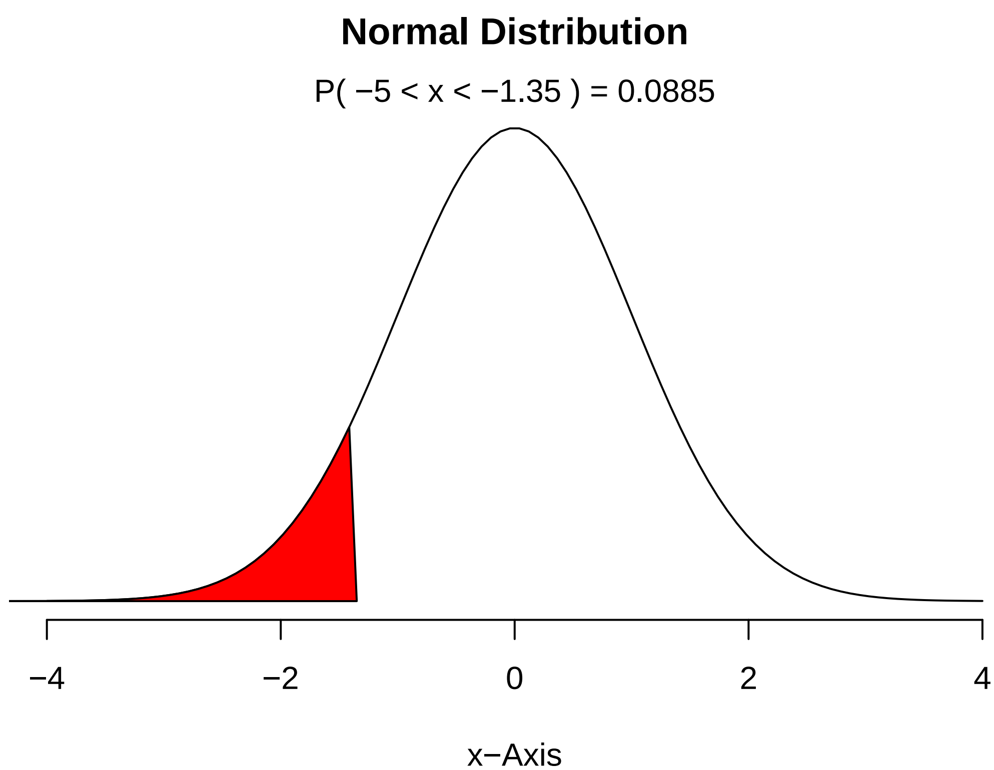
Data606 Homework Chapter 4 - Distributions of Random Variables

Homework Completed by Chad Bailey

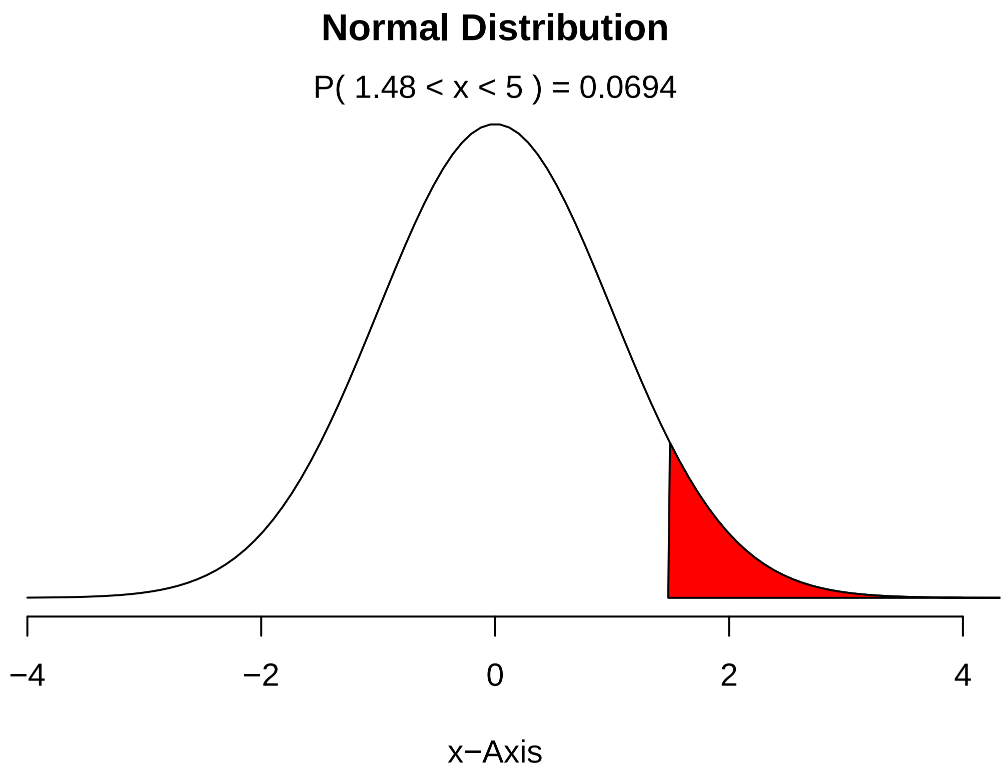
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$
(b) $Z > 1.48$
(c) $-0.4 < Z < 1.5$
(d) $|Z| > 2$

