HW_4_Normal Distribution

Salma Elshahawy

9/28/2019

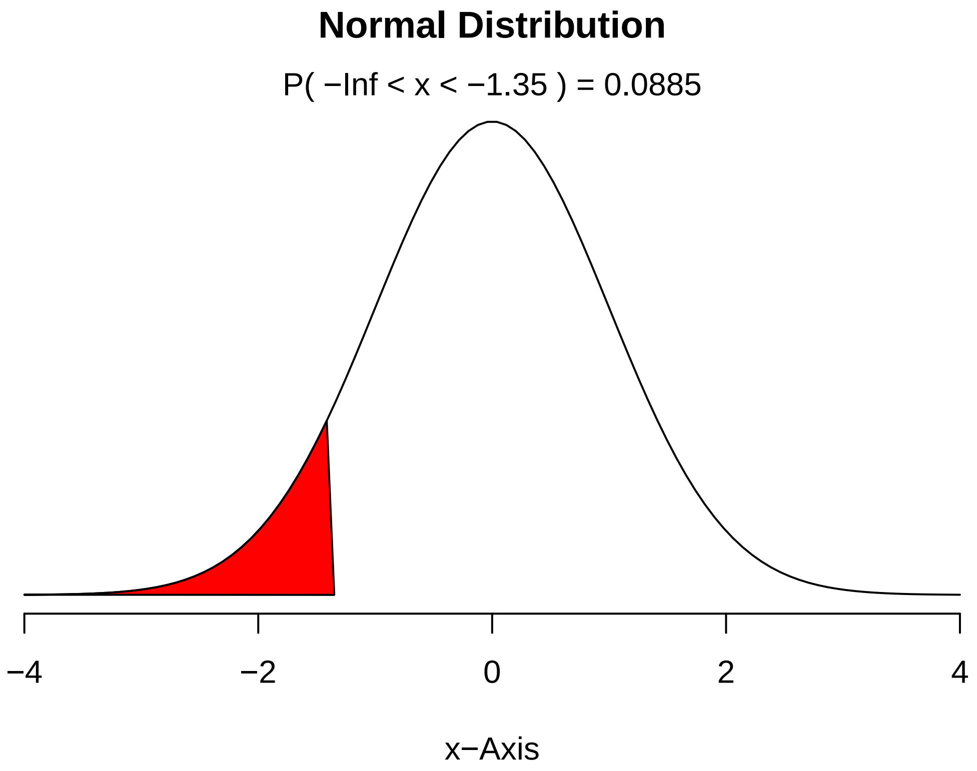
Chapter_4 Distributions of random variables

Area under the curve, Part I. What percent of a standard normal distribution N(= 0, = 1) is found in each region? Be sure to draw a graph. (a) Z <−1.35 => 8.85% (b) Z >1.48 => 6.94% (c) −0.4< Z <1.5 => 58.9% (d) |Z|>2

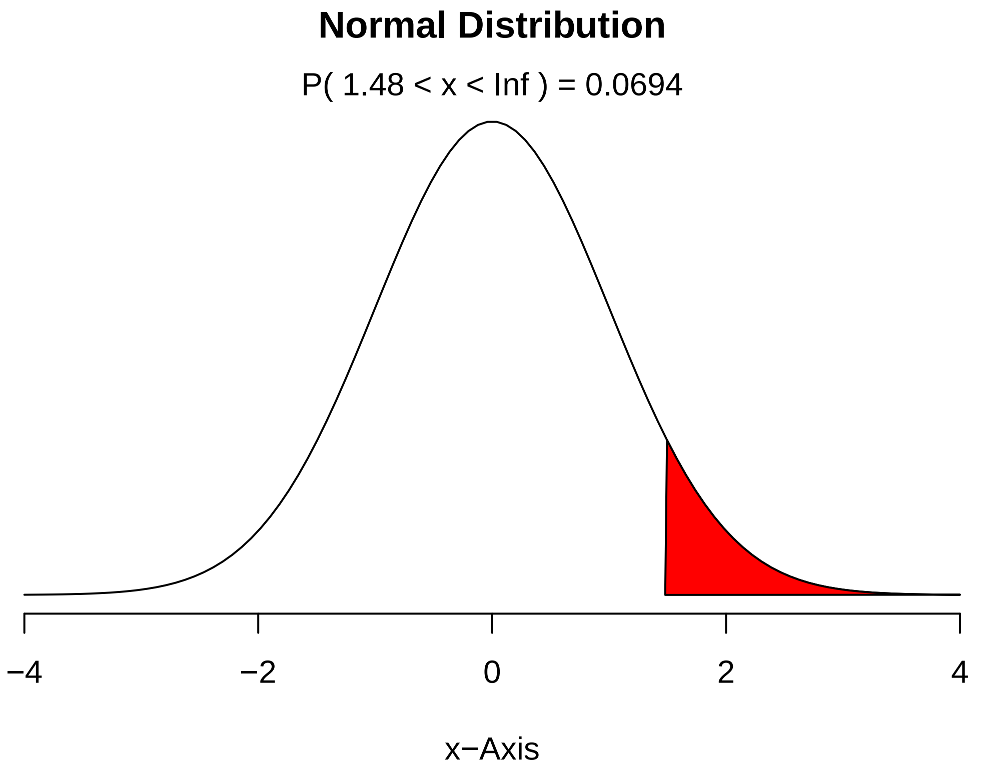
```r
library(utils)
library(DATA606)
```

```
##
## 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.
```

