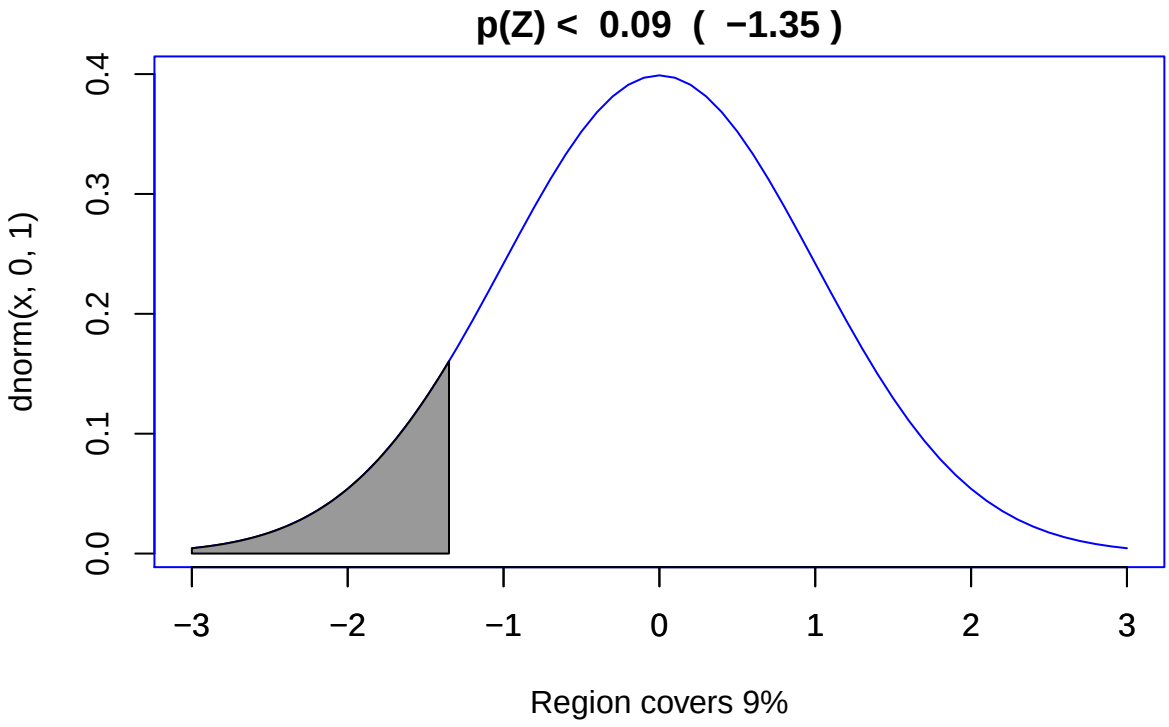
Chapter 4 - Distributions of Random Variables

Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$

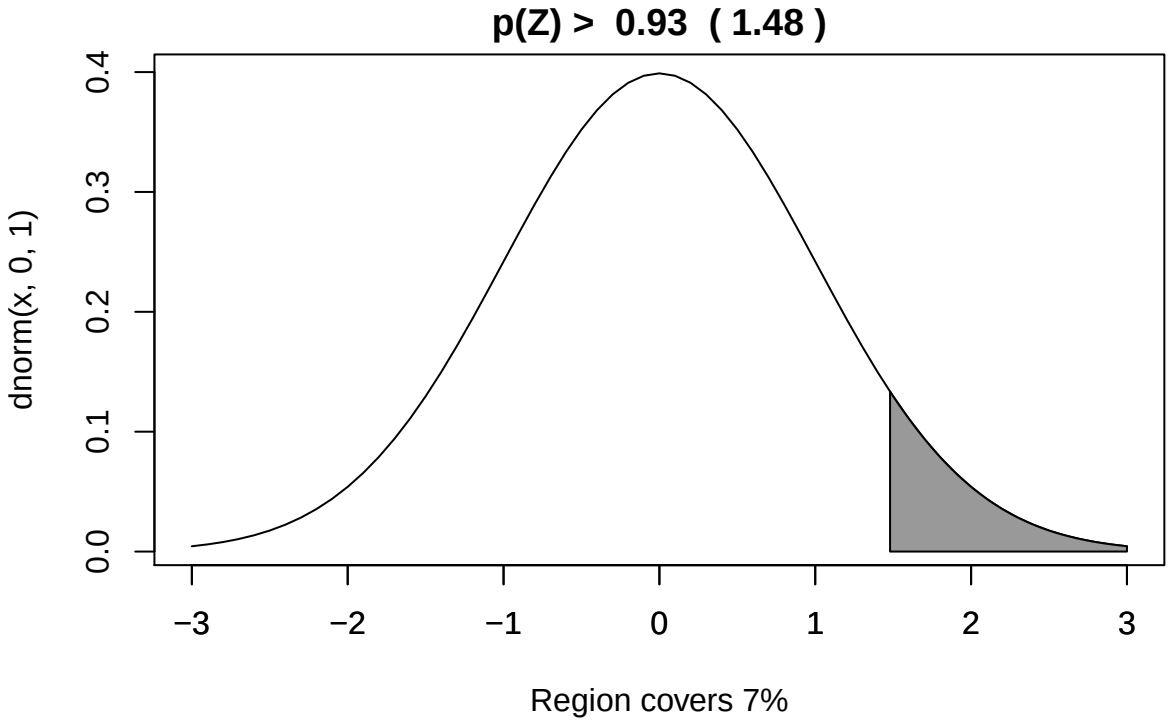
In these questions, we will use pnorm() to return the overall percentage in which the Z-score corresponds to. And we will use dnorm() to draw a graph of the section below or above the Z-score on the normalized curve. Top title displays the pnorm() result. Bottom label answers the question.

```
# save off some parameters
par( col = "blue", xaxp=c(-3,1,3),yaxp=c(0,1,4) )
x<-seq(-3, 3,0.1)

z1<- -3
z2<- -1.35
z_title<-paste("p(Z) < ", format(pnorm(z2), digits=0, nsmall=2),
               " ( ", format(z2,digits=0,nsmall=2), ")")

plot(x = x,  y = dnorm(x,0,1),type = "l", xlab ="Region covers 9%")

cord.1x<- c(z1,seq(z1,z2,0.01),z2)
cord.1y<- c(0,dnorm(seq(z1,z2,0.01)),0)
title(main=z_title, cex.main = 1.2, line=0.5, cex.lab=2)
polygon(cord.1x,cord.1y,col='grey60')
Axis(side=1,at=seq(-3, 3, by = 1))
```



