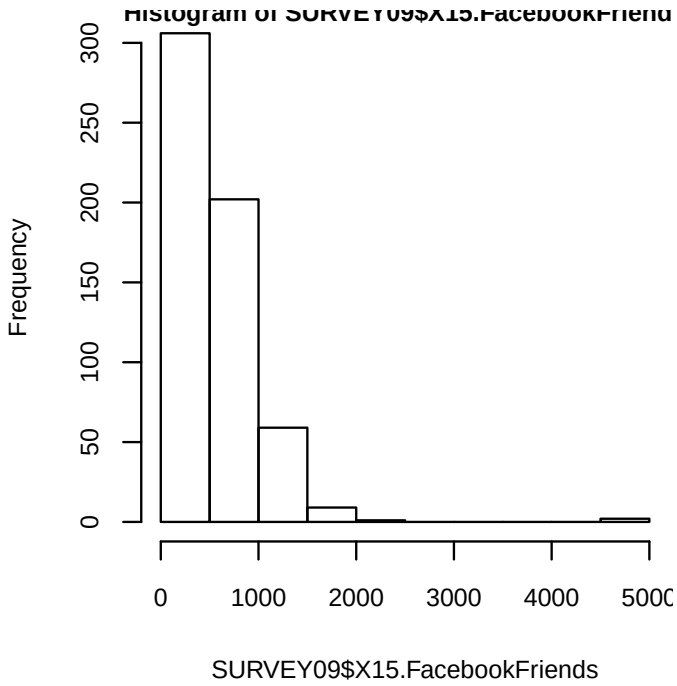
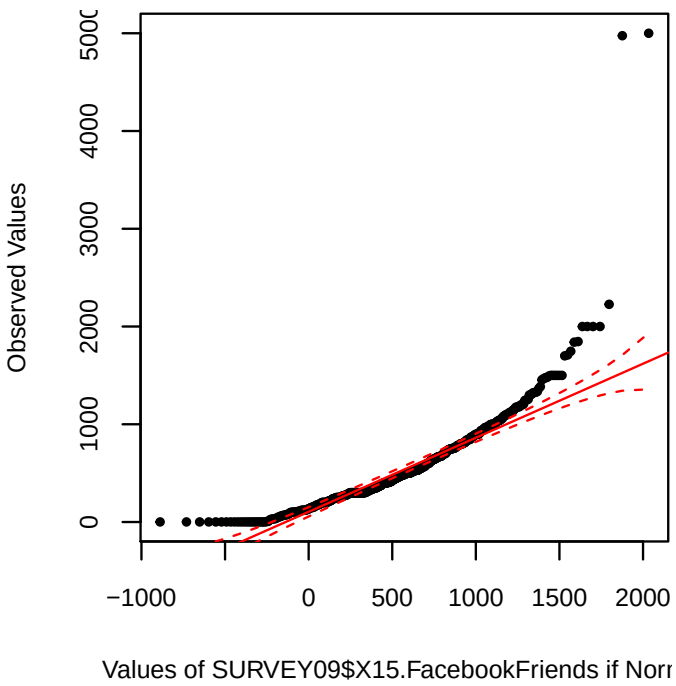
QQ plot example3 (not Normal)

```
hist(SURVEY09$X15.FacebookFriends); qq(SURVEY09$X15.FacebookFriends)
```

