Continuous Probability Distributions and Sampling Distributions

Blake Pappas

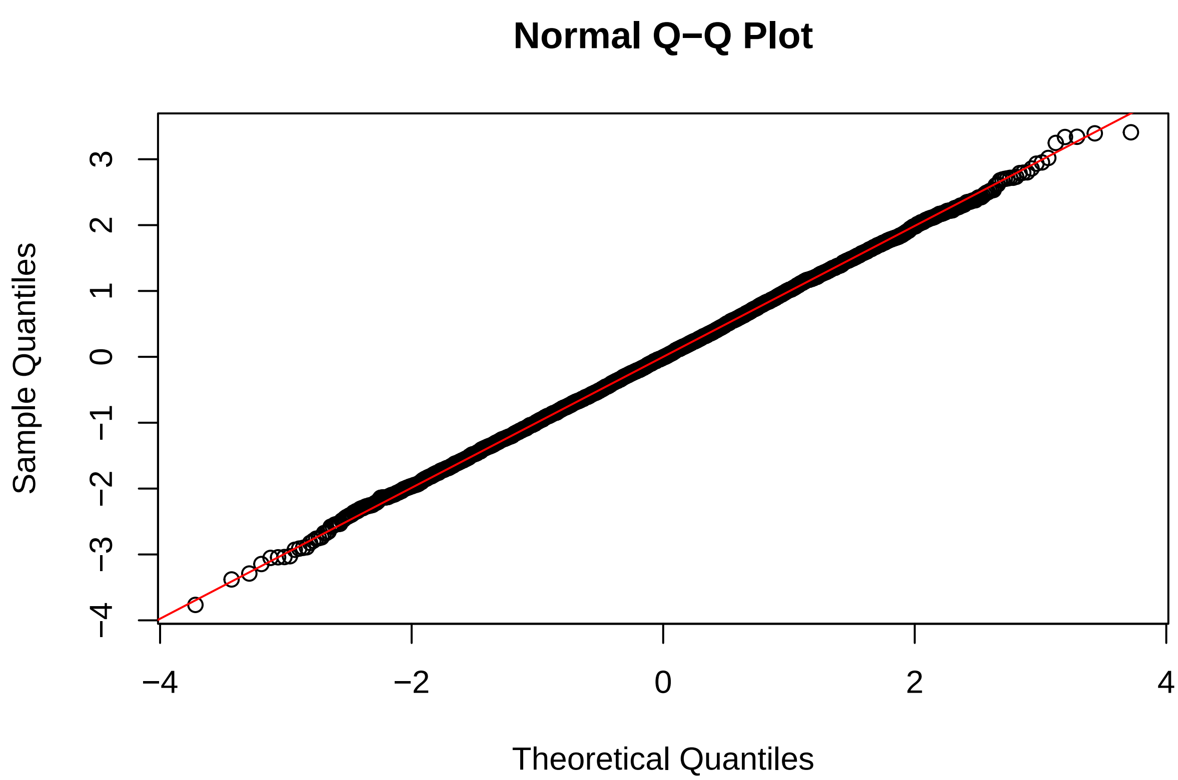
10/5/2021

Normal Quantile Plots

Very often in statistics we assume that observed data values are realizations of random variables from a normal distribution. A normal quantile plot is used to check whether that assumption is reasonable.

