Precept 8 Exercise Answers

Your Name

Fall 2019

Objectives

- dnorm(x, mean = 0, sd = 1, ...), pnorm(q, mean = 0, sd = 1, ...), qnorm(p, mean = 0, sd = 1, ...), rnorm(n, mean = 0, sd = 1,);
- Constructing confidence intervals for different scenarios;
- Learn how to checking Normality for a sample;
- CLT v.s. LLN

Load `ggplot2` package

```
library(ggplot2)
```

Demo

Norm distribution

The following functions are related to X ~ Normal(mean, sd) random variables:

- dnorm(x, mean = 0, sd = 1, ...) – Density at X = x: P(X = x)
- pnorm(q, mean = 0, sd = 1, ...) – P(X <= q)

- qnorm(p, mean = 0, sd = 1, ...) – Given p, find the quantile q such that the P(X <= q) = p
- rnorm(n, mean = 0, sd = 1,) – Generate a sample of size n, drawn from Normal(m = size, p = prob)

Example 1 Using a Normal distribution to approximate the distribution of a discrete random variable

The average verbal SAT score (i.e., for the Evidence-Based Reading and Writing part) for graduating seniors in 2018 was 536 and the SD was 102 (https://reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf).

(SAT benchmark scores are set by the College Board to indicate likelihood of success in college. The verbal SAT benchmark is associated with a 75% chance of earning at least a C in first-semester, credit-bearing, college-level courses in history, literature, social science, or writing.)

You may assume that the histogram for the verbal SAT scores in 2018 follows the Normal curve closely. Answer the following questions about the students who took the SAT in 2018.

(a) The benchmark score for the verbal SAT in 2018 is 480. What percentage of the students miss the benchmark in 2018?

```
pnorm(480, mean = 536, sd = 102)
[1] 0.291496
# Note that if X ~ Normal(mean = 536, sd = 102), P(X = 480) = 0.
```

Aside: Note that the SAT scores have discrete values but the Normal distribution is a continuous distribution. For example if $X \sim Normal(mean = 536, sd = 102)$ the chance that $P(X = 480) = 0$ but $P(score = 480)$ could be non-zero for the SAT score dataset. In order to make a better estimate on $P(score = 480)$ we can make a continuity correction; this means that we will calculate $P(479.5 \leq X \leq 480.5)$ for $P(score = 480)$. If we made this continuity correction our answer would change to

```
pnorm(479.5, mean = 536, sd = 102)
[1] 0.2898163
```

However, you can see that the difference between the two answers are very small: .17%

```
pnorm(480, mean = 536, sd = 102) - pnorm(479.5, mean = 536, sd = 102)
[1] 0.001679735
```

This is because the SAT verbal score ranges from 200 to 800 in 2018; i.e., there are many bins for the histogram. Thus, the area under the Normal curve that correspond to a particular bin of width 1 is very small. Therefore, we will not do the continuity correction here.

(b) What percentage of the students score above the average verbal SAT score?

50% because the Normal distribution is symmetric around the mean.

Alternatively, you can do the calculation:

```
1 - pnorm(536, mean = 536, sd = 102)
[1] 0.5
```

(b) What percentage of the students receive verbal SAT scores that are between 530 and 670?

```
pnorm(670, mean = 536, sd = 102) - pnorm(530, mean = 536, sd = 102)
[1] 0.4289844
```

(b) What percentage of the students receive verbal SAT scores that are less or equal to 400 or greater or equal to 700?