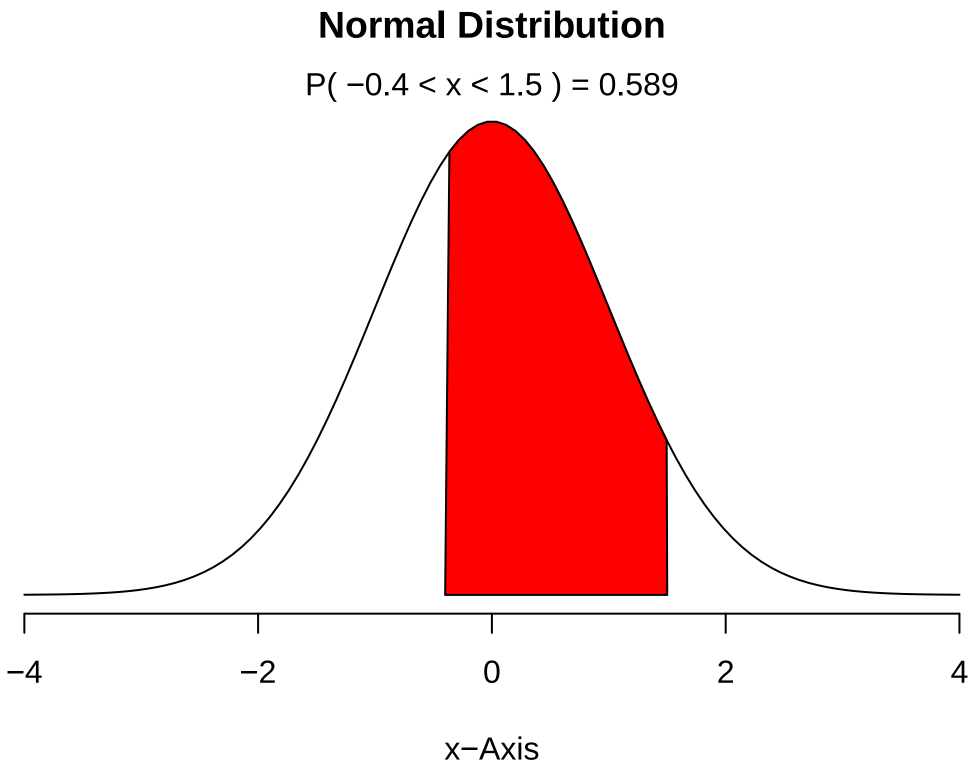
```r
normalPlot(mean = 0, sd = 1,bounds = c(-Inf, -1.35), tails = FALSE)
```



