# R Assignment 3

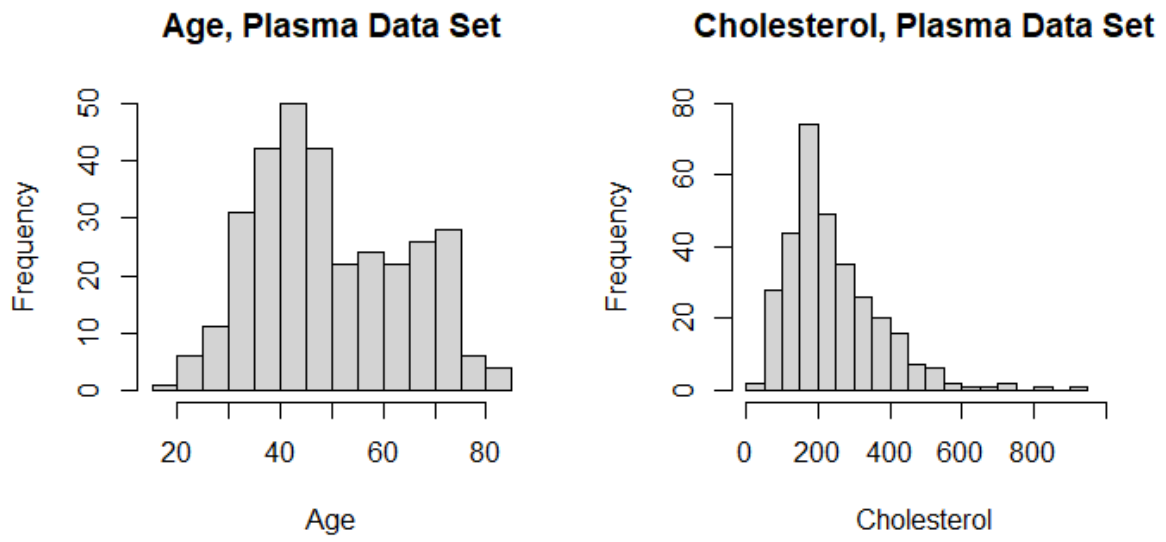
*Code segments provided for each question, full R script for the assignment at the bottom*

## EXPLORATORY DATA ANALYSIS FOR AGE AND CHOLESTEROL

> 1. Fill in the following table with summary statistics for the variables: *age* and *cholesterol*. In addition, create a single graph for the variables: *age* and *cholesterol*. Based on these results, describe the shape, center, and spread for the two variables.

| | Mean | Median | Q1 | Q3 | IQR | Range | Min | Max | SD |
|---|---|---|---|---|---|---|---|---|---|
| **Age** | 50.15 | 48.00 | 39.00 | 62.50 | 14.5 | 64 | 19.00 | 83.00 | 14.58 |
| **Cholesterol** | 242.5 | 206.3 | 155.0 | 308.9 | 153.9 | 863 | 37.7 | 900.7 | 131.99 |

```
# create single graph for variables
par(mfrow= c(1,2))    # set two plots in one graph
hist(plasma$age, main = "Age, Plasma Data Set", xlab = "Age")
hist(plasma$cholesterol,
     main = "Cholesterol, Plasma Data Set", xlab = "Cholesterol",
     breaks = 15,
     xlim = c(0, 1000), ylim = c(0, 80))
par(mfrow=c(1,1))
```

