Worksheet 3

Foundations of Bayesian Methodology

Wenje Tu Lea Bührer Jerome Sepin Zhixuan Li
Elia-Leonid Mastropietro Jonas Raphael Füglistaler

Spring Semester 2022

Exercise 3 (Conjugate Bayes: analytical derivation)

Assumptions:

$$\begin{aligned} y_1, \cdots, y_n \mid m &\overset{i.i.d}{\sim} \mathcal{N}(m, \kappa^{-1}) \\ m &\sim \mathcal{N}(\mu, \lambda^{-1}) \end{aligned}$$

3 (a)

The prior predictive distribution of one future observation y assuming that no observations have been collected yet:

$$\begin{aligned} f(y) &= \int_{-\infty}^{\infty} f(y \mid m) f(m) \mathrm{d}m \\ &= \int_{-\infty}^{\infty} \sqrt{\frac{\kappa}{2\pi}} \exp\left(-\frac{\kappa}{2}(y-m)^2\right) \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(m-\mu)^2\right) \mathrm{d}m \\ &= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\kappa(y^2 - 2ym + m^2) + \lambda(m^2 - 2m\mu + \mu^2)\right)\right) \mathrm{d}m \\ &= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left((\kappa+\lambda)\left(m - \frac{\kappa y + \lambda\mu}{\kappa+\lambda}\right)^2 - \frac{(\kappa y + \lambda\mu)^2}{\kappa+\lambda} + \kappa y^2 + \lambda\mu^2\right)\right) \mathrm{d}m \\ &= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left((\kappa+\lambda)\left(m - \frac{\kappa y + \lambda\mu}{\kappa+\lambda}\right)^2 + \frac{\kappa\lambda(y-\mu)^2}{\kappa+\lambda}\right)\right) \mathrm{d}m \\ &= \frac{\sqrt{\kappa\lambda}}{2\pi} \exp\left(-\frac{\kappa\lambda(y-\mu)^2}{2(\kappa+\lambda)}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{\kappa+\lambda}{2}\left(m - \frac{\kappa y + \lambda\mu}{\kappa+\lambda}\right)^2\right) \mathrm{d}m \\ &= \frac{\sqrt{\kappa\lambda}}{2\pi} \sqrt{\frac{2\pi}{\kappa+\lambda}} \exp\left(-\frac{\kappa\lambda(y-\mu)^2}{2(\kappa+\lambda)}\right) \underbrace{\int_{-\infty}^{\infty} \sqrt{\frac{\kappa+\lambda}{2\pi}} \exp\left(-\frac{\kappa+\lambda}{2}\left(m - \frac{\kappa y + \lambda\mu}{\kappa+\lambda}\right)^2\right) \mathrm{d}m}_{=1} \\ &= \frac{\sqrt{\kappa\lambda}}{2\pi} \sqrt{\frac{2\pi}{\kappa+\lambda}} \exp\left(-\frac{\kappa\lambda(y-\mu)^2}{2(\kappa+\lambda)}\right) \\ &= \sqrt{\frac{1}{2\pi\left(\frac{1}{\lambda} + \frac{1}{\kappa}\right)}} \exp\left(-\frac{(y-\mu)^2}{2\left(\frac{1}{\lambda} + \frac{1}{\kappa}\right)}\right) \end{aligned}$$

The prior predictive distribution of one future observation y is

$$\mathcal{N}(\mu, \lambda^{-1} + \kappa^{-1})$$

3 (b)

The posterior predictive distribution of one future observation y_{n+1} given that $y_1, \cdots, y_n$ have been observed:

$$\begin{aligned} f(y_{n+1} \mid y_1, \cdots, y_n) &= \int_{-\infty}^{\infty} f(y_{n+1}, m \mid y_1, \cdots, y_n) \mathrm{d}m \\ &= \int_{-\infty}^{\infty} f(y_{n+1} \mid m, y_1, \cdots, y_n) f(m \mid y_1, \cdots, y_n) \mathrm{d}m \\ &= \int_{-\infty}^{\infty} f(y_{n+1} \mid m) f(m \mid y_1, \cdots, y_n) \mathrm{d}m \end{aligned}$$

$$f(m \mid y_1, \cdots, y_n) = \sqrt{\frac{n\kappa + \lambda}{2\pi}} \exp\left(-\frac{n\kappa + \lambda}{2} \left(m - \frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda} \right)^2 \right)$$

