Meeting 5:
Bayesian inference -Continuous probability models – Conjugate families

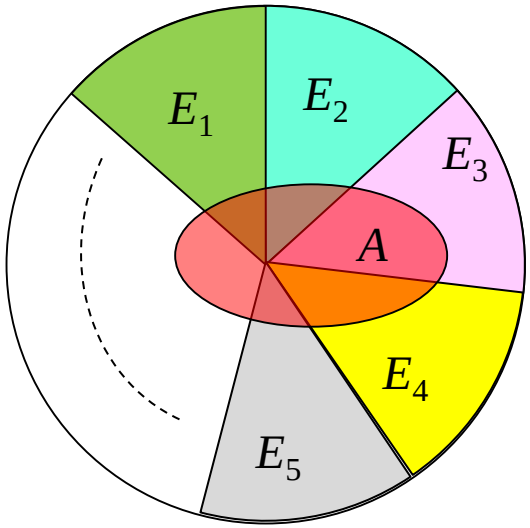
Bayes' theorem as a general probabilistic "rule"

Discrete events:

Let $E_1,\ldots,E_M$ be mutually exclusive events with $\bigcup_{j=1}^{M} E_j = S$ (sample space)

Then for any event A in the sample space

$$P(E_i|A) = \frac{P(A|E_i)\cdot P(E_i)}{\sum_{i=1}^{M} P(A|E_j)\cdot P(E_j)} \propto P(A|E_i)\cdot P(E_i)$$



Conditional probability density functions:

$$f_{\tilde{y}|\tilde{x}=x}(y|x) = \frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)} = \frac{f_{\tilde{x}|\tilde{y}}(x|y)\cdot f_{\tilde{y}}(y)}{f_{\tilde{x}}(x)}$$

Interpretation:

For every value x (of $\tilde{x}$)there is a conditional distribution of $\tilde{y}$

For each value y given the value x the conditional density of $\tilde{y}$ can be expressed in terms of the conditional density of $\tilde{x}$ given $\tilde{y}$ =y evaluated at x and the marginal densities of $\tilde{x}$ and $\tilde{y}$ evaluated at x and y respectively

Bayes' theorem for probability density functions

Of interest when there is a sample of observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ each having a (common) continuous probability distribution characterized by one or several parameters (θ). The probability density function of this distribution can be written

$$f(y \mid \theta)$$

Hence, expressed as a *conditional* probability density function.

The goal is to make inference about θ using all available observations $\Rightarrow$ We should use the likelihood of θ that can be obtained from the sample:

$$L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

$$f_{\tilde{\theta}|\tilde{y}}(\theta|y) = \frac{f_{\tilde{y}|\tilde{\theta}}(y|\theta) \cdot f_{\tilde{\theta}}(\theta)}{f_{\tilde{y}}(y)} \Rightarrow f_{\tilde{\theta}|\tilde{y}}(\theta|\boldsymbol{y}) = \frac{L(\theta; \boldsymbol{y}) \cdot f_{\tilde{\theta}}(\theta)}{f_{\tilde{y}}(y)}$$

$$\Rightarrow f_{\tilde{\theta}|\tilde{y}}(\theta|y) = \frac{L(\theta; \boldsymbol{y}) \cdot f_{\tilde{\theta}}(\theta)}{\int L(z; \boldsymbol{y}) \cdot f_{\tilde{\theta}}(z) dz} \propto L(\theta; \boldsymbol{y}) \cdot f_{\tilde{\theta}}(\theta)$$

The exponential class of distributions

A (family) of probability distribution(s) belong(s) to the *k*-parameter exponential class of distributions if the probability density (or mass) function can be written:

$$f\left(\boldsymbol{x}\middle|\boldsymbol{\theta}\right)=e^{\sum_{j=1}^{k}A_j\left(\boldsymbol{\theta}\right)B_j\left(\boldsymbol{x}\right)+C\left(\boldsymbol{x}\right)+D\left(\boldsymbol{\theta}\right)}$$

where

- $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_k)$
- $A_1(\boldsymbol{\theta}),\ldots,A_k(\boldsymbol{\theta})$ and $D(\boldsymbol{\theta})$ are functions of the parameter $\boldsymbol{\theta}$ only (and not of $\boldsymbol{x}$)
- $B_1(\boldsymbol{x}),\ldots,B_k(\boldsymbol{x})$ and $C(\boldsymbol{x})$ are functions of $\boldsymbol{x}$ only (and not of $\boldsymbol{\theta}$)

Examples

Two parameter Gamma distribution (univariate), shape and rate parameterization:

$$f\left(x|\boldsymbol{\theta}\right)= f\left(x|\alpha,\beta\right)=\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}x^{\alpha-1}e^{-\beta x} \qquad ; x\geq 0$$

$$=e^{\left(\alpha-1\right)\left(\ln x\right)-\beta x+\alpha\ln\beta-\ln\Gamma\left(\alpha\right)}=e^{\alpha\ln x-\beta x-\ln x+\alpha\ln\beta-\ln\Gamma\left(\alpha\right)}$$

