ST740 – Assignment 1 – Solution

TA: Jianing Chu

1. Assume $Y_1, ..., Y_n|\theta \sim \text{Uniform}(0, \theta)$ independent over i.

(a). Identify a conjugate family of prior distributions for θ and derive the posterior.

Solution: The likelihood is

$$f(\boldsymbol{Y}|\theta) = \frac{1}{\theta^n}\mathbb{I}\{\theta > Y_{(n)}\}$$

Consider a Pareto prior distribution, $\theta \sim \text{Pareto}(\alpha, \beta)$, with density

$$\pi(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}}\mathbb{I}(\theta > \beta).$$

We have

$$\begin{aligned} p(\theta|\boldsymbol{Y}) &\propto f(\boldsymbol{Y}|\theta)\pi(\theta) \\ &\propto \frac{1}{\theta^{\alpha+n+1}}\mathbb{I}\left[\theta > \max\{Y_{(n)}, \beta\}\right]. \end{aligned}$$

The posterior distribution is $\text{Pareto}(\alpha + n, \max\{Y_{(n)}, \beta\})$.

(b). Now assume you observe $n = 50$ samples as below

```
> set.seed(919)
> Y <- runif(50,0,10)
> range(Y)
[1] 0.05161189 9.75337425
```

Use an uninformative prior and summarize the posterior in a table and plot.

Solution: Let $\beta < Y_{(n)}$ and α be small enough, e.g., $\alpha = 0.01$. Thus the posterior distribution is $\text{Pareto}(0.01 + 50, Y_{(n)})$.

```
set.seed(919)
n <- 50
Y <- runif(n,0,10)
Y_max <- max(Y)
alpha <- 0.01
# posterior mean
(alpha+n) * Y_max / (alpha+n-1)
```

```
## [1] 9.952382
```

```
# posterior sd
sqrt(Y_max^2 * (alpha+n) / (alpha+n-1)^2 / (alpha+n-2))
```

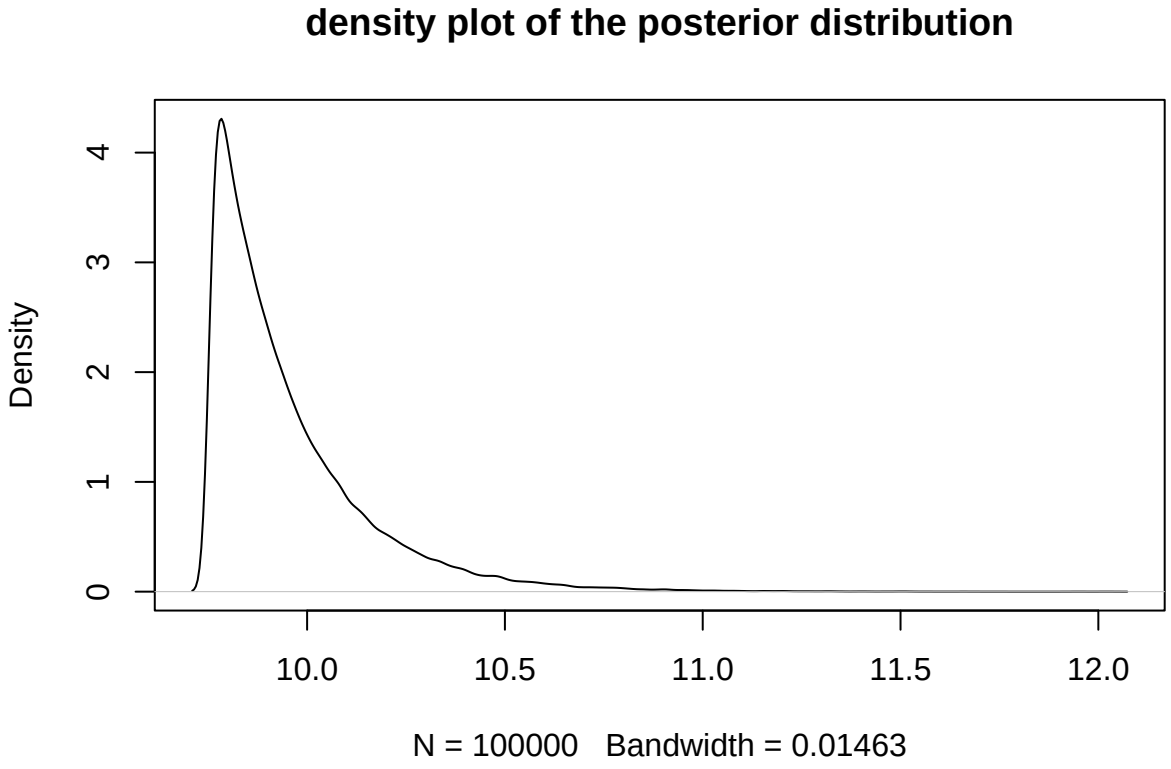
```
## [1] 0.2031107
```

```r
# 95% credible set
library(Pareto)
qPareto(c(0.025,0.975),Y_max,alpha+n)
```

```
## [1]  9.758313 10.500009
```

