Chapter 4 - Distributions of Random Variables

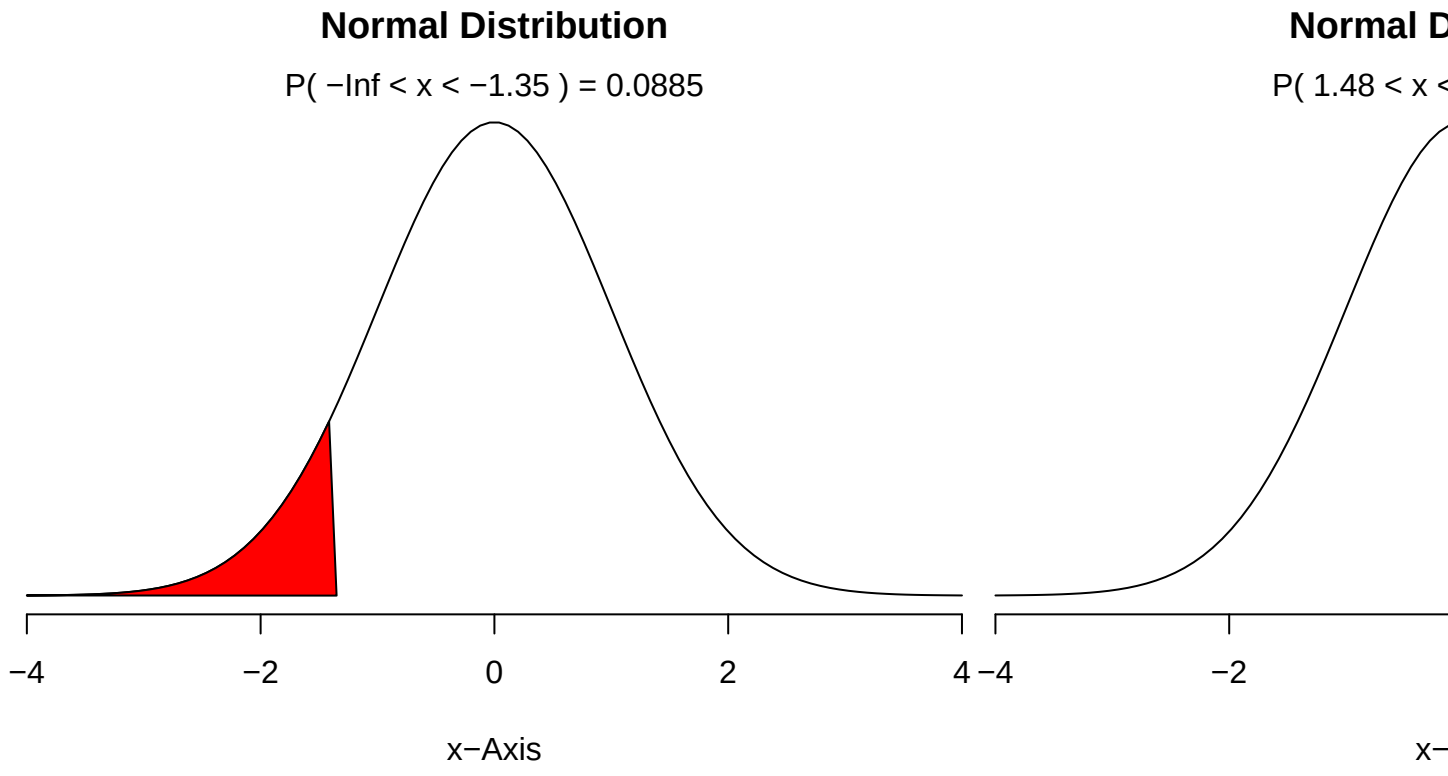
Chunjie Nan

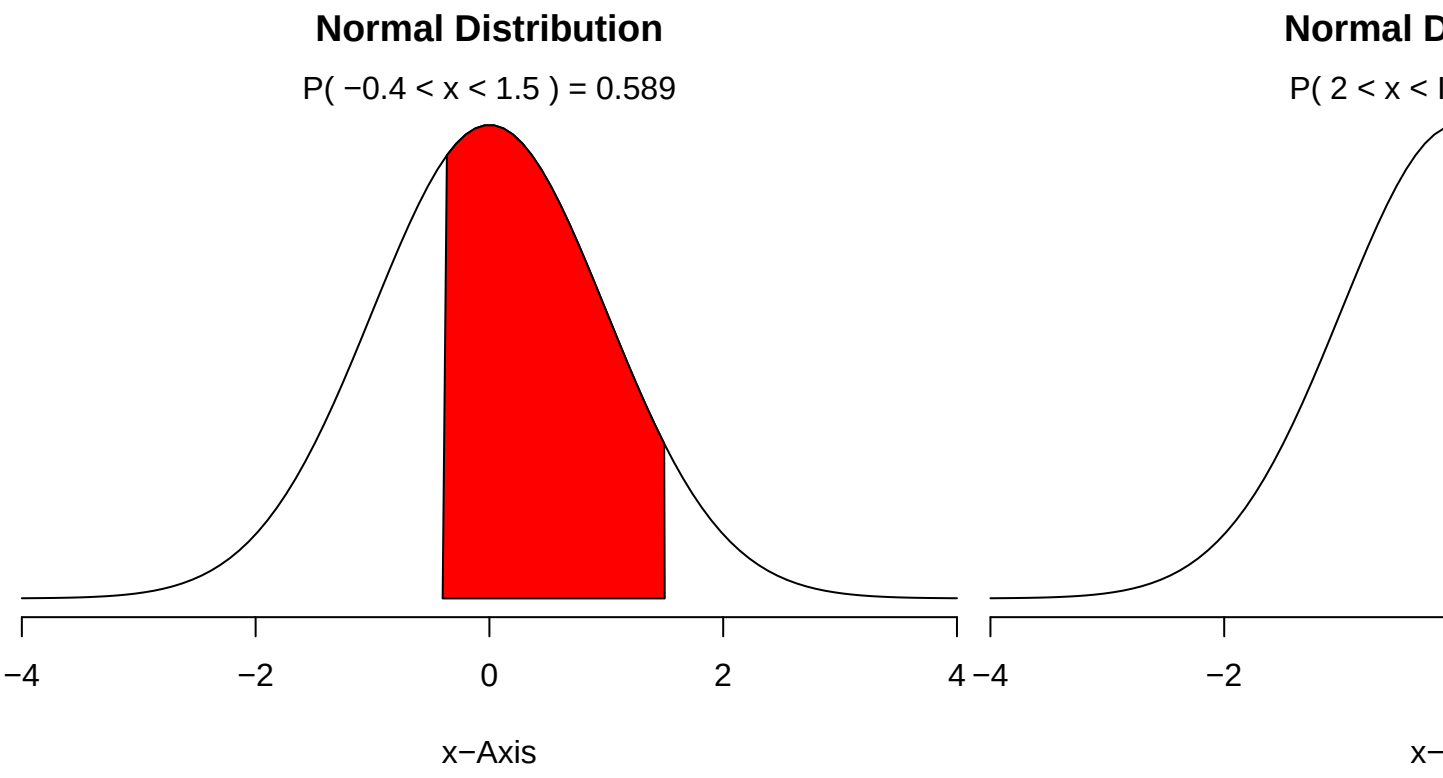
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

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

```
## Loading required package: shiny

## Loading required package: openintro

## Please visit openintro.org for free statistics materials

##
## Attaching package: 'openintro'

## The following objects are masked from 'package:datasets':
##
##     cars, trees

## Loading required package: OIdata

## Loading required package: RCurl

## Loading required package: maps

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:openintro':
##
##     diamonds

## Loading required package: markdown
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.

##
## Attaching package: 'DATA606'

## The following object is masked from 'package:utils':
##
##     demo
```




Normal Distribution
P(−0.4 < x < 1.5) = 0.589
−4
−2
0
2
4
x−Axis
Normal D
P(2 < x < I
−4
−2
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(μ=4313,sd=583) Women: N(μ=5621,sd=807)

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

```
z_for_leo <- (4948-4313)/583
z_for_leo
```

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

```
z_for_mary<-(5513-5261)/807
z_for_mary
```

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

Answer: Leo is 1.089 above the mean, while Mary is 0.31 above the mean.

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning. Answer: Since the shorter time means faster, so Mary did better in this case.

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

```
pnorm(1.09,lower.tail = FALSE)
```

```
## [1] 0.1378566
```

Answer: 13.79% faster than in his group.

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

```
pnorm(0.31,lower.tail=FALSE)
```

```
## [1] 0.3782805
```

Answer: 37.82% 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.

