ST4234 Notes

Contents

1 Important Distributions

Beta distribution: $X \sim Beta(\alpha, \beta)$. Then $\pi(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

- Binomial: probability is a *parameter*
- Beta: probability is a *random variable*
- Can use normal approximation when α and β are both large

Gamma distribution: $X \sim Gamma(\alpha, \frac{1}{\beta})$. Then $\pi(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

- Interpretation: α is the *shape*, $\frac{1}{\beta}$ is the *rate*
- $E[X] = \frac{\alpha}{\beta}$, $Var(X) = \frac{\alpha}{\beta^2}$
- $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\ dx = (\alpha-1)!$
- $\Gamma(\frac{1}{2}+n) = \frac{(2n)!}{4^n n!}\sqrt{\pi}$
- Can use normal approximation when α is large

Poisson distribution: $X \sim Po(\lambda)$. Then $\pi(x) = \frac{e^{-\lambda}\lambda^x}{x!}$.

Exponential distribution: $X \sim Exp(\frac{1}{\lambda})$. Then $\pi(x) = \lambda e^{-\lambda x}$ where $x > 0$.

- $E[X] = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$

Normal distribution: $X \sim N(\mu, \frac{1}{\tau})$. Then $\pi(x) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x-\mu)^2}$.

Standard central t-distribution: $X \sim t_v$. Then $\pi(x) = \frac{1}{B(v/2,1/2)} \frac{1}{\sqrt{v}} (1 + \frac{x^2}{v})^{-\frac{v+1}{2}}$

- $E[X] = 0$ for $v > 1$
- $Var(X) = \frac{v}{v-2}$ for $v > 2$

Non-central t-distribution: $X \sim t_v(m, \frac{1}{c})$. Then $\pi(x) = \frac{1}{B(v/2,1/2)} \sqrt{\frac{c}{v}} (1 + c\frac{(x-m)^2}{v})^{-\frac{v+1}{2}}$

- $E[X] = m$ for $v > 1$
- $Var(X) = \frac{1}{c} \cdot \frac{v}{v-2}$ for $v > 2$

Dirichlet distribution: $X \sim Dirichlet(\alpha_1, \alpha_2, \alpha_3)$. Then $\pi(x_1, x_2, x_3) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} x_3^{\alpha_3-1}$, where $\sum_{i=1}^N x_i = 1$

- Generalisation of the *beta* distribution

Pareto distribution: $X \sim Pareto(m, a)$ where $m > 0$ and $a > 0$. Then $\pi(x) = \frac{am^a}{x^{a+1}}$, where $x > m$

- $E[X] = \frac{am}{a-1}$ for $a > 1$ — ∞ for $\alpha \leq 1$
- $Var(X) = \frac{am^2}{(a-1)^2(a-2)}$ for $a > 2$ — ∞ for $\alpha \leq 2$
- $F(x) = 1 - (\frac{m}{x})^a$ for $x > m$
- Mode is at m

2 Introduction

- Prior density: $\pi(\theta)$
- Likelihood function (model density): $f(x|\theta)$
- Posterior density: $\pi(\theta|x)$

Bayesian formulation:

- Posterior $\propto$ Prior $\times$ Likelihood i.e. $\pi(\theta|x) \propto \pi(\theta) \times f(x|\theta)$

2.1 Conditional Probability

Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

- $P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$
- $P(A|B) \propto P(A \cap B)$

Conditional density function: $f(x|y) = \frac{f(x,y)}{f(y)}$

- where $f(x, y)$ is joint pdf, $f(x)$ and $f(y)$ are marginal pdfs
- $f(x, y) = f(x|y) \times f(y) = f(y|x) \times f(x)$
- $f(x|y) \propto f(x, y)$

2.1.1 Axioms of Conditional Probability

1. $0 \leq P(A|H) \leq 1$
2. $P(H|H) = 1$
3. Area Rule: $P(A_1 \cup A_2|H) = P(A_1|H) + P(A_2|H)$ if A_1 and A_2 are disjoint
4. Product Rule: $P(A_1 \cap A_2|H) = P(A_1|H) \times P(A_2|A_1 \cap H)$ (you can "ignore" the H)

Corollaries:

- $P(A_1 \cup \ldots \cup A_k|H) = P(A_1|H) + \ldots + P(A_k|H)$ if the A events are mutually disjoint
- $P(A_1 \ldots A_k|H) = P(A_1|H)P(A_2|A_1H)P(A_3|A_1A_2H) \ldots P(A_k|A_1 \ldots A_{k-1}H)$

2.2 Bayes' Theorem

Let $A_1 \ldots A_k$ be a partition of Ω, i.e. *disjoint* (mutually exclusive) and *exhaustive*.

$$\begin{aligned} P(B) &= P(A_1B) + P(A_2B) + \ldots + P(A_kB) \\ &= P(A_1) \times P(B|A_1) + P(A_2) \times P(B|A_2) + \ldots + P(A_k) \times P(B|A_k) \end{aligned}$$

Bayes' Theorem: $P(A_i|B) \propto P(A_i) \times P(B|A_i)$

- $P(A_i)$ is the *prior probability*
- $P(A_i|B)$ is the *posterior probability*
- Normalization constant: reciprocal of $P(B) = P(A_1) \times P(B|A_1) + \ldots$ as above.

2.2.1 Example: Drawing Cards

Draw without replacement from $\{1, 2, 2, 3, 3, 3\}$. What is the probability that the 2nd draw is a 2?

$$\begin{aligned}
P(X_2 = 2) &= P(X_2 = 2, X_1 = 1) + P(X_2 = 2, X_1 = 2) + P(X_2 = 2, X_1 = 3) \\
&= \frac{2}{5} \times \frac{1}{6} + \frac{1}{5} \times \frac{2}{6} \times \frac{2}{5} \times \frac{3}{6} \\
&= \frac{1}{15} + \frac{1}{15} + \frac{1}{5} \\
&= \frac{1}{3}
\end{aligned}$$

2.2.2 Example: Bags of Apples

Bag 1: 10% are bad apples Bag 2: 20% are bad apples Bag 3: 40% are bad apples

Suppose we pick one of the bags at random, and pick an apple from it.

1. Suppose the apple is bad. What is the probability that we picked Bag 1?
2. Suppose the apple is good. What is the probability that we picked Bag 1?

1. $P(B = 1|A = bad) = \frac{P(B=1) \times P(A=bad|B=1)}{P(A=bad)} = \frac{1/3 \times 0.1}{1/3 \times 0.1 + 1/3 \times 0.2 + 1/3 \times 0.4} = \frac{1}{7}$
2. $P(B = 1|A = good) = \frac{P(B=1) \times P(A=good|B=1)}{P(A=good)} = \frac{1/3 \times 0.9}{1/3 \times 0.9 + 1/3 \times 0.8 + 1/3 \times 0.6} = \frac{9}{23}$

3 Lecture 2: Bayesian Theorems for Random Variables

3.1 Conditional Random Variables

- Conditional distribution function: $F(y|x) \equiv P(Y \leq y|X = x)$
- Conditional p.m.f: $\pi(y|x) \equiv P(Y = y|X = x)$
- Conditional p.d.f: $\pi(y|x) \equiv \frac{d}{dy}F(y|x)$
- Independence: X and Y are independent if $\pi(y|x) = \pi(y)$

Conditional mean and variance

- $E(Y|X = x) = \sum_y y \cdot \pi(y|x)$ ($Y|X = x$ is discrete)
- $E(Y|X = x) = \int y \cdot \pi(y|x)\ dy$ ($Y|X = x$ is continuous)
- $Var(Y|X = x) = \sum_y [y - E(Y|X = x)]^2 \cdot \pi(y|x)$ ($Y|X = x$ is discrete)
- $Var(Y|X = x) = \int [y - E(Y|X = x)]^2 \cdot \pi(y|x)\ dy$ ($Y|X = x$ is continuous)

Joint density

- $\pi(x, y) = \pi(y|x) \cdot \pi(x) = \pi(x|y) \cdot \pi(y)$

3.2 Bayes' Theorem for Two Random Variables

$$\boxed{\pi(x|y) \propto \pi(x) \cdot \pi(y|x)}$$

- Posterior $\propto$ Prior $\times$ Likelihood, where normalisation constant is Marginal $= \pi(y)$
- *Marginal density* $\pi(y) = \int \pi(x) \cdot \pi(y|x)\ dx$ (continuous) OR $\sum_x \pi(x) \cdot \pi(y|x)$ (discrete)
- Kernel: form of the posterior, that *ignores constant factors* (e.g. normalisation constant)

Table 6.14 The simplified table for finding posterior distribution given $Y = 3$

π	*prior*	*likelihood*	*prior* $\times$ *likelihood*	*posterior*
.4	$\frac{1}{3}$	.1536	.0512	$\frac{.0512}{.2497} = .205$
.5	$\frac{1}{3}$	.2500	.0833	$\frac{.0833}{.2497} = .334$
.6	$\frac{1}{3}$	.3456	.1152	$\frac{.1152}{.2497} = .461$
marginal $P(Y = 3)$			.2497	1.000

$$\boxed{\pi(y) \propto \frac{\pi(y|x)}{\pi(x|y)}}$$

- Proof: $RHS = \frac{\pi(x,y)/\pi(x)}{\pi(x,y)/\pi(y)} = \frac{\pi(y)}{\pi(x)} \propto \pi(y)$

3.2.1 Example: Beta as conjugate family for Binomial observations

Let $X \sim U(0, 1)$ and $Y|X \sim Bin(n, X)$. What is the conditional density of X given $Y = y$, i.e. $\pi(x|y)$?

- $\pi(x) = 1$ for $x \in [0, 1]$ (definition of uniform)
- $\pi(y|x) = \frac{n!}{y!(n-y)!} x^y (1 - x)^{n-y}$ (definition of binomial)
- Then $\pi(x|y) \propto \pi(x) \cdot \pi(y|x) \propto x^y (1 - x)^{n-y} = x^{a_n - 1}(1 - x)^{b_n - 1}$
- $\therefore X|Y = y \sim Beta(a_n, b_n)$ where $a_n = y + 1$ and $b_n = n - y + 1$

3.2.2 Example: Gamma as conjugate family for Poisson observations

Let $X \sim Gamma(\alpha, \frac{1}{\beta})$ and $Y|X \sim Po(X)$. What is the conditional density of X given $Y = y$, ie. $\pi(x|y)$?

- $\pi(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$ (definition of gamma)
- $\pi(y|x) = \frac{e^{-x} x^y}{y!}$ (definition of poisson)
- Then $\pi(x|y) \propto \pi(x) \cdot \pi(y|x) \propto \beta^\alpha x^{\alpha-1} e^{-\beta x} \cdot \frac{e^{-x} x^y}{y!} \propto x^{\alpha_n - 1} e^{-\beta_n x}$
- $\therefore X|Y = y \sim Gamma(\alpha_n, 1/\beta_n)$ where $\alpha_n = \alpha + y$ and $\beta_n = \beta + 1$

3.2.3 Example: Gamma-Normal conjugate family

Let $\tau \sim Gamma(\alpha, \frac{1}{\beta})$; $\mu|\tau \sim N(m, \frac{1}{\tau t})$ where t is known.

(i) Find the conditional distribution of $\tau|\mu$.

- $\pi(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}$, where $\tau \geq 0$
- $\pi(\mu|\tau) = \sqrt{\frac{\tau t}{2\pi}} e^{-\frac{\tau t}{2}(\mu-m)^2}$
- Then $\pi(\tau|\mu) \propto \pi(\mu|\tau) \cdot \pi(\tau) \propto \sqrt{\tau} e^{-\frac{\tau t}{2}(\mu-m)^2} \tau^{\alpha-1} e^{-\beta\tau} \propto \ldots \propto \tau^{\alpha_n - 1} e^{-\beta_n \tau}$
- $\therefore \tau|\mu \sim Gamma(\alpha_n, 1/\beta_n)$ - where $\alpha_n = \alpha + \frac{1}{2}$ and $\beta_n = \beta + \frac{t}{2}(\mu - m)^2$

(ii) Find the marginal distribution of μ.

- $\pi(\mu) \propto \frac{\pi(\mu|\tau)}{\pi(\tau|\mu)} \propto \frac{\tau^{1/2} e^{-\frac{\tau t}{2}(\mu-m)^2}}{\beta_n^{\alpha_n} \tau^{\alpha_n - 1} e^{-\beta_n \tau}} \propto \beta_n^{-\alpha_n} = (\beta + \frac{t}{2}(\mu-m)^2)^{-(\alpha+\frac{1}{2})} \propto (1 + \frac{\alpha t}{\beta} \frac{(\mu-m)^2}{2\alpha})^{-(2\alpha+1)/2}$
- $\therefore \mu \sim t_{2\alpha}(m, (\frac{\alpha t}{\beta})^{-1})$

3.3 Bayes' Theorem for Several Random Variables

Product rule: $\pi(y, x_1, \ldots, x_k) = \pi(y) \cdot \pi(x_1|y) \cdot \pi(x_2|y, x_1) \cdot \ldots \cdot \pi(x_k|y, x_1, \ldots, x_{k-1})$

Bayes' Theorem for several random variables: $\pi(y|x_1, \ldots, x_k) \propto \pi(y, x_1, \ldots, x_k)$

- $\pi(y|x_1) \propto \pi(y) \cdot \pi(x_1|y)$
- $\pi(y|x_1, x_2) \propto \pi(y|x_1) \cdot \pi(x_2|y, x_1)$ — to understand this, see that x_1 is always on the RHS
- $\ldots$
- $\pi(y|x_1, \ldots, x_k) \propto \pi(y|x_1, \ldots, x_{k-1}) \cdot \pi(x_k|y, x_1, \ldots, x_{k-1})$
- This is called *sequential updating*

3.3.1 Example: Joint given Conditional (Beta → Dirichlet)

What is the joint distribution of $(X_1 \ldots X_k)$, given the following?

- Let $\alpha_1 \ldots \alpha_k = 0$, $\alpha_{i+} = \alpha_i + \ldots + \alpha_k$.
- Let $X_1 \sim Beta(\alpha_1, \alpha_{2+})$
- Let $(\frac{X_2}{1-X_1}|X_1) \sim Beta(\alpha_2, \alpha_{3+})$
- Let $(\frac{X_{k-1}}{1-X_1-\ldots-X_{k-2}}|X_1, \ldots, X_{k-2}) \sim Beta(\alpha_{k-1}, \alpha_{k+})$

If $X \sim f(x)$, then $Y = aX \sim \frac{1}{a} f(\frac{y}{a})$.

