HYpothesis Testing using Ranomization Distributions

by Craig W. Slinkman

October 22, 2015

# Q1:

A certain chemical pollutant in the Genesee River has been constant for several years with mean µ = 34 ppm (parts per million). A group of factory representatives whose companies discharge liquids into the river is now claiming that they have lowered the average with improved filtration devices. A group of environmentalists will test to see if this is true at the 5% level of significance. They have drawn a random sample of size 25 from the river. The sample values are given below in the vector .

## [1] 29 19 30 32 24 13 30 31 25 37 31 38 29 26 39 46 37 32 32 33 27 39 29  
## [24] 47 33

## The null null hypothesis and alternative hypothothesis

The envronment group wants to test the claim that the level of the pollutant is lower than 34 ppm. So are null and alternative hypothesis are:

Perform a hypothesis test (6 step procedure outlined in class) at the 4% level of significance and state your decision.

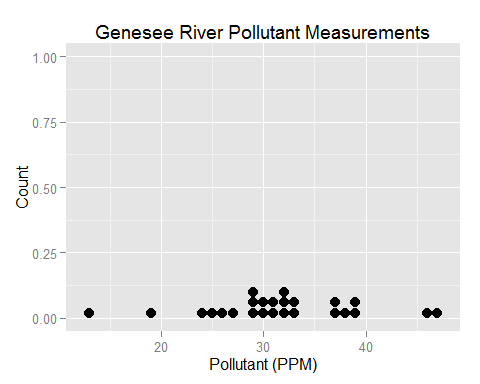
## Graphics and descriptive statistics

The first set in statistical analysis is to draw a suitable grpahic. When analyzing quanitative data with the small sample sizes the dotplot is the ideal graphic.

require( ggplot2 )

## Loading required package: ggplot2

ggplot( data.frame( PPM ),  
 aes( x=PPM )) +  
 geom\_dotplot( binwidth = 1 ) +  
 xlab( "Pollutant (PPM)" ) +  
 ylab( "Count" ) +  
 ggtitle( "Genesee River Pollutant Measurements")



ybar <- round( mean(PPM), 1)   
 s <- round( sd( PPM ), 1)

## Creating the ranomization distribution for the hypothesis test

We must adjust the sample data so that it centered around the null hypothesis value. We do this by subtracting the sample mean from each sample value and then adding the null hypothesis value to each observation.

In R we do this as follows.

muNH <- 34 # NH value  
adjusted <- PPM - ybar + 34 # Adjust sample mean to NH value.  
adjusted

## [1] 31.5 21.5 32.5 34.5 26.5 15.5 32.5 33.5 27.5 39.5 33.5 40.5 31.5 28.5  
## [15] 41.5 48.5 39.5 34.5 34.5 35.5 29.5 41.5 31.5 49.5 35.5

cat( "mean(Adjusted)=", round(mean(adjusted),1 ))

## mean(Adjusted)= 34

We now compute the ranndomization distribution of the adjusted values. We will use 5,000 replications because this is an inportant isssue.

require( simpleboot )

## Loading required package: simpleboot  
## Loading required package: boot  
## Simple Bootstrap Routines (1.1-3 2008-04-30)

reps <- 5000  
bootOut <- one.boot( adjusted,   
 mean,  
 reps )  
rd <- bootOut$t

# Computing the p-value

Because we have a lowe tail hypothesis we compute the proportion of observations whose value is less that the $${y} and divide this number by the number of replications . We do this below:

lowerTail <- adjusted[ rd < ybar ]  
pValue <- round( length( lowerTail ) / reps, 4 )  
xx <- round( 10000 \* pValue )

The p-value for the null hypotheis is 0.0416.

## Statistical conclusion

This is evidence that the null hypothesis is incorrect.

\*At the 5% or 0.05 level of significance we would reject the null hypothesis.

\*There is insufficient evidence to reject the null hypothesis as the 0.01 or 1% level of significance.

## Managerial conclusion

There is some evidence that the pollution level has been reduced. At the scientific standad of the 5% level of signifcant there is evidence that the pollution level has been decreased.