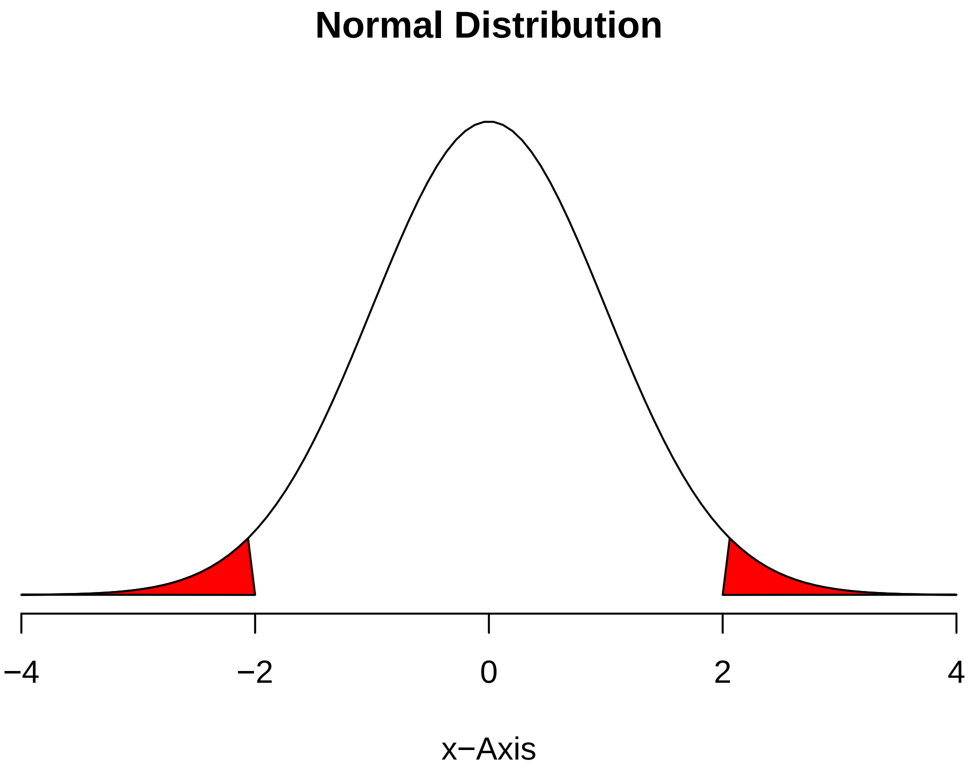
```r
normalPlot(mean = 0, sd = 1,bounds = c(1.48, Inf), tails = FALSE)
```



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



```r
normalPlot(mean = 0, sd = 1,bounds = c(-2,2), tails = TRUE)
```



Triathlon times, Part I. 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.

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.

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

```
Leo_Z <- (4948-4313)/583
round(Leo_Z, 2)
```

```
## [1] 1.09
```

```
Mary_Z <- (5513-5261)/807
round(Mary_Z, 2)
```

```
## [1] 0.31
```

The Z-score for Leo's finishing time is 1.09 which means he finished 1.09 standard deviations above the mean for his age group.

The Z-score for Mary's time is 0.31 which means she finished 0.31 standard deviations above the mean for her age group.

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

Leo ranked better than Mary for his age group since his Z-score was much higher than Mary's. The higher the Z-score the higher the percentile rank.

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

```
pnorm(Leo_Z)*100
```

```
## [1] 86.19658
```

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

```
pnorm(Mary_Z)*100
```

```
## [1] 62.25814
```

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

Yes!, I still want to calculate a Z-score, but I cannot use Z-scores to estimate probabilities and percentile ranks in a non-normal distribution. In that case I would need to know more information about the distribution than just the mean and standard deviation. If I had the total number of competitors in their age groups and the number of competitors who did better or worse than Leo and Mary then I could calculate a percentile rank.

