MichaelY_HW3_Distributions

Michael Y.

March 3, 2019

```
knitr::opts_knit$set(root.dir = "c:/users/Michael/DROPBOX/priv/CUNY/MSDS/201902-Spring/DATA606-Jason/Ho
##setwd("c:/users/Michael/DROPBOX/priv/CUNY/MSDS/201902-Spring/DATA606-Jason/Homework")
```

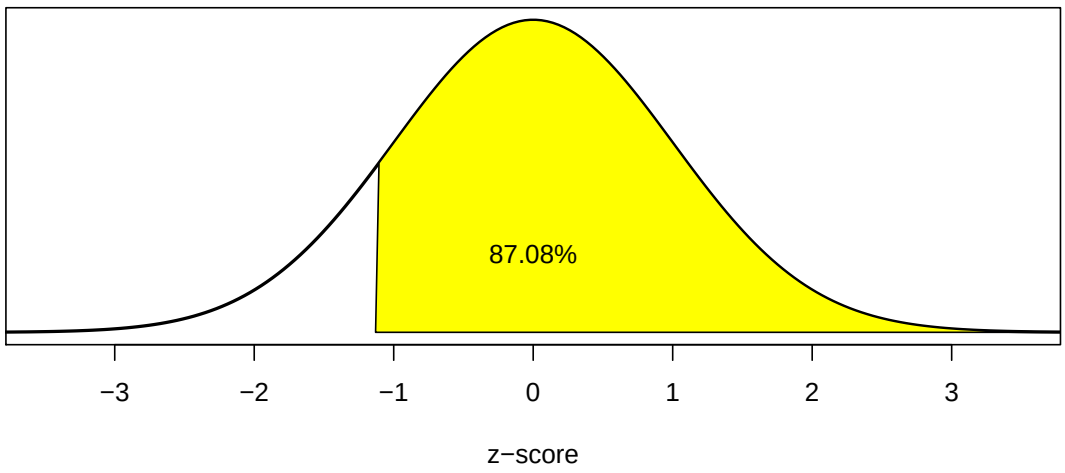
Homework - Chapter 3 - Distributions -

Exercises 3.2, 3.4, 3.18, 3.22, 3.38, 3.42 (pp.142-167)

Exercise 3.2 Area under the curve, Part II.

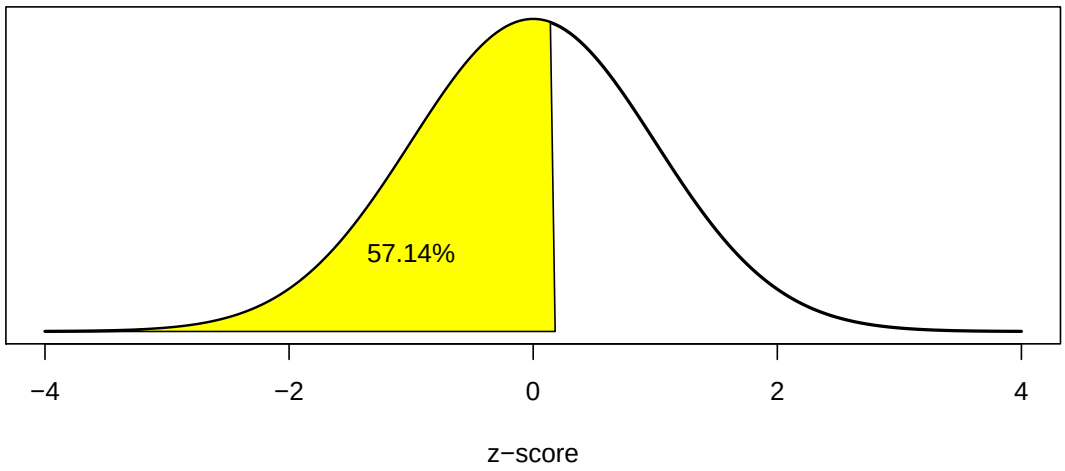
What percent of a standard normal distribution N(mu = 0, sigma = 1) is found in each region? Be sure to draw a graph.

