Module 4 Peer Review Assignment

Problem 1

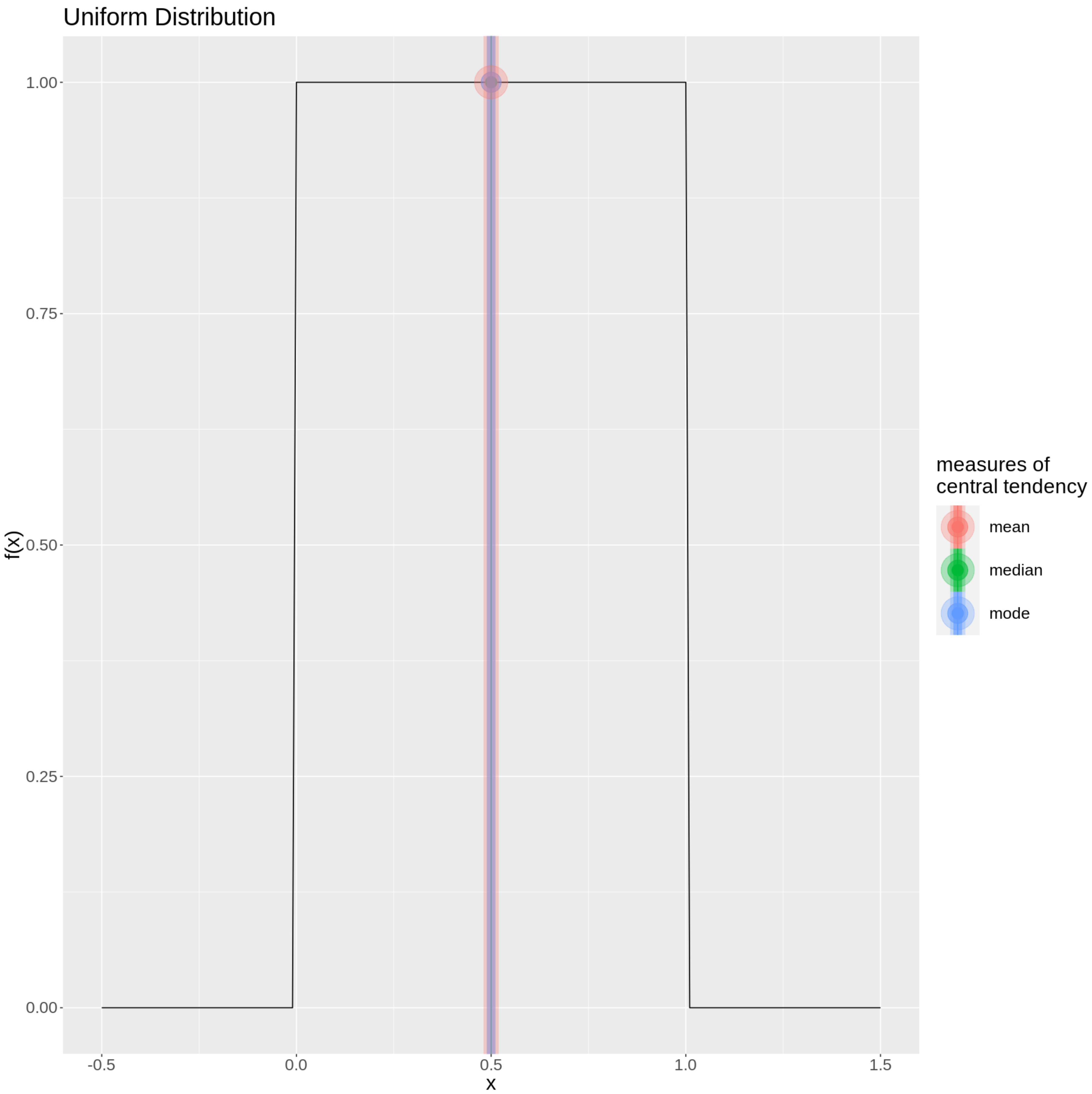
A continuous random variable with cumulative distribution function F has the median value m such that $F(m) = 0.5$. That is, a random variable is just as likely to be larger than its median as it is to be smaller. A continuous random variable with density f has the mode value x for which $f(x)$ attains its maximum. For each of the following three random variables, (i) state the density function, (ii) compute the median, mode and mean for the random variable, and (iii) Provide at least one graph for the density function using values of the parameter(s) that you select. Indicate the median, mode, and mean values on your graph. (The purpose of this problem is to see the relative locations of the median, mode, and mean for the different random variables).

a) W which is uniformly distributed over the interval $[a, b]$, for some value $a, b \in \mathbb{R}$.

