Continuous Random Variables I

COSC/DATA 405/505

Probability Density Function (pdf)

Calculation of Probabilities: Cumulative distribution function (cdf)

Expected Value (E)

Variance (Var)

Probability Density Function (pdf)

A continuous function $f(x)$ is a probability density function if it is always nonnegative, and the area under its graph is exactly 1.0. That is,

$$f(x) \geq 0, \text{ for all } x$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

All probability density functions have these two properties.

The pdf completely characterizes the probability model.

The pdf is highest at values of x that are most probable.

Uniform Random Variables

The function

$$f_U(x) = \begin{cases} \frac{1}{b-a}, & \textbf{for } x \in [a,b] \\ 0, & \textbf{otherwise} \end{cases}$$

is an example of a pdf since

$$f_U(x) \geq 0$$

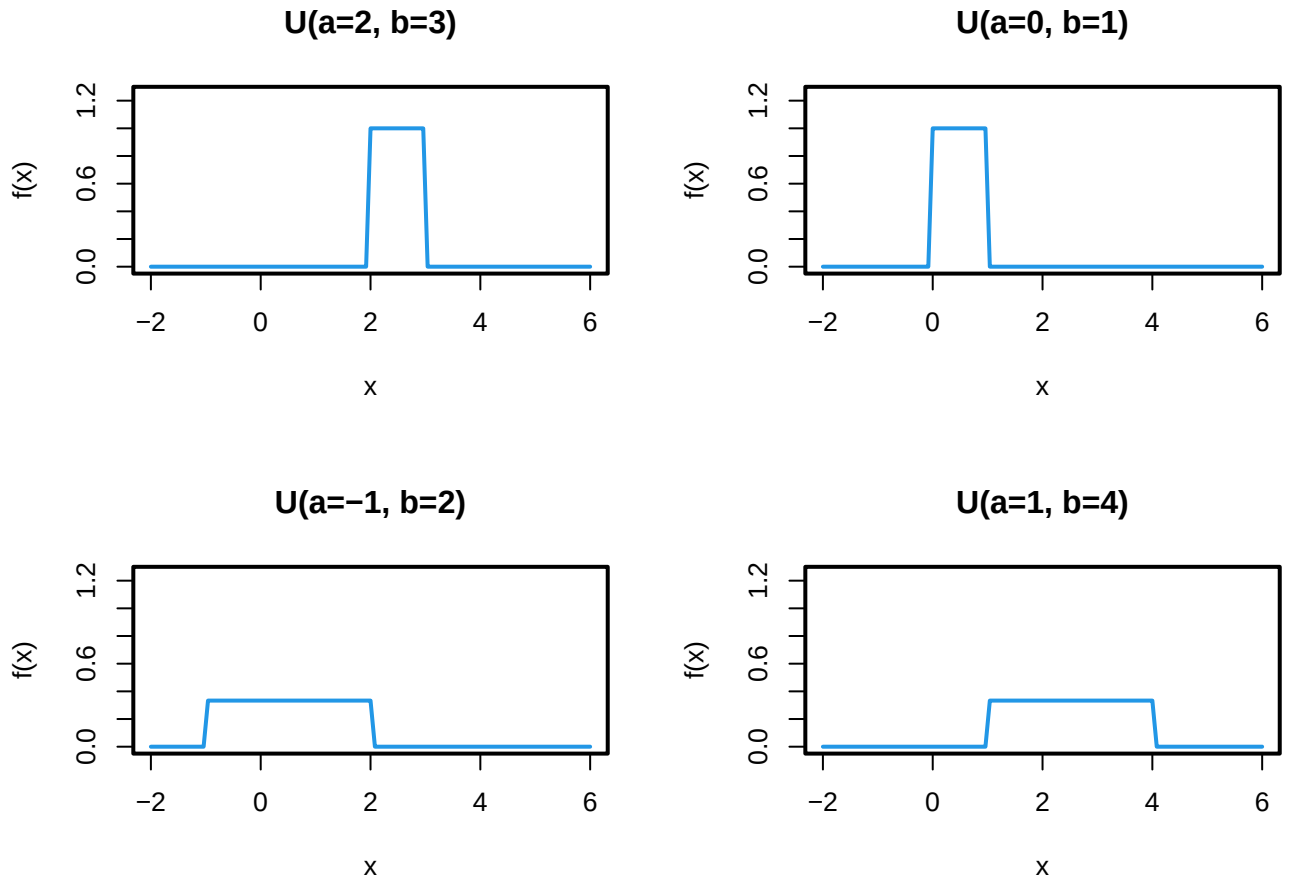
and

$$\int_{-\infty}^{\infty} f_U(x)dx = \int_a^b f_U(x)dx = 1.$$

$f_U(x)$ **is the uniform density function.**

The uniform distribution is a possible model for measurement error but its most important function is as a building block for almost all other distributions.

Picturing Some Examples of the Uniform pdf



The area under the blue curve is 1 in all cases. This represents the probability that the random variable takes a value in the interval $[a, b]$.

Calculation of Probabilities

The probability that a random variable X with density function $f(x)$ takes a value in an interval $[a_1, b_1]$ is calculated as

$$P(a_1 \leq X \leq b_1) = \int_{a_1}^{b_1} f(x)dx.$$

Such probabilities are also expressed in terms of the *cumulative distribution function* (cdf):

$$F(y) = P(X \leq y) = \int_{-\infty}^{y} f(x)dx.$$

$$P(a_1 \leq X \leq b_1) = F(b_1) - F(a_1).$$

Note also that the probability density function can be recovered from the cumulative distribution function by differentiation:

$$f(x) = F'(x).$$

Evaluation of Probabilities in R

The `punif()` function can be used to calculate the probability that a uniform random variable is less than a given value, i.e. the cumulative distribution function at the given value.

To calculate $F(x) = P(X \leq x)$ the use `punif(x, a, b)`. This explicitly evaluates the uniform cumulative distribution function $F(x) = \frac{x-a}{b-a}$, when x lies in $[a, b]$.

