ADM 2304 - Assignment 1

Renad Gharz

16/05/2022

```
## Loading required package: pps
## Loading required package: sampling
## Loading required package: survey
## Loading required package: grid
## Loading required package: Matrix
## Loading required package: survival
##
## Attaching package: 'survival'
## The following objects are masked from 'package:sampling':
##
##     cluster, strata
##
## Attaching package: 'survey'
## The following object is masked from 'package:graphics':
##
##     dotchart
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

1. Real Estate

The dataset RealEstate contains information on the listing of 1,047 real estate properties in a certain region.

1.1.a

> Treating the data in column **Living Area** [**sq ft**] as the population, use software to find the population mean and the population standard deviation. Is the population data reasonably normal? Examine a boxplot and a histogram of the data in column **Living Area** [**sq ft**] to justify your answer. From here on, assume that the population standard deviation is not known.

Mean and Standard Deviation

```r
##Mean
realestate_mean <- round(mean(
  realestate$`Living Area [sq ft]`), 2)
realestate_mean
```

```
## [1] 1807.3
```

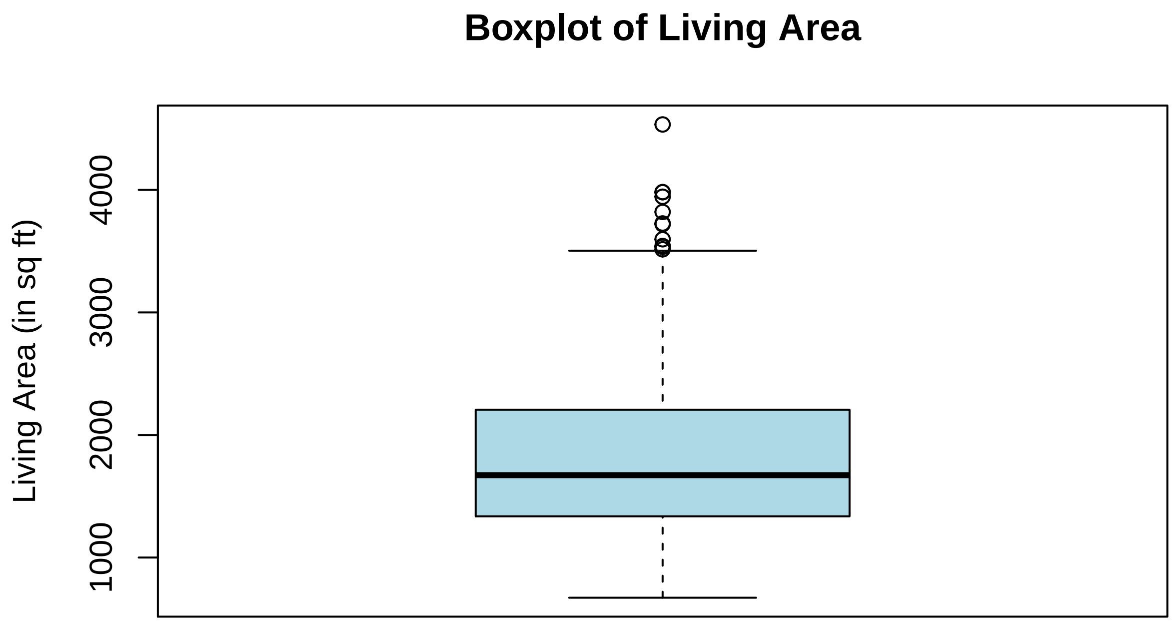
```r
##Standard deviation
realestate_sd <- round(sd(
  realestate$`Living Area [sq ft]`), 2)
realestate_sd
```