In [118]:

```r
# Your Code Here

# from the results explained in the next section

options(repr.plot.width = 15, repr.plot.height = 15)

f <- function(x) {
    return (ifelse((x >= a) && (x <= b), 1 / (b - a), 0))
}
f <- Vectorize(f)

plot_graph <- function(x, f, title) {
    ggplot() + geom_line(aes(x, f(x))) +
               geom_point(aes(median, f(median), col='median'), size=5) +
               geom_vline(aes(xintercept=median, col='median')) +
               geom_point(aes(mode, f(mode), col='mode'), size=9, alpha=0.6) +
               geom_vline(aes(xintercept=mode, col='mode'), lwd=4, alpha=0.6) +
               geom_point(aes(mean, f(mean), col='mean'), size=15, alpha=0.3) +
               geom_vline(aes(xintercept=mean, col='mean'), lwd=7, alpha=0.3) +
               labs(col='measures of\ncentral tendency') +
               #scale_shape(solid = FALSE) +
               theme(text = element_text(size=20)) + ggtitle(title)
}

# parameters
a <- 0
b <- 1

median <- (a+b)/2
mode <- (a+b)/2 # can be any point in between [a,b], just one is shown in the figure
mean <- (a+b)/2

print(c(median, mode, mean))

library(ggplot2)

x <- seq(a-0.5,b+0.5,0.01)
plot_graph(x, f, 'Uniform Distribution')
```

```
[1] 0.5 0.5 0.5
```



(i) $f(x) = \left\{ \begin{array}{ll} \frac{1}{b-a}, & \text{for } a \leq x \leq b \\ 0, & \text{otherwise} \end{array} \right\}$

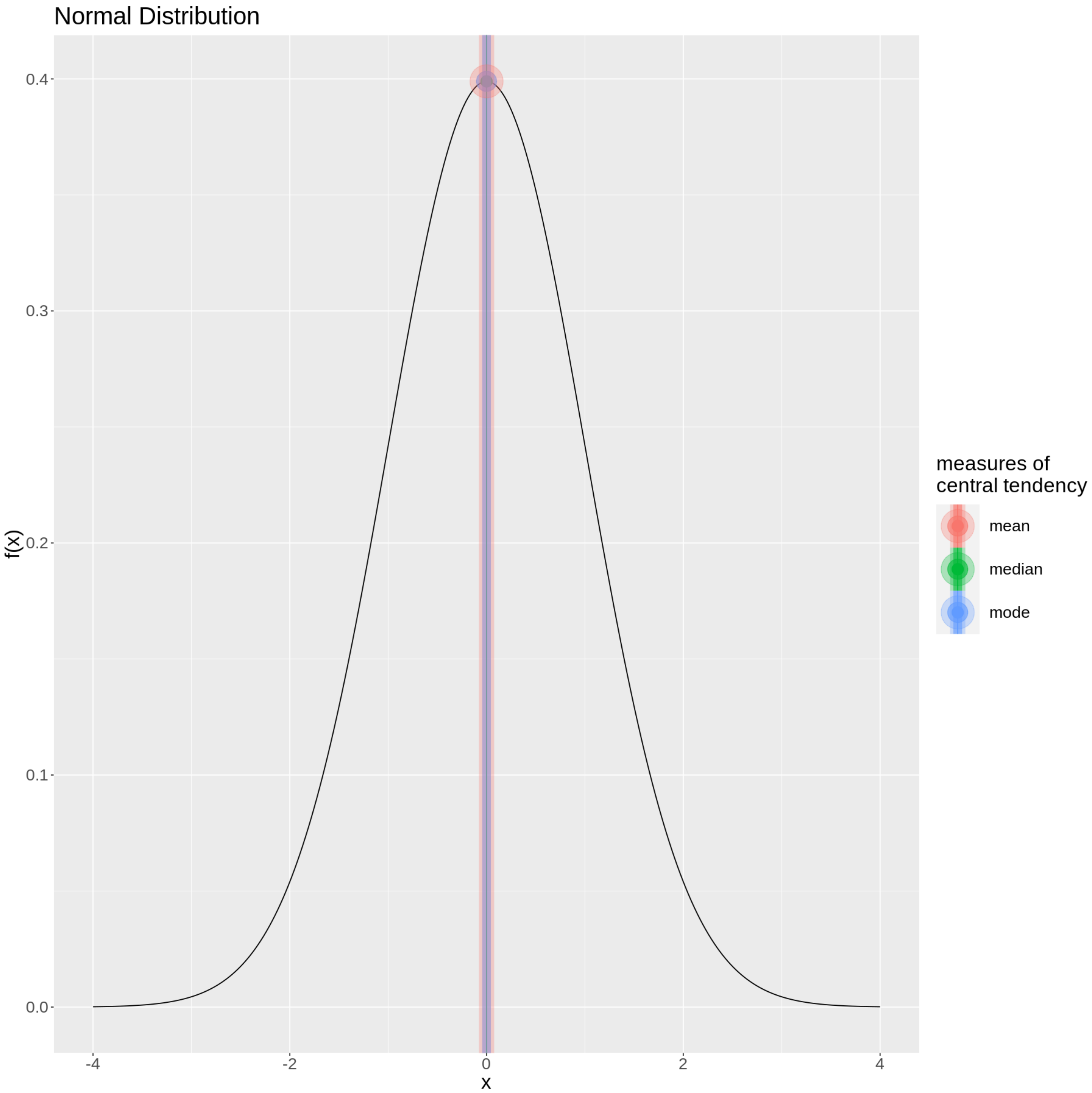
(ii)

