Group Work on Tests and Intervals

Please write the NetIDs of your group members:

- Name: Ying Mufang ; NetID: mying4
- Name: Cai Naiqing ; NetID: ncai5
- Name: Meng Yuhan ; NetID: meng46
- Name: Pan Hongwei ; NetID: hpan55
- Name: Zhang Xueqian ; NetID: xzhang2278
- Name: Li Zihao ; NetID: zli873
- Name: Wei Haoxiang ; NetID: hwei64
- Name: Zhai Yibo ; NetID: yzhai28

Revise this tests.Rmd (tests.Rmd) file to answer the questions below. Include reasonable labels (titles, axis labels, legends, etc.) with each graph. Please use **boldface** (by enclosing text in `** ... **`) when writing your answers so that we can find them easily.

At the end of class, one person from each group should submit a completed copy of this file. (Please don't submit multiple copies.)

We'll grade this exercise by opening your "tests.Rmd" file, clicking "Knit HTML", and reading the output. We'll assign points as follows:

- No submission: 0/5
- Poor work: 3/5
- Good work: 5/5 (even if incomplete–there's more to do here than can be done in 75 minutes)

John Wayne

(From p. 166 of Whitlock & Schulter's *The Analysis of Biological Data*)

In 1956, John Wayne played Genghis Khan in a movie called *The Conqueror*. The movie was filmed downwind of 11 above-ground nuclear bomb tests. Of the 220 people who worked on this movie, 91 had been diagnosed with cancer by the early 1980s, including Wayne, his co-stars, and the director. (Wayne died of cancer in 1979.) According to large-scale epidemiological data, only about 14% of people of this age group, on average, should ahve been stricken with cancer within this time frame. Are these data strong evidence for an increased cancer risk for people associated with this film? Run and analyze an appropriate test.

Hypothesis:

H0: p=0.14

H1: p>0.14

Parameter p means the proportion of people in this age group who died from cancer to the 220 people who worked on this movie

Run the proportion test

```
x=91 # number of success
n=220 # number of total trials
out=prop.test(x, n, p = 0.14, alternative = "greater", conf.level = 0.95) # one-sided
 proportion test
cat("X-squared is:", out$statistic, "\n")
```

```
## X-squared is: 134.5549
```

```
cat("p.value is:", out$p.value, "\n")
```

```
## p.value is: 2.065698e-31
```

```
cat("Confidence Interval is:[", out$conf.int[1],",", out$conf.int[2], "]", "\n")
```

```
## Confidence Interval is:[ 0.3581924 , 1 ]
```

Since p.value is: 2.066e-31<0.05, we should reject the null hypothesis. Thus these data are strong evidence for an increased cancer risk for people associated with this film.

Repetitive Stress Injuries

(From p. 367 of Devore's *Probability and Statistics for Engineering and the Sciences*)