- If $Y = \frac{X}{c} \sim Beta(a, b)$, then we have $f_Y(y) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}$

- Then $X = cY$, then we have $f_X(x) = \frac{1}{c} f_Y(\frac{X}{c}) = \frac{1}{c}\frac{(x/c)^{a-1}(1-x/c)^{b-1}}{B(a,b)}$
- $X_1 \sim Beta(\alpha_1, \alpha_{2+})$
  - So $\pi(x_1) = \frac{\Gamma(\alpha_{1+})}{\Gamma(\alpha_1)\Gamma(\alpha_{2+})} x_1^{\alpha_1 - 1}(1-x_1)^{\alpha_{2+}-1}$
- $\frac{X_2}{1-X_1}|X_1 \sim Beta(\alpha_2, \alpha_{3+})$ and $X_2|X_1 = (1-X_1)(\frac{X_2}{1-X_1}|X_1)$
  - So $\pi(x_2|x_1) = \frac{1}{1-X_1}\frac{\Gamma(\alpha_{2+})}{\Gamma(\alpha_2)\Gamma(\alpha_{3+})}(\frac{x_2}{1-x_1})^{\alpha_2-1}(1-\frac{x_2}{1-x_1})^{\alpha_{3+}-1} = \frac{\Gamma(\alpha_{2+})}{\Gamma(\alpha_2)\Gamma(\alpha_{3+})}\frac{x_2^{\alpha_2-1}(1-x_1-x_2)^{\alpha_{3+}-1}}{(1-x_1)^{\alpha_{2+}-1}}$
  - Then $\pi(x_1, x_2) = \pi(x_1)\cdot\pi(x_2|x_1) = \frac{\Gamma(\alpha_{1+})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_{3+})} x_1^{\alpha_1-1}x_2^{\alpha_2-1}(1-x_1-x_2)^{\alpha_{3+}-1}$
- $\frac{X_3}{1-X_1-X_2} \sim Beta(\alpha_3, \alpha_{4+}) \dots$
  - So $\pi(x_3|x_1,x_2) = \frac{\Gamma(\alpha_{3+})}{\Gamma(\alpha_3)\Gamma(\alpha_{4+})}\frac{x_3^{\alpha_3-1}(1-x_1-x_2-x_3)^{\alpha_{4+}-1}}{(1-x_1-x_2)^{\alpha_{3+}-1}}$
  - Then $\pi(x_1,x_2,x_3) = \pi(x_1)\cdot\pi(x_2|x_1)\cdot\pi(x_3|x_1,x_2) = \frac{\Gamma(\alpha_{1+})}{\Pi_{i=1}^3\Gamma(\alpha_i)\cdot\Gamma(\alpha_{4+})}\Pi_{i=1}^3 x_i^{\alpha_i-1}\cdot(1-\Pi_{i=1}^3 x_i)^{\alpha_{4+}-1}$
  - Notice that when $k=3$, we have $\pi(x_1,x_2,x_3) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha+3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}x_3^{\alpha_3-1}$, therefore it $\sim Dirichlet(\alpha_1, \alpha_2,$
- In general, $\pi(x_1,\dots,x_k) \sim Dirichlet(\alpha_1,\dots,\alpha_k)$

3.3.2 Example: Conditional given Joint (Multinomial → Binomial)

Let $(X_1,\dots,X_k) \sim Multinomial(n, p_1,\dots,p_k)$ where $\sum_{i=1}^k p_i = 1$ and $\sum_{i=1}^k x_i = n$. Given $X_1,\dots,X_{k-2}$, what is the distribution of X_{k-1}?

$\pi(x_1,\dots,x_k) = \frac{n!}{x_1!\dots x_k!}p_1^{x_1}\dots p_k^{x_k}$

- Set $n^* = n - \sum_{i=1}^{k-2} x_i$ and $p^* = 1 - \sum_{i=1}^{k-2} p_i$
- Then $\pi(x_{k-1}|x_1,\dots,x_{k-2}) \propto \pi(x_1,\dots,x_k) \propto \frac{1}{x_{k-1}!(n^*-x_{k-1})!}p_{k-1}^{x_k-1}(p^*-p_{k-1})^{n^*-x_{k-1}} \propto \frac{1}{x_{k-1}!(n^*-x_{k-1})!}\tilde{p}^{x_{k-1}}(1-\tilde{p})^{n^*-x_{k-1}}$ where $\tilde{p} = p_{k-1}/p^*$
  - $(n^*-x_{k-1})! = x_k!$ and $(p^*-p_{k-1}) = p_k$; we keep them because x_k is dependent on x_{k-1}
- Then $x_{k-1}|x_1,\dots,x_{k-2} \sim Bin(n^*, \tilde{p})$

3.4 Overview of Bayesian Inference

3.4.1 Problem Setup

Suppose we have parameter θ, and a random quantity X which has model density $\pi(X|\theta)$. Suppose we obtain n i.i.d. observations, $X_1,\dots,X_n$. How can we infer θ?

3.4.2 Frequentist Approach

- MLE estimator $\hat{\theta}$ maximises the likelihood function $L(\theta|X_1,\dots,X_n) = \prod_{i=1}^n \pi(X_i|\theta)$ (or the loglikelihood function $\ell(\theta)$ etc.)
- Standard error of $\hat{\theta}$ — depends on variance of $\hat{\theta}$, through the probability distribution of $\hat{\theta} = f(X_1,\dots,X_n)$
- $(1-\alpha)$ CI: means that approximately $(1-\alpha)$ of the intervals constructed will include true parameter θ,

if we repeatedly get realisations of n observations from $\pi(X|\theta)$ over long period of time

3.4.3 Bayesian Approach

- Assume that θ has a prior density $\pi(\theta)$
- Update the prior to get posterior density $\pi(\theta|x_1, \ldots, x_n)$ using Bayes' Theorem:
- $\pi(\theta|x_1, \ldots, x_n) \propto \pi(\theta) \times \prod_{i=1}^{n} \pi(x_i|\theta)$ — prior × likelihood function
- Estimate θ using *posterior mean*
- *Posterior variance*: $Var(\theta|x_1, \ldots, x_n)$
- $(1-\alpha)$ Credible set/highest density region (HDR): there is $(1-\alpha)$ chance that interval contains true parameter θ

Improper prior densities: do *not* integrate to 1!

- But these we can "deflate" the improper prior to get a proper posterior
- One common improper prior is the *flat* prior (uniform)

4 Lecture 3: Bayesian Inference for a Normal Population

$X \sim N(\mu, \frac{1}{\tau})$, where $\tau = \frac{1}{\sigma^2}$ is the *precision parameter*

4.1 Normal Population with Known Variance $\tau = r \Rightarrow$ Estimate μ

Likelihood function

$$\begin{aligned}
L(\mu|x_1,\ldots,x_n) &\propto \prod_{i=1}^{n}\sqrt{\frac{r}{2\pi}}e^{-\frac{r}{2}(\mu-x_i)^2} \\
&\propto \prod_{i=1}^{n} e^{-\frac{r}{2}[(\mu-\bar{x})+(\bar{x}-x_i)]^2} \\
&\propto e^{-\frac{r}{2}\sum_{i=1}^{n}[(\mu-\bar{x})^2+(\mu-\bar{x})(\bar{x}-x_i)+(\bar{x}-x_i)^2]} \text{ (ignore middle term, because summation of } (\bar{x}-x_i)=0) \\
&\propto e^{-\frac{nr}{2}(\mu-\bar{x})^2}
\end{aligned}$$

- Prior: assume $\mu \sim N(m, \frac{1}{t})$
- Observations: n iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ from $X \sim N(\mu, \frac{1}{r})$, where r is known
- Posterior: $(\mu|\mathbf{x}) \sim N(m_n, \frac{1}{t_n})$ where $t_n = t + nr$ and $m_n = \frac{tm+nr\bar{x}}{t+nr} = w_n m + (1-w_n)\bar{x}$, where $w_n = \frac{t}{t+nr}$ — weighing factor
  - (Proof omitted: use identity below)
- Estimate of population mean, $\hat{\mu} = E[\mu|\mathbf{x}] = m_n$
  - Weighing factor: trade-off between prior information and observed data
  - If $nr >> t$, then w_n is very small, and posterior mean is close to $\bar{x}$

Useful identity

- $r(\mu-a)^2 + t(\mu-b)^2 = (r+t)(\mu-\bar{m})^2 + \frac{(a-b)^2}{t^{-1}+r^{-1}}$ where $\bar{m} = \frac{r}{t+r}a + \frac{t}{t+r}b$ — weighted average of a and b
- Implication: to minimize the LHS, we minimize the $(r+t)(\mu-\bar{m})^2$ term on the RHS (the other term is a constant, can ignore)

4.1.1 Example

Find a) the posterior distribution of θ and b) the posterior probability of $\theta \geq 15$:

- Suppose we don't know θ, the mean of a Normal population
- Suppose we know its variance is 1
- We observe 10 observations: 14.5, 15.1, 15.3, 15.5, 16.3, 16.5, 17.3, 17.3, 17.6, 18.0

i) Suppose $\theta \sim N(0, 1)$:

- Then $\theta|\mathbf{x} \sim N(m_n, \frac{1}{t_n})$, where $m_n = \frac{tm+nr\bar{x}}{t+nr} = 14.8546$ and $t_n = t + nr = 11$
- $P(\theta \geq 15|\mathbf{x}) = 1 - \Phi(\sqrt{t_n}(15 - m_n)) = 1 - \Phi(\sqrt{11}(15 - 14.8546)) = 0.3148$

ii) Suppose $\theta \sim Exp(\frac{1}{\lambda})$, where $\theta \geq 0$ and $\lambda = \frac{1}{16}$:

- $\pi(\theta|\mathbf{x}) \propto \pi(\theta) \times L(\theta|x_1,\ldots,x_n) \propto e^{-\frac{\theta}{\lambda}}e^{-\frac{nr}{2}(\theta-\bar{x})^2} \propto e^{-\frac{nr}{2}(\theta^2-2\bar{x}\theta+\bar{x}^2)-\frac{\theta}{\lambda}} \propto e^{-\frac{nr}{2}(\theta^2-2[\bar{x}-1/(nr\lambda)]\theta)} \propto e^{-\frac{nr}{2}(\theta-m_n)^2}$ where $m_n = \bar{x} - 1/(nr\lambda)$ and $\theta \geq 0$

- Then $\theta|\mathbf{x} \sim TN(m_n, \frac{1}{nr})$ where $\theta \geq 0$, a truncated normal
- $P(\theta \geq 15|\mathbf{x}) = \frac{P(W\geq 15)}{P(W\geq 0)} = \frac{1-\Phi(\sqrt{nr}(15-m_n))}{1-\Phi(\sqrt{nr}(0-m_n))} = \ldots = 0.3228$, where $W \sim N(m_n, \frac{1}{nr})$

4.2 Normal Population with Known Mean $\mu = h \Rightarrow$ Estimate τ

Likelihood function

$$\begin{aligned} L(\tau|x_1, \ldots, x_n) &\propto \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x_i-h)^2} \\ &\propto \tau^{\frac{n}{2}} e^{-[\frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2]\tau} \end{aligned}$$

Loglikelihood function

$$\ell(\tau|x_1, \ldots, x_n) \propto \frac{n}{2}\log\tau - [\frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2]\tau$$

- Maximising loglikelihood function by taking derivative, we have $\frac{d}{d\tau}\ell(\tau|x_1, \ldots, x_n) = \frac{n}{2\tau} - \frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2$
- Hence frequentists will take sample precision $\hat{\tau} = \frac{1}{n^{-1}\sum_{i=1}^{n}(x_i-h)^2}$, maximises loglikelihood function

Bayesian Inference

- Prior: $\tau \sim Gamma(\alpha, \frac{1}{\beta})$
- Observations: n iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ from $X \sim N(h, \frac{1}{\tau})$, where h is known
- Posterior: $(\tau|\mathbf{x}) \sim Gamma(\alpha_n, \frac{1}{\beta_n})$, where $\alpha_n = \alpha + \frac{n}{2}$ and $\beta_n = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2$
  - $\pi(\tau|x_1, \ldots, x_n) = \pi(\tau) \times L(\tau|x_1, \ldots, x_n) \propto \tau^{\alpha-1}e^{-\beta\tau} \times \tau^{n/2}e^{-[\frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2]\tau} \propto \tau^{\alpha+n/2-1}e^{-[\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2]\tau}$
- Estimate of population precision, $\hat{\tau} = E[\tau|\mathbf{x}] = \frac{\alpha_n}{\beta_n} = \frac{\alpha+\frac{n}{2}}{\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2}$
  - Incidentally, this is a weighted average of prior mean and MLE: $w_n \times \frac{\alpha}{\beta} + (1-w_n) \times \frac{n}{\sum_{i=1}^{n}(x_i-h)^2}$ where $w_n = \frac{\beta}{\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-h)^2}$
- As we deflate prior (letting α, $\beta \to 0$), posterior mean converges to $\frac{n}{\sum_{i=1}^{n}(x_i-h)^2}$; approximately inverse of variance $\frac{\sum_{i=1}^{n}(x_i-h)^2}{n}$
  - (By Taylor Expansion, we have $E[\frac{1}{X}] \approx \frac{1}{E[X]}$) — OK approximation when n is large

4.3 Normal Population with Unknown Mean and Variance

Let (μ, τ) be $Gamma - Normal(\alpha, \frac{1}{\beta}; m, \frac{1}{t})$, where $\tau \sim Gamma(\alpha, \frac{1}{\beta})$ and $(\mu|\tau) \sim N(m, \frac{1}{\tau t})$

- Then $\pi(\mu, \tau) \propto \tau^{\alpha-1}e^{-\beta\tau} \cdot \sqrt{\tau}e^{-\frac{\tau t}{2}(\mu-m)^2}$

Likelihood function

$$\begin{aligned} L(\mu, \tau|x_1, \ldots, x_n) &\propto \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x_i-\mu)^2} \\ &\propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum_{i=1}^{n}[(\mu-\bar{x})^2+2(\mu-\bar{x})(x_i-\bar{x})+(x_i-\bar{x})^2]} \text{ (can ignore middle term, can't ignore last term which depend} \\ &\propto \tau^{\frac{n}{2}} e^{-\frac{n\tau}{2}(\mu-\bar{x})^2-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{aligned}$$

From this, frequentists will derive MLEs of (μ, τ) to be $(\bar{x}, \frac{n}{\sum_{i=1}^n (x_i - \bar{x})^2})$