Denote:

$$\mu_{\text{post}} = \frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda}$$

$$\lambda_{\text{post}} = n\kappa + \lambda$$

$$f(y_{n+1} \mid y_1, \cdots, y_n) = \int_{-\infty}^{\infty} \sqrt{\frac{\kappa}{2\pi}} \exp\left(-\frac{\kappa}{2} (y_{n+1} - m)^2 \right) \sqrt{\frac{\lambda_{\text{post}}}{2\pi}} \exp\left(-\frac{\lambda_{\text{post}}}{2} (m - \mu_{\text{post}})^2 \right) \mathrm{d}m$$

Repeating the same derivation steps as for the prior predictive distribution, we obtain the posterior predictive distribution:

$$y_{n+1} \mid y_1, \cdots, y_n \sim \mathcal{N}(\mu_{\text{post}}, \lambda_{\text{post}}^{-1} + \kappa^{-1})$$

Exercise 4 (Conjugate Bayesian analysis in practice)

4 (a)

Plot the prior predictive distribution for one observation y and compute its expectation and standard deviation. Estimate P[y > 200] for one future observation of Height.

Prior predictive distribution:

$$y \sim \mathcal{N}(\mu, \lambda^{-1} + \kappa^{-1}) = \mathcal{N}(161, 970)$$

```
height <- c(166,168,168,177,160,170,172,159,175,164,175,167,164)
n      <- length(height) # sample size
y.bar  <- mean(height)   # sample mean
kappa  <- 1/900          # precision for data
mu     <- 161            # prior mean
lambda <- 1/70           # prior precision

## Compute posterior mean and precision
mu.post     <- (kappa*n*y.bar + lambda*mu) / (n*kappa + lambda)
lambda.post <- n*kappa + lambda
```

So the expectation for y equals $\mu = 161$ and the variance is $\lambda^{-1} + \kappa^{-1} = 900 + 70 = 970$.

Prior predictive distribution:

```r
curve(dnorm(x, mean=mu, sd=sqrt(1/kappa)), ylim=c(0, 0.015), xlim=c(71, 251),
      col=2, lwd=2, xlab="Height", ylab="Density", main="Prior Predictive Distribution")
```

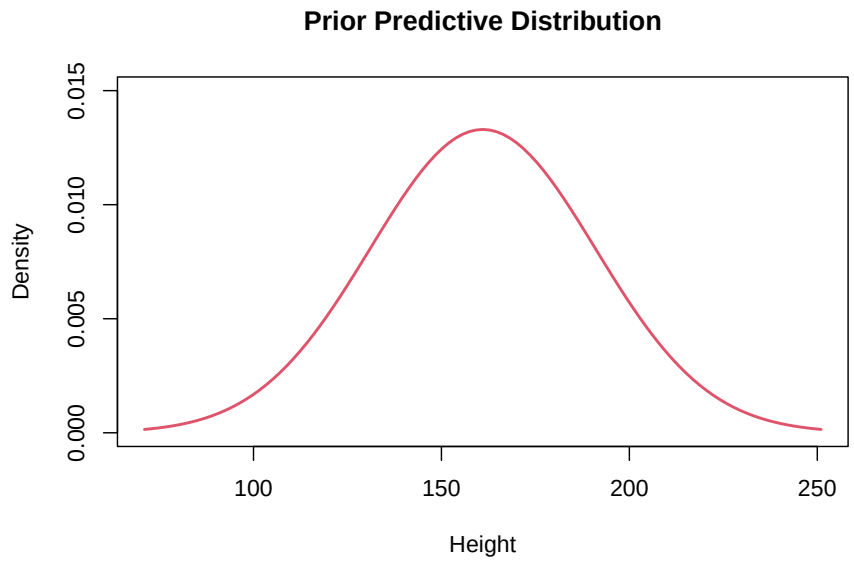


Figure 1: Prior Predictive Distribution

```r
## P[y > 200] for one future observation of Height
pnorm(200, mean=mu, sd=sqrt(1/lambda + 1/kappa), lower.tail=F)
```

```
## [1] 0.1052459
```

The probability that an observation of Height larger than 200 cm is made equals 0.105. So it can be concluded that we expect around 10% of future observation to be larger than 200 cm.

