DS606 Homework 3 Jagdish Chhabria

Jagdish Chhabria

March 1, 2019

Chapter 3 - Distributions of Random Variables Graded: 3.2 (see normalPlot), 3.4, 3.18 (use qqnormsim from lab 3), 3.22, 3.38, 3.42

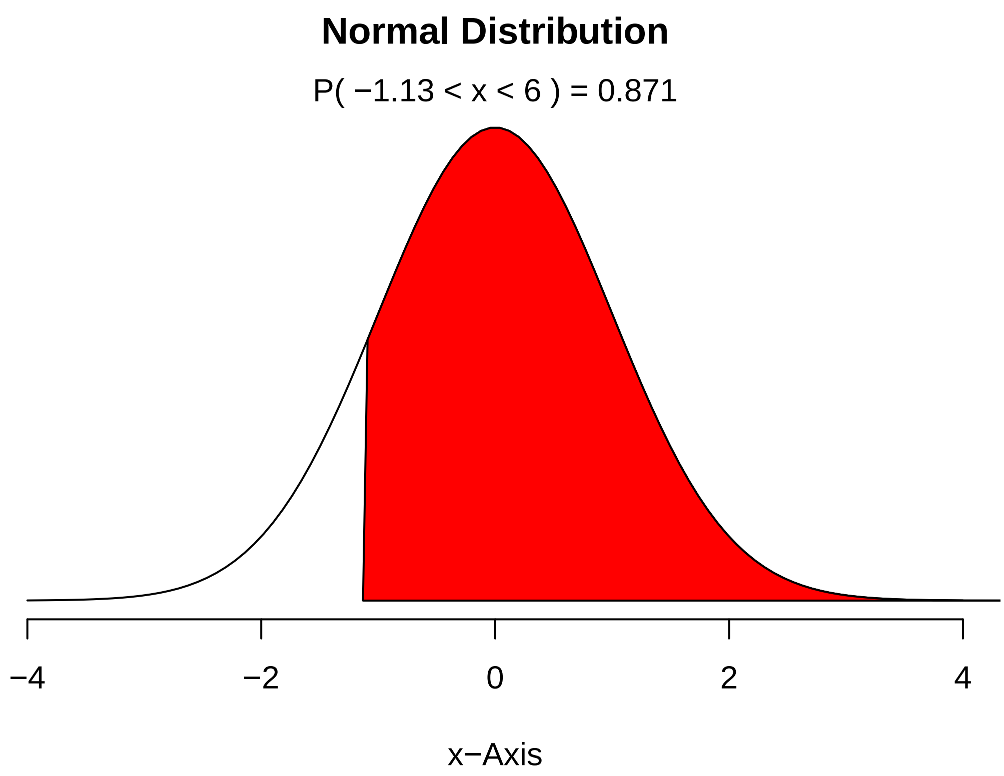
3.2 Area under the curve, Part II. What percent of a standard normal distribution N(?? = 0,! = 1) is found in each region? Be sure to draw a graph.

(a) Z > ???1.13 (b) Z < 0.18 (c) Z > 8 (d) |Z| < 0.5

```
#(a) Z > ???1.13
1-pnorm(-1.13)
```

```
## [1] 0.8707619
```

