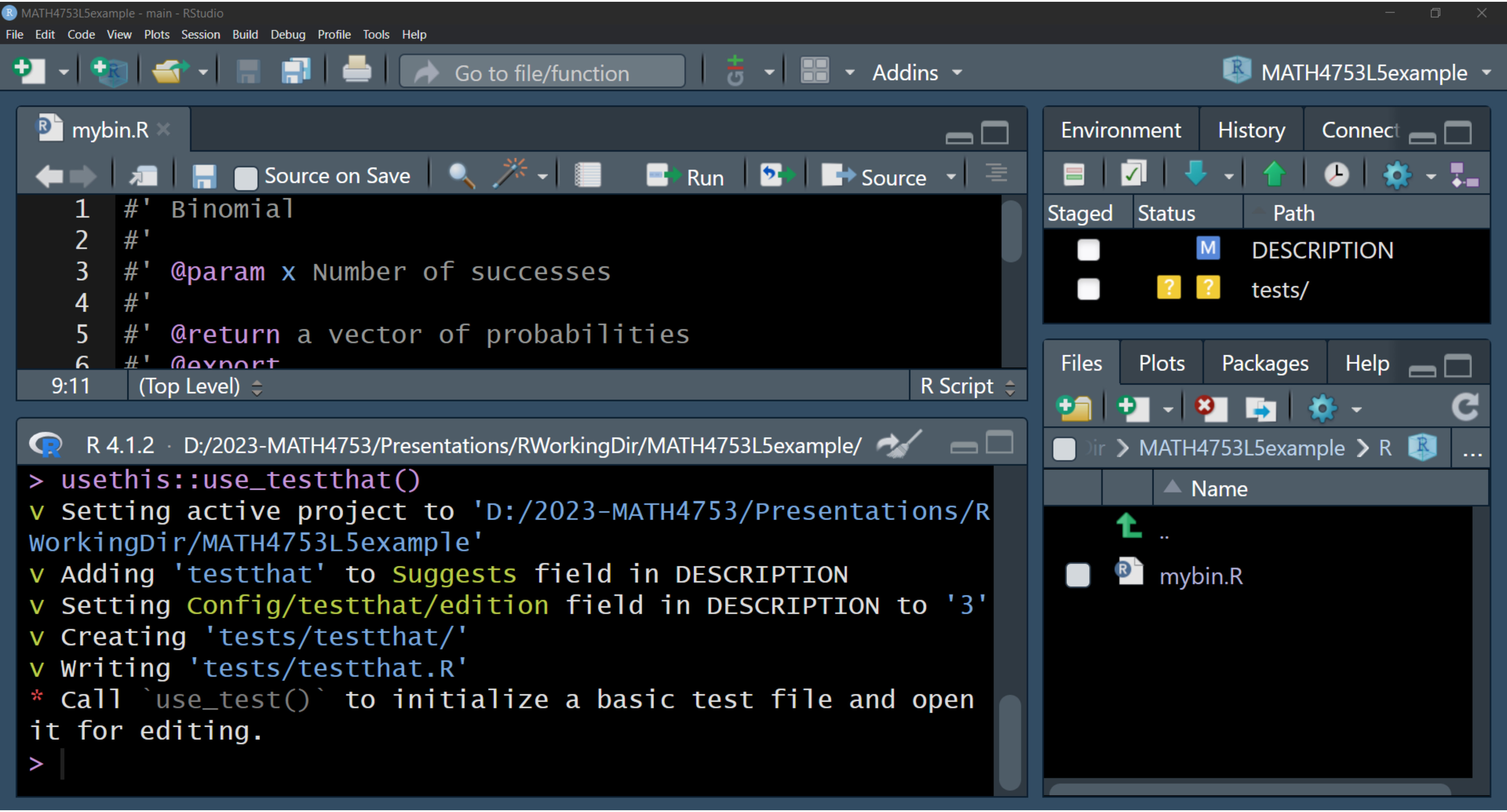
Chapter 5

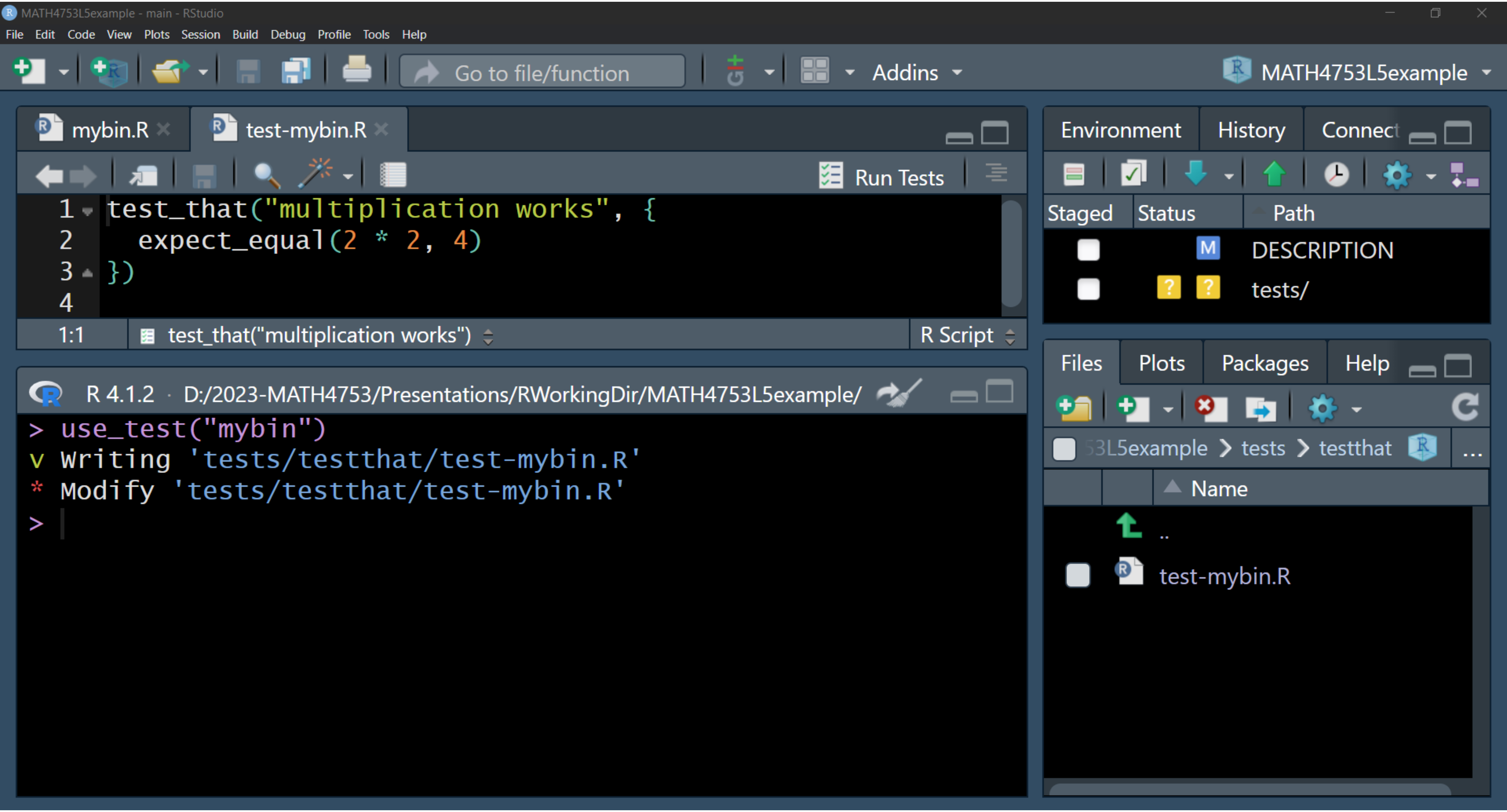
Dr Wayne Stewart

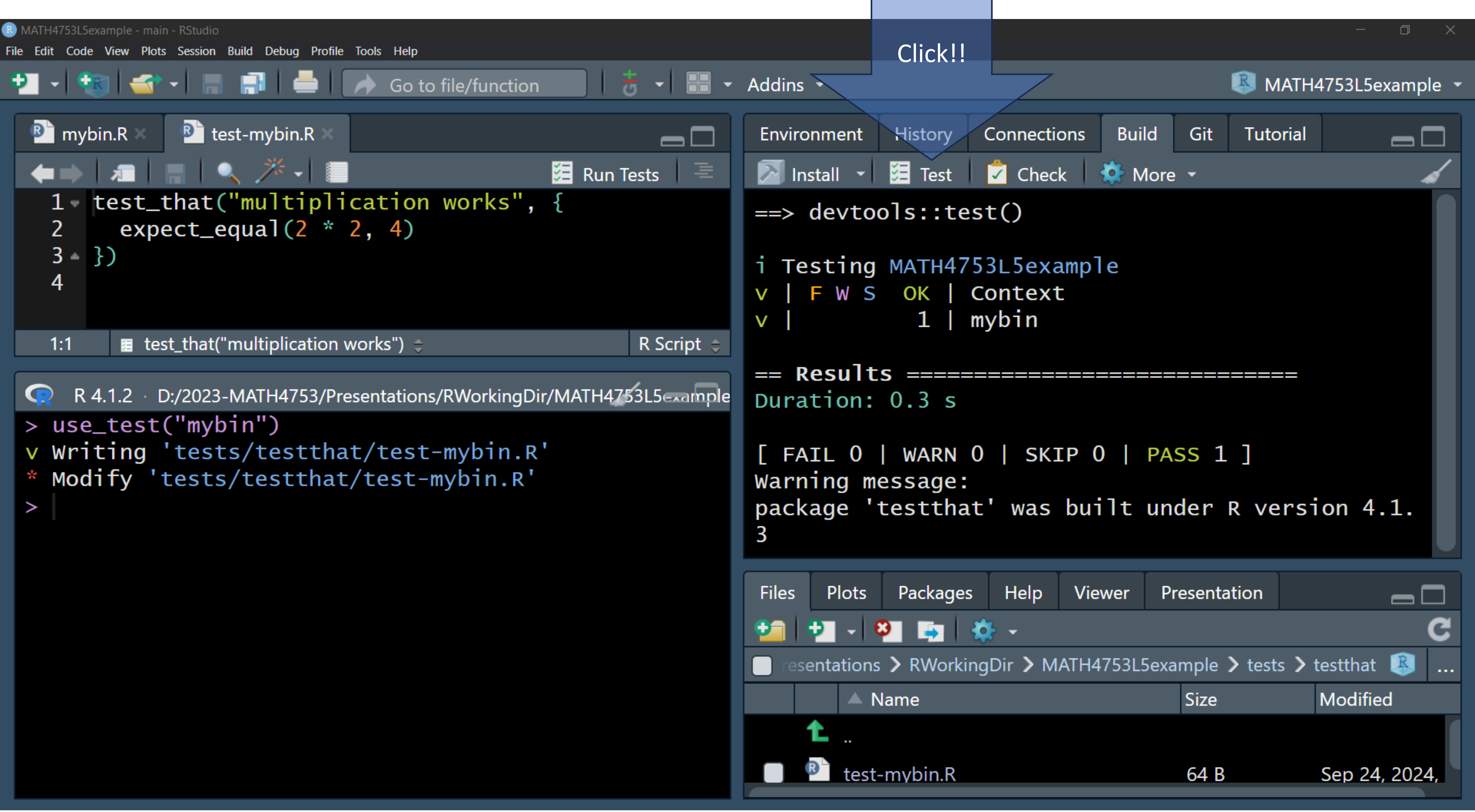
Setting up tests for functions

```
usethis::use_testthat()
```



