|  |
| --- |
| library("FinCal")  > library("psych")  > ?Orange  > data\_frame<-Orange  > View(data\_frame) #вивід усіх даних  > #дані за різні роки для 5-х дерев; номер дерева - факторна змінна, причому  > #номер рівня впорядкований за зростанням максимального діаметру  > #так: 1-ше дерево - фактор 2-го рівня, 2-ге - 4-го, 3-тє - 1-го, 4-те - 5-го, 5-те - 3-го  > #Характеристики положення центра  > #Математичне сподівання  > #порахуємо мат.сподівання для діаметру стовбуру дерев для кожного віку  > #це можна зробити за допомогою однієї функції, щоб не дублювати рядки  > #для кожного віку окремо  > mean\_all<-aggregate(x=data\_frame$circumference, by=list(data\_frame$Tree),FUN=mean)  > colnames(mean\_all)<-c("Tree (factor level)", "Mean")  > mean\_all  Tree (factor level) Mean  1 3 94.00000  2 1 99.57143  3 5 111.14286  4 2 135.28571  5 4 139.28571  > #Середнє геометричне (для випадкових величин, які з йм 1 додатні); наші дані невід'ємі  > #вказує на темпи зростання  > geomean\_all<-aggregate(x=data\_frame$circumference, by=list(data\_frame$Tree),geometric.mean)  > colnames(geomean\_all)<-c("Tree (factor level)", "Geometric mean")  > geomean\_all  Tree (factor level) Geometric mean  1 3 83.42813  2 1 88.82051  3 5 94.28934  4 2 116.06545  5 4 117.11644  > #Cереднє гармонічне (для в.в. з позитивними значеннями)  > harm\_mean\_1004<-harmonic.mean(data\_frame$circumference[data\_frame$age==1004])  > harm\_mean\_1004  [1] 130.3672  > #Мода, вбудованої функції немає  > mode <- function(data\_frame) {  + uniq <- unique(data\_frame) #прибираємо значення, що дублюються  + uniq[which.max(tabulate(match(data\_frame,uniq)))]  + }  > #тут match повертає вектор позицій першого входження ел-ів x у множині значень х  > #tabulate рахує кількість повторів отриманих індексів, тоді максимальне з них  > #буде індексом моди  > mode\_<-mode(data\_frame$circumference)  > mode\_  [1] 30  > #бо на початку більшість дерев мали діаметр 30  > #Медіана (центральний ел-т статистики)  > median\_<-median(data\_frame$circumference)  > median\_  [1] 115  > #Характеристики розсіювання значень змінної  > #Дисперсія  > variation\_all<-aggregate(x=data\_frame$circumference, by=list(data\_frame$age),var)  > colnames(variation\_all)<-c("Age", "Variation")  > variation\_all  Age Variation  1 118 2.0  2 484 66.7  3 664 297.2  4 1004 672.7  5 1231 854.3  6 1372 1078.3  7 1582 1107.7  > #Стандартне відхилення  > #цікаво подивитись відхилення для кожного року окремо і подивитись чи змінюється воно  > sd\_all<-aggregate(x=data\_frame$circumference,by=list(data\_frame$age),sd)  > colnames(sd\_all)<-c("Age", "Standart deviation")  > sd\_all  Age Standart deviation  1 118 1.414214  2 484 8.167007  3 664 17.239490  4 1004 25.936461  5 1231 29.228411  6 1372 32.837479  7 1582 33.282127  > #бачимо, що зі збільшенням віку діаметр дерев почав суттєво варіюватись,  > #тобто вони мали різний темп росту  > #Коефіцієнт варіації (оскільки мат.спод.<>0)  > #має сенс рахувати для певного віку  > #як і взагалі всі характеристики  > #характеризує ступінь різноманітності(виділяють низький, середній та високий)  > #порахуємо для віку 1582  > coef\_var\_1582<-coefficient.variation(sd=sd\_1582,avg=mean\_1582)  Error in coefficient.variation(sd = sd\_1582, avg = mean\_1582) :  object 'sd\_1582' not found  > coef\_var\_1582  Error: object 'coef\_var\_1582' not found  > #майже 19%, що каже про середню змінюваність  > #Ймовірнісне відхилення (половина інтерквартильної широти)  > prob\_dev <- function(data) {  + IQR(data)/2  + }  > prob\_dev\_all<-aggregate(x=data\_frame$circumference,by=list(data\_frame$age),prob\_dev)  > colnames(prob\_dev\_all)<-c("Age", "IQR of circumference")  > prob\_dev\_all  Age IQR of circumference  1 118 1.0  2 484 5.5  3 664 15.0  4 1004 