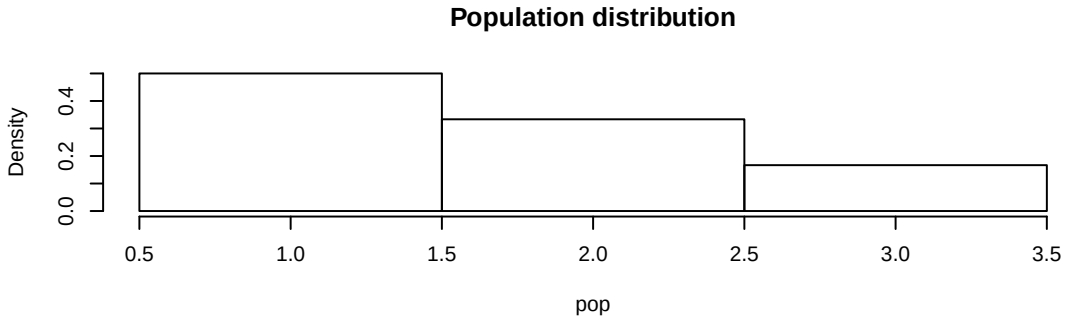
```
pnorm(400, mean = 536, sd = 102) + (1 - pnorm(700, mean = 536, sd = 102))
[1] 0.145146
```

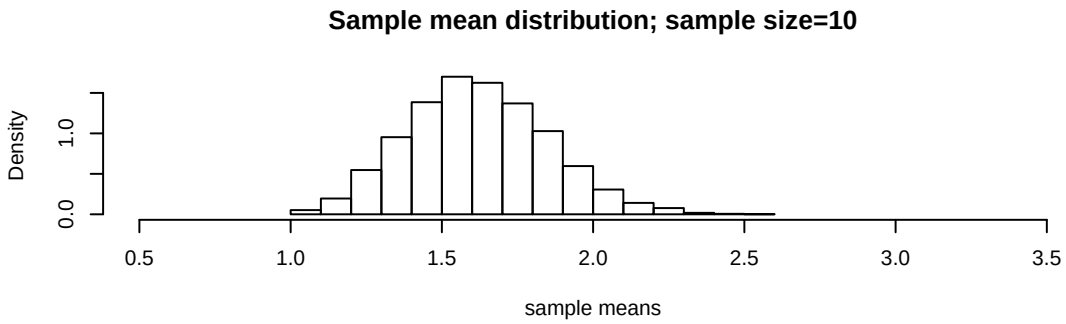
The Central Limit Theorem

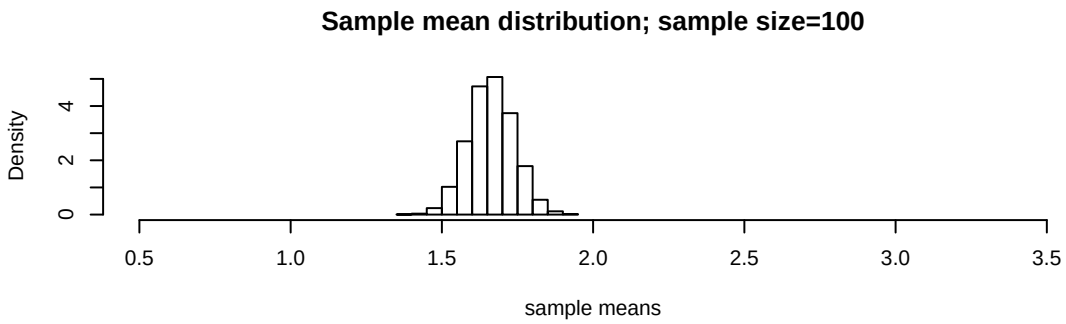
CLT v.s. LLN

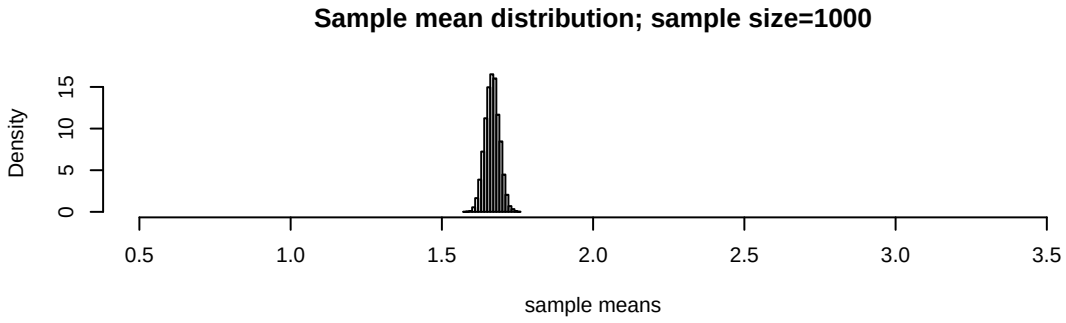
Preceptor: please explain the difference between CLT and LLN. (You do not need to explain the code used to produce the graphs.)