Heights of female college students

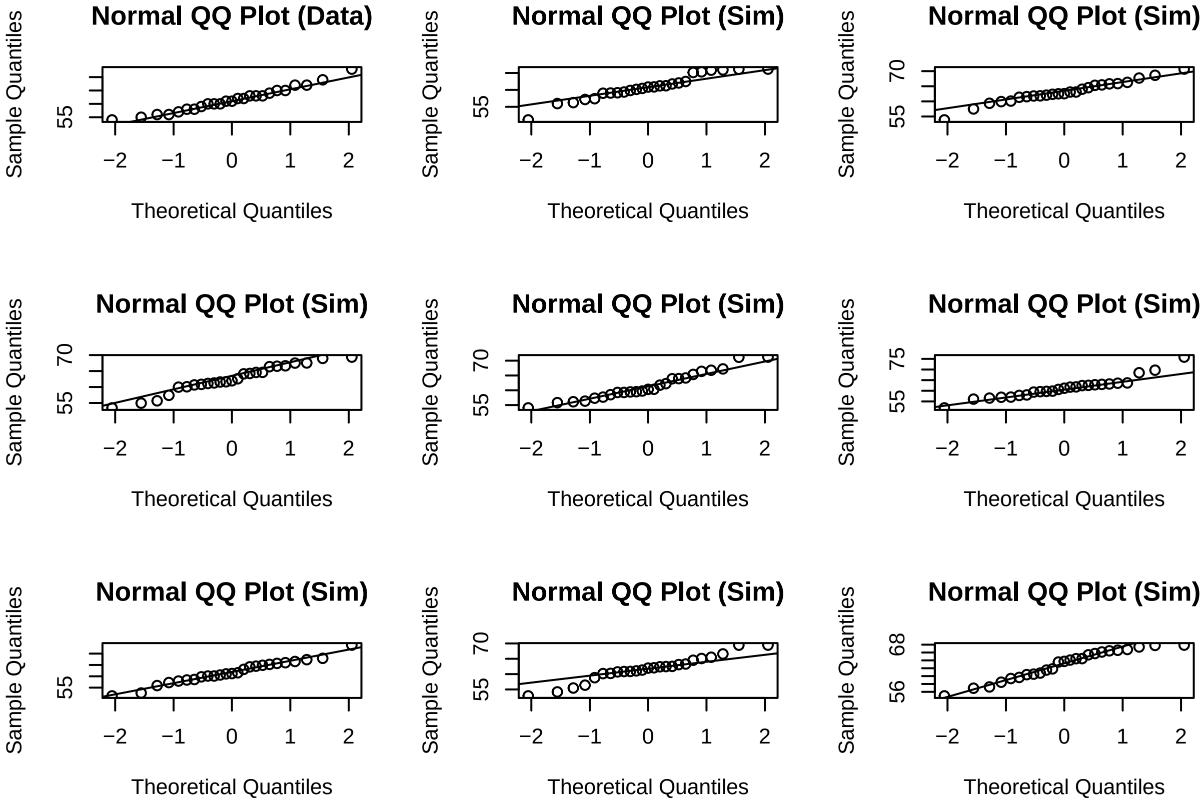
```
height <- c(54,55,56,56,57,58,58,59,60,60,60,61,61,62,62,63,63,63,64,65,65,67,67,69,73)
height
```

```
##  [1] 54 55 56 56 57 58 58 59 60 60 60 61 61 62 62 63 63 63 64 65 65 67 67
## [24] 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 heights of female college students in this dataset do seem to follow the 68-95-99.7% Rule since they appear to be normally distributed. In fact they follow a normal pattern even more closely than the simulations.

```r
qqnormsim(height)
```



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

HistVsQQPlot Yes the data represented by the graphs above appear to follow a normal distribution. It's a little harder to tell based on the histogram than on the normal probability plot. In the histogram the shape created by the tops of the bars roughly follows the normal curve, but in the normal probability plot, it's much clearer that the data follows a straight line.

Defective rate 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?

```r
p <- .02
n <- 10
P <- (1-p)^(n-1)*p
round(P, 3)
```

```
## [1] 0.017
```

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

```
round((1-p)^100, 3)
```

```
## [1] 0.133
```

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

```
EV <- 1/p
EV
```

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

```
sd <- sqrt((1-p)/p^2)
round(sd, 2)
```

```
## [1] 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 <- .05
EV <- 1/p
EV
```

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

```
sd <- sqrt((1-p)/p^2)
round(sd, 1)
```

```
## [1] 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?

Increasing the probability of an event (that is a Bernoulli random variable) makes both the expected value and the standard deviation lower.

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.

```
# binomial distribution
p = .51
P = round(dbinom(2,3,p),3)
P
```

```
## [1] 0.382
```

```
# double check
n <- 3
k <- 2
p <- .51
factorial(n)/(factorial(k)*factorial(n-k))*p^k*(1-p)^(n-k)
```

```
## [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.

```
# the addition rule for disjoint outcomes
P <- .51*.51*.49 + .51*.49*.51 + .49*.51*.51
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).

```
# number of ways you can get k successes in n trials
fc=function(n,k){factorial(n)/(factorial(k)*factorial(n-k))}
fc(8, 3)
```

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

Serving in volleyball 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?

```
# first find the probability of exactly 2 successes in 9 trials
p = .15
P = dbinom(2,9,p)
# then multiply by the probability that she will be successful on her 10th try
round(P * .15,3)
```

```
## [1] 0.039
```

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

The probability that her 10th serve will be successful is still 0.15. The gambler's fallacy makes you think that the outcome of prior trials might have an affect on the next trial, but if the trials are independent, each trial has exactly the same probability as every other trial.

(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?

Part (b) only asks about the probability of the 10th trial whereas part (a) asks about the joint probability of all of the first 10 trials.