Lab 04

Jyosna Philip

2023-12-15

## AIM

Test for proportion of single population (large samples)

## DATASET DESCRIPTION

This dataset contains information on bone tumors. The data was collected from patients at the Memorial Sloan Kettering Cancer Center (MSKCC).

Bone\_Tumor <- read.csv("C:/Users/jyosn/Downloads/archive (1)/Bone Tumor Dataset.csv")

## ANALYSIS

head(Bone\_Tumor)

## Patient.ID Sex Age Grade Histological.type  
## 1 STS\_001 Female 63 High pleiomorphic leiomyosarcoma  
## 2 STS\_002 Female 54 Intermediate pleiomorphic leiomyosarcoma  
## 3 STS\_003 Male 22 Intermediate synovial sarcoma  
## 4 STS\_004 Female 54 Intermediate synovial sarcoma  
## 5 STS\_005 Male 63 High pleiomorphic spindle cell undifferentiated  
## 6 STS\_006 Female 58 High poorly differentiated synovial sarcoma  
## MSKCC.type Site.of.primary.STS Status..NED..AWD..D.  
## 1 Leiomyosarcoma parascapusular NED  
## 2 MFH left thigh D  
## 3 MFH right buttock D  
## 4 Leiomyosarcoma right thigh NED  
## 5 MFH right parascapusular AWD  
## 6 MFH right thigh D  
## Treatment  
## 1 Radiotherapy + Surgery  
## 2 Surgery + Chemotherapy  
## 3 Radiotherapy + Surgery  
## 4 Radiotherapy + Surgery  
## 5 Radiotherapy + Surgery + Chemotherapy  
## 6 Surgery + Chemotherapy

tail(Bone\_Tumor)

## Patient.ID Sex Age Grade Histological.type  
## 495 STS\_495 Male 54 High pleomorphic sarcoma  
## 496 STS\_496 Male 31 Intermediate epithelioid sarcoma  
## 497 STS\_497 Male 41 High poorly differentiated synovial sarcoma  
## 498 STS\_498 Female 60 Intermediate pleomorphic sarcoma  
## 499 STS\_499 Male 43 High synovial sarcoma  
## 500 STS\_500 Female 73 Intermediate pleiomorphic leiomyosarcoma  
## MSKCC.type Site.of.primary.STS Status..NED..AWD..D.  
## 495 MFH right thigh AWD  
## 496 MFH left thigh D  
## 497 MFH right parascapusular AWD  
## 498 Synovial sarcoma right thigh NED  
## 499 MFH right thigh D  
## 500 Leiomyosarcoma left biceps NED  
## Treatment  
## 495 Radiotherapy + Surgery + Chemotherapy  
## 496 Surgery + Chemotherapy  
## 497 Radiotherapy + Surgery + Chemotherapy  
## 498 Radiotherapy + Surgery  
## 499 Radiotherapy + Surgery + Chemotherapy  
## 500 Radiotherapy + Surgery

dim(Bone\_Tumor)

## [1] 500 9

This dataset has 9 attributes and 500 rows.

summary(Bone\_Tumor)

## Patient.ID Sex Age Grade   
## Length:500 Length:500 Min. :17.0 Length:500   
## Class :character Class :character 1st Qu.:44.0 Class :character   
## Mode :character Mode :character Median :56.0 Mode :character   
## Mean :55.4   
## 3rd Qu.:68.0   
## Max. :83.0   
## Histological.type MSKCC.type Site.of.primary.STS Status..NED..AWD..D.  
## Length:500 Length:500 Length:500 Length:500   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Treatment   
## Length:500   
## Class :character   
## Mode :character   
##   
##   
##

library(colorDF)

## colorDF: for best results, use terminals which support 256 colors.

c=colorDF(Bone\_Tumor, theme="bw")  
??colorDF

## starting httpd help server ...

## done

All columns except age is categorical, so nothing can be said about this dataset from initial analysis.

## Selecting a sample

grade\_s=sample(Bone\_Tumor$Grade,150,replace=FALSE)  
class(grade\_s)

## [1] "character"

count works when the first variable is a dataframe.

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

#Another way of taking a sample  
grade\_s=sample(1:nrow(Bone\_Tumor),150,replace=FALSE)  
grade\_s

## [1] 485 357 58 207 164 123 297 423 60 405 252 110 113 236 54 219 125 329  
## [19] 45 43 148 278 404 379 230 22 260 369 68 187 432 351 392 197 101 456  
## [37] 422 424 467 266 442 366 333 80 3 5 427 172 208 167 213 83 35 247  
## [55] 16 25 465 375 46 263 15 368 86 372 75 249 390 303 323 476 2 321  
## [73] 96 486 385 418 92 435 19 411 415 254 52 371 292 322 223 233 155 294  
## [91] 57 265 245 4 298 203 500 474 417 200 416 66 217 362 175 449 348 29  
## [109] 420 225 47 99 151 461 453 13 134 497 191 412 300 341 382 480 216 112  
## [127] 9 495 429 238 328 132 242 296 384 373 394 186 277 335 50 10 212 56  
## [145] 430 359 355 280 270 470

class(grade\_s)