4 (b)

```r
## Compute posterior mean and precision
mu.post     <- (kappa*n*y.bar + lambda*mu) / (n*kappa + lambda)
lambda.post <- n*kappa + lambda

cat(sprintf("Posterior mean is %.4f \nPosterior inversed precision is %.4f",
            mu.post, 1/lambda.post))
```

```
## Posterior mean is 164.5580
## Posterior inversed precision is 34.8066
```

Posterior distribution:

$$m \mid y_1, \cdots, y_n \sim \mathcal{N}\left(\mu_{\text{post}}, \lambda_{\text{post}}^{-1}\right) \equiv \mathcal{N}(164.558, 34.8066)$$

Posterior predictive distribution:

$$y_{n+1} \mid y_1, \cdots, y_n \sim \mathcal{N}\left(\mu_{\text{post}}, \lambda_{\text{post}}^{-1} + \kappa^{-1}\right) \equiv \mathcal{N}(164.558, 934.8066)$$

where $\mu_{\text{post}} = \dfrac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda}$ and $\lambda_{\text{post}} = n\kappa + \lambda$

```
curve(dnorm(x, mean=mu.post, sd=sqrt(1/lambda.post+1/kappa)), ylim=c(0, 0.015),
      xlim=c(61, 261), col=4, lwd=2, xlab="Height", ylab="Density",
      main="Posterior Predictive Distribution")
```

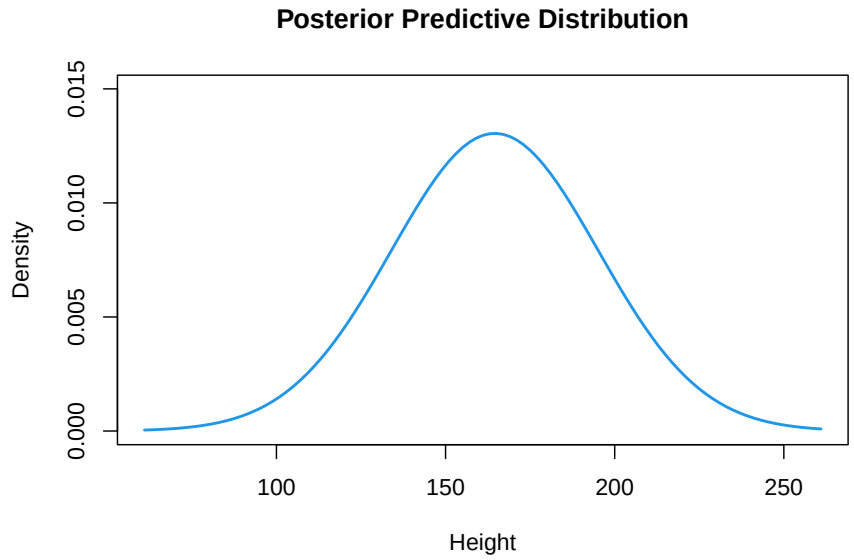


Figure 2: Posterior Predictive Distribution

Estimate for $P[y_{n+1} > 200|y_1, ..., y_n]$ for one future observation y_{n+1}:

```
pnorm(200, mean=mu.post, sd=sqrt(1/lambda.post+1/kappa), lower.tail=F)
```

