NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

R Lesson 6 - Solutions
MSPA 401 – Introduction to Statistical Analysis

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 $50.71. Find the following probabilities using R.

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

```
sprintf("%.4f", pnorm(75, mean = 81.14, sd = 20.71, lower.tail = TRUE))       result = 0.3834
```

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

```
sprintf("%.4f", pnorm(100, mean = 81.14, sd = 20.71, lower.tail = FALSE))    result = 0.1812
```

   c) What proportion of shoppers spend between $50.00 and $80.00?

```
prob_greater_than_50 <- pnorm(50, mean = 81.14, sd = 20.71, lower.tail = FALSE)
prob_greater_than_80 <-pnorm(80, mean = 81.14, sd = 20.71, lower.tail = FALSE)
sprintf("%.4f", prob_greater_than_50 - prob_greater_than_80)                  result = 0.4117
```

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))  result= 147.6800
```

   b) What is the median shoppers' purchase? That is, find the score P50 that separates the bottom 50% of shoppers' purchases from the top 50%. What is significant about this number?

```
sprintf("%.4f", qnorm(0.50, mean = 97.11, sd = 39.46, lower.tail = TRUE))  result = 97.1100
```

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 mean and standard error of the mean for the sample? Generate a second sample of size 50 from the same normal population. What is the mean and standard error of the mean for this second sample? Now repeat this process generating a sample of size 5000. Calculate the mean and standard error of the mean for this third sample and compare to the previous samples. What do you observe?

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

```
my_first_sample mean:  98.18779  std_error: 0.5006562
 my_second_sample <- rnorm(n = 50, mean = 100, sd = 4)

std_error2 <- sd(my_second_sample)/sqrt(50)
cat("\nmy_second_sample mean: ", mean(my_second_sample)," std_error:", std_error2)
my_second_sample mean:  100.5581  std_error: 0.5867295
my_third_sample <- rnorm(n = 5000, mean = 100, sd = 4)
std_error3 <- sd(my_third_sample)/sqrt(5000)
cat("\nmy_third_sample mean: ", mean(my_third_sample), " std_error:", std_error3)
my_third_sample mean:  99.99199  std_error: 0.0561008
```

The sample mean for the third sample is virtually identical to the true population mean and the standard error of the mean for the third sample is approximately one tenth that of the first two samples.

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. Then calculate the exact probability using the binomial distribution. Compare the two probabilities.

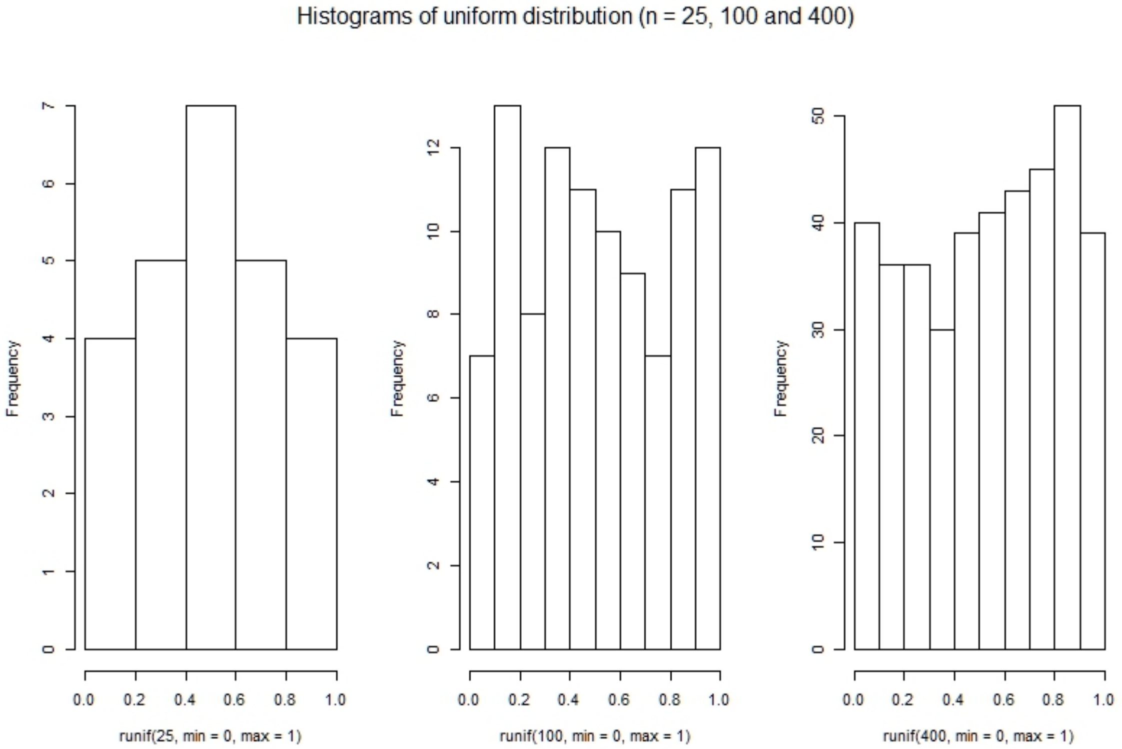
250 heads out of 600 tosses implies a binomial with n = 600 and x = 250
the biased coin has probability of heads p = 1/3
normal approximation to the binomial uses z = (x - n*p)/sqrt(n * p * (1-p))

```
sprintf("%.6f", pnorm(z, mean = 0, sd = 1, lower.tail = FALSE))              result = 0.000007
```

R provides binomial probabilties directly

```
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)  # distribution function
sprintf("%.6f", pbinom(q = x, size = n, prob = p, lower.tail = FALSE))       result = 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.

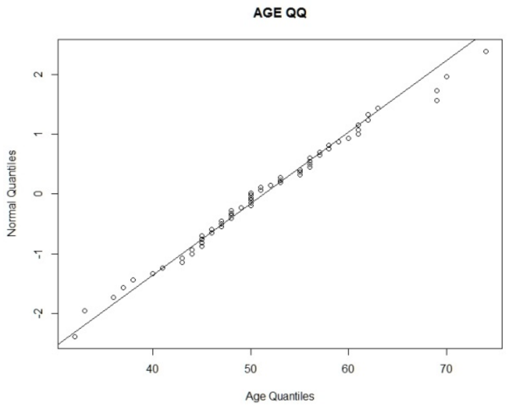


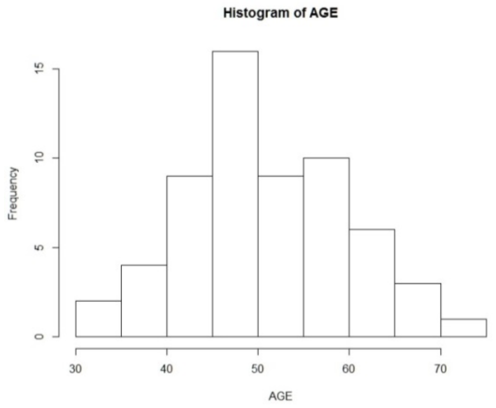
As the sample size increases, the sampling distribution becomes more uniform in shape.

6) **salaries.csv** gives the CEO age and salary for 60 small business firms. Use s**alaries.csv** and QQ plots, histograms and the Shapiro-Wilks test to answer the following questions:

a) Is the distribution of ages a normal distribution? Explain your answer.

Examination of the QQplot, histogram an dShapiro-Wilks test do not reveal extreme departures from normality.





Shapiro-Wilk normality test
data: salaries$AGE
W = 0.9873, p-value = 0.7855 n.s.