MATH4753L5example - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
MATH4753L5example
mybin.R
Source on Save
Run
Source
1 #' Binomial
2 #'
3 #' @param x Number of successes
4 #'
5 #' @return a vector of probabilities
6 #' @export
9:11
(Top Level)
R Script
R 4.1.2 · D:/2023-MATH4753/Presentations/RWorkingDir/MATH4753L5example/
> usethis::use_testthat()
v Setting active project to 'D:/2023-MATH4753/Presentations/RWorkingDir/MATH4753L5example'
v Adding 'testthat' to Suggests field in DESCRIPTION
v Setting Config/testthat/edition field in DESCRIPTION to '3'
v Creating 'tests/testthat/'
v Writing 'tests/testthat.R'
* Call `use_test()` to initialize a basic test file and open it for editing.
>
Environment
History
Connect
Staged
Status
Path
M
DESCRIPTION
?
?
tests/
Files
Plots
Packages
Help
MATH4753L5example
R
Name
..
mybin.R


MATH4753L5example - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
MATH4753L5example
mybin.R
test-mybin.R
Run Tests
test_that("multiplication works", {
expect_equal(2 * 2, 4)
})
1:1
test_that("multiplication works")
R Script
R 4.1.2 · D:/2023-MATH4753/Presentations/RWorkingDir/MATH4753L5example/
> use_test("mybin")
v Writing 'tests/testthat/test-mybin.R'
* Modify 'tests/testthat/test-mybin.R'
>
Environment
History
Connect
Staged
Status
Path
M
DESCRIPTION
?
?
tests/
Files
Plots
Packages
Help
3L5example
tests
testthat
Name
..
test-mybin.R


MATH4753L5example - main - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
MATH4753L5example
Click!!
mybin.R
test-mybin.R
Run Tests
test_that("multiplication works", {
expect_equal(2 * 2, 4)
})
1:1
test_that("multiplication works")
R Script
R 4.1.2 · D:/2023-MATH4753/Presentations/RWorkingDir/MATH4753L5example
> use_test("mybin")
v Writing 'tests/testthat/test-mybin.R'
* Modify 'tests/testthat/test-mybin.R'
>
Environment History Connections Build Git Tutorial
Install Test Check More
==> devtools::test()
i Testing MATH4753L5example
v | F W S OK | Context
v | 1 | mybin
== Results ================================
Duration: 0.3 s
[FAIL 0 | WARN 0 | SKIP 0 | PASS 1]
Warning message:
package 'testthat' was built under R version 4.1.3
Files Plots Packages Help Viewer Presentation
esentations › RWorkingDir › MATH4753L5example › tests › testthat
Name Size Modified
..
test-mybin.R 64 B Sep 24, 2024,

Look what should be made and returned

These go in CANVAS comments

- Task 7
  - Look at the function `myncurve()` below and edit it so that it will display the curve, shaded area between the curve and x axis from $-\infty$ to x=a, and calculate the area (probability, P(X<=a)) which is released to the command-line in a list.

```
myncurve = function(mu, sigma){
  curve(dnorm(x,mean=mu,sd=sigma), xlim = c(mu-3*sigma, mu + 3*sigma))
  list(mu = mu, sigma = sigma)
}
```

  - Add the completed function to your package – you can consult http://r-pkgs.had.co.nz/ for more information about building R packages.
  - Use the R package testthat and create three tests one for each component of the named list.
  - Build and install the package
  - And run the function in an R chunk (part of Lab6.Rmd) to prove the package is working by calculating `ILAS2019::myncurve(mu=10,sigma=5, a=6)`
    - Paste in the CANVAS comments the last 10 lines of check()
    - Paste in the CANVAS comments the output of the testthat "Test" button found next to "check"

```
Example: ==> devtools::test()

i Testing mytoypackage
v | F W S  OK | Context
v |         2 | myquad [0.1s]

== Results ===================
Duration: 0.4 s

[ FAIL 0 | WARN 0 | SKIP 0 | PASS 2 ]
Warning message:
package 'testthat' was built under R version 4.1.3
```

Continuous Random Variables

OBJECTIVE

To distinguish between continuous and discrete random variables and their respective probability distributions; to present some useful continuous probability distributions and show how they can be used to solve some practical problems

CONTENTS

STATISTICS IN ACTION

Super Weapons Development—Optimizing the Hit Ratio

Skills

- Plot densities
- Calculate
  - densities
  - probabilities
  - quantiles
- Know the functions
  - d-stem
  - p-stem
  - q-stem
  - r-stem

Definition 5.1

The **cumulative distribution function** $F(y_0)$ for a random variable Y is equal to the probability

$$F(y_0) = P(Y \leq y_0), -\infty < y_0 < \infty$$

n=5, p=0.5

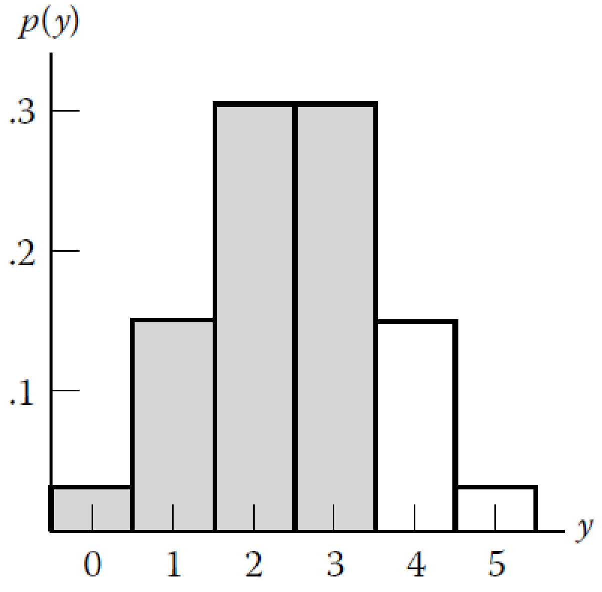


FIGURE 5.1
Probability distribution for a binomial random variable ($n = 5$, $p = 5$); shaded area corresponds to $F(3)$

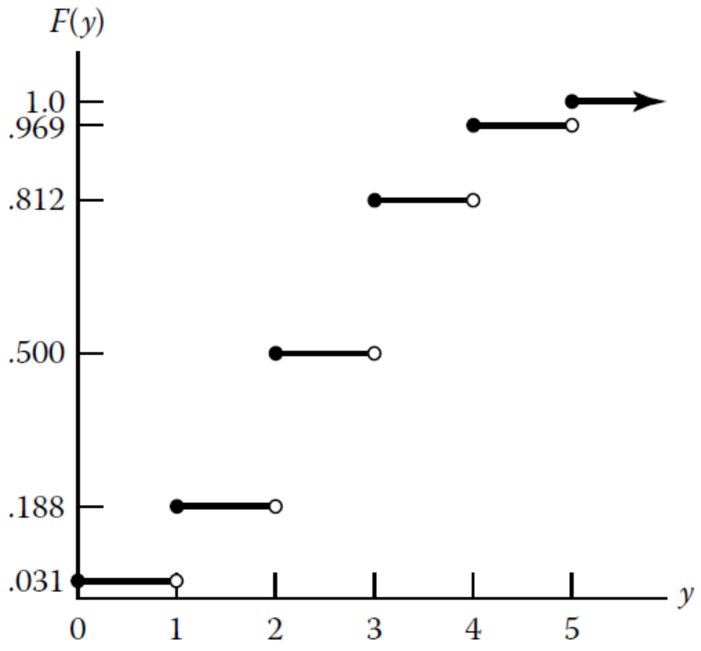


FIGURE 5.2
Cumulative distribution function $F(y)$ for a binomial random variable ($n = 5$, $p = 5$)

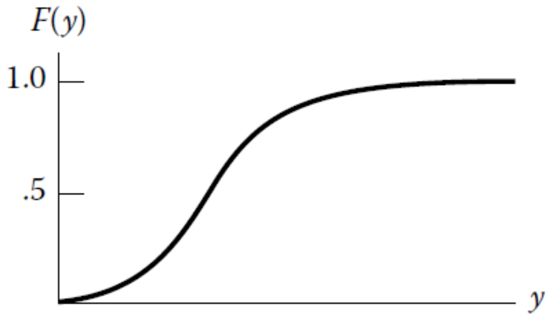


FIGURE 5.3
Cumulative distribution function for a continuous random variable

Definition 5.2

A **continuous random variable** *Y* is one that has the following three properties:

1. *Y* takes on an uncountably infinite number of values in the interval $(-\infty, \infty)$.
2. The cumulative distribution function, $F(y)$, is continuous.
3. The probability that *Y* equals any one particular value is 0.

Definition 5.3

If $F(y)$ is the cumulative distribution function for a continuous random variable Y, then the **density function $f(y)$** for Y is

$$f(y) = \frac{dF(y)}{dy}$$

Cumulative distribution function

Very important

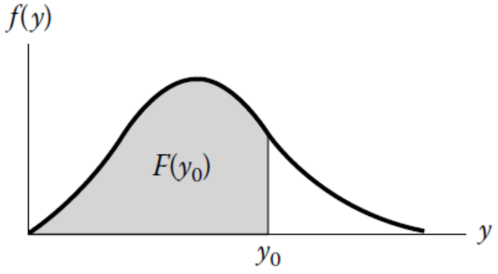


FIGURE 5.4
Density function $f(y)$ for a continuous random variable

The density function for a continuous random variable y, the model for some real-life population of data, will usually be a smooth curve, as shown in Figure 5.4. It follows from Definition 5.3 that

$$F(y) = \int_{-\infty}^{y} f(t)\, dt$$

Thus, the cumulative area under the curve between $-\infty$ and a point y_0 is equal to $F(y_0)$.

The density function for a continuous random variable must always satisfy the three properties given in the following box.

Properties of a Density Function for a Continuous Random Variable Y

1. $f(y) \geq 0$
2. $\int_{-\infty}^{\infty} f(y)\, dy = F(\infty) = 1$
3. $P(a < Y < b) = \int_{a}^{b} f(y)\, dy = F(b) - F(a)$, where a and b are constants

5.1 Let c be a constant and consider the density function for the random variable Y:

$$f(y) = \begin{cases} cy^2 & \text{if } 0 \leq y \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the value of c.
b. Find the cumulative distribution function $F(y)$.
c. Compute $F(1)$.
d. Compute $F(.5)$.
e. Compute $P(1 \leq Y \leq 1.5)$.

Make sure you can apply theory to this example

Solution

5.1.

$$f(y) = \begin{cases} cy^2 & 0 \le y \le 2 \\ 0 & \text{else} \end{cases}$$

a) Find c.

$$\int_{-\infty}^{\infty} f(t)\,dt = 1 \quad (*)$$

here $f(y) = 0$ when $-\infty < y < 0$ and $2 < y < +\infty$

So $(*)$ reduces to

$$\int_0^2 cy^2\,dy = 1 \iff c\left[\frac{1}{3}y^3\right]_0^2 = 1$$

$$c\,\frac{8}{3} = 1$$

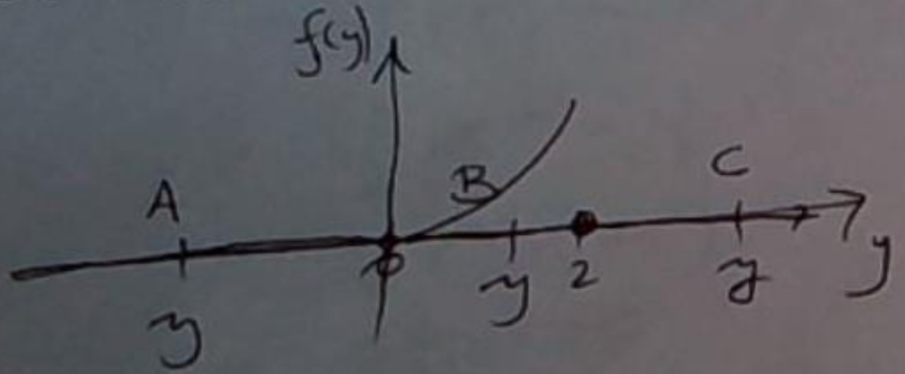
$$c = \frac{3}{8}$$

So
$$f(y) = \begin{cases} \frac{3}{8}y^2 & 0 \le y \le 2 \\ 0 & \text{else} \end{cases}$$

b) $F(y)$

$$F(y) = \int_{-\infty}^{y} f(t)\,dt$$

There are three regions



Region A $F(y) = 0$

B $F(y) = \int_0^y \frac{3}{8} t^2 dt = \frac{1}{8} y^3$

C $F(y) = 1$

$$F(y) = \begin{cases} 0 & -\infty < y < 0 \\ \frac{1}{8} y^3 & 0 \le y \le 2 \\ 1 & y > 2 \end{cases}$$