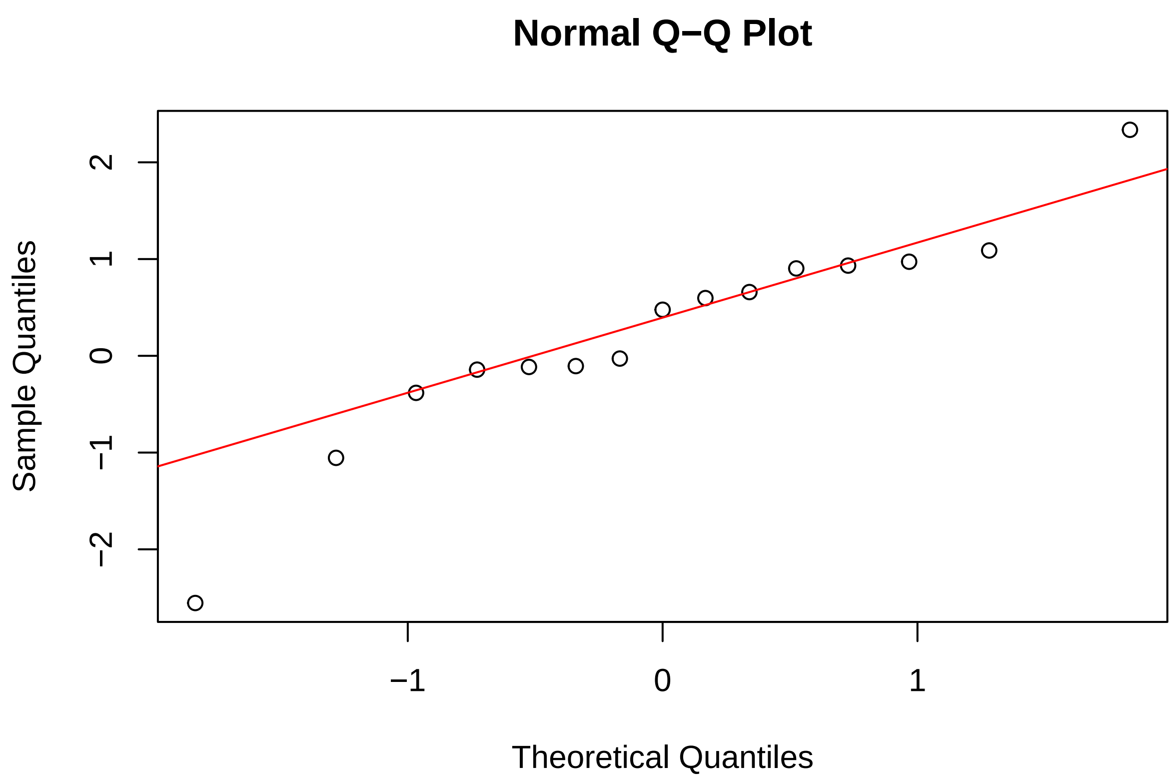
If data are from a perfect normal distribution, the normal quantile plot will show points along a perfect 45-degree, uphill line. The code below will generate random normal variables and display the quantile plot.

```
x1 <- rnorm(5e3)
qqnorm(x1)
qqline(x1, col = 'red')   # Adds the line that perfectly normal data would follow
```



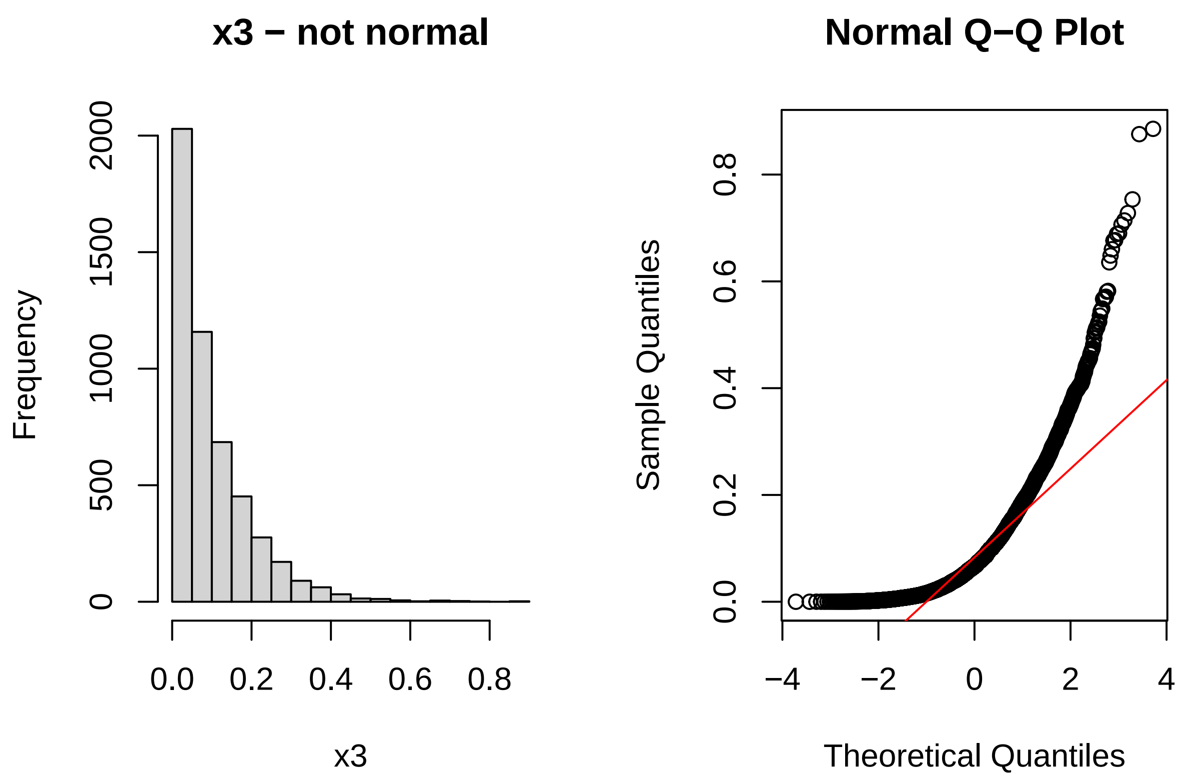
With a smaller sample, the plot will not look as perfect, but the general linear pattern will be the same.

```
x2 <- rnorm(15)
qqnorm(x2)
qqline(x2, col = 'red')
```