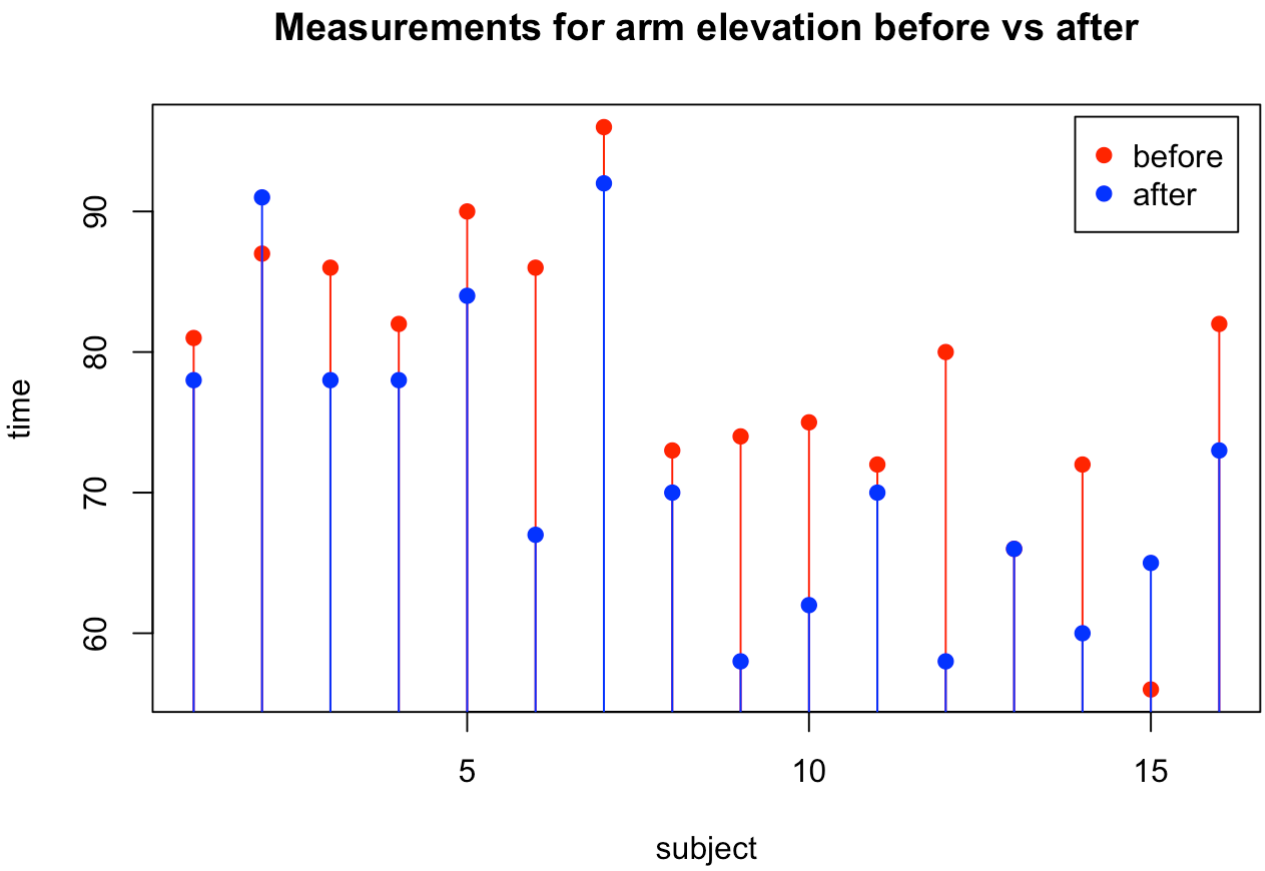
Musculoskeletal neck-and-shoulder disorders are all too common among office staff who perform repetitive tasks using computers. The article "Upper-Arm Elevation During Office Work" (Ergonomics, 1996: 1221-1230) reported on a study to determine whether more varied work conditions would have any impact on arm movement. The accompanying data was obtained from a sample of $n = 16$ subjects. Each observation is the amount of time, expressed as a proportion of total time observed, during which arm elevation was below 30°. The two measurements from each subject were obtained 18 months apart. During this period, work conditions were changed, and subjects were allowed to engage in a wider variety of work tasks.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Before	81	87	86	82	90	86	96	73
After	78	91	78	78	84	67	92	70

Subject	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Before	74	75	72	80	66	72	56	82
After	58	62	70	58	66	60	65	73

Make a suitable graph of these data.

```
data1=data.frame(x1=c(81,87,86,82,90,86,96,73,74,75,72,80,66,72,56,82),y1=c(78,91,78,
78,84,67,92,70,58,62,70,58,66,60,65,73))
plot(x=c(1:16),data1$x1,col="red",type="h",main="Measurements for arm elevation befor
e vs after",xlab="subject",ylab="time")
points(x=c(1:16),data1$x1,col="red",pch=19)
lines(x=c(1:16),data1$y1,col="blue",type="h")
points(x=c(1:16),data1$y1,col="blue",pch=19)
legend("topright",inset=.02,legend=c("before","after"),pch=c(19,19),col=c("red","blu
e"))
```



Do the data suggest that true average time during which elevation is below 30˚ differs after the change from what it was before the change? Run and analyze an appropriate test.

The hypothesis :

$H_0 : \mu_d = 0 \quad vs \quad H_1 : \mu_d \neq 0$,**where** μ_d **represent the difference between the average time during which elevation is below 30˚ before the change and after the change respectively.**

```
data1=data.frame(x1=c(81,87,86,82,90,86,96,73,74,75,72,80,66,72,56,82),
                 y1=c(78,91,78,78,84,67,92,70,58,62,70,58,66,60,65,73))
diff=data1$x1-data1$y1
t.test(diff)
```

