Francesco Caporali, Isabel Muzio

24 October 2022

Bayesian Statistics

Assignment 1

Question 1: The Galenshore distribution

Point a.

$Y|\theta \sim \text{Galenshore}\,(a, \theta)$ is such that $p(y|\theta)$ is a density in the exponential family indeed

$$p(y|\theta) = \frac{2}{\Gamma(a)}\theta^{2a}y^{2a-1}e^{-\theta^2y^2}\mathbb{1}_{\{y>0\}}, \theta > 0, a > 0.$$

Then defininining $\phi \overset{\text{def}}{=} \theta^2$ one has

$$p(y|\phi) = h(y)c(\phi)e^{\phi t(y)} \text{ with } h(y) = \frac{2y^{2a-1}}{\Gamma(a)}\mathbb{1}_{\{y>0\}}, c(\phi) = \phi^a, t(y) = -y^2,$$

hence, by the easy shape of a distribution in the exponential family, we can state that a class of conjugate priors for $p(y|\phi)$ is such that

$$p(\phi) \propto c(\phi)^{n_0}e^{\phi n_0 t_0} = \phi^{an_0}e^{\phi n_0 t_0}.$$

If ϕ has density $p(\phi)$ and we want to obtain the density of $\theta = \sqrt{\phi}$ it is sufficient to define the map $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that $f(x) = \sqrt{x}$ and recall that

$$p_\theta(\theta) = p_\phi(f(\phi)) = p_\phi(f^{-1}(\theta)\left|\frac{df^{-1}(\theta)}{d\theta}\right|.$$

Observing that $\left|\frac{df^{-1}(\theta)}{d\theta}\right| = \left|\frac{d\theta^2}{d\theta}\right| = 2\theta$

$$p(\theta) \overset{\text{def}}{=} p_\theta(\theta) = p_\phi(\theta^2)2\theta \propto \theta^{2an_0}e^{\theta^2 n_0 t_0}2\theta.$$

Remark

Observing the three parameters a, n_0 and t_0 we can say

- $a > 0$ by hypotesis;
- $n_0 > 0$ because it represents the *prior sample size* ($p(\theta)$ has the same kernel of $p(y|\theta)$ after n_0 observations);
- $t_0 < 0$ because it is the *prior guess* that we make for t, with $t(y) = -\frac{y^2}{2}, \forall y \in \mathbb{R}^+ : t_0 = \frac{\sum_{i=1}^n t(y_i)}{n} = -\frac{\sum_{i=1}^n y_i^2}{n} < 0$.

Hence we have

$$an_0 + 1 > 0 \text{ and } -n_0t_0 > 0.$$

So we can rewrite

$$p(\theta) \propto 2\theta^{2an_0+1}e^{-\left(\sqrt{-n_0t_0}\right)^2\theta^2}$$

and recognizing the kernel of a Galenshore distribution we can write explicitly

$$p(\theta) = 2\theta^{2(an_0+1)-1} e^{-\left(\sqrt{-n_0t_0}\right)^2\theta^2} \cdot \underbrace{\frac{\left(\sqrt{-n_0t_0}\right)^{2(an_0+1)}}{\Gamma(an_0+1)}}_{\text{it does not depends on }\theta} \underbrace{\mathbb{1}_{\theta>0}}_{\text{by hypotesis}} =$$

$$= \frac{2}{\Gamma(an_0+1)} \left(\sqrt{-n_0t_0}\right)^{2(an_0+1)} \theta^{2(an_0+1)-1} e^{-\left(\sqrt{-n_0t_0}\right)^2\theta^2} \mathbb{1}_{\theta>0}$$

$$\Downarrow$$

$$\theta \sim \text{Galenshore}\left(an_0+1, \sqrt{-n_0t_0}\right).$$

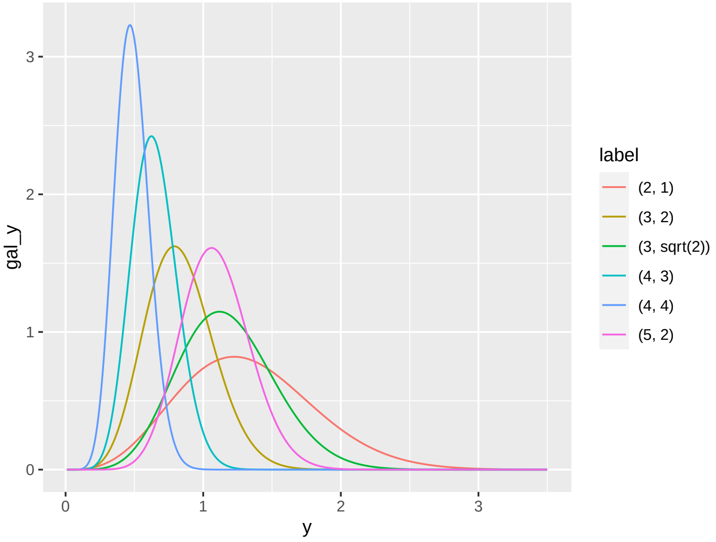
Finally we plot a few of these densities Galenshore $\left(an_0+1, \sqrt{-n_0t_0}\right)$ sampled with the following code:

- $n_0 = 1, t_0 = -1, a = 1 \implies \text{Galenshore}\,(2, 1);$
- $n_0 = 2, t_0 = -1, a = 1 \implies \text{Galenshore}\left(3, \sqrt{2}\right);$
- $n_0 = 2, t_0 = -2, a = 1 \implies \text{Galenshore}\,(3, 2);$
- $n_0 = 2, t_0 = -2, a = 2 \implies \text{Galenshore}\,(5, 2);$
- $n_0 = 3, t_0 = -3, a = 1 \implies \text{Galenshore}\,(4, 3);$
- $n_0 = 3, t_0 = -4, a = 1 \implies \text{Galenshore}\,(4, 4).$

```r
dgalenshore = function(y, a, theta) {
    (2 / gamma(a)) * theta^(2 * a) * y^(2 * a - 1) * exp(-(theta^2) * y^2)
}

y = seq(0.01, 3.5, length = 1000)
df = rbind(
    data.frame(y = y, gal_y = dgalenshore(y, 2, 1), label = "(2, 1)"),
    data.frame(y = y, gal_y = dgalenshore(y, 3, sqrt(2)), label = "(3, sqrt(2))"),
    data.frame(y = y, gal_y = dgalenshore(y, 3, 2), label = "(3, 2)"),
    data.frame(y = y, gal_y = dgalenshore(y, 5, 2), label = "(5, 2)"),
    data.frame(y = y, gal_y = dgalenshore(y, 4, 3), label = "(4, 3)"),
    data.frame(y = y, gal_y = dgalenshore(y, 4, 4), label = "(4, 4)")
)
```

Then we plot all of them at the same time:

```r
ggplot(df, aes(y, gal_y, group = label, color = label)) +
geom_line() + coord_fixed(ratio = 1)
```



Point b.

Let's define $b \overset{\text{def}}{=} an_0 + 1$ and $c = \sqrt{-n_0 t_0}$ for coinciseness.
Recalling $\theta \sim \text{Galenshore}\left(b, c\right)$ and $Y_i|\theta \sim \text{Galenshore}\left(a, \theta\right), \forall i \in 1 : n$ and defining $\text{SS}\left(y_{1:n}\right) \overset{\text{def}}{=} \sum_{i=1}^{n} y_i^2$ we have

$$\begin{aligned} p(\theta|y_{1:n}) = p(\theta)p(y_{1:n}|\theta) \propto \\ &\propto (\theta^{2b-1}e^{-c^2\theta^2})(\theta^{2na}e^{-\theta^2 \sum_{i=1}^{n} y_i^2}) \propto \\ &\propto \theta^{2(an+b)-1}e^{-(c^2+\text{SS}(y_{1:n}))\theta^2}. \end{aligned}$$

Hence we recognize the kernel of a Galenshore $\left(an + b, \sqrt{c^2 + \text{SS}\left(y_{1:n}\right)}\right)$

$$\implies \theta|Y_{1:n} \sim \text{Galenshore}\left(a(n + n_0) + 1, \sqrt{\text{SS}\left(y_{1:n}\right) - n_0 t_0}\right).$$

Point c.

$$\begin{aligned} \frac{p(\theta_a|y_{1:n})}{p(\theta_b|y_{1:n})} &= \frac{2\Gamma(a(n + n_0) + 1)}{\Gamma(a(n + n_0) + 1)2}\left(\text{SS}\left(y_{1:n}\right) - n_0 t_0\right)^{(a(n+n_0)+1)(1-1)}\left(\frac{\theta_a}{\theta_b}\right)^{2a(n+n_0)+1} e^{-(\text{SS}(y_{1:n})-n_0t_0)\left(\theta_a^2-\theta_b^2\right)} = \\ &= \left(\frac{\theta_a}{\theta_b}\right)^{2a(n+n_0)+1} e^{-\left(\sum_{i=1}^{n} y_i^2-n_0t_0\right)\left(\theta_a^2-\theta_b^2\right)}. \end{aligned}$$

Hence

$$\mathbb{P}\left(\theta \in A|Y_{1:n} = y_{1:n}\right) = \mathbb{P}\left(\theta \in A|\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} y_i^2\right), \forall A$$

and then, by definition $\text{SS}\left(Y_{1:n}\right) \overset{\text{def}}{=} \sum_{i=1}^{n} Y_i^2$ is a sufficient statistic.

Point d.

Recalling that $\theta|Y_{1:n} \sim \text{Galenshore}\left(a(n + n_0) + 1, \sqrt{\text{SS}\left(y_{1:n}\right) - n_0 t_0}\right)$ and that if $X \sim \text{Galenshore}\left(a, \theta\right) \implies \mathbb{E}\left[X\right] = \frac{\Gamma\left(a+\frac{1}{2}\right)}{\theta\Gamma(a)}$ we have

$$\mathbb{E}\left[\theta|y_{1:n}\right] = \frac{\Gamma\left(a(n + n_0) + frac32\right)}{\left(\sqrt{\text{SS}\left(y_{1:n}\right) - n_0 t_0}\right)\Gamma(a(n + n_0) + 1)}$$

Point e.

With the usual notation $b \overset{\text{def}}{=} an_0 + 1$ and $c \overset{\text{def}}{=} \sqrt{-n_0 t_0}$:

$$\begin{aligned} p(y_{n+1}|y_{1:n}) &= \int_0^\infty p(y_{n+1}|\theta)p(\theta|y_{1:n})d\theta = \\ &= \int_0^\infty \frac{2}{\Gamma(a)}\theta^{2a}y_{n+1}^{2a-1}e^{-\theta^2 y_{n+1}^2} \cdot \frac{2}{\Gamma(an + b)}\left(c^2 + \text{SS}\left(y_{1:n}\right)\right)^{an+b}\theta^{2(an+b)-1}e^{-(c^2+\text{SS}(y_{1:n}))\theta^2}d\theta = \\ &= \frac{4}{\Gamma(a)\Gamma(an + b)}y_{n+1}^{2a-1}\left(c^2 + \text{SS}\left(y_{1:n}\right)\right)^{an+b}\int_0^\infty \underbrace{\theta^{2(a+an+b)-1}e^{-(c^2+\text{SS}(y_{1:n})+y_{n+1}^2)\theta^2}}_{\text{kernel of a Galenshore}\left(a+an+b, \sqrt{c^2+\text{SS}(y_{1:n})+y_{n+1}^2}\right)} d\theta \end{aligned}$$