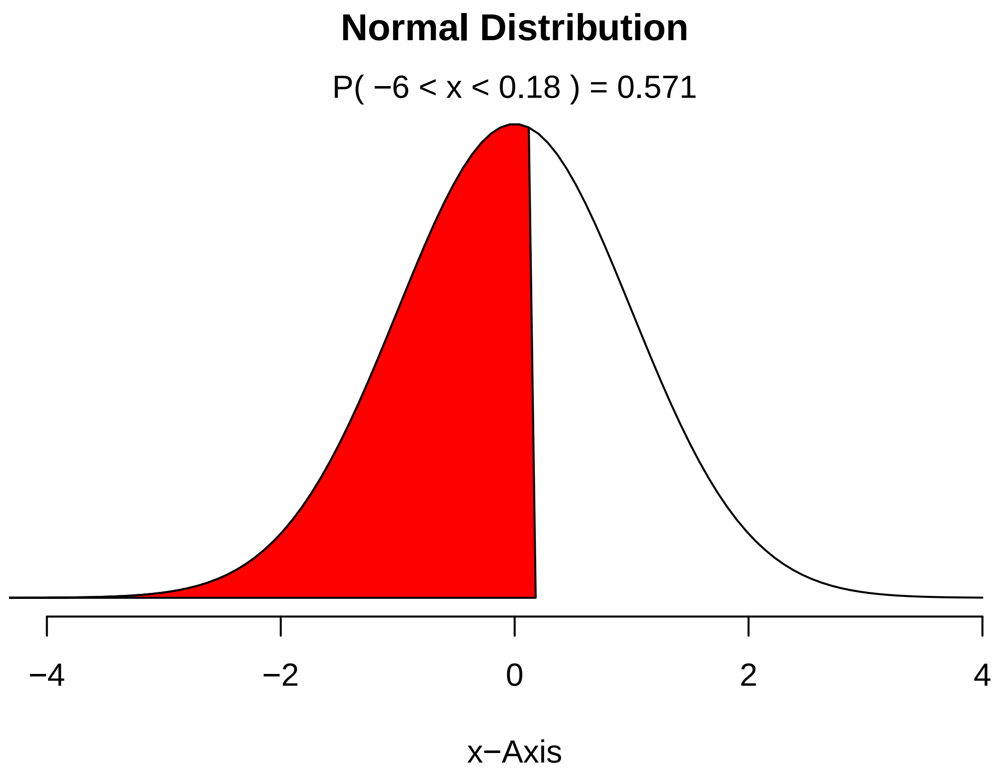
```
normalPlot(mean = 0, sd = 1, bounds = c(-1.13, 6), tails = FALSE)
```




```
# (b) Z < 0.18
pnorm(0.18)
```

```
## [1] 0.5714237
```

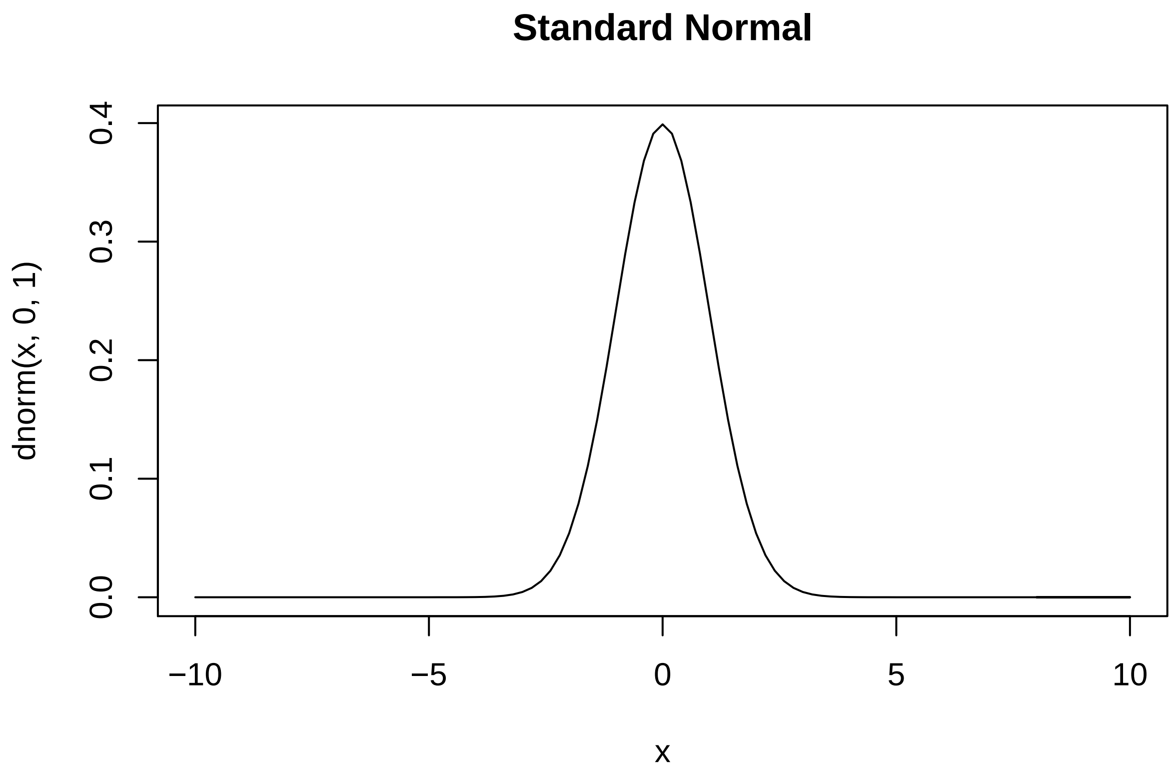
```
normalPlot(mean = 0, sd = 1, bounds = c(-6,0.18), tails = FALSE)
```




```
# (c) Z > 8
1-pnorm(8)
```

```
## [1] 6.661338e-16
```

