Data 606 - Chapter 4 - Distributions of Random Variables

Cameron Smith

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) $|Z| > 2$

ANSWERS

(a) .089
(b) .069
(c) .589
(d) .023

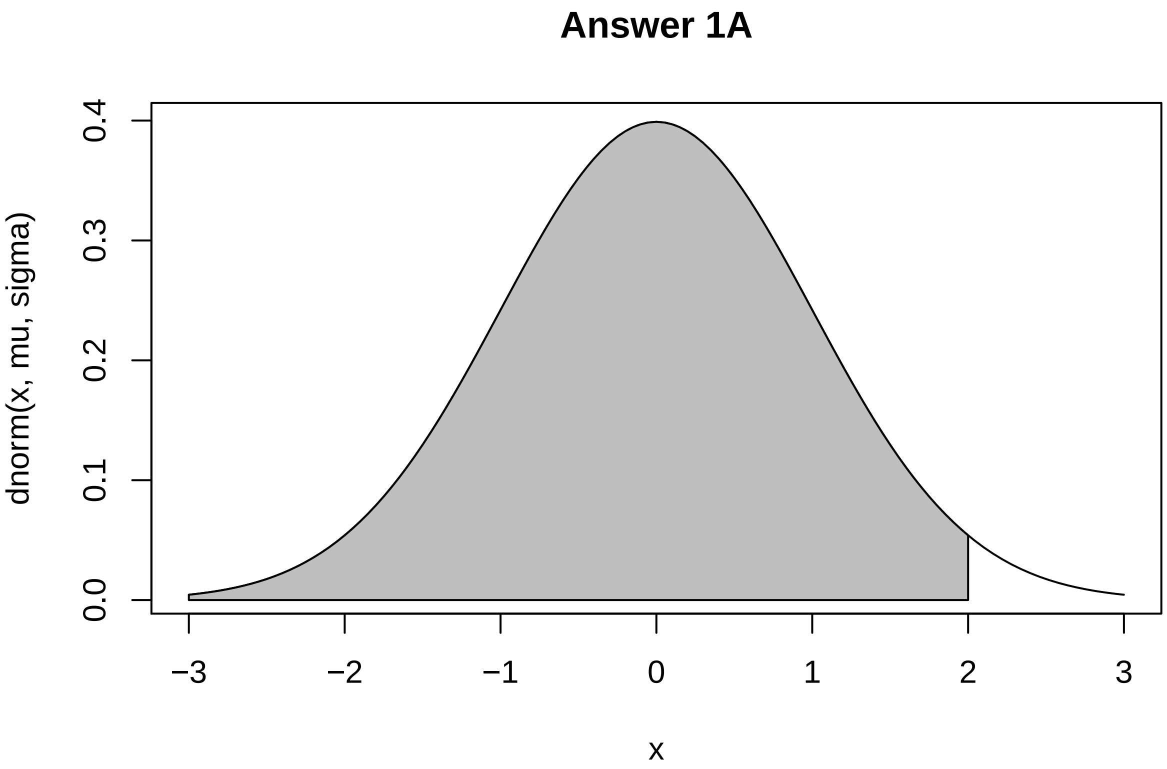
Plots for each answer are below.

```
## [1] 0.08850799
```