RAR man!!

c) $F(1) = \frac{1}{8} 1^3 = \frac{1}{8}$

d) $F(0.5) = \frac{1}{8} 0.5^3 = \frac{1}{8} * \frac{1}{8} = \frac{1}{64}$

e) $P(1 \le Y \le 1.5)$

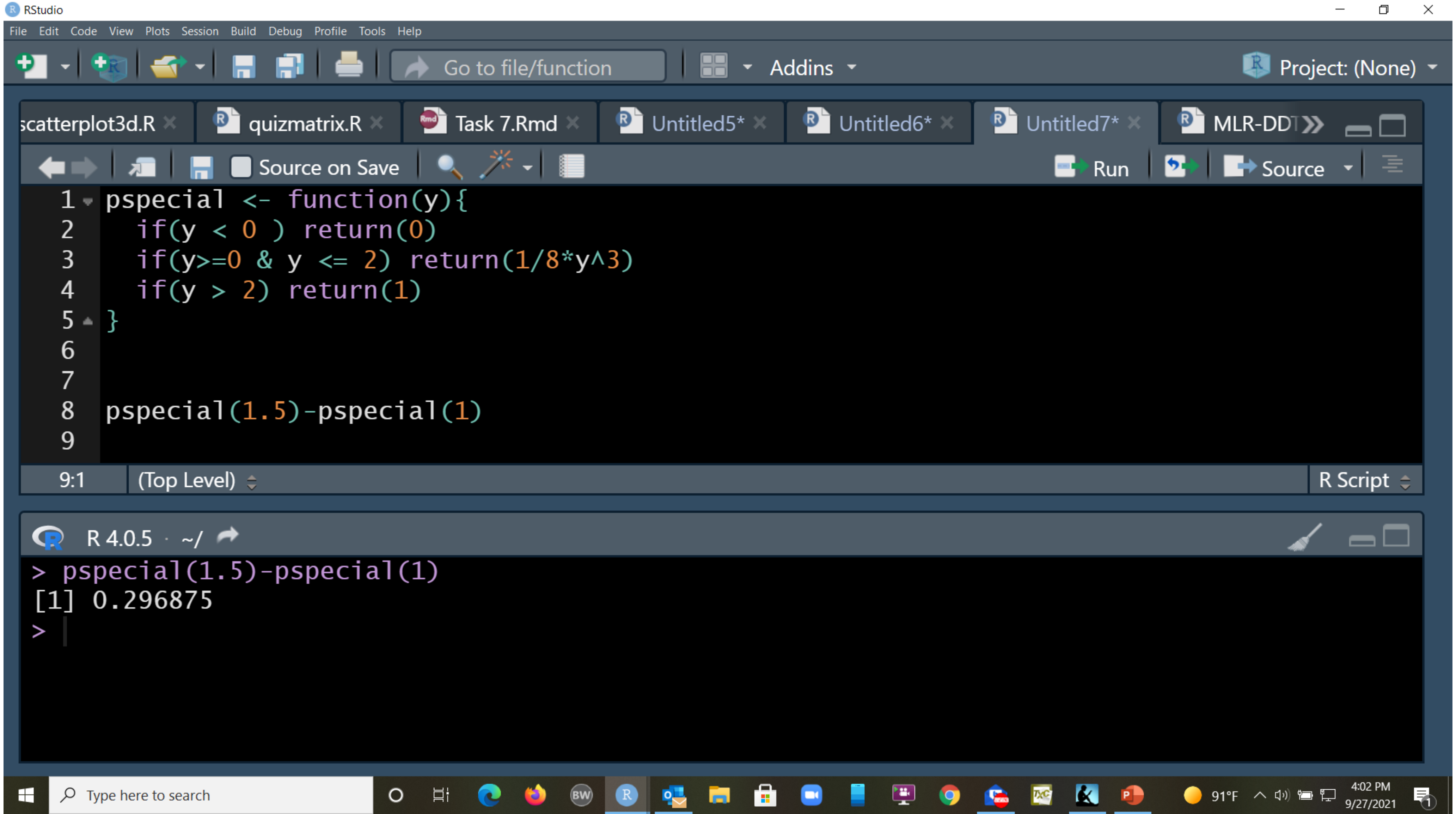
$= F(1.5) - F(1)$

$= \frac{1}{8} 1.5^3 - \frac{1}{8} 1^3$

$= \frac{1}{8}(1.5^3 - 1)$

$= \underline{0.296875}$

Make your own pstem



Expected Value

Definition 5.4

Let Y be a continuous random variable with density function $f(y)$, and let $g(Y)$ be any function of Y. Then the **expected values** of Y and $g(Y)$ are

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\, dy$$

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) f(y)\, dy$$

THEOREM 5.1

Let c be a constant, let Y be a continuous random variable, and let $g_1(Y), g_2(Y), \ldots, g_k(Y)$ be k functions of Y. Then,

$$E(c) = c$$

$$E(cY) = cE(Y)$$

$$E[g_1(Y) + g_2(Y) + \cdots + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + \cdots + E[g_k(Y)]$$

THEOREM 5.2

Let Y be a continuous random variable with $E(Y) = \mu$. Then

$$\sigma^2 = E[(Y - \mu)^2] = E(Y^2) - \mu^2$$

THEOREM 5.3

Let Y be a random variable* with mean μ and variance σ^2. Then the variances of $(c + Y)$ and cY are

$$V(c + Y) = \sigma^2_{(c+Y)} = \sigma^2 \quad \text{and} \quad V(cY) = \sigma^2_{cY} = c^2\sigma^2$$

Proof of Theorem 5.3 From Theorem 5.1, we know that $E(cY) = cE(Y) = c\mu$. Using the definition of the variance of a random variable, we can write

$$V(cY) = \sigma^2_{cY} = E[(cY - c\mu)^2] = E\{[c(Y - \mu)]^2\} = E[c^2(Y - \mu)^2]$$

Then, by Theorem 5.1,

$$\sigma^2_{cY} = c^2E[(Y - \mu)^2]$$

But, $E[(Y - \mu)^2] = \sigma^2$. Therefore,

$$\sigma^2_{cY} = c^2\sigma^2$$

5.10 *Time a train is late.* Refer to Exercise 5.5 (p. 191) The amount of time Y (in minutes) that a commuter train is late is a continuous random variable with probability density

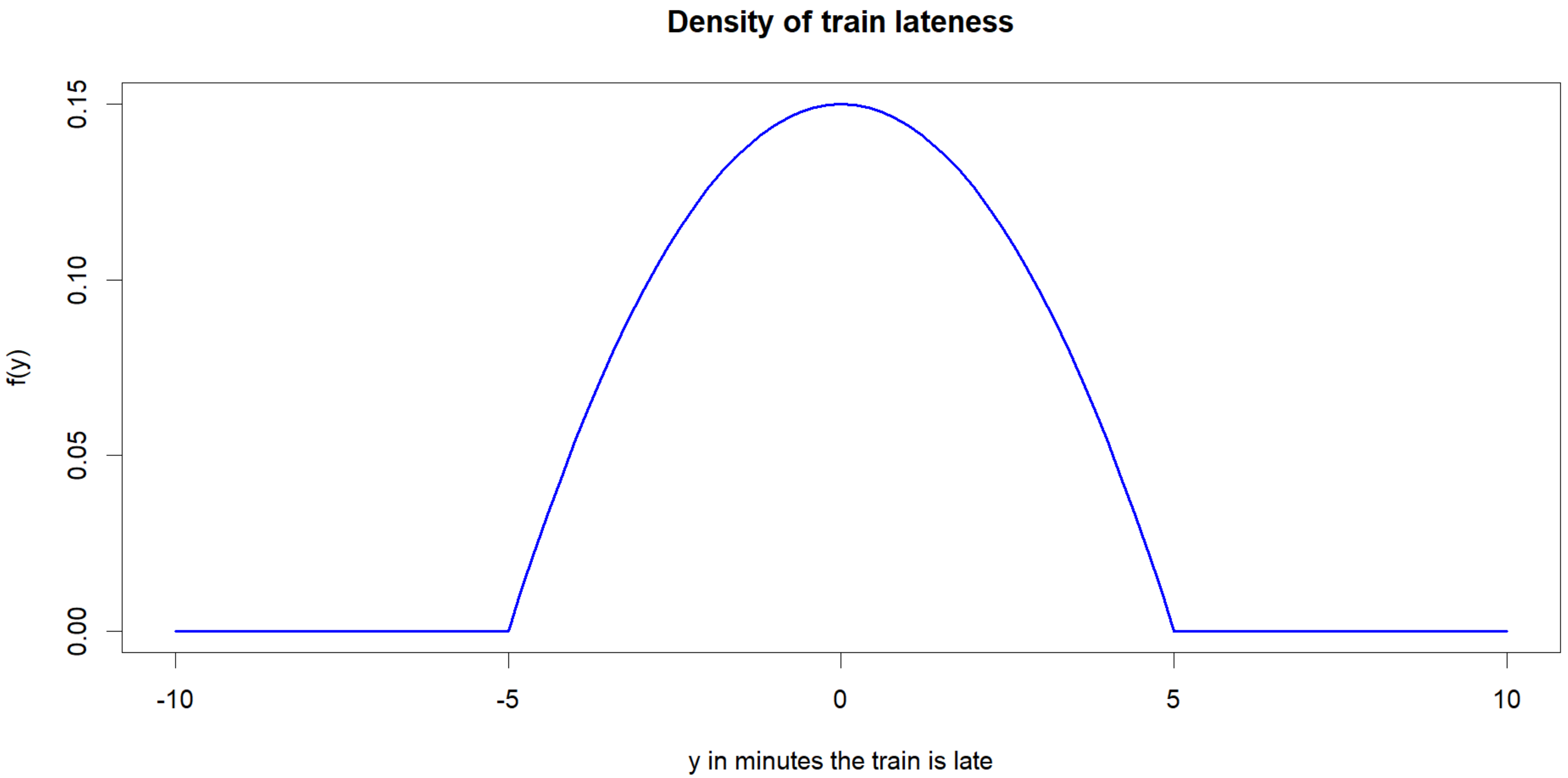
$$f(y) = \begin{cases} \frac{3}{500}(25 - y^2) & \text{if } -5 < y < 5 \\ 0 & \text{elsewhere} \end{cases}$$

a. Find the mean and variance of the amount of time in minutes the train is late.
b. Find the mean and variance of the amount of time in hours the train is late.
c. Find the mean and variance of the amount of time in seconds the train is late.

```r
# Late train MS 5.10
dtrain <- function(y){
  f <- ifelse(y > -5 & y < 5,3/500*(25-y^2),0) # vectorized
    return(f)
}

curve(dtrain(x),
      xlim = c(-10,10),
      xlab = "y in minutes the train is late",
      ylab = "f(y)",
      main = "Density of train lateness",
      lwd = 2,
      col = "Blue")

```



Density of train lateness
f(y)
0.00
0.05
0.10
0.15
-10
-5
0
5
10
y in minutes the train is late