(a) Z > -1.13



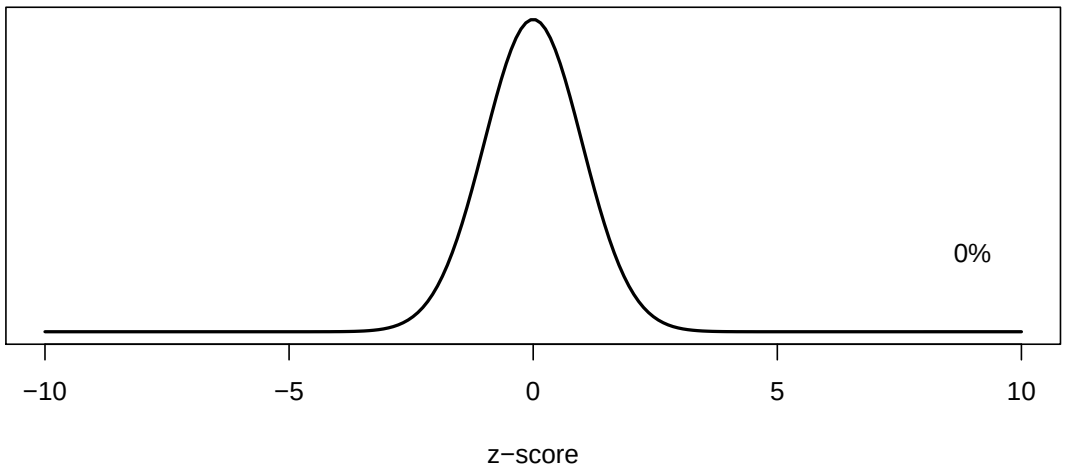
The answer is 87.08% .

(b) Z < 0.18



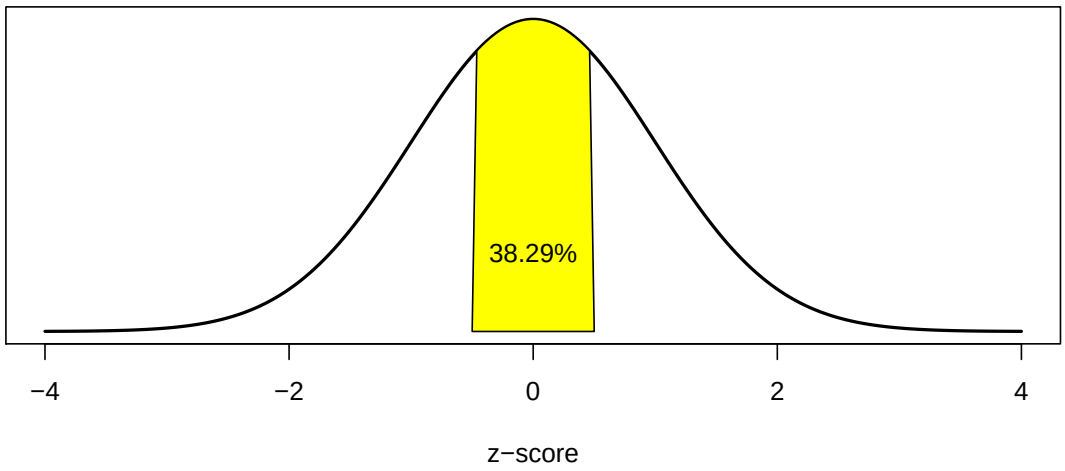
The answer is 57.14% .

(c) Z > 8



The answer is 0% . Because the result is so small, it is not easily displayed on the graph. It is the region to the right of x=8 , which is infinitesimally small.

(d) $|Z| < 0.5$



The answer is 38.29% .

Exercise 3.4 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.

$$N_{\text{men}}(\mu = 4313, \sigma = 583) = \frac{1}{583\sqrt{2\pi}} e^{-\frac{(x-4313)^2}{2(583)^2}}$$

$$N_{\text{women}}(\mu = 5261, \sigma = 807) = \frac{1}{807\sqrt{2\pi}} e^{-\frac{(x-5261)^2}{2(807)^2}}$$

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

```
leo_zscore <- (4948-4313)/583
print(paste("Leo's Z-score is " , leo_zscore))
```

```
## [1] "Leo's Z-score is  1.08919382504288"
```

```
mary_zscore <- (5513-5261)/807
print(paste("Mary's Z-score is " , mary_zscore))
```

```
## [1] "Mary's Z-score is  0.312267657992565"
```

What do these Z-scores tell you?

Leo ran more than 1 standard deviation *slower* than the men in his age group, while Mary ran only one-third of a standard deviaton slower than the women in her age group. Thus, mary ran *comparatively* faster than Leo.