$$\Downarrow$$

$$\int_0^\infty \theta^{2(a+an+b)-1} e^{-(c^2+\mathrm{SS}(y_{1:n})+y_{n+1}^2)\theta^2} = \frac{\Gamma(a+an+b)}{2}\left(\frac{1}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2}\right)^{a+an+b}$$

$$\Downarrow$$

$$\begin{aligned}
p(y_{n+1}|y_{1:n}) &= \frac{4}{\Gamma(a)\Gamma(an+b)} y_{n+1}^{2a-1}\left(c^2+\mathrm{SS}\left(y_{1:n}\right)\right)^{an+b}\frac{\Gamma(a+an+b)}{2}\left(\frac{1}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2}\right)^{a+an+b} = \\
&= \frac{2}{y_{n+1}}\frac{\Gamma(a+an+b)}{\Gamma(a)\Gamma(an+b)}\left(\frac{y_{n+1}^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2}\right)^{a}\left(\frac{c^2+\mathrm{SS}\left(y_{1:n}\right)}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2}\right)^{an+b} = \\
&= \frac{2y_{n+1}\left(c^2+\mathrm{SS}\left(y_{1:n}\right)\right)}{\left(c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2\right)^2}\cdot \\
&\cdot \underbrace{\frac{1}{B(a,an+b)}\left(\frac{y_{n+1}^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2}\right)^{a-1}\left(1-\frac{y_{n+1}^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2}\right)^{(an+b)-1}}_{\text{density of } X\sim B(a,an+b) \text{ evaluated on } \frac{y_{n+1}^2}{c^2+\mathrm{SS}(y_{1:n})+y_{n+1}^2}} = \\
&= \frac{2y_{n+1}\left(c^2+\mathrm{SS}\left(y_{1:n}\right)\right)}{\left(c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2\right)^2} p_X\left(\frac{y_{n+1}^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2}\right).
\end{aligned}$$

Now one should note that this is a differentiable transformation of $\mathbb{R}^+$ of an unknown random variable. Indeed if one try to derive, with respect to y_{n+1}, the variable of the density of X in our last expression obtains

$$\begin{aligned}
\frac{d}{dy_{n+1}}\frac{y_{n+1}^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2} &= \frac{2y_{n+1}(c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2)-y_{n+1}^2 2y_{n+1}}{\left(c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2\right)^2} = \\
&= \underbrace{\frac{2y_{n+1}(c^2+\mathrm{SS}\left(y_{1:n}\right))}{\left(c^2+\mathrm{SS}\left(y_{1:n}\right)+y_{n+1}^2\right)^2}}_{>0 \text{ indeed } y_{n+1}\in\mathbb{R}^+ \text{ and the other terms are squared}}.
\end{aligned}$$

Hence if we define $f^{-1}:\mathbb{R}^+\to\mathbb{R}^+$ such that $f^{-1}(x)=\dfrac{x^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)+x^2}$ we can state

$$\begin{aligned}
p(y_{n+1}|y_{1:n}) &= \left|\frac{d}{dy_{n+1}}f^{-1}(y_{n+1})\right| p_X(f^{-1}(y_{n+1}) = \\
&= p_{f(X)}(y_{n+1}).
\end{aligned}$$

Let's compute f explicitly

$$f^{-1}(x)=\frac{x^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)+x^2} = 1-\frac{c^2+\mathrm{SS}\left(y_{1:n}\right)}{c^2+\mathrm{SS}\left(y_{1:n}\right)+x^2}$$

hence

$$\begin{aligned}
x = 1-\frac{c^2+\mathrm{SS}\left(y_{1:n}\right)}{c^2+\mathrm{SS}\left(y_{1:n}\right)+f(x)^2} &\iff 1-x=\frac{c^2+\mathrm{SS}\left(y_{1:n}\right)}{c^2+\mathrm{SS}\left(y_{1:n}\right)+f(x)^2} \iff \\
&\iff \frac{f(x)^2}{c^2+\mathrm{SS}\left(y_{1:n}\right)}+1=\frac{1}{1-x} \iff \\
&\iff f(x)=\sqrt{\frac{x}{1-x}}\sqrt{c^2+\mathrm{SS}\left(y_{1:n}\right)}.
\end{aligned}$$

This leads us to conclude (substituting again $b = an_0 + 1$ and $c = \sqrt{-n_0 t_0}$) that

$$Y_{n+1}|Y_{1:n} \sim f\left(B(a, a(n+n_0)+1)\right), \text{ with } f : \mathbb{R}^+ \to \mathbb{R}^+, f(x) = \sqrt{\frac{x}{1-x}}\sqrt{\mathrm{SS}\left(y_{1:n}\right) - n_0 t_0}.$$

Question 2: Tumor Counts

Part 1: Tumor Counts

A cancer laboratory is estimating the rate of tumorigenesis in two strains of mice, A and B. They have tumor count data for 10 mice in strain A and 13 mice in strain B. Type A mice have been well studied, and information from other laboratories suggests that type A mice have tumor counts that are approximately Poisson-distributed with a mean of 12. Tumor count rates for type B mice are unknown, but type B mice are related to type A mice.

Point a.

Given the Poisson-Gamma models defined in the problem, the posterior distributions are:

$$\left.\begin{array}{r} Y_A \mid \theta_A \sim \mathrm{Pois}(\theta_A) \\ \theta_A \sim \mathrm{Gamma}(120, 10) \end{array}\right\} \theta_A \mid Y_A = \mathbf{y}_A \sim \mathrm{Gamma}(120 + \sum_{i=1}^{n_A} y_{A,i}\ , 10 + n_A)$$

$$\left.\begin{array}{r} Y_B \mid \theta_B \sim \mathrm{Pois}(\theta_B) \\ \theta_B \sim \mathrm{Gamma}(12, 1) \end{array}\right\} \theta_B \mid Y_B = \mathbf{y}_B \sim \mathrm{Gamma}(12 + \sum_{i=1}^{n_B} y_{B,i}\ , 1 + n_B)$$

where the observed values of the samples are:

$$\begin{aligned} \mathbf{y}_A &= (12, 9, 12, 14, 13, 13, 15, 8, 15, 6), \quad n_A = \mathrm{length}(\mathbf{y}_A) \\ \mathbf{y}_B &= (11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7), \quad n_B = \mathrm{length}(\mathbf{y}_B) \end{aligned}$$