Parametric form 1: $e^{\left(\alpha-1\right)\left(\ln x\right)-\beta x+\alpha\ln\beta-\ln\Gamma\left(\alpha\right)}$

$$A_1\left(\boldsymbol{\theta}\right)=A_1\left(\alpha,\beta\right)=\alpha-1$$
$$A_2\left(\boldsymbol{\theta}\right)=A_2\left(\alpha,\beta\right)=\beta$$
$$B_1\left(x\right)=\ln x$$
$$B_2\left(x\right)=-x$$
$$C\left(x\right)=0$$
$$D\left(\boldsymbol{\theta}\right)=D\left(\alpha,\beta\right)=\alpha\ln\beta-\ln\Gamma\left(\alpha\right)$$

Parametric form 2: $e^{\alpha \ln x - \beta x - \ln x + \alpha \ln \beta - \ln \Gamma(\alpha)}$

$$A_1(\boldsymbol{\theta}) = A_1(\alpha, \beta) = \alpha$$
$$A_2(\boldsymbol{\theta}) = A_2(\alpha, \beta) = \beta$$
$$B_1(x) = \ln x$$
$$B_2(x) = -x$$
$$C(x) = -\ln x$$
$$D(\boldsymbol{\theta}) = D(\alpha, \beta) = \alpha \ln \beta - \ln \Gamma(\alpha)$$

Canonical form: $A_j(\boldsymbol{\theta}) = \theta_j$

Poisson distribution:

$$f(x|\boldsymbol{\theta}) = f(x|\mu) = \frac{\mu^x}{x!} e^{-\mu} = e^{(\ln \mu) \cdot x - \ln x! - \mu} \quad ; x = 0,1,\ldots$$

$$\left(= e^{(\ln \mu) \cdot x - \ln \Gamma(x+1) - \mu} \right)$$

$$A_1(\boldsymbol{\theta}) = A(\mu) = \ln \mu$$
$$B_1(x) = B(x)$$
$$C(x) = -\ln x!$$
$$D(\boldsymbol{\theta}) = D(\mu) = -\mu$$

Normal distribution:

$$f\left(x|\boldsymbol{\theta}\right)=f\left(x|\mu,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-0.5}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad ;-\infty<x<\infty$$

$$=e^{-\left(1/\left(2\sigma^2\right)\right)\cdot x^2+\left(\mu/\sigma^2\right)\cdot x-0.5\ln(2\pi)-\mu^2/\left(2\sigma^2\right)-0.5\cdot\ln\sigma^2}$$

$$A_1(\boldsymbol{\theta})=A_1\left(\mu,\sigma^2\right)=\frac{1}{2\sigma^2}$$

$$A_2(\boldsymbol{\theta})=A_2\left(\mu,\sigma^2\right)=\frac{\mu}{\sigma^2}$$

$$B_1(x)=-x^2$$

$$B_2(x)=x$$

$$C(x)=-0.5\ln(2\pi)$$

$$D(\boldsymbol{\theta})=D\left(\mu,\sigma^2\right)=-\frac{\mu^2}{2\sigma^2}-0.5\cdot\ln\sigma^2$$

Bernoulli distribution:

$$f\left(x|\boldsymbol{\theta}\right)=f\left(x|p\right)=p^{x}\left(1-p\right)^{1-x} \qquad ;x=0,1$$

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

$$A_1\left(\boldsymbol{\theta}\right)=A\left(p\right)=\ln\left(\frac{p}{1-p}\right)$$

$$B_1\left(x\right)=B\left(x\right)=x$$

$$C\left(x\right)=0$$

$$D\left(\boldsymbol{\theta}\right)=D\left(p\right)=\ln\left(1-p\right)$$

Exercise: The binomial distribution belongs to the exponential class if (conditioned on) the number of trials is fixed. Why?

Conjugate families of distributions when the likelihood belongs to the exponential class

pdf (or pmf) of sample point distribution : $f\left(\boldsymbol{x}\middle|\boldsymbol{\theta}\right)=e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})B_j(\boldsymbol{x})+C(\boldsymbol{x})+D(\boldsymbol{\theta})}$

likelihood: $\prod_{i=1}^{n} f\left(\boldsymbol{x}_i\middle|\boldsymbol{\theta}\right)=\prod_{i=1}^{n} e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})B_j(\boldsymbol{x}_i)+C(\boldsymbol{x}_i)+D(\boldsymbol{\theta})}$

$$=e^{\sum_{i=1}^{n}\left(\sum_{j=1}^{k}A_j(\boldsymbol{\theta})B_j(\boldsymbol{x}_i)+C(\boldsymbol{x}_i)+D(\boldsymbol{\theta})\right)}=e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})\sum_{i=1}^{n}B_j(\boldsymbol{x}_i)+\sum_{i=1}^{n}C(\boldsymbol{x}_i)+n\cdot D(\boldsymbol{\theta})}$$

Hence the multivariate array ($\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$) with independent marginal distributions all with density $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ also belongs to the exponential class.