Posterior mean	Posterior SD	95% credible set
9.952382	0.20311071	(9.758313, 10.500009)

```r
# generate 10^5 random samples from the posterior distribution
pst_sample <- rPareto(10^5, Y_max,alpha+n)
plot(density(pst_sample), main ="density plot of the posterior distribution")
```



(c). Is the posterior senstive to the prior?

Solution: The posterior mean is $\frac{(\alpha+n)\max\{Y_{(n)},\beta\}}{\alpha+n-1}$ and the posterior variance is $\frac{(\alpha+n)[\max\{Y_{(n)},\beta\}]^2}{(\alpha+n-1)^2(\alpha+n-2)}$. When $\beta > Y_{(n)}$ or n is small, the posterior is sensitive to to the prior. When $\beta < Y_{(n)}$ and $n \to \infty$, the posterior mean converges to $Y_{(n)}$ and the posterior variance converges to 0, and thus the posterioris not sensitive to to the prior.

(d). What is the posterior predictive probability that Y_{n+1} will be a new record, i.e.,

$$\text{Prob}(Y_{n+1} > \max\{Y_1, ..., Y_n\} | Y_1, ..., Y_n).$$

Solution: Let $\tilde{\beta} = \max\{Y_{(n)}, \beta\}$.

$$\text{Prob}(Y_{n+1} > Y_{(n)} | \theta) = \frac{\theta - Y_{(n)}}{\theta}.$$

$$
\begin{aligned}
\text{Prob}(Y_{n+1} > Y_{(n)}|\boldsymbol{Y}) &= \int \text{Prob}(Y_{n+1} > Y_{(n)}|\theta)\text{Prob}(\theta|\boldsymbol{Y})d\theta \\
&= \int_{\tilde{\beta}}^{\infty} \frac{\theta - Y_{(n)}}{\theta} \frac{(\alpha+n)\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+1}} d\theta \\
&= \int_{\tilde{\beta}}^{\infty} \frac{(\alpha+n)\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+1}} d\theta - \frac{(\alpha+n)Y_{(n)}}{(\alpha+n+1)\tilde{\beta}} \int_{\tilde{\beta}}^{\infty} \frac{(\alpha+n+1)\tilde{\beta}^{\alpha+n+1}}{\theta^{\alpha+n+2}} d\theta \\
&= 1 - \frac{Y_{(n)}(\alpha+n)}{(\alpha+n+1)\tilde{\beta}} \\
&= 1 - \frac{(\alpha+n)Y_{(n)}}{(\alpha+n+1)\max\{Y_{(n)}, \beta\}}
\end{aligned}
$$

(e). Why is (d) not exactly $1/(n+1)$, or is it?