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

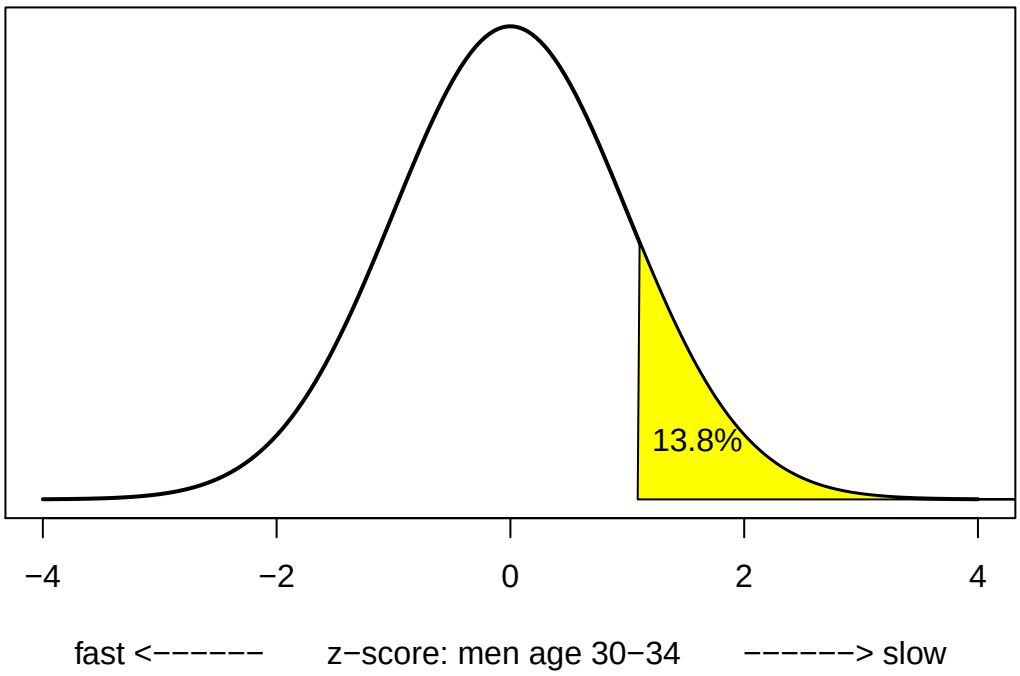
Both Leo and Mary are in the bottom half of their respective groups, as their times are above the respective mean times for their groups. Because Mary's time is closer to the mean for her group, she ranks better than does Leo, who is closer to the bottom of his respective pack.

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

```
leo_quantile = 1 - pnorm(leo_zscore)
leo_percentile = round(100*leo_quantile,2)

resultleo = pnorm(q= leo_zscore,mean = 0, sd = 1, lower.tail = F)
pctresultleo = as.character(paste0(round(100*resultleo,2),"%"))

x <- seq(-4,4,length=200)
y <- dnorm(x,mean=0, sd=1)
plot(x, y, type = "l", lwd = 2,
     xlim = c(-4,4),
     ylab='', xlab='fast <------       z-score: men age 30-34       ------> slow', yaxt='n')
lb <- leo_zscore; ub <- 10
i <- x >= lb & x <= ub
polygon(c(lb,x[i],ub), c(0,y[i],0), col="yellow")
text(1.6, .05, pctresultleo)
```



Leo finished faster than only 13.8% of his group.

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

```
mary_quantile = 1 - pnorm(mary_zscore)
mary_percentile = round(100*mary_quantile,2)

mary_quantile = 1 - pnorm(mary_zscore)
mary_percentile = round(100*mary_quantile,2)

resultmary = pnorm(q= mary_zscore,mean = 0, sd = 1, lower.tail = F)
pctresultmary = as.character(paste0(round(100*resultmary,2),"%"))

x <- seq(-4,4,length=200)
y <- dnorm(x,mean=0, sd=1)
plot(x, y, type = "l", lwd = 2,
     xlim = c(-4,4),
     ylab='', xlab='fast <------       z-score: women age 25-29       ------> slow', yaxt='n')
lb <- mary_zscore; ub <- 10
i <- x >= lb & x <= ub
polygon(c(lb,x[i],ub), c(0,y[i],0), col="yellow")
text(1.0, .05, pctresultmary)
```

Mary finished faster than 37.74% of her group.

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

Explain your reasoning.

The answer to part (b) would not change as Z-scores can be calculated for distributions that are not normal.

However, we could not answer parts (d)-(e) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

Exercise 3.18 Heights of female college students.

Below are heights of 25 female college students:

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

Note: **The supplied data set "fheights" contains only 24 of the above 25 items – it omits a repeated value ("65") which is listed in the testbook. Accordingly the results would not correspond to the values shown in the textbook if the (incorrect) supplied data set is used.**

(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.

```
heights = c(54, 55, 56, 56, 57, 58, 58, 59, 60, 60, 60, 61, 61, 62, 62, 63, 63, 63, 64, 65, 65, 67, 67,
mu = mean(heights)
sigma = sd(heights)
sd1 = c(mu-sigma,mu+sigma)
sd2 = c(mu-2*sigma,mu+2*sigma)
sd3 = c(mu-3*sigma,mu+3*sigma)

sd1heights = heights[heights>sd1[1]&heights<sd1[2]]
sd1pct = length(sd1heights)/length(heights)
print(paste0("One sd : ", as.character(round(100*sd1pct,2)),"%"))
```

```
## [1] "One sd : 68%"
```

```
sd2heights = heights[heights>sd2[1]&heights<sd2[2]]
sd2pct = length(sd2heights)/length(heights)
print(paste0("Two sd : ", as.character(round(100*sd2pct,2)),"%"))
```

```
## [1] "Two sd : 96%"
```

```
sd3heights = heights[heights>sd3[1]&heights<sd3[2]]
sd3pct = length(sd3heights)/length(heights)
print(paste0("Three sd : ", as.character(round(100*sd3pct,3)),"%"))
```

```
## [1] "Three sd : 100%"
```

Yes, the values do approximately follow the 68-96-99.7% rule.

