Chapter 4 - Distributions of Random Variables

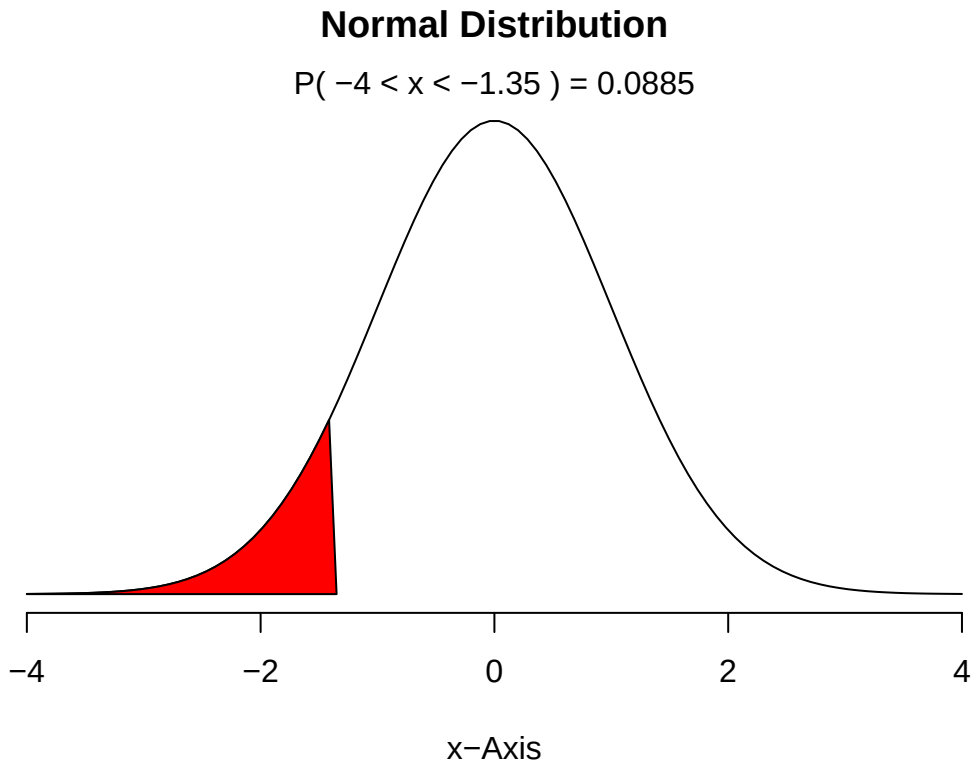
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

```
library(DATA606)
```

(a) $Z < -1.35$

8.85%

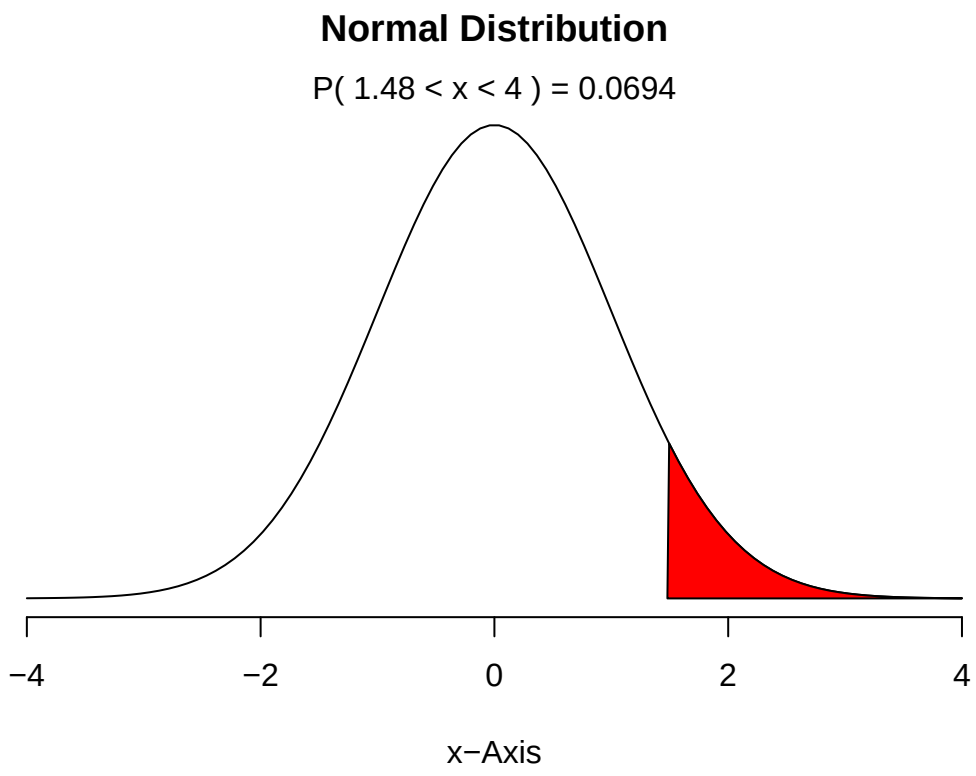
```
normalPlot(mean = 0, sd = 1, bounds = c(-4, -1.35), tails = FALSE)
```