Solution: According to (d), the posterior predictive probability is not $1/(n+1)$. When $\beta < Y_{(n)}$, $\text{Prob}(Y_{n+1} > Y_{(n)}|\boldsymbol{Y}) = \frac{1}{n+\alpha+1}$ and will approach $1/(n+1)$ as $\alpha \to 0$.

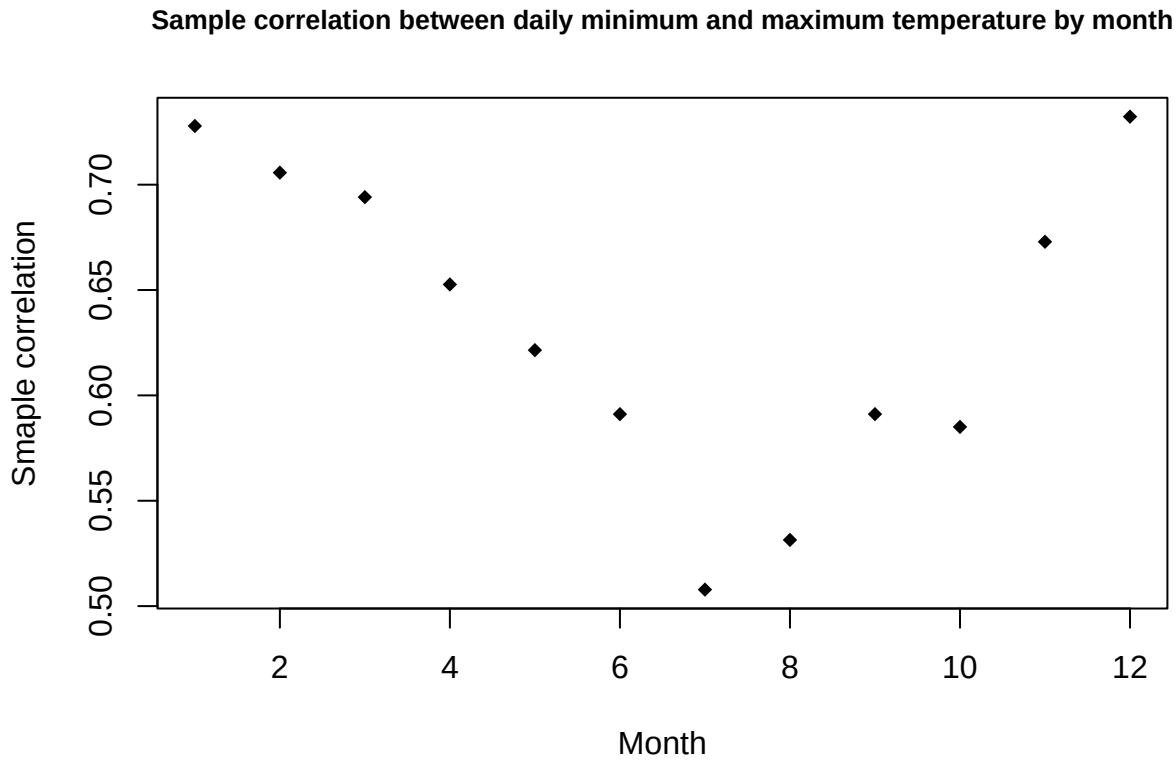
2. Download the daily weather data from RDU Airport

```
file  <- "https://www4.stat.ncsu.edu/~bjreich/ST740/RDU.csv"
dat   <- read.csv(url(file))
TMAX  <- dat[,2]/10
TMIN  <- dat[,3]/10
MONTH <- dat[,4]
```

(a). Plot the sample correlation between daily minimum (TMIN) and maximum (TMAX) temperature by month.