Proposition

- Prior: $(\mu, \tau) \sim Gamma - Normal(\alpha, \frac{1}{\beta}; m, \frac{1}{t})$
- Observations: n iid observations $\mathbf{x} = (x_1, \dots, x_n)$ from $x \sim N(\mu, \frac{1}{\tau})$, where μ, τ are both unknown
- Posterior: $(\mu, \tau|\mathbf{x}) \sim Gamma - Normal(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n})$ where:
  - $\alpha_n = \alpha + \frac{n}{2}$
  - $\beta_n = \beta + \frac{1}{2}[\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{(m-\bar{x})^2}{1/t+1/n}]$ — different
  - $t_n = t + n$
  - $m_n = \frac{t}{t+n} m + \frac{n}{t+n}\bar{x}$ — different (no r this time)
- Posterior: $(\mu|\mathbf{x}) \sim t_{2\alpha_n}(m_n, (\frac{\alpha_n t_n}{\beta_n})^{-1})$
  - i.e. t-distribution with $2\alpha_n$ degrees of freedom, location parameter m_n, precision parameter $\frac{\alpha_n t_n}{\beta_n}$
  - $\pi(\mu|\mathbf{x}) \propto \left[1 + (\frac{\alpha_n t_n}{\beta_n}) \frac{(\mu - m_n)^2}{2\alpha_n}\right]^{-(2\alpha_n+1)/2}$
  - Proportionality constant: $\frac{1}{B(\alpha_n, \frac{1}{2})} \frac{1}{\sqrt{2\alpha_n}} \sqrt{\frac{\alpha_n t_n}{\beta_n}}$
  - $E[\mu|\mathbf{x}] = m_n$
  - $Var(\mu|\mathbf{x}) = (\frac{\alpha_n t_n}{\beta_n})^{-1} \frac{2\alpha_n}{2\alpha_n - 2} = \frac{\beta_n}{t_n(\alpha_n - 1)}$

Proof of $(\mu, \tau|\mathbf{x})$ Gamma-Normal posterior:

$$\begin{aligned}
\pi(\mu, \tau|\mathbf{x}) &= \pi(\mu, \tau) \times L(\mu, \tau|\mathbf{x}) \\
&\propto \tau^{\alpha-1} e^{-\beta\tau} \cdot \sqrt{\tau} e^{-\frac{t\tau}{2}(\mu-m)^2} \times \tau^{\frac{n}{2}} e^{-\frac{n\tau}{2}(\mu-\bar{x})^2 - \frac{\tau}{2}\sum_{i=1}^n (x_i-\bar{x})^2} \text{ (just copy)} \\
&\propto \tau^{\alpha+\frac{n}{2}-\frac{1}{2}} \cdot e^{-[\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2]\tau - [\frac{t\tau}{2}(\mu-m)^2 + \frac{n\tau}{2}(\mu-\bar{x})^2]} \\
&= \tau^{\alpha+\frac{n}{2}-\frac{1}{2}} \cdot e^{-[\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2]\tau - [\frac{(t+n)\tau}{2}(\mu-m_n)^2 + \frac{\tau}{2}\frac{(m-\bar{x})^2}{1/t+1/n}]} \text{ (useful identity)} \\
&= \tau^{\alpha+\frac{n}{2}-1} e^{-[\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2 + \frac{1}{2}\frac{(m-\bar{x})^2}{1/t+1/n}]\tau} \times \sqrt{\tau} e^{-\frac{(t+n)\tau}{2}(\mu-m_n)^2}
\end{aligned}$$

Proof of $(\mu|\mathbf{x})$ t-distribution posterior:

We are given $(\tau|\mathbf{x}) \sim Gamma(\alpha_n, \frac{1}{\beta_n})$ and $(\mu|\tau, \mathbf{x}) \sim N(m_n, \frac{1}{\tau t_n})$.

From lecture 2 example 3 (about line 182 in this doc), we know that:

- $\pi(\mu) \propto \frac{\pi(\mu|\tau)}{\pi(\tau|\mu)} \propto \frac{\tau^{1/2} e^{-\frac{\tau t}{2}(\mu-m)^2}}{\beta_n^{\alpha_n} \tau^{\alpha_n - 1} e^{-\beta_n \tau}} \propto \beta_n^{-\alpha_n} = (\beta + \frac{t}{2}(\mu-m)^2)^{-(\alpha+\frac{1}{2})} \propto (1 + \frac{\alpha t}{\beta}\frac{(\mu-m)^2}{2\alpha})^{-(2\alpha+1)/2}$

$$\begin{aligned}
\pi(\mu|\mathbf{x}) &\propto [\beta_n + \frac{t_n}{2}(\mu - m_n)^2]^{-(\alpha_n+1/2)} \\
&\propto [1 + (\frac{\alpha_n t_n}{\beta_n})\frac{(\mu-m_n)^2}{2\alpha_n}]^{-(2\alpha_n+1)/2}
\end{aligned}$$

If we standardise and define $v = \sqrt{\frac{\alpha_n t_n}{\beta_n}}(\mu - m_n)$, then $v \sim t_{2\alpha_n}$ with mean 0 and variance $\frac{2\alpha_n}{2\alpha_n - 2}$.

4.3.1 Example

Suppose the following:

- Prior: $(\mu, \tau) \sim Gamma - Normal(1, \frac{1}{2}; 74, \frac{2}{3})$
- Observations: 36 observations with $\bar{x} = 82$ and $s^2 = 27$, approximately $X \sim N(\mu, \frac{1}{\tau})$
- What is the posterior distribution of (μ, τ)?
- What are the 90% prior and posterior intervals for μ?

Solving for posterior distribution of (μ, τ):

- $(\mu, \tau)|x_1, \ldots, x_n \sim Gamma - Normal(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n})$ where:
- $\alpha_n = \alpha + \frac{n}{2} = 19$
- $\beta_n = \beta + \frac{1}{2}(\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{(m-\bar{x})^2}{1/t+1/n}) = \ldots = 520.58$
- $t_n = t + n = 37.5$
- $m_n = \frac{t}{t+n}m + \frac{n}{t+n}\bar{x} = 81.68$

Solving for 90% prior interval for μ:

- $\mu \sim m + (\frac{\alpha t}{\beta})^{-1/2} t_{2\alpha}$
- 90% prior interval is $[m + (\frac{\alpha t}{\beta})^{-1/2} t_{2\alpha}(0.05), m + (\frac{\alpha t}{\beta})^{-1/2} t_{2\alpha}(0.95)] = [70.63, 77.37]$

Solving for 90% posterior interval for μ:

- $\mu|x_1, \ldots, x_n \sim m_n + (\frac{\alpha_n t_n}{\beta_n})^{-1/2} t_{2\alpha}$
- 90% posterior interval is $[m_n + (\frac{\alpha_n t_n}{\beta_n})^{-1/2} t_{2\alpha}(0.05), m_n + (\frac{\alpha_n t_n}{\beta_n})^{-1/2} t_{2\alpha}(0.95)] = [80.24, 83.12]$

5 Lecture 4: Conjugate Prior Distributions

5.1 Conjugate Family

Conjugate family: A class Π of probability distributions forms a *conjugate family* if the posterior density $\pi(\theta|x) \propto \pi(\theta) \cdot f(x|\theta)$ is in the class Π for all x, whenever the prior density $\pi(\theta)$ is in Π.

5.2 Examples of Conjugate Families

5.2.1 Normal family — for *mean* of Normal population with known variance

(Taken from before)

- Prior: assume $\mu \sim N(m, \frac{1}{t})$
- Observations: n iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ from $X \sim N(\mu, \frac{1}{r})$, where r is known
- Posterior: $(\mu|\mathbf{x}) \sim N(m_n, \frac{1}{t_n})$ where $t_n = t + nr$ and $m_n = w_n m + (1 - w_n)\bar{x}$, where $w_n = \frac{t}{t+nr}$ — weighing factor

5.2.2 Gamma family — for *precision* of Normal population with known mean

(Taken from before)

- Prior: assume $\tau \sim Gamma(\alpha, \frac{1}{\beta})$
- Observations: n iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ from $X \sim N(h, \frac{1}{\tau})$, where h is known
- Posterior: $(\tau|\mathbf{x}) \sim Gamma(\alpha_n, \frac{1}{\beta_n})$, where $\alpha_n = \alpha + \frac{n}{2}$ and $\beta_n = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - h)^2$

5.2.3 (Anti-Example) Exponential family — NOT for mean of Normal population with known variance

- Prior: exponential
- Posterior: truncated Normal distribution, not exponential

5.3 Bernoulli Distributions

Population observations: $X \sim Ber(\theta)$, where $0 < \theta < 1$ is the probability of success

Conjugate family: **Beta** family

- Prior: $\theta \sim Beta(a, b)$ i.e. $\pi(\theta) = \theta^{a-1}(1-\theta)^{b-1}$
- Likelihood: $X|\theta \sim Ber(\theta)$ i.e. $f(x|\theta) = \theta^x(1-\theta)^{1-x}$ — looks like kernel of Beta density when viewed in θ
- Posterior: $\theta|\mathbf{x} \sim Beta(a_n, b_n)$ where $a_n = a + n\bar{x}$, $b_n = b + n - n\bar{x}$, $n\bar{x} = \sum_{i=1}^{n} x_i$
  - $\pi(\theta|\mathbf{x}) \propto \pi(\theta) \cdot \prod_{i=1}^{n}[\theta^{x_i}(1-\theta)^{1-x_i}] \propto \theta^{a-1}(1-\theta)^{b-1} \cdot \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}\mathbf{I}_{0<\theta<1} = \theta^{a_n-1}(1-\theta)^{b_n-1}\mathbf{I}_{0<\theta<1}$
  - Update rule: add successes to a, add failures to b

5.3.1 Example

Question

- Prior: $\theta \sim Beta(a, b)$ with mean 0.55 and SD 0.04
- Observations: $X_1, \ldots, X_n \sim Ber(\theta)$, $n = 100$, observe 52 heads (and 48 tails)
- What is posterior distribution, and probability that θ is between 0.50 ± 0.05?
- (Note that standardised Beta RV is roughly $N(0, 1)$ by CLT)

Solving for prior a and b:

- $\mu_\theta = E(\theta) = \frac{a}{a+b} = 0.55$
- $\sigma_\theta^2 = Var(\theta) = \frac{ab}{(a+b)^2(a+b+1)} = 0.04^2$
- $\therefore a = 84.53,\ b = 69.16$

Solving for posterior a_n and b_n:

- $a_n = a + n\bar{x} = 136.53$
- $b_n = b + n - n\bar{x} = 117.16$

Solving for posterior probability that $\theta \in 0.50 \pm 0.05$:

- Using R: $P(\theta \in (0.45, 0.55)|\mathbf{x}) = pbeta(0.55, a_n, b_n) - pbeta(0.45, a_n, b_n) = 0.6436$
- Alternatively, using normal approx with CLT:
  - Posterior mean $\mu^* = E(\theta|\mathbf{x}) = \frac{a_n}{a_n+b_n} = 0.5382$
  - Posterior variance $\sigma^{*2} = Var(\theta|\mathbf{x}) = \frac{a_n b_n}{(a_n+b_n)^2(a_n+b_n+1)} = 0.0009759$
  - Then $P(\theta \in (0.45, 0.55)|\mathbf{x}) = \Phi(\frac{0.55-\mu^*}{\sqrt{\sigma^{*2}}}) - \Phi(\frac{0.45-\mu^*}{\sqrt{\sigma^{*2}}}) = \Phi(0.3785) - \Phi(-2.8227) = 0.6451$

5.4 Poisson Distributions

Population observations: $X \sim Po(\lambda)$, where $\lambda > 0$ is the mean or intensity

Conjugate family: **Gamma** family

- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$
- Likelihood: $X|\lambda \sim Po(\lambda)$ i.e. $f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ — looks like kernel of Gamma density when viewed in λ
- Posterior: $\lambda|\mathbf{x} \sim Gamma(\alpha_n, \frac{1}{\beta_n})$ where $\alpha_n = \alpha + n\bar{x}$, $\beta_n = \beta + n$, $n\bar{x} = \sum_{i=1}^n x_i$
  - $\pi(\lambda|\mathbf{x}) \propto \pi(\lambda) \cdot \prod_{i=1}^n [\lambda^{x_i} e^{-\lambda}] \propto \lambda^{\alpha-1} e^{-\beta\lambda} \cdot \lambda^{n\bar{x}} e^{-n\lambda} \mathbf{I}_{\lambda>0} = \lambda^{\alpha+n\bar{x}-1} e^{-(\beta+n)\lambda} \mathbf{I}_{\lambda>0}$
  - Update rule: add sum of observations to α, add number of observations to β

5.4.1 Example

Question

- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$ where mean is 4.4, SD is 0.4
- Observations: $X \sim Po(\lambda)$, where $n = 52$, $n\bar{x} = 257$
- What is the posterior density of λ, and what is the posterior probability that $\lambda > 5$?
- (Note that standardised Gamma RV is roughly $N(0, 1)$ by CLT)

Solving for prior α and β:

- $\mu_\lambda = E(\lambda) = \frac{\alpha}{\beta} = 4.4$
- $\sigma_\lambda^2 = Var(\lambda) = \frac{\alpha}{\beta^2} = 0.4^2$
- $\therefore \alpha = 121,\ \beta = 27.5$

Solving for posterior α_n and β_n:

- $\alpha_n = \alpha + n\bar{x} = 378$
- $\beta_n = \beta + n = 79.5$

Solving for posterior probability that $\lambda > 5$:

- Using R: $P(\lambda \geq 5|\mathbf{x}) = 1 - pgamma(5, \alpha_n, \beta_n) = 0.1579$
- Alternatively, using normal approx with CLT:
  - Posterior mean $\mu^* = E(\lambda|\mathbf{x}) = \frac{\alpha_n}{\beta_n} = 4.7547$
  - Posterior variance $\sigma^{*2} = Var(\lambda|\mathbf{x}) = \frac{\alpha_n}{\beta_n^2} = 0.0598$
  - Then $P(\lambda \geq 5|\mathbf{x}) = 1 - \Phi(\frac{5-\mu^*}{\sqrt{\sigma^{*2}}}) = 1 - \Phi(1.0031) = 0.1579$

5.5 Exponential Distributions

Population observations: $X \sim Exp(\lambda)$, where $\frac{1}{\lambda} > 0$ is the mean

Conjugate family: **Gamma** family

- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$
- Likelihood: $X|\lambda \sim Exp(\lambda)$ i.e. $f(x|\lambda) = \lambda e^{-\lambda x}$ where $x > 0$ — looks like kernel of Gamma density when viewed in λ
- Posterior: $\lambda|\mathbf{x} \sim Gamma(\alpha_n, \frac{1}{\beta_n})$ where $\alpha_n = \alpha + n$, $\beta_n = \beta + n\bar{x}$, $n\bar{x} = \sum_{i=1}^n x_i$
  - $\pi(\lambda|\mathbf{x}) \propto \pi(\lambda) \cdot \prod_{i=1}^n [\lambda e^{-\lambda x_i}] \propto \lambda^{\alpha-1} e^{-\beta\lambda} \cdot \lambda^n e^{-\lambda n\bar{x}} \mathbf{I}_{\lambda>0} = \lambda^{\alpha+n-1} e^{-(\beta+n\bar{x})\lambda} \mathbf{I}_{\lambda>0}$
  - Update rule: add number of observations to α, add sum of observations to β

5.5.1 Example

Question

- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$ where inverse of mean lifetimes of lightbulbs is 0.95, SD is 0.021 (recall that $\frac{1}{\lambda}$ is the mean of exponential)
- Observations: lifetime of lightbulb $X \sim Exp(\lambda)$ distribution. $n = 50$, $n\bar{x} = 46$
- What is the posterior density of λ, and what is the posterior probability that $\frac{1}{\lambda} \leq 0.925$?