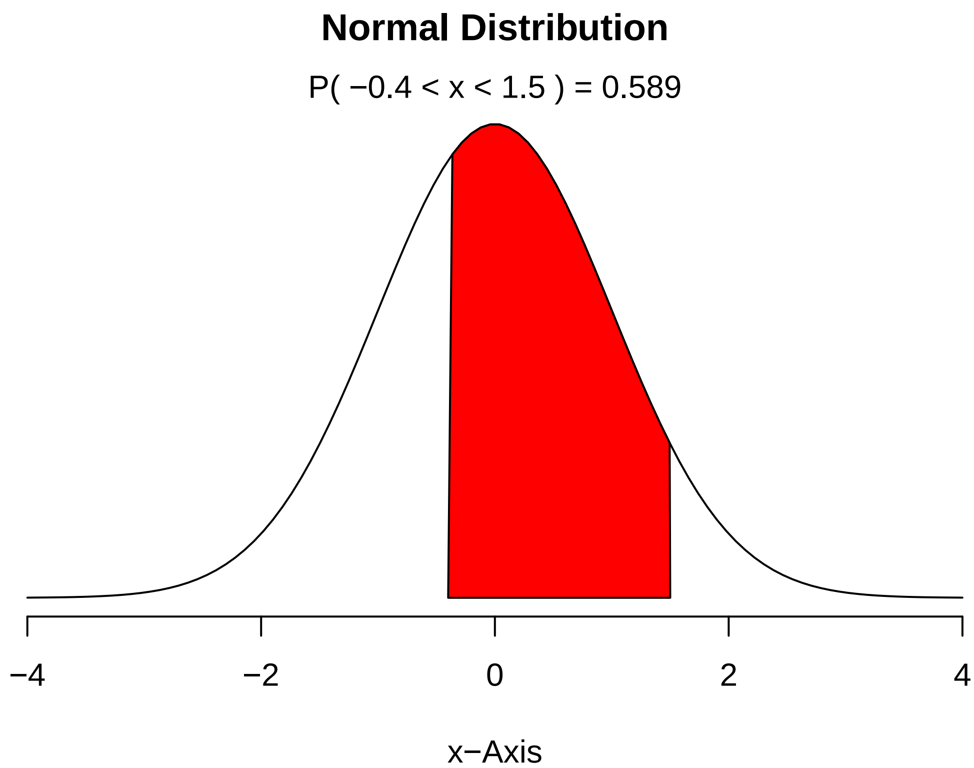
```
## Student response to question #1
##  1.a Z < -1.35, area is 0.0885
      normalPlot(mean = 0, sd = 1, bounds = c( -5, -1.35))
```



