q2 example

Daniel Detore

```r
VotingSurvey = data.frame(read.csv("VotingSurvey.csv"))
VotingSurvey$Vote = as.factor(VotingSurvey$Vote)
VotingSurvey$Gender = as.factor(VotingSurvey$Gender)
VotingSurvey$Party = as.factor(VotingSurvey$Party)
```

1

1.1

H_0 : 'Vote' and 'Gender' have no statistical association.

H_a : 'Vote' and 'Gender' have some statistical association.

In this case the statistical association is whether voter's genders make them more (or less) likely to vote.

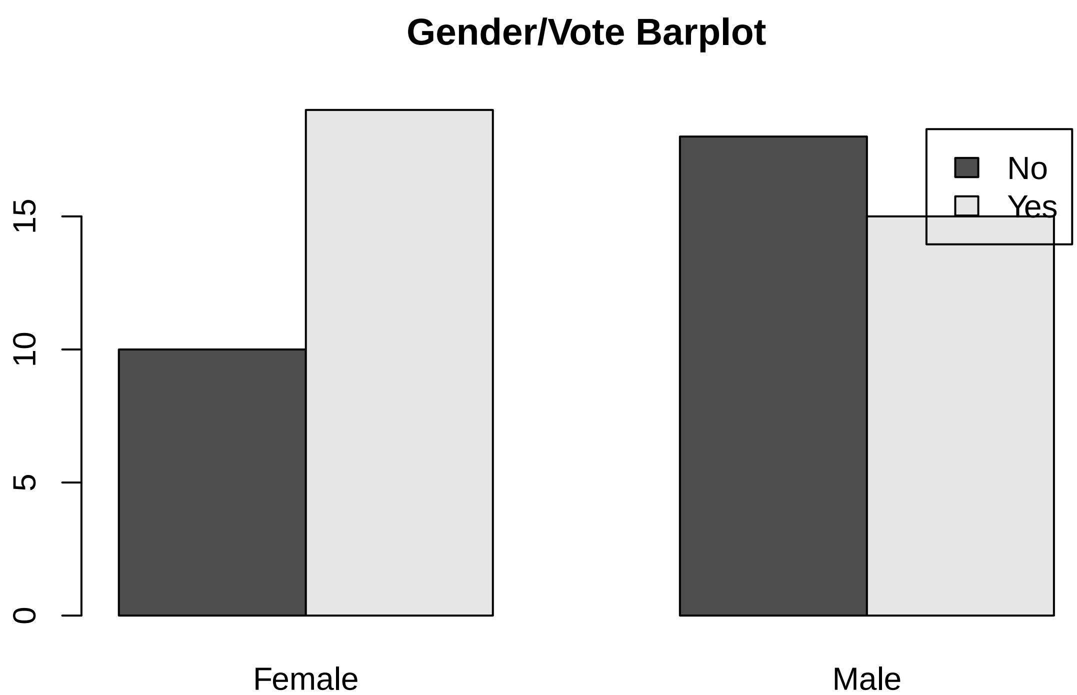
1.2

```r
table = table(VotingSurvey$Vote, VotingSurvey$Gender); table
```

```
##
##        Female Male
##   No       10   18
##   Yes      19   15
```

1.3

```r
barplot(table, legend=TRUE, beside=TRUE, main='Gender/Vote Barplot')
```



They are of different patterns. Females seem much more likely to vote and men are slightly less likely, but we should still test to see if these differences are statistically relevant.

1.4

The value of each inner cell $e_{i,j} = \frac{o_{i,+}o_{+,j}}{o_{+,+}}$.

```
o_ip <- rowSums(table)
o_pj <- colSums(table)
o_pp <- sum(table)
e <- outer(o_ip, o_pj)/o_pp
rownames(e) = c("No", "Yes")
colnames(e) = c("Female", "Male")
e
```

```
##       Female     Male
## No  13.09677 14.90323
## Yes 15.90323 18.09677
```

1.5

```
chisq.test(table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 1.764, df = 1, p-value = 0.1841
```

The testing statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$$

is observed as

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{i,j}-e_{i,j})^2}{e_{i,j}} = 1.76$$

The null distribution is χ_k^2 where degrees of freedom $k=(r-1)(c-1)=(2-1)(2-1)=1$. The p-value is approximately 0.18.

If we take $\alpha = 0.1$ then we cannot reject H_0 because $\alpha < 0.18$, thus we must assume 'Vote' and 'Gender' have no statistical association.