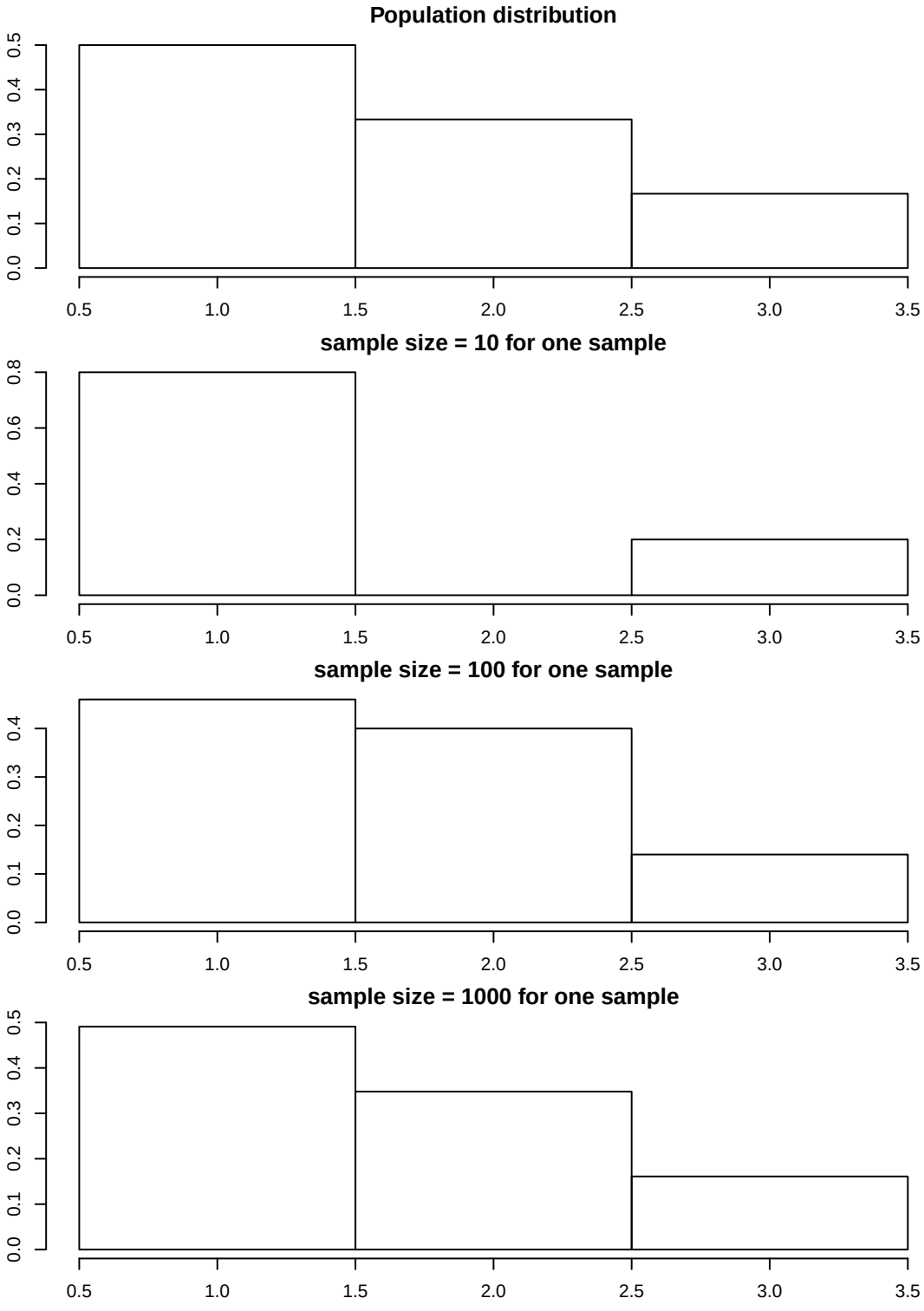
```r
# Data in population
pop = c(1, 1, 1, 2, 2, 3)
```