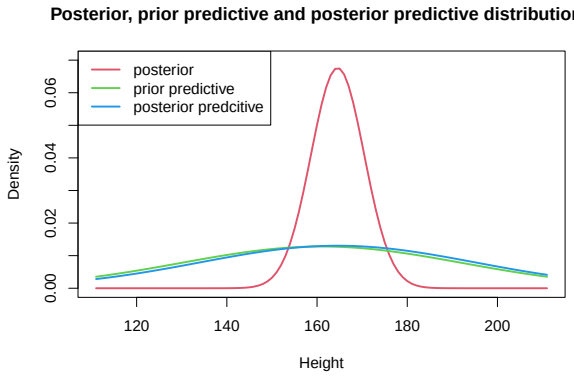
```
## [1] 0.123188
```

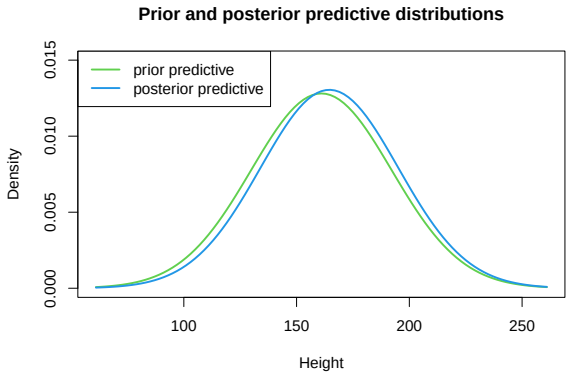
As a result we obtain a probability of 12.3% that a future observation of height will be larger than 200 cm.

4 (c)

Comparison between posterior, prior predictive, and posterior distributions

```
curve(dnorm(x, mean=164.558, sd=sqrt(34.80663)), ylim=c(0, 0.07), xlim=c(111, 211),
      col=2, lwd=2, xlab="Height", ylab="Density",
      main="Posterior, prior predictive and posterior predictive distributions")
curve(dnorm(x, mean=161, sd=sqrt(970)), col=3, lwd=2, add=TRUE)
curve(dnorm(x, mean=164.558, sd=sqrt(934.80663)), col=4, lwd=2, add=TRUE)
legend("topleft", legend=c("posterior", "prior predictive", "posterior predcitive"),
       col=2:4, lwd=2)

curve(dnorm(x, mean=161, sd=sqrt(970)), ylim=c(0, 0.015), xlim=c(61, 261), col=3, lwd=2,
      xlab="Height", ylab="Density", main="Prior and posterior predictive distributions")
curve(dnorm(x, mean=164.558, sd=sqrt(934.80663)), col=4, lwd=2, add=TRUE)
legend("topleft", legend=c("prior predictive", "posterior predictive"), col=3:4, lwd=2)
```






```
df <- data.frame(
  Mean=c(mu.post, mu, mu.post),
  Variance=c(1/lambda.post, 1/lambda+1/kappa, 1/lambda.post+1/kappa),
  SD=c(sqrt(1/lambda.post), sqrt(1/lambda+1/kappa), sqrt(1/lambda.post+1/kappa))
)
rownames(df) <- c("Posterior", "Prior predictive", "Posterior predictive")
knitr::kable(df, align="c", caption="Summary statistics")
```

Table 1: Summary statistics

	Mean	Variance	SD
Posterior	164.558	34.80663	5.899714
Prior predictive	161.000	970.00000	31.144823
Posterior predictive	164.558	934.80663	30.574608

In this example, the prior predictive and posterior predictive distributions do not differ too much. However, the posterior distribution differs dramatically from the other two predictive distributions, and this is attributed to the difference in the variance of the distributions.

The posterior distribution refers to the distribution of the parameter m while the prior (posterior) predictive distribution refers to the distribution of one future observation of `Height`. The posterior distribution has a more narrow shape (i.e. smaller variance) and it implies that we are more certain about the m. Under extreme circumstances where $\lim_{n\to\infty}(n\kappa+\lambda)^{-1} \approx 0$, there is little variance in the posterior distribution, but it does not mean that there is no variance in posterior predictive distribution. Posterior predictive distribution ensures that we are not too optimistic about the distribution of future observations.

Exercise 5 (The change-of-variables formula)

$$X \sim \text{Gamma}(a, b)$$

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$$

$$\begin{aligned} P(Y \leq y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ F_Y(y) &= F_X(g^{-1}(y)) \end{aligned}$$

By differentiating the CDFs on both sides w.r.t. y, we can get the PDF of Y.