5.10 a. $\mu = E(Y) = \int_{-5}^{5} yf(y)\,dy = \int_{-5}^{5} \frac{3y}{500}(25-y^2)\,dy = \int_{-5}^{5}\left(\frac{3y}{20} - \frac{3y^3}{500}\right)dy$

$$= \left(\frac{3y^2}{40} - \frac{3y^4}{2000}\right)\Bigg]_{-5}^{5} = \left(\frac{3(5)^2}{40} - \frac{3(5)^4}{2000}\right) - \left(\frac{3(-5)^2}{40} - \frac{3(-5)^4}{2000}\right)$$

$$= (1.875 - 0.9375) - (1.875 - 0.9375) = 0$$

$$E(Y^2) = \int_{-5}^{5} y^2 f(y)\,dy = \int_{-5}^{5} \frac{3y^2}{500}(25-y^2)\,dy = \int_{-5}^{5}\left(\frac{3y^2}{20} - \frac{3y^4}{500}\right)dy$$

$$= \left(\frac{y^3}{20} - \frac{3y^5}{2500}\right)\Bigg]_{-5}^{5} = \left(\frac{(5)^3}{20} - \frac{3(5)^5}{2500}\right) - \left(\frac{(-5)^3}{20} - \frac{3(-5)^5}{2500}\right)$$

$$= (6.25 - 3.75) - (-6.25 + 3.75) = 5$$

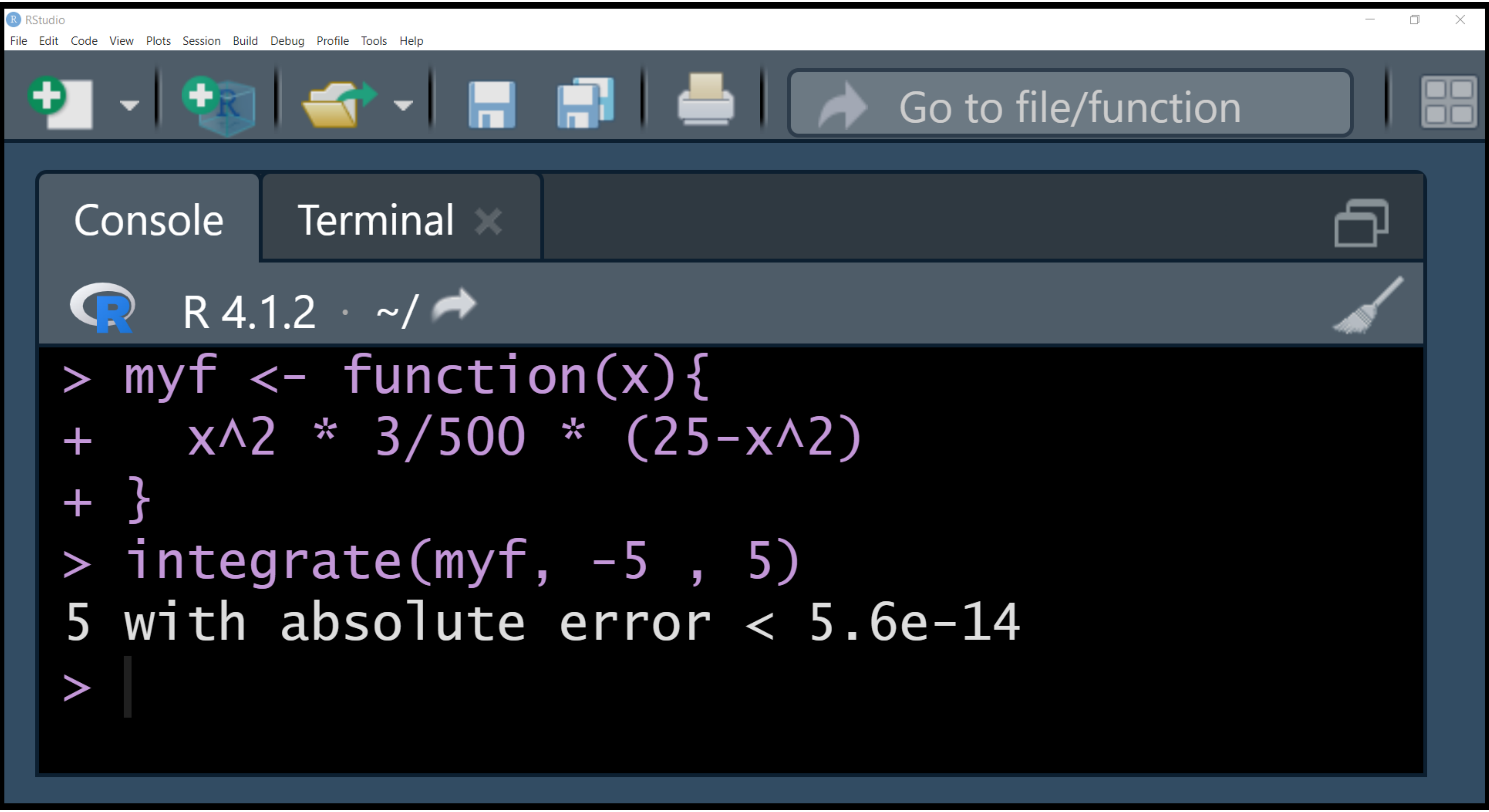
$$\sigma^2 = E(Y^2) - \mu^2 = 5 - 0^2 = 5$$

b. Let $X = \frac{1}{60}Y$. Then $\mu_X = E(X) = E\left(\frac{1}{60}Y\right) = \frac{1}{60}E(Y) = \frac{1}{60}(0) = 0$

$$\sigma_X^2 = \sigma_{1/60Y}^2 = \left(\frac{1}{60}\right)^2 \sigma_Y^2 = \frac{1}{3600}(5) = 0.0014$$

c. Let $Z = 60Y$. Then $\mu_Z = E(Z) = E(60Y) = 60E(Y) = 60(0) = 0$

$$\sigma_Z^2 = \sigma_{60Y}^2 = (60)^2\,\sigma_Y^2 = 3600(5) = 18{,}000$$


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Console
Terminal
R 4.1.2 · ~/
> myf <- function(x){
+ x^2 * 3/500 * (25-x^2)
+ }
> integrate(myf, -5 , 5)
5 with absolute error < 5.6e-14
>

The uniform

The Uniform Probability Distribution

The probability density function for a **uniform random variable**, Y, is given by

$$f(y) = \begin{cases} \dfrac{1}{b - a} & \text{if } a \leq y \leq b \\ 0 & \text{elsewhere} \end{cases}$$

$$\mu = \frac{a + b}{2} \qquad \sigma^2 = \frac{(b - a)^2}{12}$$

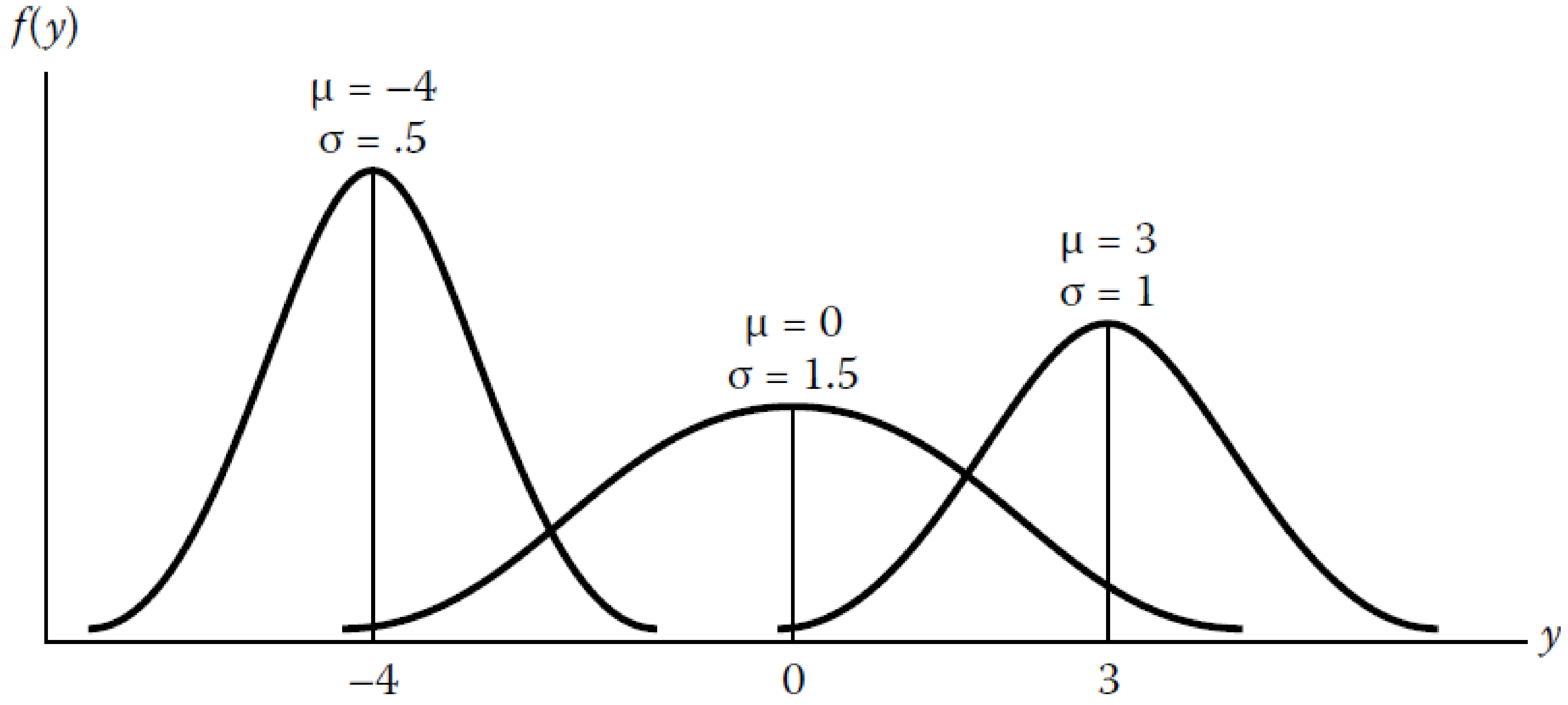
The Normal

The Normal Probability Distribution

The density function for a **normal random variable**, Y, is given by

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

The parameters μ and σ^2 are the mean and variance, respectively, of the normal random variable Y.



f(y)
μ = −4
σ = .5
μ = 0
σ = 1.5
μ = 3
σ = 1
y
−4
0
3

THEOREM 5.4

If Y is a normal random variable with mean μ and variance σ^2, then $Z = (Y - \mu)/\sigma$ is a normal random variable with mean 0 and variance 1.* The random variable Z is called a **standard normal variable**.

$$Y \sim N(\mu, \sigma^2)$$

$$Z = \frac{Y-\mu}{\sigma} = \frac{1}{\sigma}Y - \frac{\mu}{\sigma}$$

$$E(Z) = \frac{1}{\sigma}E(Y) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

$$V(Z) = V\left(\frac{1}{\sigma}Y - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}V(Y) = \frac{\sigma^2}{\sigma^2} = 1$$

?

$$Z \sim \dot{N}(0, 1)$$

MGF for a Normal dist.

$X \sim N(\mu, \sigma^2)$

then $M_X(t) = e^{\mu t + \frac{1}{2}t^2\sigma^2} = E(e^{Xt})$

$Y = aX + b$, $M_Y(t) = ?$

$$M_Y(t) \stackrel{\text{def}}{=} E(e^{Yt})$$
$$= E(e^{(aX+b)t})$$
$$= E(e^{aXt+bt})$$
$$= E(e^{bt} e^{aXt})$$
$$= e^{bt} E(e^{aXt}) \quad \text{by r. 2}$$
$$= e^{bt} E(e^{Xt'}) \quad \text{where } t' = at$$
$$= e^{bt} M_X(at)$$

$$M_Y(t) = e^{bt} M_X(at)$$

$$= e^{bt} e^{\mu at + \frac{1}{2}(at)^2\sigma^2}$$

$$= e^{bt + \mu at + \frac{1}{2}a^2\sigma^2 t^2}$$

$$= e^{(b+\mu a)t + \frac{1}{2}t^2 a^2\sigma^2}$$

$$\mu_Y = b + \mu a \qquad \sigma_Y^2 = a^2\sigma^2$$

$$Y \sim N(b + \mu a,\ a^2\sigma^2)$$

Example

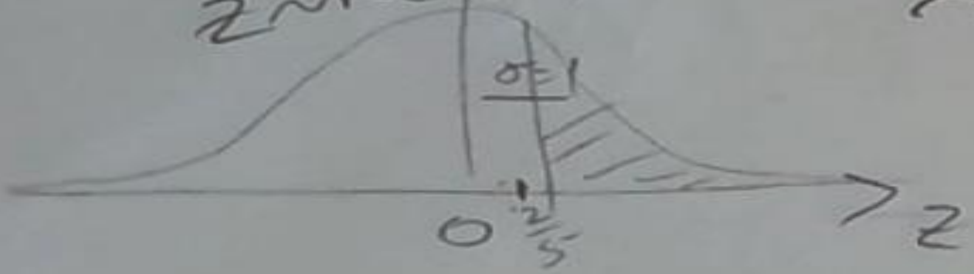
$Y \sim N(\mu = 10, \sigma = 5)$

$P(Y > 12)$

$1 - P(Y \leq 12) = 1 - \text{pnorm}(12, 10, 5)$

Book

$Z \sim N(0, 1)$



BOOK

NO!

