# Chapter 4 - Distributions of Random Variables

David Simbandumwe

**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.

```
normal_area <- function(mean = 0, sd = 1, lb, ub, acolor = "lightgray", ...) {
    x <- seq(mean - 3 * sd, mean + 3 * sd, length = 100)

    if (missing(lb)) {
        lb <- min(x)
    }
    if (missing(ub)) {
        ub <- max(x)
    }

    x2 <- seq(lb, ub, length = 100)
    plot(x, dnorm(x, mean, sd), type = "n", ylab = "")

    y <- dnorm(x2, mean, sd)
    polygon(c(lb, x2, ub), c(0, y, 0), col = acolor)
    lines(x, dnorm(x, mean, sd), type = "l", ...)
}
```

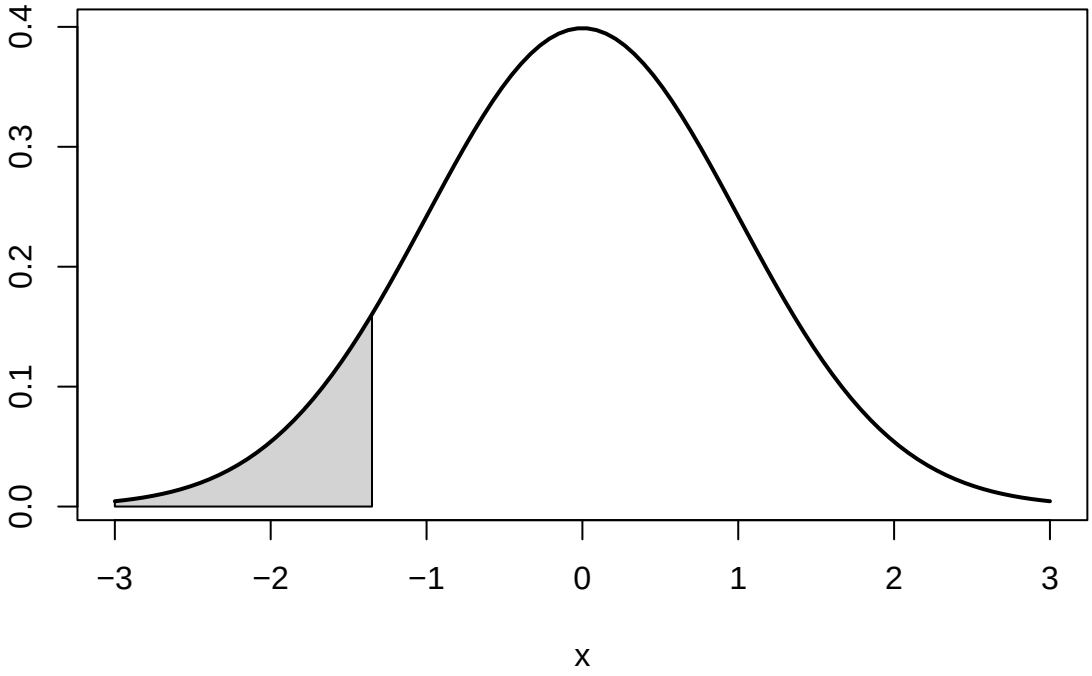
(a) $Z < -1.35$

area under the curve 8.9%

```
pnorm(-1.35)
```

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

```
normal_area(mean = 0, sd = 1, lb = -3, ub = -1.35, lwd = 2)
```