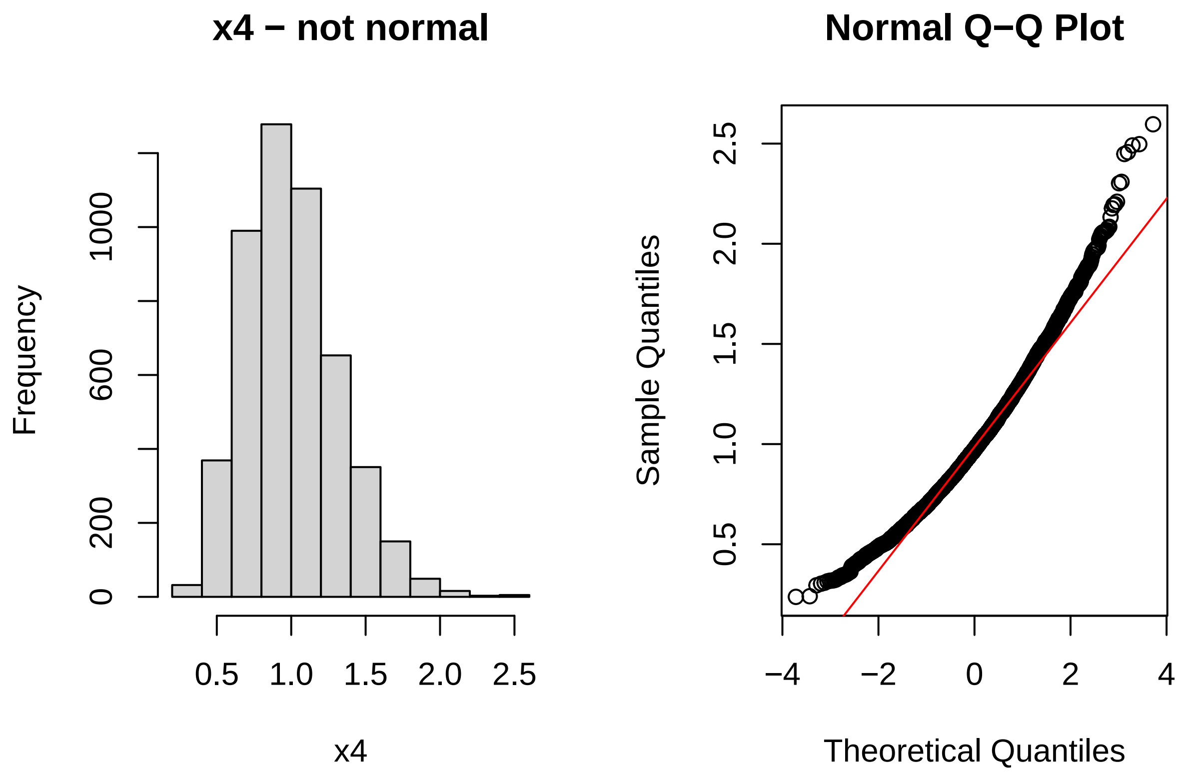


If data are not normal, you may see an S-shaped pattern in the quantile plot, or curvature in the lower or upper tail. The chunks below generate data from some skewed distributions and show their quantile plots.