$$\begin{array}{ll} Median: & F(x) = \int\limits_a^x \frac{1}{b-a} dx = \left[\frac{x}{b-a} \right]_a^x = \frac{x-a}{b-a}. \text{ Now } F(x_m) = \frac{x_m - a}{b-a} = 0.5 \implies x_m = \frac{a+b}{2} \\ \text{Mode:} & M = \underset{a \leq x \leq b}{argmax} \frac{1}{b-a} = \{x : a \leq x \leq b\}, i.e., \text{ any point in between } a \text{ and } b, \text{ since } f(x) \text{ is a constant function} \\ \text{Mean:} & \mu = \int\limits_a b \frac{x}{b-a} dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{a+b}{2} \end{array}$$

(iii) Figure above

b) X which is normal with parameters μ and σ^2, for some value $\mu, \sigma^2 \in \mathbb{R}$.

In [104]:
```
# Your Code Here

# from the results explained in the next section

f <- function(x) {
    return (exp(-(x-μ)^2/(2*σ^2))/sqrt(2*pi*σ^2))
}

# parameters
μ  <- 0
σ <- 1

median <- μ
mode <- μ # can be any point in between [a,b], just one is shown in the figure
mean <- μ

library(ggplot2)

x <- seq(-4,4,0.01)
plot_graph(x, f, 'Normal Distribution')
```



(i) $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$

(ii) $X \sim \mathcal{N}(\mu, \sigma^2)$

$$
\begin{aligned}
&Median: \quad x_m = \mu, \text{ since we can show that } \int_{-\infty}^{\mu} f(x)dx = \int_{\mu}^{\infty} f(x)dx = 0.5 \\
&\int_{\mu}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \quad \left(\text{substitute } z = \frac{x-\mu}{\sigma},\ z \sim \mathcal{N}(0,1) \text{ pdf is symmetric about } z=0, \text{ hence the integral evaluates to } \tfrac{1}{2}\right) \\
&= \frac{1}{2\sqrt{\pi}} \int_{0}^{\infty} t^{1/2-1} e^{-t} dt \quad \left(\text{substitute } t = \frac{z^2}{2}, \text{ s.t. } dz = \frac{dt}{\sqrt{2t}}\right) \\
&= \frac{1}{2\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right) \\
&= \frac{1}{2\sqrt{\pi}} \sqrt{\pi} \\
&= \frac{1}{2}
\end{aligned}
$$

$$
\begin{aligned}
&\text{Mode:} \quad M = \mu, \text{ since at an extremum point, we have the following:} \\
&\frac{df(x)}{dx} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} . \frac{-2(x-\mu)}{2\sigma^2} = 0 \\
&\implies x - \mu = 0 \\
&\implies x = \mu \quad \text{(we can show the 2nd derivative is -ive, confirming that it's a maximum, so it's mode)}
\end{aligned}
$$