Population distribution
Density
0.0
0.2
0.4
0.5
1.0
1.5
2.0
2.5
3.0
3.5
pop


Sample mean distribution; sample size=10
Density
0.0
1.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
sample means


Sample mean distribution; sample size=100
Density
0
2
4
0.5
1.0
1.5
2.0
2.5
3.0
3.5
sample means


Sample mean distribution; sample size=1000
Density
0
5
10
15
0.5
1.0
1.5
2.0
2.5
3.0
3.5
sample means


Population distribution
0.0
0.1
0.2
0.3
0.4
0.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5
sample size = 10 for one sample
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
2.5
3.0
3.5
sample size = 100 for one sample
0.0
0.1
0.2
0.3
0.4
0.5
1.0
1.5
2.0
2.5
3.0
3.5
sample size = 1000 for one sample
0.0
0.1
0.2
0.3
0.4
0.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Constructing confidence Intervals

Example 2

```
# LIZARD LENGTH DATA in cm
lizard = c(6.2, 6.6, 7.1, 7.4, 7.6, 7.9, 8, 8.3, 8.4, 8.5, 8.6,
    8.8, 8.8, 9.1, 9.2, 9.4, 9.4, 9.7, 9.9, 10.2, 10.4, 10.8, 11.3,
    11.9)
```

We would like to construct a 90% confidence interval to predict the average length (in cm) for the lizards population where the sample is drawn from. How do we construct the CI? Recall from Ch4.5 that there are different cases:

- **Case 1: When the sample size n is large:** The standardized sample means form a Normal(0, 1) approximately. If the population SD σ is known we can calculate the $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$; if the population SD σ is unknown, because the sample size is large, the sample SD will be a good estimate of the population SD, σ. **We use the Normal(0, 1) distribution to find out the value for $q_{(1-\alpha+\frac{\alpha}{2})}$.**

$$\bar{x} \pm \ q_{(1-\alpha+\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} \pm \ q_{(1-\alpha+\frac{\alpha}{2})} \frac{s}{\sqrt{n}}$$

- **Case 2: When the sample size is small, σ is unknown and the *population is normal*:**

If the population SD σ is unknown to us, with the condition that the *population is normal* (or at least close to being symmetric) the standardized sample means form a t-distribution with degree of freedom = (sample size -1) = n-1. **In this case we will use the t-distribution with degree of freedom n-1 to find out the value for $q_{(1-\alpha+\frac{\alpha}{2})}$.**

$$\bar{x} \pm \ q_{(1-\alpha+\frac{\alpha}{2})} \frac{s}{\sqrt{n}}$$

- **Case 3: When the sample size is small and the population is *not* normal:** In this case we do not know the distribution of the sample means. We will use *Bootstrap sampling* to get an approximation of the distribution of the sample means and use this approximate distribution to construct a confidence interval to estimate the population mean.

Since our sample size is a bit small, we are in either case 2 or 3. Let's check to see if it would be reasonable to assume that the population is roughly Normal.

Checking the Normality assumption for the data

Will it be reasonable to assume that the population data follow a normal distribution?

