Hypothesis Testing using Randomization 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 hypothesis and alternative hypothesis

The environment 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 at the 5% level of significance and state your decision.

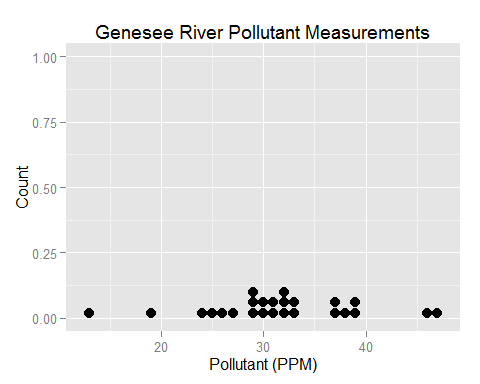
## Graphics and descriptive statistics

The first set in statistical analysis is to draw a suitable graphic. When analyzing quantitative 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 randomization distribution for the hypothesis test

We must adjust the sample data so that it centered on 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.

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)=", # For clearer output.  
 round(mean(adjusted),1 ))

## mean(Adjusted)= 34

We now compute the randomization distribution of the adjusted values. We will use 5,000 replications because this is an important issue.

###########################################################  
# We must load the simpleboot package for bootstrap #  
# functionality. #  
###########################################################  
  
require( simpleboot )

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

#  
#  
###########################################################  
# Because of the importance of the decision we use choose #  
# to use 5,000 replications. #  
###########################################################  
#  
reps <- 5000   
#  
#  
##########################################################  
# We compute 5,000 bootstrap samples of the data that #  
# was adjusted to make the null hypothesis true for the #  
# adjusted data. #  
##########################################################  
#  
bootOut <- one.boot( adjusted,   
 mean,  
 reps )  
#  
#  
##########################################################  
# We extract the bootstrap sampling distribution from #   
# bootOut list. #  
##########################################################  
#  
bsd <- bootOut$t # Bootstrap sampling distribution   
#  
#'  
##########################################################  
# We compute the usual bootstrap diagnostics: #  
# #  
# 1. Bootstrap sampling mean which should be #  
# approximately equal to the NH value. #  
# #  
# 2. We also plot the bootstrap sampling distribution #  
# along with the plotted sample statistic to give us #  
# a visual interpretation of the pVlaue. #  
##########################################################  
#  
bsMean <- round( mean( bsd ) ) # Bootstrap sampling   
 # mean

# Computing the p-value

Because we have a lower 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[ bsd < ybar ]  
pValue <- round( length( lowerTail ) / reps, 4 )  
xx <- round( 10000 \* pValue )

The p-value for the null hypothesis 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 standard of the 5% level of significance there is evidence that the pollution level has been decreased.

# Corporate 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 275 indicated that they were satisfied with their treatment. Does this data show at the 2% 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 2% level of significance?

## The null and alternative hypothesis

We want to show that the proportion of employees that are satisfied with the corporate health care provider is greater than 60% which is a proportion of 0,60. Hence our null and alternative hypothesis are

## Synthetic sample

We are not given the original sample data but summarized data. We know that 263 employees are satisfied with the provider and that 237 are not satisfied. We use these figures to construct our synthetic sample which I will call :

rm( list=ls() )  
n <- 500  
s <- 323  
f <- n - s  
  
survey <- c( rep( 1, s ),  
 rep( 0, f ))  
table(survey)

## survey  
## 0 1   
## 177 323

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

## Computing the point estimate

phat <- round( s / n )

The point estimate for the proportion is computed as follows

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

We assume that the null hypothesis is true. For proportions we can do this theoretically because under the null hypothesis 60% of the sample sizes should be true and 40% should be false. Therefore we have

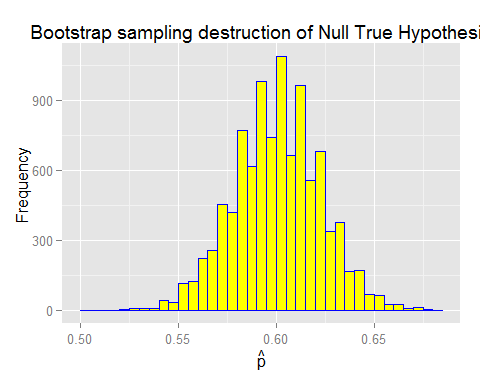
trueNH <- c( rep( 1, 300 ),  
  
 rep( 0, 200 ))  
table(trueNH)