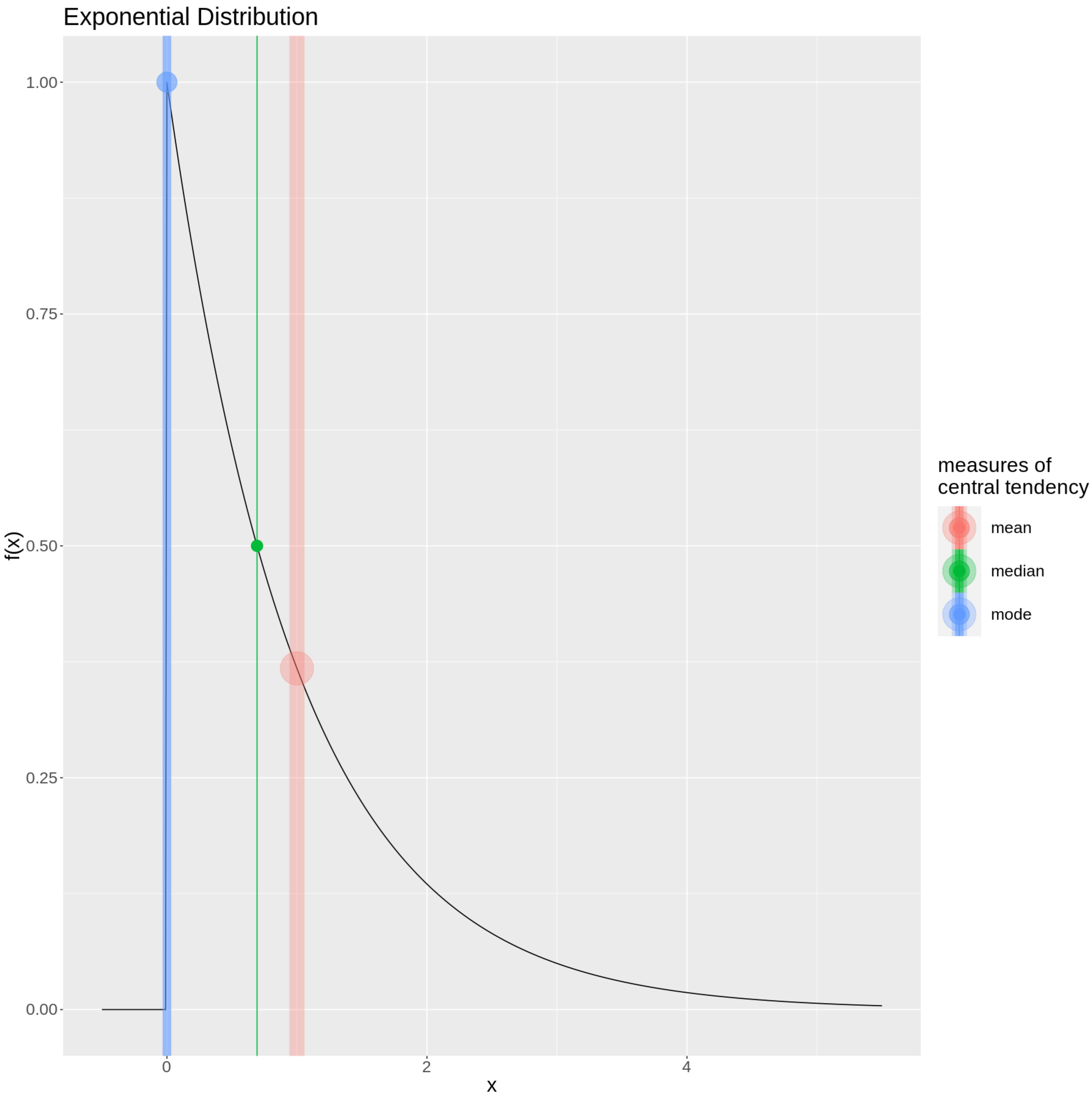
$$
\begin{aligned}
&\text{Mean:} \quad \mu = \int_{-\infty}^{\infty} x f(x) dx, \text{ since we have} \\
&E[X] = \int_{\mu}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} x e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} (\mu + \sigma z) e^{\frac{-z^2}{2}} dz \quad \left(\text{substitute } z = \frac{x-\mu}{\sigma},\ z \sim \mathcal{N}(0,1)\right) \\
&= \mu . \left(\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz\right) + \frac{\sigma}{\sqrt{2\pi}} \left(\int_{0}^{\infty} z e^{-\frac{z^2}{2}} dz\right) \\
&= \mu .1 + \frac{\sigma}{\sqrt{2\pi}} .0 \quad (1^{st} \text{ integral is a pdf: evaulates to 1, } 2^{nd} \text{ one is an odd function: evaluates to 0}) \\
&= \mu
\end{aligned}
$$

(iii) Figure above

c) Y which is exponential with rate $\lambda \in \mathbb{R}$.

In [105]:

```
# Your Code Here

# from the results explained in the next section

f <- function(x) {
    return(ifelse(x >= 0, λ*exp(-λ*x), 0))
}
f <- Vectorize(f)

# parameters
λ  <- 1
median <- log(2) / λ
mode <- 0 # can be any point in between [a,b], just one is shown in the figure
mean <- 1 / λ

library(ggplot2)

x <- seq(0-0.5,5+0.5,0.01)
plot_graph(x, f, 'Exponential Distribution')
```



(i) $f(x) = \begin{Bmatrix} \lambda e^{-\lambda x}, & \text{for } 0 \leq x < \infty \\ 0, & \text{otherwise} \end{Bmatrix}$

(ii)

$$\begin{array}{ll} Median: & F(x) = \int\limits_0^x \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^x = 1 - e^{-\lambda x}.\ \text{Now } F(x_m) = 1 - e^{-\lambda x} = 0.5 \implies x_m = \frac{\ln 2}{\lambda} \\ \text{Mode:} & M = \underset{0 \leq x < \infty}{argmax} \lambda e^{-\lambda x} = 0, \quad \text{since } f(x) \text{ is a decreasing function in } [0, \infty) \text{, since } f^{'}(x) = -\lambda^2 e^{-\lambda x\$} < 0 \text{ and } f(0) = \lambda \\ \text{Mean:} & \mu = E[X] = \int\limits_0^\infty x.\lambda e^{-\lambda x} dx = \left[-xe^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda}\right]_0^\infty \quad \text{integrating by parts} \\ & \implies \mu = \frac{1}{\lambda} \end{array}$$