```
## [1] 641.46
```

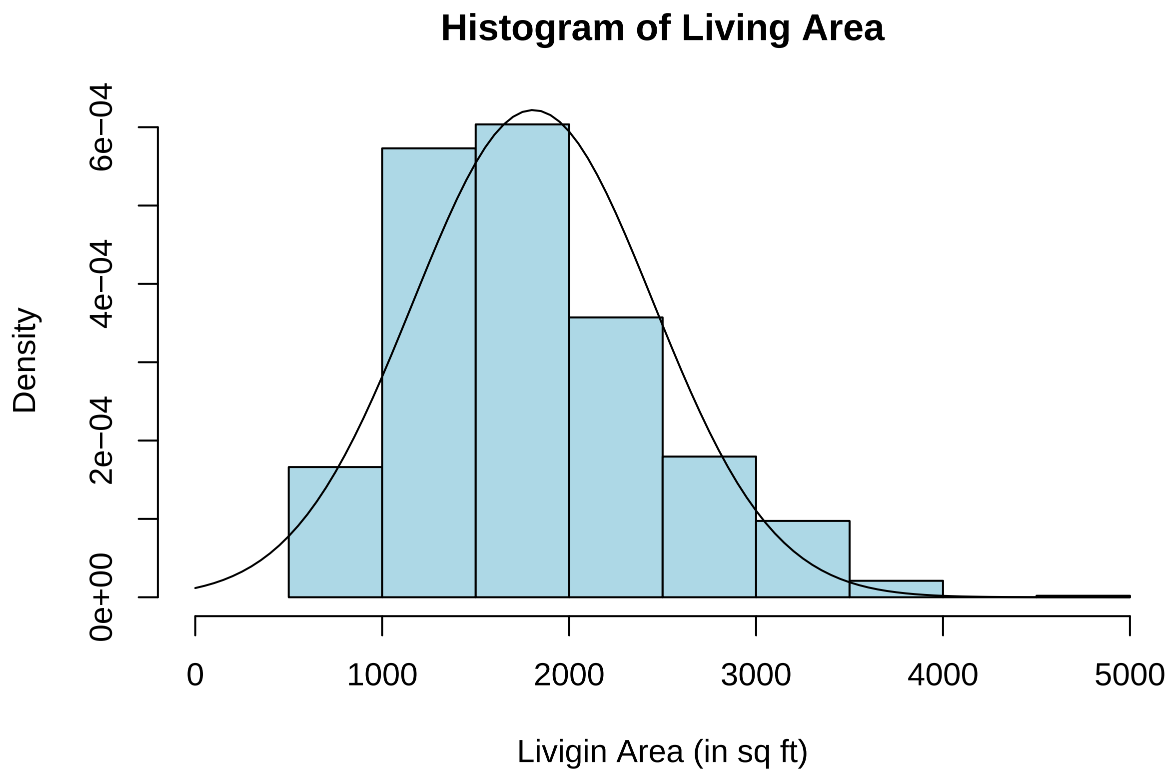
The population mean μ for **Living Area** [**sq ft**] is **1,807.3 square feet** and the population standard deviation σ is **641.5 square feet**.

Checking Normality

```r
##Boxplot
boxplot(realestate$`Living Area [sq ft]`,
        main = "Boxplot of Living Area",
        ylab = "Living Area (in sq ft)",
        col = "light blue")
```




```
##Histogram
hist(realestate$`Living Area [sq ft]`,
     main = "Histogram of Living Area",
     xlab = "Livigin Area (in sq ft)",
     xlim = c(0, 5000),
     prob = TRUE,
     col = "light blue")
##Adding normal curve to histogram
curve(dnorm(x,
            mean = realestate_mean,
            sd = realestate_sd),
      add = TRUE)
```



The population of **Living Area [sq ft]** is skewed to the right however the skew is only moderate and not that severe thus the population can still be considered relatively normal. We can tell there is skew by looking at the median line in the boxplot that is shifted to the left and at the outlier points that are above the upper fence. However, by looking at the histogram we can also see that although the bars don't perfectly align with the normal curve, they still look relatively normal thus the distribution is relatively normal.

1.1.b

> Suppose that for a family of four, the ideal property size is between 2,400 and 2,800 square feet. Use software to code the data in column **Living Area [sq ft]** and put the result in column **Coded Data**. Code living areas between 2,400 and 2,800 square feet as "1" and code living areas outside of this interval as "0". Now calculate the population proportion using software and report the result.

Coded Data Column