# Coporate health care

The Human Resources Department of a large corporation wanted to determine if a majority of its employees were satisfied with their treatment by the corporation's health care provider. A random sample of 400 employees was selected, and 251 indicated that they were satisfied with their treatment. Does this data show at the 1% level of significance that a majority of all employees is satisfied? Suppose that the Corporation's president required that more than 60% of employees should be satisfied. Does this data support that requirement at the 10% level of significance? Construct a 95% confidence interval for the proportion of all employees who are satisfied. What sample size would be required to estimate this proportion to within with 90% confidence if no prior bounds are placed on the population proportion?

## The null and alternative hypothesis

We want to show that the proprtion of employees that are staisfied with the corporate heath care provider is greater than 50% which is a proportion of of 0,50. Hence our null and alternative hypotheis are

## Synthetic sample

We are not given the orginal sample data but sumerized data. We know that 263 employees are satisified with the provider and that 249 are not satisfied. We use these figures to construct our symethetic sample which I will call :

rm( list=ls() )  
n <- 500  
s <- 263  
f <- n - s  
  
survey <- c( rep( 1, s ),  
 rep( 0, f ))  
table(survey)

## survey  
## 0 1   
## 237 263

phat <- round( sum(survey)/ n, 3 )

## Computing the point estimate

phat <- round( s / n )

The point estimate for the proprtion is computed as follows

## Adjust the sample data to make the null hypothesis true

We assume that the null hypothesis is true. For proportions we can do this thoereticall because under the null hypothesis 50% of the sample sizes should be true and 50% should be false. Therefore we have

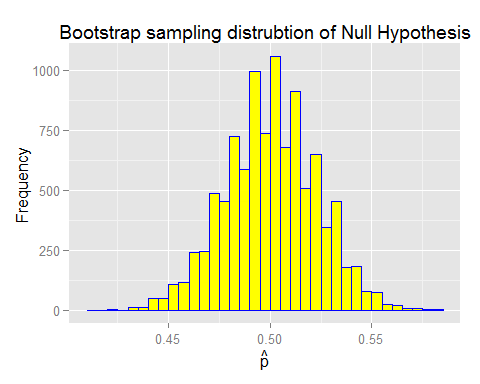
trueNH <- c( rep( 1, 250 ),  
 rep( 0, 250 ))  
table(trueNH)

## trueNH  
## 0 1   
## 250 250

## Compute the randimization distribution

We use the trueMH to compute the randomization distribution using the bootstrap.

require( simpleboot )  
  
bootOut <- one.boot( trueNH,  
 mean,  
 10000 )  
rd <- bootOut$t  
  
ggplot( data.frame( rd ),  
 aes(x=rd)) +  
 geom\_histogram( binwidth=0.005,  
 color='blue',  
 fill='yellow') +  
 geom\_segment( x=phat, y=0,  
 xend = phat, yend=1100) +  
 xlab( expression( hat(p) )) +  
 ylab( "Frequency ") +  
 ggtitle( "Bootstrap sampling distrubtion of Null Hypothesis" )



## Computing the p-value

The alternative hypothesis is an upper-tail hypothesis. Therefore we count all the randomization distribution values and divide the replicaltions.

count <- length( rd[rd>phat] )  
pValue <- round( count / 10000, 4 )

## Statistical conclusion

Because the p-value=0.1157 is not less than 0.05 there is not enough sample evidence to reject the null hypothesis.

## Managerial conclusion

There is insufficient sample evidence to conclude that more than 50% of our employess approve of our healthcare provider.

# 3: Commute Atlanta: Is there a difference mean commutetime driving home between genders

We want dtermine if there is a difference between the mean commute times in Atlanta for different genders.

## The data

The data can be accessed in the R-package in the Coomute Atlanta. We need to divide the data into two subsramples based on the sex driver. We are only interest in the variable which is the variable in the dataset.

## The hypothesis set

We define the following symbols

Parameter | Interpretation  
----------|---------------  
 | Male mean commute time  
 Female mean commute time

We can formulate out hypotheses as follows:

A more useful but equivalent formulation of our hypothesis set is

## The data set

The dataset can be found in the R-package $Lock5Data. The data frame name is . We need to create two subsamples based on gender. We are only interested in the compute time variable .