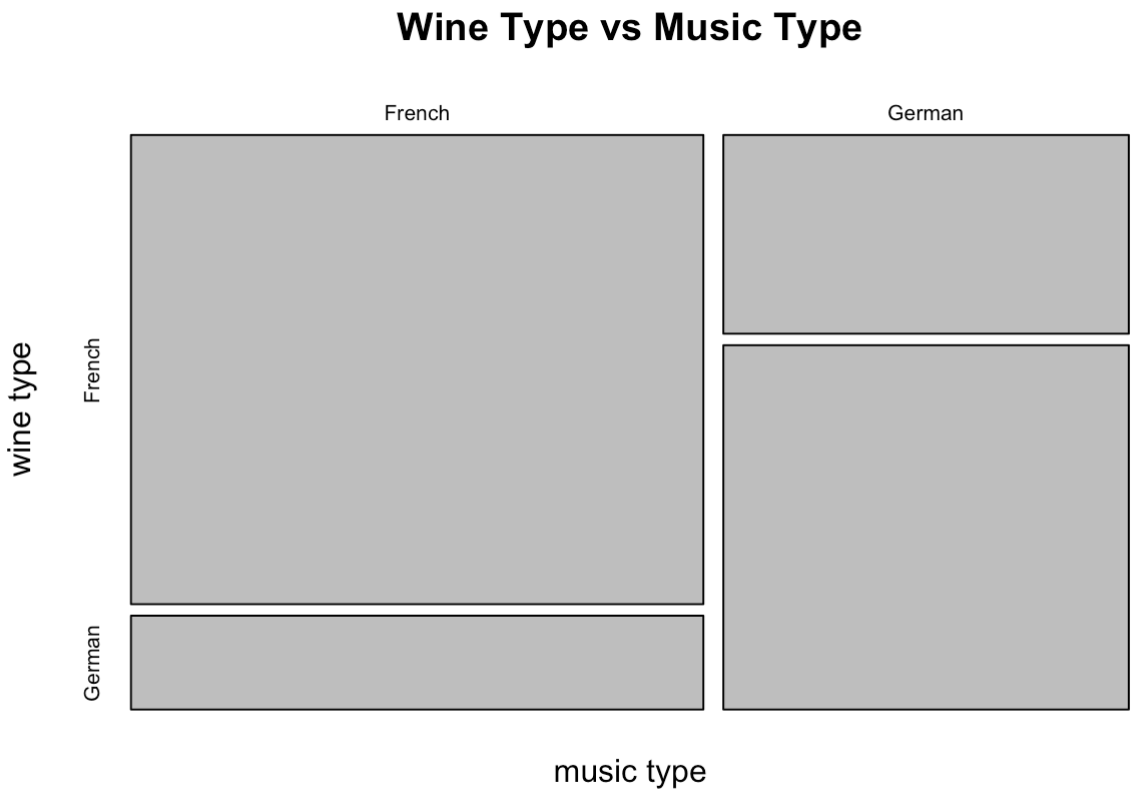
```
##
##  One Sample t-test
##
## data:  diff
## t = 3.2791, df = 15, p-value = 0.005072
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   2.362371 11.137629
## sample estimates:
## mean of x
##      6.75
```

From the result of Paired t-test,we can see that the p-value is 0.005072,**which is smaller than 0.05. Therefore we could reject** H_0.**Statistically,the true average time during which elevation isbelow 30˚ might differ after the change from what it was before the change under significant level 0.05.**

Influence of music on wine purchases

The file wine.csv (http://www.stat.wisc.edu/~jgillett/327-1/graphics/wine.csv) contains data on wine purchased from a liquor store. Each row corresponds to a bottle of wine purchased. The first column indicates which type of music was playing in the store during the purchase. The second column indicates which type of wine was purchased. Paste your graph from the group work on graphics that gives evidence about the question of whether type of music and type of wine are independent.

```
data=read.csv("wine.csv",header=F)
counts = table(data)
mosaicplot(counts, xlab='music type', ylab='wine type',main="Wine Type vs Music Type"
)
```



Based on the graph, four blocks' sizes are not similar. Therefore they are not independent.

Are these data strong evidence that music played and wine purchased are not independent? Run and analyze an appropriate test.