```r
##Creating new column with coded data
realestate['Coded Data'] <-
  ifelse(realestate$`Living Area [sq ft]` >= 2400 &
           realestate$`Living Area [sq ft]` <= 2800, 1, 0)
head(as.data.frame(realestate))
```

```
##   Price [$] Living Area [sq ft] Bathrooms Bedrooms Fireplace Lot Size [acres]
## 1    142212                1982       1.0        3         0             2.00
## 2    134865                1676       1.5        3         1             0.38
## 3    118007                1694       2.0        3         1             0.96
## 4    138297                1800       1.0        2         1             0.48
```

```
## 5    129470                2088       1.0        3          1                1.84
## 6    206512                1456       2.0        3          0                0.98
##   Age [years] Sample 1 Coded Data
## 1         133     1512          0
## 2          14     1920          0
## 3          15     2200          0
## 4          49     2843          0
## 5          29     1032          0
## 6          10     1752          0
```

Population Proportions

```r
##Proportions of coded data
##Ideal properties population proportion
pop_ideal_props <-
  round(
    length(which(realestate$`Coded Data` == 1)) /
      nrow(realestate),
    digits = 4)
pop_ideal_props
```

```
## [1] 0.0946
```

```r
##Non ideal properties population proportion
pop_non_ideal_props <-
  round(
    length(which(realestate$`Coded Data` == 0)) /
      nrow(realestate),
    digits = 4)
pop_non_ideal_props
```

```
## [1] 0.9054
```

In our population of 1,047 houses, there are 99 houses that meet the criteria for the ideal property size (between 2,400 and 2,800 square feet). Thus, the population proportion is **9.46% (0.0946)**.

1.1.c

> The data in column **Sample 1** is a simple random sample drawn from the data in column **Living Area** [**sq ft**]. Calculate manually a 95% confidence interval for the population mean based on this sample and confirm your calculations using software.

```r
##Computing the 95% C.I. of Sample 1
sample1_mean_ci <-
  MeanCI(realestate$`Sample 1`,
         conf.level = 0.95,
         na.rm = TRUE)
sample1_mean_ci
```

```
##     mean   lwr.ci   upr.ci
## 2012.100 1802.479 2221.721
```

We are 95% certain that the population mean will fall between **1,803 and 2,221 square feet**, 19 times out of 20.

1.1.d

Code the data in column **Sample 1** using the same instructions as in part b) above and put the result in column **Sample_p**. Calculate manually a 90% confidence interval for the population proportion of properties with living areas between 2,400 and 2,800 square feet and confirm your calculations using software. Assume that the required conditions are met, and you can use the normal approximation.

Coded Data for Sample 1

```
## Creating new column for Sample 1 coded data
realestate['Sample_p'] <-
  ifelse(realestate$`Sample 1` >= 2400 &
           realestate$`Sample 1` <= 2800, 1, 0)
head(as.data.frame(realestate))
```

```
##   Price [$] Living Area [sq ft] Bathrooms Bedrooms Fireplace Lot Size [acres]
## 1    142212                1982       1.0        3         0             2.00
## 2    134865                1676       1.5        3         1             0.38
## 3    118007                1694       2.0        3         1             0.96
## 4    138297                1800       1.0        2         1             0.48
## 5    129470                2088       1.0        3         1             1.84
## 6    206512                1456       2.0        3         0             0.98
##   Age [years] Sample 1 Coded Data Sample_p
## 1         133     1512          0        0
## 2          14     1920          0        0
## 3          15     2200          0        0
## 4          49     2843          0        0
## 5          29     1032          0        0
## 6          10     1752          0        0
```

Sample Proportions

```
##Ideal properties sample proportion (s1)
sample1_ideal_props <-
  round(
    length(which(realestate$`Sample_p` == 1)) /
      length(na.omit(realestate$Sample_p)),
    digits = 4)
sample1_ideal_props
```

```
## [1] 0.1
```

```
##Non ideal properties sample proportion (s1)
sample1_non_ideal_props <-
  round(
    length(which(realestate$`Sample_p` == 0)) /
      length(na.omit(realestate$Sample_p)),
    digits = 4)
sample1_non_ideal_props
```

