Midterm_jn2761

jiaying Ning

10/29/2020

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(arsenal)
```

problem 3

1) select a random sample (with replacement) of 16 men from this distribution of 2742 normally distribut4ed observations and calculate the corresponding mean

```
set.seed(8130)
population = rnorm(2742, mean = 174, sd = 7.7)
sample = sample (population, size=16, replace =F)
sample_mean = mean(sample)
```

The mean for the sample of 16 men is 173.0859471

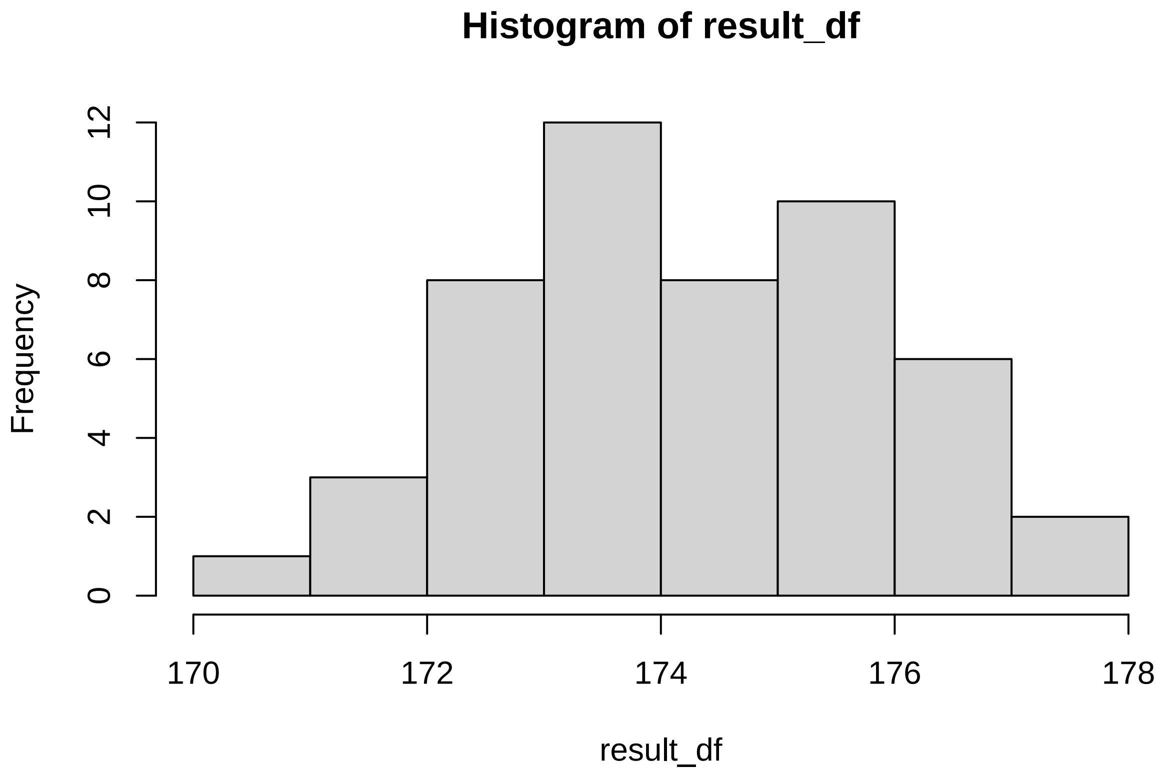
2. Repeat the step 1 50 times and plot the distribution of 50 sample means.

```
  l <- list()
result=list()
set.seed(8130)

for (i in 1:50){
l[i] <- list(sample (population, size=16, replace =F))
result[i] = list(mean(l[[i]]))
}
result_df = c(unlist(result))
result_df
```

```
##  [1] 175.9842 175.0434 170.7959 176.8533 174.4956 177.9554 172.9537 173.6712
##  [9] 175.4699 174.6011 172.8771 175.1274 176.9327 177.3581 174.2563 173.1683
## [17] 176.1784 176.1483 173.5244 173.6647 176.6682 174.1040 174.6961 173.1209
## [25] 174.2942 172.9185 172.1531 172.9259 176.1254 174.0839 173.0151 171.5019
## [33] 173.2446 173.4495 173.2195 171.7051 175.5552 171.5936 173.7085 175.3552
## [41] 173.7376 172.4736 172.4176 175.8581 175.2298 175.2760 174.0049 175.5772
## [49] 172.9777 173.2086
```

```
hist(result_df)
```



3. *Comment: the shape of the distribution in part 2 is approximatly normal with most of the value gathered around 174, the true population mean. According to central limited theorem, since our underlying population distribution is normal, even though n <30, then the shape of the samping distribution is also approximately normal.*

problem 5

```
#Import Data
vote_df=
   read.csv("./data/VoteNY2018.csv")
```

```
voted_recode=
vote_df %>%
  mutate(
        VOTED=recode(VOTED,`1`="Did not vote",`2`="Voted", `96`="Refused", `97`="Don't know", `98` ="No
        VOTED_SIMPLE = ifelse(VOTED=="Voted","YES","NO"),
         SEX = recode(SEX, `1`= "Male",`2`="Female",`9`="NIU"),
        RACESIMPLE=recode(RACESIMPLE, `1`   = "White", `2`= "Black", `3`= "American Indian or Aleut or I
        EDUSIMPLE=recode( EDUSIMPLE, `0` = "No school", `1`= "Some school but no diploma", `2` =    "Hig

          relocate(VOTED,VOTED_SIMPLE,AGE,SEX,RACESIMPLE,EDUSIMPLE)
```

```
summary(vote_df$AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   35.00   52.00   51.11   66.00   85.00
```

```
summary(vote_df$VOTED)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.666   2.000   2.000
```

I did not use the coding for AGE since there is no value of 0,99, and 99+ in the current data for age. The minimum data is 18 and the maximum data is 85.

Create a summary table of age, sex, race and education level by voting status (yes – no) in the 2018 elections. Carefully choose the summary statistics to report based on the variable type.