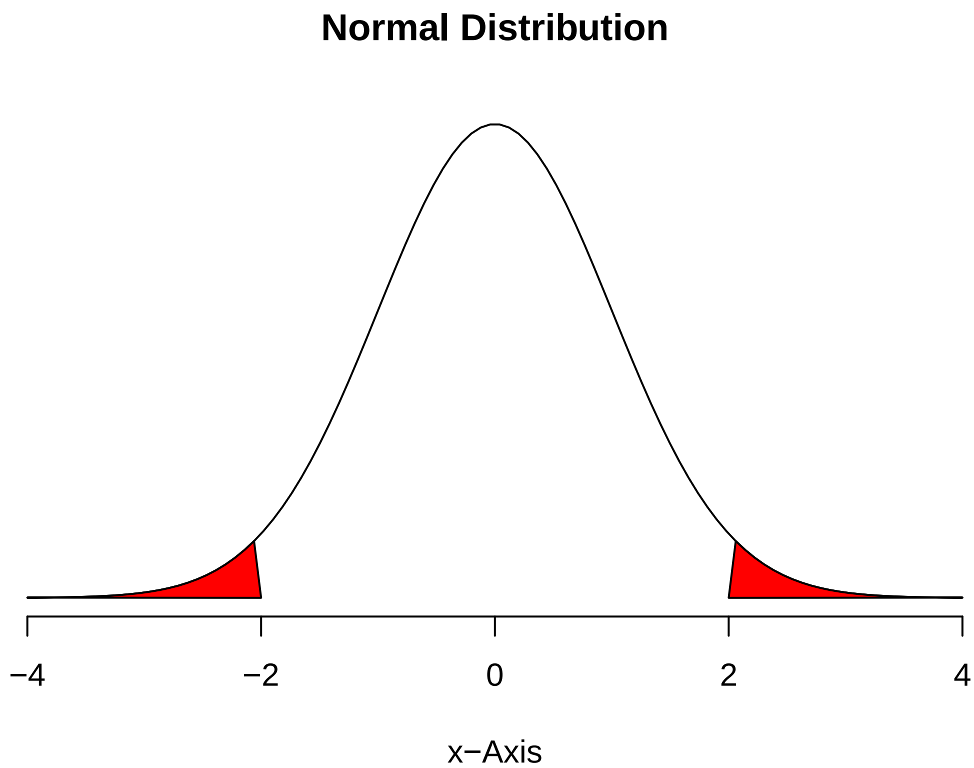

```
##  1.b Z > 1.48, area is 0.0694
normalPlot(mean = 0, sd = 1, bounds = c( 1.48, 5))
```




```r
##  1.c -0.4 < Z < 1.5, area is 0.589
normalPlot(mean = 0, sd = 1, bounds = c( -0.4, 1.5))
```




```r
## 1.d |Z| > 2, area is 0.0454 = 0.0227 + 0.0227
normalPlot(mean = 0, sd = 1, bounds = c(-2, 2), tails = TRUE)
```


Normal Distribution
−4
−2
0
2
4
x−Axis

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.
(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?
(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.
(d) What percent of the triathletes did Leo finish faster than in his group?
(e) What percent of the triathletes did Mary finish faster than in her group?
(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

```
##  2.a "Men, Ages 30 - 34": N(mu = 4313, sigma = 583)
##      "Women, Ages 25 - 29": N(mu = 5261, sigma = 807)

##  2.b z-scores
    ## Leo
      men3034_mean <- 4313
      men3034_sd <- 583
      leo_time <- 4989
      leo_zscore <- (leo_time - men3034_mean) / men3034_sd
      leo_zscore
```

```
## [1] 1.15952
```

```
     ## Mary
      women2529_mean <- 5261
      women2529_sd <- 807
      mary_time <- 4989
      mary_zscore <- (mary_time - women2529_mean) / women2529_sd
      mary_zscore
```

```
## [1] -0.3370508
```

```
    ## These z-scores tell us that Leo's time was 1.16 standard deviations
    ## above (slower than) average for his gender/age group while Mary's
    ## time was 0.34 standard deviations below (faster than) average for
    ## her gender/age group.

##  2.c Mary ranked better within her gender/age group. This is because
```

```r
##      lower times are desirable and so rankings are based on ascending
##      times. Mary's time was below (faster) than average for her group
##      while Leo's time was above (slower) than average for his group.

##  2.d Leo's time was better than 12.31% of his gender/age group.
     1-pnorm(leo_zscore)
```

```
## [1] 0.1231222
```

```r
##  2.e Mary's time was better than 63.20% of her gender/age group.
     1-pnorm(mary_zscore)
```

```
## [1] 0.6319607
```

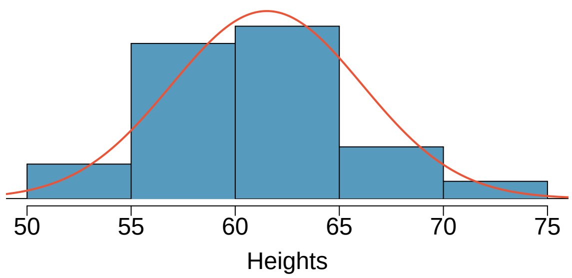
```r
##  2.f Yes, if these were not nearly normal distibutions it is likely the
##      the answers to parts (b) - (e) would change. If the distribution is
##      is not normal:
##          (i) it is not appropriate to use z-scores.
##          (ii) there could be significant skew, in which the mean and sd
##               may no longer be the best descriptors.
##          (iii) we would could not use zscore lookup tables to determine
##                the percent of times that were below their times. More
##                informaiton would be required.
```