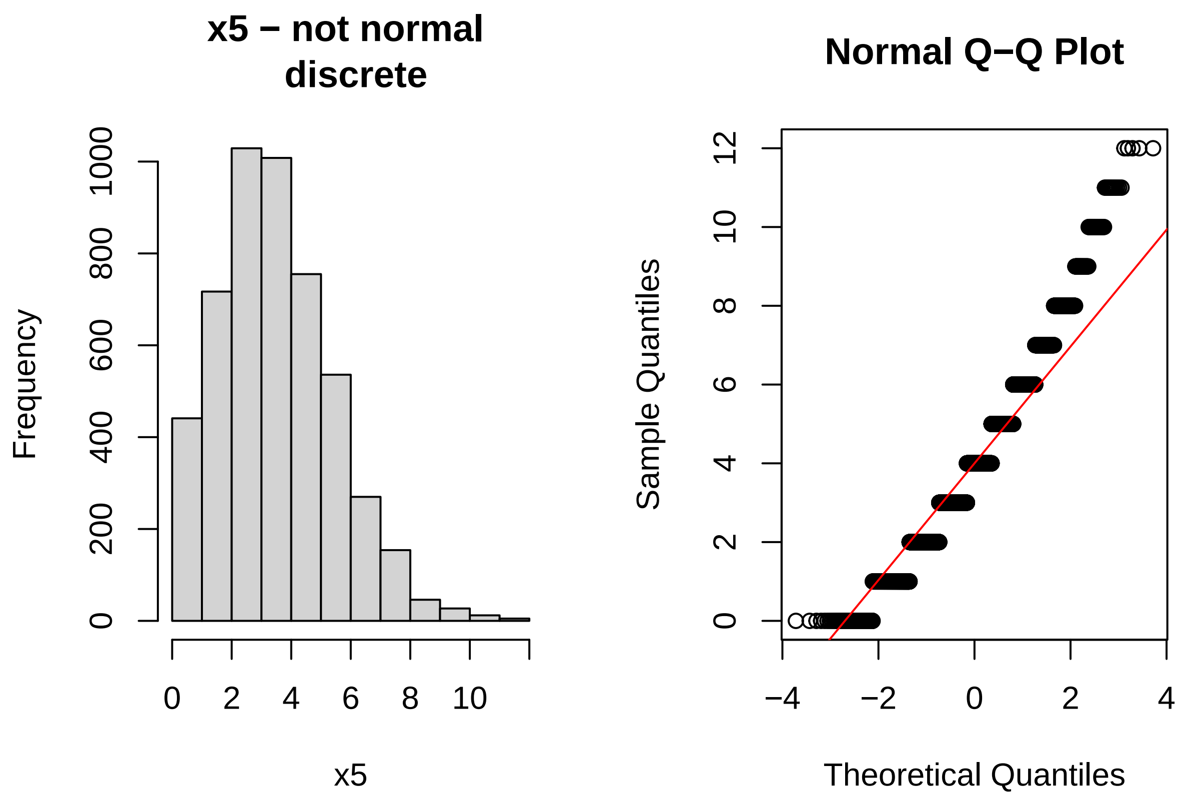
```
par(mfrow = c(1, 2))
x3 <- rgamma(5e3, 1, 10)  # Samples from a very skewed gamma distribution
hist(x3, main = 'x3 - not normal')
qqnorm(x3)
qqline(x3, col = 'red')
```




```
par(mfrow = c(1, 2))
x4 <- rgamma(5e3, 10, 10) # Samples from a nearly symmetric gamma distribution
hist(x4, main = 'x4 - not normal')
qqnorm(x4)
qqline(x4,col = 'red')
```

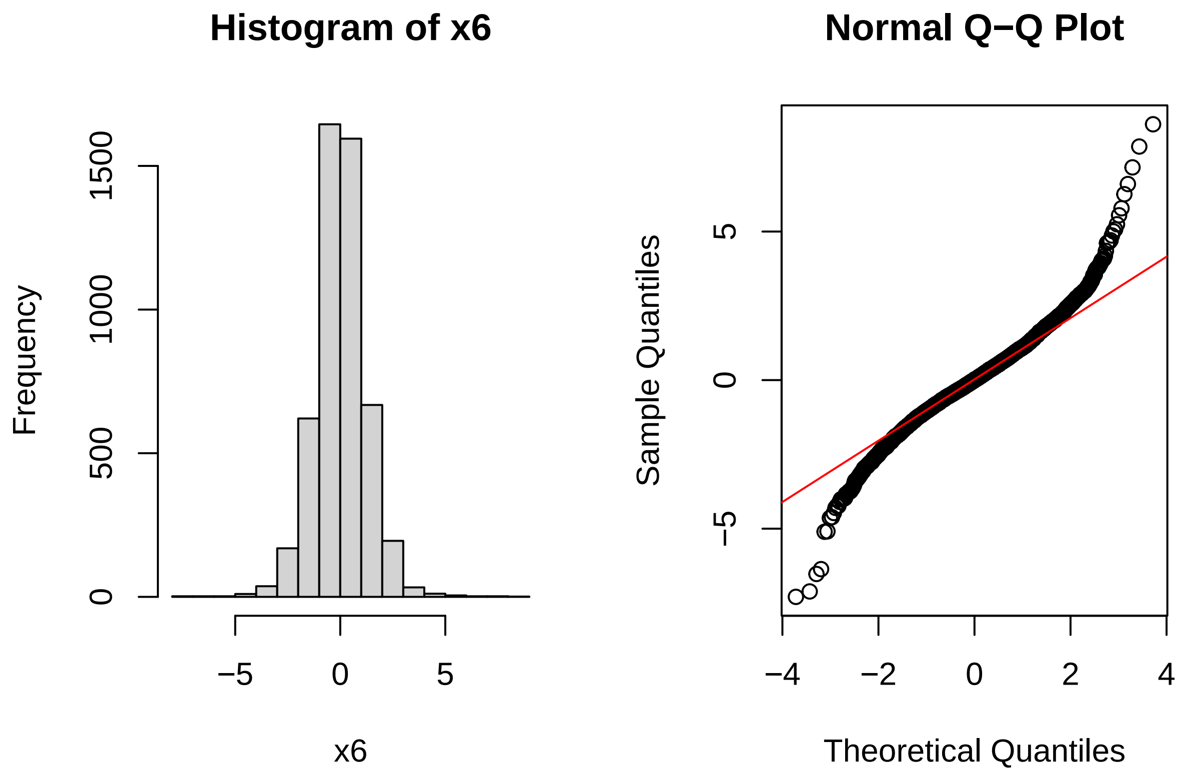



```r
par(mfrow = c(1, 2))
x5 <- rpois(5e3, 4) # Samples from a discrete distribution
hist(x5, main = 'x5 - not normal \n discrete')
qqnorm(x5)
qqline(x5,col = 'red')
```