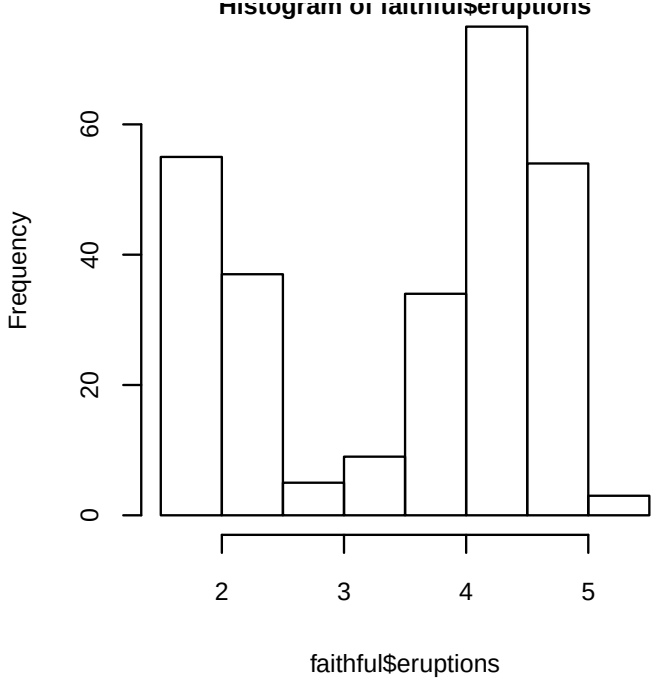


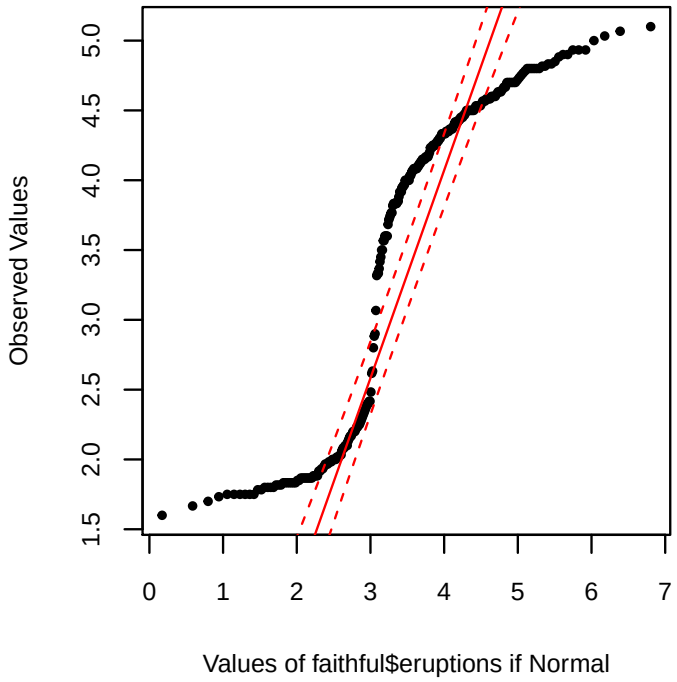


Bad match, clearly not Normal.

QQ plot example 4 (not Normal)

```
data(faithful)
hist(faithful$eruptions); qq(faithful$eruptions)
```





Bad match due to bimodality.

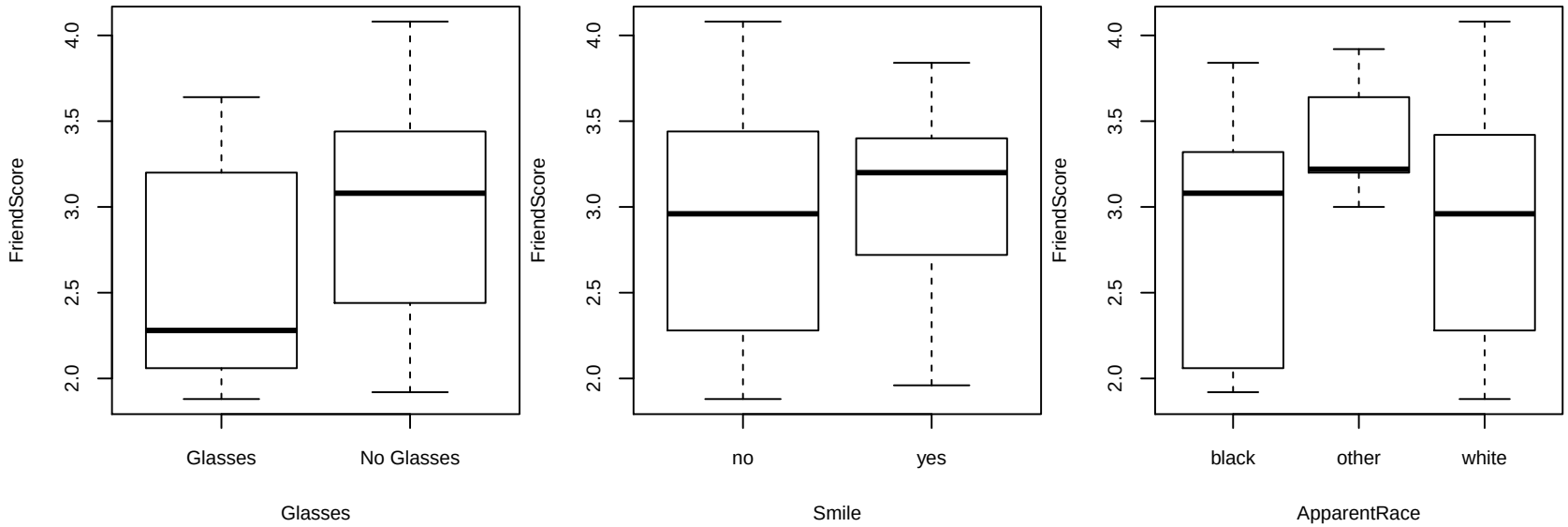
Determining Normality from QQ plots

If a Normal distribution is a good approximation for the data, then the observed data values and expected (had the data been Normal) values should closely match up and the points should fall near the diagonal red line. There is some leeway, so the distribution can be considered approximately Normal if:

- There is no systematic global curvature of the stream of points away from the diagonal red line.
- Almost all the points fall within the upper and lower dotted red bands. It is ok if a few points at the outskirts fall outside by a little.
- No points are extremely far from the upper/lower dotted bands (these are outliers, which ruin the value of the mean).