```
## [1] 0.9
```

In our sample of 40 houses, there are 4 houses that meet the criteria for the ideal property size (between 2,400 and 2,800 square feet). Thus, the sample proportion is **10% (0.1)**.

Confidence Interval

```r
##90% C.I. for Sample 1 (proportion)
sample1_prop_ci <-
  round(
    proportion.CI(
      sample1_ideal_props,
      length(na.omit(realestate$`Sample_p`)),
      conf.level = 0.90)$CI,
    digits = 4)
```

```
##
##                                  Proportion estimate: 0.1
##
##                         90% confidence interval for a proportion
##
##       (hat.p - z1<U+208B>a<U+0338>2*sqrt(hat.p*(1-hat.p)/n)  ,  hat.p + z1<U+208B>a<U+0338>2*sqrt(ha
##
##                                    (0.02198 , 0.17802)
```

```r
sample1_prop_ci
```

```
## [1] 0.022 0.178
```

We are 90% confident that **2.2% to 17.8%** of all houses in a certain region have living areas between 2,400 and 2,800 square feet.

1.1.e

> Now use software to randomly draw 19 additional samples of size n = 40 from column **Living Area** [**sq ft**]. The procedure must be repeated 19 times. Put these 19 samples in columns **Sample 2, Sample 3, . . . , Sample 19, Sample 20**. For each of these additional samples, use software to calculate the 95% confidence interval for the population mean.

Creating Samples

```r
##Creating the 19 random samples
samples <-
  as.data.frame(
    replicate(19,sample(
      realestate$`Living Area [sq ft]`,40)))
head(as.data.frame(samples))
```

```
##     V1   V2   V3   V4   V5   V6   V7   V8   V9  V10  V11  V12  V13  V14  V15
## 1 1564 1164 2334 1040 2248 1542 1508 1656 1096  864 2275 2011 3296 1456 2388
## 2 1314 2248 1248 1760 3250 2266  990 1852 1388 1164 1472 2475 2310 1281 1480
## 3 2038 3361  957 1056 2526 1150  912 1921 2412 2310 2498 2564 1056 2526 2133
## 4 1302 1540 1270 1802  990 1294 3020 3504  908 1656 3020 1480 2993 1800  960
## 5 2835 3308 2648 1536 2541 3236 2000 2028 2748 2039 1200 2872 2576 1276 2088
## 6 2216  960 1834 1586 1380 1080  784  995 2000 1623 1184 3015 2960 1568 1831
##    V16  V17  V18  V19
## 1 1743 1944 2490 1324
## 2 1628 1380 1253 1480
## 3 1445 2762 2084 1144
## 4 1480 1744 2412 1740
## 5 2068 1664 3250 1164
```

```
## 6 1110 2960 1360 3944
```

```r
#Renaming the 19 random samples
for (i in 2:20) {
  colnames(samples)[i-1] <- paste("Sample", i)
  i<- i +1
}
head(as.data.frame(samples))
```

```
##   Sample 2 Sample 3 Sample 4 Sample 5 Sample 6 Sample 7 Sample 8 Sample 9
## 1     1564     1164     2334     1040     2248     1542     1508     1656
## 2     1314     2248     1248     1760     3250     2266      990     1852
## 3     2038     3361      957     1056     2526     1150      912     1921
## 4     1302     1540     1270     1802      990     1294     3020     3504
## 5     2835     3308     2648     1536     2541     3236     2000     2028
## 6     2216      960     1834     1586     1380     1080      784      995
##   Sample 10 Sample 11 Sample 12 Sample 13 Sample 14 Sample 15 Sample 16
## 1      1096       864      2275      2011      3296      1456      2388
## 2      1388      1164      1472      2475      2310      1281      1480
## 3      2412      2310      2498      2564      1056      2526      2133
## 4       908      1656      3020      1480      2993      1800       960
## 5      2748      2039      1200      2872      2576      1276      2088
## 6      2000      1623      1184      3015      2960      1568      1831
##   Sample 17 Sample 18 Sample 19 Sample 20
## 1      1743      1944      2490      1324
## 2      1628      1380      1253      1480
## 3      1445      2762      2084      1144
## 4      1480      1744      2412      1740
## 5      2068      1664      3250      1164
## 6      1110      2960      1360      3944
```