In some of these examples, the histograms clearly indicate that the shape of the distribution does not resemble a normal curve. Sometimes, however, the quantile plot reveals patterns that are harder to spot in a histogram. Here is a sample from a t-distribution, which is bell-shaped but has fatter tails than a normal curve. The non-normality is much easier to spot in the normal quantile plot.

```
par(mfrow = c(1, 2))
x6 <- rt(5e3, 6)
hist(x6, breaks = 20)
qqnorm(x6)
qqline(x6,col = 'red')
```



Exercises

Exercise 1: Normal Probabilities

Use the functions `pnorm()` and `qnorm()` to find the following probabilities.

a. Find the 77th percentile of a normal distribution with mean 0 and standard deviation of 0.5.

```
qnorm(0.77, 0, 0.5)
```

```
## [1] 0.3694234
```

Answer: The 77th percentile of the distribution is approximately 0.3694.

b. Find the probability that a N(0, 0.5^2) random variable is less than -0.35.

```
pnorm(-0.35, 0, 0.5)
```

```
## [1] 0.2419637
```

Answer: The probability that N is less than -0.35 is approximately 0.2420.

c. Find the probability that a N(10, 3^2) random variable is greater than 17.

```r
pnorm(17, 10, 3, lower.tail = FALSE)
```

```
## [1] 0.009815329
```

Answer: The probability that N is greater than 17 is approximately 0.0100.

d. Find the probability that a N(10, 3^2) random variable is between 9 and 14.

```r
pnorm(14, 10, 3) - pnorm(9, 10, 3)
```

```
## [1] 0.5393474
```

Answer: The probability that N is between 9 and 14 is approximately 0.5393.

Exercise 2: Normal Probabilities

The age of employees at a certain company have a Normal(37, 4.7^2) distribution.

a. Find the probability that a randomly selected employee is older than 40.

```r
pnorm(40, 37, 4.7, lower.tail = FALSE)
```

```
## [1] 0.2616399
```

Answer: The probability that a randomly selected employee is older than 40 is approximately 0.2616.

b. Find the cutoff for the youngest 10% of employees; that is, 10% of employees are younger than what age?

```r
qnorm(0.1, 37, 4.7)
```

```
## [1] 30.97671
```

Answer: 10% of employees are younger than the age of 30.9767.

c. Find the probability that a randomly selected employee is between 33 and 39.

```r
pnorm(39, 37, 4.7) - pnorm(33, 37, 4.7)
```

```
## [1] 0.4674086
```

Answer: The probability that a randomly selected employee is between 33 and 39 is approximately 0.4674.

d. Find the cutoff for the oldest 5% of employees; that is, 5% of employees are older than what age?

```r
qnorm(0.95, 37, 4.7)
```

```
## [1] 44.73081
```

Answer: 5% of employees are older than 44.7308 years.

Exercise 3: Simulation to Approximate Probabilities

Simulation can be used to find approximate probabilities. It can also be used to find approximate distributions. If you take a large number of samples (no less than 500) from a probability distribution, the histogram of the samples will look very similar to the true probability density function (or pmf).