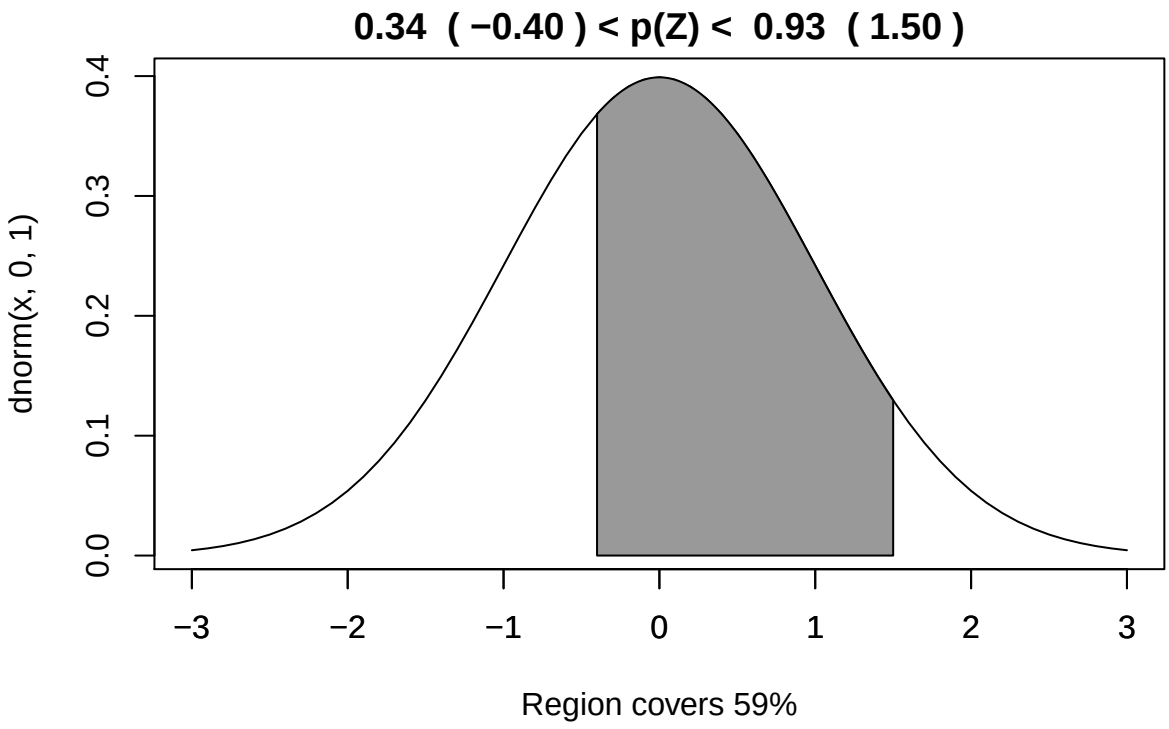
(b) $Z > 1.48$

```
z1<- 1.48
z2<- 3
z_title<-paste("p(Z) > ", format(pnorm(z1), digits=0, nsmall=2),
               " (", format(z1,digits=0,nsmall=2), ")")

plot(x = x,  y = dnorm(x,0,1),type = "l", xlab ="Region covers 7%")

cord.1x<- c(z1,seq(z1,z2,0.01),z2)
cord.1y<- c(0,dnorm(seq(z1,z2,0.01)),0)
title(main=z_title, cex.main = 1.2, line=0.5, cex.lab=2)
polygon(cord.1x,cord.1y,col='grey60')
Axis(side=1,at=seq(-3, 3, by = 1))
```



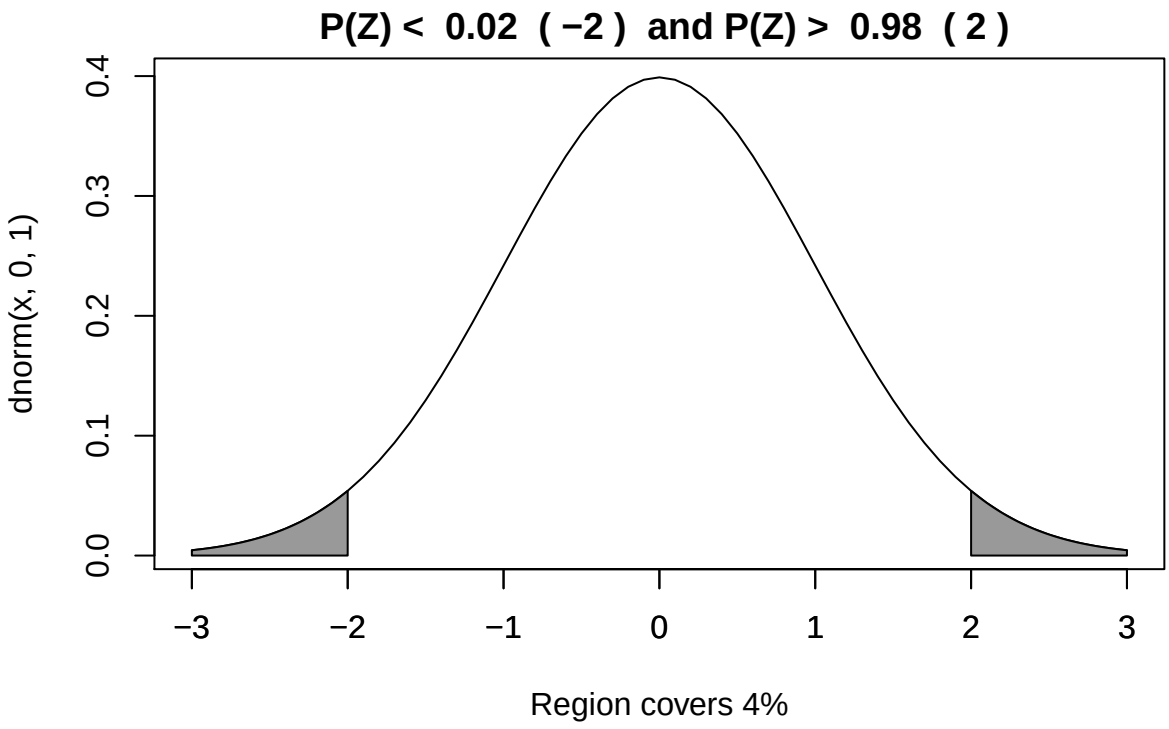
(c) $-0.4 < Z < 1.5$

```
z1<- -0.4
z2<- 1.5
z_title<-paste(format(pnorm(z1), digits=0, nsmall=2),
         " (", format(z1,digits=0,nsmall=2), ") < p(Z) < ",
         format(pnorm(z2), digits=0, nsmall=2), " (", format(z2,digits=0,nsmall=2), ")")

plot(x = x,  y = dnorm(x,0,1),type = "l", xlab ="Region covers 59%")

cord.1x<- c(z1,seq(z1,z2,0.01),z2)
cord.1y<- c(0,dnorm(seq(z1,z2,0.01)),0)
title(main=z_title, cex.main = 1.2, line=0.5, cex.lab=2)
polygon(cord.1x,cord.1y,col='grey60')
Axis(side=1,at=seq(-3, 3, by = 1))
```



