Lesson 06: Continuous Distributions

References

- Black, Chapter 6 Continuous Distributions (pp. 184-255)
- Davies, Chapter 16 Common Probability Distributions (pp. 342-362)
- Stowell, Chapter 7 Probability Distributions (pp. 87-97)
- Library Reserves: Teetor, Chapter 8 Probability (pp. 188-190)

Exercises:

1) Assume the purchases of shoppers in a store have been studied for a period of time and it is determined the daily purchases by individual shoppers are normally distributed with a mean of $81.14 and a standard deviation of $20.71. Find the following probabilities using R.

   a) What is the probability that a randomly chosen shopper spends less than $75.00

```
sprintf("%.4f", pnorm(75, mean = 81.14, sd = 20.71, lower.tail = TRUE))
```

```
## [1] "0.3834"
```

   b) What proportion of shoppers spends more than $100.00?

```
sprintf("%.4f", pnorm(100, mean = 81.14, sd = 20.71, lower.tail = FALSE))
```

```
## [1] "0.1812"
```

   c) What proportion of shoppers spends between $50.00 and $100.00?

```
prob_greater_than_50 <- pnorm(50, mean = 81.14, sd = 20.71, lower.tail = FALSE)
prob_greater_than_100 <-pnorm(100, mean = 81.14, sd = 20.71, lower.tail = FALSE)
sprintf("%.4f", prob_greater_than_50 - prob_greater_than_100)
```

```
## [1] "0.7524"
```

2) Assume that the shopper's purchases are normally distributed with a mean of $97.11 and a standard deviation of $39.46. Find the following scores using R.

   a) What weight is the 90th Percentile of the shoppers' purchases? That is, find the score P90 that separates the bottom 90% of shoppers' purchases from the top 10%.

```
sprintf("%.4f", qnorm(0.90, mean = 97.11, sd = 39.46, lower.tail = TRUE))
```

```
## [1] "147.6800"
```

   b) What is the median shoppers' purchase? (Find the score P50 that separates the bottom 50% of shoppers' purchases from the top 50%.) What is important about this number?

```r
sprintf("%.4f", qnorm(0.50, mean = 97.11, sd = 39.46, lower.tail = TRUE))
```

```
## [1] "97.1100"
```

```r
# What is important about this number?
# The normal distribution is symmetric, so its mean and median are identical.
```

3) Generate a sample of size 50 from a normal distribution with a mean of 100 and a standard deviation of 4. What is the sample mean and sample standard deviation? Calculate the standard error of the mean for this sample. Generate a second sample of size 50 from the same normal population. What is the sample mean and sample standard deviation? Calculate the standard error of the mean for this sample. Compare your results. Are the sample means and sample standard deviations of random samples of the same size taken from the same population identical? Why or why not?

   Now, repeat this process generating a sample of size 5000. Calculate the sample mean, sample standard deviation and standard error of the mean for this third sample and compare to the previous samples.

```r
set.seed(1234)  # seed the random number generator for reproducibility
my_first_sample <- rnorm(n = 50, mean = 100, sd = 4)
std_error1 <- 4/sqrt(50)
cat("\nmy_first_sample mean: ", mean(my_first_sample),
    " sample_std_dev: ", sd(my_first_sample), " std_error: ", std_error1, "\n")
```

```
##
## my_first_sample mean:  98.18779  sample_std_dev:  3.540174  std_error:  0.5656854
```

```r
my_second_sample <- rnorm(n = 50, mean = 100, sd = 4)
std_error2 <- 4/sqrt(50)
cat("\nmy_second_sample mean: ", mean(my_second_sample),
    " sample_std_dev: ", sd(my_second_sample), " std_error: ", std_error2, "\n")
```

```
##
## my_second_sample mean:  100.5581  sample_std_dev:  4.148804  std_error:  0.5656854
```

```r
my_third_sample <- rnorm(n = 5000, mean = 100, sd = 4)
std_error3 <- 4/sqrt(5000)
cat("\nmy_third_sample mean: ", mean(my_third_sample),
    " sample_std_dev: ", sd(my_third_sample), " std_error: ", std_error3, "\n")
```

```
##
## my_third_sample mean:  99.99199  sample_std_dev:  3.966926  std_error:  0.05656854
```