The consequent posterior distributions are the following:

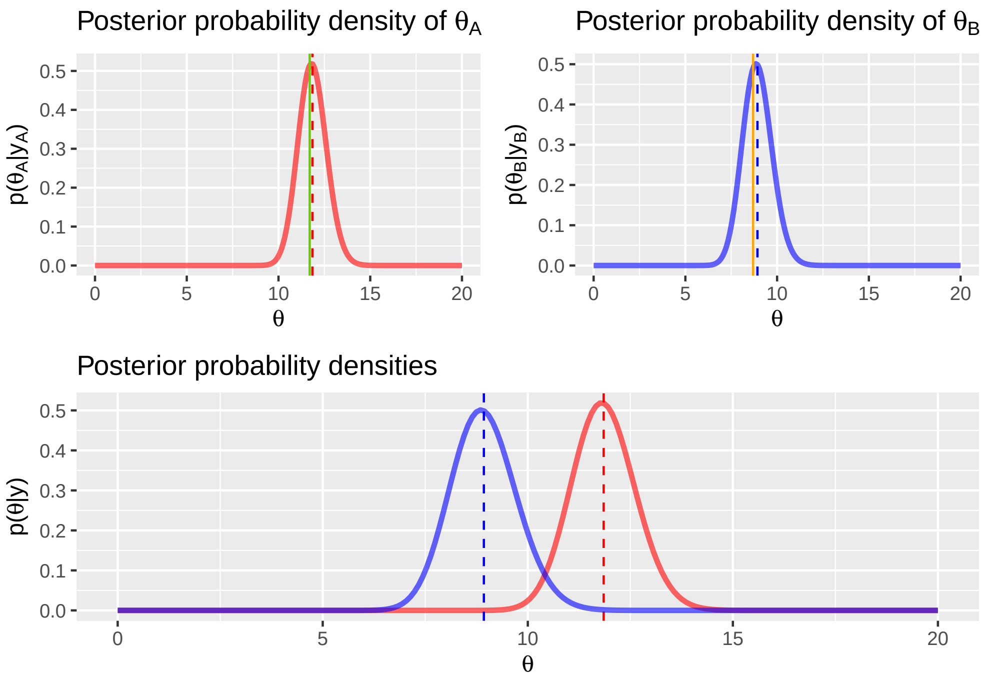
```r
load(file = "data_assignment_1.RData")
library(tidyverse)
library(gridExtra)
library(grid)
library(ggplot2)
library(lattice)

#posterior parameters

a_n = 120.0 + sum(y.a)
b_n = 10.0 + length(y.a)

c_n = 12 + sum(y.b)
d_n = 1 + length(y.b)

cat(paste(" Posterior of A : Gamma(", a_n, ",", b_n, ")\n", "Posterior of B : Gamma(", c_n, ",", d_n, ")\n"))
```

```
##  Posterior of A : Gamma( 237 , 20 )
##  Posterior of B : Gamma( 125 , 14 )
```

```r
#posterior means

mean_a = a_n / b_n
mean_b = c_n / d_n

#posterior densities

y.a.sum = sum(y.a)
n.a = length(y.a)
```

```
alpha = 0.05
gamma.values = seq(0, 20, length = 200)

post.values.a = dgamma(gamma.values, a_n, b_n)
post.values.b = dgamma(gamma.values, c_n, d_n)

post.data = data.frame(gamma.values, post.values.a, post.values.b)
```

```
#posterior density plots

post.data %>% ggplot() +
  geom_line(aes(x = gamma.values, y = post.values.a), col = "red", alpha = 0.6, size = 1.2) +
  geom_vline(xintercept = mean_a, col = "red", linetype = 2) +
  scale_color_discrete(guide = "none") -> p1

post.data %>% ggplot() +
  geom_line(aes(x = gamma.values, y = post.values.b), col = "blue", alpha = 0.6, size = 1.2) +
  geom_vline(xintercept = mean_b, col = "blue", linetype = 2) +
  scale_color_discrete(guide = "none") -> p2

p3 <- p1 +
  geom_line(aes(x = gamma.values, y = post.values.b), col = "blue", alpha = 0.6, size = 1.2) +
  geom_vline(xintercept = mean_b, col = "blue", linetype = 2) +
  xlab(expression(theta)) +
  ylab(expression(paste("p(", theta, "|", y, ")"))) +
  ggtitle("Posterior probability densities")

p1 <- p1 +
  geom_vline(xintercept = mean(y.a), col = "chartreuse3", linetype = 1) +
  xlab(expression(theta)) +
  ylab(expression(paste("p(", theta[A], "|", y[A], ")"))) +
  ggtitle(expression(paste("Posterior probability density of ", theta[A])))

p2 <- p2 +
  geom_vline(xintercept = mean(y.b), col = "orange", linetype = 1) +
  xlab(expression(theta)) +
  ylab(expression(paste("p(", theta[B], "|", y[B], ")"))) +
  ggtitle(expression(paste("Posterior probability density of ", theta[B])))

grid.arrange(arrangeGrob(p1, p2, ncol = 2), p3, nrow = 2)
```