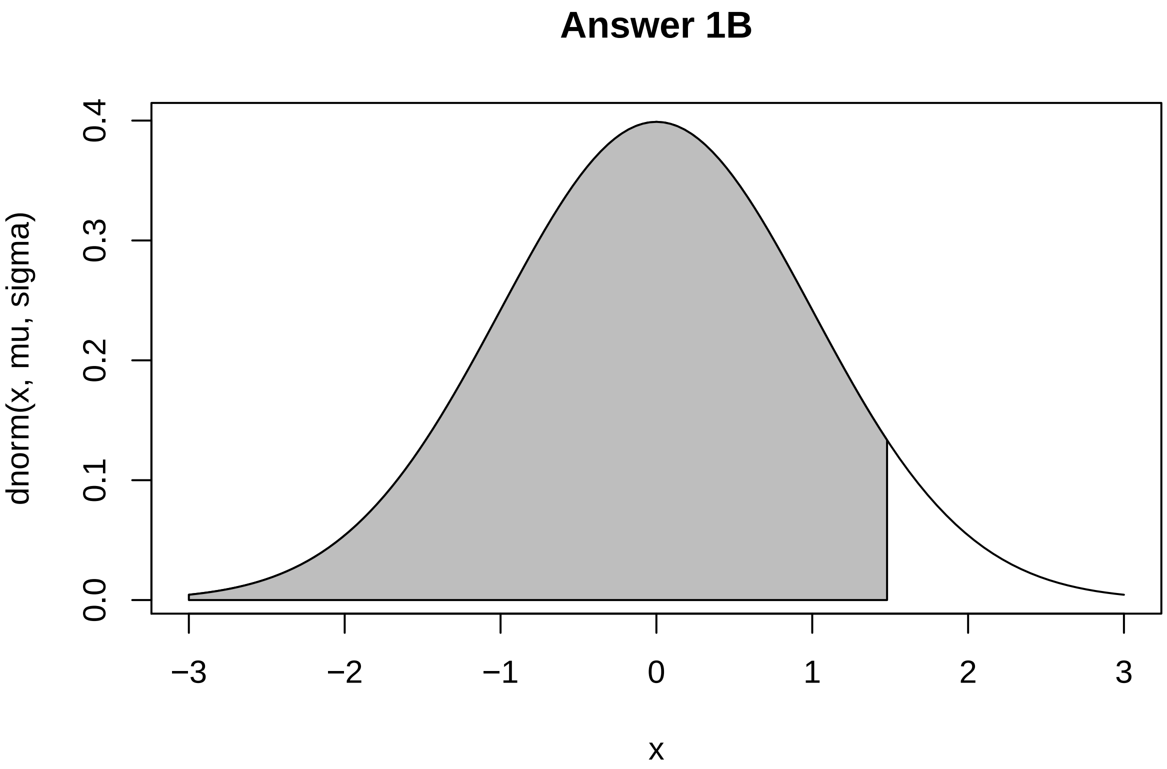
```
## [1] 0.06943662
```

```
## [1] 0.5886145
```