Now, mimic the structure of the exponential class (for the marginal distributions or the likelihood) and define the prior density as

$$p\left(\boldsymbol{\theta}\middle|\alpha_1,\ldots,\alpha_k,\alpha_{k+1}\right)=e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})\cdot\alpha_j+\alpha_{k+1}\cdot D(\boldsymbol{\theta})+K(\alpha_1,\ldots,\alpha_k,\alpha_{k+1})}$$

$$\propto e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})\cdot\alpha_j+\alpha_{k+1}\cdot D(\boldsymbol{\theta})}$$

where α_1 , … , α_{k+1} are the hyperparameters of this prior distribution and $K(\ )$ is a function of α_1 , … , α_{k+1} only .

Then

$$q\left(\boldsymbol{\theta}\middle|\mathbf{x},\alpha_1,\ldots,\alpha_k,\alpha_{k+1}\right)=q\left(\boldsymbol{\theta}\middle|x_1,\ldots,x_n;\alpha_1,\ldots,\alpha_k,\alpha_{k+1}\right)\propto\underbrace{\prod_{i=1}^{n}f\left(x_i\middle|\boldsymbol{\theta}\right)}_{\text{likelihood}}\cdot\,p\left(\boldsymbol{\theta}\middle|\alpha_1,\ldots,\alpha_k,\alpha_{k+1}\right)$$

$$=e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})\sum_{i=1}^{n}B_j(\mathbf{x}_i)+\sum_{i=1}^{n}C(\mathbf{x}_i)+n\cdot D(\boldsymbol{\theta})}\cdot e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})\cdot\alpha_j+\alpha_{k+1}\cdot D(\boldsymbol{\theta})+K(\alpha_1,\ldots,\alpha_k,\alpha_{k+1})}$$

$$=e^{\sum_{i=1}^{n}C(\mathbf{x}_i)}e^{K(\alpha_1,\ldots,\alpha_k,\alpha_{k+1})}e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})\left(\sum_{i=1}^{n}B_j(\mathbf{x}_i)+\alpha_j\right)+(n+\alpha_{k+1})\cdot D(\boldsymbol{\theta})}\propto e^{\sum_{j=1}^{k}A_j(\boldsymbol{\theta})\left(\sum_{i=1}^{n}B_j(\mathbf{x}_i)+\alpha_j\right)+(n+\alpha_{k+1})\cdot D(\boldsymbol{\theta})}$$

i.e. the posterior distribution is of the same form as the prior distribution but with hyperparameters

$$\alpha_1+\sum_{i=1}^{n}B_1(x_i),\ldots,\alpha_k+\sum_{i=1}^{n}B_k(x_i),\alpha_{k+1}+n$$

instead of

$$\alpha_1,\ldots,\alpha_k,\alpha_{k+1}$$

Some common cases (within or outside the exponential family):

Conjugate prior	*Sample distribution*	*Posterior*
Beta $\pi \sim Beta(\alpha, \beta)$	Binomial $X \sim Bin(n, \pi)$	Beta $\pi \mid x \sim Beta(\alpha + x, \beta + n - x)$
Normal $\mu \sim N(\phi, \tau^2)$	Normal, known σ^2 $X_i \sim N(\mu, \sigma^2)$	Normal $\mu \mid \bar{x} \sim N\left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\phi + \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{x}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$
Gamma $\lambda \sim Gamma(\alpha, \beta)$	Poisson $X_i \sim Po(\lambda)$	Gamma $\lambda \mid \sum x_i \sim Gamma(\alpha + \sum x_i, \beta + n)$
Pareto $p(\theta) \propto \theta^{-\alpha}; \theta \geq \beta$	Uniform $X_i \sim U(0, \theta)$	Pareto $q(\theta; \mathbf{x}) \propto \theta^{-(\alpha+n)}; \theta \geq \max\left(\beta, x_{(n)}\right)$

Exercise 4.10

Suppose that you are interested in $\tilde{p}$, the proportion of station wagons among the registered vehicles in a particular state. Your prior distribution for $\tilde{p}$ is a normal distribution with mean 0.05 and variance 0.0004. To obtain more information, a random sample of 50 registered vehicles is taken, and three are station wagons.

(a) In using Equation 4.2.3 to revise your distribution of $\tilde{p}$, what difficulties are encountered?
(b) How might you avoid such difficulties in this situation? Can they always be avoided?