Solution:

```
file  <- "https://www4.stat.ncsu.edu/~bjreich/ST740/RDU.csv"
dat   <- read.csv(url(file))
TMAX  <- dat[,2]/10
TMIN  <- dat[,3]/10
MONTH <- dat[,4]
cor_month = rep(0,12)
for (k in 1:12){
  cor_month[k] <- cor(TMIN[MONTH==k],TMAX[MONTH==k])
}
plot(x = 1:12, y = cor_month, pch =18, xlab = "Month", ylab = "Smaple correlation",
     main = "Sample correlation between daily minimum and maximum temperature by month",
     cex.main = 0.8)
```



(b). Let $Y = (T_{MIN}, T_{MAX})$ be a bivariate response and fit the model $Y|\Sigma \sim \text{Normal}(\bar{Y}, \Sigma)$ where $\bar{Y}$ is the sample mean of Y (say it is fixed and known) and Σ is the unknown 2×2 covariance matrix. Specify a conjugate family of prior distributions for Σ and derive the corresponding posterior.[1]

Solution: Suppose $\Sigma \sim \text{inverseWishart}(\Psi, \nu)$. Then

$$\pi(\Sigma) \propto \det(\Sigma)^{-(\nu+3)/2} \exp\{-\text{tr}(\Psi\Sigma^{-1})/2\}.$$

The likelihood

$$\begin{aligned}
p(\boldsymbol{Y}|\bar{Y}, \Sigma) &\propto \det(\Sigma)^{-n/2} \exp\left\{-\frac{1}{2}\sum_i (Y_i - \bar{Y})^T \Sigma^{-1} (Y_i - \bar{Y})\right\} \\
&\propto \det(\Sigma)^{-n/2} \exp\left[-\frac{1}{2}\sum_i \text{tr}\left\{(Y_i - \bar{Y})^T \Sigma^{-1} (Y_i - \bar{Y})\right\}\right] \\
&\propto \det(\Sigma)^{-n/2} \exp\left[-\frac{1}{2}\sum_i \text{tr}\left\{(Y_i - \bar{Y})(Y_i - \bar{Y})^T \Sigma^{-1}\right\}\right] \\
&\propto \det(\Sigma)^{-n/2} \exp\left[-\frac{1}{2}\text{tr}\left\{\sum_i (Y_i - \bar{Y})(Y_i - \bar{Y})^T \Sigma^{-1}\right\}\right].
\end{aligned}$$

We have

$$\begin{aligned}
p(\Sigma|\boldsymbol{Y}) &\propto \pi(\Sigma) p(\boldsymbol{Y}|\bar{Y}, \Sigma) \\
&\propto \det(\Sigma)^{-(n+\nu+3)/2} \exp\left[-\frac{1}{2}\text{tr}\left\{(\Psi + \sum_i (Y_i - \bar{Y})(Y_i - \bar{Y})^T)\Sigma^{-1}\right\}\right].
\end{aligned}$$

Thus, the posterior is $\text{inverseWishart}(\Psi + \sum_i (Y_i - \bar{Y})(Y_i - \bar{Y})^T, n + \nu)$.

[1] Recall that for vector a and square matrices B and C that $a^T B a = \text{trace}(a^T B a) = \text{trace}(B a a^T)$ and $\text{trace}(a^T B) + \text{trace}(a^T C) = \text{trace}(a^T (B + C))$