4) Assume a biased coin when flipped will generate heads one third of the time. Estimate the probability of getting at least 250 heads out of 600 flips using the normal distribution approximation. Compare to the exact probability using the binomial distribution.

```r
# Normal approximation to the binomial uses  z = (x - n*p)/sqrt(n * p * (1-p))
n <- 600
p <- 1/3
x <- 250 - 0.5  # least 250 heads implies the upper tail of the standard normal distribution
z <- (x - n*p)/sqrt(n * p * (1-p))

# R provides binomial probabilties directly
# dbinom(x, size, prob, log = FALSE)  # density function
# pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)  # distribution function
# qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)  # quantile function
# rbinom(n, size, prob)  # here n is the number of random variates to generate
sprintf("%.6f", pbinom(q = x, size = n, prob = p, lower.tail = FALSE))
```

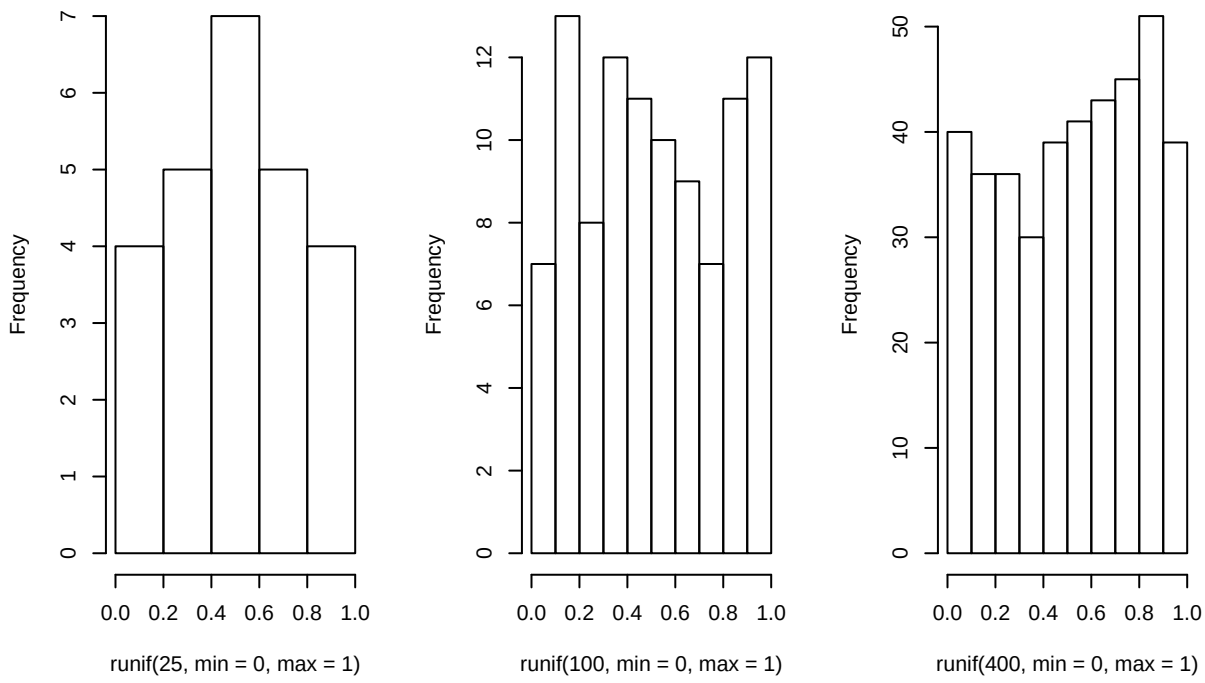
```
## [1] "0.000012"
```

```r
# The normal approximation to the binomial is very close to the binomial.
sprintf("%.6f", pnorm(z, mean = 0, sd = 1, lower.tail = FALSE))
```

```
## [1] "0.000009"
```