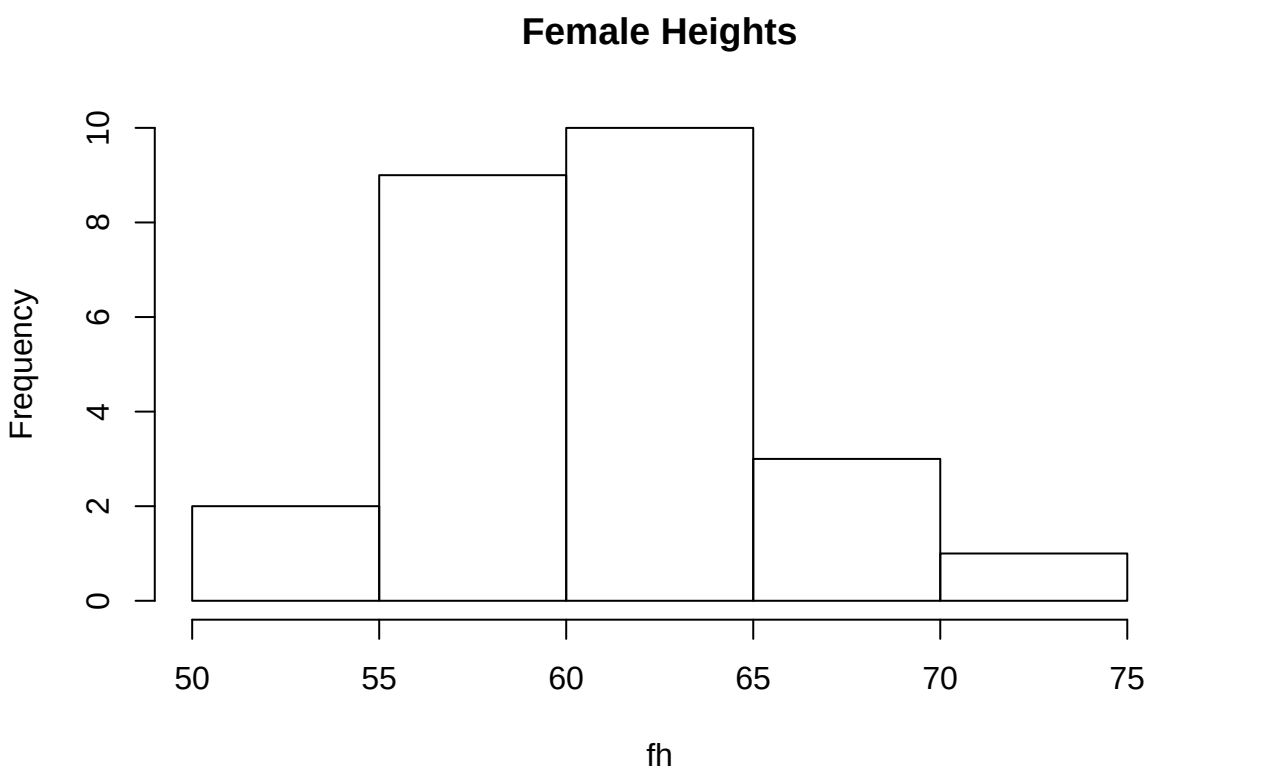
Answer: The result won't be change because all the participants shares the same distribution model. Even if the distribution is normal, the percentile will not change within the model.

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

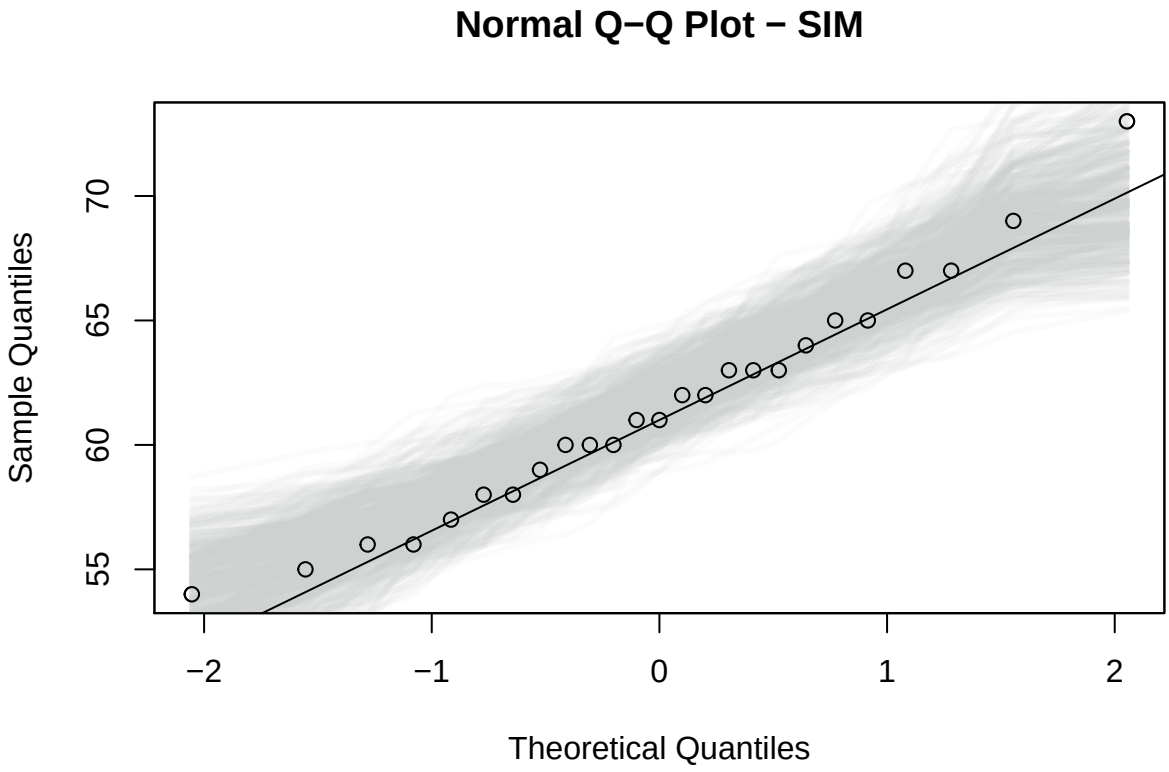
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