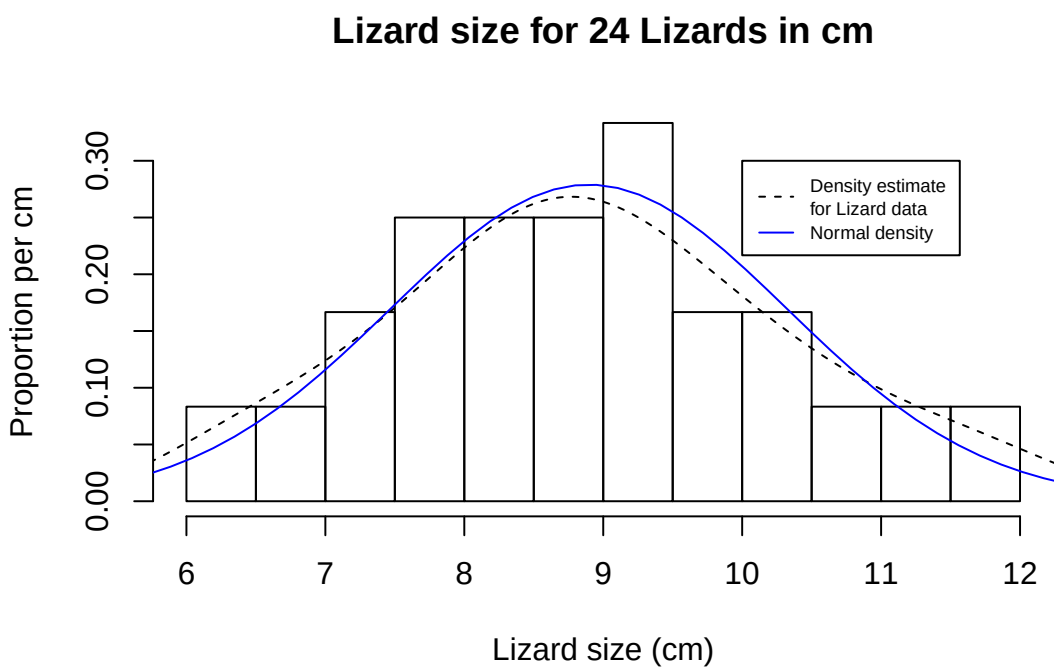
Let's look at the histogram of the data and compare it with the Normal distribution with mean and sd as the sample's.

```
# Look at the histogram
hist(lizard, breaks = 20, freq = F, xlab = "Lizard size (cm)", main = "Lizard size for 24 Lizards
    ylab = "Proportion per cm")

lines(density(lizard), lty = 2)  # This adds a line estimating
# the distribution of Lizard data; lty = 2 give a dashed line

xseq = seq(from = min(lizard) * 0.9, to = max(lizard) * 1.1, length = 50)
lines(x = xseq, y = dnorm(x = xseq, mean = mean(lizard), sd = sd(lizard)),
    col = "blue")
# This adds a Normal density with mean and SD with the same
# values as the mean and SD of the Lizard dataset

legend(x = 10, y = 0.3, legend = c("Density estimate \nfor Lizard data",
    "Normal density"), lty = c(2, 1), col = c("black", "blue"),
    cex = 0.6)
```