5) Use the uniform distribution over 0 to 1. Generate three separate simple random samples of size n = 25, n = 100, n = 400. Plot histograms for each and comment on what you observe.

```r
par(mfrow=c(1,3), oma=c(0,0,2,0))
hist(runif(25, min = 0, max = 1), main = "")
hist(runif(100, min = 0, max = 1), main = "")
hist(runif(400, min = 0, max = 1), main = "")
mtext("Histograms of uniform distribution (n = 25, 100 and 400)", side = 3,
    outer = T, line = -1)
```



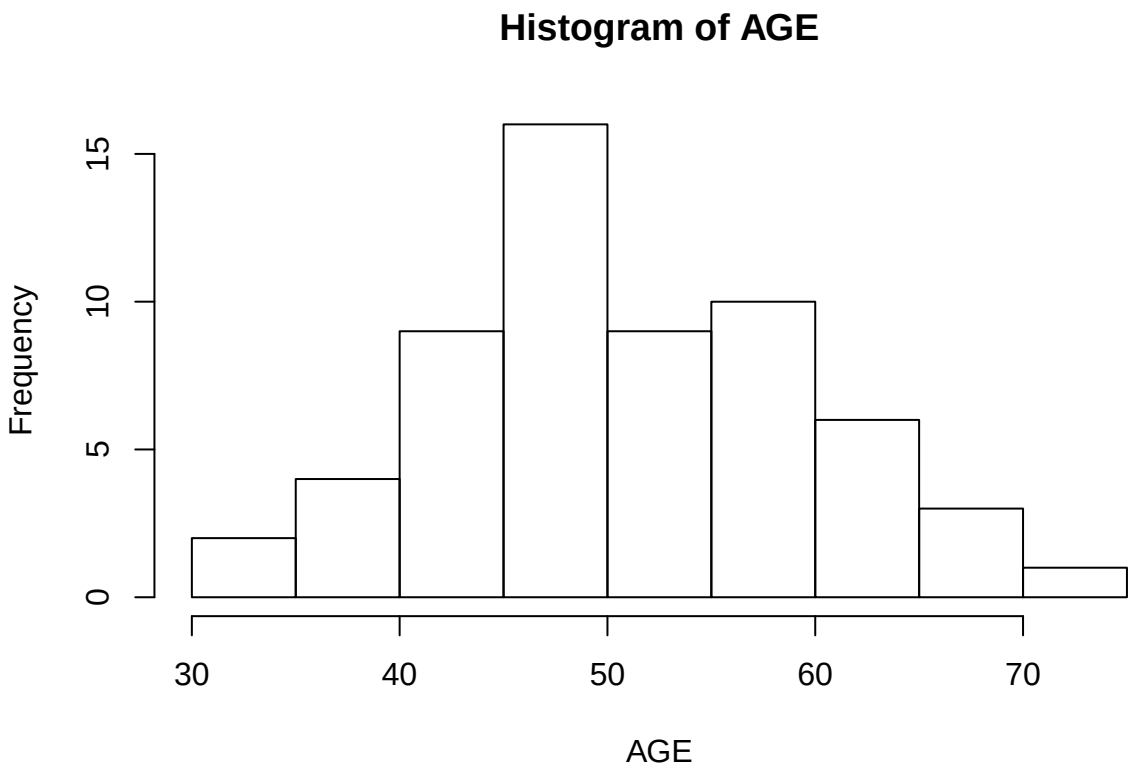

```
par(mfrow=c(1,1))
```

6) salaries.csv gives the CEO age and salary for 60 small business firms. Construct QQ plots and histograms. Is the distribution of ages a normal distribution? Explain your answer.

