```r
#PO 841
#10/29/2014
#Lab 7: T-tests and Chi-squared tests

#First, let's load the libraries we plan to use:

library(foreign)
library(car)

#Let's load the 2012 National Election Study (the file is in SPSS format so we'll use the command
 read.spss):

NES<-read.spss(file.choose())

#Let's look at the feeling thermometer for Hillary Clinton:

summary(NES$ft_hclinton)

#Let's recode the variable so that all irrelevant values (i.e., the don't-knows, refused-to-
 answers, etc.) will be coded as NA:

hrcft<-ifelse(NES$ft_hclinton %in% c(0:100), NES$ft_hclinton, NA)
summary(hrcft)
length(hrcft)
mean(hrcft,na.rm=T)


# -----------------
# ONE-SAMPLE T-TEST
# -----------------

#Let's assume that an average feeling thermometer score 2/3 of the way from 0 to 100 -- i.e., an FT
 score of 66.67 -- is a benchmark for aspiring presidential candidates one election cycle prior to
 the cycle in which they plan to run. If candidates' FT scores are less than 66.67 four years out,
 they may be in trouble public opinion-wise. So let's test whether HRC's FT score in 2012 is
 significantly different from 66.67. In this case, of course, we do NOT know the population variance
 -- hence, we're going to conduct a t-test as opposed to a z-test. H0: mu=66.67 and H1: mu =/ 66.67

t.test(hrcft, mu=66.67, alternative='two.sided')

#Of course, if we had a theoretical reason to believe that HRC's FT score would be higher or lower
 than 2/3, we could specify a different hypothesis (e.g., if we worked for a Republican consulting
 firm that has good reason to think Hillary isn't quite as formidable as everyone says).


# -------------
# PAIRED T-TEST
# -------------

#What if we wanted to assess the difference between the feeling thermometer for HRC and the FT for
 one of HRC's presumed rivals in 2016, Joe Biden? First, let's set up the Biden variable, recoding
 all non-relevant values as NA:

summary(NES$ft_dvpc)
bidenft<-ifelse(NES$ft_dvpc %in% c(0:100), NES$ft_dvpc, NA)
summary(bidenft)
length(bidenft)
```

```r
#Since both questions were asked of all survey respondents, we can do a paired t-test; we simply provide the t.test() function with two vectors -- hrcft and bidenft -- rather than one. (If the vectors were not the same length and you tried to perform a paired t-test, R would give you an error message.) Since Biden has an (IMHO undeserved!) reputation for being a bit of a buffoon/gaffe factory, let's hypothesize that Biden's FT score will be significantly lower than HRC's.

t.test(bidenft, hrcft, paired=T,alternative='less') #Re: the question in class: Put Biden first because we're hypothesizing that Biden's FT score is less than HRC's (that is, that there is a net negative difference between Biden and HRC). Alternatively:

t.test(hrcft, bidenft, paired=T,alternative='greater') #If we're assuming there's a theoretical reason for thinking Biden's approval is higher than HRC's.

t.test(hrcft, bidenft, paired=T,alternative='two.sided') #Order of the variables doesn't matter because test is two-sided.

#Note: The t.test function is, in this case, only using data on those who provided a score for BOTH HRC and Biden. If a respondent said "don't know" or refused to answer (or whatever) only for Biden, and thus got an NA only for Biden, his entire response gets dropped from this test.

#We would also get the same answer by manually calculating the difference in scores and doing a one-sample test on whether the difference is equal to zero:

ft.diff<-bidenft-hrcft
t.test(ft.diff, mu=0, alternative='less')

#OR:

ft.diff<-hrcft-bidenft
t.test(ft.diff, mu=0, alternative='greater')


# --------------------------------------
# TWO-SAMPLE T-TEST WITH EQUAL VARIANCES
# --------------------------------------

#Let's say we want to see whether there is a significant difference in HRC's feeling thermometer scores among men and women.

#First, let's look at gender:

summary(NES$gender_respondent_x)

gender<-recode(NES$gender_respondent_x, '"2. female"="female"; "1. male"="male"')

summary(gender)
length(gender)

#Now let's look at HRC FT scores among men and among women. Recall that we created a new variable named hrcft:

summary(hrcft[gender=="male"])
male<-hrcft[gender=="male"] #Tells R to give us all the values for hrcft for which gender=male. (Because vectors hrcft and gender are the same length, R can easily pull out the values for hrcft that correspond to a selected gender value.)

summary(hrcft[gender=="female"])
female<-hrcft[gender=="female"]
```

```r
#Tells R to give us all the values for hrcft for which gender=female.