For example,

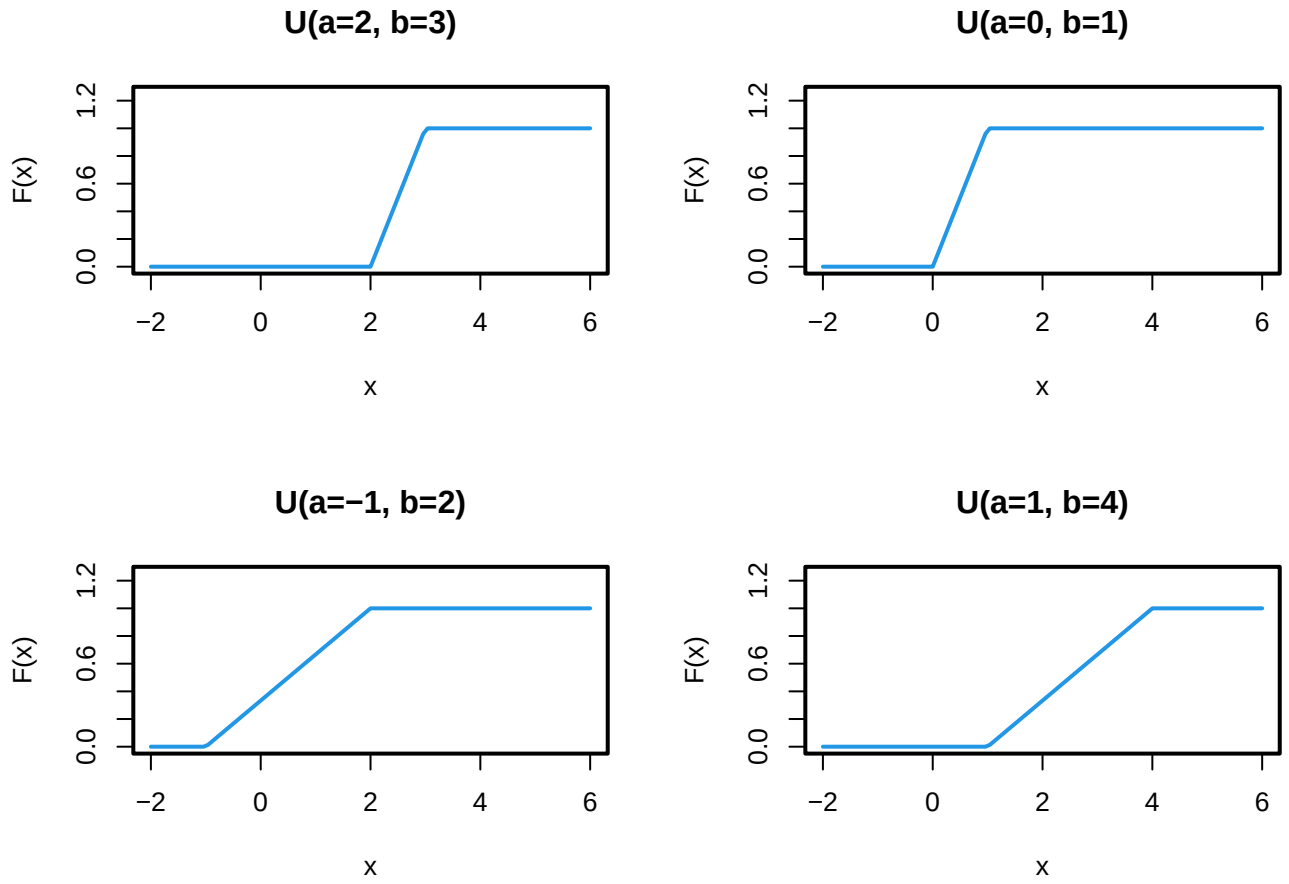
```
punif(.6, 0, 1)
```

```
## [1] 0.6
```

```
punif(2.5, 2, 3)
```

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

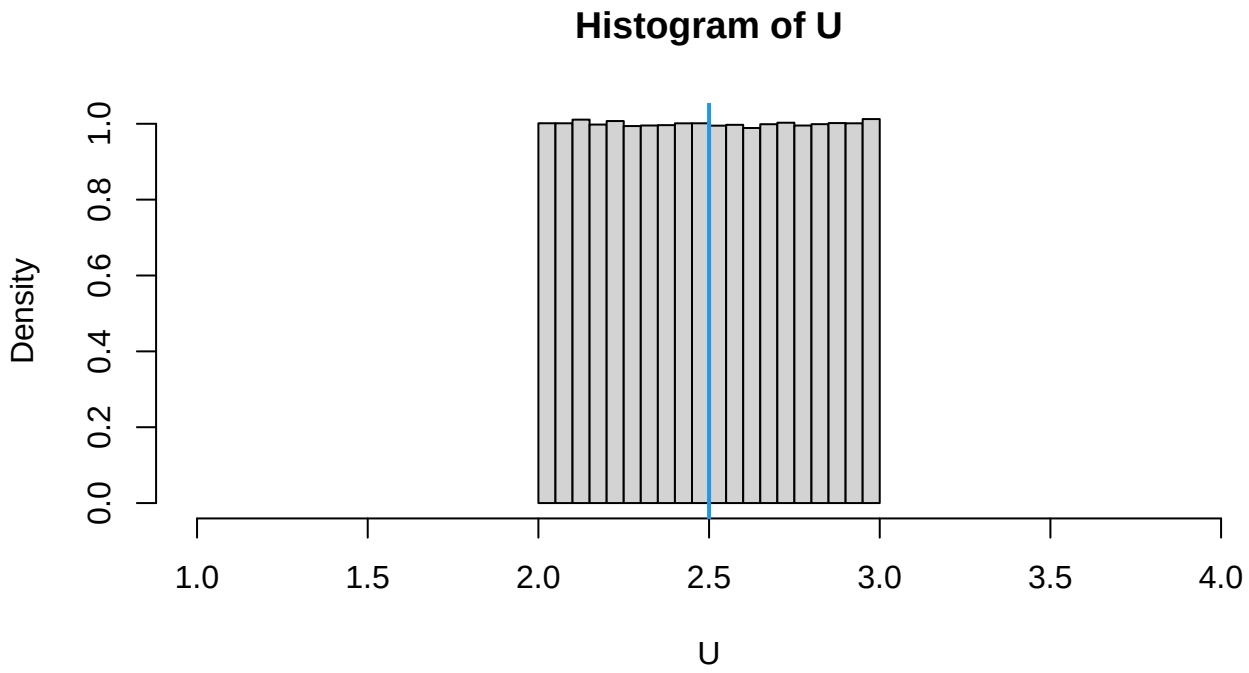
Picturing Some Examples of the Uniform cdf



From these graphs, you can read of the probability that a uniform random variable is less than the given value on the horizontal axis. e.g. in the upper left panel, $F(2.5) = 0.5$ **so the** $P(U \leq 2.5) = 0.5$**.**

Estimation of Probablities by Simulation

```
U <- runif(1000000, 2, 3)
hist(U, freq = FALSE)
abline(v = 2.5, col="blue")
```



The proportion of the area to the left of the blue line estimates the probability that U is less than 2.5.

Observe:

First 10 simulated uniforms:

```
U[1:10]
```

```
##  [1] 2.380 2.775 2.162 2.954 2.019 2.655 2.650 2.100 2.381 2.009
```

Which ones are less than 2.5?

```
U[1:10] < 2.5
```

```
##  [1]  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
```

How many are less than 2.5?

```
sum(U[1:10] < 2.5)    # FALSE is equivalent to 0 in R;  TRUE <--> 1.
```

```
## [1] 6
```

What proportion are less than 2.5?

```
sum(U[1:10] < 2.5)/10  # divide by sample size
```

```
## [1] 0.6
```

Fact: The sample mean is equal to the sum of the sample values (0's and 1's here) divided by the sample size.

Equivalent Calculation:

```
mean(U[1:10] < 2.5) # proportion less than 2.5

## [1] 0.6
```

More accurate calculation would use the larger sample size:

```
mean(U < 2.5)

## [1] 0.5004
```

Additional Examples: Assume U is Uniform on $[2, 3]$.

Estimate $P(U \leq 2.1)$ and compare with true value.

```
mean(U <= 2.1)
```

```
## [1] 0.1001
```

```
punif(2.1, 2, 3)
```

```
## [1] 0.1
```

Additional Examples: Assume U is Uniform on $[2, 3]$.

Estimate $P(U \leq 2.9)$ and compare with the true value.

```
mean(U <= 2.9)
```

```
## [1] 0.8993
```

```
punif(2.9, 2, 3)
```

```
## [1] 0.9
```

Additional Examples: Assume U is Uniform on $[2, 3]$.

Estimate $P(U > 2.9)$ and compare with the true value.

```
mean(U > 2.9)
```

```
## [1] 0.1007
```

```
1 - punif(2.9, 2, 3)
```

```
## [1] 0.1
```

Additional Examples: Assume U is Uniform on $[2, 3]$.

Estimate $P(2.1 \leq U < 2.9)$ and compare with the true value.

```
mean(U < 2.9 & U >= 2.1 )
```

```
## [1] 0.7992
```

```
punif(2.9, 2, 3) - punif(2.1, 2, 3)
```

```
## [1] 0.8
```

Expected Value

The expected value of a single (continuous) random variable X can be written as

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

where $f(x)$ is the probability density function of X.

We say $E[X]$ is the *mean* of X.

The expected value gives us a single number that, at least in a rough sense, conveys a typical value for the random variable.