## [1] "integer"

Above are serial numbers, so basicaly we get the entire row for study.

new<- Bone\_Tumor[grade\_s,3] # 3 represents col number  
new

## [1] 80 54 72 77 70 74 41 72 78 33 67 63 69 70 52 48 69 49 77 71 60 61 59 66 43  
## [26] 66 79 63 57 70 61 69 77 44 44 83 17 64 61 44 78 46 43 41 22 63 70 55 82 59  
## [51] 63 61 73 44 42 67 42 25 61 72 74 77 71 68 45 53 49 60 53 39 54 60 68 52 70  
## [76] 60 63 40 67 59 64 49 42 42 68 83 36 41 76 63 46 83 22 54 70 24 73 67 41 76  
## [101] 78 72 21 57 68 70 60 69 69 67 28 41 45 47 57 62 72 41 51 78 50 83 46 30 74  
## [126] 50 61 54 41 41 56 66 42 44 72 55 30 41 38 72 40 43 66 46 45 67 69 48 70 43

class(new)

## [1] "integer"

#Converting sample to dataframe  
df<- Bone\_Tumor[grade\_s,]  
  
class(df)

## [1] "data.frame"

Target variable=Grade of Bone tumor

count(df,Grade)

## Grade n  
## 1 High 73  
## 2 Intermediate 77

count(df,Grade,Sex)

## Grade Sex n  
## 1 High Female 27  
## 2 High Male 46  
## 3 Intermediate Female 64  
## 4 Intermediate Male 13

## Sample Proportion

#proportion=x/n  
63/150

## [1] 0.42

Sample proportion of high grade tumors is 0.42.

## Setting Hypothesis

P0:Proportion of high grade tumors

# Assuming population proportion

H0: P0 = 0.45

H1: P0! = 0.45

x=63  
n=150  
prop.test(x,n,p=0.45,alternative="two.sided",conf.level = 0.99)

##   
## 1-sample proportions test with continuity correction  
##   
## data: x out of n, null probability 0.45  
## X-squared = 0.43098, df = 1, p-value = 0.5115  
## alternative hypothesis: true p is not equal to 0.45  
## 99 percent confidence interval:  
## 0.3186680 0.5283223  
## sample estimates:  
## p   
## 0.42

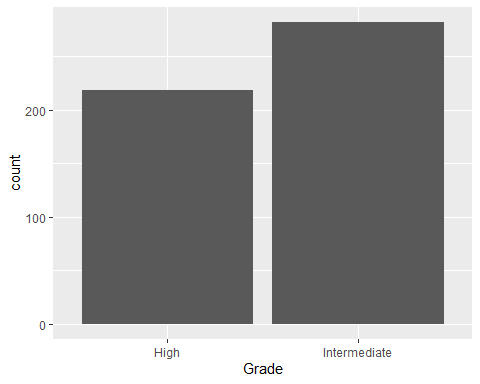
The Chi-squared value 0.43 and p-value is 0.5. Since p>0.01, we fail to reject H0

## Conclusion

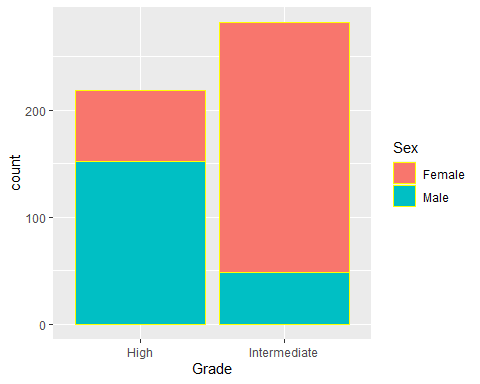
Since p>0.01, we fail to reject H0. Therefore proportion of high grade tumors in the population is 0.45. In other words, more than half of the population have intermediate grade bone tumors.

## VISUALIZATION

library(ggplot2)  
?ggplot2  
  
ggplot(Bone\_Tumor,aes(x=Grade))+geom\_bar()



ggplot(Bone\_Tumor,aes(x=Grade,fill=Sex))+geom\_bar(colour="yellow")



?geom\_bar  
?aes

## BINOM TEST

?binom.test  
binom.test(x, n, p = 0.45,alternative = "two.sided",conf.level = 0.99)

##   
## Exact binomial test  
##   
## data: x and n  
## number of successes = 63, number of trials = 150, p-value = 0.5116  
## alternative hypothesis: true probability of success is not equal to 0.45  
## 99 percent confidence interval:  
## 0.3168637 0.5283480  
## sample estimates:  
## probability of success   
## 0.42

In this test also we get the same p-value as Chi-squared test.

In conclusion, we can say that the proportion of high grade bone tumors is 0.45.

## ALTERNATIVE METHOD TO GET SAMPLE

samp<-sample(Bone\_Tumor$Grade,150,replace = FALSE)  
  
class(samp)

## [1] "character"

g<-as.data.frame(samp)  
  
  
  
count(g,samp)

## samp n  
## 1 High 59  
## 2 Intermediate 91