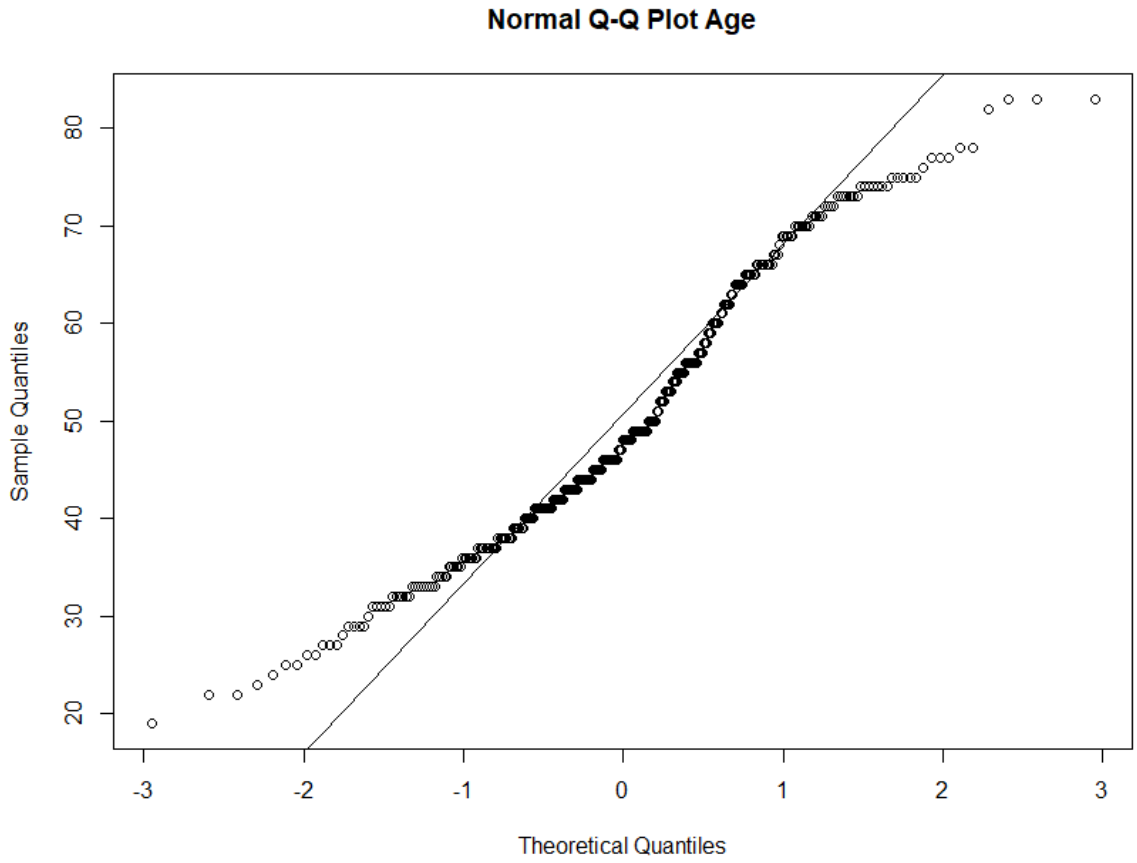


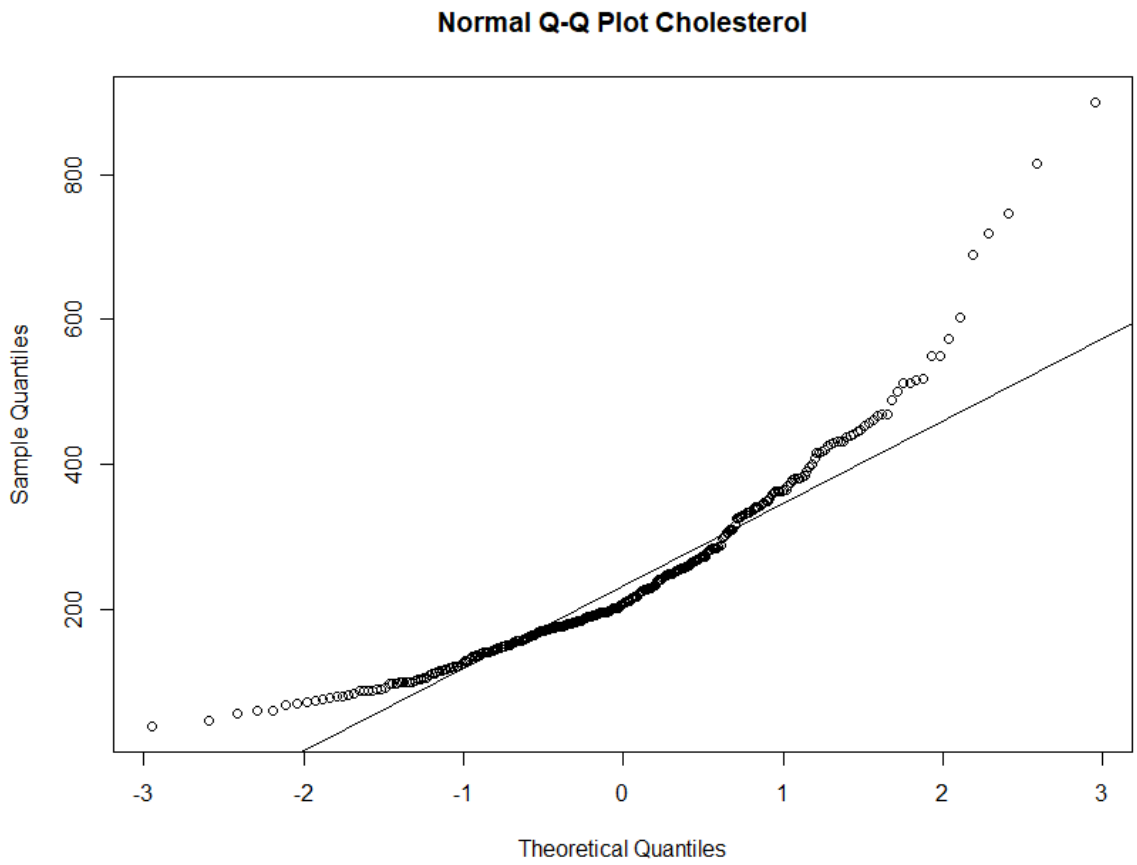

The shape of the distribution of the age variable in the plasma data set can be described as a bimodal distribution with one stronger peak on the left side. The center, described as the mean $\bar{x}$ is 50.15 with a range of 64 and a standard deviation s of 14.58, meaning that 68% of all participants of the study are between 35.57 and 64.73 years old.

The distribution of the cholesterol variable of the plasma data set is a right skewed, unimodal distribution. The center of the distribution $\bar{x}$ is 242.5 with a standard deviation of 131.99, so 68% of all participants of the study consumed between 110.51 and 374.49 mg cholesterol per day, while the maximum was 900.7 mg.

2. Create Q-Q plots for *age* and *cholesterol*. Which variable, *age* or *cholesterol*, appears to be **approximately** normally distributed? Since it can be hard to tell, I will accept any answer as long as you can support your reasoning!

```
# Q-Q plots for age and cholesterol

qqnorm(plasma$age, main = "Normal Q-Q Plot Age")
qqline(plasma$age)

qqnorm(plasma$cholesterol, main = "Normal Q-Q Plot Cholesterol")
qqline(plasma$cholesterol)
```