2

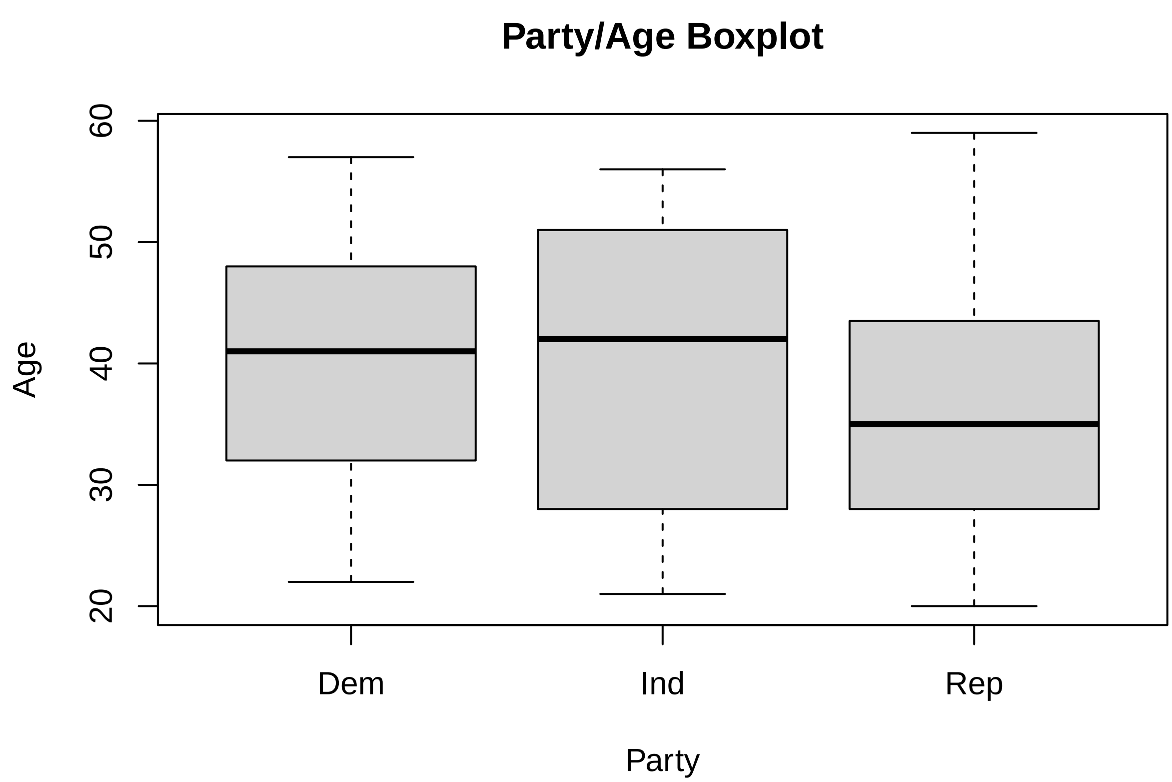
2.1

$H_0: \ \mu_1 = \mu_2 = \mu_3$; i.e. the mean age of all parties are equal.

H_a : The mean age of all parties are not all equal.

2.2

```
boxplot(xlab = "Party", ylab = "Age", VotingSurvey$Age ~ VotingSurvey$Party,
beside = TRUE, main = "Party/Age Boxplot")
```



The variances in age seem similar for all parties, but the mean for Rep looks slightly lower than the others.

2.3

```r
report = aov(Age ~ Party, VotingSurvey)
summary(report)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Party        2    127    63.4   0.504  0.607
## Residuals   59   7429   125.9
```

2.4

$$SSB = \sum_{i=1}^{k} n_i(\bar{X}_{i,.} - \bar{X}_{.,.})^2 = 127$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i,.})^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2 = 7429$$

$$SST = SSB + SSE = 7556$$

$$MSB = \frac{SSB}{k-1} = 63.4$$

$$MSE = \frac{SSE}{n-k} = 125.9$$

2.5

Our testing statistic $F = \frac{MSB}{MSE} = \frac{\frac{SSB}{k-1}}{\frac{SSE}{n-k}}$ is observed as $f = 0.504$. The null distribution is $\mathcal{F}_{k-1,n-k} = \mathcal{F}_{3-1,62-3} = \mathcal{F}_{2,59}$. The p-value is 0.607.

2.6

```r
summary(lm(Age ~ Party, data=VotingSurvey))$r.squared
```

```
## [1] 0.01678316
```

Since $\alpha = 0.1 < 0.607$, we cannot reject H_0, thus we must assume the mean age of all parties are equal.