It is sometimes called a measure of *location*, since it specifies the location of the distribution along the real axis.

Expected Value

For the density function $f_U(x)$, we have

$$E[X] = \int_a^b x/(b-a)dx = \frac{b+a}{2}. \tag{1}$$

In other words, the expected value of a uniform random variable is at the midpoint of the interval.

A commonly used alternate notation for the mean of a distribution is μ, the Greek letter which roughly translates to the letter "m".

Other types of expected value can be calculated by the appropriate integration. For continuous functions $g(x)$, we have

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

When a is a nonrandom constant, and $g(x) = ax$, we have

$$E[aX] = \int_{-\infty}^{\infty} axf(x)dx = a\int_{-\infty}^{\infty} xf(x)dx = aE[X].$$

Expected Value

It can also be shown that

$$E[X + a] = E[X] + a.$$

(Add something to a random variable, and the expected value of the variable will change by that amount.)

For example, if T is the boiling point of a liquid which is subject to random fluctuations in air pressure and with mean $E[T] = 100°C$, the expected boiling point of the temperature measurements if measured in Kelvin units is $E[T + 273] = E[T] + 273 = 373K$.

When $g(x) = x^2$, and the probability density function is as above, we have

$$E[X^2] = \int_a^b \frac{x^2}{(b-a)} dx = \frac{b^3 - a^3}{3(b-a)}$$

Variance

A feature of a distribution which is every bit as important as its location is its *scale*, or a measure of the degree of variability of the distribution.

The variance (or its square root, the standard deviation) is one way to measure the variability of a random variable.

Denoting the mean of X by μ, we have

$$V(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx.$$

An algebraically equivalent expression is

$$V(X) = E[X^2] - \mu^2.$$

Variance

For the uniform distribution $f_U(x)$, the variance is

$$V(X) = \frac{(b-a)^2}{12}.$$

A small value of $V(X)$ implies that there is more certainty about the value of X; it will tend to take values close to μ when $V(X)$ is very small.

The distribution will be more spread out when $V(X)$ is large. (i.e. when a and b are farther apart)

The standard deviation is the square root of the variance. Both quantities summarize the spread or variability in a probability distribution. Note also that

$$\textbf{Var}(aX) = a^2\textbf{Var}(X) \tag{2}$$

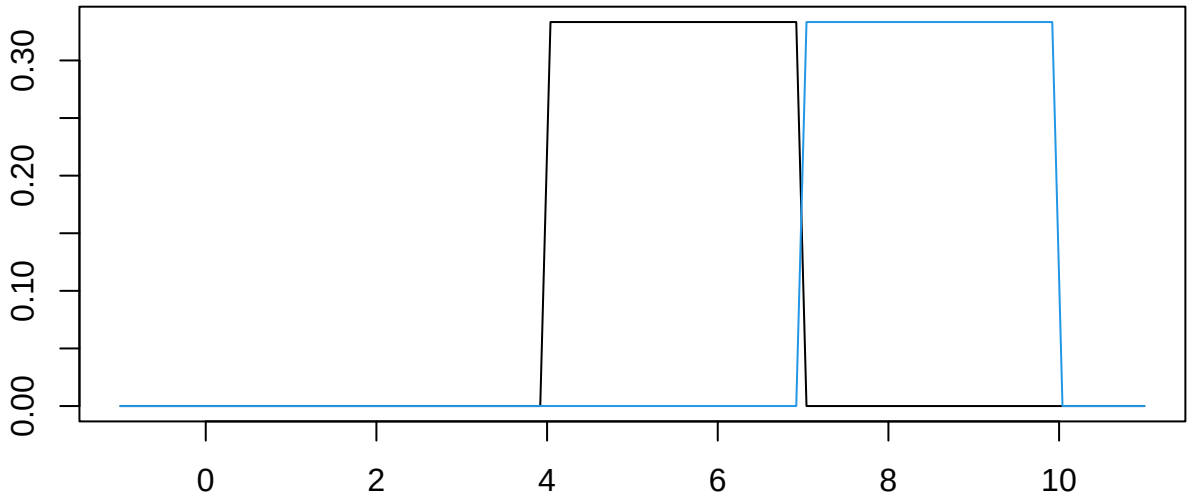
for any nonrandom constant a, and

$$\textbf{Var}(X + a) = \textbf{Var}(X). \tag{3}$$

In other words, the standard deviation of X is multiplied by a when X is. And the spread of the distribution doesn't change if it is only shifted by an amount a.

Example

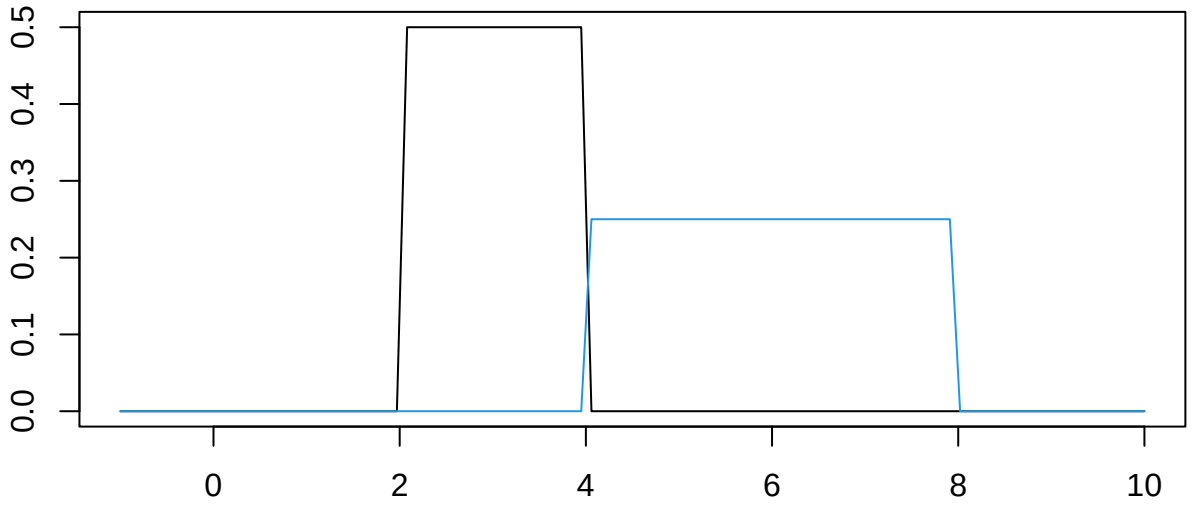
X **and** $X + 3$



No change in range of probable values in the distribution after adding 3.

Example

X **and** $2X$



The distribution becomes much more spread out after multiplying by 2**.**

Calculating the Mean and Variance from a Sample

When confronted with a sample of measurements $x_1, x_2, \ldots, x_n$, we can calculate the *sample mean* by taking the average of the sample values:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j.$$

The *sample variance* is calculated as

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2.$$

The *sample standard deviation* is the square root of this: s.

Calculating the Mean and Variance from a Sample

```
unifSample <- runif(50, 3,  7)
```

For the sample contained in `unifSample`, the sample mean, sample variance, and sample standard deviation are, respectively,

```
mean(unifSample)

## [1] 5.241

var(unifSample)

## [1] 1.655

sd(unifSample)

## [1] 1.286
```

Non-uniform Random Variables - Simulation via Inverse cdf

Weibull and Lognormal Random Variables

Distributions based on the Normal I: χ^2

The Connection between S^2 and the χ^2 Distribution

Examples:

1. **Errors can occur in the production of two-dimensional medical images.**

   A probability model for the proportions of such errors can be of use for quality assurance.

   For example, it is useful to know whether a machine is producing an unusually high proportion of errors.