```r
##Adding Sample 1 to df and shifting position
samples <-
  cbind(
    samples, "Sample 1" =
      na.omit(realestate$`Sample 1`))
samples <-
  samples %>% select("Sample 1", everything())
head(as.data.frame(samples))
```

```
##   Sample 1 Sample 2 Sample 3 Sample 4 Sample 5 Sample 6 Sample 7 Sample 8
## 1     1512     1564     1164     2334     1040     2248     1542     1508
## 2     1920     1314     2248     1248     1760     3250     2266      990
## 3     2200     2038     3361      957     1056     2526     1150      912
## 4     2843     1302     1540     1270     1802      990     1294     3020
## 5     1032     2835     3308     2648     1536     2541     3236     2000
## 6     1752     2216      960     1834     1586     1380     1080      784
##   Sample 9 Sample 10 Sample 11 Sample 12 Sample 13 Sample 14 Sample 15
## 1     1656      1096       864      2275      2011      3296      1456
## 2     1852      1388      1164      1472      2475      2310      1281
## 3     1921      2412      2310      2498      2564      1056      2526
## 4     3504       908      1656      3020      1480      2993      1800
## 5     2028      2748      2039      1200      2872      2576      1276
## 6      995      2000      1623      1184      3015      2960      1568
##   Sample 16 Sample 17 Sample 18 Sample 19 Sample 20
## 1      2388      1743      1944      2490      1324
```

```
## 2       1480       1628       1380       1253       1480
## 3       2133       1445       2762       2084       1144
## 4        960       1480       1744       2412       1740
## 5       2088       2068       1664       3250       1164
## 6       1831       1110       2960       1360       3944
```

Calculating Confidence Intervals

```r
##Computing 95% C.I.s for 19 samples
samples_ci <- list()
for (i in 1:20) {
  samples_ci[[i]] <- MeanCI(unlist(samples[i]), conf.level = 0.95)
  i + 1
}
samples_ci <- as.data.frame(t(samples_ci))
##Renaming the columns of the new df
for (i in 1:20) {
  colnames(samples_ci)[i] <- paste("Sample", i)
  i +1
}
as.list(samples_ci)
```

```
## $`Sample 1`
## $`Sample 1`[[1]]
##     mean   lwr.ci   upr.ci
## 2012.100 1802.479 2221.721
##
##
## $`Sample 2`
## $`Sample 2`[[1]]
##     mean   lwr.ci   upr.ci
## 1904.850 1724.368 2085.332
##
##
## $`Sample 3`
## $`Sample 3`[[1]]
##     mean   lwr.ci   upr.ci
## 1943.525 1687.179 2199.871
##
##
## $`Sample 4`
## $`Sample 4`[[1]]
##     mean   lwr.ci   upr.ci
## 1787.100 1561.375 2012.825
##
##
## $`Sample 5`
## $`Sample 5`[[1]]
##     mean   lwr.ci   upr.ci
## 1890.725 1660.015 2121.435
##
##
## $`Sample 6`
## $`Sample 6`[[1]]
```

```
##      mean   lwr.ci   upr.ci
## 1782.350 1574.507 1990.193
##
##
## $`Sample 7`
## $`Sample 7`[[1]]
##      mean   lwr.ci   upr.ci
## 1885.800 1651.187 2120.413
##
##
## $`Sample 8`
## $`Sample 8`[[1]]
##      mean   lwr.ci   upr.ci
## 1749.475 1528.043 1970.907
##
##
## $`Sample 9`
## $`Sample 9`[[1]]
##      mean   lwr.ci   upr.ci
## 1999.900 1783.794 2216.006
##
##
## $`Sample 10`
## $`Sample 10`[[1]]
##     mean  lwr.ci  upr.ci
## 1757.50 1536.78 1978.22
##
##
## $`Sample 11`
## $`Sample 11`[[1]]
##      mean   lwr.ci   upr.ci
## 1957.075 1724.482 2189.668
##
##
## $`Sample 12`
## $`Sample 12`[[1]]
##      mean   lwr.ci   upr.ci
## 1810.400 1610.039 2010.761
##
##
## $`Sample 13`
## $`Sample 13`[[1]]
##      mean   lwr.ci   upr.ci
## 1937.625 1763.101 2112.149
##
##
## $`Sample 14`
## $`Sample 14`[[1]]
##      mean   lwr.ci   upr.ci
## 1813.050 1630.096 1996.004
##
##
## $`Sample 15`
## $`Sample 15`[[1]]
```