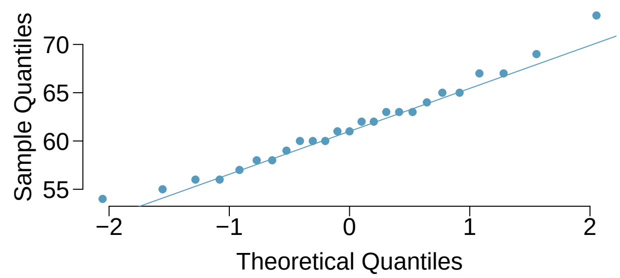
Heights of female college students Below are heights of 25 female college students.

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.

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






```
## 3.a check for 68-95-99 rule
    h_mean <- mean(heights)
    h_sd <- sd(heights)

    ##  check that roughly 68% of data are within 1 sd of mean
        length(heights[x>=h_mean-h_sd & x<=h_mean+h_sd]) / length(heights)
```

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

```
    ##  check that roughly 95% of data are within 2 sd of mean
        length(heights[x>=h_mean-2*h_sd & x<=h_mean+2*h_sd]) / length(heights)
```

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

```
    ##  check that roughly 99% of data are within 3 sd of mean
        length(heights[x>=h_mean-3*h_sd & x<=h_mean+3*h_sd]) / length(heights)
```

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

```
##  3.b Yes, this distribution appears to be nearly normal. The histogram
##      roughly follows the normal curve overlay. The QQ plot shows that most
##      of the data are near the central line.
```

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?
(b) What is the probability that the machine produces no defective transistors in a batch of 100?
(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?
(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?
(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?

```
## set the basic variables for this set of problems

  def_p <- 0.02
  def_q <- 1-def_p

## 4.a P(n=10)

    ((def_q)^(10-1))*def_p
```

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

```
## 4.b P(n=101)
    ((def_q)^(101-1))*def_p
```

```
## [1] 0.002652391
```

```
## 4.c mean and sd
    mean <- 1/def_p
    sd <- sqrt((def_q)/((def_p)^2))

    mean; sd
```

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

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

```
    ## On average, you would expect 50 transistors to be produced before
    ## one had a defect. The standard deviation is 49.50.

## 4.d defective rate 0.05
      mean2 <- 1/0.05
      sd2 <- sqrt((1-0.05)/(0.05^2))

      mean2; sd2
```

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

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

```
    ## If the defective rate is changed to 5%, you would expect, on average,
    ## for 20 transistors to be produced before one had a defect. The standard
    ## deviation is 19.49.

## 4.e  Even relatively small increases in the probability of an event results
##      meaningfully lower values for mean and standard deviation of the wait
##      time until success.
```

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.
(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.
(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).

```
##  5.a
    (factorial(3) / (factorial(2)*factorial(3-2))) * (0.51^2) * (1-0.51)^(3-2)
```

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

```
##  5.b
##    m-m-f
##    m-f-m
##    f-m-m
        0.51*0.51*(1-0.51) + 0.51*(1-0.51)*0.51 + (1-0.51)*0.51*0.51
```

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

```
##    Yes, both methods show a probability of 0.3823


##  5.c Calculating the probability that a couple who plans to have 8 kids
##      will have 3 boys using the method from part (b) would be tedious
##      because it would require the calcuaiton of 56 seperate combinations.

        factorial(8) / (factorial(3)*factorial(8-3))
```

```
## [1] 56
```

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

```
##  6.a
      comb <- (factorial(10-1) / (factorial(3-1)*factorial(10-3)))
      comb * (.15)^3 * (1-0.15)^(10-3)
```

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

```
      ## The probability that on the 10th serve she will make her 3rd successful
      ## serve is 0.0390

##  6.b The probability that her 10th serve will be successful is 0.15

##  6.c The scenario is the same but the questions are different. Part (a) is
##      considering all 10 serves as a cummulative event. While part (b) is only
##      asking about a single serve, the 10 serve. Each serve is an independent
##      event. Thus each single serve has a probability of success of 0.15
```