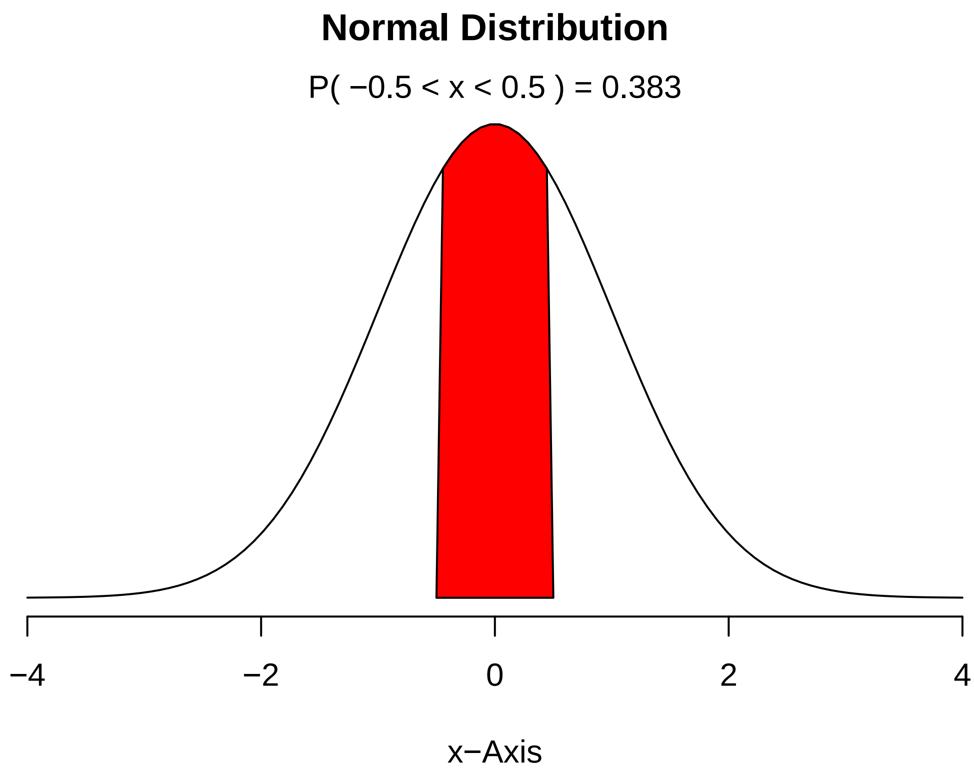
```
# This area is very, very close to 0
#pnorm(8,lower.tail = FALSE)
cord.x <- c(8,seq(8,10,0.1),10)
cord.y <- c(0,dnorm(seq(8,10,0.1)),0)
# Make a curve
curve(dnorm(x,0,1), xlim=c(-10,10), main='Standard Normal')
# Add the shaded area.
polygon(cord.x,cord.y,col='red')
```




```
# (d) |Z| < 0.5
#pnorm(0.5, lower.tail = TRUE)
#pnorm(-0.5, lower.tail=TRUE)
pnorm(0.5, lower.tail = TRUE) - pnorm(-0.5, lower.tail=TRUE)
```

```
## [1] 0.3829249
```

```
normalPlot(mean = 0, sd = 1, bounds = c(-0.5,0.5), tails = FALSE)
```



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.

```
#men30-34 ~ N(mu=4313, sigma=583)

#women25-29 ~ 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?

