# Cover Sheet

By including this statement, we, all the students listed in the table below, declare that:

* We hold a copy of this assignment if the original is lost or damaged.
* We hereby certify that no part of this assignment has been copied from any other student’s work or from any other source except where due acknowledgement is made in the assignment.
* No part of the assignment has been written for us by any other person except where collaboration has been authorised by the unit coordinator.
* We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism; this software may retain a copy on its database for future plagiarism checking.
* We hereby certify that no part of this assignment or product has been submitted by any of us in another (previous or current) assessment, except where appropriately referenced, and with prior permission from the unit coordinator for this unit.
* We hereby certify that we have read and understand what the University considers to be academic misconduct, and that we are aware of the penalties that may be imposed for academic misconduct.

|  |  |  |
| --- | --- | --- |
| Name | Student Number | Contribution (%) |
| Rachel Hardie | 18820821 |  |
| Dylan Wang | 18998014 |  |
| Bradley McInerney | 19029900 |  |
|  |  |

# Group Project

This Group Project analyses hard drive reliability data. The file HardDisks.csv contains the following information for 125864 hard drives: (1) a unique serial number; (2) the model of the hard drive; (3) the number of days the hard drive was operational; (4) the mean operating temperature of the drive; and (5) whether the hard drive was removed because it had failed. A drive that is listed as not having failed either was still operational at the end of the data collection period, or it was removed for other reasons, for instance a capacity upgrade.

#Import drives data set  
drives = read.csv("HardDisks.csv",header=TRUE) #Load a dataset from .csv file  
attach(drives)  
  
#Check if serial\_number in indeed unique  
n\_occur <- data.frame(table(serial\_number))  
n\_occur[n\_occur$Freq > 1,]

## [1] serial\_number Freq   
## <0 rows> (or 0-length row.names)

#As there are 0 rows that appear twice then it is confirmed unique in this set

## Part One - Proportion of drives that fail early

We want to investigate the proportion of drives that fail in the first year of operation. (a) [1 mark] Create a subset of the cases that is relevant for this analysis and print the number of cases in that subset.

n=nrow(drives)#total number of cases  
  
drives.failed=which(failed=="TRUE") #subset of cases that failed  
n.drives.failed=length(drives.failed)  
  
drives.failed.y1=which(failed=="TRUE" & days < 366) #subset of cases that filed in first year  
n.drives.failed.y1=length(drives.failed.y1) #number of cases in subset  
  
paste(length(drives.failed.y1),"drives have failed in their first year")

## [1] "2944 drives have failed in their first year"

testdata.n = nrow(drives) #total number of test cases  
  
#Create binomial data for cases that fail in the first year, this makes generating the boostrap much faster  
failedY1.binom = rep(NA, testdata.n)  
for (row in 1:testdata.n){  
 if(drives[row, "days"]<366 & drives[row, "failed"]==TRUE){ #If they failed when they were under a year old add a 1, else 0, to make is a bonomial  
 failedY1.binom[row] = 1  
 }else{  
 failedY1.binom[row] = 0  
 }  
}

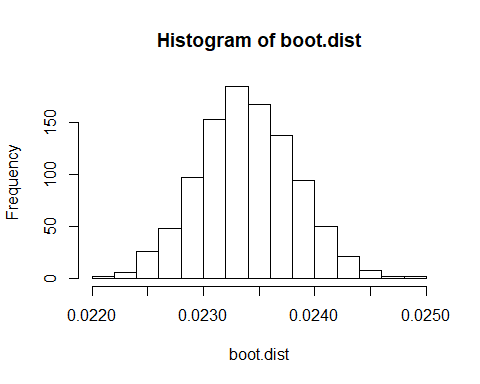
1. [1 mark] Compute a point estimate for the proportion of drives that fail in the first year of operation.

phat=n.drives.failed.y1/n  
paste("The probability of drives that fail in the first year is",phat, "(or", phat\*100,"%)")

## [1] "The probability of drives that fail in the first year is 0.0233903260662302 (or 2.33903260662302 %)"

1. [2 marks] Use bootstrapping to compute a 99% confidence interval for the proportion of drives that fail in the first year of operation