$P(Y > 12) = P\left(Z > \frac{12-10}{5}\right)$

$= P\left(Z > \frac{2}{5}\right)$

look at tables

$1 - \text{pnorm}\left(\frac{2}{5}, 0, 1\right)$

About finding probabilities with the Normal

The book is correct in its calculations but the method is now redundant.

Please use R to make calculations without converting to Z

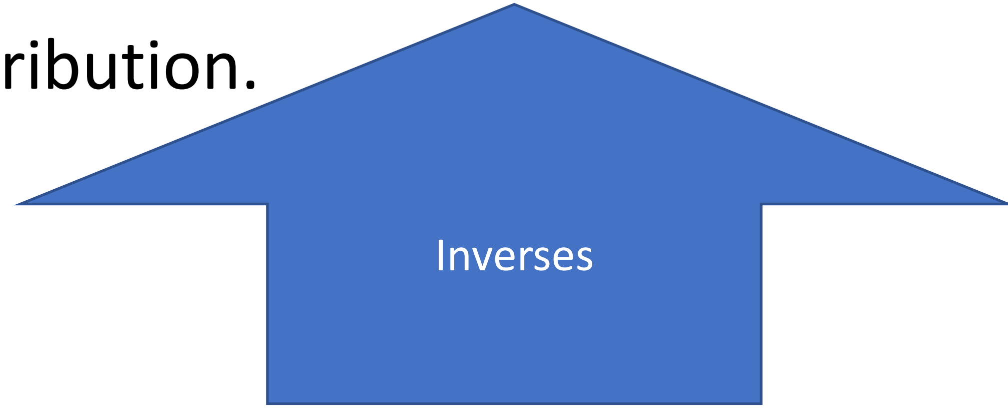
There are 4 functions you must learn

dnorm() – density (height of the Normal).

pnorm() – probability = lower tail area of the Normal up to given y.

qnorm() – quantile (y value) = the value of y with lower tail given.

rnorm() – random sample from a normal distribution.



To find probabilities use pnorm()

Say $Y \sim N(\mu = 10, \sigma = 5)$ find $P(Y > 12)$

Solution:

```
> 1-pnorm(12,10,5)
[1] 0.3445783
>
```

To find quantiles use qnorm()

Say $Y \sim N(\mu = 10, \sigma = 5)$ find y

such that $P(Y \leq y) = 0.5890$

Solution:

```
> qnorm(0.5890, 10, 5)
[1] 11.12487
>
```

Example 5.11

Finding a Value of the Normal Random Variable—Six Sigma Application

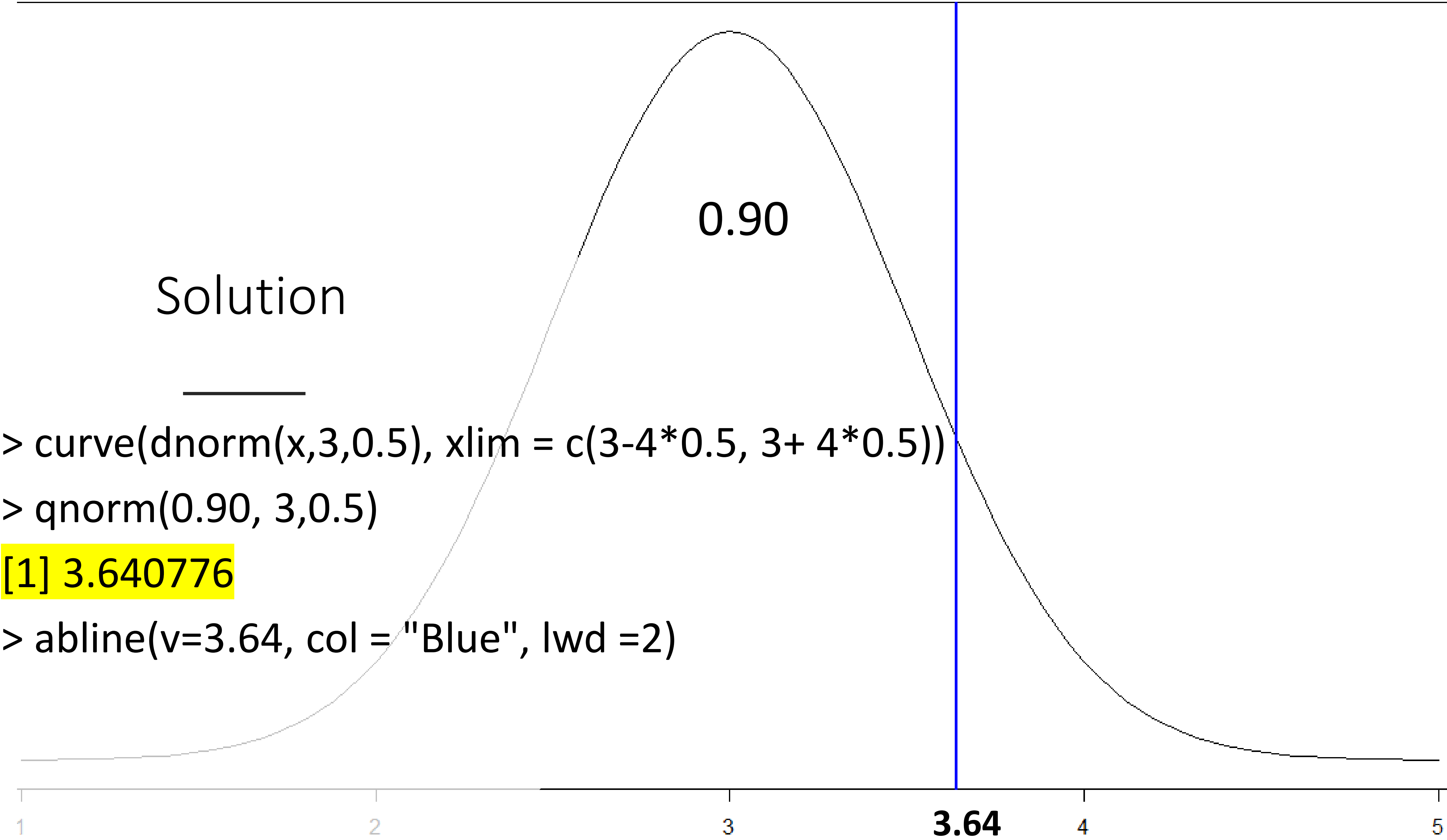
Six Sigma is a comprehensive approach to quality goal setting that involves statistics. The use of the normal distribution in Six Sigma goal setting at Motorola Corp. was demonstrated in *Aircraft Engineering and Aerospace Technology* (Vol. 76, 2004). Motorola discovered that the defect rate, Y, for parts produced on an assembly line varies according to a normal distribution with $\mu = 3$ defects per million and $\sigma = .5$ defect per million. Assume that Motorola's quality engineers want to find a target defect rate, t, such that the actual defect rate will be no greater than t on 90% of the runs. Find the value of t.

Compare with the book

You should use R for these calculations

Solution

- > curve(dnorm(x,3,0.5), xlim = c(3-4*0.5, 3+ 4*0.5))
- > qnorm(0.90, 3,0.5)
- [1] 3.640776
- > abline(v=3.64, col = "Blue", lwd =2)



Assignment 2

Bayes' rule.

Suppose, a particular test for whether someone has been using cannabis is 90% sensitive, meaning the true positive rate (TPR) = 0.90. Therefore, it leads to 90% true positive results (correct identification of drug use) for cannabis users.

The test is also 80% specific, meaning true negative rate (TNR) = 0.80. Therefore, the test correctly identifies 80% of non-use for non-users, but also generates 20% false positives, or false positive rate (FPR) = 0.20, for non-users.

Assuming 0.05 prevalence, meaning 5% of people use cannabis, what is the probability that a random person who tests positive is really a cannabis user?

Solution: What we want $P(U|+)$!

Notice the preliminary calcs

$$P(U) = 0.05, P(U^c) = 0.95$$
$$P(+|U) = 0.90, P(+|U^c) = 1 - 0.80 = 0.20$$

$$P(U|+) = \frac{P(+|U)P(U)}{P(+|U)P(U) + P(+|U^c)P(U^c)} = \frac{0.90X0.05}{0.90X0.05 + 0.20X0.95}$$

$$= \frac{0.045}{0.045 + 0.19} = 0.19$$

Assignment 2 Bayes

c. First, we find $P(I)=\frac{100}{1000}=0.1,\ P\left(I^{c}\right)=1-P(I)=1-0.5=0.5,$

$P\left(P \mid I^{c}\right)=1-P\left(P^{c} \mid I^{c}\right)=1-0.99=0.01$

$$P(I \mid P)=\frac{P(P \mid I)P(I)}{P(P \mid I)P(I)+P\left(P \mid I^{c}\right)P\left(I^{c}\right)}=\frac{0.5(0.1)}{0.5(0.1)+0.01(0.9)}$$

$$=\frac{0.05}{0.05+0.009}=\frac{0.05}{0.059}=0.8475$$

Determining Whether the Data Are from an Approximately Normal Distribution

1. Construct either a **histogram** or a **stem-and-leaf** display for the data. If the data are approximately normal, the shape of the graph will be similar to the normal curve, Figure 5.12 (i.e., mound-shaped and symmetric around the mean with thin tails).
2. Find the **interquartile range, IQR**, and **standard deviation**, s, for the sample, then calculate the ratio IQR/s. If the data are approximately normal, then $IQR/s \approx 1.3$.
3. Construct a **normal probability plot** for the data. (See the following example.) If the data are approximately normal, the points will fall (approximately) on a straight line.