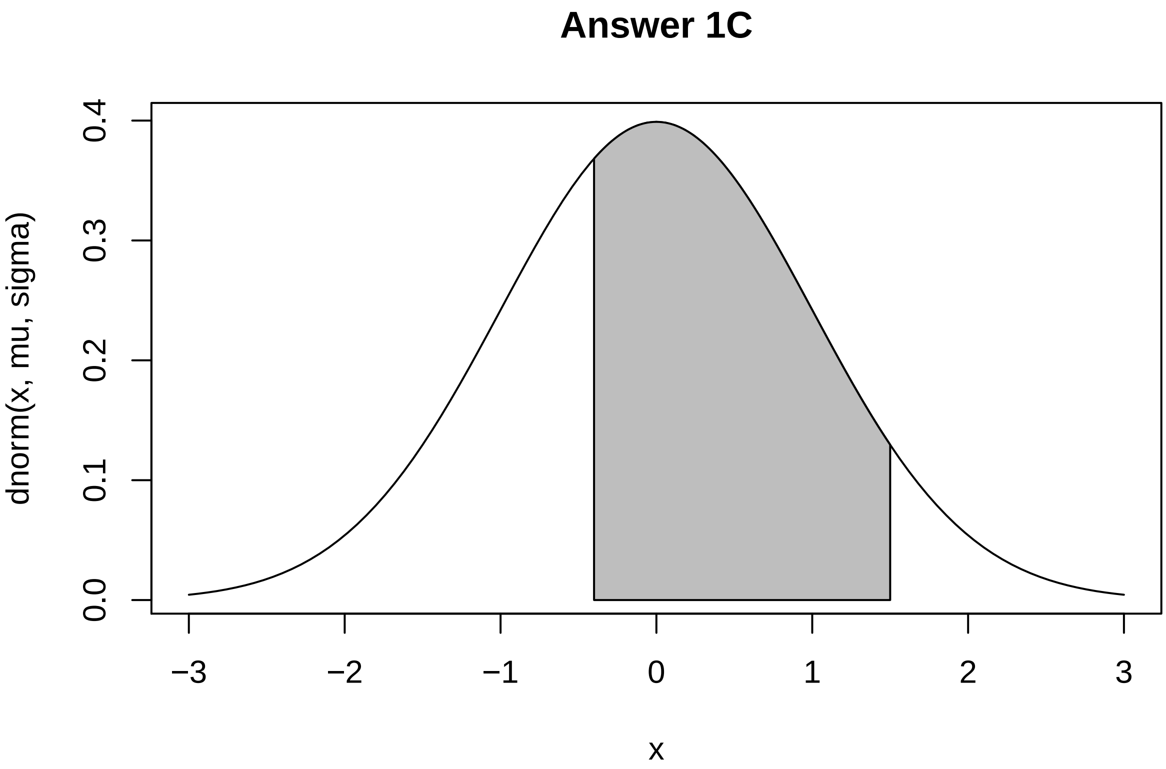
```
## [1] 0.02275013
```

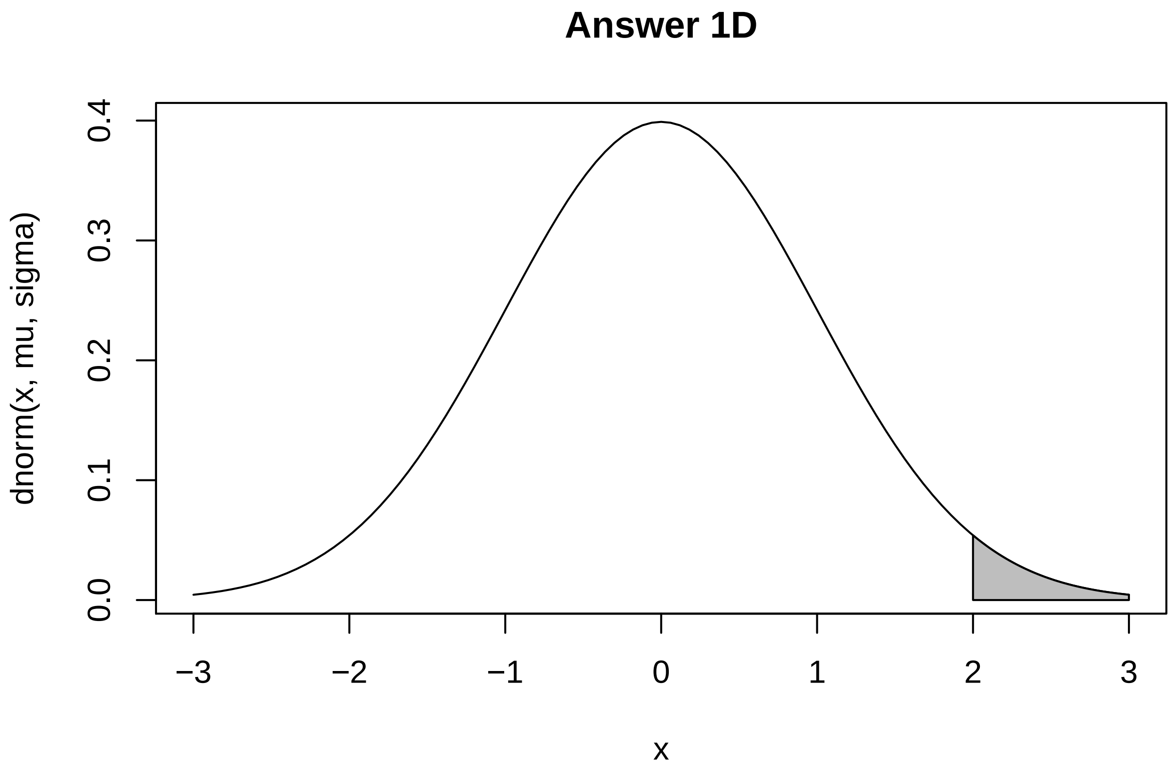


Answer 1A



Answer 1B


Answer 1C
dnorm(x, mu, sigma)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x


Answer 1D
dnorm(x, mu, sigma)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

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.

