**Очищення даних**

EDA показало наявність викидів в оригінальному датасеті. Особливо виділяється розмах значень tot\_chole:

A graph with numbers and a number

Description automatically generated

Діаграма частот значень tot\_chole оригінального датасету

A graph with numbers and a line

Description automatically generated

Box plot tot\_chole оригінального датасету

Як виявилось, кількість спостережень, для яких tot\_chole > 500, є малою – всього лиш 61 спостереження:

big\_tot\_chole\_count <- nrow(data %>%

+ filter(tot\_chole > 500))

> big\_tot\_chole\_count

[1] 61

Було вирішено видалити ці 61 спостереження.

Графіки tot\_chole після видалення:

A blue graph on a white background

Description automatically generated

Діаграма частот значень tot\_chole чистого датасету

A graph with a bar graph

Description automatically generated

Box plot tot\_chole чистого датасету

Таким чином, розподіл вирівнявся у порівнянні з початковим (хоча все ще наявне відхилення вліво). При цьому видалення відповідних спостережень не має значущого впливу на характер даних:

A grid with dots and lines

Description automatically generated

Матриця кореляцій оригінального датасету

A grid with dots and lines

Description automatically generated

Матриця кореляцій очищеного датасету

**Довірчі інтервали для середніх значень**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **name** | **mean** | **standard\_deviation** | **confidence\_interval\_norm\_a** | **confidence\_interval\_norm\_b** |
| age | 47.61452822803556 | 14.18143239760604 | 47.586611202626315 | 47.642445253444805 |
| height | 162.24050322611885 | 9.282932834572376 | 162.22222919959253 | 162.25877725264516 |
| weight | 63.283771351096156 | 12.51408744869878 | 63.25913659718105 | 63.308406105011265 |
| waistline | 81.23305652063384 | 11.850382483398542 | 81.20972831092816 | 81.25638473033952 |
| sight\_left | 0.9808368741954878 | 0.6059612464003457 | 0.9796440020625128 | 0.9820297463284628 |
| sight\_right | 0.9784308028778836 | 0.6047867022605942 | 0.9772402429075784 | 0.9796213628481889 |
| hear\_left | 1.0314965237005742 | 0.17465544213706122 | 1.0311527036773709 | 1.0318403437237775 |
| hear\_right | 1.030475625552314 | 0.17189209291353677 | 1.0301372453527 | 1.030814005751928 |
| SBP | 122.43204671920459 | 14.542805706063477 | 122.40341830812206 | 122.46067513028713 |
| DBP | 76.05226958167387 | 9.889232704609686 | 76.03280201641303 | 76.07173714693471 |
| BLDS | 100.42162589126829 | 24.170510958360314 | 100.3740447478647 | 100.46920703467188 |
| tot\_chole | 195.5132747023053 | 38.054265230965484 | 195.4383625312638 | 195.58818687334679 |
| HDL\_chole | 56.92946118367693 | 15.220790071277209 | 56.89949811850445 | 56.95942424884941 |
| LDL\_chole | 113.00705045173734 | 34.95017868182186 | 112.93824886675199 | 113.0758520367227 |
| triglyceride | 132.04616537742967 | 100.4162789672993 | 131.8484897313571 | 132.24384102350226 |
| hemoglobin | 14.229770479499317 | 1.5848706873330576 | 14.22665056368206 | 14.232890395316574 |
| urine\_protein | 1.0941839069328265 | 0.437596771436844 | 1.0933224706656928 | 1.0950453431999603 |
| serum\_creatinine | 0.8604647104159857 | 0.48053998001597714 | 0.8595187377909266 | 0.8614106830410448 |
| SGOT\_AST | 25.986995654121323 | 23.483270964269703 | 25.94076738508467 | 26.033223923157976 |
| SGOT\_ALT | 25.75357314351891 | 26.306113918022902 | 25.70178793381105 | 25.80535835322677 |
| gamma\_GTP | 37.12669325844057 | 50.36653709788572 | 37.02754361991925 | 37.225842896961886 |

**Довірчі інтервали для медіан (bootstrap)**

# boot median confidence interval fail

> calc\_median <- function(x, i) {

+ return(median(x[i]))

+ }

> boot\_medians\_by\_column <- bind\_rows(lapply(names(data\_all\_numeric), function(column\_name) {

+ boot\_result <- boot(data\_all\_numeric[[column\_name]], statistic = calc\_median, R = 500)

+ boot\_ci\_result <- boot.ci(boot\_result, type = "basic", conf = 0.95)

+ tibble(name = column\_name,

+ median\_t0 = boot\_result$t0,

+ median\_ci = boot\_ci\_result$t0)

+ })

+ )