```
z.leo<-(4948-4313)/583
cat("The z score for Leo is:",z.leo,"\n")
```

```
## The z score for Leo is: 1.089194
```

```
z.mary<-(5513-5261)/807
cat("The z score for Mary is:",z.mary,"\n")
```

```
## The z score for Mary is: 0.3122677
```

```r
print("The 2 z-scores indicate that Mary placed relatively better in the womens 25-29 group that Leo di
```

```
## [1] "The 2 z-scores indicate that Mary placed relatively better in the womens 25-29 group that Leo di
```

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

The 2 z-scores indicate that Mary placed relatively better in the womens 25-29 group that Leo did in the mens 30-34 group. This is because Mary's time is less than 1 standard deviation above the mean, while Leo's time is more than 1 standard deviation above the mean. Given that this is a race, a lower time indicates a better performance.
(d) What percent of the triathletes did Leo finish faster than in his group?

```r
righttail.leo<-pnorm(1.089194,0,1, lower.tail = FALSE)
cat("The percent of triathletes that Leo finished faster than in his group is:",righttail.leo,"\n")
```

```
## The percent of triathletes that Leo finished faster than in his group is: 0.1380342
```

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

```r
righttail.mary<-pnorm(0.3122677,0,1, lower.tail = FALSE)
cat("The percent of triathletes that Mary finished faster than in her group is:",righttail.mary,"\n")
```

```
## The percent of triathletes that Mary finished faster than in her group is: 0.3774185
```

(f) If the distributions of finishing times are not nearly normal, would your answers to parts

(g) • (e) change? Explain your reasoning.

Yes, because while z-scores can be calculated for any distribution, they can only be used to make inferences about area under the curve i.e. probability of observing a given value and it's percentile if the distribution is normal or nearly normal. So in this case if the distribution was not nearly normal, making inferences about relatively better performance and percentile above or below the finish times of Leo and Mary would be incorrect.

3.18 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.

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

```
## [1] 61.52
```

```r
(heights.sd<-sd(heights))
```

```
## [1] 4.583667
```

```r
(heights.plus.1sd<-heights.mean+heights.sd)
```

```
## [1] 66.10367
```

```r
(heights.minus.1sd<-heights.mean-heights.sd)
```

```
## [1] 56.93633
```

```r
(heights.area.1sd<-pnorm(heights.plus.1sd,heights.mean,heights.sd) - pnorm(heights.minus.1sd,heights.mea
```

```
## [1] 0.6826895
```

```
#heights.plus.1sd - heights.minus.1sd
(heights.plus.2sd<-heights.mean+2*heights.sd)
```

