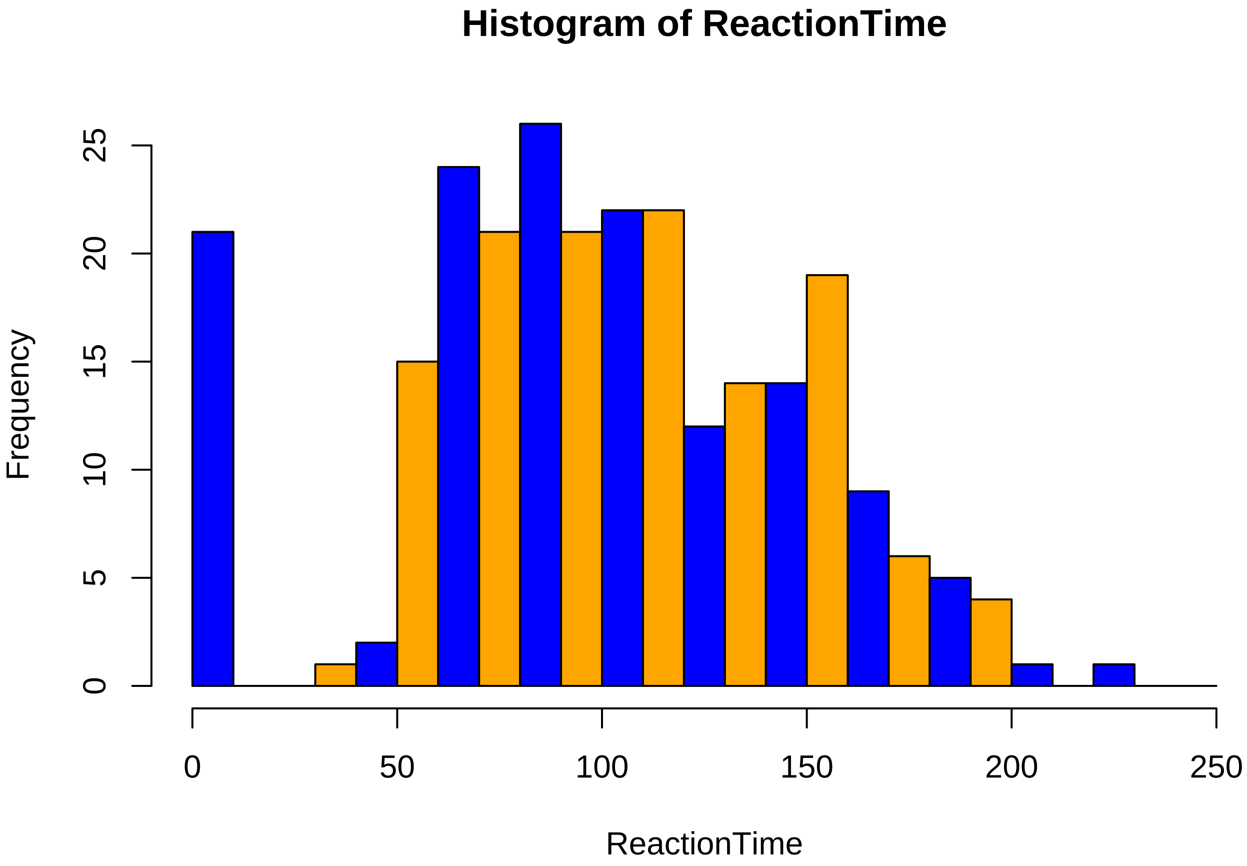
# Stat 311: HW 9, due Th 5/27/10 in your Quiz Section

Fritz Scholz

Your returned assignment should show your name and student ID number. It should be printed or written clearly.

1. The data set `ReactionTime` contains the reaction times of emergency medical teams (EMT) for a particular shift of a particular Shoreline Fire Station. You can find the data set at this homework link. Download it to the directory from which you start your R session. Read this data set into R via `ReactionTime <- scan("ReactionTime")`. The reaction time (measured in seconds) is the time elapsed between the emergency call being issued to the response teams and the time that they respond: we are on our way, i.e., moving in their vehicle.

a) Make a histogram of the data vector `ReactionTime` and comment on whether it looks normal or not. Try to explain any abnormality by the possibility that sometimes EMTs receive their calls while being on the road? Try `hist(ReactionTime,col=c("blue","orange"),breaks=seq(0,250,10))` and show your plot.