2. **Probability models are also of use in reliability: what is the probability that an individual or a machine will survive for a given amount of time? Did this component burn out unusually early?**

An approximate model for the proportion of pixels in an image that have been incorrectly classified is

$$f(x) = (\alpha + 1)x^{\alpha}, \quad 0 \leq x \leq 1$$

where α is an *unknown parameter*.

The function $f(x)$ is another example of a *probability density function* (pdf), since it is nonnegative and it integrates to 1.

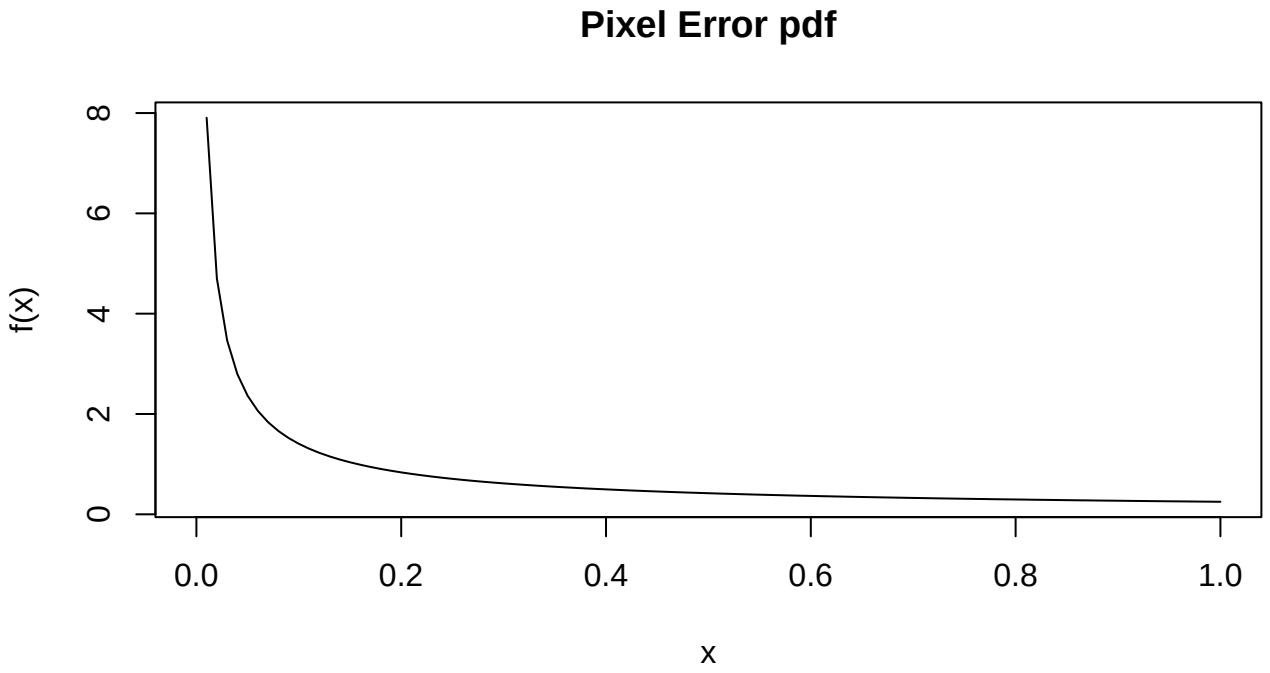
The density is highest at values of x that are most probable.

In this case, we might expect to observe error proportions which are close to 0, and we would not expect to see many near 1.

Visualizing the pdf

The density curve can be plotted using the `curve()` function, which takes a function of x as its first argument.

```
alpha <- -0.75  #  alpha is set to -0.75
curve((alpha+1)*x^alpha, ylab="f(x)",
      main="Pixel Error pdf")
```



Recall: the probability that a random variable X with density function $f(x)$ takes a value in an interval $[a, b]$ is calculated as

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

The *cumulative distribution function* (cdf) is

$$F(y) = P(X \leq y) = \int_{-\infty}^{y} f(x)dx = \int_0^y (\alpha+1)x^\alpha dx = y^{\alpha+1}, \quad y \in [0, 1].$$

Probabilities of Large Error Proportions

For example, we may interested in knowing whether an observed error proportion v is unusually large.

We can check this by calculating the probability that the error proportion X exceeds y.

$$P(X > y) = 1 - F(y) = 1 - y^{\alpha+1}.$$

Note that we are assuming $y \in (0, 1)$ here. If $y \geq 1$, the probability would be 0.

If we know that the value of α is -0.5, then The cumulative distribution function is

$$F(y) = y^{0.5} \quad \textbf{so } P(X > y) = 1 - y^{0.5}.$$

Simulating from the Model

We can use the same procedure as we used to simulate exponential random variables: invert the cdf at a sequence of random variables U.

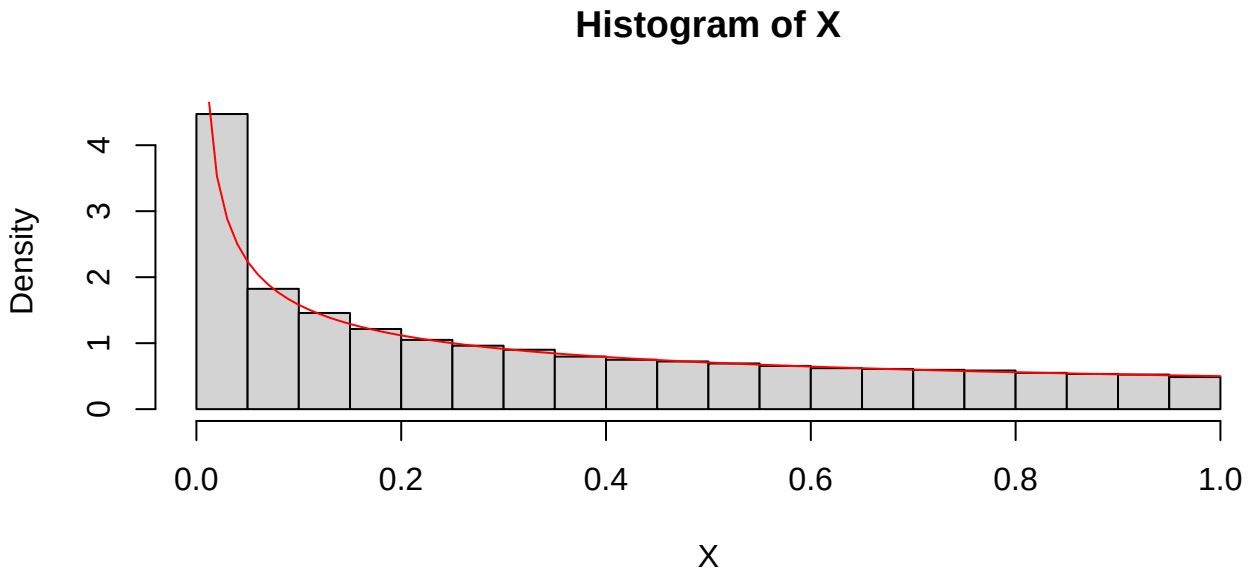
First, let's verify that this makes sense. The mathematics was worked out earlier, to show that if U is a uniform random variable, then when $F^{-1}(U)$ is a random variable with cdf $F(x)$.

We can also use simulation to show this:

```
U <- runif(100000) # simulate lots of uniforms
X <- U^2 # apply inverse cdf (assume alpha = -.5)
```

Plot the histogram and overlaid density curve

```
hist(X, freq = FALSE)
alpha <- -0.5
curve((alpha + 1)*x^(alpha), 0, 1, add = TRUE, col = "red")
```



The density curve matches the relative frequency histogram closely. Exercise: try this for other values of alpha.