Hypothesis:

H0:row and column variables are independent

H1:row and column variables are not independent

```
chitest=chisq.test(counts)
cat("The p-value of the chi-square test is",chitest$p.value)
```

```
## The p-value of the chi-square test is 2.478964e-05
```

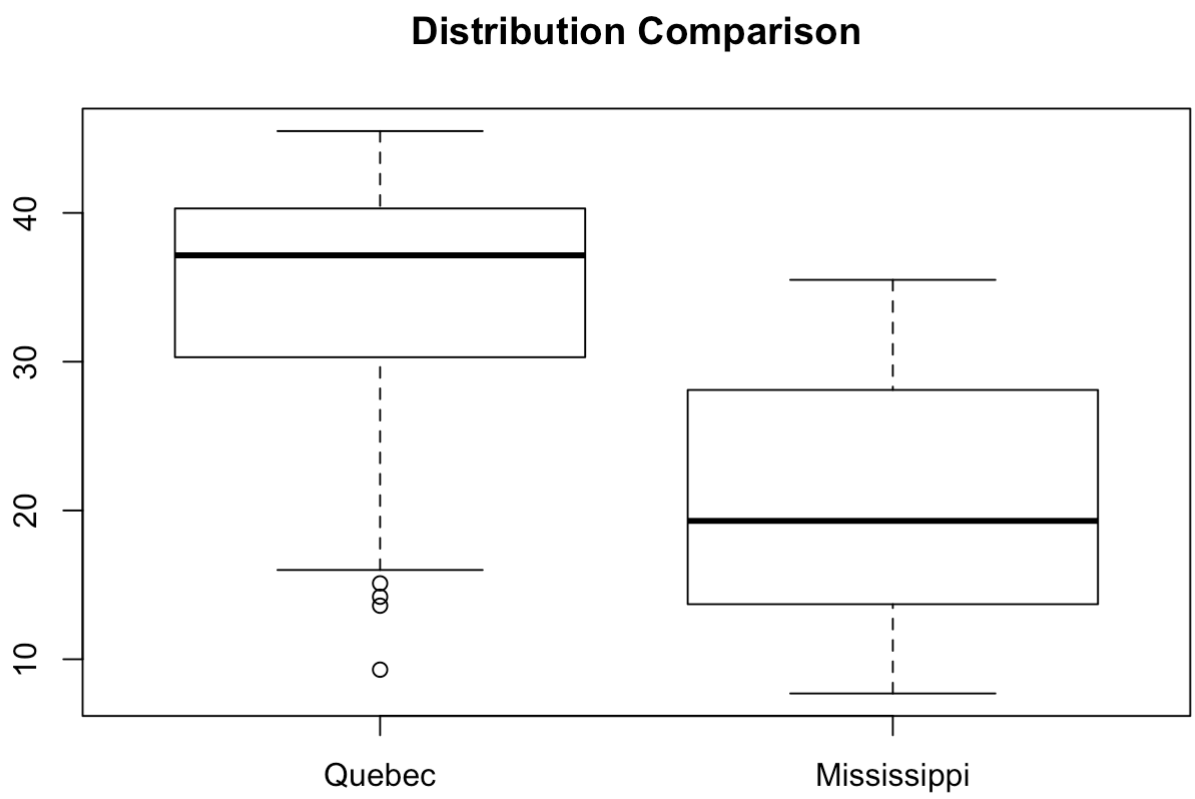
Since the p-value of Chi-square test equals 2.479e-05<0.05, we could reject the null hypothesis that row and column variables are independent. Thus, these data are strong evidence that music played and wine purchased are not independent.

Plants

Look at the built-in data frame `CO2` . Describe the data set (in English and the kind of language used in an introductory statistics course, not in R language). Mention whether it's from an experiment or an observational study and mention which are independent/explanatory variables and which are dependent/response variables.

Make a graph that helps with comparing the distribution of uptake for the two grasses, Quebec and Mississippi.

```
boxplot(uptake~Type, data = CO2,main = "Distribution Comparison")
```



The CO2 data comes from an experiment. And uptake is the dependent variable and others are independent variables.

Are these data strong evidence that the population mean uptake for Quebec grass is different from the population mean uptake for Mississippi grass? Run and analyze an appropriate test.

```
t.test(CO2$uptake~CO2$Type)
```

```
##
##  Welch Two Sample t-test
##
## data:  CO2$uptake by CO2$Type
## t = 6.5969, df = 78.533, p-value = 4.451e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.839475 16.479572
## sample estimates:
##      mean in group Quebec mean in group Mississippi
##                  33.54286                  20.88333
```

According to the T-test, with 95% confidence, true difference in means is not equal.

Find a 99% confidence interval for the difference between the true mean uptake for Quebec grass and the true mean uptake for Mississippi grass.

```
t.test(CO2$uptake~CO2$Type,conf.level = 0.99)$conf.int
```

```
## [1]  7.593543 17.725505
## attr(,"conf.level")
## [1] 0.99
```