Solving for prior α and β:

- $\mu_\lambda = E(\lambda) = \frac{\alpha}{\beta} = 0.95$
- $\sigma_\lambda^2 = Var(\lambda) = \frac{\alpha}{\beta^2} = 0.021^2$
- $\therefore \alpha = 2046$, $\beta = 2154$

Solving for posterior α_n and β_n:

- $\alpha_n = \alpha + n = 2096$
- $\beta_n = \beta + n\bar{x} = 2200$

Solving for posterior probability that $\frac{1}{\lambda} \leq 0.925$:

- Using R: $P(\frac{1}{\lambda} \leq 0.925|\mathbf{x}) = P(\lambda \geq \frac{1}{0.925}|\mathbf{x}) = 1 - pgamma(\frac{1}{0.925}, \alpha_n, \beta_n) = 1.6912 \times 10^{-9}$
- Alternatively, using normal approx with CLT:
  - Posterior mean $\mu^* = E(\lambda|\mathbf{x}) = \frac{\alpha_n}{\beta_n} = 0.9529$
  - Posterior variance $\sigma^{*2} = Var(\lambda|\mathbf{x}) = \frac{\alpha_n}{\beta_n^2} = 0.0004431$
  - Then $P(\lambda \geq \frac{1}{0.925}|\mathbf{x}) = 1 - \Phi(\frac{1/0.925-\mu^*}{\sqrt{\sigma^{*2}}}) = 1 - \Phi(6.1612) = 3.6098 \times 10^{-10}$

5.6 Uniform Distributions

Population observations: $X \sim U(0, \theta)$

Conjugate family: **Pareto** family

- Prior: $\theta \sim Pareto(m, a)$ i.e. $\pi(\theta) = \frac{am^a}{\theta^{a+1}} \mathbf{I}_{\theta > m}$
- Likelihood: $X|\theta \sim U(0, \theta)$ i.e. $f(x|\theta) = \frac{1}{\theta} \mathbf{I}_{0<x<\theta}$ — looks like kernel of Pareto density when viewed in λ
- Posterior: $\theta|\mathbf{x} \sim Pareto(m_n, a_n)$ where $m_n = \max(m, x_{\max})$, $a_n = a + n$, $x_{\max} = \max_{i=1}^{n} x_i$
  - $\pi(\theta|\mathbf{x}) \propto \frac{am^a}{\theta^{a+1}} \mathbf{I}_{\theta>m} \times \prod_{i=1}^{n} [\frac{1}{\theta} \mathbf{I}_{0<x_i<\theta}] \propto \frac{1}{\theta^{a+1}} \mathbf{I}_{\theta>m} \times \frac{1}{\theta^n} \mathbf{I}_{0<\max x_i<\theta} = \frac{1}{\theta^{a_n+1}} \mathbf{I}_{\theta>m_n}$
  - Update rule: set m to maximum of itself and observations, add number of observations to a

5.6.1 Example

Question

- Prior: $\theta \sim Pareto(m, a)$ where $m = 0.01$, $a = 1.7$
- Observations: $X \sim U(0, \theta)$ — sample is $\{0.2, 0.58, 0.1, 1.5, 2.4, 1.77\}$ i.e. $n = 6$, $x_{\max} = 2.4$
- What is the posterior density of θ, and what is the posterior probability that $\theta > 4$?

Solving for posterior m_n and a_n:

- $m_n = \max(m, x_{\max}) = 2.4$
- $a_n = a + n = 7.7$

Solving for posterior probability that $\theta > 4$:

- $P(\theta > 4|\mathbf{x}) = 1 - (1 - (\frac{m_n}{4})^{a_n}) = 0.01958$ (recall that in Pareto distribution, $F(x) = 1 - (\frac{m}{x})^a$ for $x > m$)

5.7 Multinomial Distributions

Population observations: $(X_1, \ldots, X_k) \sim Multinomial(n; p_1, \ldots, p_k)$ — multivariate generalisation of Binomial

- $f(x_1, \ldots, x_k|p_1, \ldots, p_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \; \mathbf{I}_{\sum_{i=1}^{k} x_i = n}$

Conjugate family: **Dirichlet** family — multivariate generalisation of Beta

- Prior: $(p_1, \ldots, p_k) \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$
  - $\pi(p_1, \ldots, p_k) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \ldots p_k^{\alpha_k - 1}$ where $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i = 1$
- Likelihood: M observations of $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$
- Posterior: $(p_1, \ldots, p_k|\mathbf{x}_1, \ldots, \mathbf{x}_M) \sim Dirichlet(\alpha_{n1}, \ldots, \alpha_{nk})$ where $\alpha_{ni} = \alpha_i + m_i$ and $m_j = \sum_{i=1}^{M} x_{ij}$
  - $\pi(p_1, \ldots, p_k|\mathbf{x}_1, \ldots, \mathbf{x}_M) \propto \pi(p_1, \ldots, p_k) \times \prod_{i=1}^{M} f(x_{i1}, \ldots, x_{ik}|p_1, \ldots, p_k) \propto p_1^{\alpha_1 + m_1 - 1} \ldots p_k^{\alpha_k + m_k - 1} \; \mathbf{I}_{0<p_i<1; \sum_{i=1}^{k}}$
  - Update rule: add sum of x across each category to the α for that category

6 Lecture 5: Predictive Distributions

6.1 Introduction

Idea: we not only want to estimate the unknown parameters, but make predictions/forecasts for new observations

- We want to know $P(a < X_{n+1} < b|\mathbf{X} = \mathbf{x})$ i.e. $F_{X_{n+1}}(b) - F_{X_{n+1}}(a)$
- Prediction for new observation depends on posterior distribution of θ

6.1.1 ($\star$) Proposition: Predictive Distribution

Proposition: predictive distribution of X_{n+1} based on $\mathbf{x}$ is:

$$\boxed{F_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) \text{ i.e. } P(X_{n+1} \leq x|\mathbf{X} = \mathbf{x}) = E[F(x|\theta)|\mathbf{X} = \mathbf{x}]}$$

$$\boxed{f_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) = E[f(x|\theta)|\mathbf{X} = \mathbf{x}] = \int_{\theta\in\Theta} f(x|\theta) \cdot \pi(\theta|\mathbf{X} = \mathbf{x})\ d\theta}$$

Proof of proposition

$$\begin{aligned}
F_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) &= P(X_{n+1} \leq x|\mathbf{X} = \mathbf{x}) \\
&= E[\mathbf{I}_{X_{n+1}\leq x}|\mathbf{X} = \mathbf{x}] \quad \text{— using fact that } P(X \leq x) = E[\mathbf{I}(X \leq x)] \\
&= E[E(\mathbf{I}_{X_{n+1}\leq x}|\theta, \mathbf{X} = \mathbf{x})|\mathbf{X} = \mathbf{x}] \\
&= E[E(\mathbf{I}_{X_{n+1}\leq x}|\theta)|\mathbf{X} = \mathbf{x}] \quad \text{— by independence assumption} \\
&= E[F(x|\theta)|\mathbf{X} = \mathbf{x}]
\end{aligned}$$

6.1.2 Useful Facts: Double Expectation Formula

Double expectation formula:

- Standard form: $E(Y) = E[E(Y|Z)]$ — inside $E(Y|Z)$ is expectation wrt Y yielding a function of Z, outer $E[E(Y|Z)]$ is expectation wrt Z
- Conditional form: $E(Y|X) = E[E(Y|Z, X)|X]$ — have to introduce conditional on X in both expectations

$$\begin{aligned}
E(Y) &= \int y \cdot f(y)\ dy \\
&= \int y \cdot \left[\int f(y, z)\ dz\right]\ dy \\
&= \int y \cdot \left[\int f(y|z) \cdot f(z)\ dz\right]\ dy \\
&= \int \left[\int y \cdot f(y|z)\ dz\right] \cdot f(z)\ dy \\
&= \int E(Y|z) \cdot f(z)\ dz \\
&= E[E(Y|z)]
\end{aligned}$$

Application of double expectation to variance: $Var(Y) = E[Var(Y|Z)] + Var(E[Y|Z])$

$$
\begin{aligned}
Var(Y) &= E(Y^2) - [E(Y)]^2 \\
&= E[E(Y^2|Z)] - (E[E(Y|Z)])^2 \\
&= E[E(Y^2|Z) - (E(Y|Z))^2 + (E(Y|Z))^2] - (E[E(Y|Z)])^2 \\
&= E[Var(Y|Z) + (E(Y|Z))^2] - (E[E(Y|Z)])^2 \\
&= E[Var(Y|Z)] + E[(E(Y|Z))^2] - (E[E(Y|Z)])^2 \\
&= E[Var(Y|Z)] + Var(E[Y|Z])^2
\end{aligned}
$$

6.2 Bernoulli Distributions

Previously known facts

- Prior: $\theta \sim Beta(a, b)$
- Likelihood: $X|\theta \sim Ber(\theta)$ where $f(x|\theta) = \theta^x(1-\theta)^{1-x}$
- Posterior: $\theta|\mathbf{X} = \mathbf{x} \sim Beta(a_n, b_n)$, where $a_n = a + n\bar{x}$ and $b_n = b + n - n\bar{x}$

Predictive density (mass in this case cos discrete)

- $P(X_{n+1} = 1|\mathbf{X} = \mathbf{x}) = E[f(1|\theta)|\mathbf{X} = \mathbf{x}] = E[\theta|\mathbf{X} = \mathbf{x}] = \frac{a_n}{a_n+b_n}$
- $P(X_{n+1} = 0|\mathbf{X} = \mathbf{x}) = E[f(0|\theta)|\mathbf{X} = \mathbf{x}] = E[1-\theta|\mathbf{X} = \mathbf{x}] = 1 - \frac{a_n}{a_n+b_n} = \frac{b_n}{a_n+b_n}$

Predictive variance

$$
\begin{aligned}
Var(X_{n+1}|\mathbf{X} = \mathbf{x}) &= P(X_{n+1} = 1|\mathbf{X} = \mathbf{x}) \times P(X_{n+1} = 0|\mathbf{X} = \mathbf{x}) \\
&= \frac{a_n}{a_n + b_n} \times \frac{b_n}{a_n + b_n} \\
&= \dots \\
&= (a_n + b_n)Var(\theta|\mathbf{X} = \mathbf{x}) + Var(\theta|\mathbf{X} = \mathbf{x})
\end{aligned}
$$

In general, predictive distribution's variance has two components:

- Error in estimating the parameter (from the posterior distribution)
- Uncertainty due to randomness of future value (from the model)

6.2.1 Example

Suppose $X_i = 1$ if the sun rises on the i-th day, and suppose we have a uniform prior. After observing sunrises on 500 days, how certain are you that the sun will rise tomorrow?

- Prior: $\theta \sim Beta(a, b)$ with $a = 1$ and $b = 1$
- Likelihood: $x|\theta \sim Ber(\theta)$
- Posterior: $\theta|\mathbf{X} = \mathbf{x} \sim Beta(a_n, b_n)$ with $a_n = 501$ and $b_n = 1$
- Predictive probability: $P(X_{n+1} = 1|\mathbf{X} = \mathbf{x}) = \frac{a_n}{a_n+b_n} = \frac{501}{502}$

6.3 Exponential Distributions

Previously known facts

- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$

- Likelihood: $X|\lambda \sim Exp(\lambda)$ where $f(x|\lambda) = \lambda e^{-\lambda x}$
- Posterior: $\lambda|\mathbf{X} = \mathbf{x} \sim Gamma(\alpha_n, \frac{1}{\beta_n})$ where $\alpha_n = \alpha + n$, $\beta_n = \beta + n\bar{x}$

Predictive density: **Pareto**

$$\begin{aligned}
f_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) &= E[f(x|\lambda)|\mathbf{X} = \mathbf{x}] \\
&= E[\lambda e^{-\lambda x}|\mathbf{X} = \mathbf{x}] \\
&= \int_0^\infty \lambda e^{-\lambda x} \cdot \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \lambda^{\alpha_n - 1} e^{-\beta_n \lambda} \, d\lambda \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \int_0^\infty \lambda^{(\alpha_n+1)-1} e^{-(x+\beta_n)\lambda} \, d\lambda \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \cdot \frac{\Gamma(\alpha_n + 1)}{(x+\beta_n)^{\alpha_n+1}} \\
&= \frac{\alpha_n \beta_n^{\alpha_n}}{(x+\beta_n)^{\alpha_n+1}}
\end{aligned}$$

Very similar to Pareto distribution!

- (To see this, if we transform to let $Y_{n+1} = X_{n+1} + \beta_n$, then $f_{Y_{n+1}}(y|\mathbf{X} = \mathbf{x}) = \frac{\alpha_n \beta_n^{\alpha_n}}{y^{\alpha_n+1}}$ which is Pareto density)
- $X_{n+1} + \beta_n|\mathbf{X} = \mathbf{x} \sim Pareto(\beta_n, \alpha_n)$
- $P(X_{n+1} + \beta_n \le x + \beta_n|\mathbf{X} = \mathbf{x}) = 1 - \left(\frac{\beta_n}{x+\beta_n}\right)^{\alpha_n}$
- i.e. $P(X_{n+1} \le x|\mathbf{X} = \mathbf{x}) = 1 - \left(\frac{\beta_n}{x+\beta_n}\right)^{\alpha_n}$

6.3.1 Example

Suppose we have prior $\lambda \sim Gamma(1, 1)$, observations have distribution $X|\lambda \sim Exp(\lambda)$ (mean is $\frac{1}{\lambda}$).