[1] "All values of t are equal to 45 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 160 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 60 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 81 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 120 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 76 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 96 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 193 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 55 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 111 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 14.3 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 0.8 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 23 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 20 \n Cannot calculate confidence intervals"

[1] "All values of t are equal to 23 \n Cannot calculate confidence intervals"

boot\_samples <- boot(data\_all\_numeric$age, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 45 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$height, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 160 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$weight, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 60 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$waistline, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 81 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$sight\_left, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$sight\_right, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$hear\_left, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$hear\_right, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$SBP, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 120 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$DBP, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 76 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$BLDS, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 96 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$tot\_chole, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 193 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$HDL\_chole, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 55 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$LDL\_chole, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 111 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$triglyceride, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

> boot\_ci

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 100 bootstrap replicates

CALL :

boot.ci(boot.out = boot\_samples, conf = 0.95, type = "basic")

Intervals :

Level Basic

95% (105, 106 )

Calculations and Intervals on Original Scale

Some basic intervals may be unstable

boot\_samples <- boot(data\_all\_numeric$hemoglobin, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 14.3 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$urine\_protein, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 1 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$serum\_creatinine, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 0.8 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$SGOT\_AST, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 23 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$SGOT\_ALT, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 20 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

boot\_samples <- boot(data\_all\_numeric$gamma\_GTP, statistic = calc\_median, R = 100)

> boot\_ci <- boot.ci(boot\_samples, type = "basic", conf = 0.95)

[1] "All values of t are equal to 23 \n Cannot calculate confidence intervals"

> boot\_ci

NULL

За допомогою bootstrap вдалось обчислити довірчий інтервал медіани лише для однієї характеристики – triglyceride. Це пов’язано з наступними факторами:

1. Велика кількість спостережень – 991284 спостереження
2. Відносна одноманітність даних – значення характеристик в датасеті округлені до 1 знаку після коми.

Враховуючи ці фактори, доходимо до висновку, що використання bootstrap для визначення довірчих інтервалів медіан є недоцільним.

**Довірчі інтервали для медіан (no bootstrap)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| name | median | standard\_deviation | confidence\_interval\_norm\_a | confidence\_interval\_norm\_b |
| age | 45 | 14.18143239760604 | 44.972082974590755 | 45.027917025409245 |
| height | 160 | 9.282932834572376 | 159.9817259734737 | 160.0182740265263 |
| weight | 60 | 12.51408744869878 | 59.97536524608489 | 60.02463475391511 |
| waistline | 81 | 11.850382483398542 | 80.97667179029432 | 81.02332820970568 |
| sight\_left | 1 | 0.6059612464003457 | 0.998807127867025 | 1.001192872132975 |
| sight\_right | 1 | 0.6047867022605942 | 0.9988094400296947 | 1.0011905599703053 |
| hear\_left | 1 | 0.17465544213706122 | 0.9996561799767966 | 1.0003438200232033 |
| hear\_right | 1 | 0.17189209291353677 | 0.9996616198003859 | 1.000338380199614 |
| SBP | 120 | 14.542805706063477 | 119.97137158891746 | 120.02862841108254 |
| DBP | 76 | 9.889232704609686 | 75.98053243473916 | 76.01946756526084 |
| BLDS | 96 | 24.170510958360314 | 95.95241885659641 | 96.04758114340359 |
| tot\_chole | 193 | 38.054265230965484 | 192.9250878289585 | 193.0749121710415 |
| HDL\_chole | 55 | 15.220790071277209 | 54.97003693482752 | 55.02996306517248 |
| LDL\_chole | 111 | 34.95017868182186 | 110.93119841501465 | 111.06880158498535 |
| triglyceride | 106 | 100.4162789672993 | 105.8023243539274 | 106.1976756460726 |
| hemoglobin | 14.3 | 1.5848706873330576 | 14.296880084182744 | 14.303119915817257 |
| urine\_protein | 1 | 0.437596771436844 | 0.9991385637328661 | 1.0008614362671338 |
| serum\_creatinine | 0.8 | 0.48053998001597714 | 0.7990540273749409 | 0.8009459726250592 |
| SGOT\_AST | 23 | 23.483270964269703 | 22.953771730963346 | 23.046228269036654 |
| SGOT\_ALT | 20 | 26.306113918022902 | 19.94821479029214 | 20.05178520970786 |
| gamma\_GTP | 23 | 50.36653709788572 | 22.90085036147868 | 23.09914963852132 |

Варто зауважити, що довірчий інтервал значення медіани для triglyceride, обчислений за допомогою bootstrap, співпадає із інтервалом, обчисленим без bootstrap.

**Доведення гіпотез**

Для перевірки всіх гіпотез вважатимемо достатньою достовірність 95%.

