Math 3070 § 1.	**Mark Twain Example:**	Name: Example
Treibergs	**Pooled t-Test.**	June 20, 2011

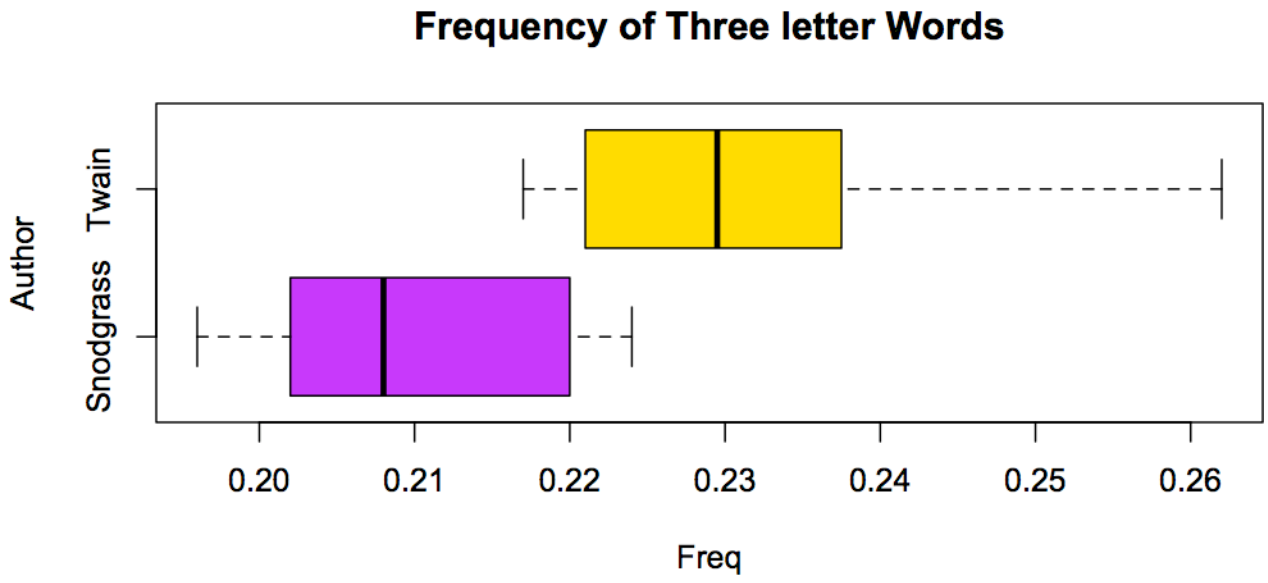
Data File Used in this Analysis:

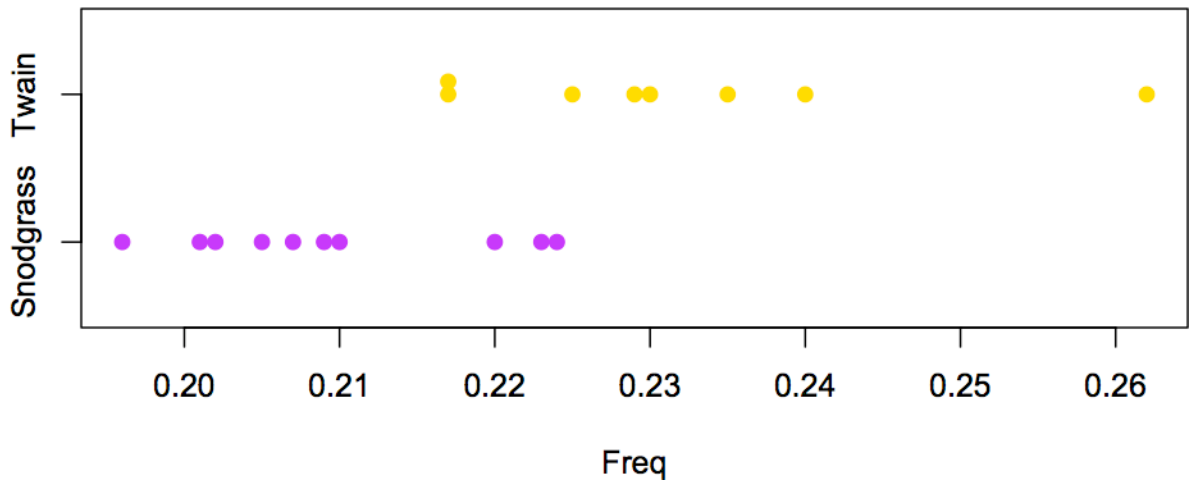
```
# Math 3070 - 1         Mark Twain Example         June 21, 2011
# Treibergs
#
# From Larsen and Marx, "An Introduction to Mathematical Statistics and its
# Applications, 4th ed.," Pearson 2006
#
# An article of Claude Brinegar in the Journal of the American Statistical
# Association, 1963, tried to establish the authorship of writings.
# Ten essays appeared in the newspaper, "New Orleans Daily Crescent" by
# someone who called himself Quintus Curtius Snodgrass, but were suspected
# to be by Mark Twain (also a pseudonym!) Studies have shown that authors
# are very consistent in the frequecy they use words of a certain length.
# A statistical test is whether the frequencies in the Snodgrass articles
# agree with the frequencies in articles known to be by Twain.
#
# Test whether the mean frequency of the two authors is equal.
#
# The data represents observed frequencies of three letter words
#
Freq  Author
.225   Twain
.262   Twain
.217   Twain
.240   Twain
.230   Twain
.229   Twain
.235   Twain
.217   Twain
.209   Snodgrass
.205   Snodgrass
.196   Snodgrass
.210   Snodgrass
.202   Snodgrass
.207   Snodgrass
.224   Snodgrass
.223   Snodgrass
.220   Snodgrass
.201   Snodgrass
```