- (i) Suppose we have collected no data. What is the predictive probability that the new observation is <8?
- (ii) Suppose we collect 10 observations with sum = 98. What is the new predictive probability that the new observation is <8?
- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$ where $\alpha = 1$ and $\beta = 1$
- Likelihood: $X|\lambda \sim Exp(\lambda)$
- Posterior with data in part (ii): $\lambda|\mathbf{X} = \mathbf{x} \sim Gamma(\alpha_n, \frac{1}{\beta_n})$ where $\alpha_n = \alpha + n = 11$ and $\beta_n = \beta + n\bar{x} = 99$
- Answer for (i): $X_{n+1} + \beta \sim Pareto(\beta, \alpha)$, so $P(X_{n+1} < 8) = P(X_{n+1} + \beta < 8 + \beta) = 1 - \left(\frac{\beta}{8+\beta}\right)^{\alpha} = \frac{8}{9}$
- Answer for (ii): $X_{n+1} + \beta_n|\mathbf{X} = \mathbf{x} \sim Pareto(\beta_n, \alpha_n)$, so $P(X_{n+1} < 8|\mathbf{X} = \mathbf{x}) = P(X_{n+1} + \beta_n < 8 + \beta_n|\mathbf{X} = \mathbf{x}) = 1 - \left(\frac{\beta_n}{8+\beta_n}\right)^{\alpha_n} = 0.5746$

6.4 Normal Distribution with Known Variance: $N(\mu, \frac{1}{r})$ with r known

Previously known facts

- Prior: $\mu \sim N(m, \frac{1}{t})$
- Likelihood: $X|\mu \sim N(\mu, \frac{1}{r})$
- Posterior: $\mu|\mathbf{X} = \mathbf{x} \sim N(m_n, \frac{1}{t_n})$ where $m_n = \frac{tm + nr\bar{x}}{t + nr}$ and $t_n = t + nr$

- Predictive: $X_{n+1}|\mathbf{X}=\mathbf{x} \sim N(m_n, t_n^{-1}+r^{-1})$

$$
\begin{aligned}
f_{X_{n+1}}(x|\mathbf{X}=\mathbf{x}) &= E[f(x|\mu)|\mathbf{X}=\mathbf{x}] \\
&= \int_{-\infty}^{\infty} f(x|\mu)\cdot\pi(\mu|\mathbf{X}=\mathbf{x})\,d\mu \text{ (since } \mu \text{ is the RV in the expectation)} \\
&= \int_{-\infty}^{\infty} \sqrt{\frac{r}{2\pi}} e^{-\frac{r}{2}(x-\mu)^2}\cdot\sqrt{\frac{t_n}{2\pi}} e^{-\frac{t_n}{2}(\mu-m_n)^2}\;d\mu \\
&= \frac{\sqrt{t_n r}}{2\pi}\int_{-\infty}^{\infty} e^{-[\frac{r}{2}(\mu-x)^2+\frac{t_n}{2}(\mu-m_n)^2]}\;d\mu \\
&= \frac{\sqrt{t_n r}}{2\pi}\int_{-\infty}^{\infty} e^{-[\frac{r+t_n}{2}(\mu-\bar{m}_n)^2+\frac{1}{2(r^{-1}+t_n^{-1})}(x-m_n)^2]}\;d\mu \text{ (using useful identity)} \\
&\text{where } \bar{m}_n = \frac{rx+t_n m_n}{r+t_n} \\
&= \frac{\sqrt{t_n r}}{2\pi} e^{-\frac{1}{2(r^{-1}+t_n^{-1})}(x-m_n)^2}\cdot\int_{-\infty}^{\infty} e^{-\frac{r+t_n}{2}(\mu-\bar{m}_n)^2}\;d\mu \\
&\text{Since the integral on the right } = \sqrt{\frac{2\pi}{t_n+r}} \text{ by normal density,} \\
&= \sqrt{\frac{t_n r}{2\pi(t_n+r)}} e^{-\frac{1}{2(t_n^{-1}+r^{-1})}(x-m_n)^2} \\
&= \sqrt{\frac{(t_n^{-1}+r^{-1})^{-1}}{2\pi}} e^{-\frac{(t_n^{-1}+r^{-1})^{-1}}{2}(x-m_n)^2}
\end{aligned}
$$

Remarks

- Note that both posterior and predictive have the same mean
- But predictive is *more variable* than posterior: has additional term r^{-1}

6.4.1 Example

Problem setup

- Prior: $\theta \sim N(m, \frac{1}{t})$ with $m=0$, $t=1$
- Likelihood: $x|\theta \sim N(\mu, \frac{1}{r})$ with $r=1$
- Observations: $n=10$, $\bar{x}=16.34$

Results

- Posterior distribution: $\theta|\mathbf{X}=\mathbf{x} \sim N(m_n, \frac{1}{t_n})$ with $m_n = \frac{tm+nr\bar{x}}{t+nr} = 14.8546$ and $t_n = t+nr = 11$
- Predictive distribution: $X_{n+1}|\mathbf{X}=\mathbf{x} \sim N(m_n, \frac{1}{t_n}+\frac{1}{r})$ with $m_n = 14.8546$ and $\frac{1}{t_n}+\frac{1}{r} = \frac{12}{11}$
- 95% posterior interval: $m_n \pm z_{0.975}\times\sqrt{\frac{1}{t_n}} = 14.8546 \pm 1.96\times\sqrt{\frac{1}{11}} = [14.26, 15.45]$
- 95% predictive interval: $m_n \pm z_{0.975}\times\sqrt{\frac{1}{t_n}+\frac{1}{r}} = 14.8546 \pm 1.96\times\sqrt{\frac{12}{11}} = [12.81, 16.90]$

6.5 Normal Distribution with Unknown Mean and Variance: $N(\mu, \frac{1}{\tau})$ with μ and τ unknown

Previously known facts

- Prior: $(\mu, \tau) \sim Gamma - Normal(\alpha, \frac{1}{\beta}; m, \frac{1}{t})$, i.e. $\tau \sim Gamma(\alpha, \frac{1}{\beta})$ and $\mu \sim N(m, \frac{1}{\tau t})$
- Likelihood: $X|(\mu, \tau) \sim N(\mu, \frac{1}{\tau})$
- Posterior: $(\mu, \tau)|\mathbf{X} = \mathbf{x} \sim N(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n})$
  - $\alpha_n = \alpha + \frac{n}{2}$
  - $\beta_n = \beta + \frac{1}{2}[\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{(m-\bar{x})^2}{1/t+1/n}]$
  - $t_n = t + n$
  - $m_n = \frac{t}{t+n} m + \frac{n}{t+n} \bar{x}$
- Marginal posterior: $\mu|\mathbf{X} = \mathbf{x} \sim t_{2\alpha_n}(m_n, (\frac{\alpha_n t_n}{\beta_n})^{-1})$
  - Density: $\pi(\mu|\mathbf{x}) \propto \left[1 + (\frac{\alpha_n t_n}{\beta_n}) \frac{(\mu - m_n)^2}{2\alpha_n}\right]^{-(2\alpha_n+1)/2}$
  - Density's proportionality constant: $\frac{1}{B(\alpha_n, \frac{1}{2})} \frac{1}{\sqrt{2\alpha_n}} \sqrt{\frac{\alpha_n t_n}{\beta_n}}$

Predictive density

$$
\begin{aligned}
f_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) &= E[f(x|\mu, \tau)|\mathbf{X} = \mathbf{x}] \\
&= \int_0^\infty \int_{-\infty}^\infty f(x|\mu, \tau) \times \pi(\mu, \tau|\mathbf{X} = \mathbf{x}) \; d\mu \; d\tau \\
&= \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x-\mu)^2} \times \sqrt{\frac{\tau t_n}{2\pi}} e^{-\frac{\tau t_n}{2}(\mu - m_n)^2} \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \tau^{\alpha_n - 1} e^{-\beta_n \tau} \; d\mu \; d\tau \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \frac{\sqrt{t_n}}{2\pi} \int_0^\infty \left[\int_{-\infty}^\infty e^{-\frac{\tau}{2}[(\mu-x)^2 + t_n(\mu - m_n)^2]} \; d\mu\right] \tau^{\alpha_n} e^{-\beta_n \tau} \; d\tau \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \frac{\sqrt{t_n}}{2\pi} \int_0^\infty S(\tau) \cdot \tau^{\alpha_n} e^{-\beta_n \tau} \; d\tau \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \frac{\sqrt{t_n}}{2\pi} \int_0^\infty \sqrt{\frac{2\pi}{(t_n + 1)\tau}} e^{-\frac{(x-m_n)^2}{2(t_n^{-1}+1)}\tau} \cdot \tau^{\alpha_n} e^{-\beta_n \tau} \; d\tau \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \sqrt{\frac{t_n}{2\pi(1 + t_n)}} \int_0^\infty \tau^{\alpha_n + \frac{1}{2} - 1} e^{-[\beta_n + \frac{t_n}{2(1+t_n)}(x - m_n)^2]} \; d\tau \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \sqrt{\frac{t_n}{2\pi(1 + t_n)}} \frac{\Gamma(\alpha_n + \frac{1}{2})}{[\beta_n + \frac{t_n}{2(1+t_n)}(x - m_n)^2]^{(2\alpha_n+1)/2}} \\
&= \frac{\Gamma((2\alpha_n + 1)/2)}{\Gamma(2\alpha_n/2)} \sqrt{\frac{1}{2\alpha_n \pi} \left(\frac{\alpha_n}{\beta_n} \frac{t_n}{1 + t_n}\right)} \left[1 + \frac{1}{2\alpha_n} \left(\frac{\alpha_n}{\beta_n} \frac{t_n}{1 + t_n}\right) (x - m_n)^2\right]^{-(2\alpha_n+1)/2} \\
&= \frac{1}{Beta(\alpha_n, \frac{1}{2})} \sqrt{\frac{\alpha_n t_n}{\beta_n(1 + t_n)}} \frac{1}{\sqrt{2\alpha_n}} \left[1 + \frac{\alpha_n t_n}{\beta_n(1 + t_n)} \frac{(x - m_n)^2}{2\alpha_n}\right]^{-\frac{2\alpha_n+1}{2}}
\end{aligned}
$$

$$
\begin{aligned}
\text{where } S(\tau) &= \int_{-\infty}^{\infty} e^{-\frac{\tau}{2}[(\mu-x)^2+t_n(\mu-m_n)^2]}\ d\mu \\
&= \int_{-\infty}^{\infty} e^{-\frac{\tau}{2}[(t_n+1)(\mu-\frac{t_nm_n+x}{t_n+1})^2+\frac{(x-m_n)^2}{t_n^{-1}+1}]}\ d\mu \\
&= \int_{-\infty}^{\infty} e^{-\frac{(t_n+1)\tau}{2}(\mu-\frac{t_nm_n+x}{t_n+1})^2-\frac{(x-m_n)^2}{2(t_n^{-1}+1)}\tau}\ d\mu \\
&= \int_{-\infty}^{\infty} e^{-\frac{(t_n+1)\tau}{2}(\mu-\frac{t_nm_n+x}{t_n+1})^2}\ d\mu\ \cdot e^{\frac{(x-m_n)^2}{2(t_n^{-1}+1)}\tau} \\
&= \sqrt{\frac{2\pi}{(t_n+1)\tau}} e^{-\frac{(x-m_n)^2}{2(t_n^{-1}+1)}\tau}
\end{aligned}
$$

Hence $X_{n+1}|\mathbf{X}=\mathbf{x} \sim t_{2\alpha_n}\left[m_n, (\frac{\alpha_n t_n}{\beta_n(1+t_n)})^{-1}\right]$, i.e. $X_{n+1}|\mathbf{X}=\mathbf{x} = m_n + (\frac{\alpha_n t_n}{\beta_n(1+t_n)})^{-1/2}\ t_{2\alpha_n}$

Remarks

- Compare the posterior and predictive distributions: $\mu|\mathbf{X}=\mathbf{x} \sim t_{2\alpha_n}\left[m_n, (\frac{\alpha_n t_n}{\beta_n})^{-1}\right]$ while $X_{n+1}|\mathbf{X}=\mathbf{x} \sim t_{2\alpha_n}\left[m_n, (\frac{\alpha_n t_n}{\beta_n(1+t_n)})^{-1}\right]$
- Note that both share the same mean, but $X_{n+1}|\mathbf{X}=\mathbf{x}$ has *more variance* than $\mu|\mathbf{X}=\mathbf{x}$
  - $Var(\mu|\mathbf{x}) = (\frac{\alpha_n t_n}{\beta_n})^{-1}\frac{2\alpha_n}{2\alpha_n-2} = \frac{\beta_n}{t_n(\alpha_n-1)}$
  - $Var(X_{n+1}|\mathbf{x}) = (\frac{\alpha_n t_n}{\beta_n(1+t_n)})^{-1}\frac{2\alpha_n}{2\alpha_n-2} = \frac{\beta_n}{t_n(\alpha_n-1)} + \frac{\beta_n}{\alpha_n-1} = Var(\mu|\mathbf{x}) + \frac{\beta_n}{\alpha_n+1}$

6.5.1 Example

Problem setup

- Prior: $(\mu,\tau) \sim Gamma-Normal(\alpha, \frac{1}{\beta}; m, \frac{1}{t})$ with $\alpha=1$, $\beta=2$, $m=74$, $t=\frac{3}{2}$
- Likelihood: $x|\mu,\tau \sim N(\mu, \frac{1}{\tau})$
- Observations: $n=36$, $\bar{x}=82$, $s^2=27 \to \sum_{i=1}^{n}(x_i-\bar{x})^2 = (n-1)s^2 = 35\times 27$

Results

- Posterior distribution of (μ,τ): $(\mu,\tau)|\mathbf{X}=\mathbf{x} \sim Gamma-Normal(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n})$ where $\alpha_n=19$, $\beta_n=520.58$, $t=37.5$, $m_n=81.68$
- Posterior distribution of μ: $\mu|\mathbf{X}=\mathbf{x} \sim t_{2\alpha_n}\left(m_n, (\frac{\alpha_n t_n}{\beta_n})^{-1}\right)$
- Predictive distribution: $X_{n+1}|\mathbf{X}=\mathbf{x} \sim t_{2\alpha_n}\left(m_n, (\frac{\alpha_n t_n}{\beta_n(1+t_n)})^{-1}\right)$
- Predictive probability that new observation greater than 82, $P(X_{n+1}\geq 82|\mathbf{X}=\mathbf{x})$:
  - Using R: $1-pt(\sqrt{\frac{\alpha_n t_n}{\beta_n(1+t_n)}}(\bar{x}-m_n), 2\alpha_n) = 0.4761$
  - Using standardisation: $1-P(X_{n+1}\leq 82|\mathbf{X}=\mathbf{x}) = 1-P\left(t_{2\alpha_n}\leq (82-m_n)\times\sqrt{\frac{\alpha_n t_n}{\beta_n(1+t_n)}}\right) = 1-P(t_{2\alpha_n}\leq 0.060334) = 0.4761$
- 90% predictive interval: $m_n \pm t_{2\alpha_n,0.95}\times(\frac{\alpha_n t_n}{\beta_n(1+t_n)})^{-1/2} = 81.68\pm 8.94 = [72.74, 90.62]$
- 90% posterior interval for μ: $m_n \pm t_{2\alpha_n,0.95}\times(\frac{\alpha_n t_n}{\beta_n})^{-1/2} = 81.68\pm 1.44 = [80.24, 83.12]$

7 Lecture 6: Hypothesis Testing: One-Sample Problem

7.1 Introduction

Hypothesis test: procedure to make a decision about parameter θ, on choosing between two hypotheses $\{\theta \in \Theta_1\}$ or $\{\theta \in \Theta_2\}$ (disjoint subsets of parameter space Θ)

Bayesian approach to hypothesis test: compare posterior probabilities of events $\{\theta \in \Theta_1\}$ and $\{\theta \in \Theta_2\}$, i.e. $P(\theta \in \Theta_1|\boldsymbol{X})$ vs $P(\theta \in \Theta_2|\boldsymbol{X})$

- Prior probabilities must be >0

7.2 Test between Two Parameter Values: $\{\theta = \theta_1\}$ or $\{\theta = \theta_2\}$

Prior: two-point distribution

- $P(\theta = \theta_1) = p$
- $P(\theta = \theta_2) = 1 - p$

Posterior: two-point distribution

- $P(\theta = \theta_1|\boldsymbol{x}) \propto p \cdot \pi(\boldsymbol{x}|\theta_1)$
- $P(\theta = \theta_2|\boldsymbol{x}) \propto (1 - p) \cdot \pi(\boldsymbol{x}|\theta_2)$
- Normalisation constant: $p\pi(\boldsymbol{x}|\theta_1) + (1 - p)\pi(\boldsymbol{x}|\theta_2)$

Prior and Posterior Odds on θ_1 against θ_2

- Prior odds: $O = \frac{P(\theta=\theta_1)}{P(\theta=\theta_2)}$
- Posterior odds: $O_n = \frac{P(\theta=\theta_1|\boldsymbol{x})}{P(\theta=\theta_2|\boldsymbol{x})}$
- Then $P(\theta = \theta_1|\boldsymbol{x}) = \frac{O_n}{1+O_n}$, $P(\theta = \theta_2|\boldsymbol{x}) = \frac{1}{1+O_n}$

Hypothesis test

- In favour of θ_1 if $P(\theta = \theta_1|\boldsymbol{x}) > 0.5 \Leftrightarrow O_n > 1$
- In favour of θ_2 if $P(\theta = \theta_2|\boldsymbol{x}) > 0.5 \Leftrightarrow O_n < 1$
- In general, we can calculate either posterior probabilities or posterior odds to conduct hypothesis test

7.2.1 Example

Ahmad believes that probability of stock going up (vs down) in any given day is 0.5. Jamal believes that probability of stock going up (vs down) in any given day is 0.75. Is Ahmad's claim more favourable if there are 62 up-days in the past 100 days?