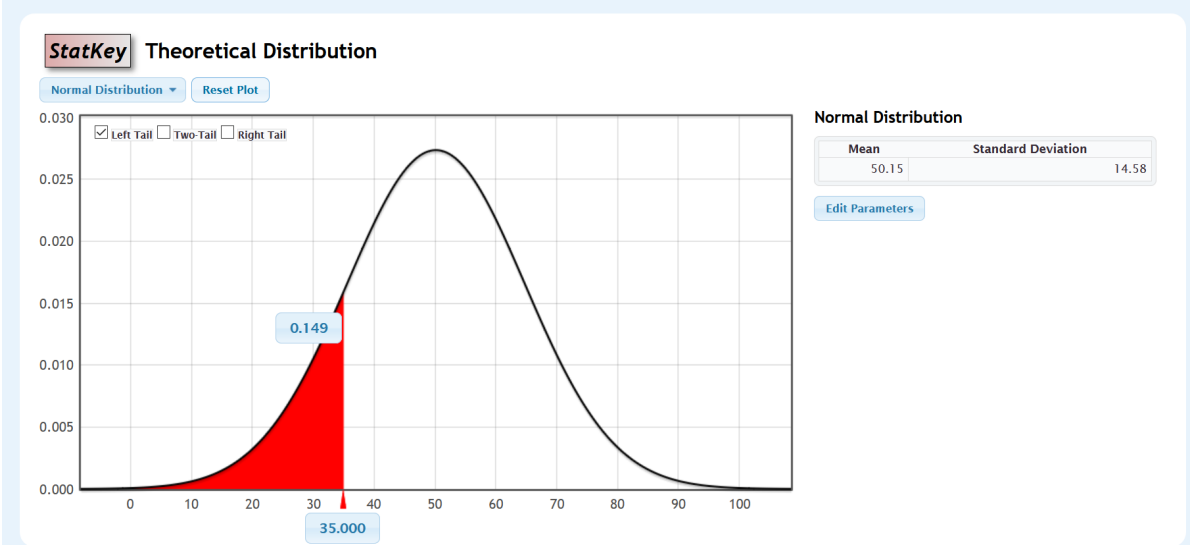





The straight line in each graph shows the perfect normal distribution. While the dots of the age variable seem to follow a stretched *S*-shape, it appears to be approximately normal distributed. Even if there is some deviation for the dots in the bottom left and top right corner, the overall distribution seems to follow approximately the normal distribution.

The distribution of cholesterol follows more of a curve than the normal distribution. One might argue that the distribution for cholesterol approximately follows the normal distribution for some parts, but overall, the distribution follows a curve while the normal distribution is a line.

> 3. Using the Normal Distribution applet, what is the probability a patient from the population of elective surgical procedures of non-cancerous organs is a Millennial (i.e., age ≤ 35)? Show your work by taking a screenshot of the applet image. Confirm your result in R by pasting your code and result here!



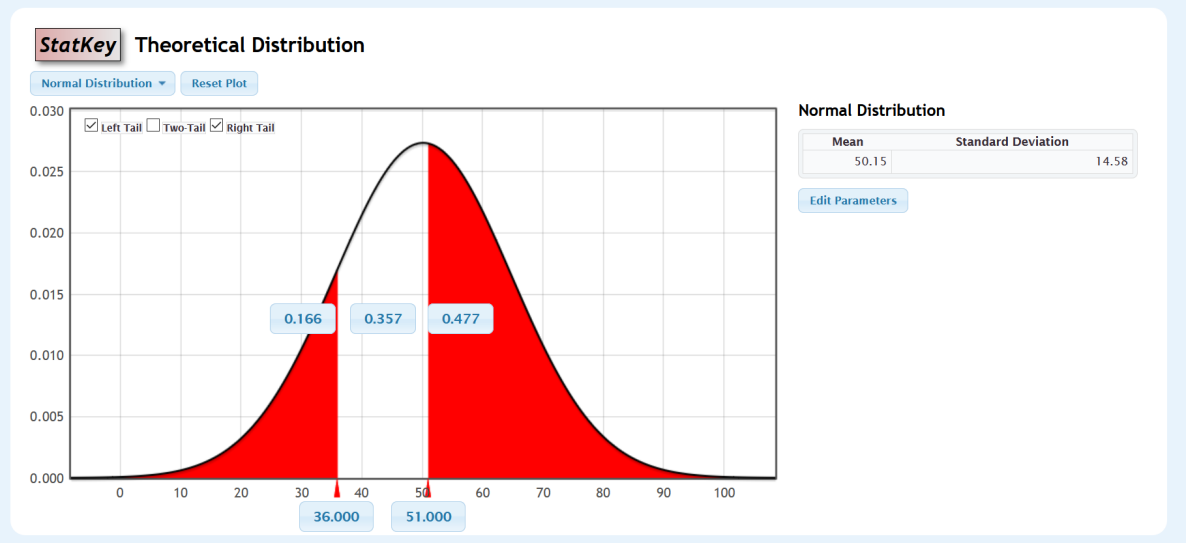
```
# probability patient being age <= 35
pnorm(35, mean = 50.15, sd = 14.58)

[1] 0.1493804
```

As seen in the Normal Distribution applet, the probability a patient from the population of elective surgical procedures of non-cancerous organs is a Millennial (i.e., age ≤ 35) is 14.9%.

This result is confirmed by the R code.

> 4. Using the Normal Distribution applet, what is the probability a patient from the population of elective surgical procedures of non-cancerous organs is a GenXer (i.e., 35 < age ≤ 51)? Show your work by taking a screenshot of the applet image. Confirm your result in R by pasting your code and result here!

As the range is exclusive for the lower limit and the age variable in this data set is a numerical discrete variable, for the following questions the next higher number will be used to calculate the probability which uses inclusive limits for the range.