ANSWERS

(a) Shorthand for the men group: N(mu = 4313, sigma = 483) Shorthand for the women group: N(mu = 5261, sigma = 807)

(b) Leo's Z score: 1.09, Mary's S Score: .312

(c) Mary scored better. Her Z score was .31, versus Leo's of 1.08, indicating her score was the least far (i.e. least amount of time from) the mean.

(d) Leo finished faster than 13.8% of others in his group

(e) Mary finished faster than 31.2% of others in her group

(f) Yes, if the distributions were not nearly normal then the answers would change. If not a normal distribution then the Z scores and associated probabilities would be invalid as they could not be predicted based on ths information.

```
# Create function to calculate Z Score
zscore = function(q, m=0, s=1){
  (q-m) / s
}

#LeoZscore
zscore(4948, 4313, 583)
```

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

```
1 - pnorm(4948, 4313, 583)
```

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

```
#Mary Z Score
zscore(5513, 5261, 807)
```

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

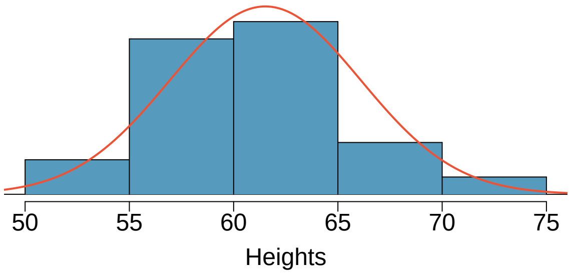
```
1 - pnorm(5513, 5261, 807)
```

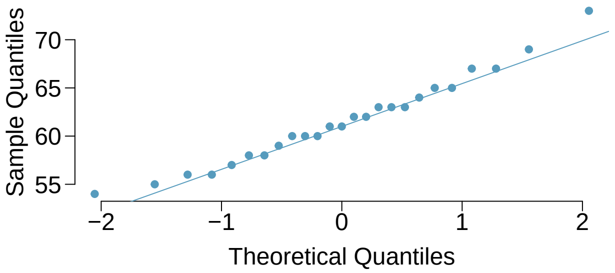
```
## [1] 0.3774186
```

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.





```
# Use the DATA606::qqnormsim function
```

ANSWERS

(a) Yes the data approximately follow the 68-95-99.7% rule. Please see code below for specific percentages.
(b) Yes, the data appears to follow a normal distribution based on the histogram and QQ plot. The QQ plot in particular makes it clear that the majority of data points are on or near the theoretical values.

```
heightdata <- data.frame(
  student_id = c(1:25),
  heights = c(54, 55, 56, 56, 57, 58, 58, 59, 60, 60, 60, 61, 61, 62, 62, 63, 63, 63, 64, 65, 65, 67, 6
)

heightmean <- mean(heightdata$heights)
heightsd <- sd(heightdata$heights)

# Percentage within 1 SD
num1 <- heightdata %>% filter(between(heights, heightmean - 1*heightsd, heightmean + 1*heightsd)) %>%
  nrow() / nrow (heightdata)

# Percentage within 2 SD
num2 <- heightdata %>% filter(between(heights, heightmean - 2*heightsd, heightmean + 2*heightsd)) %>%
  nrow() / nrow (heightdata)

# Percentage within 3 SD
num3 <- heightdata %>% filter(between(heights, heightmean - 3*heightsd, heightmean + 3*heightsd)) %>%
  nrow() / nrow (heightdata)

# Detailed answer for part (a)
cat("% within 1 standard deviation - Normal = 68%:    ", num1, "\n")
```

```
## % within 1 standard deviation - Normal = 68%:     0.68
```