- Hypotheses: $H_0 : \theta = \frac{1}{2}$ (for Ahmad), $H_1 : \theta = \frac{3}{4}$ (for Jamal)
- Prior: uniform i.e. $P(\theta = \frac{1}{2}) = \frac{1}{2}$ and $P(\theta = \frac{3}{4}) = \frac{1}{2}$
- Model density: $X \sim Ber(\theta)$, where $X = 1$ means stock goes up, $X = 0$ means stock goes down
- Likelihood: $L(\theta|\boldsymbol{x}) \propto \theta^{n\bar{x}}(1 - \theta)^{n-n\bar{x}}$
- Observations: $n\bar{x} = 62$, $n = 100$
- Posterior
  - $P(\theta = \frac{1}{2}|\boldsymbol{x}) \propto P(\theta = \frac{1}{2}) \cdot L(\theta = \frac{1}{2}|\boldsymbol{x}) = \frac{1}{2} \cdot (\frac{1}{2})^{62}(1 - \frac{1}{2})^{38}$
  - $P(\theta = \frac{3}{4}|\boldsymbol{x}) \propto P(\theta = \frac{3}{4}) \cdot L(\theta = \frac{3}{4}|\boldsymbol{x}) = \frac{1}{2} \cdot (\frac{3}{4})^{62}(1 - \frac{3}{4})^{38}$

- $O_n = \frac{P(\theta=\frac{1}{2}|\boldsymbol{x})}{P(\theta=\frac{3}{4}|\boldsymbol{x})} = \frac{1/2^{100}}{3^{62}/4^{100}} = 3.3226$

- Therefore H_0 (Ahmad's belief) is more favourable

7.3 Test between Two Parameter Subsets: $\{\theta \in \Theta_1\}$ or $\{\theta \in \Theta_2\}$

- We assume that Θ_1 and Θ_2 are *disjoint*
- Prior: $\pi(\theta)$
- Posterior: $\pi(\theta|\boldsymbol{x}) \propto \pi(\theta) \cdot \pi(\boldsymbol{x}|\theta) = \pi(\theta) \cdot L(\theta|\boldsymbol{x})$
- Suppose we use *conjugate prior* for θ, so posterior distribution of θ is nice and in same parametric family

7.3.1 Case 1: $\Theta_1 \cup \Theta_2 = \Theta$, i.e. span entire parametric space

Posterior

- $P(\theta \in \Theta_1|\boldsymbol{x}) = \int_{\theta\in\Theta_1} \pi(\theta|\boldsymbol{x})\ d\theta$
- $P(\theta \in \Theta_2|\boldsymbol{x}) = \int_{\theta\in\Theta_2} \pi(\theta|\boldsymbol{x})\ d\theta$
- Can be obtained easily in many ways, e.g. looking up statistical tables, normal approximation

7.3.2 Case 2: $\Theta_1 \cup \Theta_2 \neq \Theta$, i.e. do NOT span parametric space

Posterior: probabilities need to be re-normalised

- $P(\theta \in \Theta_1|\boldsymbol{x}, \theta \in \Theta_1 \cup \Theta_2) = \frac{P(\theta\in\Theta_1|\boldsymbol{x})}{P(\theta\in\theta_1|\boldsymbol{x})+P(\theta\in\theta_2|\boldsymbol{x})}$
- $P(\theta \in \Theta_2|\boldsymbol{x}, \theta \in \Theta_1 \cup \Theta_2) = \frac{P(\theta\in\Theta_2|\boldsymbol{x})}{P(\theta\in\theta_1|\boldsymbol{x})+P(\theta\in\theta_2|\boldsymbol{x})}$

7.3.3 Prior and Posterior Odds

- Prior odds: $O = \frac{\int_{\Theta_1} \pi(\theta)\ d\theta}{\int_{\theta_2} \pi(\theta)\ d\theta}$ — defined implicitly by choice of prior for θ
- Posterior odds: $O_n = \frac{\int_{\Theta_1} \pi(\theta|\boldsymbol{x})\ d\theta}{\int_{\theta_2} \pi(\theta|\boldsymbol{x})\ d\theta}$

7.3.4 Mixture Priors

$$\boxed{\begin{aligned} P(\theta \in \Theta_1|\boldsymbol{x}) \propto P(\theta \in \Theta_1) \cdot \int_{\Theta_1} \pi(\theta|\theta \in \Theta_1) \cdot f(\boldsymbol{x}|\theta)\ d\theta \\ P(\theta \in \Theta_2|\boldsymbol{x}) \propto P(\theta \in \Theta_2) \cdot \int_{\Theta_2} \pi(\theta|\theta \in \Theta_2) \cdot f(\boldsymbol{x}|\theta)\ d\theta \end{aligned}}$$

- Re-express the posterior: $P(\theta \in \Theta_2|\boldsymbol{x}) \propto \int_{\Theta_2} \pi(\theta) \cdot f(\boldsymbol{x}|\theta)\ d\theta \propto P(\theta \in \Theta_2) \cdot \int_{\Theta_2} \pi(\theta|\theta \in \Theta_2) \cdot f(\boldsymbol{x}|\theta)\ d\theta$
- $\pi(\theta|\theta \in \Theta_2)$: proper density restricted over Θ_2 by a re-normalisation with $\frac{\pi(\theta)}{P(\theta\in\Theta_2)}$
- $\pi(\theta)$: has total mass $P(\theta \in \Theta_2) = \int_{\Theta_2} \pi(\theta)\ d\theta$

Consequences of mixture priors:

- We can choose $P(\theta \in \Theta_1)$ and $P(\theta \in \Theta_2)$ arbitrarily, as long as they add up to 1
- We can assume different prior densities $\pi(\theta|\theta \in \Theta_1)$ and $\pi(\theta|\theta \in \Theta_2)$ over the different subsets Θ_1 and Θ_2

7.3.5 Example 1 (non-mixture vs mixture prior)

Ahmad believes that probability of stock going up (vs down) in any given day is <0.54. Jamal believes that probability of stock going up (vs down) in any given day is >0.70. Is Ahmad's claim more favourable if there are 62 up-days in the past 100 days?

- (i) Assume a uniform prior
- (ii) Assume a mixture prior: $P(\theta < 0.54) = P(\theta > 0.7) = \frac{1}{2}$, and both $\pi(\theta|\theta < 0.54)$ and $\pi(\theta|\theta > 0.7)$ are uniform
- Hypotheses: $H_0 : \theta < 0.54$ (for Ahmad), $H_1 : \theta > 0.70$ (for Jamal)
- Prior (i): uniform i.e. $\theta \sim Uniform(0, 1)$
- Prior (ii): mixture where $P(\theta < 0.54) = P(\theta > 0.7) = \frac{1}{2}$, and $\theta|\theta < 0.54 \sim Uniform(0, 0.54)$, $\theta|\theta > 0.7 \sim Uniform(0.7, 1)$
- Model density: $X \sim Ber(\theta)$, where $X = 1$ means stock goes up, $X = 0$ means stock goes down
- Likelihood: $L(\theta|\boldsymbol{x}) \propto \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$
- Observations: $n\bar{x} = 62$, $n = 100$
- Posterior (i): favours H_0
  - $\theta|\boldsymbol{x} \sim Beta(a_n, b_n)$ where $a_n = 63$ and $b_n = 39$
  - Using R:
    - $P(\theta < 0.54|\boldsymbol{x}) \propto \int_0^{0.54} \theta^{62}(1-\theta)^{38}\, d\theta \propto pbeta(0.54, 62+1, 38+1) = 0.05531$
    - $P(\theta > 0.70|\boldsymbol{x}) \propto \int_{0.70}^{1} \theta^{62}(1-\theta)^{38}\, d\theta \propto pbeta(0.70, 62+1, 38+1) = 0.03970$
    - $O_n = \frac{P(\theta<0.54|\boldsymbol{x})}{P(\theta>0.70|\boldsymbol{x})} = \frac{0.05531}{0.03970} = 1.3933$
  - Using normal approximation: $\mu^* = \frac{a_n}{a_n+b_n} = 0.6176$ and $\sigma^{*2} = \frac{a_n b_n}{(a_n+b_n)^2(a_n+b_n+1)} = 0.002293$
    - $P(\theta < 0.54|\boldsymbol{x}) \propto \Phi(\frac{0.54-\mu^*}{\sigma^*}) = 0.05245$
    - $P(\theta > 0.70|\boldsymbol{x}) \propto 1 - \Phi(\frac{0.70-\mu^*}{\sigma^*}) = 0.04273$
    - $O_n = \frac{P(\theta<0.54|\boldsymbol{x})}{P(\theta>0.70|\boldsymbol{x})} = \frac{0.05245}{0.04273} = 1.2274$
  - Therefore H_0 (Ahmad's belief) is more favourable
- Posterior (ii): favours H_1
  - Using R:
    - $P(\theta < 0.54|\boldsymbol{x}) \propto \frac{1}{2}\int_0^{0.54} \frac{1}{0.54-0}\theta^{62}(1-\theta)^{38}\, d\theta \propto pbeta(0.54, 62+1, 38+1)/1.08 = 0.05121$
    - $P(\theta > 0.70|\boldsymbol{x}) \propto \frac{1}{2}\int_{0.70}^{1} \frac{1}{1-0.70}\theta^{62}(1-\theta)^{38}\, d\theta \propto (1 - pbeta(0.70, 62+1, 38+1))/0.60 = 0.06617$
    - $O_n = \frac{P(\theta<0.54|\boldsymbol{x})}{P(\theta>0.70|\boldsymbol{x})} = \frac{0.05121}{0.6617} = 0.7739$
  - Using normal approx on $\theta|\boldsymbol{x} \sim Beta(a_n, b_n)$ with $a_n = 63$ and $b_n = 39$
    - So $\mu^* = \frac{a_n}{a_n+b_n} = 0.6176$ and $\sigma^{*2} = \frac{a_n b_n}{(a_n+b_n)^2(a_n+b_n+1)} = 0.002293$
    - $P(\theta < 0.54|\boldsymbol{x}) \propto \frac{1}{2}\int_0^{0.54} \frac{1}{0.54-0}\theta^{62}(1-\theta)^{38}\, d\theta \propto \Phi(\frac{0.54-\mu^*}{\sigma^*})/1.08 = 0.04856$
    - $P(\theta > 0.70|\boldsymbol{x}) \propto \frac{1}{2}\int_{0.70}^{1} \frac{1}{1-0.70}\theta^{62}(1-\theta)^{38}\, d\theta \propto (1 - \Phi(\frac{0.70-\mu^*}{\sigma^*}))/0.60 = 0.07122$
    - $O_n = \frac{P(\theta<0.54|\boldsymbol{x})}{P(\theta>0.70|\boldsymbol{x})} = \frac{0.04856}{0.07122} = 0.6818$

– Therefore H_1 (Jamal's belief) is more favourable

7.3.6 Example 2 (simple)

Normal observations

- Hypothesis (i): $H_0 : \theta > 77$ vs $H_1 : \theta \leq 77$
- Hypothesis (ii): $H_0 : \theta < 77$ vs $H_1 : \theta > 80$
- Prior: θ is flat, i.e. $\pi(\theta) \propto 1$
- Model: $X \sim N(\theta, \frac{1}{r})$ where $r = \frac{1}{200}$
  - $L(\theta|\boldsymbol{X} = \boldsymbol{x}) \propto e^{-\frac{nr}{2}(\theta - \bar{x})^2}$
- Observations: $n = 100$, $\bar{x} = 78.5$
- Posterior: $\theta|\boldsymbol{X} = \boldsymbol{x} \sim N(\bar{x}, \frac{1}{nr})$
  - $\pi(\theta|\boldsymbol{X} = \boldsymbol{x}) \propto \pi(\theta) \cdot L(\theta|\boldsymbol{X} = \boldsymbol{x}) \propto e^{-\frac{nr}{2}(\theta - \bar{x})^2}$

For hypothesis (i):

- $P(\theta > 77|\boldsymbol{X} = \boldsymbol{x}) = 1 - \Phi(\sqrt{nr}(77 - \bar{x})) = 0.8556$
- $P(\theta \leq 77|\boldsymbol{X} = \boldsymbol{x}) = \Phi(\sqrt{nr}(77 - \bar{x})) = 0.1444$
- Hence we favour $H_0 : \theta > 77$

For hypothesis (ii):

- $P(\theta < 77|\boldsymbol{X} = \boldsymbol{x}) = \Phi(\sqrt{nr}(77 - \bar{x})) = 0.1444$
- $P(\theta > 80|\boldsymbol{X} = \boldsymbol{x}) = 1 - \Phi(\sqrt{nr}(80 - \bar{x})) = 0.1444$
- Hence we do not favour more that H_0 compared to H_1

7.4 Test between a Parameter Point and a Set: $\{\theta = \theta_1\}$ or $\{\theta \in \Theta_2\}$

Point null hypothesis, i.e. test whether parameter is a certain value or not: $H_0 : \theta = \theta_1$ vs $H_1 : \theta \neq \theta_1$

We must assume a *mixture prior*: because any continuous prior distribution assumes 0 probability at any point

- Mixture form: $p^{\delta_{\theta_1}} + (1 - p)Y$
  - δ_{θ_1} is random variable of $\{\theta = \theta_1\}$ with probability 1
  - Y is random variable defined on Θ_2 with density $\pi(\theta|\theta \in \Theta_2)$

$$\boxed{\begin{cases} \theta = \theta_1 & \text{with probability } \ P(\theta = \theta_1) = p \\ \theta \in \Theta_2 & \text{with probability } \ P(\theta \in \Theta_2) = (1 - p), \text{ where } (\theta|\theta \in \Theta_2) \text{ has proper prior density} \end{cases}}$$

Posterior probabilities

$$\boxed{\begin{aligned} &P(\theta = \theta_1|\boldsymbol{x}) \propto P(\theta = \theta_1) \cdot f(\boldsymbol{x}|\theta_1) \\ &P(\theta \in \Theta_2|\boldsymbol{x}) \propto P(\theta \in \Theta_2) \cdot \int_{\Theta_2} \pi(\theta|\theta \in \Theta_2) \cdot f(\boldsymbol{x}|\theta) \, d\theta \end{aligned}}$$

Normalisation constant: $p\pi(\boldsymbol{x}|\theta_1) + (1 - p) \int_{\Theta_2} \pi(\theta|\theta \in \Theta_2)\pi(\boldsymbol{x}|\theta) \, d\theta$

7.4.1 Example

Ahmad believes that probability of stock going up (vs down) in any given day is 0.5. Initially, we believe in Ahmad's claim with 99% certainty. After observing 62 up-days across 100 days, do we still believe in Ahmad's claim?