(iii) Figure above

Problem 2

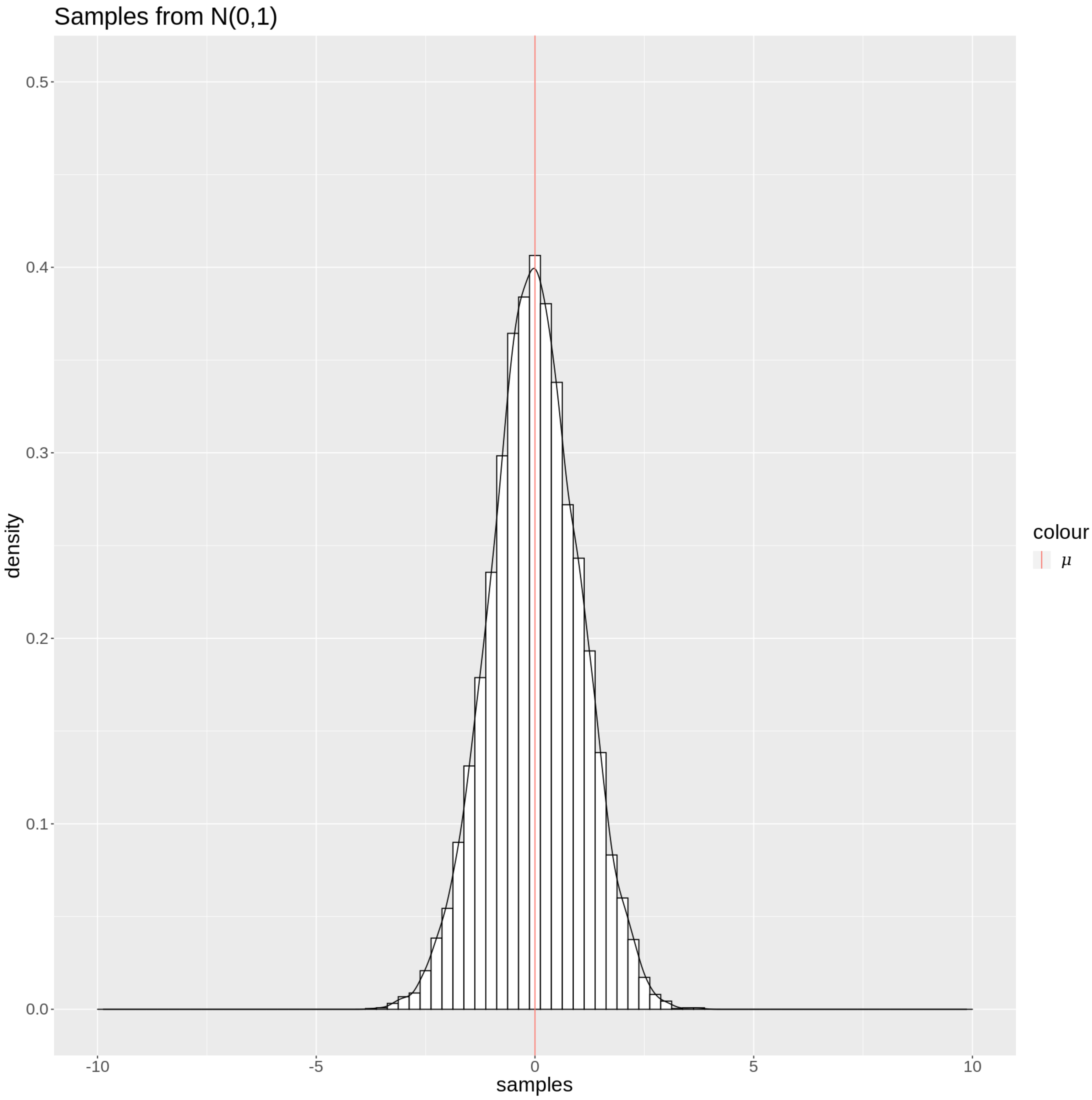
For this problem, we're going to visualize what's happening when we go between different normal distributions.

Part A)

Draw at least 10000 samples from the standard normal distribution $N(0, 1)$ and store the results. Make a density histogram of these samples. Set the x-limits for your plot to $[-10, 10]$ and your y-limits to $[0, 0.5]$ so we can compare with the plots we'll generate in **Parts B-D**.

In [113]:
```
# Your Code Here
set.seed(1)
samples = rnorm(10000, 0, 1)
ggplot() + geom_histogram(aes(samples, y = ..density..), colour = 1, fill = "white", binwidth=0.25) +
            geom_density(aes(samples)) + xlim(-10,10) + ylim(0,0.5) +
            geom_vline(aes(xintercept=0, color='μ', linetype="dotted")) +
            ggtitle('Samples from N(0,1)') +
            theme(text = element_text(size=20)) +
            guides(linetype="none")
```

```
Warning message:
“Removed 2 rows containing missing values (geom_bar).”
```