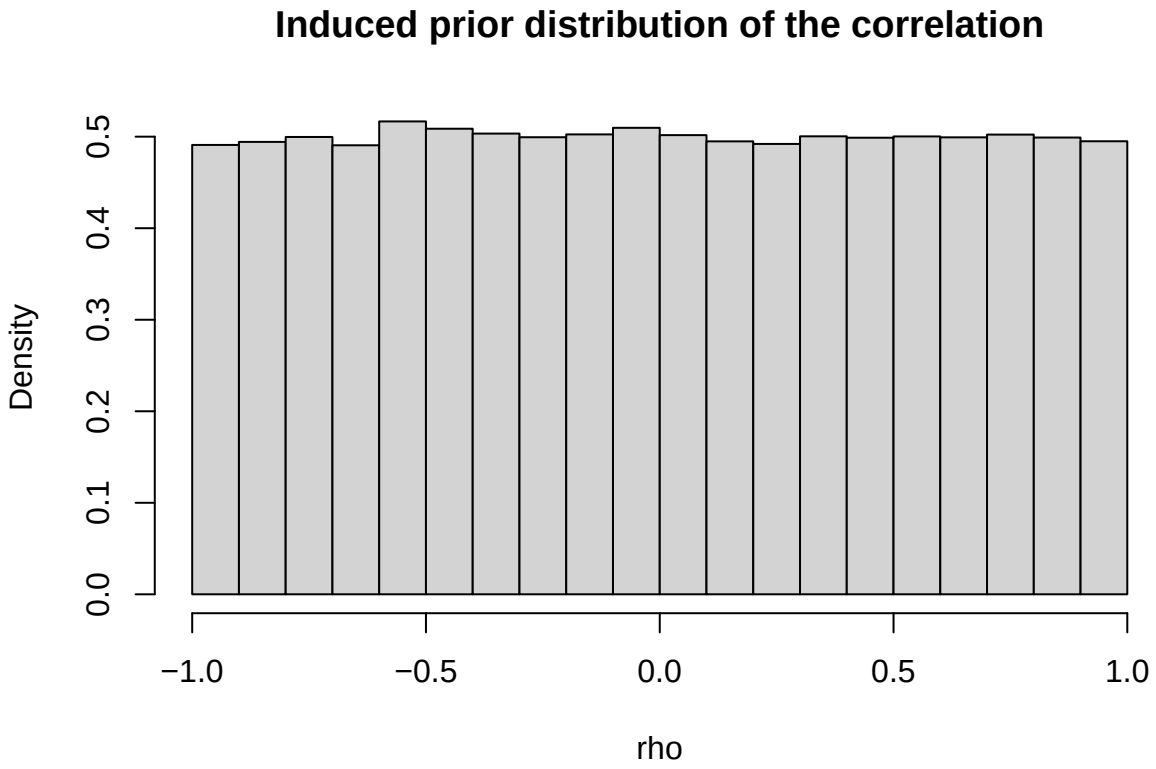
(c). Select an uninformative prior distribution and summarize the induced prior distribution on the correlation $\rho = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$ in figure.

Solution: Let the prior be inverseWishart$(\boldsymbol{I}_2, 3)$.

```
library(MCMCpack)
p <- 2
nu <- p+1
Psi <- diag(p)
S <- 10^5
rho <- rep(0,S)

for(s in 1:S){
Sigma <- riwish(nu,Psi)
rho[s] <- Sigma[1,2]/sqrt(Sigma[1,1]*Sigma[2,2])
}

hist(rho, freq =FALSE,
     main = "Induced prior distribution of the correlation", xlab = 'rho')
```