Posterior probability density of θ_A

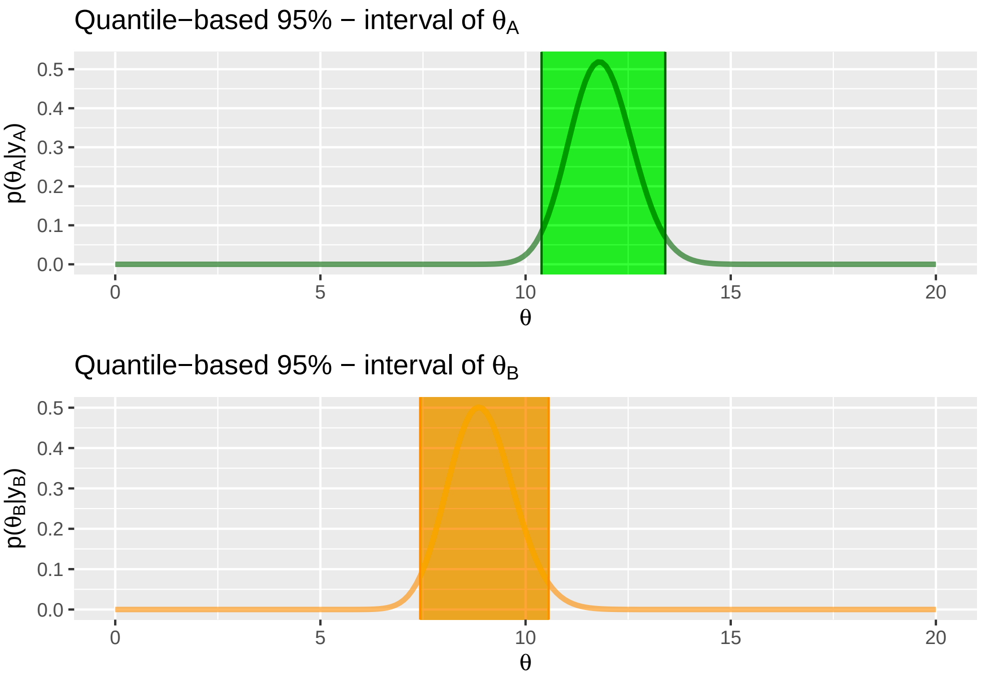
Posterior probability density of θ_B

Posterior probability densities

Using the posterior distributions, the 95% quantile-based intervals are:

```r
#quantile intervals

quant.interval.a = qgamma(c(alpha / 2, 1 - alpha / 2), shape = a_n, rate = b_n)
quant.interval.b = qgamma(c(alpha / 2, 1 - alpha / 2), shape = c_n, rate = d_n)
```

```r
#quantile-based intervals plots

post.data %>% ggplot() +
  geom_line(aes(x = gamma.values, y = post.values.a), col = "darkgreen", alpha = 0.6, size = 1.2) +
  geom_vline(xintercept = quant.interval.a[1], col = "darkgreen") +
  geom_vline(xintercept = quant.interval.a[2], col = "darkgreen") +
  geom_rect(aes(xmin = quant.interval.a[1], xmax = quant.interval.a[2], ymin = -Inf, ymax = Inf), fill = "green", alpha = 0.002) +
  xlab(expression(theta)) +
  ylab(expression(paste("p(", theta[A], "|", y[A], ")"))) +
  ggtitle(expression(paste("Quantile-based 95% - interval of ", theta[A]))) +
  scale_color_discrete(guide = "none") -> q1

post.data %>% ggplot() +
  geom_line(aes(x = gamma.values, y = post.values.b), col = "darkorange", alpha = 0.6, size = 1.2) +
  geom_vline(xintercept = quant.interval.b[1], col = "darkorange") +
  geom_vline(xintercept = quant.interval.b[2], col = "darkorange") +
  geom_rect(aes(xmin = quant.interval.b[1], xmax = quant.interval.b[2], ymin = -Inf, ymax = Inf), fill = "orange", alpha = 0.002) +
  xlab(expression(theta)) +
  ylab(expression(paste("p(", theta[B], "|", y[B], ")"))) +
  ggtitle(expression(paste("Quantile-based 95% - interval of ", theta[B]))) +
  scale_color_discrete(guide = "none") -> q2


grid.arrange(q1, q2, nrow = 2)
```



Point b.

Suppose we now consider a new model, where the prior for θ_B is dependent on a parameter n_0 in the following way:

$$\theta_B \sim \text{Gamma}(12 \times n_0, n_0)$$

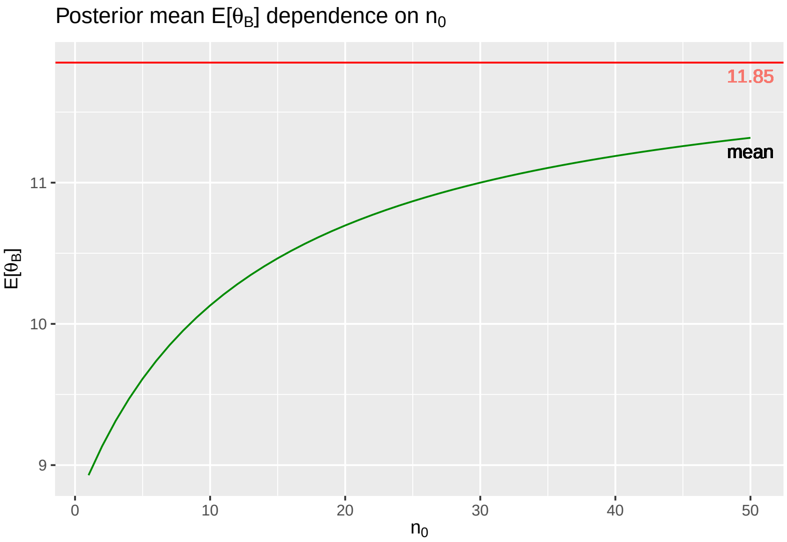
Therefore the new posterior is the following:

$$\left.\begin{array}{l} Y_B \mid \theta_B \sim \text{Pois}(\theta_B) \\ \theta_B \sim \text{Gamma}(12 \times n_0, n_0) \end{array}\right\} \theta_B \mid Y_B = \mathbf{y}_B \sim \text{Gamma}(12 \times n_0 + \sum_{i=1}^{n_B} y_{B,i} \ , n_0 + n_B)$$