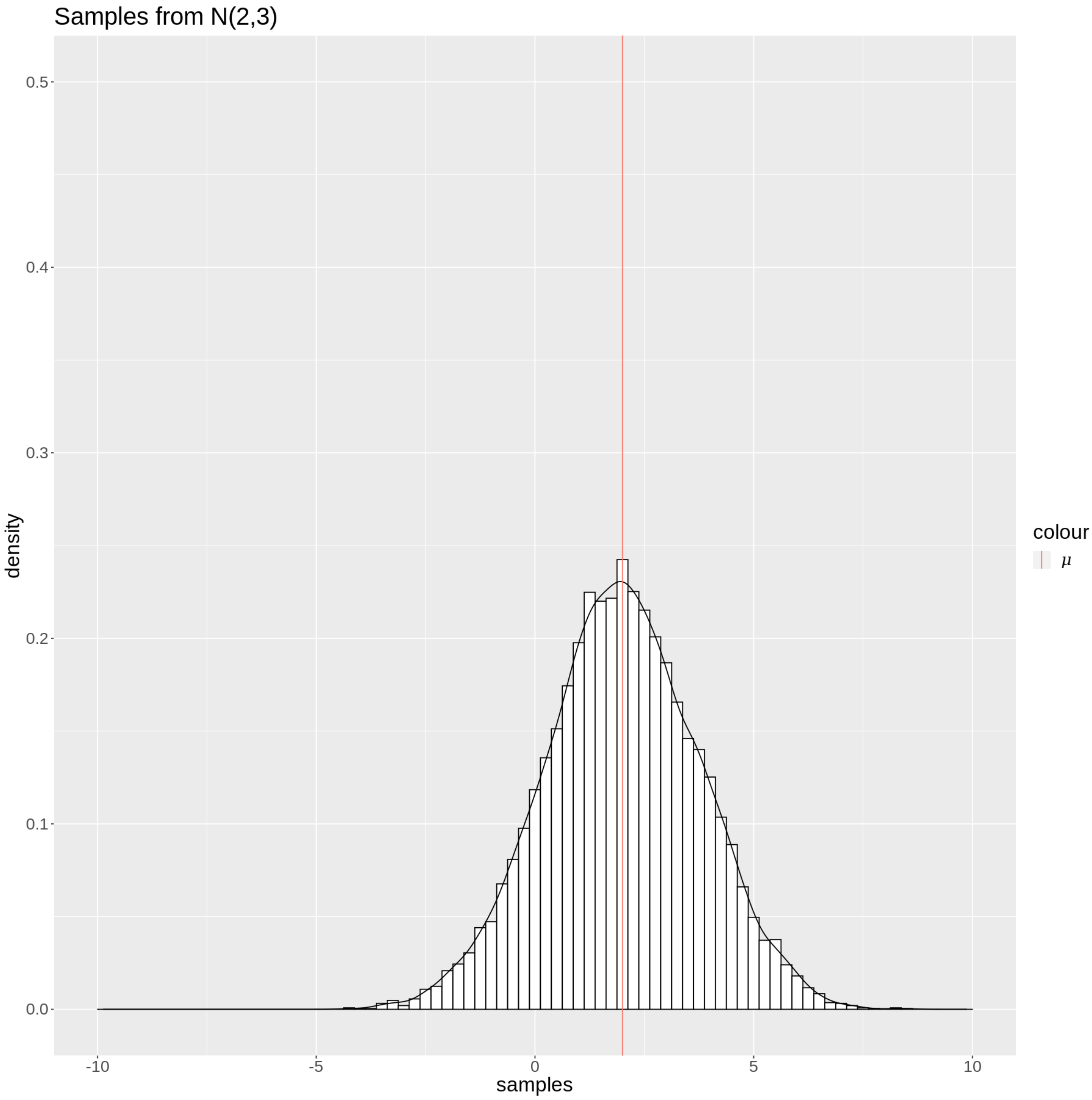


Part b) Now generate 10000 samples from a $N(2, 3)$ distribution and plot a histogram of the results, with the same x-limits and y-limits. Does the histogram make sense based on the changes to parameters?

Note: Be careful with the parameters for `rnorm`. It may help to check the documentation.

In [114]:
```
# Your Code Here
set.seed(1)
samples = rnorm(10000, 2, sqrt(3))
ggplot() + geom_histogram(aes(samples, y = ..density..), colour = 1, fill = "white", binwidth=0.25) +
            geom_density(aes(samples)) + xlim(-10,10) + ylim(0,0.5) +
            geom_vline(aes(xintercept=2, color='μ', linetype="dotted")) +
            ggtitle('Samples from N(2,3)') +
            theme(text = element_text(size=20)) +
            guides(linetype="none")
```

```
Warning message:
“Removed 2 rows containing missing values (geom_bar).”
```