General Simulation Method: Inverse CDF

If you have a way of calculating the inverse function of the cdf, the following method can be used to convert uniform numbers to the flavour you are targetting:

```
U <- runif(N) # simulate N uniforms
X <- Finv(U) # tranform to the target distribution
                          # using the inverse of the cdf
X
```

Another Example

Suppose X is a random variable with cdf $F(x) = \sin(x)$ for $x \in [0, \pi/2]$.

1. **Is $F(x)$ a true cdf?**

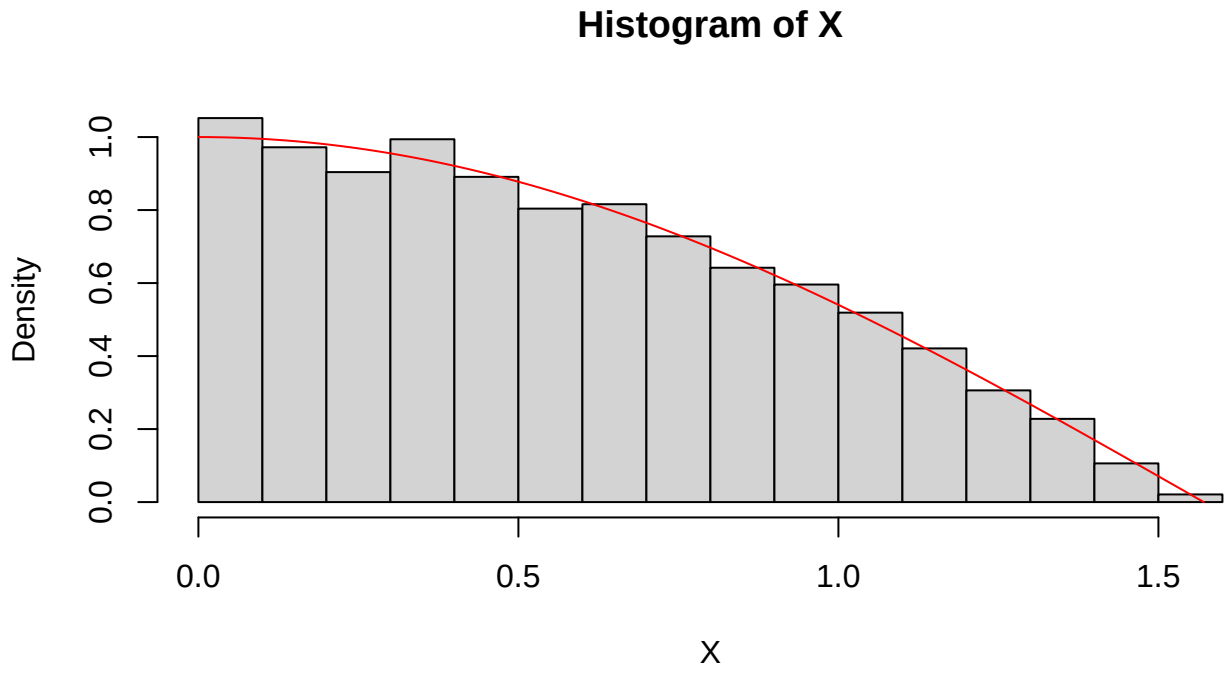
2. **Simulate 10000 random variates from this distribution.**

```
U <- runif(10000)
X <- asin(U)
```

Note that the pdf is $f(x) = \cos(x)$ for $x \in [0, \pi 2/]$, and 0, otherwise.

Plot the histogram and overlaid density curve

```
hist(X, freq = FALSE)
curve(cos(x), 0, pi/2, add = TRUE, col = "red")
```



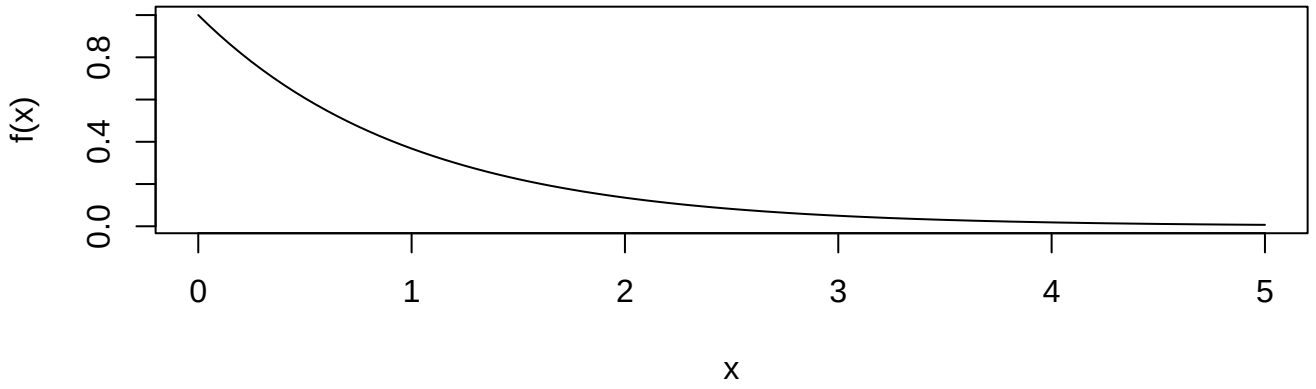
The density curve matches the relative frequency histogram closely.

Recall the exponential distribution, a simple model for a lifetime distribution:

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

and $f(x) = 0$, otherwise.

```
curve(dexp(x), 0, 5, ylab="f(x)")
```



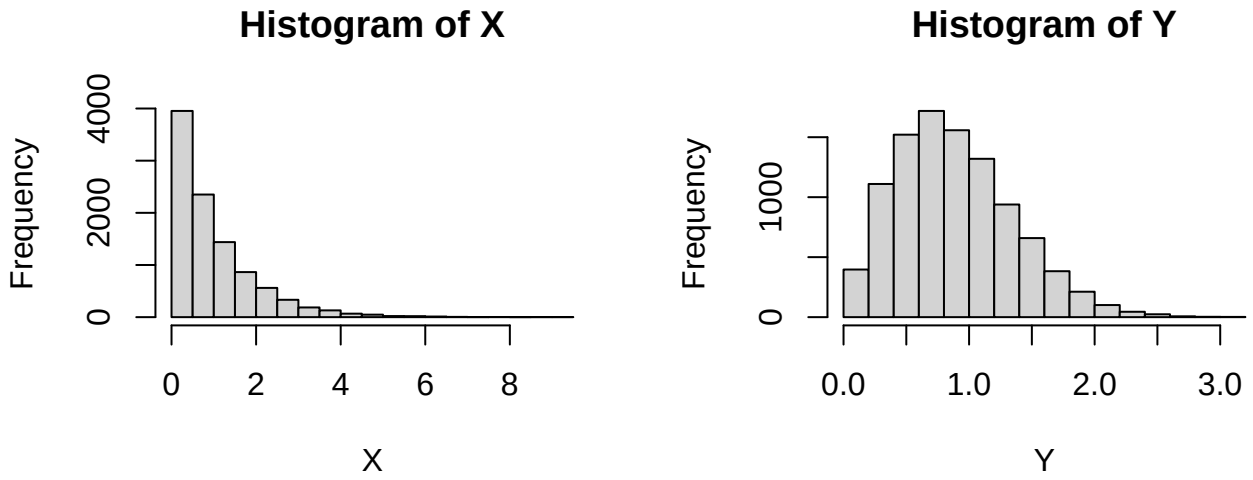
The density is highest near 0. When we simulate from this distribution we get a lot of unrealistically low values:

```
X <- rexp(9); X

## [1] 2.584 1.308 1.329 1.232 0.597 1.266 0.459 1.228 8.319
```

If we take the square root of X, the behaviour is different:

```
sqrt(X)

## [1] 1.607 1.144 1.153 1.110 0.773 1.125 0.677 1.108 2.884
```

```
X <- rexp(10000)
Y <- sqrt(X)
par(mfrow=c(1, 2))
hist(X); hist(Y)
```



The Weibull Distribution

In general, a Weibull random variable is defined as a power of an exponential random variable.