The posterior mean therefore depends on the parameter n_0 with the following evolution:

```
mean.b = rep(0, 50)

n_0 = 1:50

for (i in n_0){
  a = 12 * i + sum(y.b)
  b = i + length(y.b)
  mean.b[i] = a / b
}

mean_varying <- data.frame(n_0, mean.b)
```

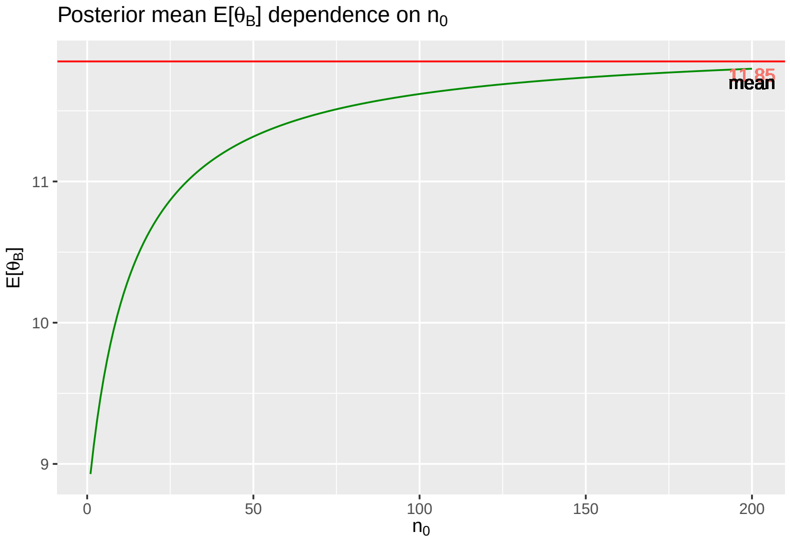
```
mean_varying %>% ggplot(aes(x = n_0, y = mean.b)) +
  geom_line(col = "green4") +
  xlab(expression(n[0])) +
  ylab(expression(paste("E[", theta[B], "]"))) +
  ggtitle(expression(paste("Posterior mean ", "E[", theta[B], "]", " dependence on ", n[0]))) +
  geom_hline(yintercept = mean_a, col = "red") +
  scale_color_discrete(guide = "none") +
  geom_text(aes(50, mean_a, label = mean_a, vjust = +1.5, col = "red")) +
  geom_text(aes(50, mean.b[50], label = "mean", vjust = +1.5))
```



As we can observe, the more n_0 grows, the more the posterior mean of θ_B approaches that of θ_A, as we can observe if we extend the interval of variation of n_0 itself:

```
mean.b = rep(0, 200)

n_0 = 1:200

for (i in n_0){
  a = 12 * i + sum(y.b)
  b = i + length(y.b)
  mean.b[i] = a / b
}

mean_varying_2 <- data.frame(n_0, mean.b)
```

```
mean_varying_2 %>% ggplot(aes(x = n_0, y = mean.b)) +
  geom_line(col = "green4") +
  xlab(expression(n[0])) +
  ylab(expression(paste("E[", theta[B], "]"))) +
  ggtitle(expression(paste("Posterior mean ", "E[", theta[B], "]", " dependence on ", n[0]))) +
  geom_hline(yintercept = mean_a, col = "red") +
  scale_color_discrete(guide = "none") +
  geom_text(aes(200, mean_a, label = mean_a, vjust =  +1.5, col = "red")) +
  geom_text(aes(200, mean.b[200], label = "mean", vjust = +1.5))
```



Therefore we can deduce that a large n_0 is required in order to have a posterior expectation close to that of θ_A.

Point c.

The independency assumption we have made in the model can be justified by the lack of information about the interrelations between the populations of mice. Even though we know that some form of dependency exists between the two parameters since the types of mice are related, we lack any additional information about the complexity of this relation, and therefore we can a priori assume the two parameters to be independent, since we need this to be able to compute the posterior distributions in a sensible way. Since the results seem to be consistent with this assumption, we justify the model through its own computability.

Part 2: Tumor Counts Comparison

Point d.

We can use the posterior distributions to obtain an approximation of $p\left(\theta_B < \theta_A \mid \mathbf{y}_A, \mathbf{y}_B\right)$ through Monte Carlo sampling. We start by obtaining that of the original distributions:

```
s = 10000

#Monte Carlo estimate with original p(theta[B])
sample.a = rgamma(s, a_n, b_n)
sample.b = rgamma(s, c_n, d_n)

mc1 = sum(sample.a > sample.b) / s

sprintf("The Monte Carlo estimate given the original prior of theta[B] is: %f", mc1)
```

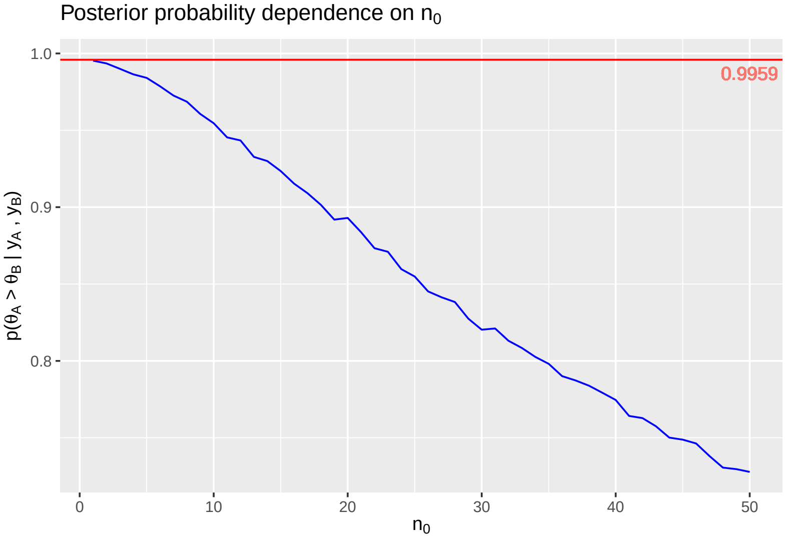
```
## [1] "The Monte Carlo estimate given the original prior of theta[B] is: 0.995900"
```

Point e.