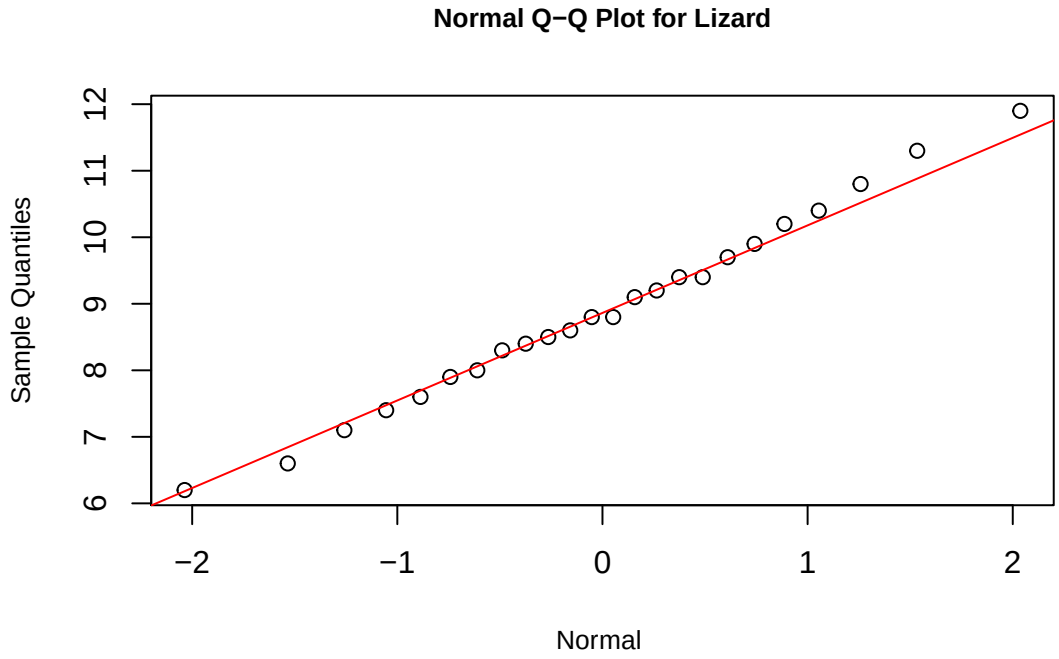


Let's look at the qq-plots for the Lizard data too.

We use qq-plot to compare the sample quantiles with the quantiles from N(0, 1) and check the assumption that the population is approximately normally distributed.

```r
# Compare to the quantiles of Normal(0,1) distribution

qqnorm(y = lizard, main = "Normal Q-Q Plot for Lizard", xlab = "Normal ",
    ylab = "Sample Quantiles", cex.lab = 0.8, cex.main = 0.8)
qqline(y = lizard, col = "red")
```



There is no sign indicating that the population data are not Normal.

Constructing a CI for case 2 with unknown population SD

Since we do not know the population SD, the sample size is a bit small (24), and since it seems to be reasonable to assume that the population is Normal, we can use the t-distribution to find q.

```r
# Set the degree of confidence for the interval
p = 0.9

my.q = qt(p + (1 - p)/2, df = length(lizard) - 1)  # do not use `q`
# as your variable name; q() is a function to terminate your R
# session; see `?q` or `?quit`
lizard.se = sd(lizard)/sqrt(length(lizard))

# a p*100-percent confidence Interval for the average length of
# the lizards is
mean(lizard) + c(-1, 1) * my.q * lizard.se
[1] 8.395575 9.396092
```

```
# Can also do everything in one line
mean(lizard) + c(-1, 1) * qt(p + (1 - p)/2, df = length(lizard) -
    1) * sd(lizard)/sqrt(length(lizard))
[1] 8.395575 9.396092
```

Exercises

Please fill in any missing axis labels and graph titles yourself.

Question 1

For the population in the Central Limit Theorem section the population mean and SD are

```
mean(pop)
[1] 1.666667
sd(pop)
[1] 0.8164966
```

Let's draw a sample of size 60 from the population.

```
set.seed(4321)
sampl.60 = sample(pop, size = 60, replace = T)
```

Let's assume that you do not know the population mean or the population SD. You can calcualte the sample mean and sample SD.