20.5  5 1231 26.0  6 1372 30.5  7 1582 29.0  > #Розмах вибірки  > range<-range(data\_frame$circumference)  > range  [1] 30 214  > #на мою думку також варто подивитись розмах по кожному з років  > #щоб наочно побачити що він зростав, про що нам сказало середньоквадратичне відхилення  > range\_all<-aggregate(x=data\_frame$circumference,by=list(data\_frame$age),range)  > colnames(range\_all)<-c("Age", "Range")  > range\_all  Age Range.1 Range.2  1 118 30 33  2 484 49 69  3 664 75 112  4 1004 108 167  5 1231 115 179  6 1372 139 209  7 1582 140 214  > #бачимо, що різниця достатньо суттєва, хоча це все апельсинові дерева  > #Інтервал концентрації розподілу  > #якби розподіл був нормальний, то цей інтервал містив би 99,7% усіх спостережень  > conc\_interval<- function(data){  + left<-(mean(data)-3\*sd(data))  + right<-(mean(data)+3\*sd(data))  + return (paste(left,right, sep=" "))  + }  > conc\_int<-aggregate(x=data\_frame$circumference,by=list(data\_frame$age),conc\_interval)  > colnames(conc\_int)<-c("Age", "Interval")  > conc\_int  Age Interval  1 118 26.7573593128807 35.2426406871193  2 484 33.2989796130855 82.3010203869145  3 664 41.4815313451761 144.918468654824  4 1004 56.390617532331 212.009382467669  5 1231 57.9147674918975 233.285232508103  6 1372 74.8875642367929 271.912435763207  7 1582 75.953617992438 275.646382007562  > #Квантилі  > quantile(data\_frame$circumference)  0% 25% 50% 75% 100%  30.0 65.5 115.0 161.5 214.0  > #з застосуванням функції describeBy, вона рахує основні статистики по групованим даним  > #порахуємо основні оцінки для кожного віку  > describeBy(x=data\_frame$circumference, group =data\_frame$age)  Descriptive statistics by group  group: 118  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 5 31 1.41 30 31 0 30 33 3 0.42 -2 0.63  ---------------------------------------------------------------  group: 484  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 5 57.8 8.17 58 57.8 10.38 49 69 20 0.18 -1.91 3.65  ---------------------------------------------------------------  group: 664  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 5 93.2 17.24 87 93.2 17.79 75 112 37 0.16 -2.19 7.71  ---------------------------------------------------------------  group: 1004  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 5 134.2 25.94 125 134.2 25.2 108 167 59 0.23 -2.12 11.6  ---------------------------------------------------------------  group: 1231  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 5 145.6 29.23 142 145.6 40.03 115 179 64 0.08 -2.17 13.07  ---------------------------------------------------------------  group: 1372  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 5 173.4 32.84 174 173.4 47.44 139 209 70 0 -2.18 14.69  ---------------------------------------------------------------  group: 1582  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 5 175.8 33.28 177 175.8 47.44 140 214 74 0 -2.15 14.88  > #Розвідувальний аналіз  > #Гістограма  > library(ggplot2)  > ggplot(data\_frame, aes(x=circumference))+  + geom\_histogram(fill="blue",col="black",binwidth = 20, alpha=0.5)+  + labs(title="Hist for Circumference")  > #ящик з вусами; для кожного віку окремо  > boxplot(circumference~age, data=data\_frame, main="Boxplot")  #ймовірнісний графік  plot(data\_frame$circumference,pnorm(data\_frame$circumference, mean(data\_frame$circumference),  sd(data\_frame$circumference)) , main="P-P-plot") |
|  |
| |  | | --- | |  | |

> #перевірка нормальності

> #нульова гіпотеза: дані мають нормальний розподіл

> #при p>0.05 нульова гіпотеза не відхиляється

> shapiro.test(data\_frame$circumference)

Shapiro-Wilk normality test

data: data\_frame$circumference

W = 0.94591, p-value = 0.08483

> #p=0.08, отже, гіпотеза про нормальність не відхиляється