(d) $|Z| > 2$

```
z1<- -3
z2<- -2
z_title<-paste(" P(Z) < ", format(pnorm(z2), digits=0, nsmall=2)," (", z2, ")")

plot(x = x,  y = dnorm(x,0,1),type = "l", xlab ="Region covers 4%")
cord.1x<- c(z1,seq(z1,z2,0.01),z2)
cord.1y<- c(0,dnorm(seq(z1,z2,0.01)),0)
polygon(cord.1x,cord.1y,col='grey60')


z1<- 2
z2<- 3

z_title<-paste(z_title, " and P(Z) > ", format(pnorm(z1), digits=0, nsmall=2)," (", z1, ")")

cord.1x<- c(z1,seq(z1,z2,0.01),z2)
cord.1y<- c(0,dnorm(seq(z1,z2,0.01)),0)
polygon(cord.1x,cord.1y,col='grey60')
title(main=z_title, cex.main = 1.2, line=0.5, cex.lab=2)
Axis(side=1,at=seq(-3, 3, by = 1))
```



We can review the results with the Z score lookup from pnorm()

```
z_score_df <- data.frame(Z_Score=-3,cdf_percent=round(pnorm(-3),4))

for (i in c(-2,-1.35, -1,-0.4,0,1,1.48,2,3)) {
    z_score_df <- rbind(z_score_df,data.frame(Z_Score=i,cdf_percent=round(pnorm(i),4)))
}

knitr::kable(z_score_df, caption='Z Score on the CDF')
```

Table 1: Z Score on the CDF

Z_Score	cdf_percent
-3.00	0.0013
-2.00	0.0228
-1.35	0.0885
-1.00	0.1587
-0.40	0.3446
0.00	0.5000
1.00	0.8413
1.48	0.9306
2.00	0.9772
3.00	0.9987

Triathlon times, Part I (4.4, p. 142) In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the *Men, Ages 30 - 34* group while Mary competed in the *Women, Ages 25 - 29* group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

- The finishing times of the *Men, Ages 30 - 34* group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the *Women, Ages 25 - 29* group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

(a) Write down the short-hand for these two normal distributions.

Men, Ages 30 - 34 : $N(\mu = 4313, \sigma = 583)$ Women, Ages 25 - 29 : $N(\mu = 5261, \sigma = 807)$

(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

```
time_leo<-4948
time_mary<-5513
time_mean_leo_group<-4313
time_sd_leo_group<-583

time_mean_mary_group<-5261
time_sd_mary_group<-807

z_score_leo<-(time_leo-time_mean_leo_group)/time_sd_leo_group
z_score_mary<-(time_mary-time_mean_mary_group)/time_sd_mary_group

paste("Leo Z-Score = ", round(z_score_leo,3))
```

```
## [1] "Leo Z-Score =  1.089"
```

```
paste("Mary Z-Score = ", round(z_score_mary,3))
```

```
## [1] "Mary Z-Score =  0.312"
```

Z-Score $= (\bar{x} - \mu)/\sigma$

A higher Z-Score refelects a worse time. The Z-Score is the number of standard deviations above the mean.