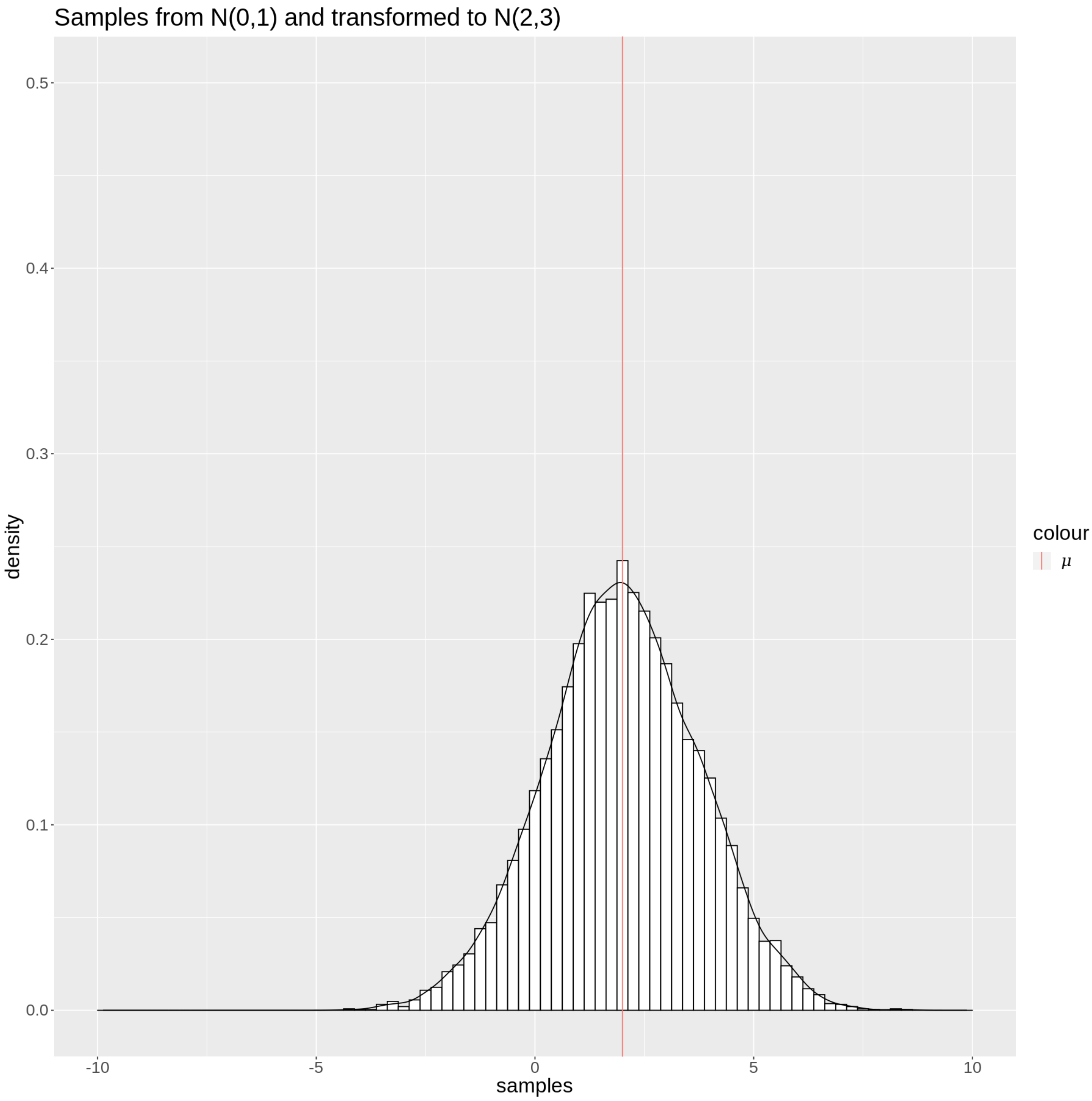
Yes, as can be seen from the above figures, the density now has its bick at $x \approx 2$, which is expected, since the population mean μ has shifted to 2 now. Also, the variance of the distribution has been increased to 3, so it's more spread out along the x axis (flatter-tailed and shorter height this time, as opposed to taller and skinny last time with lower varaince).

Part c)

Suppose we are only able to sample from the standard normal distribution $N(0, 1)$. Could we take those samples and perform a simple transformation so that they're samples from $N(2, 3)$? Try this, and plot another histogram of the transformed data, again with the same axes. Does your histogram based of the transformed data look like the histogram from **Part B**?

In [115]:
```
# Your Code Here
# x = μ + zσ
μ = 2
σ = sqrt(3)
set.seed(1)
z = rnorm(10000, 0, 1)
x = μ + z*σ
ggplot() + geom_histogram(aes(samples, y = ..density..), colour = 1, fill = "white", binwidth=0.25) +
            geom_density(aes(x)) + xlim(-10,10) + ylim(0,0.5) +
            geom_vline(aes(xintercept=2, color='μ', linetype="dotted")) +
            ggtitle('Samples from N(0,1) and transformed to N(2,3)') +
            theme(text = element_text(size=20)) +
            guides(linetype="none")
```

```
Warning message:
“Removed 2 rows containing missing values (geom_bar).”
```