Definition 5.5

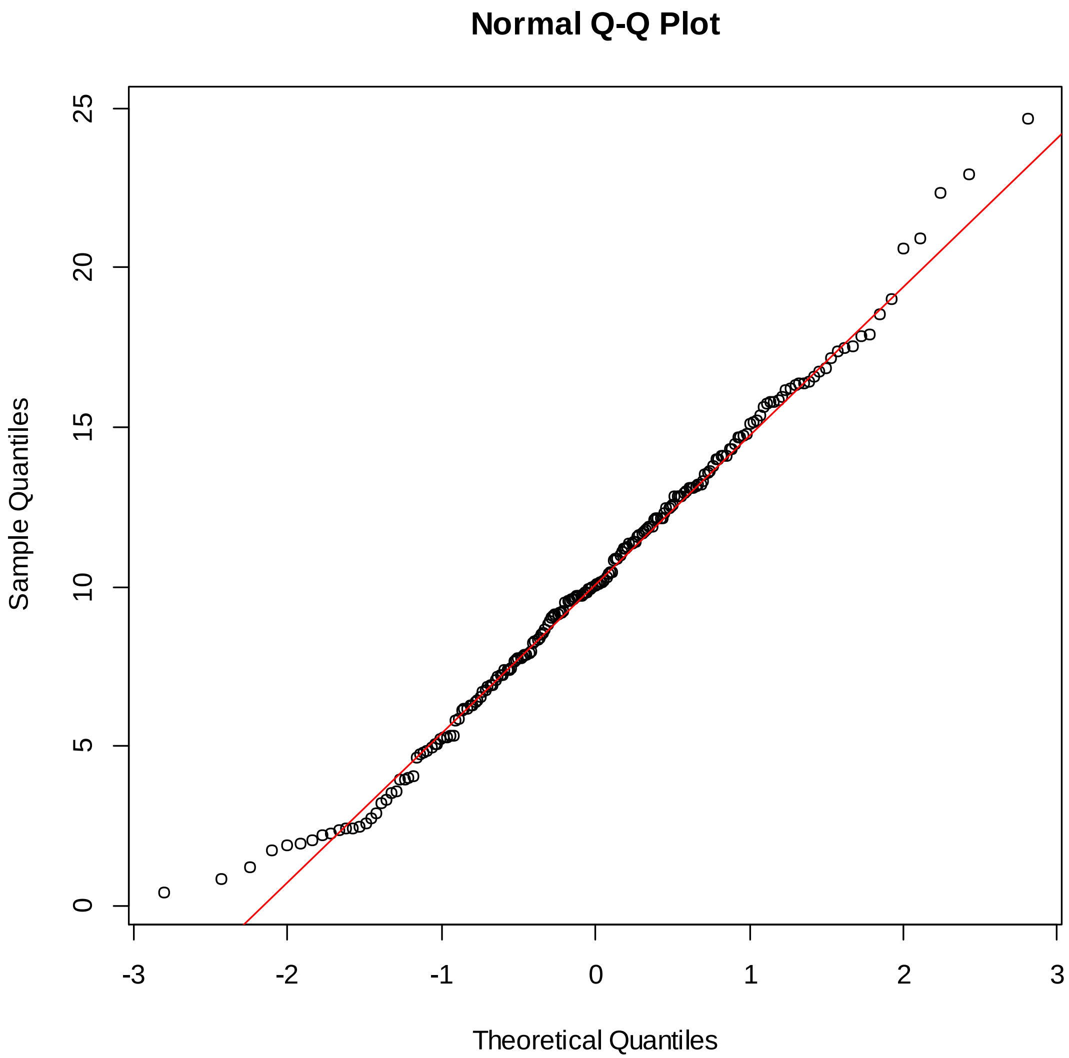
A **normal probability plot** for a data set is a scatterplot with the ranked data values on one axis and their corresponding expected *Z*-scores from a standard normal distribution on the other axis. (*Note*: Computation of the expected standard normal *Z*-scores is beyond the scope of this text. Therefore, we will rely on available statistical software packages to generate a normal probability plot.)

How to test for Normality

Inter-quartile range to standard deviation

$$\frac{IQR}{\sigma} \approx 1.3$$

QQplot



Proof

$$X_{0.75} = \mu + Z_{0.75}\sigma$$
$$X_{0.25} = \mu + Z_{0.25}\sigma$$
$$IQR = X_{0.75} - X_{0.25} = \mu + Z_{0.75}\sigma - \mu - Z_{0.25}\sigma$$
$$IQR = Z_{0.75}\sigma - Z_{0.25}\sigma = (Z_{0.75} - Z_{0.25})\sigma$$
$$\frac{IQR}{\sigma} = \frac{(Z_{0.75} - Z_{0.25})\sigma}{\sigma} = Z_{0.75} - Z_{0.25} = qnorm(0.75) - qnorm(0.25)$$

```
> qnorm(0.75)-qnorm(0.25)
[1] 1.34898
```

The Gamma Probability Distribution

The probability density function for a gamma-type random variable Y is given by

$$f(y) = \begin{cases} \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & \text{if } 0 \le y < \infty;\ \alpha > 0;\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}\,dy$$

The mean and variance of a gamma-type random variable are, respectively,

$$\mu = \alpha\beta \qquad \sigma^2 = \alpha\beta^2$$

It can be shown (proof omitted) that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ and that $\Gamma(\alpha) = (\alpha - 1)!$ when α is a positive integer. Values of $\Gamma(\alpha)$ for $1.0 \le \alpha \le 2.0$ are presented in Table 6 of Appendix B.

the intersection at positive integers, both are valid analytic continuations of the factorials to the non-integers

Main definition [edit]

The notation $\Gamma(z)$ is due to Legendre.[1] If the real part of the complex number z is strictly positive ($\Re(z) > 0$), then the integral

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

converges absolutely, and is known as the **Euler integral of the second kind**. (Euler's integral of the first kind is the beta function.[1]) Using integration by parts, one sees that:

$$\begin{aligned}\Gamma(z+1) &= \int_0^\infty x^{z} e^{-x}\, dx \\ &= \Big[-x^z e^{-x}\Big]_0^\infty + \int_0^\infty z x^{z-1} e^{-x}\, dx \\ &= \lim_{x\to\infty}\left(-x^z e^{-x}\right) - \left(-0^z e^{-0}\right) + z\int_0^\infty x^{z-1} e^{-x}\, dx.\end{aligned}$$

Recognizing that $-x^z e^{-x} \to 0$ as $x \to \infty$,

$$\begin{aligned}\Gamma(z+1) &= z\int_0^\infty x^{z-1} e^{-x}\, dx \\ &= z\Gamma(z).\end{aligned}$$

We can calculate $\Gamma(1)$:

$$\begin{aligned}\Gamma(1) &= \int_0^\infty x^{1-1} e^{-x}\, dx \\ &= \Big[-e^{-x}\Big]_0^\infty \\ &= \lim_{x\to\infty}\left(-e^{-x}\right) - \left(-e^{-0}\right) \\ &= 0 - (-1) \\ &= 1.\end{aligned}$$

Given that $\Gamma(1) = 1$ and $\Gamma(n+1) = n\Gamma(n)$,

$$\Gamma(n) = 1\cdot 2\cdot 3\cdots (n-1) = (n-1)!$$

for all positive integers n. This can be seen as an example of proof by induction.

The identity $\Gamma(z) = \frac{\Gamma(z+1)}{z}$ can be used (or, yielding the same result, analytic continuation can be used) to uniquely extend the integral formulation for $\Gamma(z)$ to a meromorphic function defined for all complex numbers z, except integers less than or equal to zero.[1] It is this extended version that is commonly referred to as the gamma function.[1]

Alternative definitions [edit]

Check parameterization with R ?dgamma

Details

If `scale` is omitted, it assumes the default value of `1`.

The Gamma distribution with parameters `shape` = *a* and `scale` = *s* has density

f(x)= 1/(s^a Gamma(a)) x^(a-1) e^-(x/s)

for *x ≥ 0*, *a > 0* and *s > 0*. (Here *Gamma(a)* is the function implemented by R's `gamma()` and defined in its help. Note that *a = 0* corresponds to the trivial distribution with all mass at point 0.)

The mean and variance are *E(X) = a*s* and *Var(X) = a*s^2*.

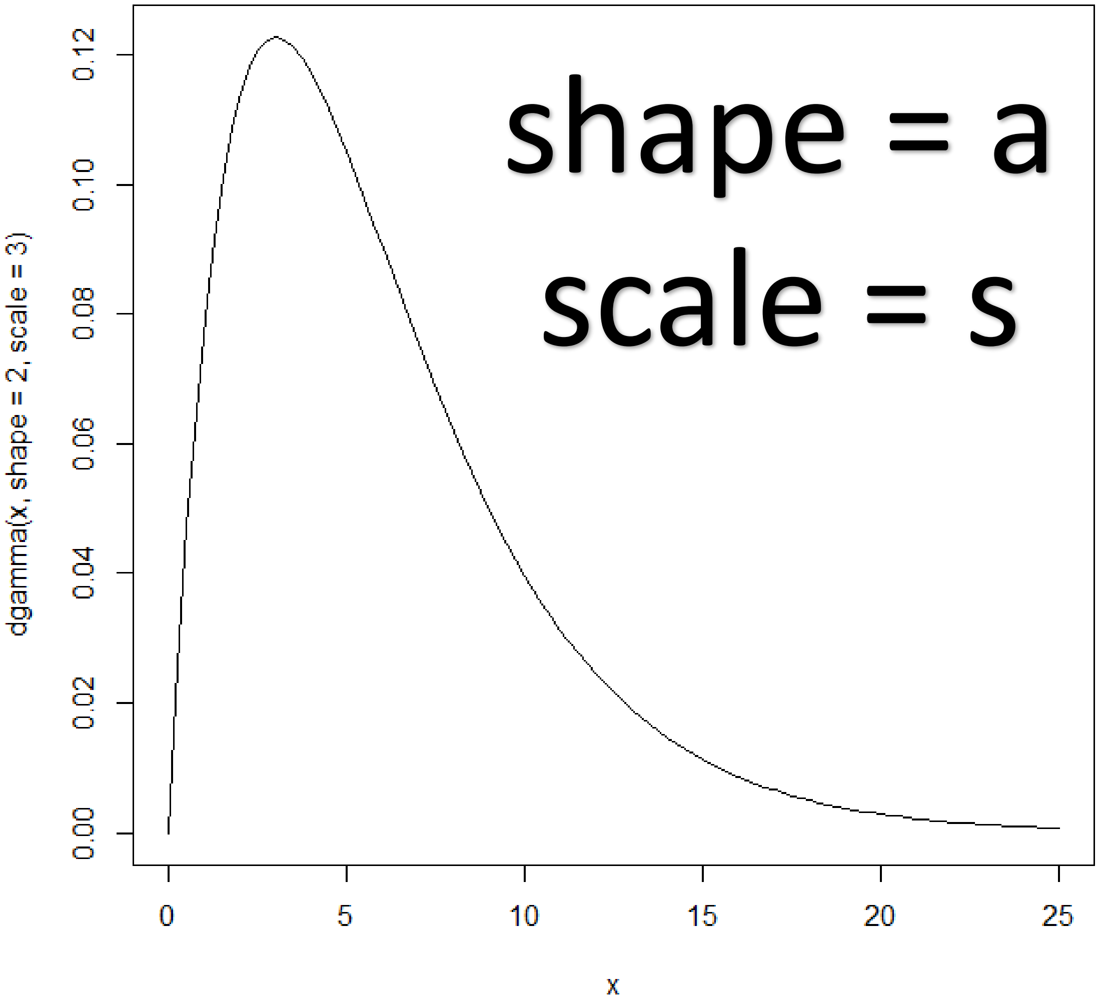
The cumulative hazard *H(t) = - log(1 - F(t))* is

```
-pgamma(t, ..., lower = FALSE, log = TRUE)
```

Note that for smallish values of `shape` (and moderate `scale`) a large parts of the mass of the Gamma distribution is on values of *x* so near zero that they will be represented as zero in computer arithmetic. `rgamma` may well return values which will be represented as zero. (This will also happen for very large values of `scale` since the actual generation is done for `scale = 1`.)

$$\alpha = a$$
$$\beta = s$$

- dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
- pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)
- qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,log.p = FALSE)
- rgamma(n, shape, rate = 1, scale = 1/rate)