R Session:

```
R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.31 (5538) powerpc-apple-darwin8.11.1]

[Workspace restored from /Users/andrejstreibergs/.RData]

> tt <- read.table("M3074TwainData.txt",header=TRUE)
> tt
    Freq    Author
1  0.225     Twain
2  0.262     Twain
3  0.217     Twain
4  0.240     Twain
5  0.230     Twain
6  0.229     Twain
7  0.235     Twain
8  0.217     Twain
9  0.209 Snodgrass
10 0.205 Snodgrass
11 0.196 Snodgrass
12 0.210 Snodgrass
13 0.202 Snodgrass
14 0.207 Snodgrass
15 0.224 Snodgrass
16 0.223 Snodgrass
17 0.220 Snodgrass
18 0.201 Snodgrass
```

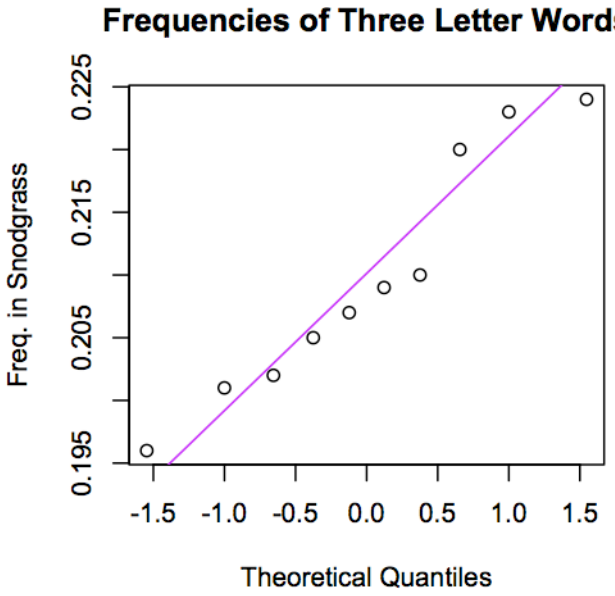
```
> attach(tt)
> ######################## PICK OFF TWAIN AND SNODGRASS DATA ################
> FT <- Freq[Author=="Twain"]
> FS <- Freq[Author=="Snodgrass"]

> ######################## PLOT COMPARATIVE BOXPLOTS AND STRIPCHARTS ########
> layout(matrix(1:2,ncol=1))
> plot(Freq ~ Author,horizontal=T,main="Frequency of Three letter Words",col=c("purple","gold"))
> stripchart(Freq ~ Author,method="stack",pch=19,col=c("purple","gold"),ylim=c(0.5,2.5))
>
> # THE AUTHORS SEEM TO HAVE CONSIDERABLY DIFFERENT THREE WORD FREQUENCIES.
>
```