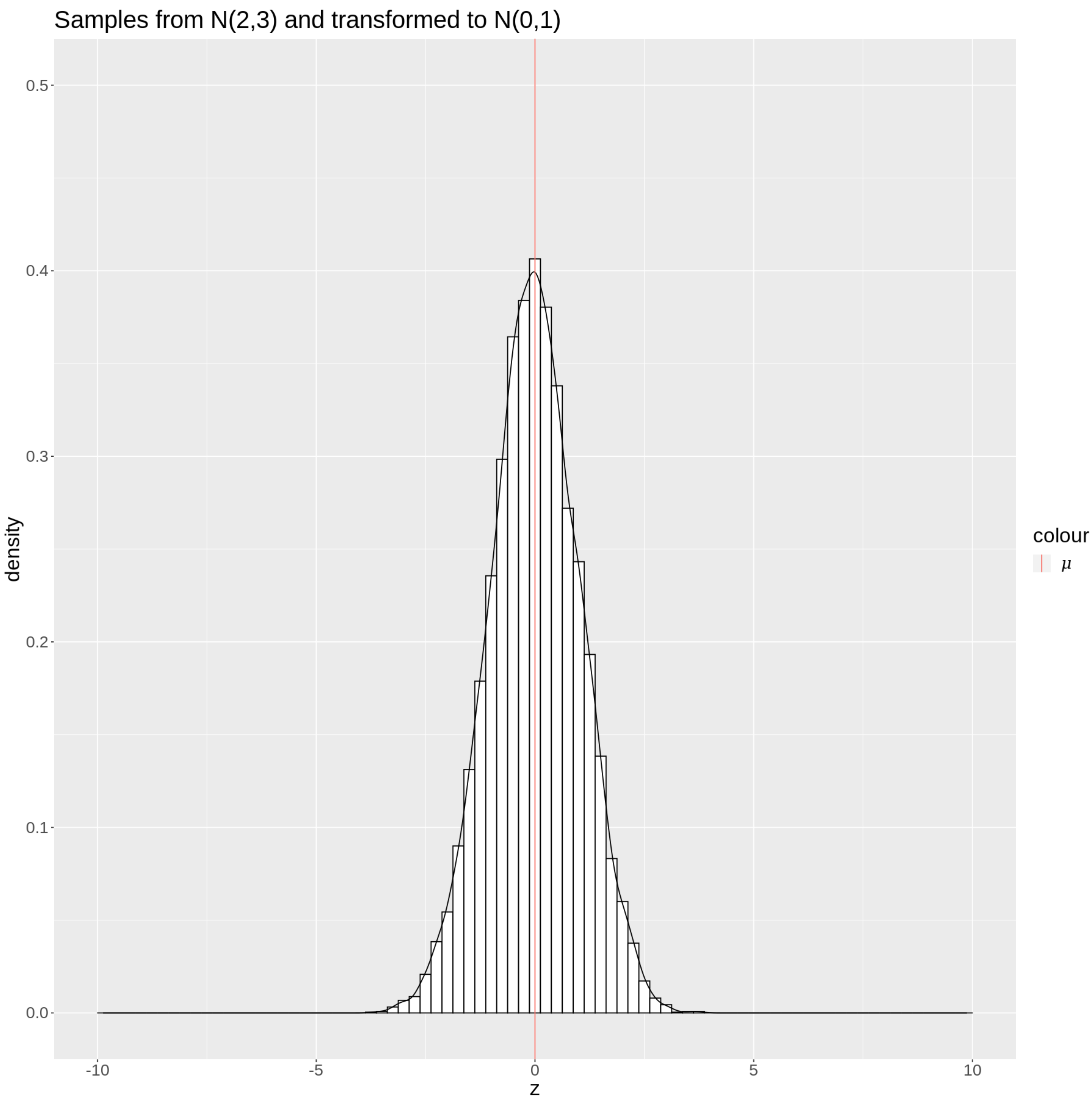
Yes, we can. Using the transformation $x = \mu + z\sigma$ we can convert standard normal r.v. $Z \sim \mathcal{N}(0, 1)$ to a normal r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$, as shown in the figure above.

Part d)

But can you go back the other way? Take the $N(2, 3)$ samples from **Part B** and transform them into samples from $N(0, 1)$? Try a few transformations and make a density histogram of your transformed data. Does it look like the plot of $N(0, 1)$ data from **Part A**?

In [117]:
```
# Your Code Here
# Your Code Here
# z = (x - μ) / σ
μ = 2
σ = sqrt(3)
set.seed(1)
x = rnorm(10000, μ, σ)
z = (x - μ) / σ
ggplot() + geom_histogram(aes(z, y = ..density..), colour = 1, fill = "white", binwidth=0.25) +
            geom_density(aes(z)) + xlim(-10,10) + ylim(0,0.5) +
            geom_vline(aes(xintercept=0, color='μ', linetype="dotted")) +
            ggtitle('Samples from N(2,3) and transformed to N(0,1)') +
            theme(text = element_text(size=20)) +
            guides(linetype="none")
```

```
Warning message:
“Removed 2 rows containing missing values (geom_bar).”
```



Yes, we can. Using the transformation $z = \frac{x-\mu}{\sigma}$ we can convert normal r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ to a standard normal r.v. $Z\sim \mathcal{N}(0,1), as shown in the figure above.