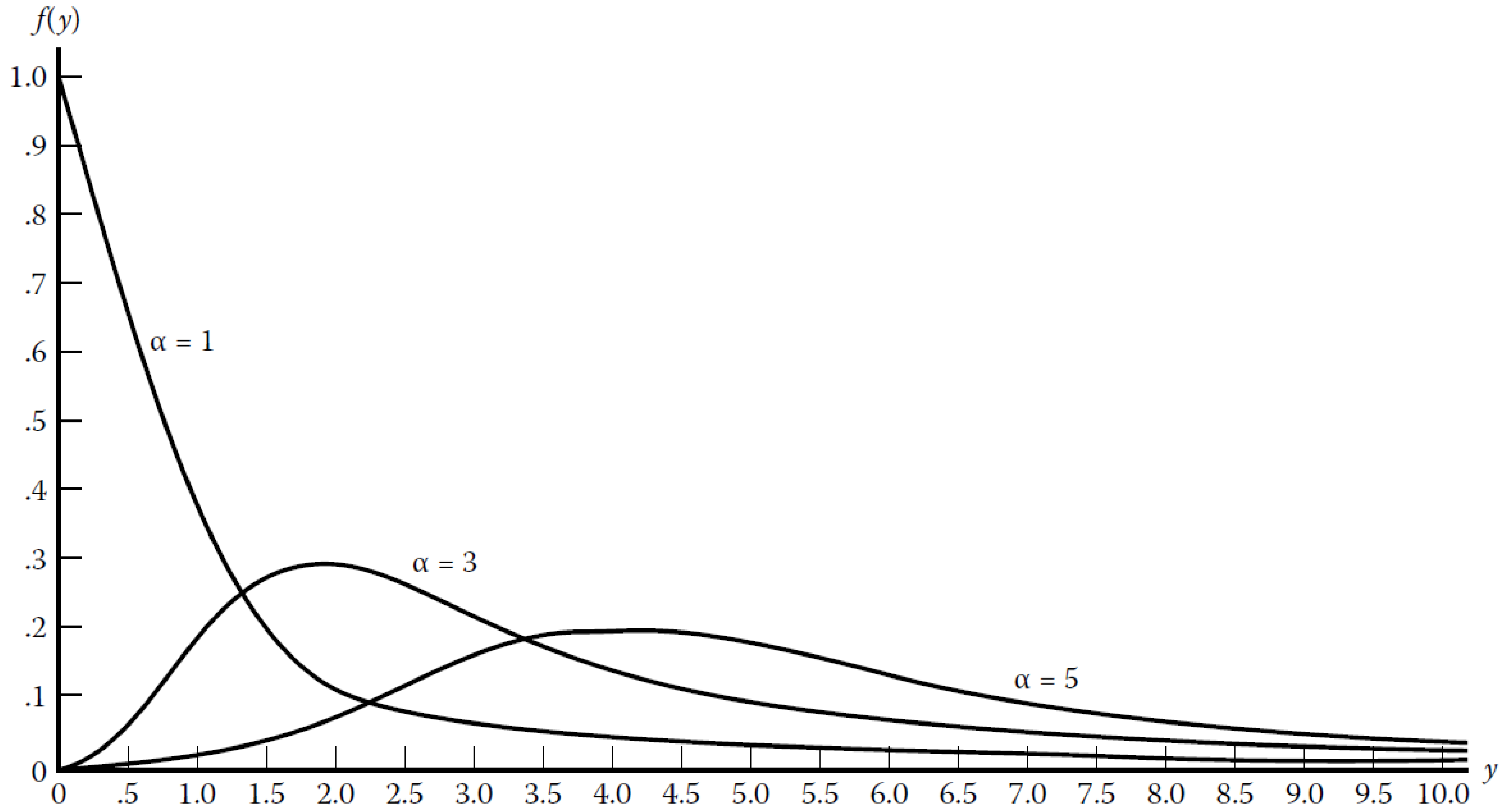


FIGURE 5.19

Graphs of gamma density functions for $\alpha = 1, 3$, and 5; $\beta = 1$

The Chi-Square Probability Distribution

A **chi-square random variable** is a gamma-type random variable Y with $\alpha = \nu/2$ and $\beta = 2$:

$$f(y) = c(y)^{(\nu/2)-1}e^{-y/2} \quad (0 \le y^2 < \infty)$$

where

$$c = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}$$

The mean and variance of a chi-square random variable are, respectively,

$$\mu = \nu \qquad \sigma^2 = 2\nu$$

The parameter ν is called the **number of degrees of freedom** for the chi-square distribution.

In R

$x=y$

$n=V$

?dchisq

- The chi-squared distribution with df= $n \geq 0$ degrees of freedom has density

- $f_n(x) = 1 / (2^{(n/2)} \Gamma(n/2)) x^{(n/2-1)} e^{(-x/2)}$

- for $x > 0$, where $f_0(x) := \lim_{n \to 0} f_n(x) = \delta_0(x)$, a point mass at zero, is not a density function proper, but a "δ distribution".
- The mean and variance are n and $2n$.

The Exponential Probability Distribution

An **exponential distribution** is a gamma density function with $\alpha = 1$:

$$f(y) = \frac{e^{-y/\beta}}{\beta} \qquad (0 \leq y < \infty)$$

with mean and variance

$$\mu = \beta \qquad \sigma^2 = \beta^2$$

In R the density is:

f(x) = λ {e}^{- λ x}

$$\lambda = \frac{1}{\beta}$$

$$\beta = \frac{1}{\lambda}$$

If $Y \sim Exp(\beta = 3)$ Find $P(Y \leq 4)$

Answer:

First notice that $\lambda = 1/3$

So $P(Y \leq 4) = pexp(4, \frac{1}{3})$

```
> pexp(4,1/3)
[1] 0.7364029
```

Example 5.13

Gamma Distribution Application—Customer Complaints

From past experience, a manufacturer knows that the relative frequency distribution of the length of time Y (in months) between major customer product complaints can be modeled by a gamma density function with $\alpha = 2$ and $\beta = 4$. Fifteen months after the manufacturer tightened its quality control requirements, the first complaint arrived. Does this suggest that the mean time between major customer complaints may have increased?

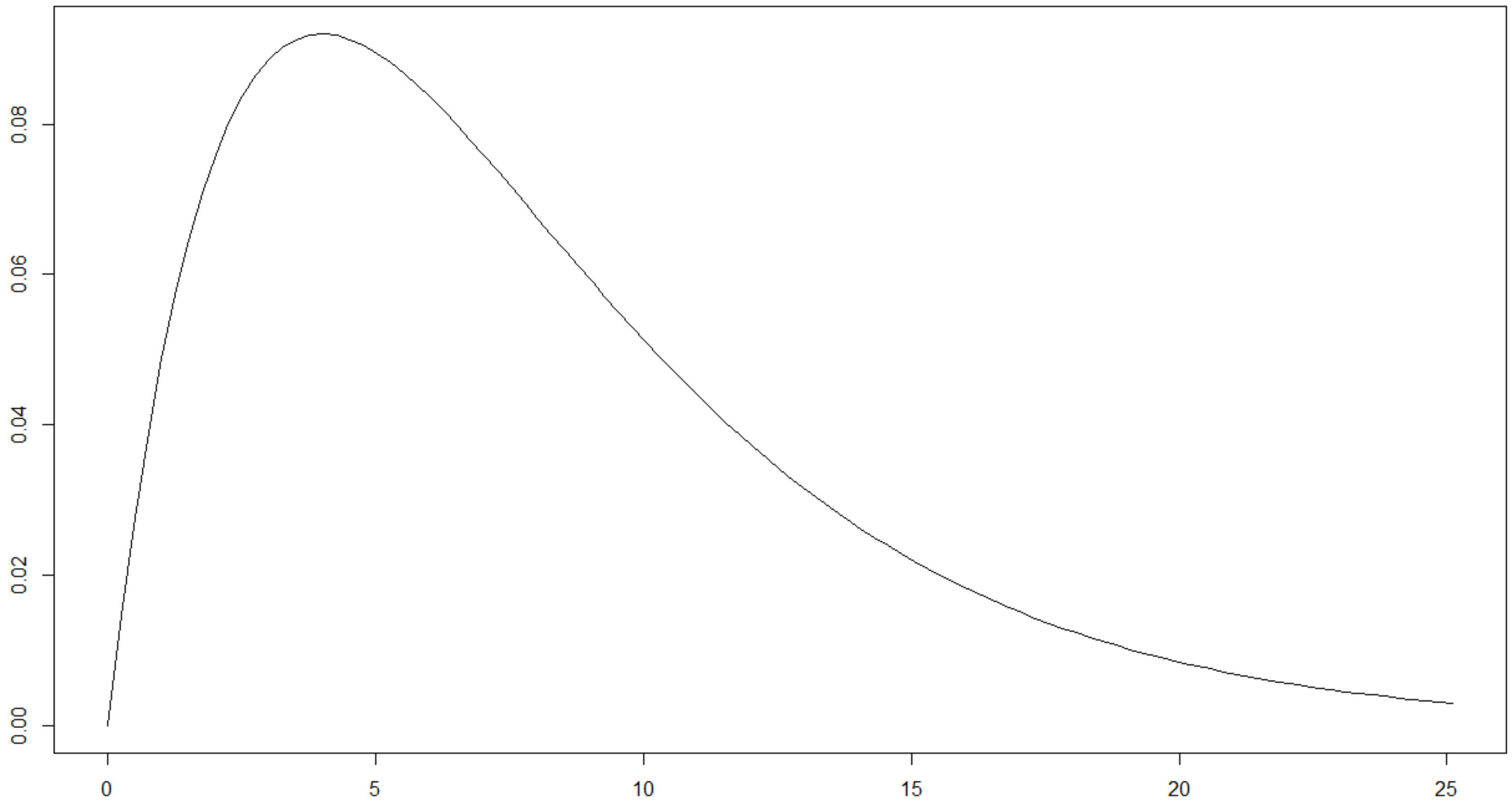
To solve this we need to restate the problem.
With $\alpha = 2, \beta = 4$ would 15 or more months till the first complaint be unlikely?
We can calculate this is R
mean = 2*4, sigma2 = 2*4^2, sigma = 5.7 (mean and variance of gamma)

```
> curve(dgamma(x,shape = 2, scale = 4), xlim = c(0, 8+3*5.7)) # – see plot next slide
> 1-pgamma(15, shape=2,scale =4)
[1] 0.1117093
```

$$P(Y \geq 15) = 0.11$$

This is quite likely so

Time between customer complaints



Time between complaints

The Weibull Probability Distribution

The probability density function for a Weibull random variable, Y is given by

$$f(y) = \begin{cases} \frac{\alpha}{\beta} y^{\alpha-1} e^{-y^{\alpha}/\beta} & \text{if } 0 \leq y < \infty;\ \alpha > 0;\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\mu = \beta^{1/\alpha} \Gamma\left(\frac{\alpha + 1}{\alpha}\right)$$

$$\sigma^2 = \beta^{2/\alpha}\left[\Gamma\left(\frac{\alpha + 2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha + 1}{\alpha}\right)\right]$$

Compare with R documentation ?dweibull

Details

The Weibull distribution with `shape` parameter *a* and `scale` parameter *b* has density given by

f(x) = (a/b) (x/b)^(a-1) exp(- (x/b)^a)

for *x > 0*. The cumulative distribution function is *F(x) = 1 - exp(- (x/b)^a)* on *x > 0*, the mean is *E(X) = b Γ(1 + 1/a)*, and the *Var(X) = b^2 * (Γ(1 + 2/a) - (Γ(1 + 1/a))^2)*.