That is, if X is exponential, λ, then $Y = X^{1/\beta}$ is Weibull with parameters β and λ. β controls the shape of the distribution and λ controls the scale.

The cdf of Y is

$$F(y) = P(Y \leq y) = P(X \leq y^{1/\beta}) = 1 - e^{-\lambda y^{1/\beta}}$$

where we used the exponential cdf of X in the middle of the above derivation.

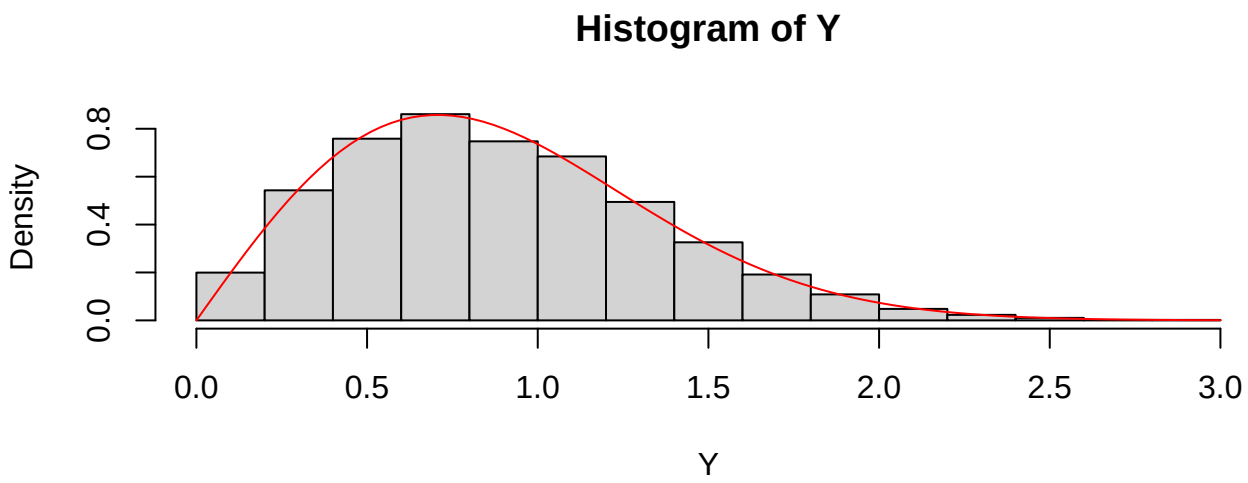
Differentiating $F(y)$ gives you the pdf of the Weibull.

Note:

```
dweibull(x, shape = 2, scale = 1) # Weibull pdf with beta = 2, lambda = 1
pweibull(x, shape = 2, scale = 1) # cdf
rweibull(n, shape = 2, scale = 1) # rng
```

Simulating and comparing with the density curve:

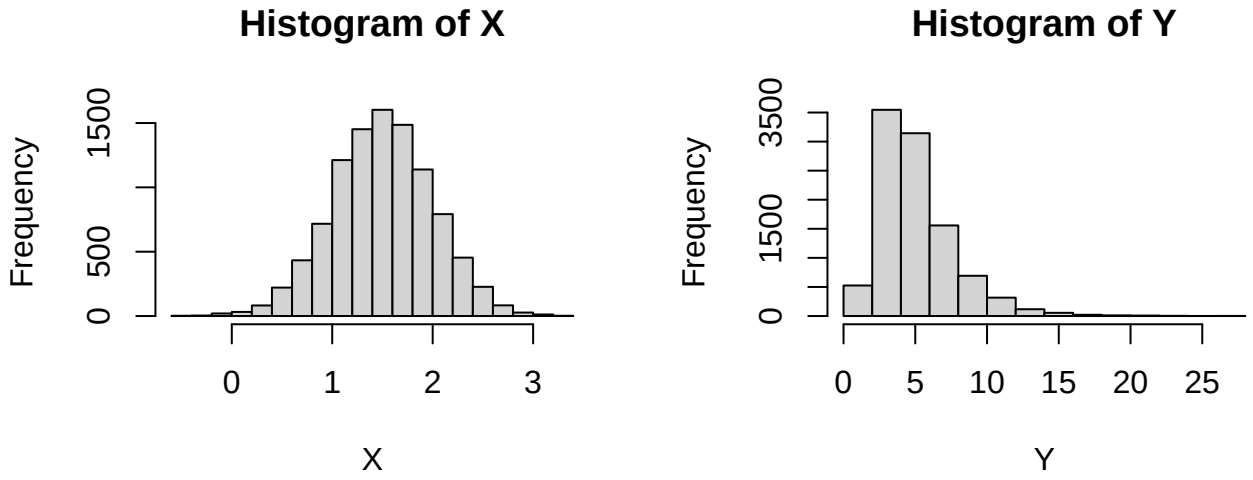
```
Y <- rweibull(10000, shape = 2, scale = 1)
hist(Y, freq = FALSE)
curve(dweibull(x, shape = 2, scale = 1), 0, 3, col = "red", add = TRUE)
```



The Lognormal Distribution

If we take the exponential of X, where X is a normal random variable, we obtain a lognormal random variable, another model for survival times:

```
X <- rnorm(10000, mean = 1.5, sd = 0.5)
Y <- exp(X)
par(mfrow=c(1, 2))
hist(X); hist(Y)
```



Example: Liver Transplant Waiting Times

Data in the `transplant` data slide in the *survival* package relate to waiting times for liver transplant patients.

We consider the males who have type B blood here:

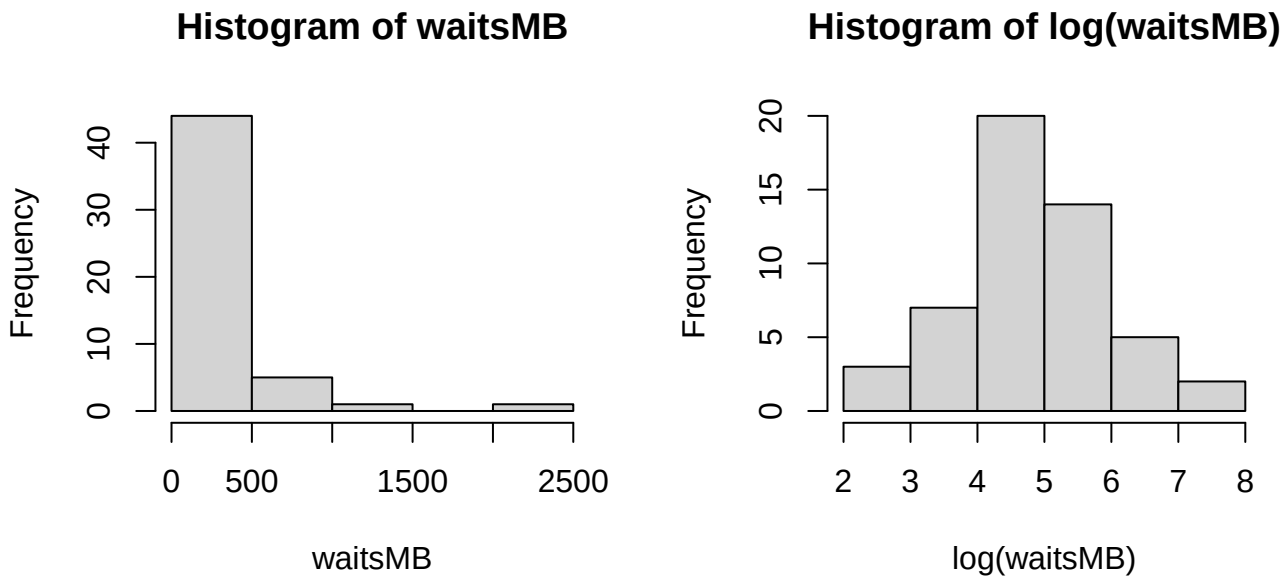
```
library(survival)
waitsMB <-  subset(transplant, sex=="m" & abo=="B" )$futime
```

These data are well approximated by the lognormal distribution.