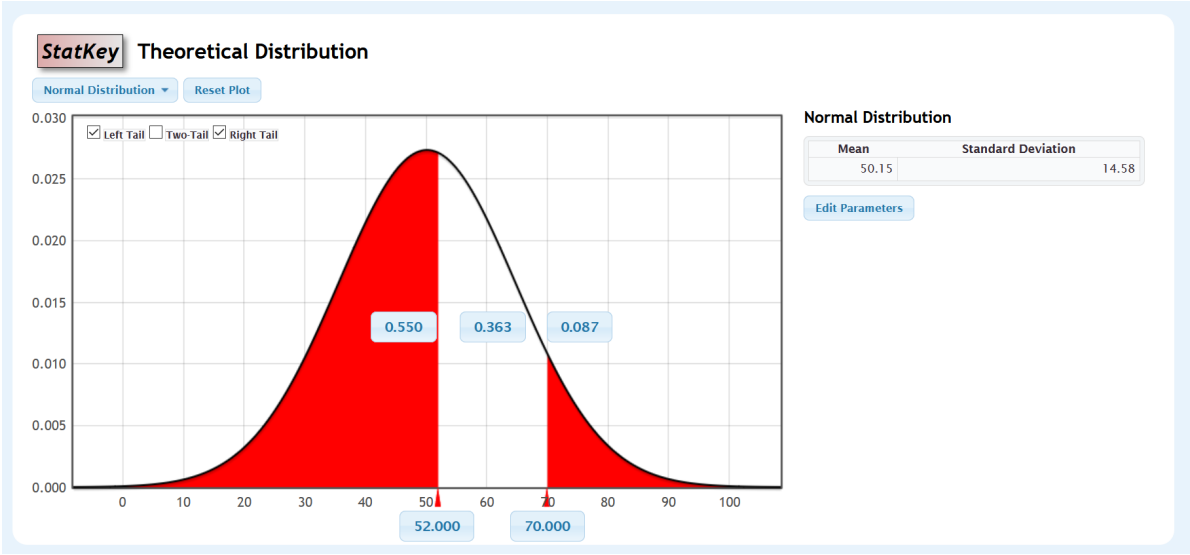



```
pnorm(51, mean = 50.15, sd = 14.58) - pnorm(36, mean = 50.15, sd = 14.58)

[1] 0.357348
```

As seen in the Normal Distribution applet, the probability a patient from the population of elective surgical procedures of non-cancerous organs is a GenXer (i.e., 35 < age ≤ 51) is 35.7%.

This result is confirmed by the R code.

> 5. The Normal Distribution applet, what is the probability a patient from the population of elective surgical procedures of a non-cancerous organ is a Baby Boomer (i.e., 51 < age ≤ 70)? Show your work by taking a screenshot of the applet image. Confirm your result in R by pasting your code and result here!

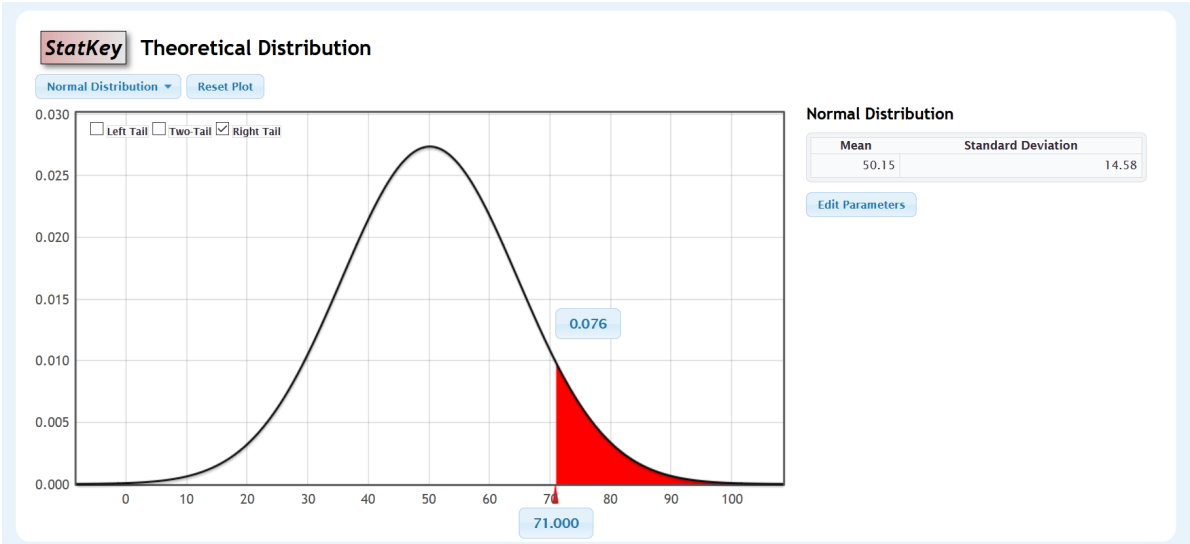



```
pnorm(70, mean = 50.15, sd = 14.58) - pnorm(52, mean = 50.15, sd = 14.58)

0.3628301
```