##Generating a Bootstrap Distribution  
b = 1000 #number of bootstrap statistics  
boot.dist = rep(NA, b)  
for (i in 1:b) {  
 boot.sample = sample(failedY1.binom, replace=TRUE)  
 boot.dist[i] = sum(boot.sample)/length(boot.sample) #As we are testing for proportion  
}  
hist(boot.dist)



CI = quantile(boot.dist, c(0.005,0.995)) #Quartiles for a 99% CI  
cat("The 99% confidence interval for the proportion of drives that fail in the first year of operation is", CI[1], CI[2])

## The 99% confidence interval for the proportion of drives that fail in the first year of operation is 0.02230975 0.02454236

quantile(boot.dist, c(0.005, 0.995)) ##99% confidence interval

## 0.5% 99.5%   
## 0.02230975 0.02454236

## Part 2 Temperature and time to failure

For this part, we only consider the drives that failed. We want to analyse whether the mean operating temperature of the drive and the time to failure of the drive are associated.

1. [1 mark] Compute and interpret the correlation between the mean operating temperature and the number of days until failure.

failed.meantemp=meantemp[drives.failed]  
failed.days=days[drives.failed]  
  
cor.test(failed.meantemp,failed.days)

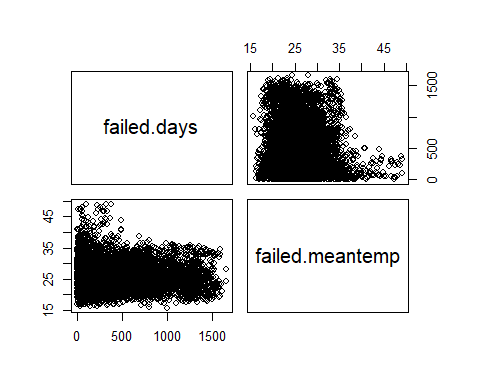
##   
## Pearson's product-moment correlation  
##   
## data: failed.meantemp and failed.days  
## t = -3.759, df = 7331, p-value = 0.0001719  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.06668266 -0.02099337  
## sample estimates:  
## cor   
## -0.04386095

1. [2 marks] Use randomisation to test at a significance level of 5% whether there is evidence that a higher mean op- erating temperature is associated to earlier failure.

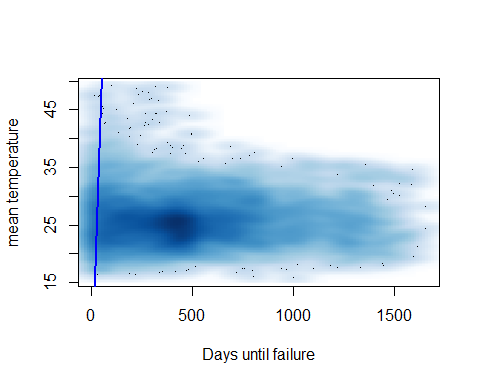
#paired t-test  
  
t.test(failed.days,   
 failed.meantemp,   
 paired=TRUE,   
 conf.level=0.95)

##   
## Paired t-test  
##   
## data: failed.days and failed.meantemp  
## t = 111.38, df = 7332, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 463.7567 480.3727  
## sample estimates:  
## mean of the differences   
## 472.0647

pairs(failed.days ~ failed.meantemp)



smoothScatter(failed.days, failed.meantemp,  
 pch = ".",  
 xlab="Days until failure",  
 ylab="mean temperature")  
  
abline(0,1, col="blue", lwd=2)



1. [1 mark] Interpret your findings, comparing the results from parts (a) and (b). Discuss, in particular, whether there is evidence that a higher operating temperature causes drives to fail earlier.

## Part 3 Three 2TB drive models

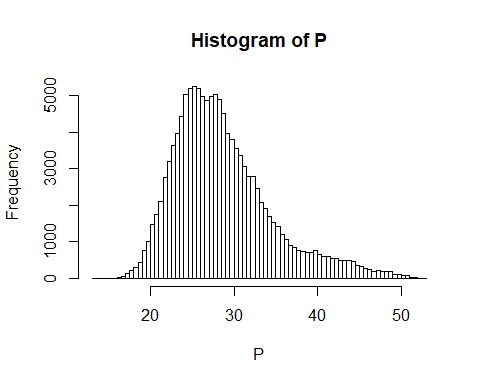
For this part, we only consider 2TB drives with the following model identifiers: “Hitachi HDS723020BLA642” “ST320005XXXX” (Seagate) “WDC WD20EFRX” (Western Digital)

1. [2 marks] At a significance level of 1%, test whether there is evidence for a difference in the mean operating temperature between the three drive models and conduct a pairwise t-test. Discuss your findings. Be specific about any differences between the models that can be inferred from the data.

