Worksheet 6

Andrey Sumadic BSIT 2B

2023-12-09

```
#1.
fertilizer_levels <- c(10,10,10, 20,20,50,10,20,10,50,20,50,20,10)
df <- data.frame(fertilizer_levels)
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
describe(df)
```

```
## df
##
##  1  Variables      14  Observations
## ---------------------------------------------------------------------------
## fertilizer_levels
##        n  missing distinct     Info     Mean      Gmd
##       14        0        3     0.87    22.14    16.15
##
## Value         10    20    50
## Frequency      6     5     3
## Proportion 0.429 0.357 0.214
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------
```

```
summary(df)
```

```
##  fertilizer_levels
##  Min.   :10.00
##  1st Qu.:10.00
##  Median :20.00
##  Mean   :22.14
##  3rd Qu.:20.00
##  Max.   :50.00
```

```
#2.
df$fertilizer_levels <- factor(df$fertilizer_levels, ordered = TRUE,
                               levels = c(10, 20, 50))

df$fertilizer_levels
```

```
##  [1] 10 10 10 20 20 50 10 20 10 50 20 50 20 10
```

```
## Levels: 10 < 20 < 50
```

```
# 3.
exercise <- c("l", "n", "n", "i", "l", "l", "n", "n", "i", "l")
exercise_levels <- factor(exercise,
                          levels = c("n", "l", "i"),
                          ordered = TRUE)
str(exercise_levels)
```

```
##  Ord.factor w/ 3 levels "n"<"l"<"i": 2 1 1 3 2 2 1 1 3 2
```

```
# 4.
state <- c("tas", "sa", "qld", "nsw", "nsw", "nt", "wa", "wa", "qld",
           "vic", "nsw", "vic", "qld", "qld", "sa", "tas", "sa", "nt",
           "wa", "vic", "qld", "nsw", "nsw", "wa", "sa", "act", "nsw",
           "vic", "vic", "act")
state_factor <- factor(state)
str(state_factor)
```

```
##  Factor w/ 8 levels "act","nsw","nt",..: 6 5 4 2 2 3 8 8 4 7 ...
```

```
#State_factor is now a factor variable with 7 levels: act, nsw, nt, qld, sa, tas, vic
#Applying the factor() function organized the state codes
#into a factor variable with known levels for further analysis.
```

```
# 5.
state_factor <- factor(state)
incomes <- c(60, 49, 40, 61, 64, 60, 59, 54,
             62, 69, 70, 42, 56, 61, 61, 61, 58, 51, 48,
             65, 49, 49, 41, 48, 52, 46, 59, 46, 58, 43)
incmeans <- tapply(incomes, state_factor, mean)
incmeans
```

```
##      act      nsw       nt      qld       sa      tas      vic       wa
## 44.50000 57.33333 55.50000 53.60000 55.00000 60.50000 56.00000 52.25000
```

```
#We can interpret that on average, tax accountants from Tasmania (TAS) earned
#the highest mean income of $60,000,
#while those from the Australian Capital Territory (ACT)
#earned the lowest mean of $45,500.
```

```
#6.
stdError <- function(x) sqrt(var(x)/length(x))
incster <- tapply(incomes, state_factor, stdError)
incster
```

```
##      act      nsw       nt      qld       sa      tas      vic       wa
## 1.500000 4.310195 4.500000 4.106093 2.738613 0.500000 5.244044 2.657536
```

```
#The standard error for Tasmania is 0, since there was no variation in the single observation.
#The highest standard error is for the Northern Territory (NT), at $7.0711,
#indicating higher variation around the mean for that state.
#The lowest non-zero standard error is for South Australia (SA), at $5.6568,
#showing least variation around the mean income for that state grouping.
```

```
#7.
data("Titanic")
titanic_df <- as.data.frame(Titanic)
survived_df <- subset(titanic_df, Survived == "Yes")
```

```r
not_survived_df <- subset(titanic_df, Survived == "No")
head(survived_df)
```

```
##    Class    Sex   Age Survived Freq
## 17   1st   Male Child      Yes    5
## 18   2nd   Male Child      Yes   11
## 19   3rd   Male Child      Yes   13
## 20  Crew   Male Child      Yes    0
## 21   1st Female Child      Yes    1
## 22   2nd Female Child      Yes   13
```

```r
head(not_survived_df)
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

```r
breast_cancer_data <- read.csv("/cloud/project/Worksheet#6/breastcancer_wisconsin.csv")
#8.
# a.)
length(breast_cancer_data)
```