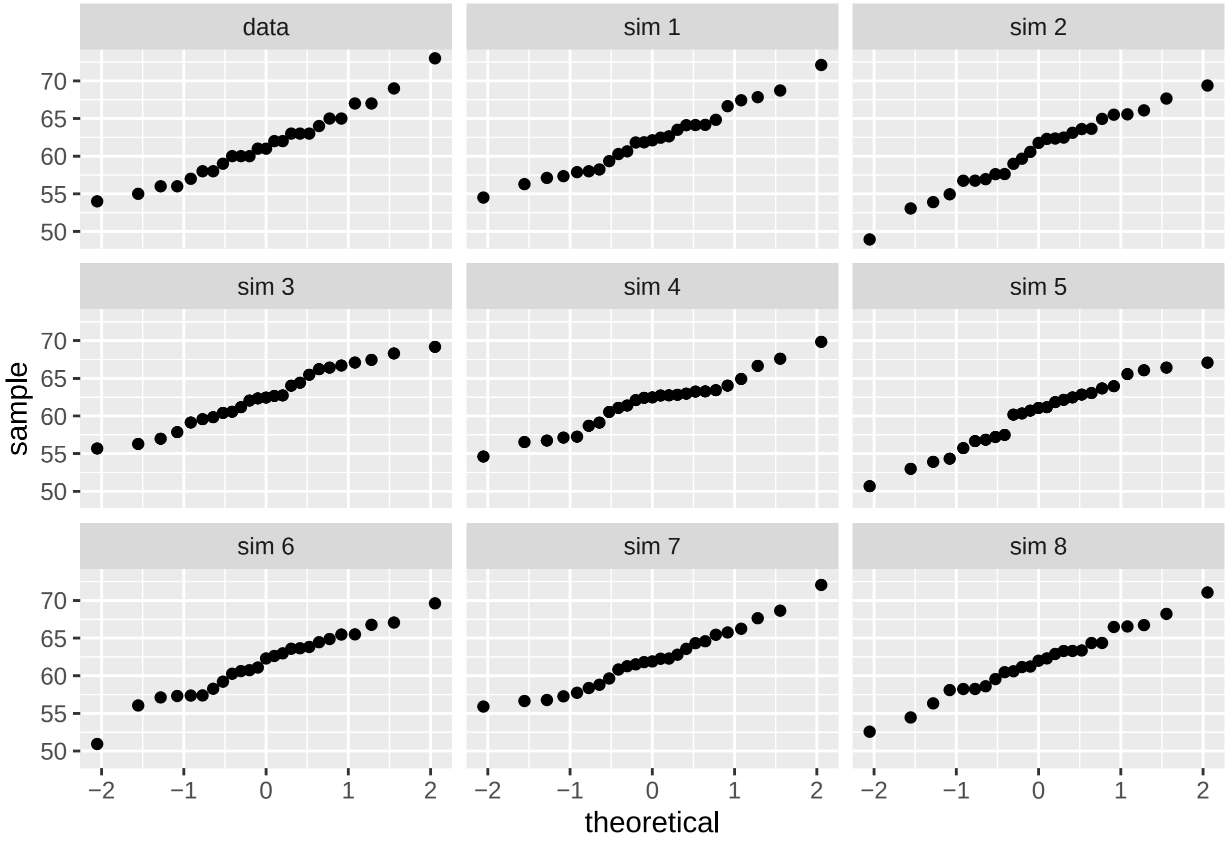
```
cat("% within 2 standard deviations - Normal = 95%:   ", num2, "\n")
```

```
## % within 2 standard deviations - Normal = 95%:    0.96
```

```
cat("% within 3 standard deviations- Normal = 99.7%:  ", num3,"\n")
```

```
## % within 3 standard deviations- Normal = 99.7%:   1
```

```
# QQ plots based on simualated data to compare with real data
qqnormsim(heightdata$heights, heightdata)
```



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?

**ANSWERS*

Calculations for each are detailed in the code block below.

(a) 1.6%
(b) 13%
(c) The mean is 50, so we'd expect that many trials until the first failure. The standard deviation is nearly the same - 49.49.
(d) The mean is 20, so we'd expect that many trials until the first failure. The standard deviation is 19.49.
(e) It seems that increasing the probability decreases the mean and the standard deviation.

```
n <- 10
p <- .02

# A - probability of 10th being the first to be defective
(1-p)^(n-1)*p
```

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

```
# B - no defective models in 100 trials
(1-p)^100
```

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

```
# C
1/p #the mean
```

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

```
sqrt((1-p)/(p^2)) #the standard deviation
```

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

```
# D
p2 <- .05
1/p2 #the mean
```

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

```
sqrt((1-p2)/(p2^2)) #the standard deviation
```

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

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).

