LAB 4

STAT 131A

Welcome to the lab 4. In this lab, you will

1) implement a permutation test;

2) implement a T-test;

3) illustrate the difference between a T-distribution and the standard normal distribution;

We will continue to use the rent price dataset from the lab 2.

Read in data:

```r
craigslist <- read.csv("craigslist.csv",
                       header = TRUE)
```

Replicate

To repeat things in R, there is a more convenient and efficient way than using for loop. The function `replicate` is designed for evaluating an expression repeatedly. For example, to get a vector of length 50 where each element is generated independently from the sum of five normal distributed samples:

```r
set.seed(20170209)
example1 <- replicate(50, sum(rnorm(5)))
```

The equivalence of the above code using for loops will be:

```r
# create a vector of length 0 to store the results
example2 <- c()
# calculate for 50 times
for (i in 1:50){
  # concatenate the calculated sum of normal samples to the end of `results` vector
  example2 <- c(example2, sum(rnorm(5)))
}
```

The first argument of `replicate` function is the number of replications. And the second argument is an expression which will be evaluated repeatedly, which usually involves random sampling and simulation. (Otherwise, you would get a vector of one identical number. That's useless!) Sometimes, the things you want to replicate can not finish on one line, and then you may need to use the big bracket such as the following. Elements in the `example3` vector would be the value of the last expression in the bracket. (Since we set the same seed for `example1` and `example3`, you may want to compare the value of them to see what happened.)

```r
set.seed(20170209)
example3 <- replicate(50, {
  a <- rnorm(5)
  b <- 5
  sum(a) + b      # return the value
})
```

Permutation test

Exercise 1

You will do a permutation test to compare the average one bedroom apartment rent price in Berkeley and Palo Alto.

   (a) Subset the dataset to only consider one bedroom apartments in Berkeley and Palo Alto.

```
# insert code here save the data frame of one bedroom postings
# in Berkeley and Palo Alto as
# 'subset'

subset = NULL
```

   (b) Calculate the number of postings for Berkeley in the data frame `subset`.

```
# insert code here save the number of postings for Berkeley as
# 'no.berkeley'

no.berkeley = NULL
```

   (c) Calculate the observed statistics, i.e., the absolute difference between the mean of Berkeley and mean of Palo Alto one bedroom rent price.

```
# insert code here save the observed statistics as
# 'stat.obs'

stat.obs <- NULL
```

(d) Calculate the permuted statistics (absolte mean difference), repeat for 1000 times. HINT: use `sample` function to sample from a group of observations. You can use either for loop or the `replicate` function introduced above. However, it is a good practive to try `replicate`.

```r
# leave seed set as is.
# insert code after set seed to save the simulated statistics as
# 'stat.bootstrap'
set.seed(20172828)

stat.bootstrap <- NULL
```

(e) Calculate the p-value of our observed statistic using our approximation of the sampling distribution from part (d). We are testing if there is any difference between the two means. Our null hypothesis is that there is no difference between the two means. Our alternative hypothesis is that there is a difference.

```r
# insert code here save the observed statistics as
# 'p.value'

p.value <- NULL
```

T-test

With the function `t.test`, implementing t-tests is easy in R. We will illustrate using a simulated example:

```r
# Generate 100 samples from N(0, 1)
group1 <- rnorm(100, mean = 0, sd = 1)
# Generate 100 samples from N(0.2, 1)
group2 <- rnorm(100, mean = 0.2, sd = 1)
# perform t test
t.test(group1, group2)
```

```
##
##  Welch Two Sample t-test
##
## data:  group1 and group2
## t = -1.0893, df = 197.4, p-value = 0.2774
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4252694  0.1226339
## sample estimates:
##    mean of x    mean of y
## 0.008502307 0.159820020
```

`t.test` usually print the results. To obtain the value of the t statistics or p-value, it is generally not wise to copy and paste from the printed output. Imagine you need to do 1000 t-test simultaneously, it is impossible to copy and paste every time. We need to figure out the output of `t.test`:

```r
ttest.result <- t.test(group1, group2)
```

If you check the `ttest.result` object in your Environment window, you will find that it is a list of 9. And each element in the list stores some information.

```r
names(ttest.result)
```

```
##  [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
##  [6] "null.value"  "stderr"      "alternative" "method"      "data.name"
```

To get the p-value, you can use either dollar sign or double square bracket, which are two ways usually used to extract elements from lists.

```r
ttest.result$p.value     # with dollar sign, you can do tab completion
```

```
## [1] 0.2773637
```

```r
ttest.result[['p.value']]
```

```
## [1] 0.2773637
```

Exercise 2.

Now apply t-test to compare the mean rent of one bedroom listings in Palo Alto and Berkeley.

```r
# insert code here and save the t-statistic as
# 'rent.1b.tstat'

rent.1b.tstat <- NULL
```

```r
# insert code here and save the p value as
# 'rent.1b.pvalue'

rent.1b.pvalue <- NULL
```

T-distribution vs Normal distribution

The T-distribution has mean = 0 and variance = n/(n-2). It is parameterized by just one parameter, the degrees of freedom. The degrees of freedom for practical purposes is equal to the sample size minus 1 for a one sample T-test. For a two sample T-test, as done above, the degrees of freedom is calculated through a formula (that you don't need to know).

The T-distribution is similar in shape to the standard normal distribution, but the relative difference in the tail regions of the distributions is significant. This implies that for hypothesis tests, the p-value we find will be quite different if we use the wrong distribution as our sampling distribution for the test statistic.

The following code finds the tail probability for a T-score (or T-statistic / T-value) of -1, with 30 degrees of freedom.

```
pt(-1, df=30)
```

```
## [1] 0.1626543
```

We can see that this different from that of a Z-score (standard normal distribution) of -1

```
pnorm(-1)
```

```
## [1] 0.1586553
```

The difference is larger when we change the degrees of freedom to 10

```
pt(-1, df=10)
```

```
## [1] 0.1704466
```

Exercise 3.

    (a) Find the tail probability of a T-score of -2 with 30 degrees of freedom. Divide this by the tail probability of a Z-score of -2.

```
#insert code here
#save final answer (ratio) as e3a

e3a = NULL
```

    (b) Repeat part (a) but for 10 degrees of freedom.

```
#insert code here
#save final answer (ratio) as e3b

e3b = NULL
```

It is common to run a hypothesis test with a significance level of 5 percent, or 0.05. This corresponds to rejecting the null hypothesis if we are more than 2 standard errors from the mean, or a Z-score of less than -2 or more than 2. However, if we do not know the true standard error, and we use the sample standard deviation to estimate it, then we must use the T-distribution. You can see from above that in this case our true P-value is much greater than we would otherwise think it is under the standard normal distribution. This is the effect of the uncertainty that comes from not knowing the true population standard deviation.