```
##     mean   lwr.ci   upr.ci
## 1635.950 1452.847 1819.053
##
##
## $`Sample 16`
## $`Sample 16`[[1]]
##     mean   lwr.ci   upr.ci
## 1888.675 1672.250 2105.100
##
##
## $`Sample 17`
## $`Sample 17`[[1]]
##     mean   lwr.ci   upr.ci
## 1783.675 1595.804 1971.546
##
##
## $`Sample 18`
## $`Sample 18`[[1]]
##     mean   lwr.ci   upr.ci
## 1816.300 1628.552 2004.048
##
##
## $`Sample 19`
## $`Sample 19`[[1]]
##     mean   lwr.ci   upr.ci
## 1941.175 1689.883 2192.467
##
##
## $`Sample 20`
## $`Sample 20`[[1]]
##     mean   lwr.ci   upr.ci
## 1713.875 1484.589 1943.161
```

1.1.f

> Now count the number of confidence intervals, obtained from all the 20 samples, that contain the true value of the population mean from part a). Is this what you might expect? Explain your answer.

Of the confidence intervals computed, 19 of them contain the true value of the population while only Sample 3 does not contain it.

This result is not surprising and is expected because we are building 95% confidence intervals around the population mean from a). With 95% C.I.s we are saying that we are 95% confident that the population mean will be included within our sample, 19 times out of 20. Thus, the fact that there is one sample among the 20 samples that does not contain the population mean is expected.

1.2.a

> Using the data in column **Sample 1**, manually test the hypothesis that the population mean is not equal to 2,000 square feet. Use a 5% significance level and the critical value approach. Confirm your results using software. Is your conclusion supported by the confidence interval from part c)? Explain your answer.

Defining Hypotheses

$$\begin{cases} H_0 : \mu = 2000 & \text{population mean of living are is equal to 2,000 sq. ft.} \\ H_A : \mu \neq 2000 & \text{population mean of living area is not equal to 2,000 sq. ft} \end{cases}$$

Hypothesis Test and Statistic

```r
##Hypothesis test
a1_q1_p2_a_ht <-
  t.test(realestate$`Sample 1`, mu = 2000)
a1_q1_p2_a_ht
```

```
##
##  One Sample t-test
##
## data:  realestate$`Sample 1`
## t = 0.11676, df = 39, p-value = 0.9077
## alternative hypothesis: true mean is not equal to 2000
## 95 percent confidence interval:
##  1802.479 2221.721
## sample estimates:
## mean of x
##    2012.1
```

```r
##Test statistic
round(a1_q1_p2_a_ht$statistic,3)
```

```
##     t
## 0.117
```

Critical Value

```r
##Critical value
a1_q1_p2_a_ct <-
  round(
    qt(0.05/2, 39,
       lower.tail = FALSE),
    digits = 3)
a1_q1_p2_a_ct
```