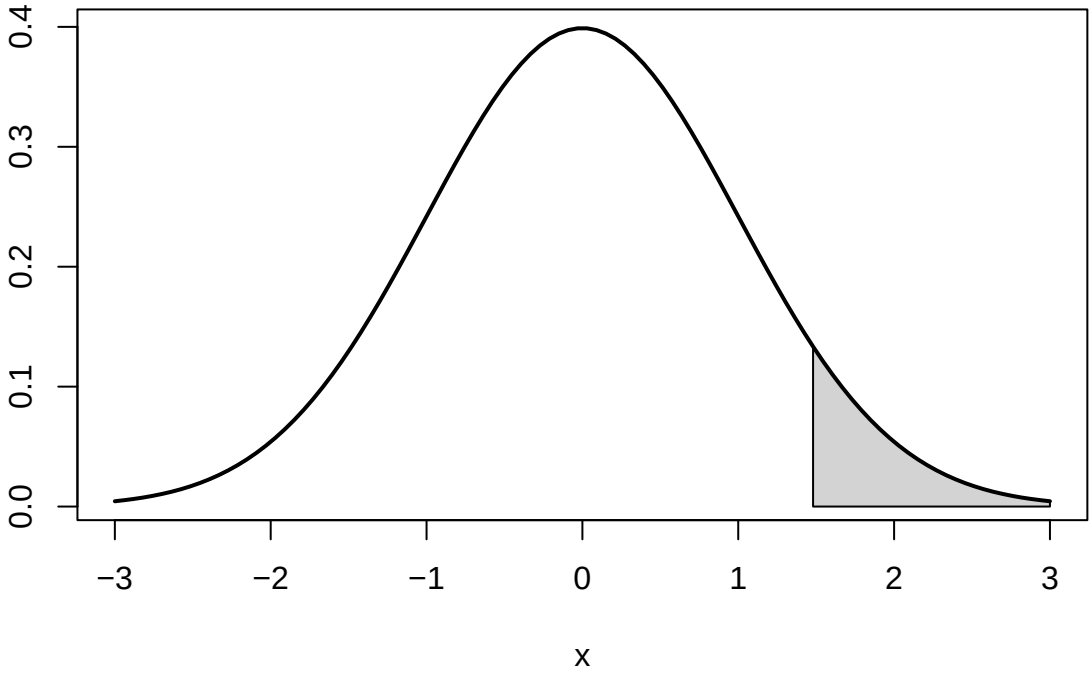

(b) $Z > 1.48$

area under the curve 6.9%

```
1 - pnorm(1.48)
```

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

```
normal_area(mean = 0, sd = 1, lb = 1.48, ub = 3, lwd = 2)
```