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

Leos score would be considered worse. He performed worse than 85% which can be deduced from pnorm(1.08).

```
z_cdf_pct_leo<-pnorm(z_score_leo)
z_cdf_pct_mary<-pnorm(z_score_mary)


sprintf("Leo did worse than = %.2f%s",  round(z_cdf_pct_leo,2),'%')
```

```
## [1] "Leo did worse than = 0.86%"
```

```
sprintf("Mary did worse than = %.2f%s",  round(z_cdf_pct_mary,2),'%')
```

```
## [1] "Mary did worse than = 0.62%"
```

(d) What percent of the triathletes did Leo finish faster than in his group?

```
sprintf("Leo did better than = %.2f%s",  round((1 - z_cdf_pct_leo) * 100,2),'%')
```

```
## [1] "Leo did better than = 13.80%"
```

(e) What percent of the triathletes did Mary finish faster than in her group?

```
sprintf("Mary did better than = %.2f%s",  round((1 - z_cdf_pct_mary)*100,2),'%')
```

```
## [1] "Mary did better than = 37.74%"
```

(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

In a normal distribution, the Z-Score and the Percentage given are accurate responses to the question, what percentage did Leo or Mary do better or worse than?

If the distribution were not normal than the standard deviation above and below the mean would not be the same, so in that case, a more analytical approach must be taken.

Heights of female college students Below are heights of 25 female college students.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

(a) The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.

```
heights <- c(54, 55, 56, 56, 57, 58, 58, 59, 60, 60, 60, 61,
             61, 62, 62, 63, 63, 63, 64, 65, 65, 67, 67, 69, 73)

sd_heights<-sd(heights)
sd_plus1<-mean(heights) + sd_heights
sd_minus1<-mean(heights) - sd_heights

sd_plus2<-sd_plus1 + sd_heights
sd_minus2<-sd_minus1 - sd_heights

sd_plus3<-sd_plus2 + sd_heights
sd_minus3<-sd_minus2 - sd_heights

sd_1_pct<-length(subset(heights, heights > sd_minus1 & heights < sd_plus1))/length(heights)
sd_2_pct<-length(subset(heights, heights > sd_minus2 & heights < sd_plus2))/length(heights)
sd_3_pct<-length(subset(heights, heights > sd_minus3 & heights < sd_plus3))/length(heights)



sprintf("The percentage of heights that fall with in 1 standard deviations (%.2f andd %.1f) is %.2f%s",
          sd_plus1, sd_minus1, sd_1_pct * 100,'%' )
```

```
## [1] "The percentage of heights that fall with in 1 standard deviations (66.10 andd 56.9) is 68.00%"
```