As seen in the Normal Distribution applet, the probability a patient from the population of elective surgical procedures of non-cancerous organs is a Baby Boomer (i.e., 51 < age ≤ 70) is 39.0%.

This result is the rounded value of the same result calculated from the R code.

> 6. Using the Normal Distribution applet, what is the probability a patient from the population of elective surgical procedures of a non-cancerous organ is a Silent Generation (i.e., age > 70)? Show your work by taking a screenshot of the applet image. Confirm your result in R by pasting your code and result here!




```
pnorm(71, mean = 50.15, sd = 14.58, lower.tail = FALSE)

[1] 0.0763526
```

As seen in the Normal Distribution applet, the probability a patient from the population of elective surgical procedures of non-cancerous organs is a Silent Generation (i.e., age > 70) is 7.6%.

This result is confirmed by the R code.

> 7. Suppose you converted the age distribution into a Standard Normal distribution. Find the Z-score for a patient that is 45 years old and interpret what this value means in the context of the problem.

```
z_score_age_45 = (45 - 50.15) / 14.58

z_score_age_45
[1] -0.3532236
```

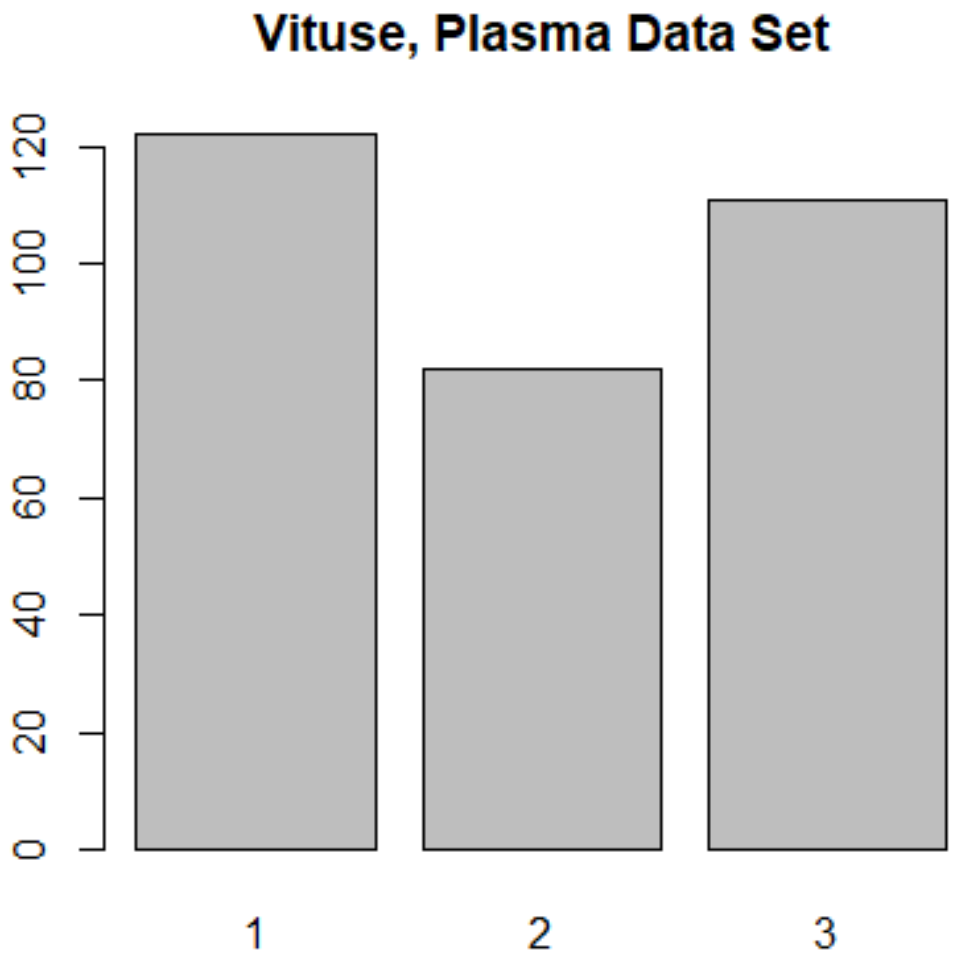
The Z-score for a patient that is 45 years old is -0.353. This means that age 45 is 0.353 standard deviations below the mean.