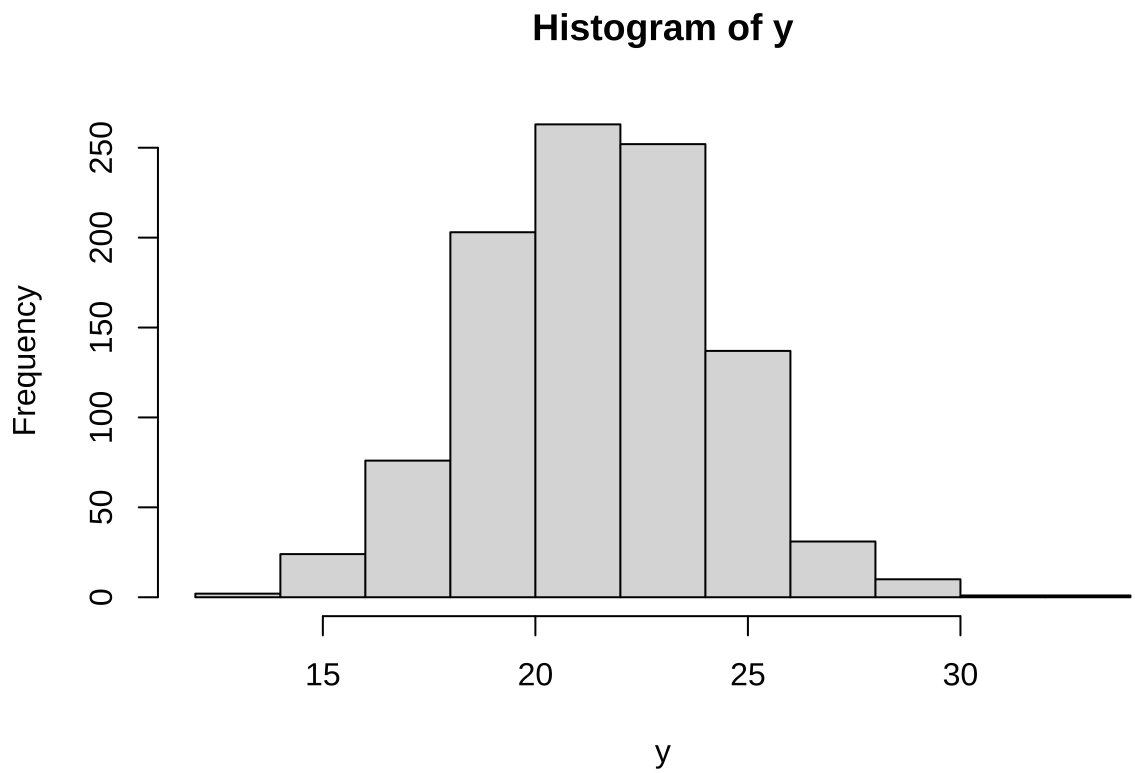
a. The `rnorm()` function can be used to take samples from a normal distribution. Here is an example of using `rnorm` to draw n = 5 samples from a N(-4, 0.8^2) distribution.

```r
y <- rnorm(n = 5, mean = -4, sd = 0.8)
print(y)
```

```
## [1] -4.458742 -3.643861 -3.897232 -3.535779 -4.529340
```

Now use `rnorm` to draw 1000 samples from the N(21.4, 2.8^2) distribution. Make a histogram of the samples. Does the shape of the histogram resemble a normal curve? Then calculate the mean and standard deviation of the samples. Do these numbers come close to the true mean and standard deviation?

```r
y <- rnorm(n = 1000, mean = 21.4, sd = 2.8)
hist(y)
```




```r
mean(y)
```

```
## [1] 21.47085
```

```r
sd(y)
```

```
## [1] 2.78349
```

Answer: Yes, the shape of the histogram does resemble that of a normal curve. Yes, these numbers do come close to the true mean and standard deviation. However, they are not exact. A larger sample size would bring them closer to the true mean and standard deviation.

b. Using the samples from part (a), calculate the proportion of samples that are greater than 25. Compare this to the true probability that a N(21.4, 2.8^2) random variable is greater than 25, which you can find using `pnorm()`.

```r
pnorm(25, mean(y), sd(y), lower.tail = FALSE)
```

```
## [1] 0.1024192
```

```r
pnorm(25, 21.4, 2.8, lower.tail = FALSE)
```

```
## [1] 0.0992714
```

Answer: The sample's probability is less than the true probability. Again, a larger sample size would bring this calculated probability closer to the true probability.