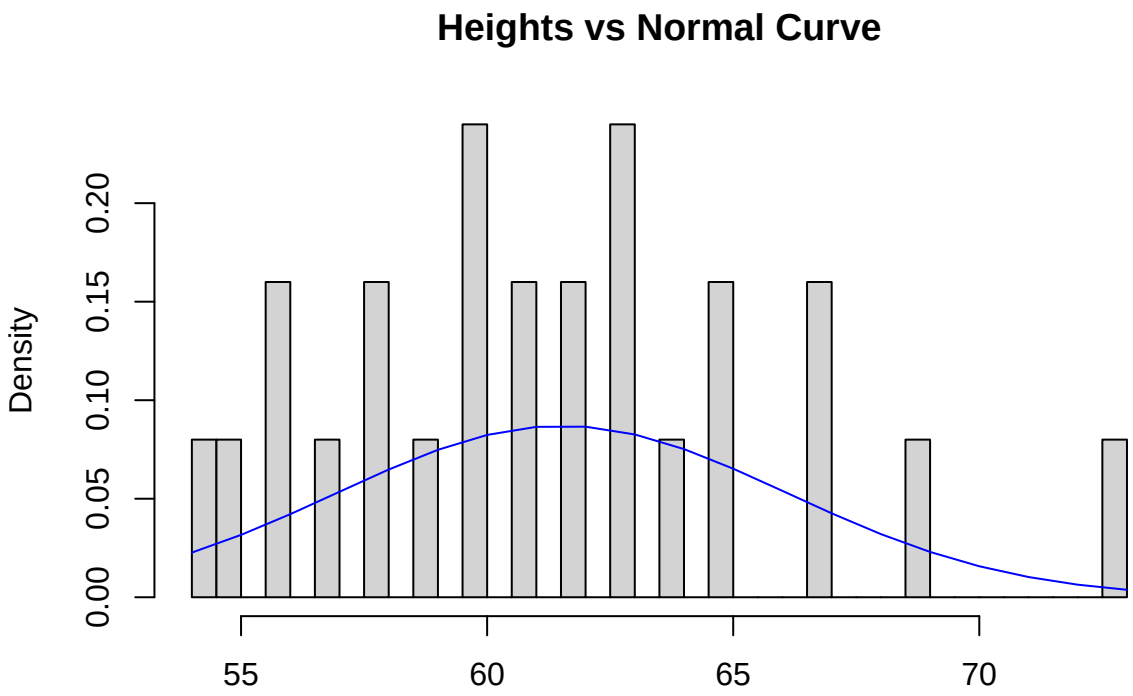
```
sprintf("The percentage of heights that fall with in 2 standard deviations (%.2f andd %.1f) is %.2f%s",
          sd_plus2, sd_minus2, sd_2_pct * 100,'%' )
```

```
## [1] "The percentage of heights that fall with in 2 standard deviations (70.69 andd 52.4) is 96.00%"
```

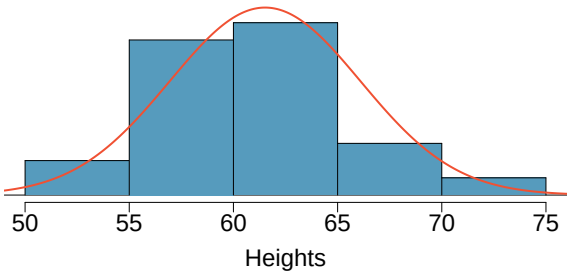
```
sprintf("The percentage of heights that fall with in 3 standard deviations (%.2f andd %.1f) is %.2f%s",
          sd_plus3, sd_minus3, sd_3_pct * 100,'%' )
```

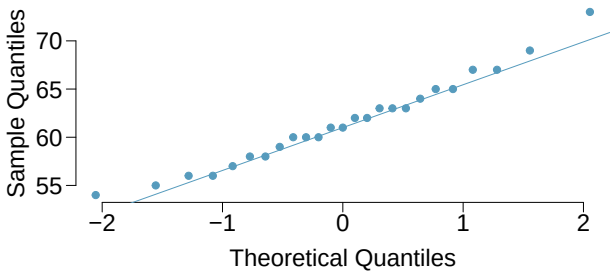
```
## [1] "The percentage of heights that fall with in 3 standard deviations (75.27 andd 47.8) is 100.00%"
```

```
hist(heights, main="Heights vs Normal Curve", probability = TRUE,   breaks=30, xlab="")

# now create a totally seperate set of x,y values
x_values<-seq(min(heights),max(heights),by =1)
y_values<-dnorm(x = x_values, mean = mean(heights), sd = sd(heights))

lines(x = x_values, y = y_values, col = "blue")                    # add the line
```