#As we can see, the samples of male and female respondents are different lengths, which sets us up to perform a two-sample t-test:

length(male)
length(female)

#If we want to do a two-sample t-test -- in this case, using subsamples of the survey -- we need to provide the t.test() function with the two different vectors we examined above. We also need to specify whether we want to assume equal variance or not -- in this case, it is a fair assumption that the samples of male and female responses (from the same survey) will have equal variances. H0: mu=0; H1: mu=/0

t.test(male, female, paired=F, var.equal=T, alternative='two.sided')


# ---------------------------------------
# TWO-SAMPLE T-TEST WITH UNEQUAL VARIANCES
# ---------------------------------------

#In what scenario might we expect between-group differences in population variances? Well, it makes sense that the variance of FT scores for HRC might be different for those people who were paying very close attention to the 2012 election versus those who paid very little attention to the election (those who paid a lot of attention might hold altogether stronger opinions in both directions, pro and con).

#Let's examine interest in the 2012 campaign:

summary(NES$interest_attention)
payattention<-recode(NES$interest_attention,'"1. always"="5"; "2. most of the time"="4"; "3. about half the time"="3"; "4. some of the time"="2"; "5. never"="1"; else=NA')
summary(payattention)

#Let's now assume that scores of 4 and 5 ("always" and "most of the time") denote high levels of interest in the campaign; and that scores of 1 and 2 ("never" and "sometimes") denote low levels of interest. Everything else (including those who pay attetion about half of the time) we can classify as NA:

interest<-recode(payattention, '5="high"; 4="high"; 2="low"; 1="low"; else=NA')
summary(interest)

#Test the null hypothesis that there is no difference in HRC FT scores between high-information and low-information voters against the alternative that there IS a difference:

t.test(hrcft[interest=="high"], hrcft[interest=="low"], paired=F, var.equal=F,
alternative='two.sided')

#If we had assumed equal variances, we would have gotten a slightly lower t-score (and slightly higher p-value), though this would not have changed the outcome:

t.test(hrcft[interest=="high"], hrcft[interest=="low"], paired=F, var.equal=T,
alternative='two.sided')


# -----------------
# CHI-SQUARED TESTS
# -----------------
```

```
#Remember that a chi-squared test allows us to test for a relationship between categorical
 variables. Specifically, the test tells us how likely it is that observed differences in
 frequencies between particular categories came about by chance. Our H0 here would be that there is
 no significant difference in frequencies between categories and therefore no relationship between
 the specified variables, while our H1 would be that there IS a significant difference in
 frequencies between catgeories, and thus a significant relationship between the variables under
 consideration. Let's look at the chi-squared distribution for a moment:

hist(rchisq(10000, df=10), xlab="Chi-squared statistic", probability=T)
curve(dchisq(x, df=10), add=T, col="red")

#To perform a chi-squared test, let's refer to the Afrobarometer data from the previous lab.

afro<-read.dta(file.choose(),convert.factors=F)
ghana<-afro[afro$COUNTRY==2,] #Check codebook; Ghana=2
summary(ghana)

#Recall the examples from Monday's class and from the previous lab using Party ID. We'll recode the
 Party ID variable again, except this time we won't include an "other" category and will make the
 values that we previously coded as "other" equal to NA. (Chi-squared tests encounter problems when
 there is a category with only a small number of observations.)

partyID<-recode(ghana$Q87A, '0="None"; 120="NPP"; 121="NDC"; else=NA')
table(partyID)

#Last time we looked at whether there appeared to be relationships between income and party ID, and
 region and party ID. Now, let's examine whether or not there seems to be a relationship between
 whether or not the respondent is a government worker and the respondent's party ID:

govworker<-ghana$Q91==1
tab1<-table(govworker,partyID)
prop.table(tab1,margin=1) #Remember, margin=1 means display proportions by row.

#Just by eyeballing this table, there doesn't seem to be a strong relationship between the two
 variables. How can we know for sure? We can perform a chi-squared test! The function syntax is
 simple; just list the two variables you would use to make a table (order doesn't matter).

#Our H0 is that there is no relationship between being a government worker and party ID (i.e., no
 difference in frequencies between "government worker" and "party ID") ); our alternative H1 is that
 there IS a relationship between the two variables.

chisq.test(govworker,partyID)

#Last week we suspected by looking at our table comparing region with party ID that there is a
 relationship between which region of Ghana respondents live in and their party ID. We can use the
 chi-squared test to confirm that we were right; there IS a relationship between region and party
 ID.

region<-recode(ghana$REGION,'120="Ashanti"; 128="Volta"; else="Other"')
tab2<-table(region,partyID)
prop.table(tab2,margin=1)

chisq.test(region,partyID)
```