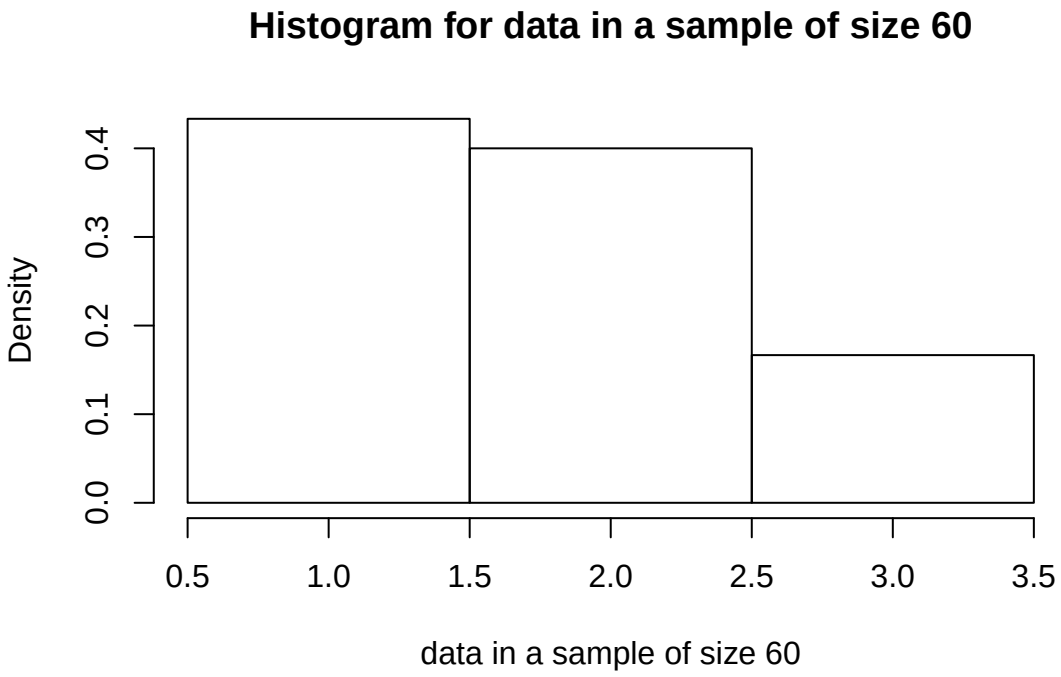
```
mean(sampl.60)
[1] 1.733333
sd(sampl.60)
[1] 0.7333847
```

What distribution does the sample mean follow in this case?

Part a

Here is the histogram of the sample data. Fill in the blanks for the following questions.

```
hist(sampl.60, main = "Histogram for data in a sample of size 60",
    freq = F, xlab = "data in a sample of size 60", breaks = c(0.5,
        1.5, 2.5, 3.5))
```



The distribution for the data points in the sample is skewed, just like population distribution; this is a result of the _____.

Even though the sample distribution is skewed, the distribution of the sample mean will still be Normally distributed; this is a result of the _____.

The SD of the sample mean distribution is about _____.

Part b

Construct a 80% CI to predict the population mean with your sample data.

Part c (no action required)

Instead of 60 now draw a sample of size 6000 from the population and again let's assume that you do not know the population mean or the population SD.

```
set.seed(4321)
sampl.6000 = sample(pop, size = 6000, replace = T)
mean(sampl.6000)
[1] 1.6635
sd(sampl.6000)
[1] 0.7416374
```

Part d

Does the magnitude of the sample SD (sample SD refers to the SD of the data in the sample) change when the sample size is increased from 60 to 6000? Does the magnitude of the estimated SD(sample mean) change when the sample size is increased from 60 to 6000?

Part e

Construct a 80% CI to predict the population mean.

Part f

For the length of the CI for the sample of size 60 is about _____ times of that for the sample of size 6000.

Question 2

(6.10 in OpenIntro) **Life rating in Greece**. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered suffering.

Part a

Describe the population parameter of interest. What is the value of the point estimate of this parameter?

Part b

Which case (see the beginning of the demo section) are we in? Check if the conditions required for constructing a confidence interval based on these data are met.

Part c

Construct a 95% confidence interval for the proportion of Greeks who are suffering. Note that if you have a sample of 0-1 numbers, the SD for the sample data can be obtained with this formula:

$Sample\_SD = \sqrt{p(1-p)}\sqrt{n/(n-1)}$

where p is the proportion of 1's in the sample and n is the sample size. For example, for a sample of size 100 that has 40 1's and 60 0's the sample SD is

```
sqrt(0.4 * (1 - 0.4)) * sqrt(100/(100 - 1))
[1] 0.492366
```

We can verify this result by recreating the sample with 40 1's and 60 0's and calculate its SD:

```
test.sampl = rep(0:1, times = c(60, 40))

sd(test.sampl)
[1] 0.492366
```

Part d

Without doing any calculations and with the same sample data, describe what would happen to the confidence interval if we decided to use a higher confidence level.