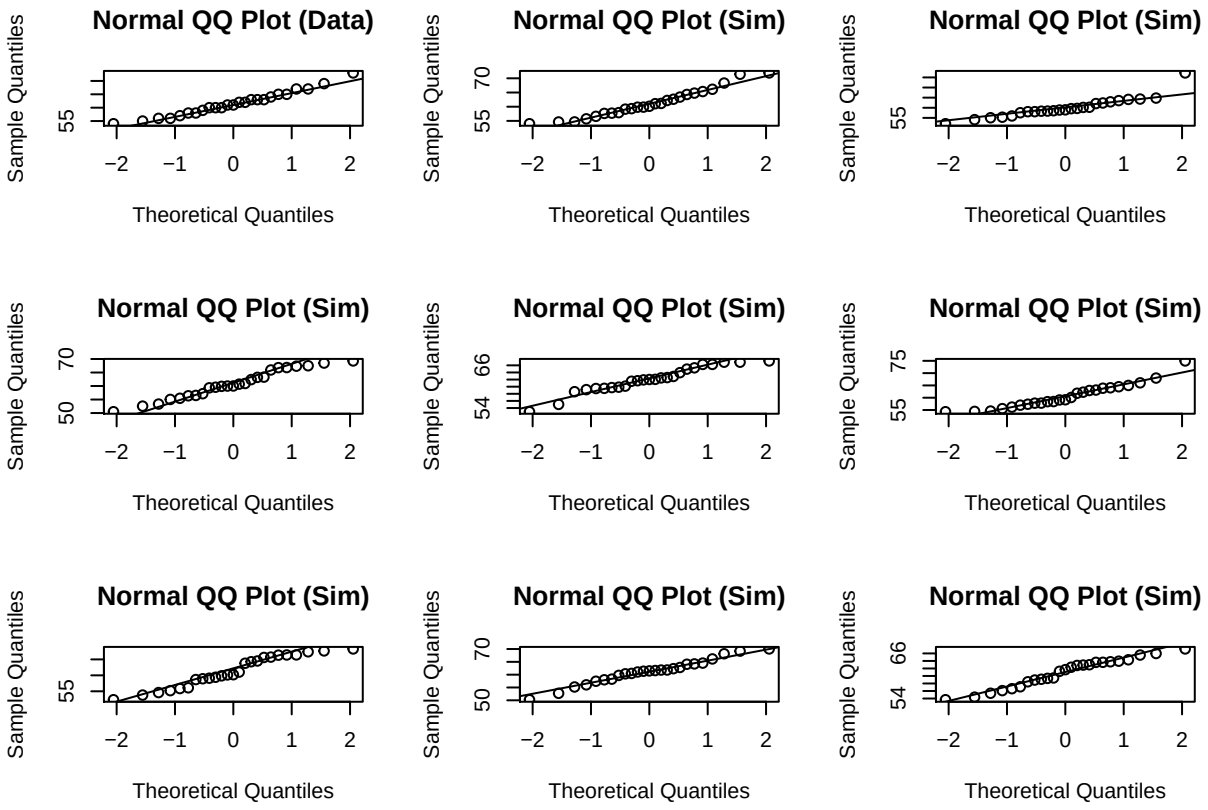
(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.





The qqnorm and qqline function generate a comparison between the heights vector and a normalized curve. If the sample distribution were normally distributed the plot would be a centered diagonal line.

```
# Use the DATA606::qqnormsim function

DATA606::qqnormsim(heights)
```



The qqnormsim function generates 8 plots by calling rnorm() to generate 8 distributions. A data frame is populated and then ggplot is called with stat_qq() to generate the qqplot and facet_wrap to display the grids side by side.

Defective rate. (4.14, p. 148) A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

(a) What is the probability that the 10th transistor produced is the first with a defect?

```
noquote(sprintf("                    .98^9 * .02 = %.7f ",.98^9 * .02))
```

```
## [1]                    .98^9 * .02 = 0.0166750
```

(b) What is the probability that the machine produces no defective transistors in a batch of 100?

```
noquote(sprintf("                    .98^10 = %.7f ",.98^10))
```

```
## [1]                    .98^10 = 0.8170728
```

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

The Expected Value or Mean is $\frac{1}{50}$ of the sample size so you would expect to have at least 1 defect by the 50th transistor.*

$$E(x) = \sum n * P(x)$$

The standard deviation of a binomial variable with probability of x is

$$\mu = \sqrt{n * P(x) * (1 - P(x))}$$

Note : a Bernoulli variable is how we describe a sample size of 1.

```
sprintf("Given n = %d, standard deviation = %.2f", 1, sqrt(1 * .98 * .02))
```

```
## [1] "Given n = 1, standard deviation = 0.14"
```

```
sprintf("Given n = %d, standard deviation = %.2f", 100, sqrt(100 * .98 * .02))
```

```
## [1] "Given n = 100, standard deviation = 1.40"
```

```
sprintf("Given n = %d, standard deviation = %.2f", 500, sqrt(500 * .98 * .02))
```

```
## [1] "Given n = 500, standard deviation = 3.13"
```

```
sprintf("Given n = %d, standard deviation = %.2f", 1000, sqrt(1000 * .98 * .02))
```

```
## [1] "Given n = 1000, standard deviation = 4.43"
```

The standard deviation increases as sample sizes increases but it increases at a lower rate than n increases.