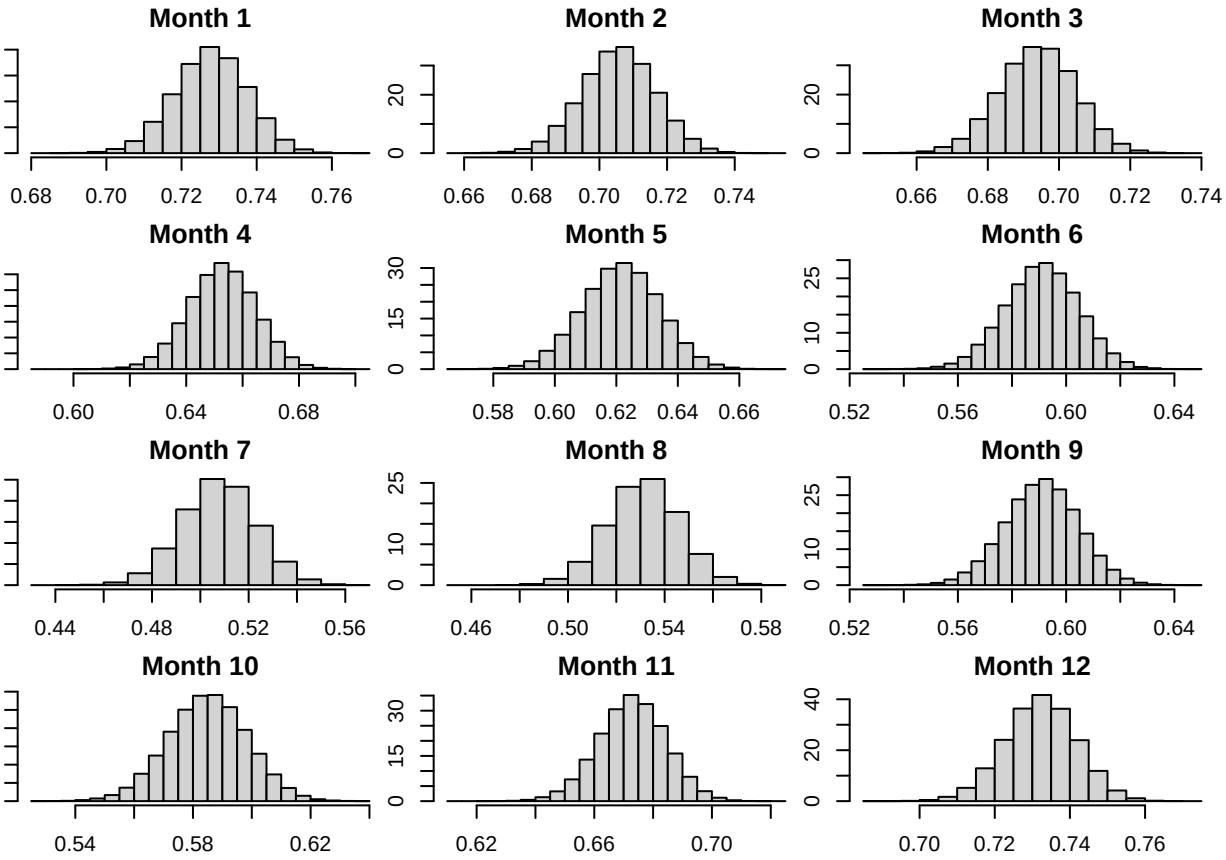


(d). Fit the model to the temperature data separately by month (including a monthly $\bar{Y}$) and plot the posterior distribution of the correlation between T_{MIN} and T_{MAX} (ρ) by month. Is there a statistically significant correlation between these variables? Are there statistically signficant differences by month?