(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.

```
fh <- 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,
hist(fh, main="Female Heights")
```




```
qqnormSim((fh))
```




```
pnorm(61.52+4.58, mean=61.52, sd=4.58)
```

```
## [1] 0.8413447
```

```
pnorm(61.52+2*4.58,mean=61.52,sd=4.58)
```

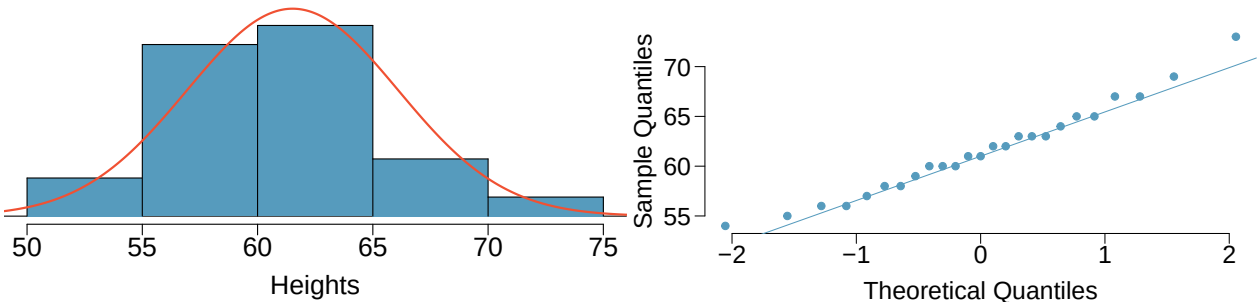
```
## [1] 0.9772499
```

```
pnorm(61.52+3*4.58,mean=61.52,sd=4.58)
```

```
## [1] 0.9986501
```

Answer: It doesn't follow the 68-95-99.7 rule because the distribution is left skewed.

(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below. Answer: No,even thought the outliers are stay in the gray area, the histogram shows that the distribution left skewed.



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?

```
pgeom(9,0.02)
```

```
## [1] 0.1829272
```

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

```
pgeom(100,0.02)
```

```
## [1] 0.8700328
```

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

```
1/0.02
```

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

```
sqrt((1-0.02)/0.02^2)
```

```
## [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?

```
1/0.05
```

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

```
sqrt((1-0.05)/0.05^2)
```

```
## [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?

Answer: Increase the probability of success, decrease the wait time for success and decreases the spread in the distribution.

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.

```
dbinom(2,3,0.51)
```

```
## [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.

```
(0.51^2)*0.49*3
```

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

Answer: It is confirmed that the result from(a) an (b) are matched with the number 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).

Answer: the senario of a couple who plans to have 8 kids will have 3 is way more tedious than the part b since part b in the question is structured with simple combinations.

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?

```
choose(9,2)*(0.15^3)*(0.85^7)
```

```
## [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?

Answer: Each attempt has the same probability because they are independent. so, the probability is 15% as well.

(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?

Answer: The reason is becuase for part(b), we don't care about the probabilities of other attempts, we just focus on the 10th of independent attempt. However, the part(a) we put more consideration on the 3rd success serve that exactly happens on the 10th attempt. This is how the question of the two different parts made the results different.