```
## [1] 2.023
```

Validating Test

```r
##Validating results
a1_q1_p2_a_ht$statistic > a1_q1_p2_a_ct
```

```
##     t
## FALSE
```

Because the t-statistic is not greater than the critical value ($0.117 \ngtr 2.023$), we fail to reject the null hypothesis H_0 meaning that we there is not sufficient evidence to prove that the population mean of living area is 2,000 sq. ft.

1.2.b

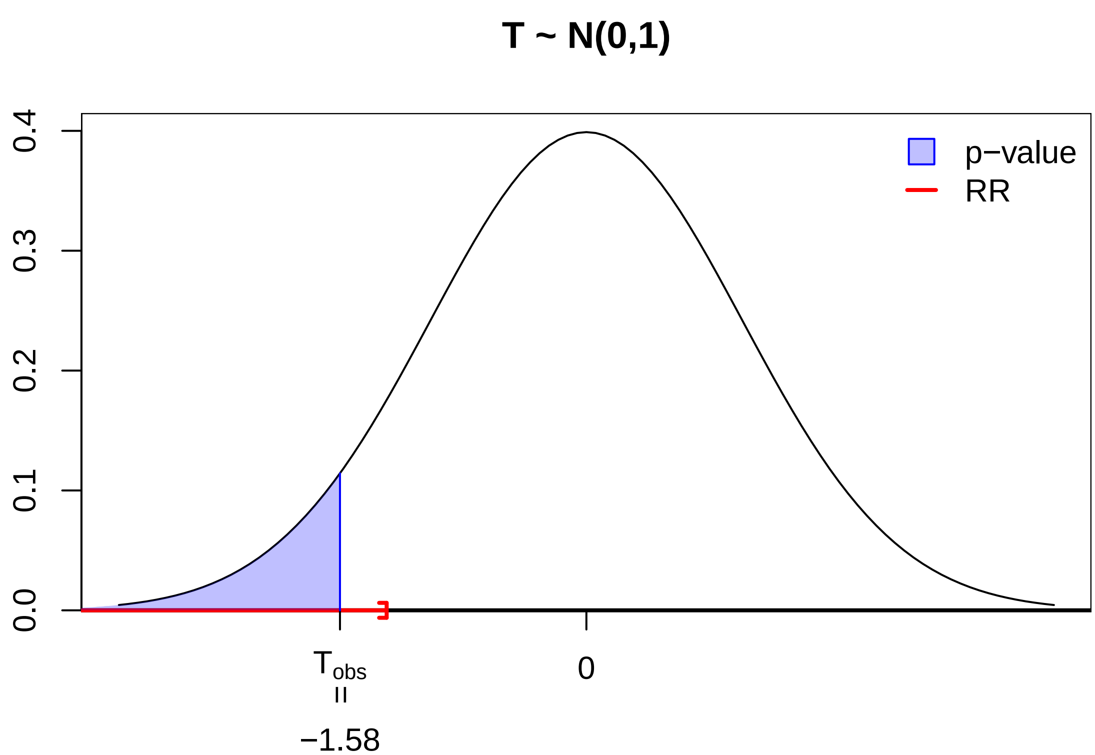
Using the data in column Sample_p, manually test the hypothesis that the population proportion of properties that are ideal for a family of four is less than 20%. Use a 10% significance level. Calculate the p-value manually (i.e., using a normal distribution table) and explain how it confirms the conclusion reached by using the critical value approach. Assume that the normal approximation is reasonable in this case. Check your results using software.

Defining Hypotheses

$$\begin{cases} H_0 : p \nless 0.2 & \text{population proportion of properties is not less than 20\%} \\ H_A : p < 0.2 & \text{population proportion of properties is less than 20\%} \end{cases}$$

Hypothesis Test and Statistic

```
##Hypothesis test
a1_q1_p2_b_ht <-
  proportion.test(sample1_ideal_props,
                  length(na.omit(
                    realestate$`Sample 1`)),
                  p0 = 0.2,
                  alternative = "less",
                  alpha = 0.1,
                  plot = TRUE)
```




```
a1_q1_p2_b_ht
```

```
##
##  Test for a proportion
##
## H0: p = 0.2
## H<U+2090>: p < 0.2
## T = (hat.p - p0) / sqrt(p0 * (1 - p0) / n)
## T ~ N(0,1)
## a = 0.1
## T_obs = -1.58114
## RR = (-8, -1.28155]
## p-value = 0.05692
```

```
##Test statistic
a1_q1_p2_b_ht$statistic
```

```
##         T
## -1.581139
```

Critical Value

```
##Critical value
a1_q1_p2_b_ct <-
  round(qnorm(0.05),
        digits = 3)
a1_q1_p2_b_ct
```