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \cdot \dfrac{d}{dy} g^{-1}(y) & g(\cdot) \text{ is monotonically increasing} \\ -f_X(g^{-1}(y)) \cdot \dfrac{d}{dy} g^{-1}(y) & g(\cdot) \text{ is monotonically decreasing} \end{cases}$$

Therefore:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$$

$$Y = \frac{1}{X} \implies X = \frac{1}{Y}$$

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{y}\right)^{a-1} \exp(-\frac{b}{y}) \cdot \left| -\frac{1}{y^2} \right| \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{y}\right)^{a+1} \exp(-\frac{b}{y}) \end{aligned}$$

$$Z = \sqrt{\frac{1}{X}} \implies X = \frac{1}{Z^2}$$

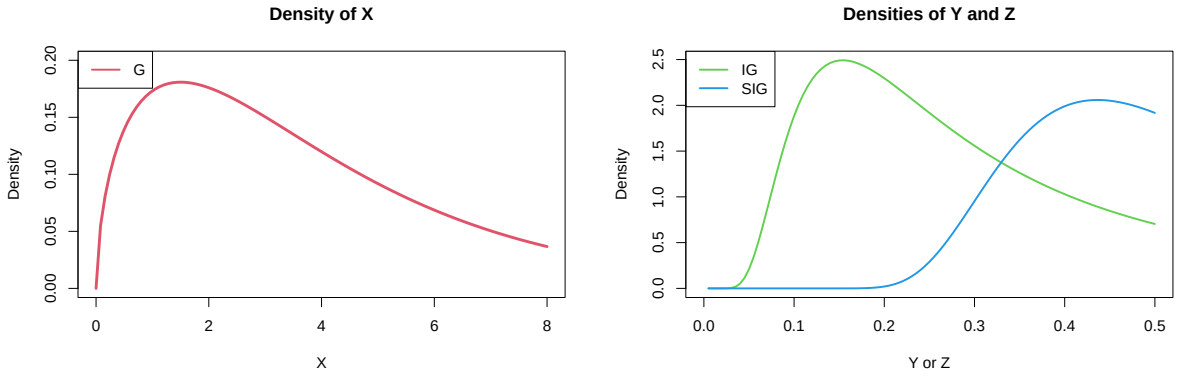
$$\begin{aligned} f_Z(z) &= f_X(g^{-1}(z)) \cdot \left| \frac{d}{dz} g^{-1}(z) \right| \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{z^2}\right)^{a-1} \exp(-\frac{b}{z^2}) \cdot \left| -\frac{2}{z^3} \right| \\ &= \frac{b^a}{\Gamma(a)} 2 \left(\frac{1}{z}\right)^{2a+1} \exp(-\frac{b}{z^2}) \end{aligned}$$

```r
## Define inverse-gamma distribution function
dinvgamma <- function(x, a, b) {
  return(
    (b^a)/gamma(a) * (1/x)^(a+1) * exp(-b/x)
  )
}

## Define square root inverse-gamma distribution function
dsqrtinvgamma <- function(x, a, b) {
  return(
    2 * (b^a)/gamma(a) * (1/x)^(2*a+1) * exp(-b/x^2)
  )
}
```

```r
a <- 1.6
b <- 0.4
```

```r
curve(dgamma(x, shape=a, rate=b), xlim=c(0, 8), ylim=c(0, 0.2), col=2, lwd=3,
      xlab="X", ylab="Density", main="Density of X")
legend("topleft", legend="G", col=2, lwd=2)

curve(dinvgamma(x, a, b), xlim=c(0, 0.5), col=3, lwd=2,
```

```
     main="Densities of Y and Z", y="Density", xlab="Y or Z")
curve(dsqrtinvgamma(x, a, b), add=TRUE, col=4, lwd=2)
legend("topleft", legend=c("IG", "SIG"), col=c(3, 4), lwd=2)
```



By overlaying the densities of Y and Z, we see that the shape of densities is nearly flat when Y and Z are very close to 0, which implies that it is very unlikely that a random draw from Inverse Gamma distribution or from Square root Inverse Gamma distribution would be very close to 0. Recall that a Gamma prior is used for the precision $\frac{1}{\sigma^2}$, an Inverse Gamma prior is used for the variance σ^2, and a Square root Inverse Gamma prior is used for the standard deviation σ. It also implies that the prior about the variance (or the standard deviation) **cannot** be infinitesimally small (i.e. very close to 0).