Example: Liver Transplant Waiting Times

Raw data on left. Log of raw data on right.

```
par(mfrow=c(1, 2))
 hist(waitsMB); hist(log(waitsMB))
```

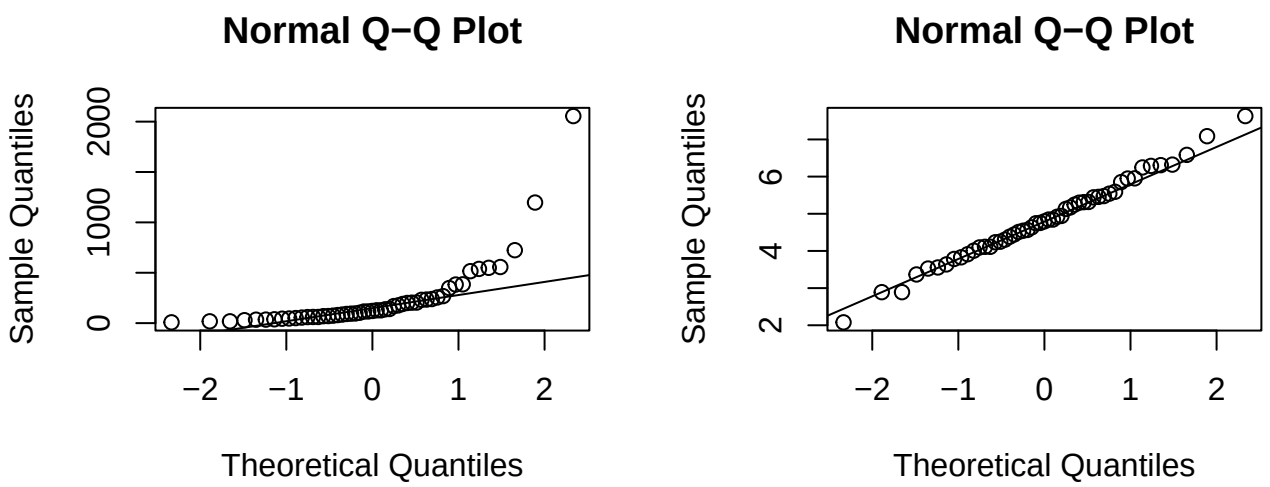


The log plot looks very normal.

QQ-Plot: A Better Way of Checking Normality

Look for a straight line. If you see it, you have normality. If not, you don't.

```
par(mfrow=c(1, 2))
qqnorm(waitsMB); qqline(waitsMB)
qqnorm(log(waitsMB)); qqline(log(waitsMB))
```



The plot on the right looks much straighter than the one on the left.

Construction starts with the standard normal random variable

- **Let Y be a normal random variable with mean μ and standard deviation σ**

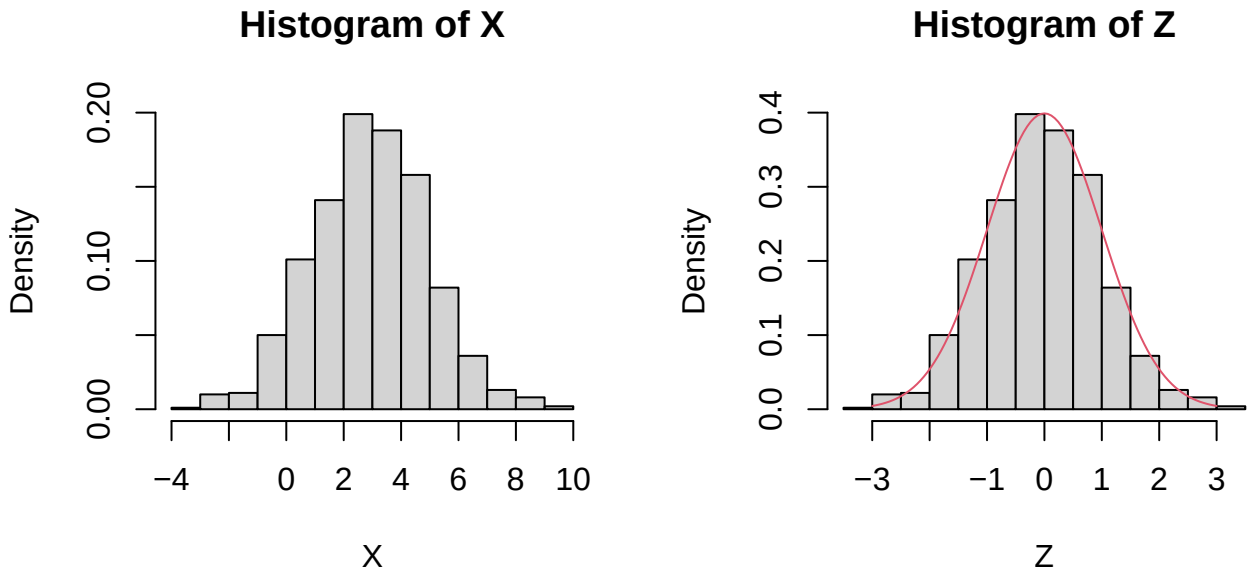
- $$Z = \frac{Y - \mu}{\sigma} \tag{4}$$

  is a standard normal random variable.

Transforming Normal to Standard Normal

Check standardization by simulation:

```
X <- rnorm(1000, mean =3, sd = 2); Z <- (X-3)/2
par(mfrow=c(1, 2))
hist(X, freq=FALSE); hist(Z, freq=FALSE)
curve(dnorm(x), -3, 3, col=2, add=TRUE)
```



The distribution of `Z` is identical to that of `X`, therefore normal. N(0,1) pdf curve matches.

The χ^2 Random Variables

- **Squaring Z leads to a $\chi^2\chi^2$ random variable on 1 degree of freedomdegree of freedom.**

- **Note that**

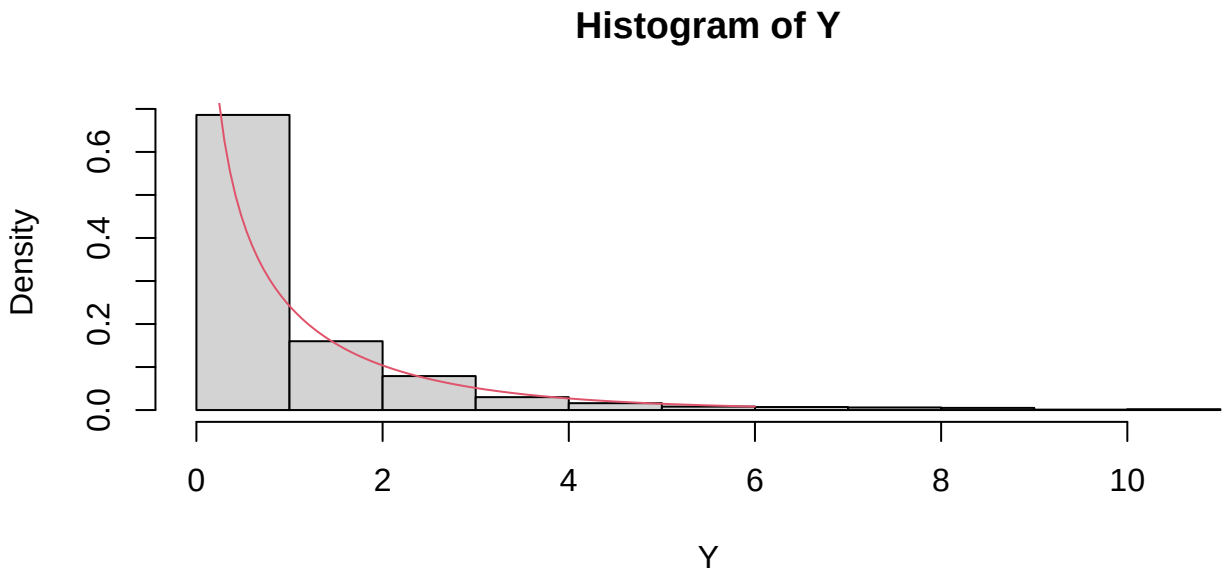
$$E[Z^2] = 1 \tag{5}$$

  $\rightsquigarrow$ a χ^2 random variable on 1 degree of freedom has expected value 1.