- Hypothesis: $H_0 : \theta = \frac{1}{2}$ vs $H_1 : \theta \neq \frac{1}{2}$
- Prior: $P(\theta = \frac{1}{2}) = 0.99$, $P(\theta \neq \frac{1}{2}) = 0.01$ where $\pi(\theta|\theta \neq \frac{1}{2}) \propto 1$ (flat prior)
- Model density: $X \sim Ber(\theta)$
  - $L(\theta|\boldsymbol{x}) \propto \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$
- Observations: $n = 100$, $n\bar{x} = 62$
- Posterior
  - $P(\theta = \frac{1}{2}|\boldsymbol{x}) \propto 0.99 \cdot \frac{1}{2}^{62}\frac{1}{2}^{38} = 7.8097 \times 10^{-31}$
  - $P(\theta \neq \frac{1}{2}|\boldsymbol{x}) \propto 0.01 \cdot \int_0^1 \theta^{62}(1-\theta)^{38}\, d\theta = 0.01 \cdot Beta(63, 39) = 1.71462 \times 10^{-32}$
  - $(O_n = 44.72)$
  - Hence H_0 is more favourable

7.4.2 Example 2

Normal observations

- Hypothesis: $H_0 : \theta = 78$ vs $H_1 : \theta \neq 78$
- Mixture prior:
  - $P(\theta = 78) = \frac{1}{2}$
  - $P(\theta \neq 78) = \frac{1}{2}$, where $\theta|\theta \neq 78 \sim N(m, \frac{1}{t})$ with $m = 79$, $t = \frac{1}{3}$
- Model density: $X \sim N(\theta, \frac{1}{r})$ where $r = 1$
  - $L(\theta|\boldsymbol{X} = \boldsymbol{x}) \propto e^{-\frac{nr}{2}(\theta-\bar{x})^2}$
- Observations: $n = 100$, $\bar{x} = 78.5$
- Posterior: $\pi(\theta|\boldsymbol{X} = \boldsymbol{x}) \propto \pi(\theta) \cdot L(\theta|\boldsymbol{X} = \boldsymbol{x}) \propto e^{-\frac{n}{2}(\theta-\bar{x})^2}$
  - Hence $\theta|\boldsymbol{X} = \boldsymbol{x} \sim N(\bar{x}, \frac{1}{n})$
  - $P(\theta = 78|\boldsymbol{x}) \propto \frac{1}{2}e^{-\frac{nr}{2}(\theta-\bar{x})^2} = 1.8633 \times 10^{-6}$
  - $P(\theta \neq 78|\boldsymbol{x}) \propto \frac{1}{2}\int_{-\infty}^{\infty} \pi(\theta|\theta \neq 78) \cdot L(\theta|\boldsymbol{x})\, d\theta = \frac{1}{2}\int_{-\infty}^{\infty} \sqrt{\frac{t}{2\pi}} e^{-\frac{t}{2}(\theta-m)^2} \cdot e^{-\frac{nr}{2}(\theta-\bar{x})^2}\, d\theta = \frac{1}{2}\sqrt{\frac{t}{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[(t+nr)(\theta-\bar{m})}$
  - $= \ldots = 0.02756$
  - See L6 slide 43 for more details
  - Hence we favour $H_1 : \theta \neq 78$

7.5 Hypothesis Tests with Nuisance Parameters

Often, our parameters of interest are restricted to a lower-dimensional subset Θ of the full parameter set $\tilde{\Theta}$

- E.g. Normal population: often only μ is of interest and not σ^2

To deal with them, construct a prior on full parameter $\tilde{\Theta}$ as usual, then apply the double expectation formula to integrate out nuisance/unwanted parameter

7.5.1 Illustration

Suppose X depends on θ and λ. How to choose between $\{\theta \in \Theta_1\}$ and $\{\theta \in \Theta_2\}$ without any knowledge about the value of λ?

Prior: let $P(\theta \in \Theta_1) = p$ and $P(\theta \in \Theta_2) = 1 - p$ Model density: get it by integrating over λ, i.e. $\pi(\boldsymbol{x}|\theta) = \int \pi(\boldsymbol{x}|\theta, \lambda) \cdot \pi(\lambda|\theta)\ d\lambda$ Posterior probabilities:

- $P(\theta \in \Theta_i|\boldsymbol{x}) \propto P(\theta \in \Theta_i) \cdot [\int L(\theta, \lambda|\boldsymbol{x}) \cdot \pi(\lambda|\theta \in \Theta_i)\ d\lambda]$

7.5.2 Example

Model: $X \sim N(\mu, \frac{1}{\tau})$ where μ and τ are both unknown Hypothesis: $H_0 : \mu = \mu_1$ vs $H_1 : \mu = \mu_2$ Prior: $\tau|\mu = \mu_1 \sim Gamma(\alpha_1, \frac{1}{\beta_1})$ and $\tau|\mu = \mu_2 \sim Gamma(\alpha_2, \frac{1}{\beta_2})$ Likelihood: $L(\mu, \tau|\boldsymbol{x}) \propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum_{r=1}^{n}(\mu - x_r)^2}$ Posterior:

- $P(\mu = \mu_i|\boldsymbol{x}) \propto P(\mu = \mu_i) \cdot \int L(\mu_i, \tau|\boldsymbol{x}) \cdot \pi(\tau|\mu = \mu_i)\ d\tau \propto p_i \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \int \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum_{r=1}^{n}(\mu_i - x_r)^2} \cdot \tau^{\alpha_i - 1} e^{-\beta_i \tau}\ d\tau$
- $= p_i \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \int \tau^{\alpha_i + \frac{n}{2} - 1} e^{-[\beta_i + \frac{1}{2}\sum_{r=1}^{n}(\mu_i - x_r)^2]\tau}\ d\tau = \frac{p_i \beta_i^{\alpha_i} \Gamma(\alpha_i + \frac{n}{2})}{\Gamma(\alpha_i)[\beta_i + \frac{1}{2}\sum_{r=1}^{n}(\mu_i - x_r)^2]^{\alpha_i + \frac{n}{2}}}$

8 Lecture 7: Bayesian Computation

Numerical approximation techniques: approximate any expectation/probability when answers are NOT readily available in known forms

8.1 Monte Carlo Integration

Monte Carlo Integration: take M random samples of target RVs, through direct simulation

- Target: suppose we want to calculate the integral $\gamma_g = E[g(\theta)] = \int_\Theta g(\theta)p(\theta)\ d\theta$
- Approximation: use the average $\widetilde{\gamma_g} = \frac{1}{M}\sum_{i=1}^M g(\theta_i)$, assuming $\theta_1, \ldots, \theta_M$ are iid
- Reason: by LLN, $\widetilde{\gamma_g}$ converges to $E[g(\theta)]$ as $M \to \infty$

Monte Carlo Integration in Bayesian Inference

- Suppose we have M samples of θ_i, each drawn directly from posterior $\pi(\theta|\boldsymbol{x})$
- ($\star$) $E[g(\theta)|\boldsymbol{x}] = \int_\Theta g(\theta)\pi(\theta|\boldsymbol{x})\ d\theta \approx \frac{1}{M}\sum_{i=1}^M g(\theta_i)$
  - $p(\theta) \Rightarrow \pi(\theta|\boldsymbol{x})$
  - $\gamma_g \Rightarrow E[g(\theta)|\boldsymbol{x}]$
- Example: Posterior mean: approximate with $\widetilde{\gamma_g} = \frac{1}{M}\sum_{i=1}^M \theta_i$
- Example: Posterior k-th moment: approximate with $\widetilde{\gamma_g} = \frac{1}{M}\sum_{i=1}^M \theta_i^k$

8.1.1 Quality of Monte Carlo Approximation

Quality of approximation improves as we increase M. Measure quality of approximation using standard error of $\widetilde{\gamma_g}$, i.e. the square root of its variance:

- Variance of $\widetilde{\gamma_g}$ is $Var[\widetilde{\gamma_g}] = Var[\frac{1}{M}\sum_{i=1}^M g(\theta_i)] = \frac{1}{M}Var[g(\theta)]$
- Substituting $\widetilde{\gamma_h}$ into the above, we get estimated standard error of $\widetilde{\gamma_g}$ to be $\sqrt{\frac{1}{M(M-1)}\sum_{i=1}^M [g(\theta_i) - \widetilde{\gamma_g}]^2}$

Working for γ_h:

- Target: let $\gamma_h = Var[g(\theta)] = E[h(\theta)] \equiv E\{g(\theta) - E[g(\theta)]\}^2$
- Approximation: $\widetilde{\gamma_h} = \frac{1}{M-1}\sum_{i=1}^M [g(\theta_i) - \widetilde{\gamma_g}]^2$

8.1.2 Example 1

Model: $X \sim N(\mu, \sigma^2)$ Given 5 independent samples -1.75, -1.17, -0.5, 0.33, 1.3, give estimates for:

1. μ, σ, $P(X > 0.3)$
2. mean of $Y \equiv X^2/(X-1)$
3. $P(X^2 > 0.3)$
4. variance of estimate for $P(X > 0.3)$

Solution

1. $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \ldots = -0.358$; $\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2 = \ldots = 1.2083$; $P(\widehat{X > 0.3}) = \frac{1}{n}\sum_{i=1}^n I(x_i > 0.3) = \frac{2}{5} = 0.4$
2. $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \frac{x_i^2}{x_i - 1} = \ldots = 0.7119$
3. $P(\widehat{X^2 > 0.3}) = \frac{1}{n}\sum_{i=1}^n I(x_i^2 > 0.3) = \frac{3}{5} = 0.6$

4. (See lecture 7 slide 17)

8.1.3 Example 2

Given 5 independent samples -1.75, -1.17, -0.5, 0.33, 1.3 from posterior of mean μ of population X, give estimates for:

1. Bayes estimate for population mean μ
2. Posterior variance of μ
3. Posterior probability that $\mu > 0.3$

Solution

1. $\hat{E}(\mu|x_1, \dots, x_n) = \bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i = \dots = -0.358$
2. $V\hat{a}r(\mu|x_1, \dots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2 = \dots = 1.2083$
3. $P(\widehat{\mu > 0.3}) = \frac{1}{n}\sum_{i=1}^{n} I(\mu_i > 0.3) = \frac{2}{5} = 0.4$

8.1.4 Example 3

Model: $X \sim N(\mu, \frac{1}{\tau})$ where μ and τ are unknown, $(\mu, \tau) \sim Gamma-Normal(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n})$ How to approximate the Bayes estimate for coefficient of variation CV $= \mu\sqrt{\tau}$? Given M independent pairs $(\mu_1, \tau_1), \dots, (\mu_M, \tau_M)$ — drawn first from $\tau_i \sim Gamma(\alpha_n, \frac{1}{\beta_n})$ then $\mu_i|\tau_i \sim N(m_n, \frac{1}{\tau_i t_n})$

Solution

- $E[\mu\sqrt{\tau}|\boldsymbol{x}] \approx \frac{1}{M}\sum_{i=1}^{M} \mu_i\sqrt{\tau_i}$

8.2 Importance Sampling

(More details: see section 20.2 SIR, p450)

Problem: what if we cannot sample directly from the posterior distribution? Especially when there are no closed-form expressions for posterior density.

Importance sampler: indirect sampling procedure from target density

- Target: $E[g(\theta)|\boldsymbol{x}] = \int_\Theta g(\theta) \cdot \pi(\theta|\boldsymbol{x})\ d\theta$
- Problem: we are unable to sample from posterior density $\pi(\theta|\boldsymbol{x})$
- Solution: sample from *importance density* $h(\theta)$ instead

Importance density and weight function

- Importance density: $h(\theta)$
- Importance weight function: $\omega(\theta) = \frac{\pi(\theta)\pi(\boldsymbol{x}|\theta)}{h(\theta)}$
- We can sample θ_i instead from $h(\theta)$ whereby $\omega(\theta)h(\theta) = \pi(\theta)\pi(\boldsymbol{x}|\theta)$
- Note: it's OK to have something proportional to $\omega(\theta)$ instead of the actual, it'll cancel out on numerator and denominator

Hence $E[g(\theta)|\boldsymbol{x}] = \int_\Theta g(\theta)\pi(\theta|\boldsymbol{x})\ d\theta = \frac{\int_\Theta g(\theta)\pi(\theta)\pi(\boldsymbol{x}|\theta)\ d\theta}{\int_\Theta \pi(\theta)\pi(\boldsymbol{x}|\theta)\ d\theta} = \frac{\int_\Theta g(\theta)\omega(\theta)h(\theta)\ d\theta}{\int_\Theta \omega(\theta)h(\theta)\ d\theta} \approx \frac{\sum_{i=1}^{M} g(\theta_i)\omega(\theta_i)}{\sum_{i=1}^{M}\omega(\theta_i)}$ where $\theta_i \sim h(\theta)$.

Prior density as importance density

- Let $h(\theta) = \pi(\theta)$
- Then importance weight function is $\pi(\boldsymbol{x}|\theta)$

- Hence $E[g(\theta)|\boldsymbol{x}] \approx \frac{\sum_{i=1}^{M} g(\theta_i)\pi(\boldsymbol{x}|\theta)}{\sum_{i=1}^{M} \pi(\boldsymbol{x}|\theta)}$

Likelihood-based density as importance density

- Let $h(\theta) \propto L(\theta|\boldsymbol{x})$
- Then importance weight function is $\pi(\theta)$
- Hence $E[g(\theta)|\boldsymbol{x}] \approx \frac{\sum_{i=1}^{M} g(\theta_i)\pi(\theta_i)}{\sum_{i=1}^{M} \pi(\theta_i)}$

Choosing a good importance density $h(\theta)$

- Accuracy of approximation depends on how good importance density $h(\theta)$ can approximate target density
- Choosing prior density $\pi(\theta)$ as importance density might not be good if it doesn't carry much information (esp. flat prior)
- Then in that case we might favour likelihood as importance density instead

8.2.1 Example

Suppose we have 5 importance samples from the *prior* of θ, $(-1.75, -1.17, -0.5, 0.33, 1.3)$. Using importance sampling, estimate the i) posterior mean θ, ii) probability that posterior has a positive mean.