(b) $Z > 1.48$

6.94%

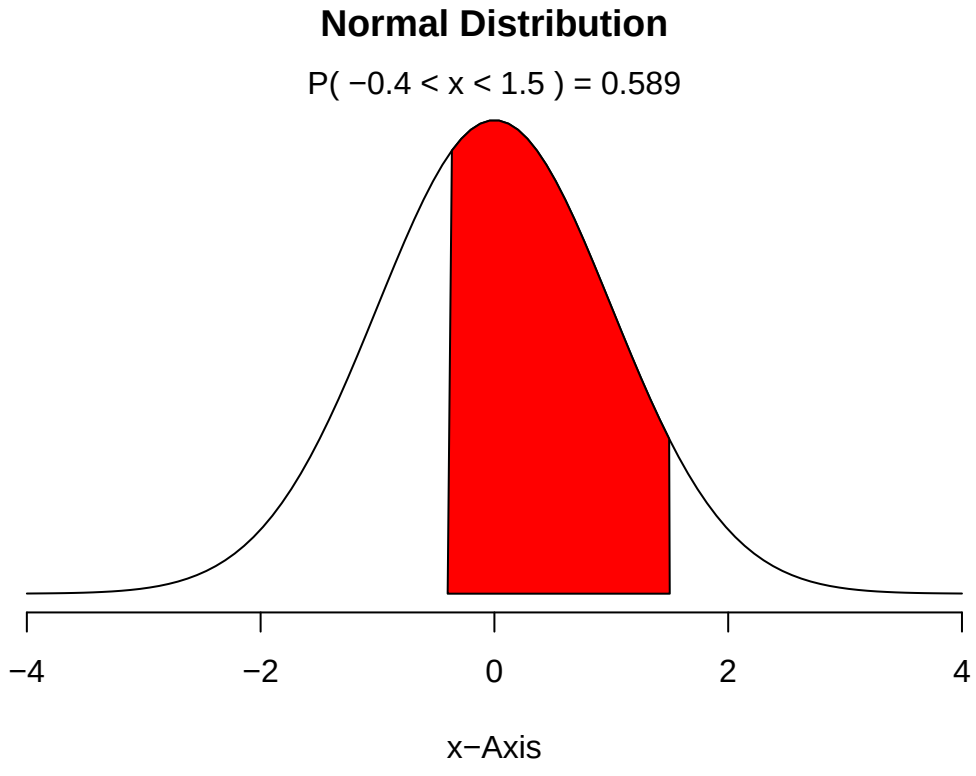
```
normalPlot(mean = 0, sd = 1, bounds = c(1.48, 4), tails = FALSE)
```



(c) $-0.4 < Z < 1.5$

58.9%

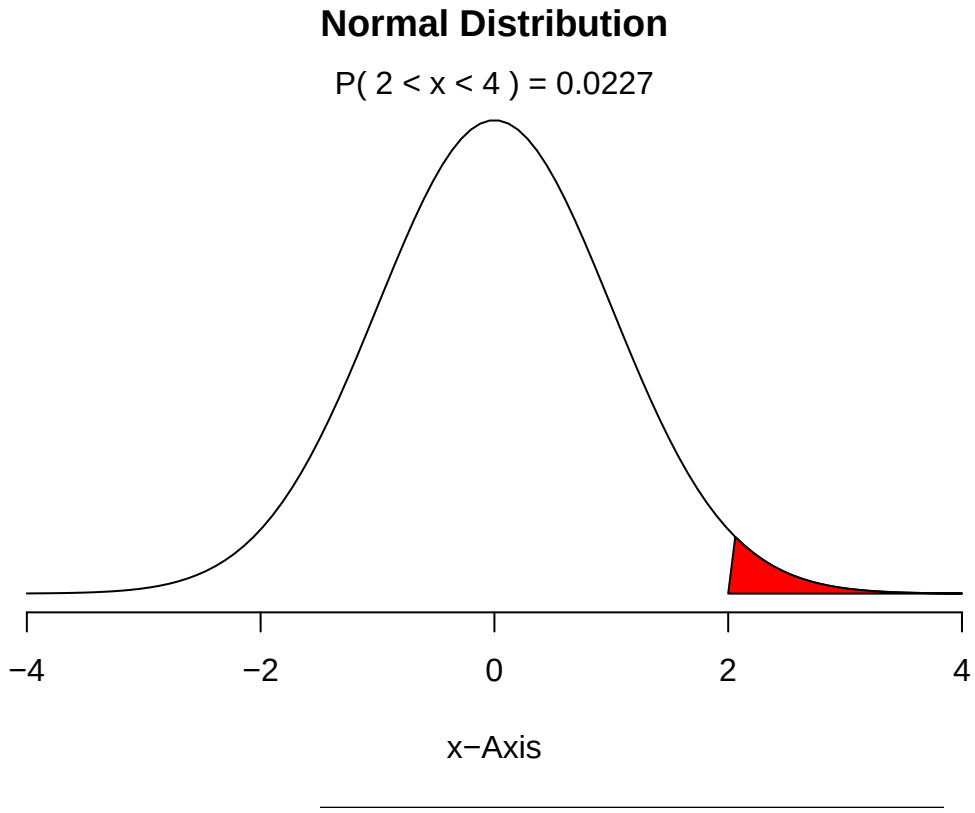
```
normalPlot(mean = 0, sd = 1, bounds = c(-0.4, 1.5), tails = FALSE)
```