The chunk below shows how to the carry out these tasks.

rm(list=ls())  
set.seed( 14925 )  
  
require( Lock5Data )

## Loading required package: Lock5Data

data( CommuteAtlanta )  
head( CommuteAtlanta )

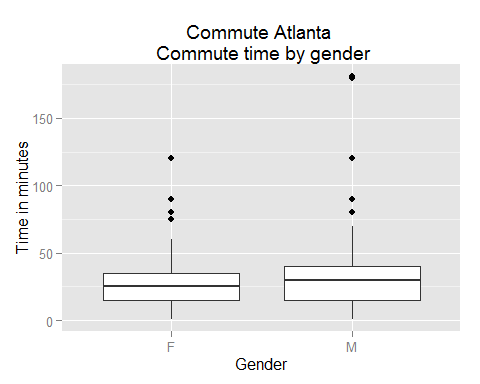
## City Age Distance Time Sex  
## 1 Atlanta 19 10 15 M  
## 2 Atlanta 55 45 60 M  
## 3 Atlanta 48 12 45 M  
## 4 Atlanta 45 4 10 F  
## 5 Atlanta 48 15 30 F  
## 6 Atlanta 43 33 60 M

Males <- CommuteAtlanta[CommuteAtlanta$Sex=="M",4]  
Females <- CommuteAtlanta[CommuteAtlanta$Sex=="F",4]

## Graphical and describptive statistics

We should always polt the data when it is qunatitative. Because we are comparing nultiple sub-populations we use comparitive boxplots. Note that we need the orginal data frame to plot this data.

require( ggplot2 )  
   
ggplot( CommuteAtlanta,  
 aes( x=Sex, y=Time)) +  
 geom\_boxplot() +  
 xlab( "Gender" ) +  
 ylab( "Time in minutes" ) +  
 ggtitle( "Commute Atlanta \n Commute time by gender")



barM <- round( mean( Males), 1 )  
barF <- round( mean( Females ), 1 )  
barDif <- barM - barF

We have the following sample statistics:

## Adjusting the sample data to make the null hypothesis true

We the null hypothis value is zero to adjust the sample to make the null hypotheis true all we need to do is to subtatct the means of each sub-sample:

adjM <- Males - barM  
adjF <- Females - barF  
round( mean( adjM ), 1 )

## [1] 0

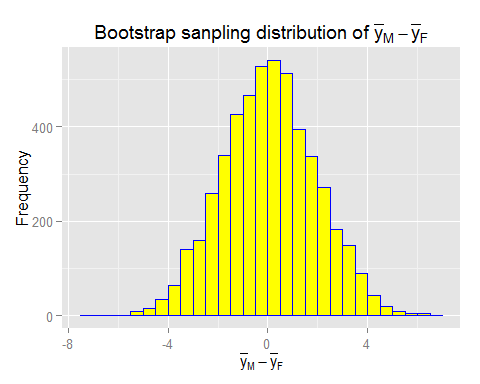
round( mean( adjF ), 1 )

## [1] 0

## Creating the ranomized distribution

Because we have two populations we must use the R-function in the package.

require( simpleboot )  
  
bootOut <- two.boot( adjM,   
 adjF,  
 mean,  
 5000 )  
rd <- bootOut$t  
  
ggplot( data.frame( rd ),  
 aes( x = rd )) +  
 geom\_histogram( binwidth=0.5,  
 color="blue",  
 fill="yellow") +  
 xlab( expression(bar(y)[M]-bar(y)[F]) ) +  
 ylab( "Frequency" ) +  
 ggtitle( expression( paste("Bootstrap sanpling distribution of ",   
 bar(y)[M]-bar(y)[F]) ))



## Compute the pvalue using the sample statistic

The computation of the p-value is more complex in two tail tests than in single tail tests. We must include areas from both tails. If the sampling distribution is syetric we can approximate the p-vlaue by multiply in the p-value as if we had a one-tail by two in the tail that test statstic is found.

count <- length( rd[rd>=barDif])  
pValue <- round( 2 \* count/5000, 4 )

## Statistical conclusion

The p-value of 0.0148 is less than 0.05 so we shoulds reject the sample data is inconsistent with the null hypothesis.

## Magaerial Conclusion

There is mildly strong evidence that the null hypotheis that the mean commute time is the same for both males and females is incorrect.