Equation 4.2.3:

$$f(\theta|y) = \frac{f(\theta) f(y|\theta)}{\int_{-\infty}^{\infty} f(\theta) f(y|\theta) d\theta}$$

(a)

$$f(\theta) = f(p) = \left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(p-\mu)^2}{2\sigma^2}}$$

$$\mu = 0.05$$

$$\sigma^2 = 0.0004$$

$$f(y|\theta) = f(y|p) \approx \left\langle \begin{array}{l} \text{Approximate with a} \\ \text{binomial sampling} \\ \text{model} \end{array} \right\rangle \approx \binom{50}{y} \cdot p^y (1-p)^{50-y}$$

$$\Rightarrow f(p) f(y|p) = \left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \cdot \binom{50}{y} \cdot p^y (1-p)^{50-y}$$

With $y = 3$:

$$f(p) f(3|p) = \left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \cdot \binom{50}{3} \cdot p^3 (1-p)^{50-3}$$

$$f(\theta|y) = f(p|y) = \frac{f(p)f(y|p)}{\int_{-\infty}^{\infty} f(p)f(y|p)dp} = \frac{\left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \cdot \binom{50}{y} \cdot p^y (1-p)^{50-y}}{\int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(q-\mu)^2}{2\sigma^2}} \cdot \binom{50}{y} \cdot q^y (1-q)^{50-y} dq}$$

$$= \frac{\left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \cdot p^y (1-p)^{50-y}}{\int_{-\infty}^{\infty} \left(2\pi\sqrt{\sigma^2}\right)^{-0.5} e^{-\frac{(q-\mu)^2}{2\sigma^2}} \cdot q^y (1-q)^{50-y} dq} = \frac{\left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \cdot p^y (1-p)^{50-y}}{E\left(\tilde{p}^y (1-\tilde{p})^{50-y}\right)}$$

Difficulties?

$E\left(\tilde{p}^y (1-\tilde{p})^{50-y}\right)$ must be developed into a mean of a polynomial in $\tilde{p}$.

(b)

Feasible when $y = 3$ using the moment generating function of $\tilde{p}$

$$M(t) = e^{\mu t + 0.5\sigma^2 t^2}$$

(b) cont.

But here we may approximate the binomial likelihood with a normal distribution:

$$f(p|y) \tilde{\propto} \frac{\left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \cdot \left(2\pi \cdot 50\, p(1-p)\right)^{-0.5} e^{-\frac{(y-50p)^2}{2\cdot 50\, p(1-p)}}}{\int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-0.5} e^{-\frac{(q-\mu)^2}{2\sigma^2}} \cdot \left(2\pi \cdot 50q(1-q)\right)^{-0.5} e^{-\frac{(y-50q)^2}{2\cdot 50q(1-q)}} dq}$$

With $y = 3$ approximate further $50p(1–p)$ and $50q(1–q)$ with $50 \cdot \frac{3}{50} \cdot \frac{47}{50} = \frac{141}{50}$

Using completion of squares the posterior $f(p \mid y = 3)$ is approximately normally distributed.

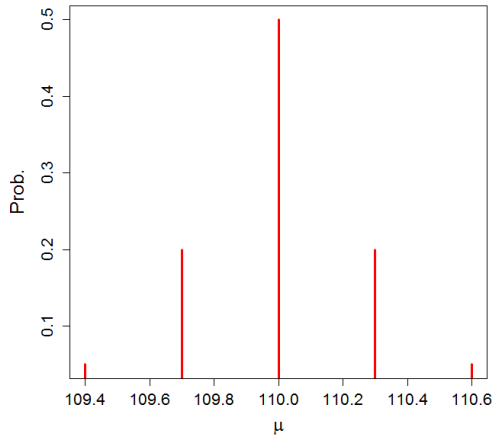
Can only be used when the normal approximation is valid.

Exercise 4.27

27. A production manager is interested in the mean weight of items turned out by a particular process. He feels that the weight of items from the process is normally distributed with mean $\tilde{\mu}$ and that $\tilde{\mu}$ is either 109.4, 109.7, 110.0, 110.3, or 110.6. The production manager assesses prior probabilities of $P(\tilde{\mu} = 109.4) = 0.05$, $P(\tilde{\mu} = 109.7) = 0.20$, $P(\tilde{\mu} = 110.0) = 0.50$, $P(\tilde{\mu} = 110.3) = 0.20$, and $P(\tilde{\mu} = 110.6) = 0.05$. From past experience, he is willing to assume that the process variance is $\sigma^2 = 4$. He randomly selects five items from the process and weighs them, with the following results: 108, 109, 107.4, 109.6, and 112. Find the production manager's posterior distribution and compute the means and the variances of the prior and posterior distributions.

Prior distribution of $\tilde{\mu}$:

$$\tilde{\mu} = \begin{cases} \mu_1 = 109.4 & \text{with prob. } 0.05\ (= p(\mu_1)) \\ \mu_2 = 109.7 & \text{with prob. } 0.20\ (= p(\mu_2)) \\ \mu_3 = 110.0 & \text{with prob. } 0.50\ (= p(\mu_3)) \\ \mu_4 = 110.3 & \text{with prob. } 0.20\ (= p(\mu_4)) \\ \mu_5 = 110.6 & \text{with prob. } 0.05\ (= p(\mu_5)) \end{cases}$$



Discretized normal distribution?