```
## [1] 11
```

```r
head(breast_cancer_data)
```

```
##        id clump_thickness size_uniformity shape_uniformity marginal_adhesion
## 1 1000025               5               1                1                 1
## 2 1002945               5               4                4                 5
## 3 1015425               3               1                1                 1
## 4 1016277               6               8                8                 1
## 5 1017023               4               1                1                 3
## 6 1017122               8              10               10                 8
##   epithelial_size bare_nucleoli bland_chromatin normal_nucleoli mitoses class
## 1               2             1               3               1       1     2
## 2               7            10               3               2       1     2
## 3               2             2               3               1       1     2
## 4               3             4               3               7       1     2
## 5               2             1               3               1       1     2
## 6               7            10               9               7       1     4
```

```r
summary(breast_cancer_data)
```

```
##        id           clump_thickness  size_uniformity  shape_uniformity
##  Min.   :   61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1171710   Median : 4.000   Median : 1.000   Median : 1.000
##  Mean   : 1071704   Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
##  3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##  Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
##  marginal_adhesion epithelial_size  bare_nucleoli      bland_chromatin
##  Min.   : 1.000    Min.   : 1.000   Length:699         Min.   : 1.000
##  1st Qu.: 1.000    1st Qu.: 2.000   Class :character   1st Qu.: 2.000
```

```
##  Median : 1.000    Median : 2.000   Mode  :character   Median : 3.000
##  Mean   : 2.807    Mean   : 3.216                     Mean   : 3.438
##  3rd Qu.: 4.000    3rd Qu.: 4.000                     3rd Qu.: 5.000
##  Max.   :10.000    Max.   :10.000                     Max.   :10.000
##  normal_nucleoli     mitoses           class
##  Min.   : 1.000   Min.   : 1.000   Min.   :2.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
##  Median : 1.000   Median : 1.000   Median :2.00
##  Mean   : 2.867   Mean   : 1.589   Mean   :2.69
##  3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
##  Max.   :10.000   Max.   :10.000   Max.   :4.00
```

```
str(breast_cancer_data)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ id               : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 10
##  $ clump_thickness  : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ size_uniformity  : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ shape_uniformity : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ marginal_adhesion: int  1 5 1 1 3 8 1 1 1 1 ...
##  $ epithelial_size  : int  2 7 2 3 2 7 2 2 2 2 ...
##  $ bare_nucleoli    : chr  "1" "10" "2" "4" ...
##  $ bland_chromatin  : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ normal_nucleoli  : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ mitoses          : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ class            : int  2 2 2 2 2 4 2 2 2 2 ...
```

```
#d.)
#d.1
install.packages("psych")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:Hmisc':
##
##     describe
```

```
se_mean_clump_thickness <- sd(breast_cancer_data$clump_thickness) / sqrt(length(breast_cancer_data$clump
```

```
#d.2
mean_marginal_adhesion <- mean(breast_cancer_data$marginal_adhesion)
sd_marginal_adhesion <- sd(breast_cancer_data$marginal_adhesion)
cv_marginal_adhesion <- (sd_marginal_adhesion / mean_marginal_adhesion) * 100
```

```
#d.3
install.packages("naniar")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(naniar)
```

```r
null_values_bare_nucleoli <- sum(is.na(breast_cancer_data$bare_nucleoli))
```

```r
#d.4
mean_bland_chromatin <- mean(breast_cancer_data$bland_chromatin)
sd_bland_chromatin <- sd(breast_cancer_data$bland_chromatin)
```

```r
#d.5
ci_mean_uniformity_cell_shape <- t.test(breast_cancer_data$shape_uniformity)$conf.int
```

```r
#9.
install.packages("AppliedPredictiveModeling")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
library(AppliedPredictiveModeling)
data(abalone)
head(abalone)
```

```
##   Type LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
## 1    M        0.455    0.365  0.095      0.5140        0.2245        0.1010
## 2    M        0.350    0.265  0.090      0.2255        0.0995        0.0485
## 3    F        0.530    0.420  0.135      0.6770        0.2565        0.1415
## 4    M        0.440    0.365  0.125      0.5160        0.2155        0.1140
## 5    I        0.330    0.255  0.080      0.2050        0.0895        0.0395
## 6    I        0.425    0.300  0.095      0.3515        0.1410        0.0775
##   ShellWeight Rings
## 1       0.150    15
## 2       0.070     7
## 3       0.210     9
## 4       0.155    10
## 5       0.055     7
## 6       0.120     8
```

```r
summary(abalone)
```

```
##  Type      LongestShell      Diameter          Height        WholeWeight
##  F:1307   Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020
##  I:1342   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415
##  M:1528   Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995
##           Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287
##           3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530
##           Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255
##  ShuckedWeight    VisceraWeight     ShellWeight         Rings
##  Min.   :0.0010   Min.   :0.0005   Min.   :0.0015   Min.   : 1.000
##  1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.: 8.000
##  Median :0.3360   Median :0.1710   Median :0.2340   Median : 9.000
##  Mean   :0.3594   Mean   :0.1806   Mean   :0.2388   Mean   : 9.934
##  3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:11.000
##  Max.   :1.4880   Max.   :0.7600   Max.   :1.0050   Max.   :29.000
```