Обчислення статистики критерія для перевірки гіпотези:

# hypos

calc\_statistics\_criteria <- function(x, y) {

mean\_x <- mean(x)

mean\_y <- mean(y)

mean\_d <- mean\_x - mean\_y

sd\_x <- sd(x)

sd\_y <- sd(y)

nx <- length(x)

ny <- length(y)

se = sqrt(((sd\_x \* sd\_x)/nx) + ((sd\_y \* sd\_y)/ny))

sc = mean\_d / se

ci\_low = mean\_d - se \* 1.64

return(c(statistics\_criteria=sc, standard\_error=se, confidence\_interval\_low=ci\_low))

}

Всі альтернативні гіпотези є гіпотезами типу >. Відкидання нульової гіпотези проводиться незалежно згідно 2 форм тесту Волда:

1. (див. слайд 27 лекції 5)
2. (див. слайд 34 лекції 5)

*Гіпотеза 1: рівень гемоглобіну збільшується з вагою*

H0: hemoglobin(obese) = hemoglobin(not\_obese)

Ha: hemoglobin(obese) > hemoglobin(not\_obese)

> # hypo: weight (body mass index) influence on hemoglobin

> # h0: hemoglobin(obese) = hemoglobin(not\_obese)

> # ha: hemoglobin(obese) > hemoglobin(not\_obese)

> data <- data %>% mutate(mbi = data$weight / ((0.01 \* data$height)^2))

> x <- (data %>% filter(mbi > 25))$hemoglobin

> y <- (data %>% filter(mbi <= 25))$hemoglobin

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

2.180586e+02 3.276411e-03 7.090761e-01

> c(mean\_hemoglobin\_not\_obese = mean(y), mean\_hemoglobin\_obese = mean(x))

mean\_hemoglobin\_not\_obese mean\_hemoglobin\_obese

13.98845 14.70290

Статистика критерію ~= 218 > z(1-a) = 1.64

t0 = 0 < 0.709

Отже різниця:

“mean hemoglobin (no obese): 13.98845”

“mean hemoglobin (obese): 14.70290”

Є статистично значущою, **гіпотезу підтверджено.**

*Гіпотеза 2.1: Вживання алкоголю впливає на SBP*

H0: SBP(Alc) = SBP(NoAlc)

Ha: SBP(Alc) > SBP(NoAlc)

> # hypo: Alcohol influences on SBP

> # h0: SBP(Alc) = SBP(NoAlc)

> # ha: SBP(Alc) > SBP(NoAlc)

> x <- (data %>% filter(DRK\_YN == "Y"))$SBP

> y <- (data %>% filter(DRK\_YN == "N"))$SBP

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

33.00968331 0.02919659 0.91588787

Статистика критерію ~= 33 > 1.64

t0 = 0 < 0.915

Отже різниця:

"mean SBP (no drinker): 121.950355053334"

"mean SBP (drinker): 122.914125338334"

є статистично значущою, **гіпотеза підтверджена.**

*Гіпотеза 2.2: Вживання алкоголю впливає на DBP*

H0: DBP(Alc) = DBP(NoAlc)

Ha: DBP(Alc) > DBP(NoAlc)

> # hypo: Alcohol influence on DBP

> # h0: DBP(Alc) = DBP(NoAlc)

> # ha: DBP(Alc) > DBP(NoAlc)

> x <- (data %>% filter(DRK\_YN == "Y"))$DBP

> y <- (data %>% filter(DRK\_YN == "N"))$DBP

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

100.93139320 0.01976421 1.96241556

Статистика критерію 100.93 > 1.64

t0 = 0 < 1.96

Отже різниця:

"mean DBP (no drinker): 75.0552556162157"

"mean DBP (drinker): 77.0500844698583"

є статистично значущою, **гіпотеза підтверджена**

*Гіпотеза 3: вживання алкоголю впливає на загальний рівень холестерину.*

H0: tot\_chole(Alc) = tot\_chole(NoAlc)

Ha: tot\_chole(Alc) > tot\_chole(NoAlc)

> # hypo: Alcohol influence on tot\_chole

> # h0: tot\_chole(Alc) = tot\_chole(NoAlc)

> # ha: tot\_chole(Alc) > tot\_chole(NoAlc)

> x <- (data %>% filter(DRK\_YN == "Y"))$tot\_chole

> y <- (data %>% filter(DRK\_YN == "N"))$tot\_chole

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

19.52149038 0.07642627 1.36661560

Статистика критерію 19.52 > 1.64

t0 = 0 < 1.3666

Отже різниця:

"mean tot\_chole (no drinker): 194.767596870771"

"mean tot\_chole (drinker): 196.259551552853"