ANSWERS

(a) There's a 38% chance that they would have two boys in 3 trials (births)
(b) Combinations are G-B-B, B-B-G, B-G-B, all with 38% chance. Details in code below.
(c) There are many more combinations to account for and thus much more manual effort would be required.

```
p_boy <- .51
p_girl <- 1 - p_boy
n <- 3
k <- 2

# A - Calculate probability of success that two will be boys in 3 trials
calc_a <- dbinom(k, n, p_boy)
cat("Number calculated with function: ", calc_a, "\n")
```

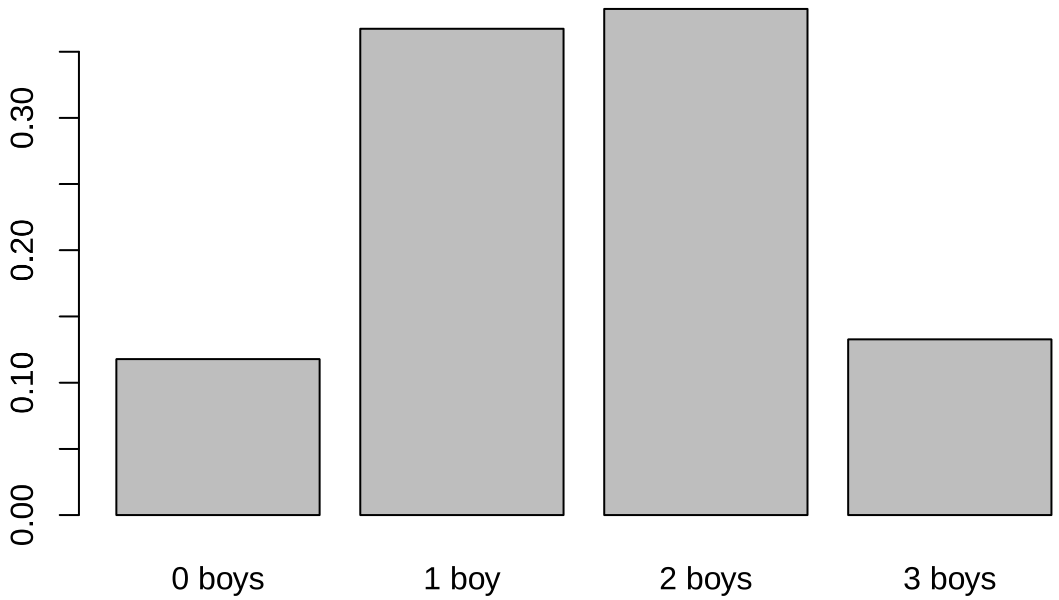
```
## Number calculated with function:  0.382347
```

```
# B
# Possible combinations: G-B-B, B-B-G, B-G-B

# Adding probability of all possible scenarios equals answer calculated above
calc_b <- (p_girl * p_boy * p_boy) +
(p_boy * p_boy * p_girl) +
(p_boy * p_girl * p_boy)
cat("Number calculated manually:      ", calc_b)
```

```
## Number calculated manually:       0.382347
```

```
# Plotting the distribution
barplot(dbinom(0:n, n, p_boy), names.arg=c("0 boys", "1 boy", "2 boys", "3 boys"))
```


0.00
0.10
0.20
0.30
0 boys
1 boy
2 boys
3 boys

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?

ANSWERS

(a) 3.9%
(b) 15%, because the outcomes are independent.
(c) In part a no events had occurred yet, it was a blank slate.

In part b the event (two serves, i.e. 'successes') had already occurred, so all that had to be calculated was the odds of the next event being a success.

```r
# A
# Need to break it down into 2 parts: probability of two successful serves in 9 tries, and then 1 more
p <- .15
n <- 9
k <- 2
twoinnine <- dbinom(2, 9, .15)
successfultenth <- .15
threeinten <- twoinnine * successfultenth
cat("The probability that on the 10th try she will have 3 successes is: ", threeinten)
```

```
## The probability that on the 10th try she will have 3 successes is:  0.03895012
```