2.7

Using the box plot from 2.2, we may wat to check if the mean age for Rep party is lower than the average mean age of Dem and Ind parties, i.e. whether $\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \mu_3 = 0$.

2.8

```r
contrasts(VotingSurvey$Party) = c(1/2, 1/2, -1)
go = aov(Age ~ Party, VotingSurvey)
summary(go, split = list(Party = list(`mu3 vs 1/2 mu1 + 1/2 mu2` = 1)))
```

```
##                                 Df Sum Sq Mean Sq F value Pr(>F)
## Party                            2    127    63.4   0.504  0.607
##   Party: mu3 vs 1/2 mu1 + 1/2 mu2  1    126   125.5   0.997  0.322
## Residuals                       59   7429   125.9
```

We have observed testing statistic $f = 0.997$ which gives p-value 0.322. Since the p-value $0.322 > \alpha = 0.1$, we cannot reject H_0 and thus we must that the mean age of the Rep party is equal to the average mean ages of the Ind and Dem parties.

2.9

We can do multiple comparison using Fischer's Least Significant Difference.

```r
library(agricolae)
```

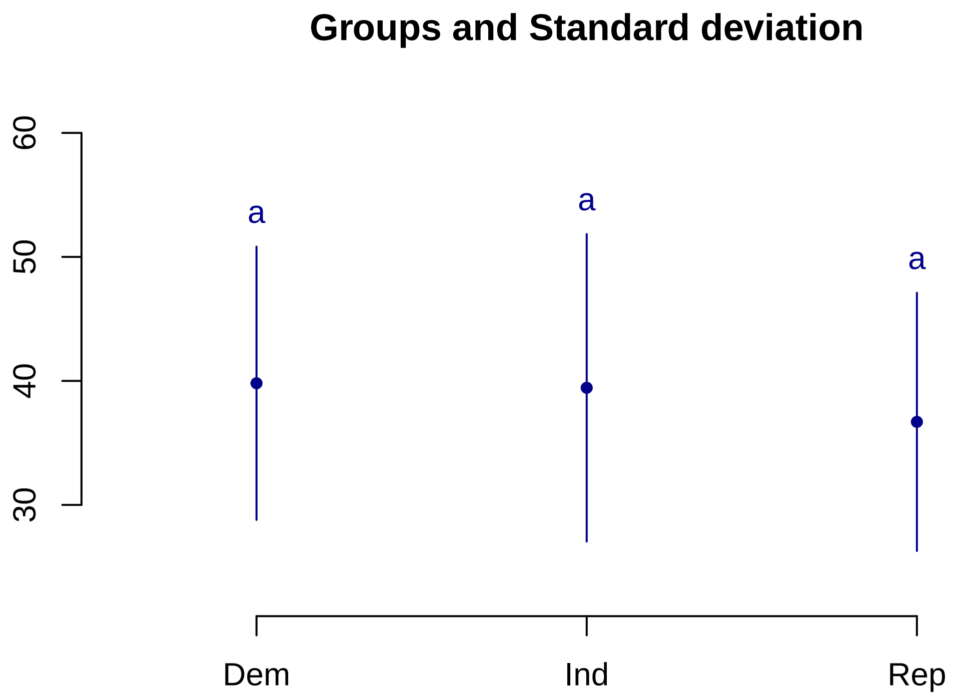
```
## Warning: package 'agricolae' was built under R version 4.4.3
```

```r
comparison = LSD.test(go, "Party", p.adj = "none")
comparison
```

```
## $statistics
##    MSerror Df     Mean      CV
##   125.9077 59 38.54839 29.1085
##
## $parameters
##         test p.ajusted name.t ntr alpha
##   Fisher-LSD      none  Party   3  0.05
##
## $means
##          Age      std  r       se      LCL      UCL Min Max  Q25 Q50  Q75
## Dem 39.80952 11.02098 21 2.448592 34.90990 44.70915  22  57 32.0  41 48.0
## Ind 39.44444 12.40124 18 2.644782 34.15225 44.73664  21  56 28.5  42 50.0
## Rep 36.69565 10.41168 23 2.339711 32.01390 41.37740  20  59 28.0  35 43.5
##
## $comparison
## NULL
##
## $groups
##          Age groups
## Dem 39.80952      a
## Ind 39.44444      a
## Rep 36.69565      a
##
## attr(,"class")
## [1] "group"
```

And we can visualize the comparison as such:

```r
plot(comparison, variation = "SD")
```



Because these functions put all of the parties in the "a" group, we can conclude that all parties' actual mean ages are equal.