Part e

Without doing any calculations, describe what would happen to the confidence interval if we used a larger sample.

Question 3

This dataset *Newcomb.RData* includes data on the amount of time (in seconds) taken for a beam of light to travel from the Naval Observatory in Washington D.C. to the Washington Monument and back, a distance of 7.44373km. The data was collected in 1880 by Simon Newcomb (working with Ablert Michelson).

```
Newcomb = data.frame(Time = c(2.4828e-05, 2.4826e-05, 2.4833e-05,
    2.4824e-05, 2.4834e-05, 2.4756e-05, 2.4827e-05, 2.4816e-05,
    2.484e-05, 2.4798e-05, 2.4829e-05, 2.4822e-05, 2.4824e-05, 2.4821e-05,
    2.4825e-05, 2.483e-05, 2.4823e-05, 2.4829e-05, 2.4831e-05, 2.4819e-05,
    2.4824e-05, 2.482e-05, 2.4836e-05, 2.4832e-05, 2.4836e-05, 2.4828e-05,
    2.4825e-05, 2.4821e-05, 2.4828e-05, 2.4829e-05, 2.4837e-05,
    2.4825e-05, 2.4828e-05, 2.4826e-05, 2.483e-05, 2.4832e-05, 2.4836e-05,
    2.4826e-05, 2.483e-05, 2.4822e-05, 2.4836e-05, 2.4823e-05, 2.4827e-05,
    2.4827e-05, 2.4828e-05, 2.4827e-05, 2.4831e-05, 2.4827e-05,
    2.4826e-05, 2.4833e-05, 2.4826e-05, 2.4832e-05, 2.4832e-05,
    2.4824e-05, 2.4839e-05, 2.4828e-05, 2.4824e-05, 2.4825e-05,
    2.4832e-05, 2.4825e-05, 2.4829e-05, 2.4827e-05, 2.4828e-05,
    2.4829e-05, 2.4816e-05, 2.4823e-05), Series = as.factor(c(1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,
    2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3,
    3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
    3, 3, 3, 3, 3)))
str(Newcomb)
'data.frame':   66 obs. of  2 variables:
 $ Time  : num  2.48e-05 2.48e-05 2.48e-05 2.48e-05 2.48e-05 ...
 $ Series: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
head(Newcomb)
        Time Series
1 2.4828e-05      1
2 2.4826e-05      1
3 2.4833e-05      1
4 2.4824e-05      1
5 2.4834e-05      1
6 2.4756e-05      1
```

Part a

Use ggplot() to generate a histogram for the time data. A density plot is a smoothed version of a histogram. Superimpose a density plot on top of the histogram by using the `geom_density()` function (you can just add the following layer to your code that produce the histogram).

```
# DENSITY PLOT

geom_density(color = "blue") + labs(x = "Time (in seconds)", y = "Proportion per second",
    title = "Distribution for the Amount of Time (in seconds) \nTaken for a Beam of Light to Trav
    subtitle = "from the Naval Observatory to the Washington Monument and Back")
```

Part b

We already saw on the histogram that the sample distribution has some outliers with small values. Use qq-plot to compare the quantile of the data with the quantile from N(0, 1) to further confirm that the distribution of time is not normally distributed if we include the outliners.

Part c

Determine the 90% confidence interval for the average time. To construct your CI in a correct way does the population data need to be Normal?

Part d (Optional)

Use a bootstrap sampling with 10,000 simulations to determine the 90% confidence intervals of the standard deviation of *Time*.

Part e (Optional)

Generate a density plot of the bootstrap sample SD's above.