Data: $\mathbf{y}$ = (108.0 109.0 107.4 109.6 112.0) $\sim N\left(\tilde{\mu}, \sigma^2 \approx 4\right)$

Sample point density: $f\left(y|\tilde{\mu}=\mu\right)=\left(2\pi\sigma^2\right)^{-0.5}e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad ;\sigma^2=4$

Likelihood: $\mathcal{L}\left(\mu;\boldsymbol{y}\right)=\prod_{j=1}^{n=5}f\left(y_j|\tilde{\mu}=\mu\right)=\prod_{j=1}^{n=5}\left(2\pi\sigma^2\right)^{-0.5}e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}$

$$=\left(2\pi\sigma^2\right)^{-0.5n}e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2}=\left(8\pi\right)^{-2.5}e^{-\frac{1}{8}\sum_{j=1}^{5}(y_j-\mu)^2}$$

$$=\left(8\pi\right)^{-2.5}e^{-\frac{1}{8}\left[(108-\mu)^2+(109-\mu)^2+(107.4-\mu)^2+(109.6-\mu)^2+(112-\mu)^2\right]}$$

Posterior distribution of $\tilde{\mu}$:

$$q(\mu|\boldsymbol{y}) = \frac{\mathcal{L}(\mu;\boldsymbol{y})\cdot p(\mu)}{\sum_{i=1}^{5}\mathcal{L}(\mu_i;\boldsymbol{y})\cdot p(\mu_i)} = \frac{\left(2\pi\sigma^2\right)^{-0.5n} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2}\cdot p(\mu)}{\sum_{i=1}^{5}\left(2\pi\sigma^2\right)^{-0.5n} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu_i)^2}\cdot p(\mu_i)}$$

$$= \frac{e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2}\cdot p(\mu)}{\sum_{i=1}^{5} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu_i)^2}\cdot p(\mu_i)} = \frac{e^{-\frac{1}{8}\sum_{j=1}^{n}(y_j-\mu)^2}\cdot p(\mu)}{\begin{aligned}
&e^{-\frac{1}{8}\left[(108-109.4)^2+(109-109.4)^2+(107.4-109.4)^2+(109.6-109.4)^2+(112-109.4)^2\right]}\cdot 0.05 \\
&+e^{-\frac{1}{8}\left[(108-109.7)^2+(109-109.7)^2+(107.4-109.7)^2+(109.6-109.7)^2+(112-109.7)^2\right]}\cdot 0.20 \\
&+e^{-\frac{1}{8}\left[(108-110.0)^2+(109-110.0)^2+(107.4-110.0)^2+(109.6-110.0)^2+(112-110.0)^2\right]}\cdot 0.50 \\
&+e^{-\frac{1}{8}\left[(108-110.3)^2+(109-110.3)^2+(107.4-110.3)^2+(109.6-110.3)^2+(112-110.3)^2\right]}\cdot 0.20 \\
&+e^{-\frac{1}{8}\left[(108-110.6)^2+(109-110.6)^2+(107.4-110.6)^2+(109.6-110.6)^2+(112-110.6)^2\right]}\cdot 0.05
\end{aligned}}$$

$$\frac{e^{-\frac{1}{8}\sum_{j=1}^{n}(y_j-\mu)^2}\cdot p(\mu)}{e^{-\frac{1}{8}\cdot 12.92}\cdot 0.05 + e^{-\frac{1}{8}\cdot 13.97}\cdot 0.20 + e^{-\frac{1}{8}\cdot 15.92}\cdot 0.50 + e^{-\frac{1}{8}\cdot 18.77}\cdot 0.20 + e^{-\frac{1}{8}\cdot 22.52}\cdot 0.05} \approx \frac{e^{-\frac{1}{8}\sum_{j=1}^{n}(y_j-\mu)^2}\cdot p(\mu)}{0.1353}$$

$\Rightarrow$

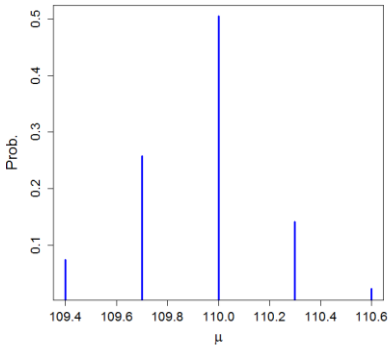
$$q(\mu_1|\boldsymbol{y}) = \underline{\underline{q(109.4|\boldsymbol{y})}} \approx \frac{e^{-\frac{1}{8}\left[(108-109.4)^2+(109-109.4)^2+(107.4-109.4)^2+(109.6-109.4)^2+(112-109.4)^2\right]}\cdot 0.05}{0.1353} \approx \underline{\underline{0.0735}}$$

$$q(\mu_2|\boldsymbol{y}) = \underline{\underline{q(109.7|\boldsymbol{y})}} \approx \frac{e^{-\frac{1}{8}\left[(108-109.7)^2+(109-109.7)^2+(107.4-109.7)^2+(109.6-109.7)^2+(112-109.7)^2\right]}\cdot 0.20}{0.1353} \approx \underline{\underline{0.2578}}$$

