior distribution is symmetric and the expectation is finite, this is the same as the mode.

Ex. 7.2.2. Bernoulli model - $X_1, \dots, X_n \sim \text{Ber}(\theta)$, $\theta \in [0,1]$ unknown; $\theta \sim \text{Beta}(\alpha, \beta)$

The posterior distribution of θ is $\text{Beta}(n\bar{x}+\alpha,\ n(1-\bar{x})+\beta)$

Then $\mathbb{E}(\theta|x_1, \dots, x_n) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \int_0^1 \theta^{n\bar{x}+\alpha}(1-\theta)^{n(1-\bar{x})+\beta-1}\,d\theta =$

$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \cdot \frac{\Gamma(n\bar{x}+\alpha+1)\Gamma(n(1-\bar{x})+\beta)}{\Gamma(n+\alpha+\beta+1)} =$

$= \frac{n\bar{x}+\alpha}{n+\alpha+\beta}$; when $\alpha=\beta=1$ (uniform prior), the posterior expectation is $\mathbb{E}(\theta|x_1, \dots, x_n) = \frac{n\bar{x}+1}{n+2}$.

To determine the posterior mode, we have to maximize $\ln \theta^{n\bar{x}+\alpha-1}(1-\theta)^{n(1-\bar{x})+\beta-1}$; direct approach yields

$\hat{\theta} = \frac{n\bar{x}+\alpha-1}{n+\alpha+\beta-2}$ which is negative whenever $\alpha \geq 1$, $\beta \geq 1$.

This last restriction implies that the mode of the prior is in $(0,1)$ and not $=0$ or 1. When $\alpha=\beta=1$, the mode is $\hat{\theta}=\bar{x}$ — the same as MLE. When n is large, both are approx. the same and close to the MLE.

7.2.3 Location normal model

Now, $X_1, \dots = X_n \sim N(\mu, \sigma_0^2)$, μ is unknown, σ_0^2 is known.

The prior on mean μ is $N(\mu_0, \tau_0^2)$

The posterior distribution of μ is

$$N\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma^2}\bar{x}\right), \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\right)$$

It is symmetric about its mode, so the posterior mode and the mean are the same: $\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)$ – the weighted average of the prior mean and the sample mean.

Remarks: 1) When $n \to \infty$, we get $\bar{x}$ – MLE

2) When $\sigma_0^2 \to \infty$ (diffuse prior), get $\bar{x}$ – MLE again

3) The ratio of the sampling variance of $\bar{x}$ to the posterior variance of μ is $\frac{\sigma_0^2}{n}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right) = 1 + \frac{\sigma_0^2}{n\tau_0^2} > 1$.

The closer $\frac{\sigma_0^2}{n}$ is to 0, the larger this ratio is

4) As $\tau_0^2 \to 0$, posterior mean $\to \mu_0$

Numerical example: Average weight μ of adult cows in a large herd. Farmer weighs 10 randomly chosen cows from the herd; $\bar{Y} = 410$ kg / 925 lbs; let $\sigma_0 = 20$ kg and the prior distribution is normal $\Rightarrow \bar{Y}$ = MLE. Prior belief of the farmer: the weight should be ≈ 420 kg and unlikely to differ from it by more than 30 kg $\to$ Prior $N(420, 10^2)$ due to the empirical three sigma rule; the posterior Bayes is then $\frac{100}{400/10+400}410 + \frac{400/10}{400/10+100}420 = 412.9$ kg

$\to$ the posterior mean is $\frac{\mu_0}{\tau_0^2}\frac{\tau_0^2\sigma_0^2}{\sigma_0^2+n\tau_0^2} + \frac{n}{\sigma_0^2}\frac{\tau_0^2\sigma_0^2}{\sigma_0^2+n\tau_0^2}\bar{y} =$

$= \mu_0 * \frac{\sigma_0^2/n}{\sigma_0^2/n+\tau_0^2} + \frac{\tau_0^2}{\sigma_0^2/n+\tau_0^2}\bar{y}$; here $\tau_0^2 = 10^2 = 100$, $\sigma_0^2 = 20^2 = 400$, $n = 10$