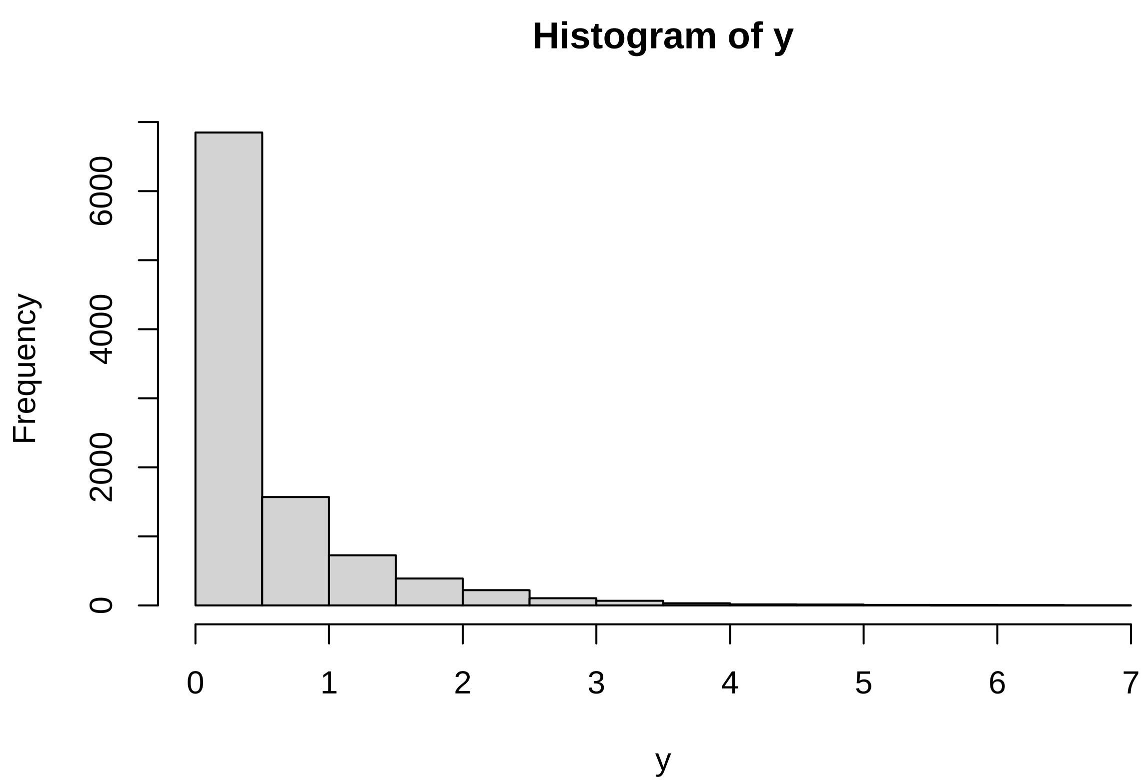
c. The code below gives an example of drawing random samples from a gamma distribution using the `rgamma` function. The code draws n = 5 samples from the Gamma distribution with shape parameter equal to 0.5 and scale parameter equal to 1. (This is sometimes called the Gamma(0.5, 1) distribution.)

```
y <- rgamma(5, shape = 0.5, scale = 1)
print(y)
```

```
## [1] 0.46789822 0.54300176 0.02677446 0.90627667 0.36409104
```

Modify the code to generate n = 10000 samples. Create a histogram and comment on the shape of the distribution. This shape will be similar to the true density curve of the Gamma(0.5, 1) distribution.

```
y <- rgamma(10000, shape = 0.5, scale = 1)
hist(y)
```




```
mean(y)
```

```
## [1] 0.4986724
```

```r
median(y)
```

```
## [1] 0.2276406
```

Answer: The shape of the gamma distribution is unimodal and extremely right-skewed.

d. Using the samples you obtained in part c, approximate the probability that a Gamma(0.5, 1) random variable is between 1 and 2.

```r
pnorm(2, mean(y), sd(y)) - pnorm(1, mean(y), sd(y))
```