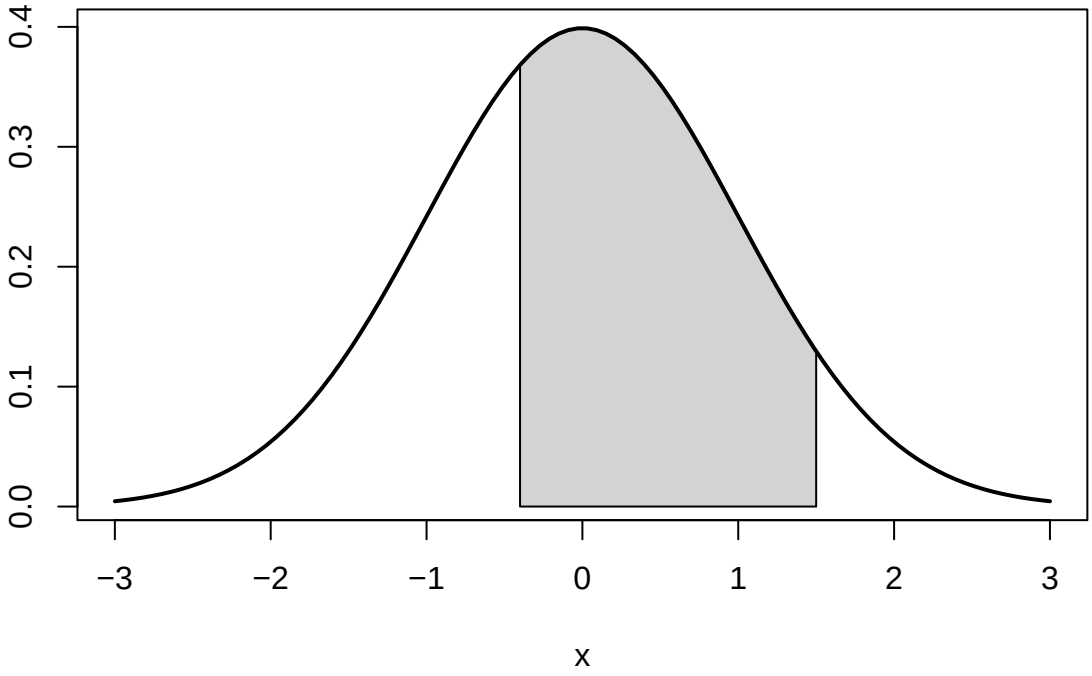

(c) $-0.4 < Z < 1.5$

area under the curve 58.8%

```
pnorm(1.5) - pnorm(-0.4)
```

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

```
normal_area(mean = 0, sd = 1, lb = -0.4, ub = 1.5, lwd = 2)
```



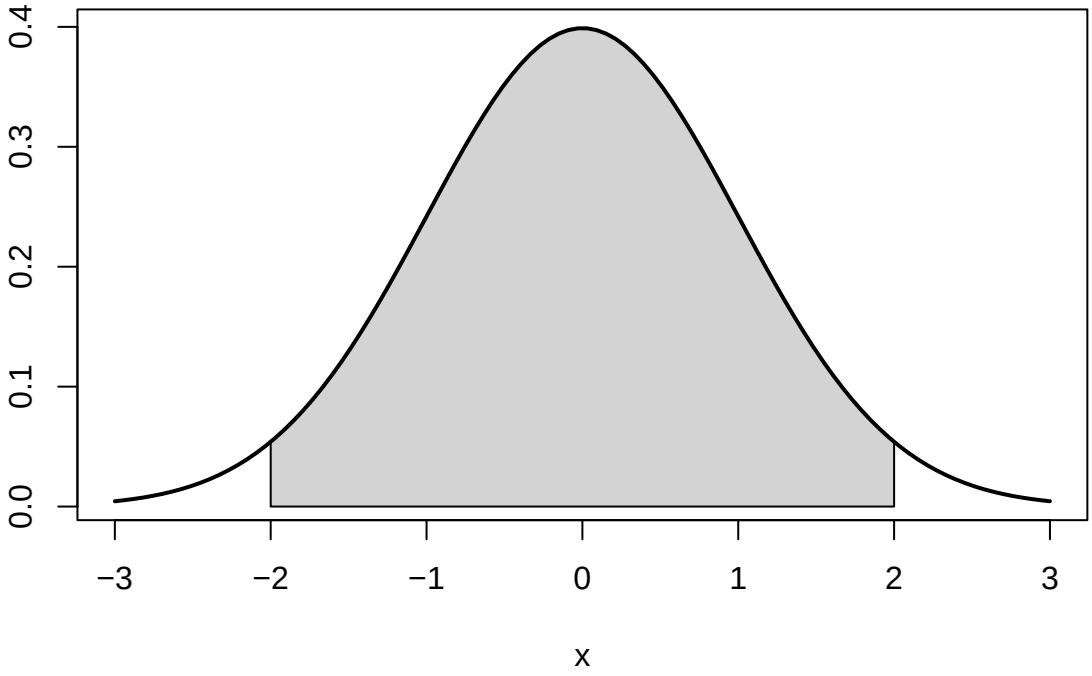

(d) $|Z| > 2$

area under the curve 4.6%

```
pnorm(-2) + 1 - pnorm(2)
```

```
## [1] 0.04550026
```

```
normal_area(mean = 0, sd = 1, lb = -2, ub = 2, lwd = 2)
```


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.

Men - N($\mu = 4313$ $\sigma = 538$) Women - 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?