## trueNH  
## 0 1   
## 200 300

## Compute the randomization distribution

Because we are using random sample data we can create the bootstrap sampling distribution the true NH synthetic sample.

set.seed( 111 ) # For reproducibility.  
  
require( simpleboot ) # Load simpleboot package.  
  
reps <- 10000  
  
bootOut <-   
 one.boot( trueNH,  
 mean,  
 reps )  
  
bsd <- bootOut$t  
  
ggplot( data.frame( bsd ),  
 aes(x=bsd)) +  
 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 destruction of Null True Hypothesis" )



## Computing the p-value

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

count <- length( bsd[bsd>phat] )  
pValue <- round( count / reps, 4 )

## Statistical conclusion

Because the p-value=0 is less than 0.02000 there is sufficient evidence to conclude that the null hypothesis is unlikely and we reject it.

## Managerial conclusion

There is sufficient evidence to conclude that the percentage of employees who are satisfied with the healthcare plan exceeds 60%.

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

We want determine 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 data frame. We need to divide the data into two subsamples 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 )

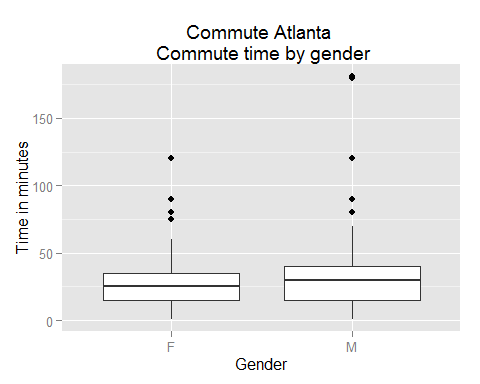
## City Age Distance Time Sex  
## 1 Atlanta 19 10 15 M  
## 2 Atlanta 55 45 60 M  
## 3 Atlanta 48 12 45 M  
## 4 Atlanta 45 4 10 F  
## 5 Atlanta 48 15 30 F  
## 6 Atlanta 43 33 60 M

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

## Graphical and descriptive statistics

We should always plot the data when it is quantitative. Because we are comparing multiple sub-populations we use comparative boxplots. Note that we need the original 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 hypothesis value is zero to adjust the sample to make the null hypothesis true all we need to do is to subtract the means of each sub-sample:

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

## [1] 0

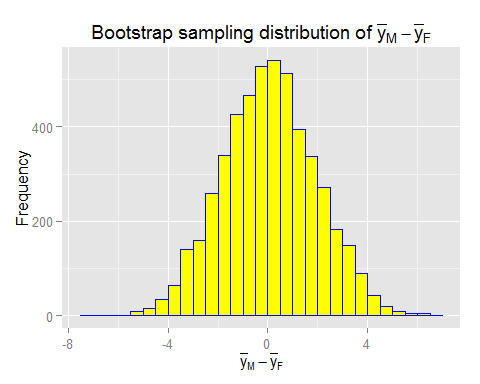
round( mean( adjF ), 1 )

## [1] 0

## Creating the randomized 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 sampling distribution of ",   
 bar(y)[M]-bar(y)[F]) ))



## Compute the p-value 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 symmetric we can approximate the p-value by multiply in the p-value as if we had a one-tail by two in the tail that test statistic 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 should reject the sample data is inconsistent with the null hypothesis.

## Managerial Conclusion

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

# Two sample proportions

The HuffPost Pollster (<http://elections.huffingtonpost.com/pollster/2016-iowa-presidential-republican-primary>) gives the following numbers for the Iowa Presidential Caucus.

n <-401 #Sample size  
nCarson <- 84 # Favor Carson  
nTrump <- 59 # Favor Trump  
#  
pCarson <- nCarson / n  
  
pTrump <- nTrump / n   
  
pDif <- pCarson - pTrump

|  |  |  |
| --- | --- | --- |
| Candidate | Count | Proportion |
| Carson | 84 | 0.209 |
| Trump | 59 | 0.147 |

Is there sufficient evidence at the 5% level of significance that Carson will finish ahead of trump provided the caucus follows the opinion polls?