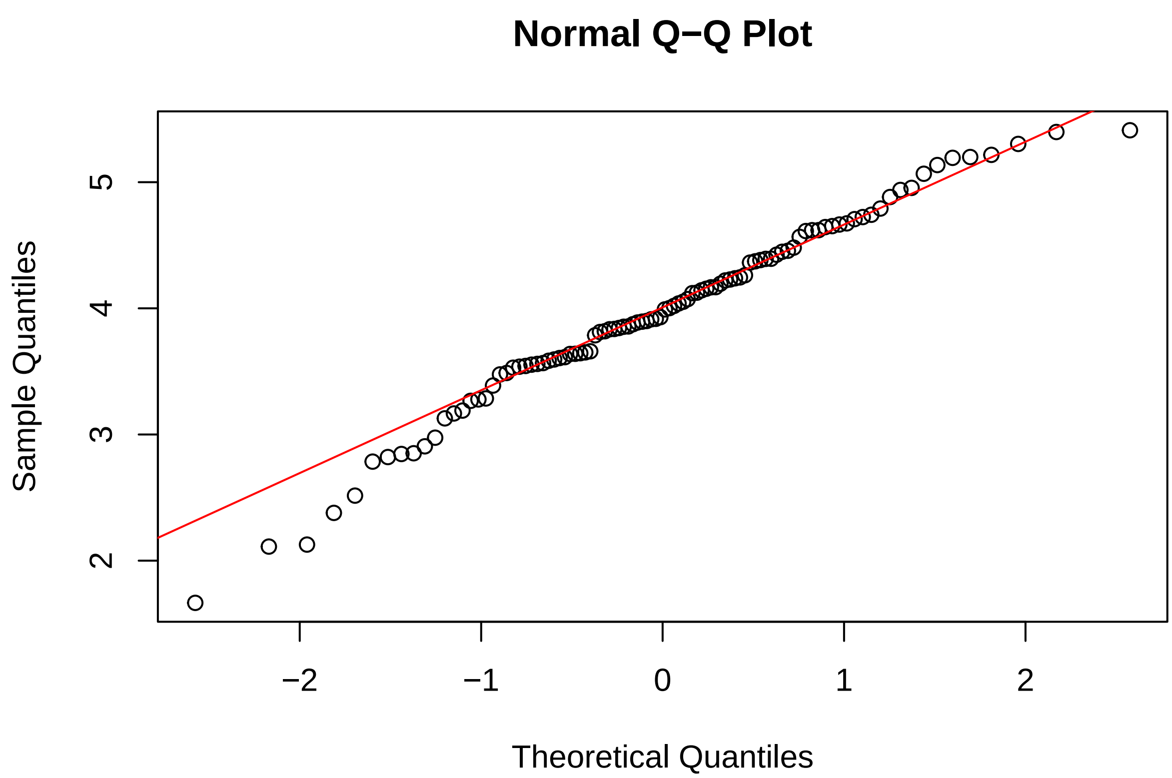
```
## [1] 0.2211221
```

Answer: The probability that a random variable is between 1 and 2 is approximately 0.22.

Exercise 4: Quantile Plots

a. Use the `rnorm` function to generate a sample of size 100 from a Normal(4, 0.75^2) distribution. Create a normal quantile plot of the sample and state what features of the plot indicate whether it is reasonable to consider the sample to be approximately normal.

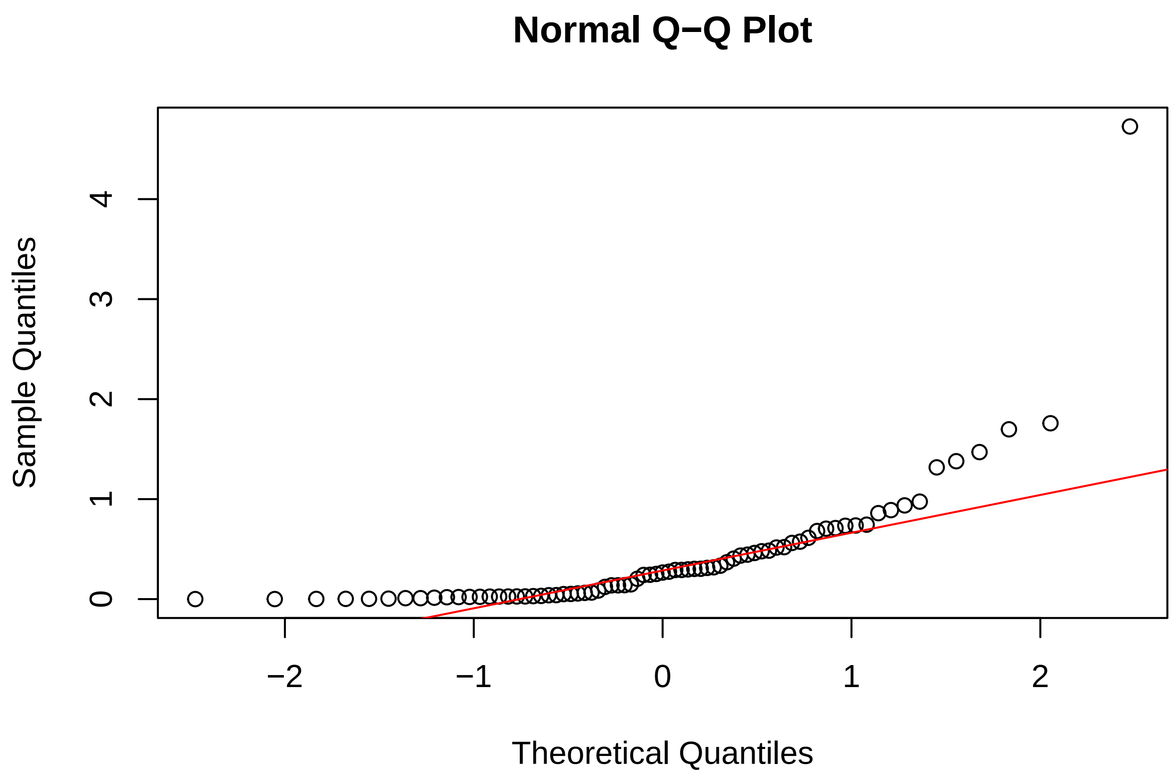
```r
x1 <- rnorm(n = 100, mean = 4, sd = 0.75)
qqnorm(x1)
qqline(x1, col = 'red')
```



Answer: It is not reasonable to consider the sample to be approximately normal because the plotted points do not seem to follow along a 45-degree, uphill line. Instead, they follow an S-shaped pattern along the trend line.

b. Use the `rgamma()` function to generate a sample of size 75 from a Gamma distribution with shape = 0.5 and scale = 1. Create a normal quantile plot of the sample and state what features of the plot indicate whether it is reasonable to consider the sample to be approximately normal.

```r
x1 <- rgamma(75, shape = 0.5, scale = 1)
qqnorm(x1)
qqline(x1, col = 'red')
```



Answer: It is also not reasonable to consider this sample to be approximately normal because of the skewed distributions in the plotting.