Even without the spike on the left the histogram seems skewed to the right, i.e., shows features that are not expected for a normal sample. The 22 lowest reaction times are

```
 [1]   0   0   0   0   0   0   0   1   1   2   2   2   4   4   5   5   6   6
 [19]   8  10  10  38
```

and it appears that the first 21 are the result of rather immediate response while the EMT was out on the road already.

b) The Department has set a goal that the population median response time $q_2$ should not exceed 2 minutes. Does this sample of 260 response times support this goal? Decide this issue by testing $H_0 : q_2 \geq 120$ seconds versus $H_1 : q_2 < 120$ seconds at level $\alpha = 0.01$. (The Department wants to be strongly convinced.) Use the sign test and deal with the issue of zeros by deleting the zero cases (see slides 19-20 for Chapter 10).

Denote the $i^{\text{th}}$ reaction time by $X_i$. Using the sign test we would reject $H_0$ when $Y = \#\{X_i - 120 > 0\}$ is too small. We have 3 cases with $X_i = 120$. After deleting these we have 257 values left and we observe $Y = y = 85$, with significance probability

$$\texttt{pbinom}(85, 257, 0.5) = \texttt{3.065703e} - \texttt{08} \ll 0.01 \quad \text{highly significant}$$

c) Construct a 99% confidence interval for $q_2$, pretending that there are no ties among the data. Try to make the actual confidence level as close as possible to .99. Which specific order statistics should be used to form the interval (assuming that they are all distinct)? Then employ the conservative step of dealing with ties and give the corresponding interval in numerical form.

```
> qbinom(.005,260,0.5)
[1] 109
> pbinom(109,260,0.5)
[1] 0.005429491
> pbinom(108,260,0.5)
[1] 0.003769801
```

Clearly $c_\alpha = 109$ will achieve a confidence level closest to 0.99, namely $\texttt{1} - \texttt{2} * \texttt{0.005429491} = \texttt{0.989141}$. When there are no ties our 98.9% confidence interval for $q_2$ would be $[X_{(c_\alpha+1)}, X_{(n-c_\alpha)}] = [X_{(110)}, X_{(151)}] = [90, 110]$. Since measurement accuracy is limited to seconds, we should widen the above interval by half a second at each end and report $[89.5, 110.5]$ as our confidence interval with coverage probability slightly higher than $\texttt{0.989141}$.

2. Text, 11.4, Problem Set A, 1.

| $\mu$ | $\mu_0$ | $\bar{x}$ | $s^2$ | $t$ | $\alpha$ | $\Delta$ | $\Delta_0$ | $\mathbf{p}$ | $\hat{\nu}$ |
|---|---|---|---|---|---|---|---|---|---|
| i. | iii. | ii. | ii. | ii. | iii. | i. | iii. | ii. | ii. |

3. Text, 11.4, Problem Set A, 2. The value 1.76 in (b) may not be the same as what you calculate in (a). For (b) you may want to write yourself an R function

```
Welchdf <- function(n1,n2,s1,s2){
# your commands for calculating the dregrees of freedom nu
# for the Welch t statistic, and return nu as output
...
nu
}
```

(a) Since none of the parameters are known, in particular it is not known whether $\sigma_1 = \sigma_2$ and $s_1 = 4.09$ and $s_2 = 1.22$ point away from that, we should use the Welch $t$ statistic

$$t_W = \frac{\hat{\Delta} - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{-0.82 - 1.39}{\sqrt{4.09^2/10 + 1.22^2/20}} = -1.671927$$

We should reject $H_0$ when $t_W$ is too low.

(b) To calculate the appropriate degrees of freedom for the approximate null distribution of $t_W$ I wrote the following function

```
Welchdf <- function(n1,n2,s1,s2){
r1 <- s1^2/n1
r2 <- s2^2/n2
nu <- (r1+r2)^2/(r1^2/(n1-1)+r2^2/(n2-1))
nu
}
```

and `Welchdf(10,20,4.09, 1.22)` returned $\overset{\frown}{9}.8094 \approx 10$.

```
> 2*pnorm(-1.76)
[1] 0.0784078
> pt(-1.76,df=28)
[1] 0.04466506
> pt(1.76,df=10)
[1] 0.9455482
> pt(-1.76,df=10)
[1] 0.05445183
> 2*pt(1.76,df=28)
[1] 1.91067
> pt(-1.76,df=9.8094)
[1] 0.05474582
```

Clearly `pt(-1.76,df=10)=0.05445183` comes closest to `pt(-1.76,df=9.8094)=0.05474582`.

(c) Since $\mathbf{p} = 0.96 \gg \alpha = 0.05$ we have no reason to reject $H_0$.