On the next slide, we check that Z^2 is χ^2 by simulation, using `dchisq()`.

The χ^2 Random Variables

```
Y <- Z^2
hist(Y, freq=FALSE)
curve(dchisq(x, df = 1), 0, 6, add=TRUE, col=2)
```



χ^2 random variables can be generated using `rchisq()`:

```
rchisq(5, df = 1)

## [1] 0.00385 6.82201 0.00743 0.29101 0.12279
```

- **If $Z_1, \ldots, Z_n$ is a sequence of n independent standard normal random variables, then**

$$X = \sum_{j=1}^{n} Z_j^2 \tag{6}$$

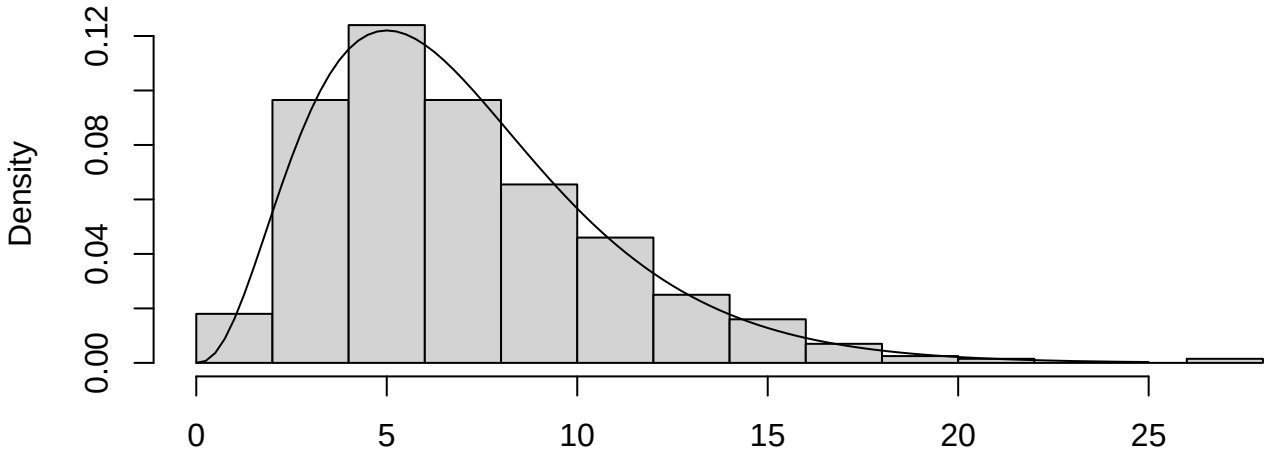
  is a $\chi^2_{(n)}$ random variable on n degrees of freedom.

-

$$E[\sum_{j=1}^{n} Z_j^2] = n \tag{7}$$

1000 simulated values of X for the case where $n = 7$

```
X <-  rchisq(1000, df = 7)
hist(X, freq = FALSE, main = " ")
curve(dchisq(x, df = 7), from = 0, to = 25, add = TRUE)
```



The χ^2 Random variables

- **if μ was known, but σ^2 was unknown, we could estimate σ^2 from a sample of Y's in an unbiased manner by using the formula**

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu)^2. \tag{8}$$

- **Unbiasedness follows from noting that**

$$E[\widehat{\sigma}^2] = \frac{1}{n}\sum_{j=1}^{n}E[(Y_j - \mu)^2] = \sigma^2. \tag{9}$$

- $\widehat{\sigma^2}/\sigma^2 = \frac{\sum(y_i-\mu)^2}{n\sigma^2}$ **is a χ^2 random variable on n degrees of freedom.**

- **Usually μ is not known. Then, S_Y^2 (which is $\widehat{\sigma^2}$ with μ replaced by $\bar{Y}$) is an unbiased estimator for σ^2, and $(n-1)S_Y^2/\sigma^2$ is a χ^2 random variable on $n-1$ degrees of freedom.**

Demonstration of Connection Between S^2 and χ^2

Let us consider a random samples of $n = 20$ normal random variables, each with mean 3 and standard deviation 2, and let us draw 1000 such samples.

We will show that $(n-1)S^2/\sigma^2$ has a χ^2 distribution on 19 degrees of freedom:

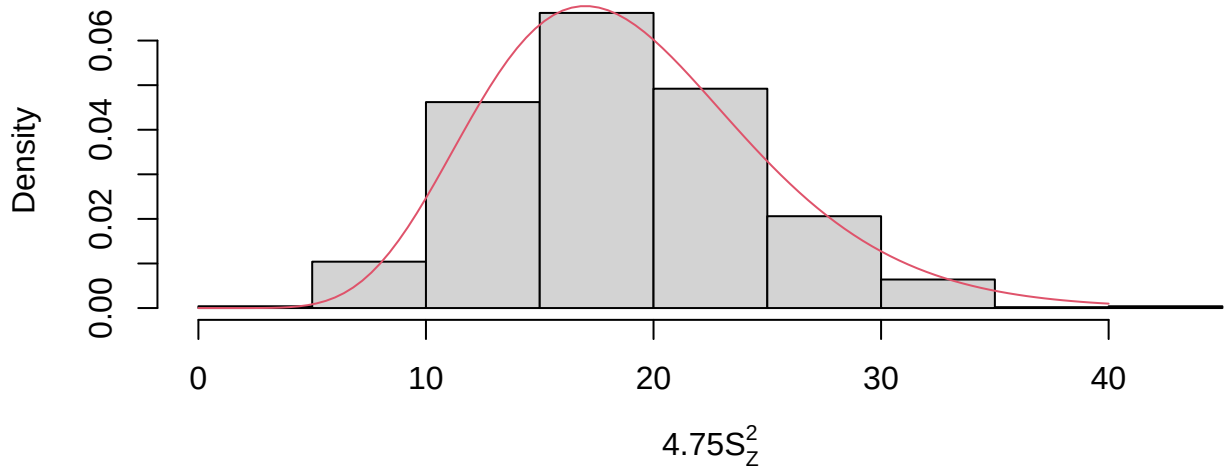
```
m <- 1000; n <- 20; sigma <- 2
# m samples of size n:
Z <- matrix(rnorm(m*n, mean = 3, sd = sigma), nrow=n)
S2z <- apply(Z, 2, var)
sqrt(S2z[1:5])  # look at the first 5 sample standard deviati

## [1] 1.28 2.41 1.86 1.77 2.27
```

These are scattered about $\sigma = 2$.

Demonstration of Connection Between S^2 and χ^2

Compare the relative frequency histogram of $(n-1)S^2/\sigma^2$ with the χ^2 density curve:

```
hist((n-1)*S2z/sigma^2, freq = FALSE, main = " ")
curve(dchisq(x, df = 19), 0, 40, col = 2, add = TRUE)
```



The histogram approximates the density curve closely. Exercise: check this result for other sample sizes, such as $n = 2, 5, 10, 50$. Try different values of the μ and σ as well.