Then we proceed to obtain the one for the case of the parametrized posterior of θ_B:

```
#Monte Carlo estimate with varying p(theta[B])
n_0 = 1:50
mc2 = rep(0, 50)

for (i in n_0){
  a = 12 * i + sum(y.b)
  b = i + length(y.b)
  sample2.b = rgamma(s, a, b)
  mc2[i] = sum(sample.a > sample2.b) / s
}

mc1_varying <- data.frame(n_0, mc2)
```

```
mc1_varying %>% ggplot(aes(x = n_0, y = mc2)) +
  geom_line(col = "blue") +
  xlab(expression(n[0])) +
  ylab(expression(paste("p(", theta[A], " > ", theta[B], " | ", y[A], " , ", y[B], ")"))) +
  ggtitle(expression(paste("Posterior probability dependence on ", n[0]))) +
  geom_hline(yintercept = mc1, col = "red") +
  scale_color_discrete(guide = "none") +
  geom_text(aes(50, mc1, label = mc1, vjust = +1.5, col = "red"))
```



The resulting Monte Carlo estimate is very sensitive to the variation of the prior of θ_B, since we have a relatively fast decay of the probability $p\,(\theta_B < \theta_A \mid \mathbf{y}_A, \mathbf{y}_B)$ with respect to the variation of the parameter n_0.

Point f.

We can also obtain an estimate for $p\,(\tilde{Y_B} < \tilde{Y_A} \mid \mathbf{y}_A, \mathbf{y}_B)$, where $\tilde{Y_A}, \tilde{Y_B}$ are samples from the posterior predictive distributions, using Monte Carlo sampling. Again we compare the results given the original prior for θ_B and the parametrized one: the fall is relatively less prominent ($\sim 20\%$ versus $\sim 40\%$ in the same interval of variation), but still strong, especially considering that the starting probability was already far lower than in the previous case.

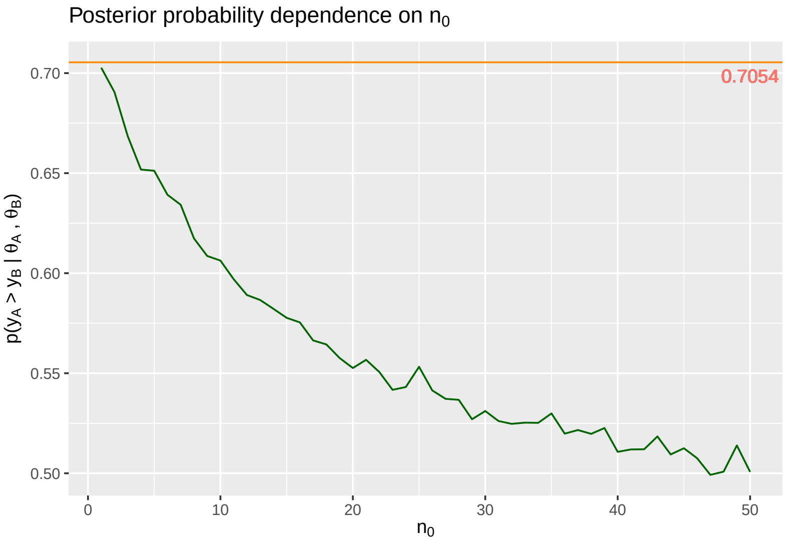
```
#Monte Carlo estimate of the posterior predictive distribution (original theta[B])

y.post.a = rep(0, s)
y.post.b = rep(0, s)

for (i in 1:s){
  y.post.a[i] = rpois(1, sample.a[i])
  y.post.b[i] = rpois(1, sample.b[i])
}
```

```
mc3 = sum(y.post.a > y.post.b) / s
sprintf("The Monte Carlo estimate given the original prior of theta[B] is: %f", mc3)
```

```
## [1] "The Monte Carlo estimate given the original prior of theta[B] is: 0.705400"
```

```
#Monte Carlo estimate of the posterior predictive distribution (varying theta[B])

mc4 = rep(0, 50)

for (i in 1:50){
  y.post2.b = rep(0, s)
  a = 12 * i + sum(y.b)
  b = i + length(y.b)
  sample2.b = rgamma(s, a, b)
  for (j in 1:s){
  y.post2.b[j] = rpois(1, sample2.b[j])
  }
  mc4[i] = sum(y.post.a > y.post2.b) / s
}

mc3_varying <- data.frame(n_0, mc4)
```

```
mc3_varying %>% ggplot(aes(x = n_0, y = mc4)) +
  geom_line(col = "darkgreen") +
  xlab(expression(n[0])) +
  ylab(expression(paste("p(", y[A], " > ", y[B], " | ", theta[A], " , ", theta[B], ")"))) +
  ggtitle(expression(paste("Posterior probability dependence on ", n[0]))) +
  geom_hline(yintercept = mc3, col = "darkorange") +
  scale_color_discrete(guide = "none") +
  geom_text(aes(50, mc3, label = mc3, vjust = +1.5, col = "orange"))
```



Part 3: Posterior predictive checks

Lastly one can investigate the adequacy of the Poisson model for the tumor count data in the following way.