## The Hypothesis set

Let be the proportion of votes who favor Carson and be the proportion of voters who favor Trump. Our hypothesis set is as follows:

## The point estimate

The point estimate of the difference in proportion is

## Creating the boot sampling distribution

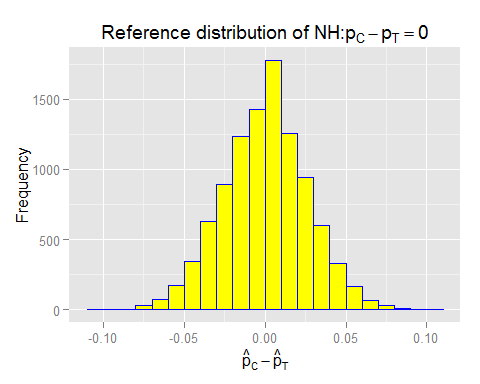
Because out results are from a random sample and not an experiment we can use the bootstrap sampling distribution. We will use 10,000 replications and the $ package function $$two.boot$ to create the bootstrap distribution.

As before we will need to center the bootstrap distribution so that difference in proportion is equal to zero. One way to do this is two create sampling at the mean proportion for both samples. We compute this sampling distribution below.

Carson <- c ( rep( 1, nCarson ), rep( 0, n - nCarson ) )  
adjCarson <- Carson - pCarson  
  
  
  
Trump <- c( rep( 1, nTrump), rep(0, n-nTrump))  
adjTrump <- Trump - pTrump  
  
reps <- 10000  
require( simpleboot )  
  
bootOut <- two.boot( adjCarson,  
 adjTrump,  
 mean,  
 reps )  
  
bsd <- bootOut$t

We compute the mean of the bootstrap sampling distribution. This value should be approximately equal to zero. We also should plot the sampling distribution which we hope will v=be symmetric and mound shaped.

bsdMean <- round( mean( bsd ), 3 )  
  
require( ggplot2 )  
  
ggplot( data.frame( bsd ),  
 aes(x=bsd )) +  
 geom\_histogram( binwidth=0.01,  
 color="blue",  
 fill="yellow") +  
 xlab( expression( hat(p)[C] - hat(p)[T]) ) +  
 ylab( "Frequency" ) +  
 ggtitle( expression(  
 paste( "Reference distribution of NH:",  
 p[C] - p[T]==0)))

 The mean of the reference distribution is equal to 0 to 3 decimal places. The reference distribution looks fairly symmetric and mound shaped. Therefore we can proceed with computing the p-value.

## Computing the p-value

The alternative hypothesis specifies upper tail test. We therefore we compute the p-value:

pValue <- round( ( length( bsd[ bsd >= pDif] )/ reps), 4)

The p-values is 0.0101.

## Statistical conclusion

Because the p-value=0.0101 is less than 0.0500. There is relatively strong evidence to reject the null hypothesis that Carson and Trump are tied.

## Subject matter conclusion

There is relatively strong evidence that Mr. Carson will win the Iowa Caucus if the Caucus is held on the same day as the survey.

# Cocaine addiction

In the statistical theory of design of experiments, randomization involves randomly allocating the experimental units across the treatment groups. For example, if an experiment compares a new drug against a standard drug, then the patients should be allocated to either the new drug or to the standard drug control using randomization.

We are going to analyze two analyze the following healthcare problem:

In a randomized experiment on treating cocaine addiction, 48 people were randomly assigned to take either Desipramine (a new drug), or Lithium (an existing drug), and then followed to see who relapsed

The research question of interest: Is Desipramine better than Lithium at treating cocaine addiction?

We will use the following notation for the experiment:

the proportion of individuals receiving Desipramine who did not relapse

the proportion of individuals receiving lithium who did not relapse.

The sample data is in the R-package . The data frame name is named . The data is given below.

rm( list=ls())  
require( Lock5Data )  
data("CocaineTreatment")  
CT <- CocaineTreatment  
head(CT)

## Drug Relapse  
## 1 Desipramine no  
## 2 Desipramine yes  
## 3 Desipramine no  
## 4 Desipramine yes  
## 5 Desipramine no  
## 6 Desipramine yes

tail(CT)

## Drug Relapse  
## 67 Placebo yes  
## 68 Placebo no  
## 69 Placebo yes  
## 70 Placebo yes  
## 71 Placebo yes  
## 72 Placebo yes

We need to compute the proportion of individuals who relapsed. We will do this as follows:

outcome <- table( CT )  
outcome

## Relapse  
## Drug no yes  
## Desipramine 14 10  
## Lithium 6 18  
## Placebo 4 20

pD <- outcome[1,1] / 24  
pL <- outcome[2,1] / 24  
pDif <- pD - pL

The point estimator of the difference in proportions is

## Hypothesis set

Our research hypothesis of interest is that using Desipramine will have a lower relapse proportion than patients who receive Lithium. Our hypothesis set is:

Equivalently we can rewrite the above hypothesis set as

## Creating the randomization distribution

Because in experiments treatments are randomized to experimental subjects our reference distribution should be created in the same way. We can do this by using the R-function $. WE will show how to do this below for a single randomization sample.