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

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

Leo z score = 1.089 Mary z score = 0.312

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

```
## [1] 0.8619658
```

```
## [1] 0.6225814
```

**Mary did better with respect to her group. For Mary only 62.3% of the competitors had a better time compared to 86.2% for Leo.**

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

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

**13.8% of the competitors in Leo's age group had a better time.**

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

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

**37.7% of the competitors had a better time than Mary**

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

**yes - i could see the situation of elite athletes posted really fast times skewing the distribution left. Then it would be difficult to predict the percentages of observations using a normal distribution**

---

**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.

**yes - the height approximates the distribution rules for a normal distribution**

```
height_df <- data.frame( height = c(54, 55, 56, 56, 57, 58, 58, 59, 60, 60, 60, 61, 61, 62, 62, 63, 63,


mean <- mean(height_df$height)
std <- sd(height_df$height)
n <- length(height_df$height)

# calc number of observations 1 standard deviation from the mean
hb <- 1 * std + mean
lb <- -1 * std + mean


height_df %>%
    filter(height > lb, height < hb) %>%
    summarise(
        count = n(),
        percent = count / n
    )
```

```
##   count percent
## 1    17    0.68
```

```
# calc number of observations 2 standard deviation from the mean
hb <- 2 * std + mean
lb <- -2 * std + mean

height_df %>%
    filter(height > lb, height < hb) %>%
    summarise(
        count = n(),
        percent = count / n
    )
```

```
##   count percent
## 1    24    0.96
```

```
# calc number of observations 3 standard deviation from the mean
hb <- 3 * std + mean
lb <- -3 * std + mean

height_df %>%
    filter(height > lb, height < hb) %>%
```

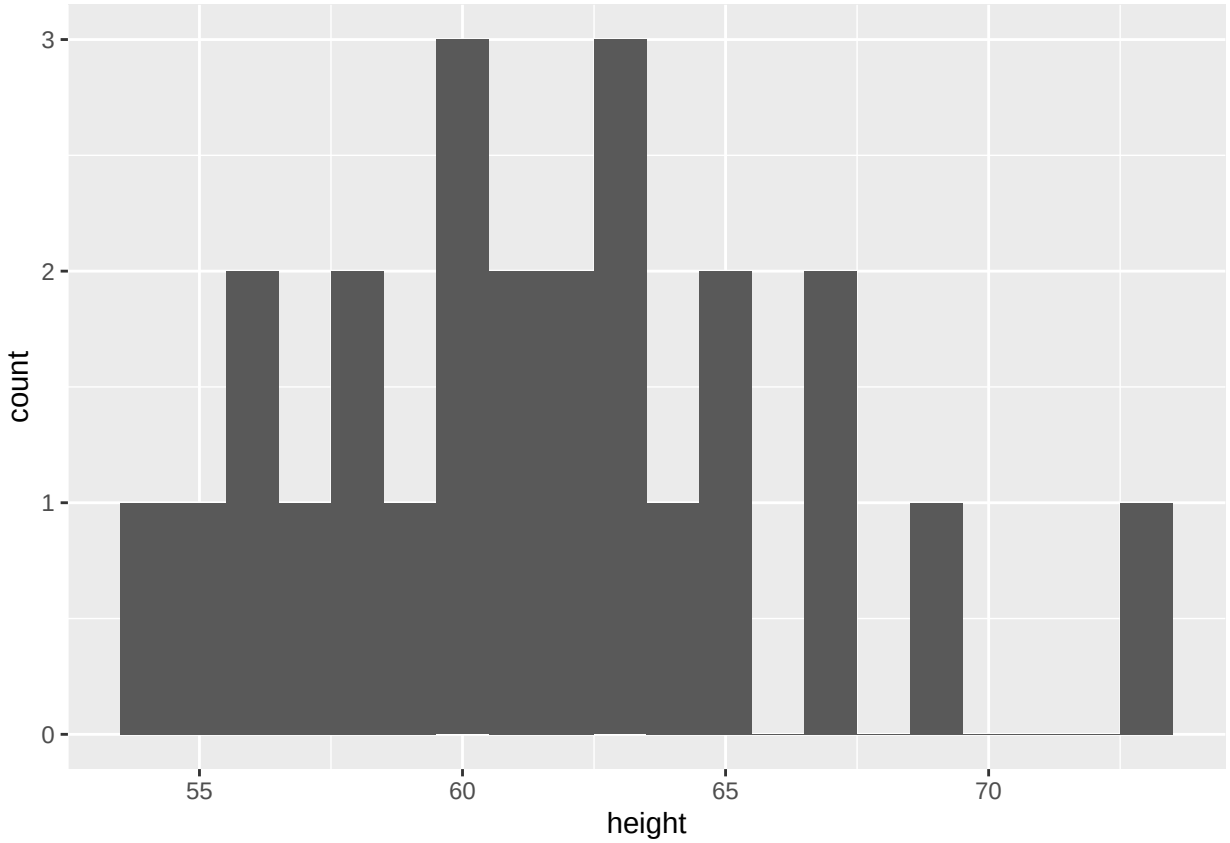
```
    summarise(
        count = n(),
        percent = count / n
    )
```