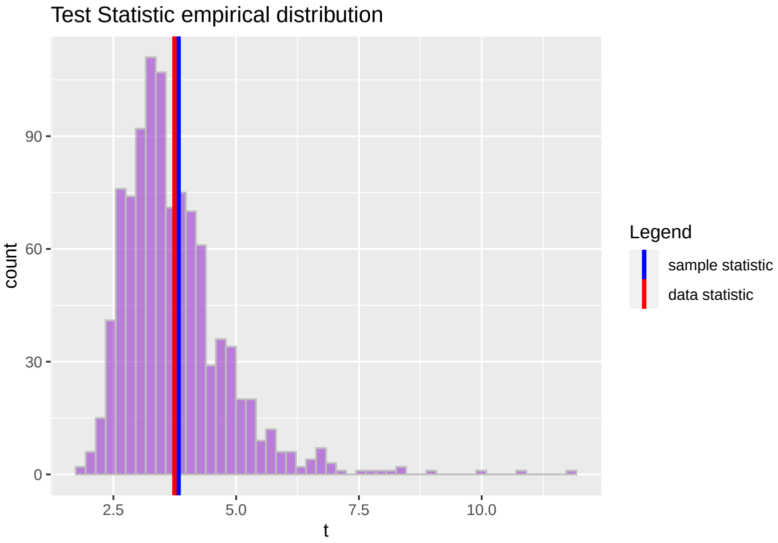
Point g.

First we generate posterior predictive datasets $\mathbf{y}_A^{(1)}, \ldots, \mathbf{y}_A^{(1000)}$, where each $\mathbf{y}_A^{(s)}$ is a sample of size $n_A = 10$ from the Poisson distribution with parameter $\theta_A^{(s)}$ and $\theta_A^{(s)}$ is itself a sample from the posterior distribution $p\,(\,\theta_A \mid Y_A = \mathbf{y}_A)$, and $\mathbf{y}_A$ is the observed data.

```
#generating the samples case theta[A]
```

```
K = 1000
sample.theta.a = rgamma(K, a_n, b_n)
t1 = rep(0, K)
for (i in 1:K){
  post.pred.a = rpois(10, sample.theta.a[i])
  t1[i] = sum(post.pred.a) / (10 * sd(post.pred.a))
}

t2 =  sum(y.a) / (10 * sd(y.a))

t.a <- data.frame(1:K, t1)

colors <- c("data mean" = "blue", "sample mean" = "red")
```

```
t.a %>% ggplot() +
  geom_histogram(aes(t1), bins = 50, fill = "darkorchid", alpha = 0.6, col = "grey") +
  geom_vline(aes(xintercept = t2, col = "sample mean"), size = 1.2)  +
  geom_vline(aes(xintercept = mean(t1), col = "data mean"), size = 1.2) +
  xlab("t") +
  ggtitle("Test Statistic empirical distribution") +
  scale_color_manual(name = "Legend", values = c("sample mean" = "blue", "data mean" = "red"), labels = c("sample statistic", "data
```



We can see that while the distribution is spread out, the expected value of the statistic is close to the one observed from the data, and the distribution itself is peaked around its mean value: this indicates that our original Poisson assumption seems to be corroborated by the data, and is therefore consistent with the observations.

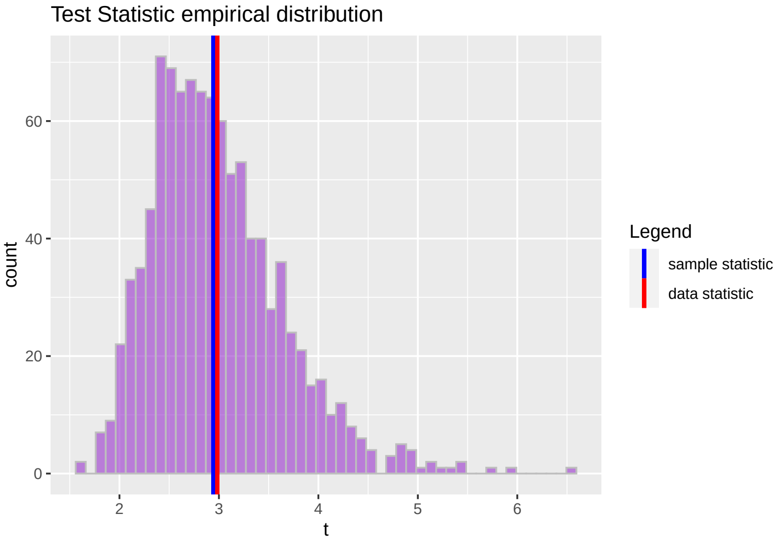
Point h.

We can repeat the same procedure with the second model, by generting posterior predictive datasets $\mathbf{y}_B^{(1)}, ...\mathbf{y}_B^{(1000)}$, where each $\mathbf{y}_B^{(s)}$ is a sample of size $n_B = 13$ from the Poisson distribution with parameter $\theta_B^{(s)}$ and $\theta_B^{(s)}$ is itself a sample from the posterior distribution $p\,(\,\theta_B \mid Y_B = \mathbf{y}_B)$, and $\mathbf{y}_B$ is the observed data.

```
#generating the samples case theta[B]

K = 1000
sample.theta.b = rgamma(K, c_n, d_n)
t3 = rep(0, K)
for (i in 1:K){
  post.pred.b = rpois(13, sample.theta.b[i])
  t3[i] = sum(post.pred.a) / (13 * sd(post.pred.b))
}
```

```
t4 =  sum(y.a) / (13 * sd(y.a))

t.b <- data.frame(1:K, t3)
```

```
t.b %>% ggplot() +
  geom_histogram(aes(t3), bins = 50, fill = "darkorchid", alpha = 0.6, col = "grey") +
  geom_vline(aes(xintercept = t4, col = "sample statistic"), size = 1.2) +
  geom_vline(aes(xintercept = mean(t3), col = "data statistic"), size = 1.2) +
  xlab("t") +
  ggtitle("Test Statistic empirical distribution") +
  scale_color_manual(name = "Legend", values = c("sample statistic" = "blue", "data statistic" = "red"), labels = c("sample statisti
```



Analogously with the case of θ_A, we have a distribution that is spread out, but with expected value close to the one observed from the data, and the distribution itself is peaked around its mean value: this indicates that our original Poisson assumption seems to be corroborated by the data, and is therefore consistent with the observations.