## EXPLORATORY DATA ANALYSIS FOR VITAMIN USE

> 8. Calculate summary statistics for the vituse variable. In addition, create a barplot for the vituse variable.

| Vitamin Use | Count | Proportion |
|---|---|---|
| **Yes, fairly often** | 122 | 0.3873016 |
| **Yes, not often** | 82 | 0.2603175 |
| **No** | 111 | 0.3523810 |

```
# summary statistics and barplot for vituse variable
t_vituse = table(plasma$vituse)
t_prop_vituse = prop.table(t_vituse)
summary(plasma$vituse)
counts_vituse = table(plasma$vituse)
barplot(counts_vituse, main = "Vituse, Plasma Data Set")

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   1.000   2.000   1.965   3.000   3.000
```




> 9. If we were to classify the categories of vitamin use as “No” and “Yes” (and ignore the qualifiers “fairly often” and “not often”), would it be appropriate to approximate the distribution of the number of patients who don’t use vitamins as a Binomial distribution? Explain your reasoning using the conditions for the binomial distribution discussed in lecture.

Whether or not a patient in the study uses vitamins can be considered as an individual random event. There is a fixed number of patients in this study, so the number of trials *n* is fixed. With the categories being changed to only "Yes" and "No", there are only two possible outcomes for this variable which can be classified as *success* and *failure*. For each trial (i. e. the patient using vitamins or not), the probability of success does not change and is the same for each trial. Therefor all conditions for a binomial distribution are met and it would be appropriate to approximate the distribution as a Binomial distribution.

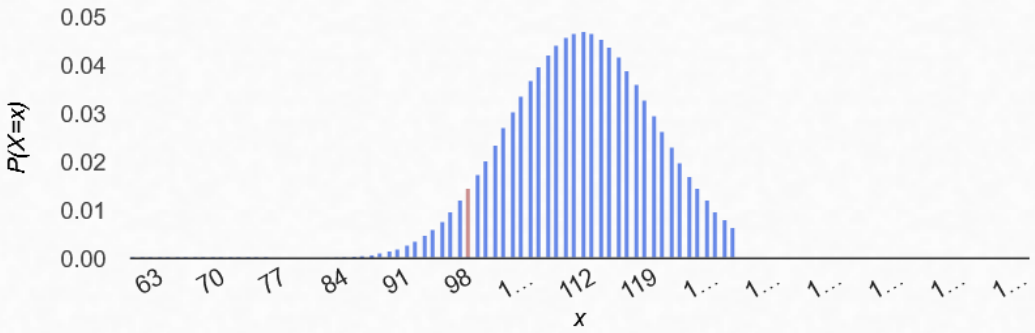
> 10. Using the Binomial Distribution applet, what is the probability that exactly 98 patients from the population of elective surgical procedures of a non-cancerous organs don’t use vitamins? Show your work by taking a screenshot of the applet. Confirm your solution in R and share the code and your result here.

**Binomial Distribution**

$X \sim Bin(n, p)$

$n =$ 315 $p =$ 0.3523810

$x =$ 98 P(X = x) = 0.01463




$\mu = E(X) = 111$ $\sigma = SD(X) = 8.479$ $\sigma^2 = Var(X) = 71.886$

```
dbinom(98, size = 315, prob = 0.3523810)

[1] 0.01463122
```

The probability of exactly 98 not using vitamins in this study is 1.46%. This result is confirmed by the R code.