(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?

The probability of first failure with P(X) could be calculated as

$$\frac{1}{P(X)}$$

or the rgeom function can simulate trials of first success/failure

```
n=1000

sample_trials<-rgeom(n,.05)
sd_trials<-sd(sample_trials)

sprintf("The mean of our trials of first defective transistor is %.2f", mean(sample_trials))
```

```
## [1] "The mean of our trials of first defective transistor is 18.90"
```

```
sprintf("The sample deviation is %.2f",sd_trials)
```

```
## [1] "The sample deviation is 18.91"
```

(e) Based on your answers to parts (c) and (d), how does increasing the probability of an event affect the mean and standard deviation of the wait time until success?

Generally the larger the sample size, the closer to actual size

```
n=100000

sample_trials<-rgeom(n,.05)
sd_trials<-sd(sample_trials)

sprintf("The new mean of our trials of first defective transistor is %.2f", mean(sample_trials))
```

```
## [1] "The new mean of our trials of first defective transistor is 18.94"
```

```
sprintf("The new sample deviation is %.2f",sd_trials)
```

```
## [1] "The new sample deviation is 19.47"
```

Male children. While it is often assumed that the probabilities of having a boy or a girl are the same, the actual probability of having a boy is slightly higher at 0.51. Suppose a couple plans to have 3 kids.

(a) Use the binomial model to calculate the probability that two of them will be boys.

```
p_boy<-.51
p_girl<-.49

p_2_boys_1_girl<-p_boy^2 * p_girl * 3      # there are 3 ways in which 2 boys can be selected
sprintf("The probability of 2 boys is %.2f",p_2_boys_1_girl)
```

```
## [1] "The probability of 2 boys is 0.38"
```

(b) Write out all possible orderings of 3 children, 2 of whom are boys. Use these scenarios to calculate the same probability from part (a) but using the addition rule for disjoint outcomes. Confirm that your answers from parts (a) and (b) match.

Child1	Child2	Child3
G	B	B
B	G	B
B	B	G

P(G and B and B) = .51 * .51 * .49 = .1275

P(B and G and B) = .51 * .51 * .49 = .1275

P(B and B and G) = .51 * .51 * .49 = .1275

(c) If we wanted to calculate the probability that a couple who plans to have 8 kids will have 3 boys, briefly describe why the approach from part (b) would be more tedious than the approach from part (a).

Part A condenses several additions into one multiplication, so it was done with one line.

Serving in volleyball. (4.30, p. 162) A not-so-skilled volleyball player has a 15% chance of making the serve, which involves hitting the ball so it passes over the net on a trajectory such that it will land in the opposing team's court. Suppose that her serves are independent of each other.

(a) What is the probability that on the 10th try she will make her 3rd successful serve?

```
# 1) the probability of a 3 in 10 permutation in any order
good_serve_perm<-.85^7 * .15^3

# 2) what are the number of ways to get 2 sucesses in 9 tries in any order (9 choose 2)
good_serve_ways<-choose(9,2)

# 3) multiply the probability of a unique 10 choose 3 result and the number of non unique 9 choose 2 wa

sprintf("The probability of hitting the 3rd success on the 10the try is %.3f",
        good_serve_ways * good_serve_perm)
```

```
## [1] "The probability of hitting the 3rd success on the 10the try is 0.039"
```

(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful?

Any individual serve has a 15% chance of sucess.

(c) Even though parts (a) and (b) discuss the same scenario, the probabilities you calculated should be different. Can you explain the reason for this discrepancy?

Probability only calculates what is unknown. part a) asked us to find probility on 10 unknown events. part b) asked about 1 unknown event.