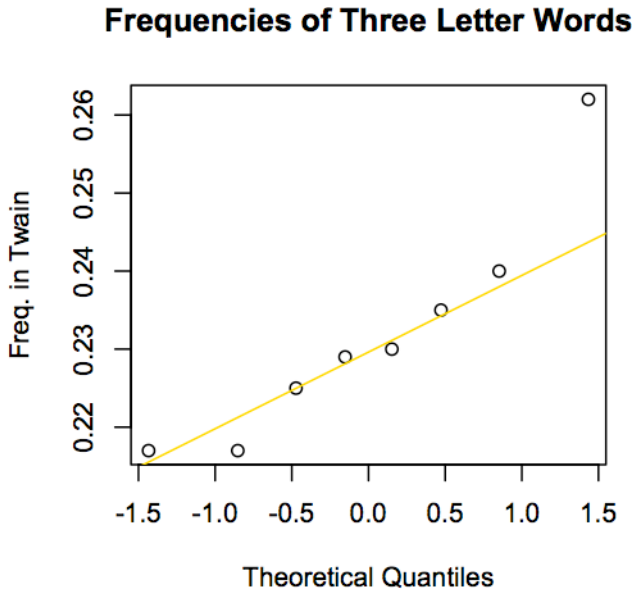




```
> ################ TEST IF VASIANCES DIFFER ############################
> var.test(FS,FT)

F test to compare two variances

data:  FS and FT
F = 0.44, num df = 9, denom df = 7, p-value = 0.2501
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.09123465 1.84688739
sample estimates:
ratio of variances
         0.4400445

> # Unable to reject huypothesis that varS/varT = 1
> # Thus, there is no significant evidence against the hypothesis that
> # sigma1 = sigma2.
>
> ###################### RUN CANNED POOLED T-TEST ########################
>
> help(t.test)
> t.test(FS,FT,var.equal=TRUE)

Two Sample t-test

data:  FS and FT
t = -3.8781, df = 16, p-value = 0.001334
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03429651 -0.01005349
sample estimates:
mean of x mean of y
 0.209700  0.231875

> # So it appears Twain and Snodgrass are not the same person:
> # Their three word frequemncies are significantly different,
> # even at the alpha = .005  level
>
> # Note that one of the hypotheses of the pooled two sample t-test
> # patently fails. The numbers are proportions, thus bounded between 0 and 1,
> # which are definitely nonnormal. However, the two sampled t-test is fairly
> # robust with regard to normality of the data. However, although the
> # f-test is recommended to alert us to possible inequality of variances,
> # the f-test is much more sensitive to nonnormality of the data.
>
```

```
> ################# QQ-PLOTS TO CHECK NEAR-NORMALITY OF THE DATA ##############
> layout(matrix(c(1,3,2,4),ncol=2))
> qqnorm(FS,ylab="Freq. in Snodgrass",main="Frequencies of Three Letter Words")
> qqline(FS,col="purple")
> qqnorm(FT,ylab="Freq. in Twain",main="Frequencies of Three Letter Words")
> qqline(FT,col="gold")
>
> # As close to normal as small n-size plots ever look. So no strong
> # indication of nonnormal.
>
```





```
> ############### POOLED T-TEST BY HAND ########################
>
> ns <- length(FS)
> nt <- length(FT)
> nu <- ns + nt - 2;nu
[1] 16
> fsbar <- mean(FS)
> ftbar <- mean(FT)
> vs <- var(FS)
> vt <- var(FT)
> sp <- sqrt(((ns-1)*vs+(nt-1)*vt)/nu)
> tp <- (fsbar-ftbar)/(sp*sqrt(1/ns+1/nt)); tp
[1] -3.878138
> alpha <- .05
> pvalue <- 2*pt(tp,nu); pvalue
[1] 0.001333821
> ta2 <- qt(alpha/2,nu,lower.tail=FALSE);ta2
[1] 2.119905
> CIhigh <-fsbar-ftbar+ta2*sp*sqrt(1/ns+1/nt)
> CIlow <-fsbar-ftbar-ta2*sp*sqrt(1/ns+1/nt)
> c(CIlow,CIhigh)
[1] -0.03429651 -0.01005349
> # So all numbers of canned test are reconstructed by hand!

> cat("\n\n Two Sample t-Test BY HAND\n\n",
+ "data: FreqSnodgrass and FreqTwain\n", "t =", tp, ", df =", nu,
+ ", p-value =", pvalue,
+ "\n", "alt. hyp.: true difference in means is not equal to 0\n",
+ 1-alpha, " confidence interval:\n", CIlow, CIhigh,
+ "\n", "sample estimates:\n",
+ "mean of FreqSnodgrass, mean of FreqTwain\n", fsbar, ftbar, "\n\n" )


 Two Sample t-Test BY HAND

 data: FreqSnodgrass and FreqTwain
 t = -3.878138 , df = 16 , p-value = 0.001333821
 alt. hyp.: true difference in means is not equal to 0
 0.95  confidence interval:
 -0.03429651 -0.01005349
 sample estimates:
 mean of FreqSnodgrass, mean of FreqTwain
 0.2097 0.231875
```