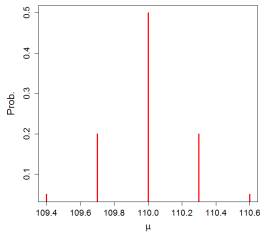
$$q(\mu_3|\boldsymbol{y}) = \underline{\underline{q(110.0|\boldsymbol{y})}} \approx \frac{e^{-\frac{1}{8}\left[(108-110.0)^2+(109-110.0)^2+(107.4-110.0)^2+(109.6-110.0)^2+(112-110.0)^2\right]}\cdot 0.50}{0.1353} \approx \underline{\underline{0.5051}}$$

$$q(\mu_4|\boldsymbol{y}) = \underline{\underline{q(110.3|\boldsymbol{y})}} \approx \frac{e^{-\frac{1}{8}\left[(108-110.3)^2+(109-110.3)^2+(107.4-110.3)^2+(109.6-110.3)^2+(112-110.3)^2\right]}\cdot 0.20}{0.1353} \approx \underline{\underline{0.1415}}$$

$$q(\mu_5|\boldsymbol{y}) = \underline{\underline{q(110.6|\boldsymbol{y})}} \approx \frac{e^{\frac{1}{8}\left[(108-110.6)^2+(109-110.6)^2+(107.4-110.6)^2+(109.6-110.6)^2+(112-110.6)^2\right]}\cdot 0.05}{0.1353} \approx \underline{\underline{0.0221}}$$



prior dist.



$$E_{\text{prior}}(\tilde{\mu}) = E(\tilde{\mu}) = 109.4 \cdot 0.05 + 109.7 \cdot 0.20 + 110.0 \cdot 0.50$$

$$+ 110.3 \cdot 0.20 + 110.6 \cdot 0.05 = \underline{110} \quad \text{(obvious?)}$$

$$Var_{\text{prior}}(\tilde{\mu}) = Var(\tilde{\mu}) = E(\tilde{\mu}^2) - (E(\tilde{\mu}))^2$$

$$= 109.4^2 \cdot 0.05 + 109.7^2 \cdot 0.20 + 110.0^2 \cdot 0.50$$

$$+ 110.3^2 \cdot 0.20 + 110.6^2 \cdot 0.05 - 110^2 = \underline{0.072}$$

$$E_{\text{posterior}}(\tilde{\mu}) = E(\tilde{\mu} | \boldsymbol{x}) = 109.4 \cdot 0.07348... + 109.7 \cdot 0.25780... + 110.0 \cdot 0.50508...$$

$$+ 110.3 \cdot 0.14148... + 110.6 \cdot 0.02213... \approx \underline{109.9}$$

$$Var_{\text{posterior}}(\tilde{\mu}) = Var(\tilde{\mu} | \boldsymbol{x}) = E(\tilde{\mu}^2 | \boldsymbol{x}) - (E(\tilde{\mu} | \boldsymbol{x}))^2$$

$$\approx 109.4^2 \cdot 0.07348... + 109.7^2 \cdot 0.25780... + 110.0^2 \cdot 0.50508...$$

$$+ 110.3^2 \cdot 0.14148... + 110.6^2 \cdot 0.02213... - 109.7^2 \approx \underline{0.066}$$

Exercise 4.28

In Exercise 27, if $\tilde{\mu}$ is assumed to be continuous and if the prior distribution for $\tilde{\mu}$ is a normal distribution with mean 110 and variance 0.4, find the posterior distribution.

Prior distribution: $\tilde{\mu} \sim N(110, 0.4) = N(m', \sigma'^2)$

Prior density: $p(\mu) = f'(\mu) = \left(2\pi\sigma'^2\right)^{-0.5} e^{-\frac{(\mu - m')^2}{2\sigma'^2}} = (2\pi \cdot 0.4)^{-0.5} e^{-\frac{(\mu - 110)^2}{0.8}}$