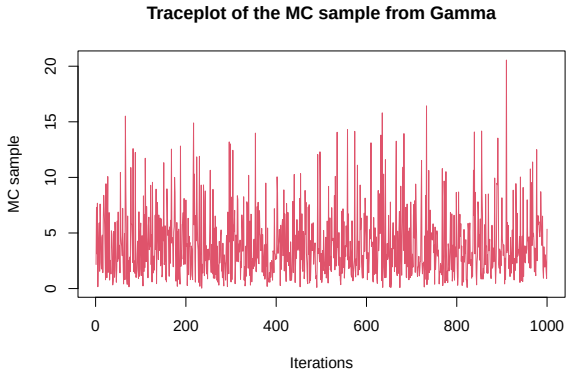
Exercise 6 (Monte Carlo: transformations of random variables)

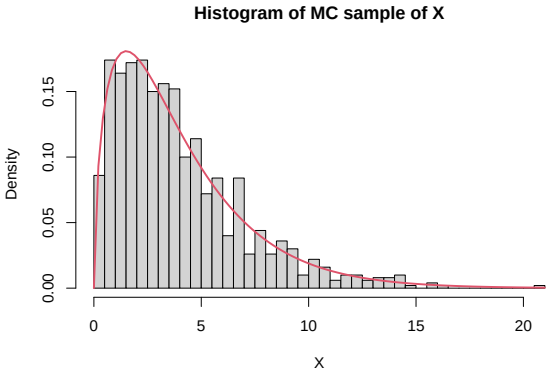
```
## Set seed for reproducible results
set.seed(44566)

## Parameters for Gamma
a <- 1.6 # shape
b <- 0.4 # rate (inverse of scale)

## MC sample size
M <- 1000
```

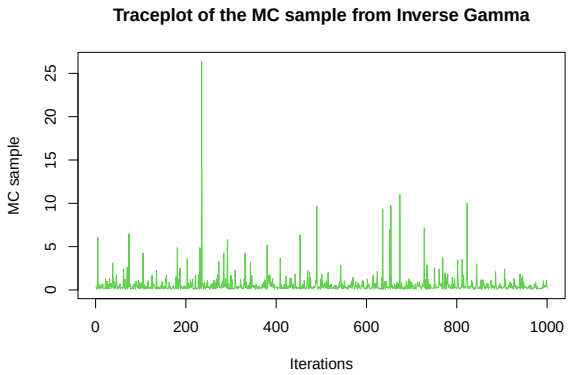
```
## Generate a MC sample of size 1000 from Gamma
mc.G <- rgamma(M, shape=a, rate=b)

## Generate a MC sample of size 1000 from Inverse Gamma
mc.IG <- 1 / mc.G

## Generate a MC sample of size 1000 from Square root Inverse Gamma
mc.SIG <- sqrt(1/mc.G)
```

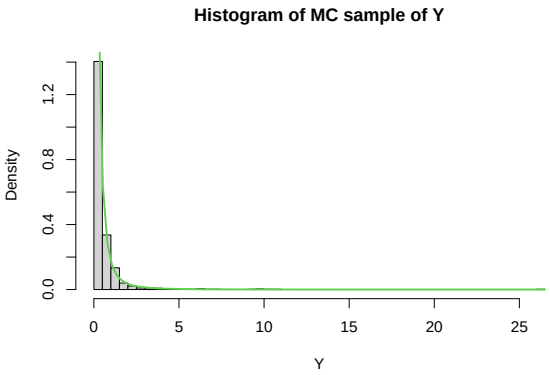
```
plot(1:M, mc.G, type="l", col=2, xlab="Iterations", ylab="MC sample",
     main="Traceplot of the MC sample from Gamma")
hist(mc.G, breaks=50, freq=FALSE, xlab="X", main="Histogram of MC sample of X")
curve(dgamma(x, a, b), add=TRUE, col=2, lwd=2)

plot(1:M, mc.IG, type="l", col=3, xlab="Iterations", ylab="MC sample",
     main="Traceplot of the MC sample from Inverse Gamma")
hist(mc.IG, breaks=50, freq=FALSE, xlab="Y", main="Histogram of MC sample of Y")
```