```
 tble = tableby( VOTED_SIMPLE ~ AGE+SEX+RACESIMPLE+EDUSIMPLE, data=voted_recode)
summary(tble, title = "Descriptive Statistics: voting in 2018",  text=T)
```

```
##
## Table: Descriptive Statistics: voting in 2018
##
## |                                      |   NO (N=883)    |  YES (N=1764)   | Total (N=2647)  | p valu
## |:-------------------------------------|:---------------:|:---------------:|:---------------:|-------
## |AGE                                   |                 |                 |                 | < 0.00
## |-  Mean (SD)                          | 46.661 (19.033) | 53.330 (17.824) | 51.105 (18.502) |
## |-  Range                              | 18.000 - 85.000 | 18.000 - 85.000 | 18.000 - 85.000 |
## |SEX                                   |                 |                 |                 |   0.0
## |-  Female                             |   442 (50.1%)   |   968 (54.9%)   |  1410 (53.3%)   |
## |-  Male                               |   441 (49.9%)   |   796 (45.1%)   |  1237 (46.7%)   |
## |RACESIMPLE                            |                 |                 |                 | < 0.00
## |-  American Indian or Aleut or Eskimo |    5 (0.6%)     |    7 (0.4%)     |    12 (0.5%)    |
## |-  Asian or Pacific Islander          |    80 (9.1%)    |    78 (4.4%)    |   158 (6.0%)    |
## |-  Black                              |   104 (11.8%)   |   250 (14.2%)   |   354 (13.4%)   |
## |-  More than one race                 |    13 (1.5%)    |    17 (1.0%)    |    30 (1.1%)    |
## |-  White                              |   681 (77.1%)   |  1412 (80.0%)   |  2093 (79.1%)   |
## |EDUSIMPLE                             |                 |                 |                 | < 0.00
## |-  Associate degree                   |    78 (8.8%)    |   180 (10.2%)   |   258 (9.7%)    |
## |-  Bachelors degree                   |   141 (16.0%)   |   448 (25.4%)   |   589 (22.3%)   |
## |-  High school graduate or GED        |   322 (36.5%)   |   399 (22.6%)   |   721 (27.2%)   |
## |-  Masters degree                     |    62 (7.0%)    |   271 (15.4%)   |   333 (12.6%)   |
## |-  No school                         |    4 (0.5%)     |    2 (0.1%)     |    6 (0.2%)     |
## |-  Professional or Doctoral degree   |   17 (1.9%)     |  106 (6.0%)     |  123 (4.6%)     |
## |-  Some college but no degree        |  142 (16.1%)    |  275 (15.6%)    |  417 (15.8%)    |
## |-  Some school but no diploma        |  117 (13.3%)    |   83 (4.7%)     |  200 (7.6%)     |
```

3. Using a 5% significance level, evaluate the associations between: voting status and race. You need to state the following: hypotheses, table of expected values, statistical test you chose and why, test statistic, critical value, decision rule and interpretation in the context of the problem.

To test the association between two categorical variable, we will construct a contingency table, calculated the expected cell counts and the chi-square statistics testing for independence.

- H0: There is no association between voting status and race, voting status and race are independent
- H1: There is association between voting status and race, voting status and race are dependent

```
table2 = table(voted_recode$RACESIMPLE,voted_recode$VOTED_SIMPLE)
```

```
table2
```

```
##
##                                        NO  YES
##   American Indian or Aleut or Eskimo    5    7
##   Asian or Pacific Islander            80   78
##   Black                               104  250
##   More than one race                   13   17
##   White                               681 1412
```

Here is the table for expected value

```
#expected_value

library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
expected_value = data.table(voted= c(883*12/2647,883*158/2647,883*354/2647,883*30/2647,883*2093/2647)
,not_voted=c(1764*12/2647,1764*158/2647,1764*354/2647,1764*30/2647.1764*2093/2647))
```

```
## Warning in as.data.table.list(x, keep.rownames = keep.rownames, check.names
## = check.names, : Item 2 has 4 rows but longest item has 5; recycled with
## remainder.
```

```
expected_value
```

```
##         voted  not_voted
## 1:   4.003022   7.996978
## 2:  52.706460 105.293540
## 3: 118.089158 235.910842
## 4:  10.007556  15.807101
## 5: 698.193804   7.996978
```

$$X^2 = sumof((originalvalue - expectedvalue)^2/expectedvalue)$$
$$df = (row - 1) * (columb - 1) = 4$$

```
chisq.test(table2)
```

```
## Warning in chisq.test(table2): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table2
## X-squared = 26.082, df = 4, p-value = 3.047e-05
```

```
qchisq(0.95,4)
```

```
## [1] 9.487729
```

- X-squared = 26.082, df = 4, p-value = 3.047e-05
- Critical value: qchisq(0.95,4) = 9.487729
- since Test Value = 26.082 > critical value = 9.487729 We reject the null hypothesis and conclude that the proportions of voting status are different among racial groups at alpha level = 0.05