Data: **y** = (108.0 109.0 107.4 109.6 112.0) $\sim N\left(\tilde{\mu}, \sigma^2 \approx 4\right)$

$$\bar{y} = \frac{108 + 109 + 107.4 + 109.6 + 112}{5} = 109.2$$

$$s^2 = \frac{1}{4}\sum_{1}^{5}\left(y_j - 109.2\right)^2 = 3.18$$

$$q(\mu|\mathbf{y}) = f''(\mu|\mathbf{y}) = \frac{\mathcal{L}(\mu;\mathbf{y})\cdot f'(\mu)}{\int_{-\infty}^{\infty}\mathcal{L}(\mu;\mathbf{y})\cdot f'(\mu)d\mu} = \left\langle \begin{matrix}\mathcal{L}(\mu;y) \text{ from} \\ \text{Exercise } 4.27\end{matrix} \right\rangle$$

$$= \frac{\left(2\pi\sigma^2\right)^{-0.5n} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2}\cdot\left(2\pi\sigma'^2\right)^{-0.5} e^{-\frac{(\mu-m')^2}{2\sigma'^2}}}{\int_{-\infty}^{\infty}\left(2\pi\sigma^2\right)^{-0.5n} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2}\cdot\left(2\pi\sigma'^2\right)^{-0.5} e^{-\frac{(\mu-m')^2}{2\sigma'^2}} d\mu} = \langle \text{"Completion of squares"} \rangle$$

$$= \left(2\pi\sigma''^2\right)^{-0.5} e^{-\frac{(\mu-m'')^2}{2\sigma''^2}}$$

where

$$m'' = \frac{\left(1/\sigma'^2\right)\cdot m' + \left(n/\sigma^2\right)\cdot m}{\left(1/\sigma'^2\right)+\left(n/\sigma^2\right)} = \frac{\left(1/\sigma'^2\right)\cdot m' + \left(n/\sigma^2\right)\cdot \bar{y}}{\left(1/\sigma'^2\right)+\left(n/\sigma^2\right)} = \frac{\left(1/0.4\right)\cdot 110 + \left(5/4\right)\cdot 109.2}{\left(1/0.4\right)+\left(5/4\right)} \approx 109.7$$

$$\sigma''^2 = \frac{\sigma^2\cdot\sigma'^2}{\sigma^2 + n\cdot\sigma'^2} = \frac{4\cdot 0.4}{4+5\cdot 0.4} \approx 0.267$$

Thus, the posterior distribution is $N(m''=109.7, \sigma''^2=0.267)$

Interpretation of prior distributions

Example Prior distribution for a proportion Has already been taken up at meeting 4!

The probably most common way of writing the density of a two-parameter beta distribution is

$$f'(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)} \quad ; 0 \le p \le 1$$

where other symbols (e.g. α and β) can substitute for (a, b)

The term $B(a, b)$ in the denominator is the *Beta function* evaluated at (a, b). This function is defined as

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$$

…but can also be expressed in terms of the *Gamma function*: $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$

The course book by Winkler uses a slightly different parameterization:

$$f'(p) = \frac{\Gamma(n)}{\Gamma(r)\Gamma(n-r)} p^{r-1}(1-p)^{n-r-1}$$

where compared to the description above $r = a$ and $n = a + b$.

Moreover, if r and n are integers the density can be written $$f'(p)=\frac{(n-1)!}{(r-1)!(n-r-1)!}p^{r-1}(1-p)^{n-r-1}$$

With this parameterization, we have $$E(\tilde{p}|r,n)=\frac{r}{n} \quad ; \quad Var(\tilde{p}|r,n)=\frac{r(n-r)}{n^2(n+1)}$$

whereas with the former parametrization, we have $$E(\tilde{p}|a,b)=\frac{a}{a+b} \quad ; \quad Var(\tilde{p}|a,b)=\frac{ab}{(a+b)^2(a+b+1)}$$

Now, $\tilde{p}$ is in almost all cases an unknown proportion (or probability interpreted in that sense) when the two-parameter beta distribution is used as its prior.

Since $E(\tilde{p}|r,n) = r/n$ we can interpret this prior distribution as the amount of information available for $\tilde{p}$ when a random sample of n units (or a run of n independent trials) contains r units (results in r outcomes) with the property (of the event) for which $\tilde{p}$ is the proportion (the probability).

Let us assume that we have prior knowledge about the proportion telling us that it would be around 20 %.

If we deem our prior knowledge not being precise, it is wise to use a beta distribution with as small parameter values as possible, but still having 0.20 as its mean.

$\Rightarrow$ Choose r = 1 and n = 5

If we deem the prior knowledge to be more precise, increase r and n keeping the proportion r/n equal to 0.20

r = 2, n = 10 r = 10, n =50 r = 200, n =1000 etc.

Example Prior distribution for a mean

Very often we have reasons to work with normally distributed data to make inference about the population mean $\tilde{\mu}$.

If the population variance is (assumed to be) known = σ^2, we can – as was demonstrated in Exercise 4.28 – use the normal distribution as a conjugate prior distribution.

From sampling theory we know that – setting aside finite population corrections – the variance of the sample mean is the population variance divided by the sample size

$$Var\left(\tilde{\bar{y}}\middle|\sigma^2, n\right) = \frac{\sigma^2}{n}$$

If σ'^2 represents the prior variance of the unknown $\tilde{\mu}$ define a new parameter n' as

$$n' = \frac{\sigma^2}{\sigma'^2}$$

Hence,

$$\sigma'^2 = \frac{\sigma^2}{n'}$$

This can be interpreted as the variance σ'^2 of a sample mean based on n' observations taken from the population with population variance σ^2.

n' then plays the role of the size of a virtual sample taken from the population on which the prior knowledge stems.