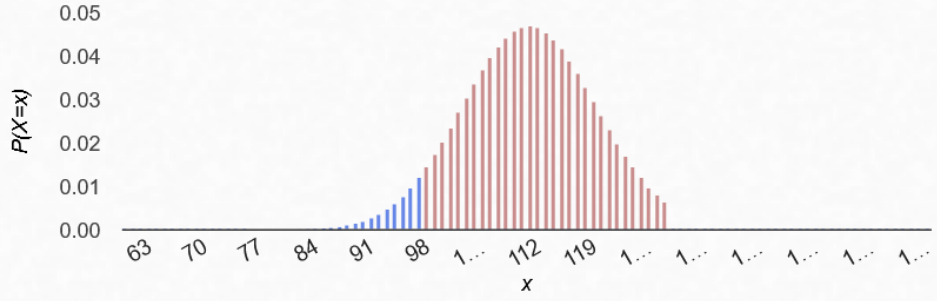
> 11. Using the Binomial Distribution applet, what is the probability that at least (greater than or equal to) 98 patients from the population of elective surgical procedures of a noncancerous organs don't use vitamins? Show your work by taking a screenshot of the applet.
> Confirm your solution in R and share the code and your result here.

**Binomial Distribution**

$X \sim Bin(n, p)$

$n =$ 315 $p =$ 0.3523810

$x =$ 98 P(X ≥ x) = 0.94544




$\mu = E(X) = 111$ $\sigma = SD(X) = 8.479$ $\sigma^2 = Var(X) = 71.886$

```
pbinom(98, size = 315, prob = 0.3523810, lower.tail = FALSE)
    + dbinom(98, size = 315, prob = 0.3523810)

[1] 0.9454389
```

The probability of at least 98 not using vitamins in this study is 94.54%. This result is confirmed by the R code.

## Complete Script for R Assignment 3

```
# read data set from plasma.csv
plasma = read.csv("plasma.csv")
names(plasma)   # see all variables

# summary for age and cholesterol
summary(plasma$age)
describe(plasma$age)
iqr_age = 62.5 - 48.0    # calculate IQR from Q3-Q1

summary(plasma$cholesterol)
describe(plasma$cholesterol)
iqr_chol = 308.9 - 155.0    # calculate IQR from Q3-Q1

# create single graph for variables
par(mfrow= c(1,2))    # set two plots in one graph
hist(plasma$age, main = "Age, Plasma Data Set", xlab = "Age")
hist(plasma$cholesterol,
     main = "Cholesterol, Plasma Data Set", xlab = "Cholesterol",
     breaks = 15,
     xlim = c(0, 1000), ylim = c(0, 80))
par(mfrow=c(1,1))


# Q-Q plots for age and cholesterol

qqnorm(plasma$age, main = "Normal Q-Q Plot Age")
qqline(plasma$age)

qqnorm(plasma$cholesterol, main = "Normal Q-Q Plot Cholesterol")
qqline(plasma$cholesterol)

# probability patient belonging to different generations
pnorm(35, mean = 50.15, sd = 14.58)
pnorm(51, mean = 50.15, sd = 14.58) - pnorm(36, mean = 50.15, sd = 14.58)
pnorm(70, mean = 50.15, sd = 14.58) - pnorm(52, mean = 50.15, sd = 14.58)
pnorm(71, mean = 50.15, sd = 14.58, lower.tail = FALSE)

# z-score for patient age 45
z_score_age_45 = (45 - 50.15) / 14.58   #(X - x_bar) / s

# summary statistics and barplot for vituse variable
t_vituse = table(plasma$vituse)
t_prop_vituse = prop.table(t_vituse)
```

```r
summary(plasma$vituse)
counts_vituse = table(plasma$vituse)
barplot(counts_vituse, main = "Vituse, Plasma Data Set")

# binomial distribution vituse
dbinom(98, size = 315, prob = 0.3523810)
pbinom(98, size = 315, prob = 0.3523810, lower.tail = FALSE) + dbinom(98, size =
315, prob = 0.3523810)
```