Solution:

```
pst_Psi <- array(0,c(12,p,p))
pst_nu <- rep(0,12)
pst_rho <- rep(0,12*S)
pst_rho_list <- vector("list", 12)
```

```
par(mfrow=c(4, 3),mar=c(2,1,2,1))
for(k in 1:12){
n <- length(which(MONTH == k))
Y_bar <- c(mean(TMIN[MONTH == k]), mean(TMAX[MONTH == k]))
Y <- matrix(c(TMIN[MONTH == k], TMAX[MONTH == k]), ncol = 2)
Y <- apply(Y, 1, function(x) x-Y_bar)
pst_Psi[k,,] <- Y %*% t(Y) + Psi
pst_nu[k] <- n + nu
for(s in 1:S){
  Sigma<- riwish(pst_nu[k],pst_Psi[k,,])
  pst_rho[(k-1)*S+s] <- Sigma[1,2]/sqrt(Sigma[1,1]*Sigma[2,2])
}
pst_rho_list[[k]] <- pst_rho[((k-1)*S+1):(k*S)]
hist(pst_rho_list[[k]], freq =FALSE, xlab = '', main = paste("Month", k))
}
```




```
for(k in 1:12){
LB <- quantile(pst_rho_list[[k]], 0.025)
UB  <-  quantile(pst_rho_list[[k]], 0.975)
cat("Month", k, "95% credible set:", sprintf("(%1.4f, %1.4f)", LB, UB),"\n")
}
```

```
## Month 1 95% credible set: (0.7085, 0.7463)
## Month 2 95% credible set: (0.6841, 0.7264)
## Month 3 95% credible set: (0.6728, 0.7144)
```

```
## Month 4 95% credible set: (0.6289, 0.6756)
## Month 5 95% credible set: (0.5965, 0.6455)
## Month 6 95% credible set: (0.5640, 0.6172)
## Month 7 95% credible set: (0.4776, 0.5372)
## Month 8 95% credible set: (0.5020, 0.5596)
## Month 9 95% credible set: (0.5639, 0.6170)
## Month 10 95% credible set: (0.5584, 0.6109)
## Month 11 95% credible set: (0.6498, 0.6948)
## Month 12 95% credible set: (0.7133, 0.7504)
```