Note that it is not necessary for n' to be integer-valued, even if it often suffices to approximate with an integer.

For the prior and posterior distribution we may thus write

$$\tilde{\mu} \sim N\left(m', \frac{\sigma^2}{n'}\right) \qquad \tilde{\mu}|\boldsymbol{y} \sim N\left(m'', \frac{\sigma^2}{n''}\right)$$

Exercise 4.28 with alternative prior parametrization

Prior distribution: $\tilde{\mu} \sim N(110, 0.4) = N(m', \sigma'^2) = N(m', \sigma^2/10)$

since $\sigma^2 = 4$

$q(\mu|\boldsymbol{y}) = \left(2\pi\sigma''^2\right)^{-0.5} e^{-\frac{(\mu - m'')^2}{2\sigma''^2}}$ [from the a bove solution]

where

$$m'' = \frac{\left(1/\sigma'^2\right)\cdot m' + \left(n/\sigma^2\right)\cdot m}{\left(1/\sigma'^2\right) + \left(n/\sigma^2\right)} = \frac{\left(n'/\sigma^2\right)\cdot m' + \left(n/\sigma^2\right)\cdot m}{\left(n'/\sigma'^2\right) + \left(n/\sigma^2\right)} = \left\langle \begin{matrix} \text{All instnaces of } 1/\sigma^2 \\ \text{can be removed} \end{matrix} \right\rangle$$

$$= \frac{n'\cdot m' + n\cdot m}{n' + n} = \frac{10\cdot 110 + 5\cdot 109.2}{10 + 5} \approx 109.7$$

$$\sigma''^2 = \frac{\sigma^2\cdot\sigma'^2}{\sigma^2 + n\cdot\sigma'^2} = \frac{\sigma^2\cdot\left(\sigma^2/n'\right)}{\sigma^2 + n\cdot\left(\sigma^2/n'\right)} = \frac{4\cdot 0.4}{4 + 5\cdot 0.4} = \frac{1.6}{6} \approx 0.267 = \frac{\sigma^2}{n''}$$

$$\Rightarrow n'' = \frac{\sigma^2}{1.6/6} = \frac{24}{1.6} = 15$$

And… $n'' = n' + n = 10 + 5$

Exercise 4.37

(c) Find the joint posterior distribution of $\tilde{\mu}$ and $\tilde{\sigma}^2$.

In Exercise 27, suppose that the production manager is unwilling to assume that $\tilde{\sigma}^2$ is known. Instead, he assesses a normal-gamma prior distribution for $\tilde{\mu}$ and $\tilde{\sigma}^2$ with parameters $m' = 110$, $n' = 10$, $v' = 4$, and $d' = 6$. Find the posterior distribution of $\tilde{\mu}$ and $\tilde{\sigma}^2$ and compute $E(\tilde{\mu}|\sigma^2)$ and $E(1/\tilde{\sigma}^2)$ from this distribution.

... and assess a continuous dis-

Normal-gamma prior distribution for $(\tilde{\mu}, \tilde{\sigma}^2)$

$$f'\left(\mu, \sigma^2 \middle| m', n', \upsilon', d'\right) = g'\left(\mu \middle| m', n', \sigma^2\right) \cdot h'\left(1/\sigma^2 \middle| \upsilon', d'\right)$$

$$m' = 110 \quad n' = 10 \quad \upsilon' = 4 \quad d' = 6$$

Sample data is as before, i.e. **y** = (108.0 109.0 107.4 109.6 112.0)

but the population variance is known unknown

$$y = \frac{108 + 109 + 107.4 + 109.6 + 112}{5} = 109.2$$

$$s^2 = \frac{1}{4}\sum_{1}^{5}\left(y_j - 109.2\right)^2 = 3.18$$

$$\Rightarrow N\left(\tilde{\mu}, \tilde{\sigma}^2\right)$$

The posterior distribution of $(\tilde{\mu}, \tilde{\sigma}^2)$ is also normal-gamma with

$$m'' = \frac{n'm' + nm}{m' + n} = \frac{10 \cdot 110 + 5 \cdot 109.2}{10 + 5} \approx 109.73$$

$$n'' = n' + n = 10 + 5 = 15$$

$$\upsilon'' = \frac{\left(d'\upsilon' + n'm'^2\right) + \left[(n-1)s^2 + nm^2\right] - n''m''^2}{d' + n}$$

$$= \frac{\left(6 \cdot 4 + 10 \cdot 110^2\right) + \left[4 \cdot 3.18 + 5 \cdot 109.2^2\right] - 15 \cdot 109.73^2}{6 + 5} \approx 4.53$$

$$d'' = d' + n = 6 + 5 = 11$$