Є статистично значущою, **гіпотезу підтверджено.**

*Гіпотеза 4: вживання алкоголю впливає на рівень гемоглобіну.*

H0: hemoglobin(Alc) = hemoglobin(NoAlc)

Ha: hemoglobin(Alc) > hemoglobin(NoAlc)

> # hypo: Alcohol influence on hemoglobin

> # h0: hemoglobin(Alc) = hemoglobin(NoAlc)

> # ha: hemoglobin(Alc) > hemoglobin(NoAlc)

> x <- (data %>% filter(DRK\_YN == "Y"))$hemoglobin

> y <- (data %>% filter(DRK\_YN == "N"))$hemoglobin

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

3.123530e+02 3.037635e-03 9.438328e-01

Статистика критерію 312 > 1.64

t0 = 0 < 0.9438.

Отже різниця:

"mean hemoglobin (no drinker): 13.7555536956403"

"mean hemoglobin (drinker): 14.7043682118831"

Є статистично значущою, **гіпотезу підтверджено.**

*Гіпотеза 5.1: Значення SBP людей, що кинули палити (2), є вищим за значення груп 1 і 3.*

H0: SBP(2) = SBP(1|3)

Ha: SBP(2) > SBP(1|3)

> # hypo: Smoking influence on SBP

> # h0: SBP(2) = SBP(1|3)

> # ha: SBP(2) < SBP(1|3)

> x <- (data %>% filter(SMK\_stat\_type\_cd == 2))$SBP

> y <- (data %>% filter(SMK\_stat\_type\_cd != 2))$SBP

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

96.32792468 0.03671778 3.47672994

Статистика критерія 96.32 > 1.64

t0 = 0 < 3.4767

Отже різниця:

"mean sbp (1s): 121.177524870646"

"mean sbp (2s): 125.344796245591"

"mean sbp (3s): 123.582846404832"

Є статистично значущою, **гіпотезу підтверджено.**

*Гіпотеза 5.2: Значення DBP людей, що кинули палити (2), є вищим за значення груп 1 і 3.*

H0: DBP(2) = DBP(1|3)

Ha: DBP(2) > DBP(1|3)

> # hypo: Smoking influence on DBP

> # h0: DBP(2) = DBP(1|3)

> # ha: DBP(2) < DBP(1|3)

> x <- (data %>% filter(SMK\_stat\_type\_cd == 2))$DBP

> y <- (data %>% filter(SMK\_stat\_type\_cd != 2))$DBP

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

95.21126989 0.02559697 2.39514091

Статистика критерія 95.21 > 1.64

t0 = 0 < 2.395

Отже різниця:

"mean dbp (1s): 74.9162075439438"

"mean dbp (2s): 78.0592885601431"

"mean dbp (3s): 77.6101595878949"

Є статистично значущою, **гіпотезу підтверджено.**

*Гіпотеза 6.1: Куріння збільшує рівень гемоглобіну.*

H0: hemoglobin(3) = hemoglobin(1)

Ha: hemoglobin(3) > hemoglobin(1)

> # hypo: Smoking influence on hemoglobin 1

> # h0: hemoglobin(3) = hemoglobin(1)

> # ha: hemoglobin(3) > hemoglobin(1)

> x <- (data %>% filter(SMK\_stat\_type\_cd == 3))$hemoglobin

> y <- (data %>% filter(SMK\_stat\_type\_cd == 1))$hemoglobin

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

4.974940e+02 3.320384e-03 1.646426e+00

Статистика критерія 497.49 > 1.64

t0 = 0 < 1.6464

Отже різниця:

"mean hemoglobin (1s): 13.6387173183764"

"mean hemoglobin (3s): 15.2905752456457"

є статистично значущою, **гіпотезу підтверджено.**

*Гіпотеза 6.2 Після припинення паління з часом гемоглобін зменшується*

H0: hemoglobin(3) = hemoglobin(2)

Ha: hemoglobin(3) > hemoglobin(2)

> # hypo: Smoking influence on hemoglobin 2

> # h0: hemoglobin(3) = hemoglobin(2)

> # ha: hemoglobin(3) > hemoglobin(2)

> x <- (data %>% filter(SMK\_stat\_type\_cd == 3))$hemoglobin

> y <- (data %>% filter(SMK\_stat\_type\_cd == 2))$hemoglobin

> calc\_statistics\_criteria(x, y)

statistics\_criteria standard\_error confidence\_interval\_low

79.456856716 0.004060616 0.315984362

Статистика критерія 79.45 > 1.64

t0 = 0 < 0.3159

Отже різниця:

"mean hemoglobin (3s): 15.2905752456457"

“mean hemoglobin (2s): 14.9679314740398”

є статично значущою, **гіпотезу підтверджено.**