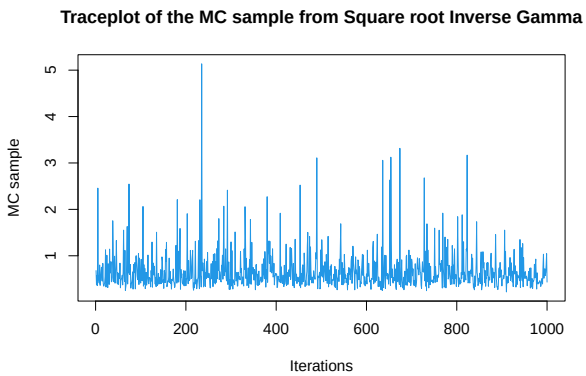
```r
curve(dinvgamma(x, a, b), add=TRUE, col=3, lwd=2)

plot(1:M, mc.SIG, type="l", col=4, xlab="Iterations", ylab="MC sample",
     main="Traceplot of the MC sample from Square root Inverse Gamma")
hist(mc.SIG, breaks=50, freq=FALSE, ylim=c(0, 2), xlab="Z",
     main="Histogram of MC sample of Z")
curve(dsqrtinvgamma(x, a, b), add=TRUE, col=4, lwd=2)
```

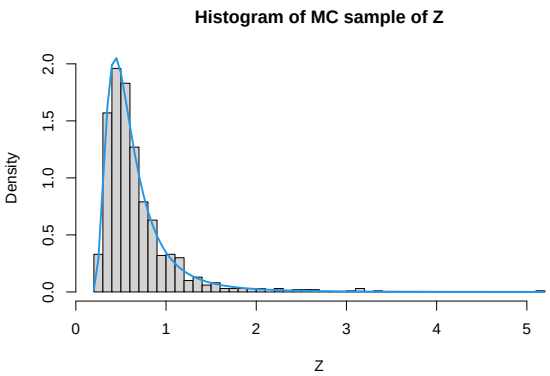













```r
## Gamma
meanG <- mean(mc.G)
medG <- median(mc.G)

## Inverse Gamma
meanIG <- mean(mc.IG)
medIG <- median(mc.IG)

## Square root Inverse Gamma
```

```r
meanSIG <- mean(mc.SIG)
medSIG <- median(mc.SIG)
```

```r
df <- data.frame(
  c(meanG, meanIG, meanSIG),
  c(medG, medIG, medSIG)
)
colnames(df) <- c("Sample Mean", "Sample Median")
rownames(df) <- c("G", "IG", "SIG")

knitr::kable(df, caption="Summary statistics", align="c")
```

Table 2: Summary statistics

	Sample Mean	Sample Median
G	3.9667004	3.2625686
IG	0.6143637	0.3065072
SIG	0.6672328	0.5536309

We know that the random variables X, Y, and Z have the following relation:

$$Y = \frac{1}{X} \implies X = \frac{1}{Y}$$
$$Z = \sqrt{\frac{1}{X}} \implies X = \frac{1}{Z^2}$$

Let us see if such a relation holds for sample medians:

```r
cat(sprintf(
  "Median of X: %.4f\nOne over median of Y: %.4f\nOne over squared median of Z: %.4f",
  medG, 1/medIG, 1/medSIG^2
))
```

```
## Median of X: 3.2626
## One over median of Y: 3.2626
## One over squared median of Z: 3.2626
```

We see that the transformation of random variables does not change the relation of medians for X, Y, and Z.

Let us further check if such a relation holds for sample means:

```r
cat(sprintf(
  "Mean of X: %.4f\nOne over mean of Y: %.4f\nOne over squared mean of Z: %.4f",
  meanG, 1/meanIG, 1/meanSIG^2
))
```

```
## Mean of X: 3.9667
## One over mean of Y: 1.6277
## One over squared mean of Z: 2.2462
```

We obtain different values for X, Y, and Z if we do back-transformation operations on their means.

Reasoning:

- Inverse Gamma Y: we conduct inverse transformation on the original data. The inverse transformation only reverses the ranking of the data but does not change the 50% quantile (median) of the data, which allows us to back transform the sample median of Y to obtain the same sample median as in the original data X.
- Square root Inverse Gamma Z: we simply do square root transformation on the top of the inverse transformation. Square root transformation is a monotonic transformation and does not change the ranking of the data, and Y and Z have exactly the same ranking (quantile) for the data. Hence, the median of Z can also be transformed back to the median of X.
- In the Gamma (shape $= a$, rate $= b$) distribution, the mean and median do not coincide. Since inverse transformation and square root transformation are nonlinear transformations, the positions of the sample means are different in X, Y, and Z. Thus the relation does not apply to sample means.