```
## [1] 70.68733
```

```
(heights.minus.2sd<-heights.mean-2*heights.sd)
```

```
## [1] 52.35267
```

```
(heights.plus.3sd<-heights.mean+3*heights.sd)
```

```
## [1] 75.271
```

```
(heights.minus.3sd<-heights.mean-3*heights.sd)
```

```
## [1] 47.769
```

```
(heights.area.2sd<-pnorm(heights.plus.2sd,heights.mean,heights.sd) - pnorm(heights.minus.2sd,heights.mea
```

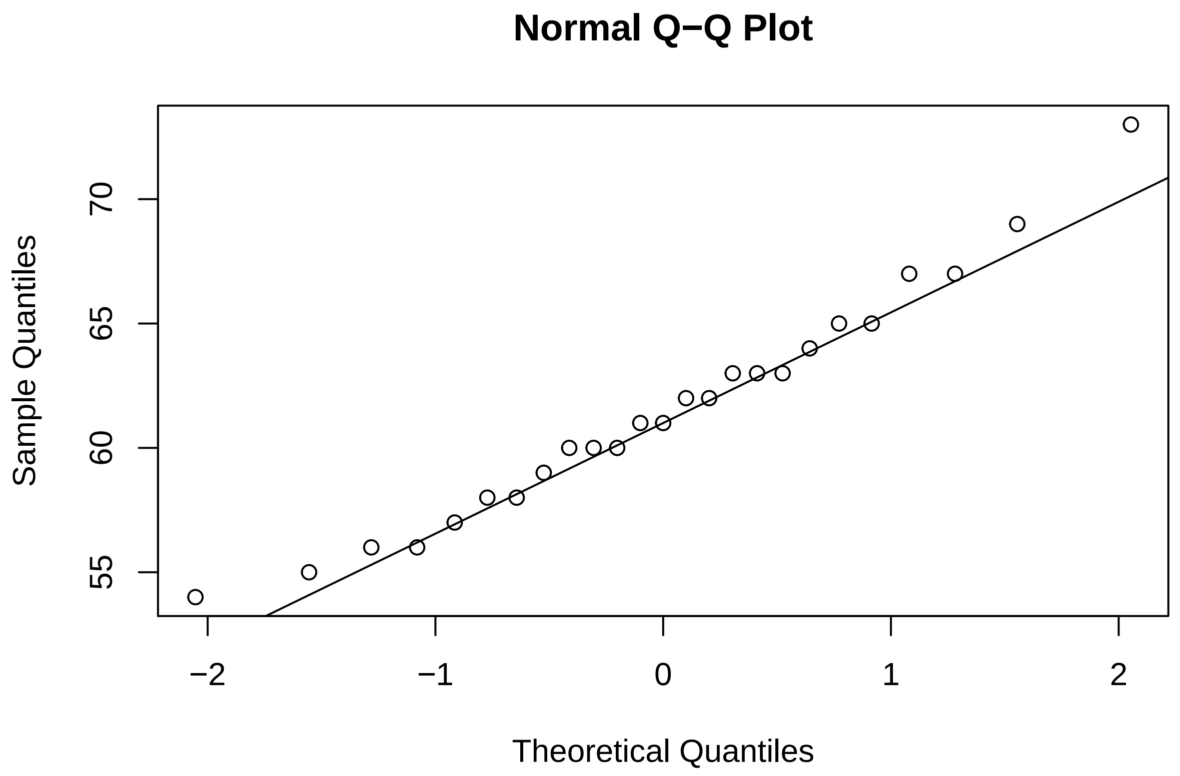
```
## [1] 0.9544997
```

```
(heights.area.3sd<-pnorm(heights.plus.3sd,heights.mean,heights.sd) - pnorm(heights.minus.3sd,heights.mea
```

```
## [1] 0.9973002
```

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

```
qqnorm(heights)
qqline(heights)
```



```r
#qqplot(heights)
```

Yes, this data appears to follow a normal distribution very closely, because the area under the curve corresponds very closely to 68-95-99 for 1sd, 2sd and 3sd away from the mean.

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?

```r
prob.defect<-0.02
prob.no.defect<-0.98
(p.10isdefect<-(0.98^9)*0.02)
```

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

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

```r
(p.100.nodefect<-0.98^100)
```

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

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

```r
(exp.val.transistors<-1/prob.defect)
```

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

```r
(sd.transistors<-sqrt((1-prob.defect)/prob.defect^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?

```r
prob.defect.2<-0.05
prob.no.defect.2<-0.95
(exp.val.transistors.2<-1/prob.defect.2)
```

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

```r
(sd.transistors.2<-sqrt((1-prob.defect.2)/prob.defect.2^2))
```

```
## [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?

The mean and standard deviation of the waiting time till first success, both decrease when the probability of success is increased.

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.

```r
p.boy<-0.51
(p.2boys<-choose(3,2)*p.boy^2*(1-p.boy))
```

```
## [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.

BBG BGB GBB

```
(p.BBG<-0.51*0.51*0.49)
```

```
## [1] 0.127449
```

```
(p.BGB<-0.51*0.49*0.51)
```

```
## [1] 0.127449
```

```
(p.GBB<-0.49*0.51*0.51)
```

```
## [1] 0.127449
```

```
(prob.2boys=p.BBG+p.BGB+p.GBB)
```

```
## [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).

Enumerating each combination where 3 out of 8 kids are boys is error prone and tedious because there are 8!/(5!)(3!) i.e. 56 possible combinations. That's why the approach in (a) is preferable.

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?

```
p.goodserve<-0.15
p.badserve<-0.85
(p.10.3<-choose(9,2)*(p.goodserve^3)*(p.badserve^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?

It would be 0.15 since the probability of a good serve is independent.

(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) involves just the 10th serve alone, while part (a) involves a particular pattern of 10 serves in which 2 out of previous 9 are good serves.