The Weibull Probability Distribution

The probability density function for a Weibull random variable, Y is given by

$$f(y) = \begin{cases} \frac{\alpha}{\beta} y^{\alpha-1} e^{-y^\alpha/\beta} & \text{if } 0 \le y < \infty;\ \alpha > 0;\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\mu = \beta^{1/\alpha}\Gamma\left(\frac{\alpha+1}{\alpha}\right)$$

$$\sigma^2 = \beta^{2/\alpha}\left[\Gamma\left(\frac{\alpha+2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha+1}{\alpha}\right)\right]$$

$$\alpha = a$$

$$\beta = b^a$$

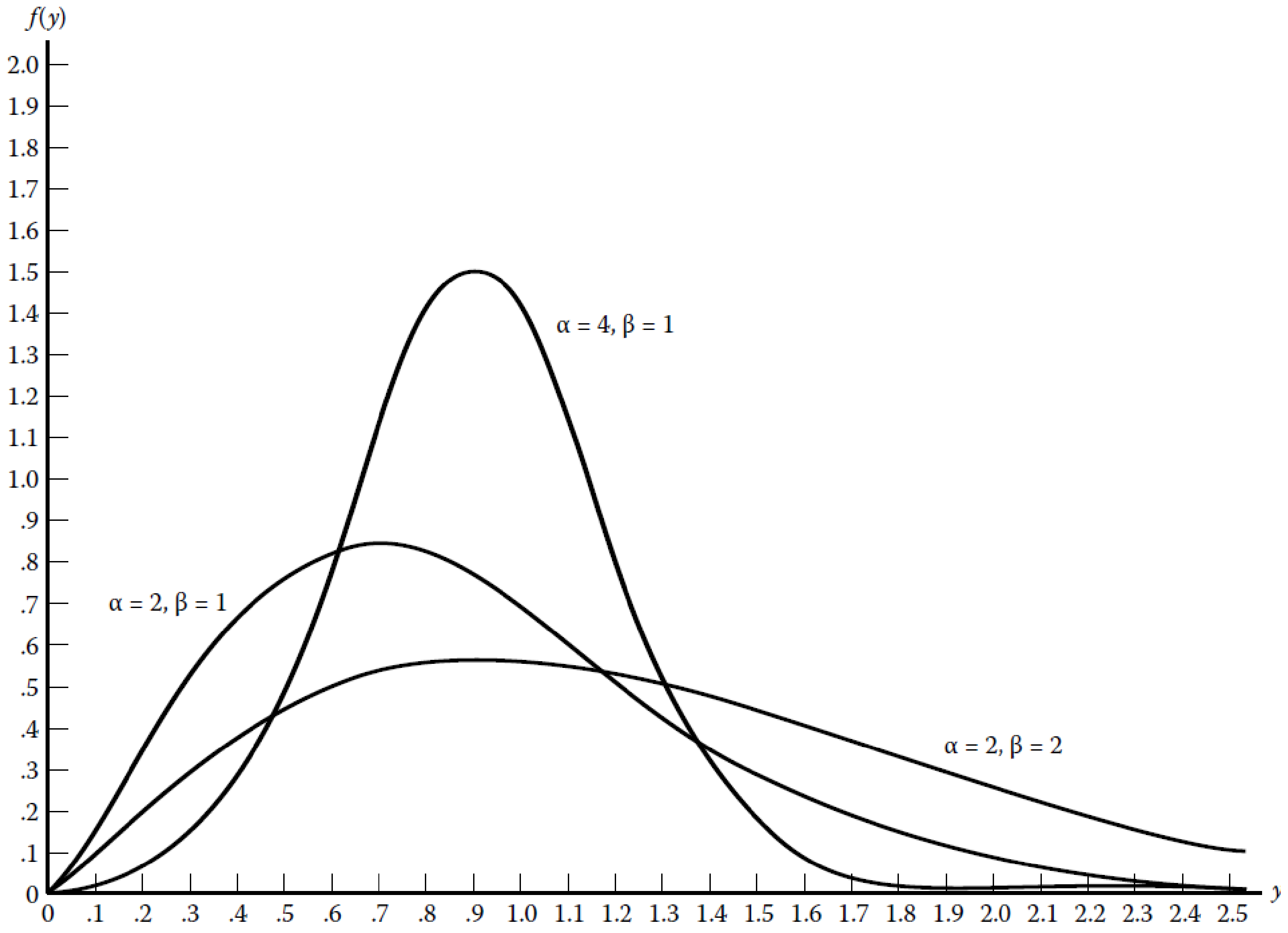


FIGURE 5.20

Graphs of Weibull density functions

Example 5.15

Weibull Distribution Application—Drill Bit Failure

The length of life Y (in hours) of a drill bit used in a manufacturing operation has a Weibull distribution with $\alpha = 2$ and $\beta = 100$. Find the probability that a drill bit will fail before 8 hours of usage.

Solution

The cumulative distribution function for a Weibull distribution is

$$F(y_0) = \int_0^{y_0} f(y)\,dy = \int_0^{y_0} \frac{\alpha}{\beta} y^{\alpha-1} e^{-y^\alpha/\beta} dy$$

By making the transformation $Z = Y^\alpha$, we have $dz = \alpha y^{\alpha-1} dy$ and the integral reduces to

$$F(y_0) = 1 - e^{-z/\beta} = 1 - e^{-y_0^\alpha/\beta}$$

To find the probability that y is less than 8 hours, we calculate

$$P(Y < 8) = F(8) = 1 - e^{-(8)^\alpha/\beta}$$

$$= 1 - e^{-(8)^2/100} = 1 - e^{-.64}$$

Interpolating between $e^{-.60}$ and $e^{-.65}$ in Table 3 of Appendix B or using a calculator with the e function, we find $e^{-.64} \approx .527$. Therefore, the probability that a drill bit will fail before 8 hours is

$$P(Y < 8) = 1 - e^{-.64} = 1 - .527 = .473$$

Using R to find $P(Y < 8)$

parameterization for R

$$\alpha = a, 2 = a,$$
$$\beta = b^{a}, 100 = b^{2}, b = 10$$

> pweibull(8, 2, 10)

[1] 0.4727076

The Beta Probability Distribution

The probability density function for a beta-type random variable Y is given by

$$f(y) = \begin{cases} \dfrac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)} & \text{if } 0 \leq y \leq 1;\ \alpha > 0;\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

where

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The mean and variance of a beta random variable are, respectively,

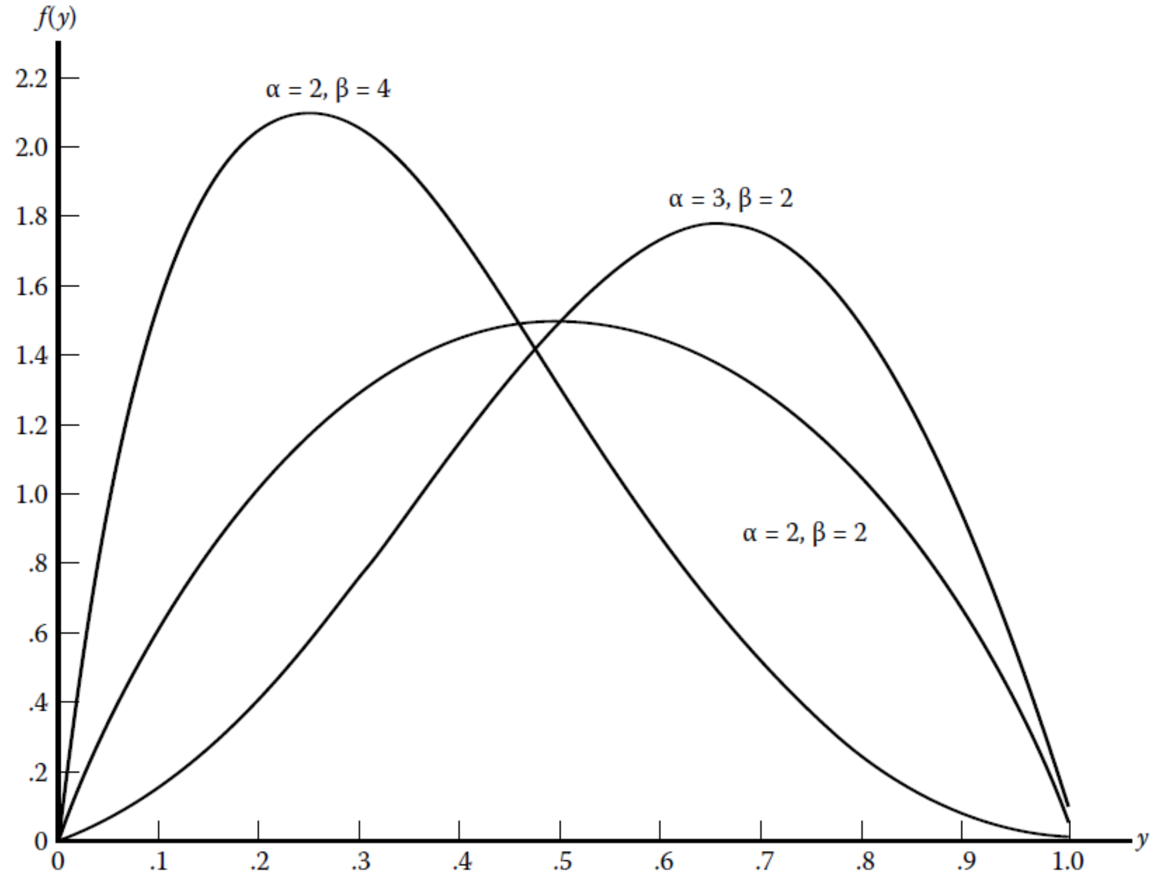
$$\mu = \frac{\alpha}{\alpha+\beta} \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

[Recall that

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$$

and $\Gamma(\alpha) = (\alpha - 1)!$ when α is a positive integer.]

FIGURE 5.21
Graphs of beta density functions



Example 5.17

Beta Distribution Application—Robotic Sensors

Infrared sensors in a computerized robotic system send information to other sensors in different formats. The percentage Y of signals sent that are directly compatible for all sensors in the system follows a beta distribution with $\alpha = \beta = 2$.

a. Find the probability that more than 30% of the infrared signals sent in the system are directly compatible for all sensors.

b. Find the mean and variance of Y.

$P(Y > 0.30)$ = 1-pbeta(0.30,2,2)
[1] 0.784

mean = 2/(2+2)=1/2
variance = 2*2/((2+2)^2(2+2+1)) = 4/(16*5)

Given without proof: BUT we don't need it!

See example

The cumulative distribution function $F(y)$ of a beta density function is called an **incomplete beta function**. Values of this function for various values of y, α, and β are given in *Tables of the Incomplete Beta Function* (1956). For the special case where α and β are integers, it can be shown that

$$F(p) = \int_0^p \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}\, dy = \sum_{y=\alpha}^{n} p(y)$$

where $p(y)$ is a binomial probability distribution with parameters p and $n = (\alpha + \beta - 1)$.

Example 5.18

Using the Binomial Distribution to Find a Beta Probability

Data collected over time on the utilization of a computer core (as a proportion of the total capacity) were found to possess a relative frequency distribution that could be approximated by a beta density function with $\alpha = 2$ and $\beta = 4$. Find the probability that the proportion of the core being used at any particular time will be less than .20.

We need $P(Y < 0.20) = pbeta(0.2, 2, 4)$

```
> pbeta(0.2,2,4)
[1] 0.26272
```

Quick Review

Key Terms

Note: Starred (*) terms are from the optional section in this chapter.

LANGUAGE LAB

Symbol	Pronunciation	Description
$f(y)$	f of y	Probability density function for a continuous random variable Y
$F(y)$	cap F of y	Cumulative distribution function for a continuous random variable Y
$\Gamma(\alpha)$	gamma of alpha	Gamma function for a positive integer α

Chapter Summary

- Properties of a **density function** for a continuous random variable Y:
  (1) $f(y) \geq 0$, (2) $F(\infty) = 1$, $P(a < Y < b) = F(b) - F(a)$
- Types of continuous random variables: **uniform, normal, gamma-type, Weibull**, and **beta-type**.
- **Uniform probability distribution** is a model for continuous random variables that are evenly distributed over a certain interval.
- **Normal** (or **Gaussian**) **probability distribution** is a model for continuous random variables that have a bell-shaped curve with thin tails.
- Descriptive methods for assessing normality: **histogram, stem-and-leaf display**, **IQR/s $\approx$ 1.3**, and **normal probability plot**.
- **Gamma-type probability distribution** is a model for continuous random variables that are lifelengths or waiting times.
- Two special types of gamma random variables are: **chi-square** random variables and **exponential** random variables.
- **Weibull probability distribution** is a model for continuous random variables that represent failure times.
- **Beta-type probability distribution** is a model for continuous random variables that fall in the interval 0 to 1.

Question 4

4 pts

Bayes' Theorem

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_i P(A_i)P(B|A_i)}$$

A particular test for whether someone has been using cannabis is 93% sensitive and 86% specific, meaning it leads to 93% true "positive" results (meaning, "Yes he used cannabis") for cannabis users and 86% true negative results for non-users. Assuming 6% of people actually do use cannabis, what is the probability that a random person who tests positive is really a cannabis user?

THE ANSWER MUST BE AS A DECIMAL (NOT A PERCENTAGE) TO 4 decimal places

Working

Show the correct formula with correct notation, calculate all needed quantities and then show substitution into the formula and final decimal answer.

Rock and Roll man!!