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

The distribution is unimodal and symmetric.

The superimposed normal curve seems to approximate the distribution pretty well.

The points on the qq-plot also seem to follow a straight line.

There is one possible outlier on the upper end that is apparent in both graphs, but it is not too extreme. (At 73 inches, she may be a recruit for the basketball team.)

We can say that the distribution is nearly normal.

Exercise 3.22 - 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?

```
prob_each_defective = 0.02
prob_each_not_defective = 1 - prob_each_defective

prob_nine_not_defective = prob_each_not_defective^9
prob_tenth_is_first_defect = prob_nine_not_defective * prob_each_defective
prob_tenth_is_first_defect
```

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

The answer is 0.016675 .

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

```
prob_100_not_defective = prob_each_not_defective^100
prob_100_not_defective
```

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

The answer is 0.1326196 .

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

```
mu02 = 1 / prob_each_defective
mu02
```

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

What is the standard deviation?

```
sigma02 = sqrt(prob_each_defective*prob_each_not_defective)
sigma02
```

```
## [1] 0.14
```

The mean is 50 and the standard deviation is 0.14 .

(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?

```
prob_defective = 0.05
prob_not_defective = 1 - prob_defective
mu05 = 1 / prob_defective
mu05
```

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

What is the standard deviation?

```
sigma05 = sqrt(prob_defective*prob_not_defective)
sigma05
```

```
## [1] 0.2179449
```

The mean is 20 and the standard deviation is 0.2179449 .

(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 causes the mean to decrease (as it is 1/p) and the standard deviation to increase (as it is sqrt(p*(1-p))) .

Exercise 3.38 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_boy = 0.51
p_girl = 1- p_boy
n = 3
k = 2

p_two_boys_of_three = choose(n,k) * (p_boy^k) * p_girl^(n-k)
p_two_boys_of_three
```

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

The answer is 0.382347 .

(b) Write out all possible orderings of 3 children, 2 of whom are boys.

(1) G,B,B

(2) B,G,B

(3) B,B,G

Use these scenarios to calculate the same probability from part (a) but using the addition rule for disjoint outcomes.

```
p_girl_first = p_girl * p_boy * p_boy
p_girl_middle = p_boy * p_girl * p_boy
p_girl_last = p_boy * p_boy * p_girl

p_sum = p_girl_first + p_girl_middle + p_girl_last
p_sum
```

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

The answer is 0.382347 .

Confirm that your answers from parts (a) and (b) match.

```
p_two_boys_of_three ==  p_sum
```

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

Confirmed.

(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).

```
choose(8,3)
```

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

The approach from part (b) would require evaluation of 56 cases, which would be quite tedious.

Part (a) would simply follow the binomial formula.

Exercise 3.42 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?

This scenario means that she has to make 2 successful serves out of the first 9 attempts, then she has to have a success on the 10th attempt for her third overall success.

```
p_success = .15
p_fail = 1 - p_success
n = 9
k = 2
p_two_of_nine_successes = choose(n,k) * p_success^k * p_fail^(n-k)
p_two_of_nine_successes
```

```
## [1] 0.2596674
```

```
### dbinom:
dbinom(x = 2, size = 9, prob = .15)
```

```
## [1] 0.2596674
```

```
p_third_success_on_tenth_trial = p_two_of_nine_successes * p_success
p_third_success_on_tenth_trial
```

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

```
### dnbinom (Negative Binomial):
dnbinom(x = 10-3, size = 3, prob = .15)
```

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

The answer is 0.0389501 .

This illustrates the Negative Binomial distribution (where above we have already subtracted 1 from each of n and k.)

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

Because each serve is independent, the probability that her 10th serve will be successful is 0.15 – the same probability for each individual serve.

(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 (a), we don't yet know when she would have successes or failures on her serves. After the first 9 serves, she has only a 0.2596674 of having had exactly two successful serves. She might have more than two, or fewer than two. If she does not have exactly two successes in

the first nine trials, then the outcome of the 10th serve is irrelevant for this event, because it cannot be her third success.

On the other hand, in part (b), we already know that she has had the requisite 2 successes in the first 9 serves. Because of the assumed independence of success on each individual serve, the probability that her next serve could be the third success is the same as the probability that any individual serve is successful.