4. Text, 12.6, Problem Set A, 1. and 2.

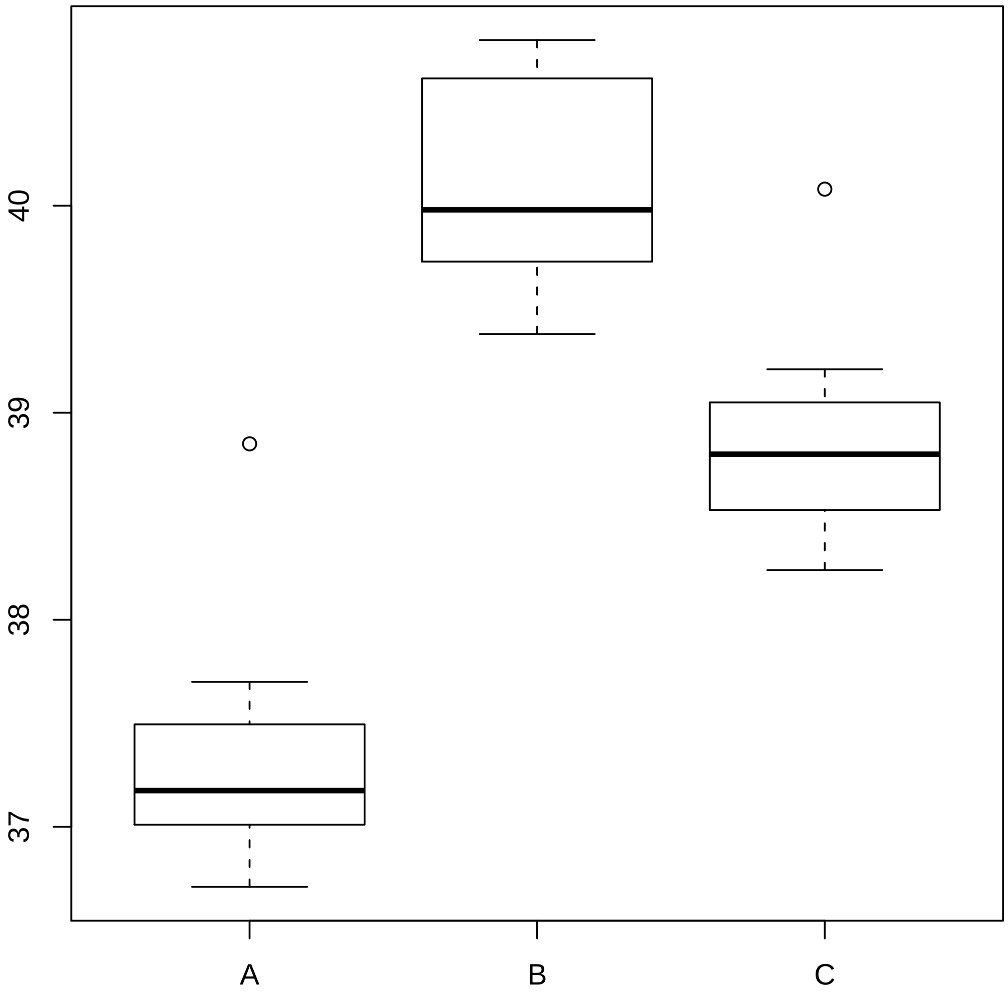
The data for this problem are on Trosset's web site

`http://mypage.iu.edu/~mtrosset/StatInfeR.html` as `salinity.dat`.

Right click on that link and save the file to the directory from which you launch your R session for this homework. Then read it into that session via `salinity <- scan("salinity.dat")`. Create a corresponding `site` vector via `site <- c(rep("A",12),rep("B",8),rep("C",10))`. Look up the documentation `?rep`. Create a data frame `salinity.site <- data.frame(salinity,site)`.

For 1. you can use `boxplot(salinity~site,data=salinity.site)` and for 2. emulate what was done on the last two slides of the ANOVA chapter.

1.



The assumption of normality and homoscedasticty does not appear to be strongly violated. We do discern asymmetry within some boxes (A and B) and outliers (A and C), but that could also be the result of sampling variation in small samples.

2.

```
> salinity <- scan("salinity.dat")
Read 30 items
> site <- c(rep("A",12),rep("B",8),rep("C",10))
> salinity.site <- data.frame(salinity,site)
> anova(lm(salinity˜site,data=salinity.site))
Analysis of Variance Table

Response: salinity
          Df Sum Sq Mean Sq F value    Pr(>F)
site       2 38.801 19.4004  66.021 4.009e-11 ***
Residuals 27  7.934  0.2939
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The significance probability is extremely small, confirming our impression from the boxplots, i.e., sampling variation alone could not easily account for these differences between sample centers. We should clearly reject the null hypothesis of equal means at level $\alpha = 0.05$.

This is your last assignment that is due. I will issue one more and you can do it on your own and compare with the solutions that I will post prior to the final.