(d) $|Z| > 2$

2.27%

```
normalPlot(mean = 0, sd = 1, bounds = c(2, 4), tails = FALSE)
```



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.

$$N(\mu = 4313, \sigma = 583) \tag{1}$$

$$N(\mu = 5261, \sigma = 807) \tag{2}$$

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

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

```
Leo <- (4948 - 4313) / 583; Leo
```

```
## [1] 1.089194
```

```
Mary <- (5513 - 5261) / 807; Mary
```

```
## [1] 0.3122677
```

These scores tell you where in the distribution of finishing times each participant sits.

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

Mary did better because her time was fewer standard deviations above the mean.

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

```
1 - pnorm(4948, mean = 4313, sd = 583, lower.tail = TRUE)
```

```
## [1] 0.1380342
```

He finished faster than 13.8% of participants in his group.

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

```
1 - pnorm(5513, mean = 5261, sd = 807, lower.tail = TRUE)
```

```
## [1] 0.3774186
```

Mary finished faster than 37.7% of her group.

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

Yes, of course. The area under the curve would be irregular in this case, meaning that the probability densities at each x value would have different values.

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)
m.height <- 61.52
s.height <- 4.58

within1 <- heights[heights >= m.height - s.height & heights <= m.height + s.height]
within2 <- heights[heights >= m.height - 2*s.height & heights <= m.height + 2*s.height]
within3 <- heights[heights >= m.height - 3*s.height & heights <= m.height + 3*s.height]

length(within1) / length(heights)
```

```
## [1] 0.68
```

```
length(within2) / length(heights)
```

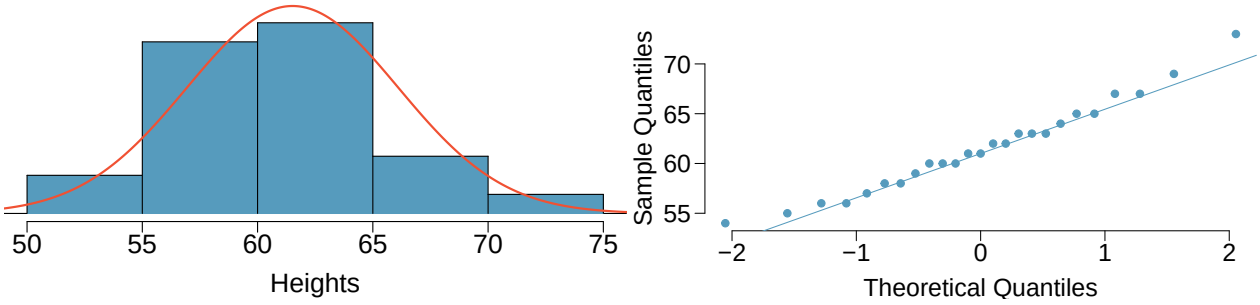
```
## [1] 0.96
```