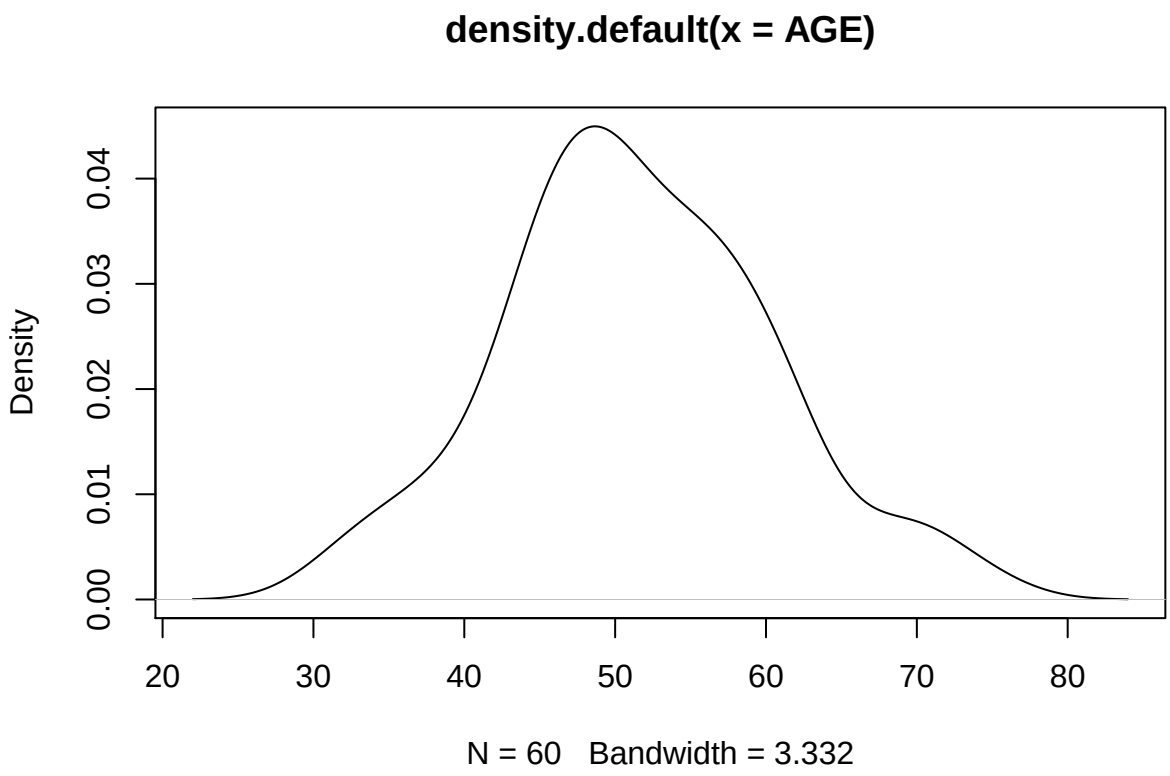
```
# Read the comma-delimited text file creating a data frame object in R,
# then examine its structure:

salaries <- read.csv("salaries.csv")
str(salaries)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ AGE: int  53 43 33 45 46 55 41 55 36 45 ...
##  $ SAL: int  145 621 262 208 362 424 339 736 291 58 ...
```

```
with(salaries, hist(AGE))
```




```
with(salaries, plot(density(AGE)))
```




```
# R provides qq plotting capabilities... see qqnorm documentation
# A straight line look to the plot suggests that the distribution
# is similar in form to a normal distribution.
qqnorm(salaries$AGE, main = "AGE QQ",xlab = "Normal Quantiles",ylab = "Age Quantiles")
qqline(salaries$AGE, distribution = qnorm, probs = c(0.25, 0.75), qtype = 7)
```


```
# Each of these graphs implies that AGE may be thought of as normally distributed.
```