```
## [1] -1.645
```

Validating Test

```
##Validating results
a1_q1_p2_b_ht$statistic > a1_q1_p2_b_ct
```

```
##    T
## TRUE
```

The critical value of $\alpha = 0.1$ on the left side is -1.645 while the z-statistic is z-stat = -1.58. Because the z-statistic is bigger than the critical value (-1.58 > -1.645), we reject the null hypothesis H_0. There is sufficient evidence to conclude that the population proportion of properties that are ideal for a family is less than 20%.

1.3.a

> Suppose you want to estimate the average living area of the real estate properties in the region. If you want to obtain a 95% confidence interval with a margin of error of ± 50 square feet, what sample size would you recommend? Assume for this exercise that the population standard deviation is 641 square feet.

```
##Recommended sample size for mu
required_n_mean <- round_any(
  MeanCIn(ci=c(realestate_mean - 50,
               realestate_mean + 50),
          sd = 641,
          conf.level = 0.95,
          norm = TRUE),
  1, f = ceiling)
required_n_mean
```

```
## [1] 632
```

The recommended sample size is 632 houses.

1.3.b

Assume that you now would like to know what proportion of the real estate properties in the region are ideal for a family of four. This population proportion is not known. To estimate this population proportion with a margin of error of ± 0.02, what sample size would you recommend? Consider a 90% confidence level.

```
required_n_prop <- round_any(
  sample.size.prop(0.02,
                   sample1_ideal_props,
                   level = 0.9)$n,
  1, f = ceiling)
required_n_prop
```

```
## [1] 609
```

The recommended sample size 609 houses.

2. Package Subscribers

Bell provides cable, phone, and internet services to customers, some of whom subscribe to *packages* consisting of multiple services. Suppose that in Ontario 25% of Bell customers are package subscribers. A local Bell representative in Ottawa wonders if the proportion of package subscribers in the city is larger than the provincial proportion. After sending a survey to 100 customers from his subscriber list at random, only 25 of them responded, and of those, 11 are package subscribers. Does this constitute sufficient evidence that the true proportion of package subscribers in the Ottawa is more than the provincial proportion? Consider a 5% significance level and clearly explain the reasoning behind your answer.

Defining Hypotheses

$$\begin{cases} H_0 : p = 0.25 & \text{true proportion of package subscribers is equal to 25\%} \\ H_A : p > 0.25 & \text{true proportion of package subscriebrs is greater than 25\%} \end{cases}$$

Checking Conditions

$$\text{Success-failure conditions} \begin{cases} np \geq 10 \rightarrow 25 \cdot 0.25 = 6.25 \not> 10 \\ n(1-p) \geq 10 \rightarrow 25 \cdot 0.75 = 18.75 > 10 \end{cases}$$

Only one of the success-failure conditions is met so we cannot use the normal distribution. We need to use the binomial distribution instead so that:

$$\hat{p} \rightarrow Bin(np, \sqrt{npq})$$

Distribution Parameters

$$\text{Distribution parameters} \begin{cases} \mu : np = 25 \cdot 0.25 = 6.25 \\ \sigma : \sqrt{npq} = \sqrt{6.25 \cdot 0.75} = 2.1651 \end{cases}$$

The binomial distribution now looks like this: $\hat{p} \rightarrow Bin(6.25, 2.1651)$

Binomial Distribution

```r
##Mean
pckg_subs_mean <- 25*0.25

##Standard deviation
pckg_subs_sd <-
  round(
    sqrt(pckg_subs_mean*0.75),
    digits = 4)

##P-value using binomial
pckg_subs_prob <-
  round(
    1- pbinom(10,
              25,
              p=0.25),
    digits = 4)
pckg_subs_prob
```

```
## [1] 0.0297
```

The p-value is 0.0297 or 2.97%.

Validating Test

```r
##Validating test
pckg_subs_prob < 0.05
```

```
## [1] TRUE
```

Since the p-value is less than our significance level ($\alpha = 0.05$); we reject the null hypothesis H_0 in favour of the alternative hypothesis H_A. There is sufficient evidence that the true proportion of Bell's package subscribers in Ottawa is more than the provincial proportion.