```
length(within3) / length(heights)
```

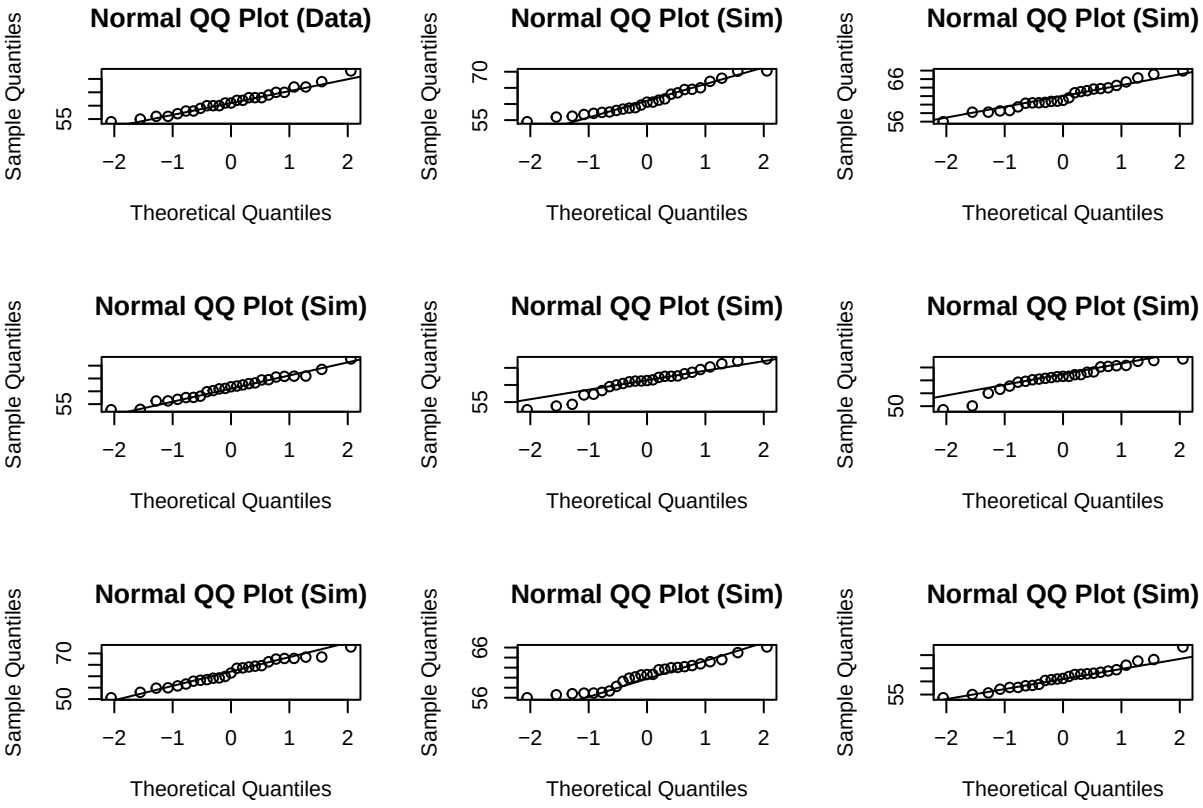
```
## [1] 1
```

Close enough for jazz!

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




```
qqnormsim(heights)
```



Judging by all representations of the data, these data do appear to follow a normal distribution. The histogram appears monomodal and symmetrical, with a mean roughly at the center. The QQ plots also show a collusion among sample data and theoretical quantities.

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?

```
p <- 0.02

(1 - p)^9 * p
```

```
## [1] 0.01667496
```

1.67% chance that the 10th is the first with a defect.

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

```
(1 - p)^100
```

```
## [1] 0.1326196
```

13.26%

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

$$\mu = \frac{1}{p} \tag{4}$$

$$\sigma^2 = \frac{1-p}{p^2} \tag{5}$$

```
1 / p
```

```
## [1] 50
```

```
sqrt((1-p)/(p^2))
```

```
## [1] 49.49747
```

(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?

```
p <- 0.05
1 / p
```

```
## [1] 20
```

```
sqrt((1-p)/(p^2))
```

```
## [1] 19.49359
```

(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?

The mean will decrease, as will the standard deviation.

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 <- 0.51
(p) * (p) * (1 - p) * 3
```

```
## [1] 0.382347
```

(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.

bbg bgb gbb

bbb ggg bgg gbg ggb

```
(p*p*(1-p)) + (p*(1-p)*p) + ((1-p)*p*p)
```

```
## [1] 0.382347
```

(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).

There are 256 combinations of rugrats in this scenario... The additive approach would be extremely tedious as it would have 256 terms!

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?

```
n <- 10
k <- 3
p <- 0.15

combos <- factorial(n-1) / (factorial(k-1) * factorial((n-1)-(k-1)))
threeInTen <- combos * p^k * (1-p)^(n-k)

threeInTen
```

```
## [1] 0.03895012
```

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

The trials are independent. p = 0.15

(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?

Because the trials are independent, the past cannot factor into the probability of one single trial being a success. However, if we are considering all the possibilities for the k^{th} success on the n^{th} trial, we can easily compute a negative binomial distribution.