- Prior: $\theta \sim N(m, t)$ where $m = 0$, $t = 1$
- Model: $X|\theta \sim N(\theta, \frac{1}{r})$ where $r = 1$
- Observations: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 16.34$ with $n = 10$
  - $L(\theta|\boldsymbol{x}) \propto e^{-\frac{nr}{2}(\theta-\bar{x})^2}$
- Importance sampling
  - $h(\theta) = \pi(\theta)$ — importance density is prior
  - $\omega(\theta) = L(\theta|\boldsymbol{x})$
  - Samples from $h(\theta)$: $(\theta_1, \ldots, \theta_M) = (-1.75, -1.17, -0.5, 0.33, 1.3)$ where $M = 5$

Solution: Importance Sampling Bayes Estimates

- For θ: $E[\theta|\boldsymbol{x}] \approx \frac{\sum_{i=1}^{M} \theta_i L(\theta_i|\boldsymbol{x})}{\sum_{i=1}^{M} L(\theta_i|\boldsymbol{x})} = NaN$ — because the likelihood values are too small!
- For $\theta > 0$: $E[I(\theta > 0)|\boldsymbol{x}] \approx \frac{\sum_{i=1}^{M} I(\theta_i > 0) L(\theta_i|\boldsymbol{x})}{\sum_{i=1}^{M} L(\theta_i|\boldsymbol{x})} = NaN$

Repeat the above, but now generate 1,000,000 samples from prior $N(0, 1)$.

Solution

- Estimate for θ: 4.7645 — bad compared to 14.85, which is the theoretical result for posterior mean!
- Estimate for $\theta > 0$: 1

```
n = 10
M = 1000000
vtheta = rnorm(M, 0, 1)                 ## Generate samples \theta_1 to \theta_M from N(0, 1) = PRIOR
w = exp(-0.5 * n * (vtheta-16.34)^2)    ## Importance weights = LIKELIHOOD
htheta = sum(vtheta * w) / sum(w)       ## Estimate for \theta
ptheta = sum((vtheta > 0) * w) / sum(w)  ## Estimate for \theta>0
```

8.2.2 Example 2

[Same setup as above, but now importance density $h(\theta) = N(16.34, \frac{1}{10})$.]

Suppose we have 5 importance samples from $N(16.34, \frac{1}{10})$, $(15.9, 16.02, 16.55, 16.6, 16.81)$. Using importance sampling, estimate the i) posterior mean θ, ii) probability that posterior has a positive mean.

Note that $N(16.34, \frac{1}{10})$ is proportional to the density of θ from the *likelihood function.*

- Observations: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 16.34$ with $n = 10$
- Likelihood function: $L(\theta|\boldsymbol{x}) \propto e^{-\frac{nr}{2}(\theta-\bar{x})^2}$
- Independent sample from **likelihood-based density**: $(\theta_1, \ldots, \theta_M) = (15.9, 16.02, 16.55, 16.6, 16.81)$ with $M = 5$

<u>Solution: Importance Sampling Bayes Estimates</u>

- For θ: $E[\theta|\boldsymbol{x}] \approx \frac{\sum_{i=1}^{M} \theta_i \pi(\theta_i)}{\sum_{i=1}^{M} \pi(\theta_i)} = 15.9154$
- For $\theta > 0$: $E[I(\theta > 0)|\boldsymbol{x}] \approx \frac{\sum_{i=1}^{M} I(\theta_i > 0)\pi(\theta_i)}{\sum_{i=1}^{M} \pi(\theta_i)} = 1$

Repeat the above, but now generate 1,000,000 samples from likelihood-based density $N(16.34, \frac{1}{10})$.

- Estimate for θ: 15.10
- Estimate for $\theta > 0$: 1

```
M = 1000000
vtheta = rnorm(M, 16.34, 1/sqrt(10))     ## Generate samples \theta_1 to \theta_M from N(16.34, 1/10)
w = exp(-0.5 * vtheta^2)                 ## Importance weights = PRIOR
htheta = sum(vtheta * w) / sum(w)        ## Estimate for \theta
ptheta = sum((vtheta > 0) * w) / sum(w)  ## Estimate for \theta>0
```

9 Lecture 8: Markov Chain Monte Carlo

Drawback of Monte Carlo integration: cannot use if you cannot draw exact samples from the posterior density (e.g. too difficult to draw exact samples, or the target density's normalisation constant is a complicated integral)

Drawback of importance sampling: can be hard to find a good importance density, that is both easy to sample and close to target density

MCMC: allows you to draw sample from exact posterior, even though only the proportional form is known

- Idea: sample a sequence of random samples, that are *correlated* with each other
- Sequence of samples constitutes a *stationary Markov chain* with a unique *stationary distribution* that coincides with the target distribution

9.1 MCMC Approximations with 2 Variables

Suppose we know joint density $\pi(x, y)$ up to a proportional constant, i.e. $\pi(x, y) = \frac{h(x,y)}{C}$

- $h(x, y)$ is the *kernel* of the joint density
- C is the normalisation constant

Sampled Markov chain: $(x_0, y_0), (x_1, y_2), \ldots, (x_M, y_M)$

- Stationary distribution of Markov chain must be *exactly identical* to desired $\pi(x, y)$

Therefore, we need to construct such a chain with the desired stationary distribution, i.e. define moves from state (x_i, y_i) to next state (x_{i+1}, y_{i+1})

9.2 Gibbs Sampler

Given present state (x_i, y_i):

- Select X and Y consecutively to form next state (x_i, y_i) using these densities:
- $(x_{i+1}|x_i, y_i)$ — has density $\pi(x_{i+1}|y_i) \propto h(x_{i+1}, y_i)$
- $(y_{i+1}|x_{i+1}, y_i)$ — has density $\pi(y_{i+1}|x_{i+1}) \propto h(x_{i+1}, y_{i+1})$
- So the transition function $k(x_{i+1}, y_{i+1}|x_i, y_i) = \pi(x_{i+1}|y_i) \cdot \pi(y_{i+1}|x_{i+1})$

Gibbs sampler

1. Start with arbitrary valid initial state (x_0, y_0)
2. For $w + M$ times, move from state (x_i, y_i) to next state (x_{i+1}, y_{i+1}) via transition function $k(x_{i+1}, y_{i+1}|x_i, y_i)$
3. Discard the first w members of the chain, keeping M members $(x_{w+1}, y_{w+1}), \ldots, (x_{w+M}, y_{w+M})$ to calculate estimates
4. w: "burn-in" period. If w is large, we can assume that (x_w, y_w) is sampled from stationary distribution $\pi(x, y)$
5. M: required Monte Carlo size

$$E[g(X, Y)] \approx \frac{1}{M} \sum_{i=1}^{M} g(x_{w+i}, y_{w+i})$$

We can treat the correlated variates (x_i, y_i) as if they were i.i.d. from $\pi(x, y)$.

9.2.1 Example: Multinomial

Setup: $(X_1, X_2, X_3) \sim Multinomial(n; p_1, p_2, p_3)$

- $\pi(x_1, x_2, x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$, noting that $p_3 = 1 - p_1 - p_2$.
- \$π(x₁|x₂) ∝ π(x₁,
- $x_1|x_2 \sim Bin(n - x_2, \frac{p_1}{1-p_2})$
- $x_2|x_1 \sim Bin(n - x_1, \frac{p_2}{1-p_1})$
- $X_1|X_2 = x_2 \sim Bin(n - x_2, \frac{p_1}{1-p_2})$
- $X_2|X_1 = x_1 \sim Bin(n - x_1, \frac{p_2}{1-p_1})$

Problem: suppose we want to find $P(X_1 > X_2)$.

Solution: using Gibbs sampling, draw samples from the conditional densities $\pi(x_1|x_2)$ and $\pi(x_2|x_1)$

- Initial state: choose some valid $(x_{1,0}, x_{2,0})$
- Step 1: sample $(x_{2,i+1}|x_{1,i}) \sim Bin(n - x_{1,i}, \frac{p_2}{1-p_1})$
- Step 2: sample $(x_{1,i+1}|x_{2,i+1}) \sim Bin(n - x_{2,i+1}, \frac{p_1}{1-p_2})$
- Repeat until we get Markov chain
- Hence $P(X_1 > X_2) \approx \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{I}_{x_{1,w+i}, x_{2,w+i}}$

9.2.2 Example: Dirichlet

Problem: suppose we want to generate (U, V, W) with Dirichlet distribution

- $f(u, v, w) = ku^4 v^3 w^2 (1 - u - v - w)$
- $\pi(u|v, w) \propto u^4 (1 - u - v - w) \propto (\frac{u}{1-v-w})^4 (1 - \frac{u}{1-v-w})(\frac{1}{1-v-w})$
  - Note that $1 - v - w$ is a constant (v and w are given)
- $\pi(u|v, w) = \frac{1}{B(5,2)} q^{5-1} (1-q)^{2-1} \frac{1}{1-v-w}$ where $q = \frac{u}{1-v-w}$
- Then $f(q|V = v, W = w) = (1 - v - w) f_{u|v,w}((1 - v - w)q) = \frac{1}{Beta(5,2)} q^{5-1} (1-q)^{2-1}$
- Let $Q = \frac{U}{1-V-W}$. Then $Q|V, W \sim Beta(5, 2)$, so $U|V, W = (1 - V - W)Q$
- Let $R = \frac{V}{1-U-W}$. Then $R|U, W \sim Beta(4, 2)$, so $V|U, W = (1 - U - W)R$
- Let $S = \frac{W}{1-U-V}$. Then $S|U, V \sim Beta(3, 2)$, so $W|U, V = (1 - U - V)S$

Solution: using Gibbs sampling, do the follows:

- Initial state: choose some valid (u_0, v_0, w_0), obeying $u_0, v_0, w_0 > 0$ and $u_0 + v_0 + w_0 < 1$
- Step 1: sample $q_{i+1} \sim Beta(5, 2)$, set $u_{i+1} = (1 - v_i - w_i) q_{i+1}$
- Step 2: sample $r_{i+1} \sim Beta(4, 2)$, set $v_{i+1} = (1 - u_{i+1} - w_i) r_{i+1}$
- Step 3: sample $s_{i+1} \sim Beta(3, 2)$, set $w_{i+1} = (1 - v_{i+1} - w_{i+1}) s_{i+1}$
- Repeat until we get Markov chain
- As $n \to \infty$, the distribution of (U_n, V_n, W_n) converges to the desired Dirichlet distribution

9.2.3 Example: Coefficient of Variation

Problem: we want to estimate the posterior CV, i.e. $E[\mu\sqrt{\tau}|\boldsymbol{x}] = \int\int \mu\sqrt{\tau} \cdot \pi(\mu,\tau|\boldsymbol{x})\ d\tau\ d\mu$

- Here, $\pi(\mu,\tau|\boldsymbol{x}) \sim Gamma - Normal(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n})$
- We need to find both conditional densities: $\mu|\tau,\boldsymbol{x}$ and $\tau|\mu,\boldsymbol{x}$
- $\mu|\tau,\boldsymbol{x} \sim N(m_n, \frac{1}{\tau t_n})$
- $\tau|\mu,\boldsymbol{x} \sim Gamma(\alpha_n + \frac{1}{2}, \frac{1}{\beta_n + \frac{t_n}{2}(\mu - m_n)^2})$
  - $\pi(\tau|\mu,\boldsymbol{x}) \propto \pi(\mu|\tau,\boldsymbol{x}) \cdot \pi(\tau|\boldsymbol{x}) \propto \tau^{1/2} e^{-\frac{\tau t_n}{2}(\mu-m_n)^2} \tau^{\alpha_n - 1} e^{-\beta_n \tau} \propto \tau^{(\alpha_n+1/2)-1} e^{-(\beta_n + \frac{t_n}{2}(\mu-m_n)^2)\tau}$

Solution: using Gibbs sampling, do the follows:

- Initial state: choose some valid (μ_0, τ_0)
- Step 1: sample $(\tau_{i+1}|\mu_i,\boldsymbol{x}) \sim Gamma(\alpha_n + \frac{1}{2}, \frac{1}{\beta_n + \frac{t_n}{2}(\mu_i - m_n)^2})$
- Step 2: sample $(\mu_{i+1}|\tau_{i+1},\boldsymbol{x}) \sim N(m_n, \frac{1}{\tau_{i+1} t_n})$
- Repeat until we get Markov chain
- Hence $E[\mu\sqrt{\tau}|\boldsymbol{x}] \approx \frac{1}{M}\sum_{i=1}^{M} \mu_{w+i}\sqrt{\tau_{w+i}}$

9.2.4 Example: Pump Failure

Suppose we have 10 pumps. Each pump i has been running for time t_i, and failed N_i times during that time.

i	N_i	t_i	N_i/t_i
1	5	94.320	0.0530
2	1	15.720	0.0636
3	5	62.880	0.0795
4	14	125.760	0.111
5	3	5.240	0.573
6	19	31.440	0.604
7	1	1.048	0.954
8	1	1.048	0.954
9	4	2.096	1.91
10	22	10.480	2.10

Assume that failure of pump i occurs according to Poisson process with rate λ_i, i.e. $N_i \sim Po(\lambda_i t_i)$

- $f(\boldsymbol{N}|\boldsymbol{\lambda}) = \prod_{i=1}^{10} \frac{(\lambda_i t_i)^{N_i} e^{-\lambda_i t_i}}{N_i!}$, where $\boldsymbol{N} = (N_1, \ldots, N_{10})$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{10})$

Assume that $\lambda_i \sim Gamma(\alpha, \frac{1}{\beta})$ where $\alpha = 1.8$

- $\pi(\lambda_i) \propto \beta^{\alpha} \lambda_i^{\alpha-1} e^{-\beta\lambda_i}$

Assume that $\beta \sim Gamma(\gamma, \frac{1}{\delta})$ where $\gamma = 0.01$ and $\delta = 1$

- $\pi(\beta) \propto \beta^{\gamma-1} e^{-\delta\beta}$
- Note: β tends to be small because $E[\beta] = 0.01$, which makes the λ_i distribution very flat — this is good, because it should not contain much a priori information, and in principle affects the results less

Posterior distribution:

- $\pi(\boldsymbol{\lambda},\beta|\boldsymbol{N}) \propto f(\boldsymbol{N}|\boldsymbol{\lambda}) \cdot \pi(\boldsymbol{\lambda}|\beta) \cdot \pi(\beta) = \ldots$
- $E[\lambda_1|\boldsymbol{N}] \approx \frac{1}{M}\sum_{i=1}^{M} \frac{N_1+\alpha}{t_1+\beta_{w+i}}$