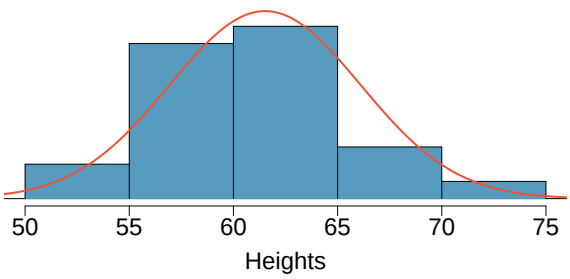
```
##   count percent
## 1    25       1
```

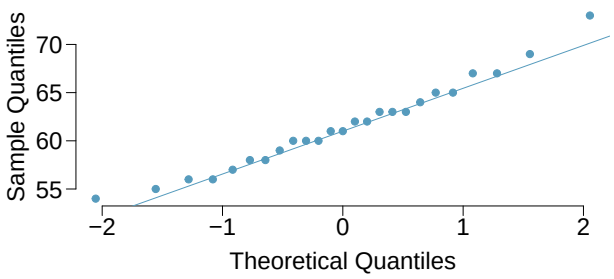
(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.

**no - based on the graph it appears that the data is right skewed**

```
ggplot(data = height_df , aes(x = height)) +
geom_histogram(bins = 20)
```

