September 26, 2020

Problem 1: Discuss the rationale behind the T-distributions. Specifically, what is the rationale for using a family of distributions rather than a single one like the z-distribution? (4 points)

T distribution which is similar to the bell-shaped distribution have higher probability for extreme values or events when even sample is too small. When the number of sample N is less than 30 then it is considered small and hence degree of freedom (determined by sample size) will be equal to one N-1. Shape of it changes with degree of freedom. The rationale behind T -distribution can be explained simply as “T-Distribution gives a higher probability to extreme events, given a small sample size”. In the world of statistics t-distribution as well as z-distribution are commonly used for statistical analysis of data. Both of these needs symmetrical or reasonably normally distributed data. The outlier i.e. extremes in data will bring inaccuracy while testing. When population standard deviation is known then it would be appropriate to use Z-test but if there is no standard deviation of population given, with the help of size, mean and standard deviation of sample t-test can be done. Hence, t-test assumes that population standard deviation is unknown, observations in data are independent, the sample is random and the null hypothesis is true.

For the following questions please use the heartValve.csv data set (found on Canvas). A description of the variables can be found below.

Problem 2: You are evaluating the data in the heartValve.csv data set and you’ve become concerned that the data might be biased towards older subjects. However, this is just a hunch and you have no published work to back this up. You do know, however, that the population of people receiving heart valve replacement have an average age of 48.97 years old at the time of surgery. Please assess this hypothesis using a traditional approach where the p-value is calculated from a mathematically defined null distribution (e.g. not generated via bootstrapping or resampling). Be sure to test all assumptions of the test you’re using – if possible. All tests should be carried out in R. (8 Points)

Given Hypothesis:

H0: surgery age == 48.97 H1: surgery age!= 48.97

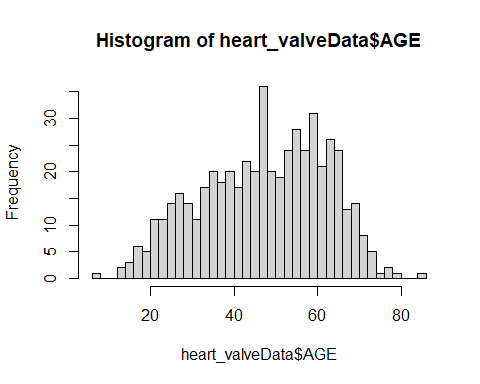
library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts -------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(infer)  
library(moderndive)  
# Read data set  
heart\_valveData <- read.csv("heartValve.csv", header = T)  
hist(heart\_valveData$AGE,30)



K=which(heart\_valveData$AGE!="NA") # collecting not null values  
clean\_hvage<-heart\_valveData[K,]  
# ks.test  
ks.test(clean\_hvage$AGE,"pnorm",mean(clean\_hvage$AGE),sd(clean\_hvage$AGE))

## Warning in ks.test(clean\_hvage$AGE, "pnorm", mean(clean\_hvage$AGE),  
## sd(clean\_hvage$AGE)): ties should not be present for the Kolmogorov-Smirnov test

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: clean\_hvage$AGE  
## D = 0.06783, p-value = 0.01581  
## alternative hypothesis: two-sided

#here  
proMn=48.97 # given  
# shapiro.test  
shapiro.test(clean\_hvage$AGE)

##   
## Shapiro-Wilk normality test  
##   
## data: clean\_hvage$AGE  
## W = 0.98095, p-value = 2.264e-06

# one sample t-test  
t.test(clean\_hvage$AGE, alternative = "two.sided", mu = proMn, var.equal = T)

##   
## One Sample t-test  
##   
## data: clean\_hvage$AGE  
## t = -1.9727, df = 525, p-value = 0.04905  
## alternative hypothesis: true mean is not equal to 48.97  
## 95 percent confidence interval:  
## 46.42694 48.96469  
## sample estimates:  
## mean of x   
## 47.69582

## calculate by hand just to show you the computations   
ageMn = mean(clean\_hvage$AGE)

Problem 3: Please repeat the above steps using a bootstrapping approach. (8 Points)

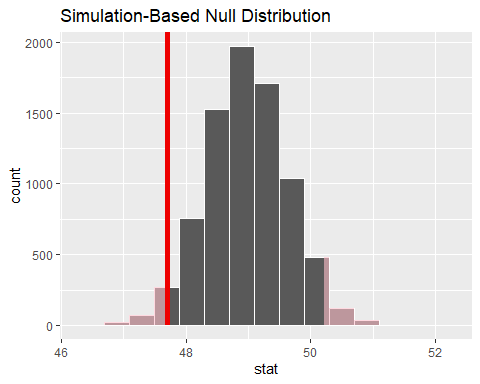
# bootstrapping approach  
number\_rept=8000 # number of repetition  
mean\_age = data.frame(heart\_valveData$AGE) # creating data frame  
null\_distn\_one\_mean <-mean\_age %>%  
 specify(response =heart\_valveData.AGE) %>% # working on the variables  
 hypothesize(null = "point", mu = proMn) %>% # Hypothesize the Null  
 generate(reps = number\_rept) %>% # Generating replicates  
 calculate(stat = "mean") # Get Summary stats

## Warning: Removed 2 rows containing missing values.

## Setting `type = "bootstrap"` in `generate()`.

## Look at the Null Dist with Xbar  
null\_distn\_one\_mean %>%  
 visualize(obs\_stat = ageMn, direction = "two\_sided")

## Warning: `visualize()` should no longer be used to plot a p-value. Arguments  
## `obs\_stat`, `obs\_stat\_color`, `pvalue\_fill`, and `direction` are deprecated. Use  
## `shade\_p\_value()` instead.



## get p-value  
pvalue <- null\_distn\_one\_mean %>%  
 get\_pvalue(obs\_stat = ageMn, direction = "two\_sided")  
pvalue

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0.0488

## Double check, does the p-value make sense?  
k = which(ageMn<=null\_distn\_one\_mean$stat)  
exact\_p\_value = (1 + length(k)) / (number\_rept + 1)  
exact\_p\_value

## [1] 0.975628

Conclusion: Since p-value obtained from various tests here without bootstrapping show less than 0.05 which is statistically significant but in t-test above t score is negative whereas degree of freedom is high which needs further evaluation to check accuracy.

Finally, on applying bootstrapping approach exact p-value is about 0.97 i.e. about 1 and hence it can be inferred that the age mean in significantly smaller. And p-value < alpha where alpha =1.0.05=0.95 (confidence level) and hence null hypothesis is rejected and alternative hypothesis is accepted.