#Hypothisis test  
#H0: The mean of the mean operating temperatures of all 3 tested drives will be the same  
#HA: The mean of the mean operating temperatures for atleast 1 tested drive with differ from the others  
  
#subsetting data  
P = drives[,"meantemp"] #The data the subsets came from  
mt1=drives[model=="Hitachi HDS723020BLA642", "meantemp"]  
mt2=drives[model=="ST320005XXXX", "meantemp"]  
mt3=drives[model=="WDC WD20EFRX", "meantemp"]  
  
paste("Using an F-distribution to compute p-values requires all data that the subsets came from being normally distributed and all populations having the same variance")

## [1] "Using an F-distribution to compute p-values requires all data that the subsets came from being normally distributed and all populations having the same variance"

hist(P, breaks=100)



paste("Via a histogram of the data the subsets came from, we can assume it is normally distributed")

## [1] "Via a histogram of the data the subsets came from, we can assume it is normally distributed"

paste("As all pf the data that the subsets came from is the same set AND the sample standard deviations(", sd(mt1), sd(mt2), sd(mt3), ") differ by less then factors of 2, we can assume that the population variances are the same.")

## [1] "As all pf the data that the subsets came from is the same set AND the sample standard deviations( 4.52455965340018 3.71624390804005 4.6232421415749 ) differ by less then factors of 2, we can assume that the population variances are the same."

MT = c(mt1, mt2, mt3) #combined set  
N=length(MT) #length of combined set  
k = 3 #number of groups  
  
SST=(N-1)\*sd(MT)^2  
SSE=(length(mt1)-1)\*sd(mt1)^2 + (length(mt2)-1)\*sd(mt2)^2 + (length(mt3)-1)\*sd(mt3)^2  
SSG=SST-SSE  
fstat = ((SSG/(k-1)) / (SSE/(N-k)))  
  
pval=pf(fstat,k-1,N-k,lower.tail=FALSE)  
print(pval)

## [1] 0.0124111

paste("At a signifigance level of 1% (0.01) and a p-value of", pval, ", as the p value is greater than the our signifigance level, we do not reject the null hypothisis 'there is no difference in mean operating temperatures between the 3 tested divice models'.")

## [1] "At a signifigance level of 1% (0.01) and a p-value of 0.0124111018427071 , as the p value is greater than the our signifigance level, we do not reject the null hypothisis 'there is no difference in mean operating temperatures between the 3 tested divice models'."

paste("So there is statistical evidence on a 1% signifigance that there is no difference in mean operating temperatures between the 3 tested divice models.")

## [1] "So there is statistical evidence on a 1% signifigance that there is no difference in mean operating temperatures between the 3 tested divice models."

#Pariwise T-test  
MT.df = drives[model=="Hitachi HDS723020BLA642" | model=="ST320005XXXX" | model=="WDC WD20EFRX", ] #get MT but as a dataframe  
pairwise.t.test(MT.df$meantemp, MT.df$model) #Use a paired t test to see the individual differences

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: MT.df$meantemp and MT.df$model   
##   
## Hitachi HDS723020BLA642 ST320005XXXX  
## ST320005XXXX 0.0098 -   
## WDC WD20EFRX 0.0274 0.1423   
##   
## P value adjustment method: holm

#Discussion  
cat()

### Discussion:

This whole test is under the assumption that the 3 chosen models only come in (or are only recoded in) 2TB versions as there is no data in the data frame to determine the sizes of the drives.  
 The sizes in subsets for each drive varied greatly with the subset lengths: 11, 18, 167.",  
 The pairwise t-test shows that the higest difference in means was between the models 'WDC WD20EFRX' and 'ST320005XXXX'  
 It also shows that the lowest was between 'Hitachi HDS723020BLA642' and 'WDC WD20EFRX'."

1. [3 marks] At a significance level of 1%, test whether there is evidence for a difference in the proportion of failed drives between the three drive models. Discuss your findings. Be specific about any differences between the models that can be inferred from the data.

The end

#leave this as the last chunk  
detach(drives)