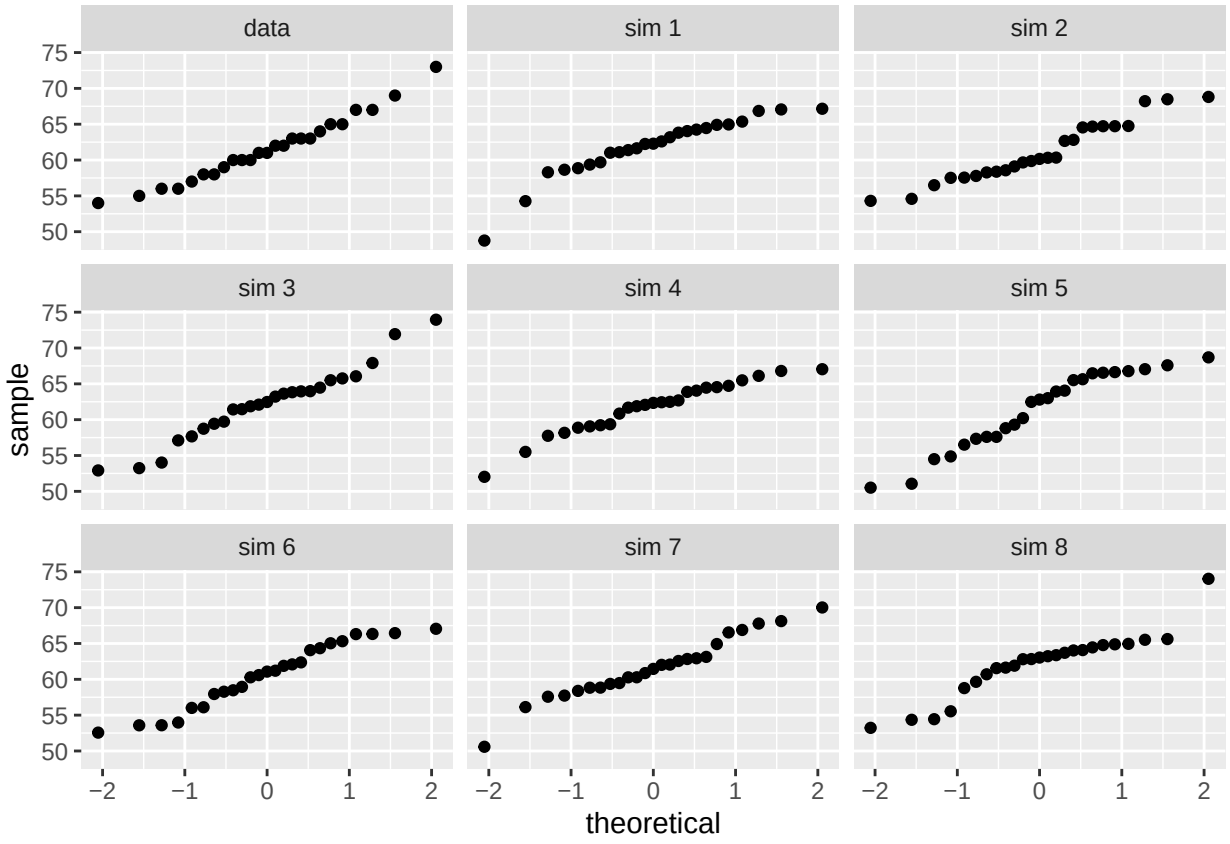


```r
# Use the DATA606::qqnormsim function

qqnormsim(sample = height, data = height_df)
```




---

**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?

```
prob <- 0.02
n <- 10
prob * (1-prob)^(n-1)
```

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

**1.7%**

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

```
prob <- 0.02
n <- 100
(1 - prob)^n
```

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

**13.3%**

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

```
p <- 0.02
k <- 1

k/p
```

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

```
(((1-p)/p^2))^.5
```

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

**average transistors before defect 50 standard deviation 49.5**

(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
k <- 1

k/p
```

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

```
(((1-p)/p^2))^.5
```

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

**average transistors before defect 20 standard deviation 19.5**

(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?

**as expected the increased probability reduces the number of transistores produced before you would expect a defect. It also reduces the variability of the process.**

---

**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
n <- 3
k <- 2

factorial(n) / factorial(k) / factorial(n-k) *p^k * (1-p)^(n-k)
```

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

**38.2%**

(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.

1) bbg
2) bgb
3) gbb

**yes the answers from a and b are equal**

```
b <- .51
g <- 1 - b

b*b*g + b*g*b + g*b*b
```

```
## [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).

```
p <- 0.51
n <- 8
k <- 3

factorial(n) / factorial(k) / factorial(n-k)
```

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

**to use the approach in b we would need to identify all the different permutations of 3 out of 8 kids being boys. the calculation of all 56 permutations that would be required**

---

**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?

```
p <- 0.15
n <- 9
k <- 2

factorial(n) / factorial(k) / factorial(n-k) *p^k * (1-p)^(n-k) * p
```

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

**3.9%**

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

**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?

**in part b we are assuming that the she has already successfully served 2 balls into the court and we are only concerned with the probably of a successful 10th serve. Since the serves are independent we can ignore the events that happended before the 10th attempt**