#  
###########################################################  
# We need to change the strings to numeric o's and 1's so #  
# we can use the mean function to compute proportions. #  
###########################################################  
#  
y <- ifelse(CT$Relapse[1:48]=="yes",1, 0 )  
#  
#  
###########################################################  
# We create an index to assign treatments to the random #  
# Groups D and L. #  
###########################################################  
#  
All <- 1:48  
#  
#  
###########################################################  
# We ranomly assign 24 of the values in y to group D. #  
###########################################################  
#  
D <- sample.int( 48, 24, replace=FALSE)  
D

## [1] 36 30 14 40 37 17 13 42 9 27 18 34 24 29 46 7 3 44 43 4 35 1 19  
## [24] 31

y[D]

## [1] 1 0 0 1 0 1 0 1 0 1 1 1 0 1 1 1 0 0 1 1 1 0 0 0

#  
#  
###########################################################  
# We assign any objects not assigned to group D to group #  
# group L. #  
###########################################################  
L <- All[-D]  
L

## [1] 2 5 6 8 10 11 12 15 16 20 21 22 23 25 26 28 32 33 38 39 41 45 47  
## [24] 48

y[L]

## [1] 1 0 1 0 1 0 0 1 1 0 0 1 0 0 1 1 1 1 1 1 1 0 1 1

#  
#  
##########################################################  
# We compute the difference between the completely #  
# randomized groups. #  
##########################################################  
#  
dif <- mean(y[D]) - mean( y[L])  
dif

## [1] -0.08333333

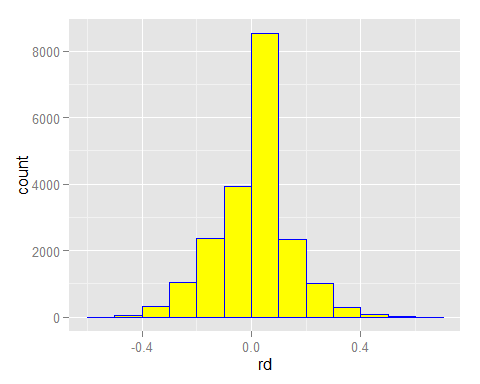
In practice we need to do this a large number of times. The code below will create randomization distribution with 20,000 replcations using the drug data.

reps <- 20000  
rd <- rep( NA, reps)  
  
outcome <- ifelse(CT$Relapse[1:48]=="yes", 1, 0 )  
trials <- 1:48  
  
for( i in 1:reps){  
 D <- sample.int( 48, 24, replace=FALSE )  
 L <- trials[-D]  
 rd[i]= mean( y[D] - y[L])  
}

The mean of the randomization should be approximately equal to zero. Its value is -0.001. This value is close to zero.

We plot the randomization distribution below.

require( ggplot2 )  
  
ggplot( data.frame(rd),  
 aes(x=rd)) +  
 geom\_histogram( binwidth=0.10,  
 color="blue",  
 fill="yellow")



## Computing the p-value

WE have a lower tail test so the p-value is computed for all the randomization values less that the variable .

pValue <- round( length( rd[rd >= pDif] ) / reps, 4 )

## Statistical conclusion

because the p-value of 0.004 is much smaller than 0.0500 we conclude that the bull hypothesis is incorrect.

## Clinical conclusion

There is very strong sample evidence to conclude that Desipramine reduces relapses more than Lithium.

# Using a confidence interval to test a hypotheis

Suppose a craft brewer brews a beer called Winter Warmer. The beer should have an alcholoic content of 11%. A random sample of size 15 is drawn from 10 random retail outlets and the alcololic content by volume, is measured.

The sample data is given below.

## [1] 11.9 11.9 11.2 11.5 12.0 11.9 11.3 11.4 11.4 11.8

## The null and the alternative hypotheses

Because we do not apriori know if the ABV is higher or lower we use a two-tailed hypothesis test.