There is a statistically significant correlation between T_{MIN} and T_{MAX} by month.

```
for(k in 1:11){
  for(l in (k+1):12){
    pst_difference <- pst_rho_list[[k]]-pst_rho_list[[l]]
    LB <- quantile(pst_difference,0.025)
    UB <- quantile(pst_difference,0.975)
    if (LB < 0 & UB >0){
      cat("Month", k, "and Month", l, "difference 95% credible set:",
          sprintf("(%1.4f, %1.4f)", LB, UB), "significant\n")
    }
    else{
    cat("Month", k, "and Month", l, "difference 95% credible set:",
        sprintf("(%1.4f, %1.4f)", LB, UB),"\n")
    }
  }
}
```

```
## Month 1 and Month 2 difference 95% credible set: (-0.0060, 0.0507) significant
## Month 1 and Month 3 difference 95% credible set: (0.0059, 0.0619)
## Month 1 and Month 4 difference 95% credible set: (0.0451, 0.1054)
## Month 1 and Month 5 difference 95% credible set: (0.0755, 0.1376)
## Month 1 and Month 6 difference 95% credible set: (0.1042, 0.1698)
## Month 1 and Month 7 difference 95% credible set: (0.1848, 0.2557)
## Month 1 and Month 8 difference 95% credible set: (0.1622, 0.2314)
## Month 1 and Month 9 difference 95% credible set: (0.1045, 0.1697)
## Month 1 and Month 10 difference 95% credible set: (0.1108, 0.1754)
## Month 1 and Month 11 difference 95% credible set: (0.0256, 0.0847)
## Month 1 and Month 12 difference 95% credible set: (-0.0308, 0.0221) significant
## Month 2 and Month 3 difference 95% credible set: (-0.0180, 0.0416) significant
## Month 2 and Month 4 difference 95% credible set: (0.0215, 0.0844)
## Month 2 and Month 5 difference 95% credible set: (0.0520, 0.1170)
## Month 2 and Month 6 difference 95% credible set: (0.0807, 0.1485)
## Month 2 and Month 7 difference 95% credible set: (0.1616, 0.2347)
## Month 2 and Month 8 difference 95% credible set: (0.1387, 0.2103)
## Month 2 and Month 9 difference 95% credible set: (0.0810, 0.1487)
## Month 2 and Month 10 difference 95% credible set: (0.0870, 0.1544)
## Month 2 and Month 11 difference 95% credible set: (0.0021, 0.0637)
## Month 2 and Month 12 difference 95% credible set: (-0.0548, 0.0015) significant
## Month 3 and Month 4 difference 95% credible set: (0.0100, 0.0727)
## Month 3 and Month 5 difference 95% credible set: (0.0403, 0.1048)
## Month 3 and Month 6 difference 95% credible set: (0.0692, 0.1372)
## Month 3 and Month 7 difference 95% credible set: (0.1497, 0.2227)
```

```
## Month 3 and Month 8 difference 95% credible set: (0.1272, 0.1983)
## Month 3 and Month 9 difference 95% credible set: (0.0694, 0.1369)
## Month 3 and Month 10 difference 95% credible set: (0.0754, 0.1426)
## Month 3 and Month 11 difference 95% credible set: (-0.0094, 0.0518) significant
## Month 3 and Month 12 difference 95% credible set: (-0.0662, -0.0103)
## Month 4 and Month 5 difference 95% credible set: (-0.0026, 0.0652) significant
## Month 4 and Month 6 difference 95% credible set: (0.0263, 0.0970)
## Month 4 and Month 7 difference 95% credible set: (0.1071, 0.1830)
## Month 4 and Month 8 difference 95% credible set: (0.0844, 0.1585)
## Month 4 and Month 9 difference 95% credible set: (0.0264, 0.0973)
## Month 4 and Month 10 difference 95% credible set: (0.0327, 0.1029)
## Month 4 and Month 11 difference 95% credible set: (-0.0525, 0.0122) significant
## Month 4 and Month 12 difference 95% credible set: (-0.1093, -0.0496)
## Month 5 and Month 6 difference 95% credible set: (-0.0058, 0.0668) significant
## Month 5 and Month 7 difference 95% credible set: (0.0747, 0.1524)
## Month 5 and Month 8 difference 95% credible set: (0.0522, 0.1281)
## Month 5 and Month 9 difference 95% credible set: (-0.0058, 0.0670) significant
## Month 5 and Month 10 difference 95% credible set: (0.0006, 0.0725)
## Month 5 and Month 11 difference 95% credible set: (-0.0846, -0.0181)
## Month 5 and Month 12 difference 95% credible set: (-0.1415, -0.0802)
## Month 6 and Month 7 difference 95% credible set: (0.0432, 0.1234)
## Month 6 and Month 8 difference 95% credible set: (0.0203, 0.0991)
## Month 6 and Month 9 difference 95% credible set: (-0.0378, 0.0378) significant
## Month 6 and Month 10 difference 95% credible set: (-0.0316, 0.0435) significant
## Month 6 and Month 11 difference 95% credible set: (-0.1169, -0.0472)
## Month 6 and Month 12 difference 95% credible set: (-0.1739, -0.1088)
## Month 7 and Month 8 difference 95% credible set: (-0.0649, 0.0184) significant
## Month 7 and Month 9 difference 95% credible set: (-0.1232, -0.0430)
## Month 7 and Month 10 difference 95% credible set: (-0.1170, -0.0376)
## Month 7 and Month 11 difference 95% credible set: (-0.2025, -0.1277)
## Month 7 and Month 12 difference 95% credible set: (-0.2599, -0.1893)
## Month 8 and Month 9 difference 95% credible set: (-0.0991, -0.0205)
## Month 8 and Month 10 difference 95% credible set: (-0.0928, -0.0144)
## Month 8 and Month 11 difference 95% credible set: (-0.1783, -0.1051)
## Month 8 and Month 12 difference 95% credible set: (-0.2354, -0.1670)
## Month 9 and Month 10 difference 95% credible set: (-0.0316, 0.0432) significant
## Month 9 and Month 11 difference 95% credible set: (-0.1165, -0.0470)
## Month 9 and Month 12 difference 95% credible set: (-0.1740, -0.1090)
## Month 10 and Month 11 difference 95% credible set: (-0.1225, -0.0533)
## Month 10 and Month 12 difference 95% credible set: (-0.1796, -0.1152)
## Month 11 and Month 12 difference 95% credible set: (-0.0888, -0.0302)
```