Side-by-side boxplots

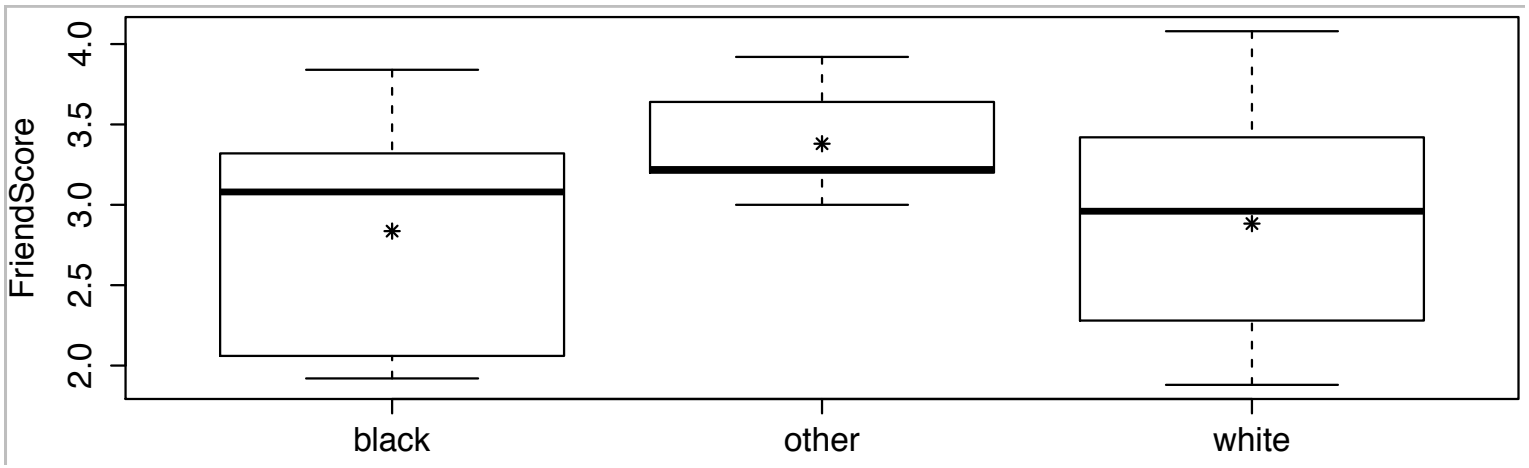
Recall that side-by-side boxplots are convenient ways of showing the overall distribution of y for two or more levels of x. The *shape* of the overall distribution remains hidden, but the median values are represented by the bar through the box.



Boxplots in associate

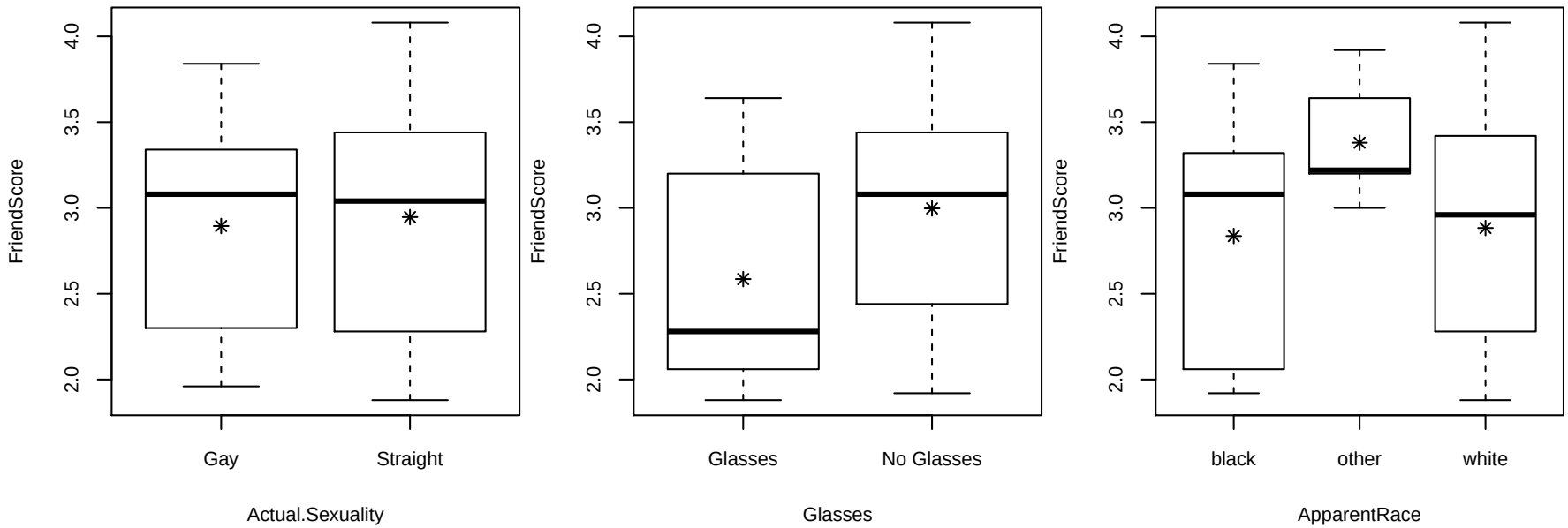
R's default presentation of boxplots is somewhat lacking because the average is not displayed. We will be analyzing association using the associate command in package regclass, which does show the averages for easy comparison.

```
associate(FriendScore~ApparentRace,data=FFRIEND)
```



What to look for in the boxplots

Look to see if the means and/or medians noticeably differ between levels of x.
If they do, we suspect an association.

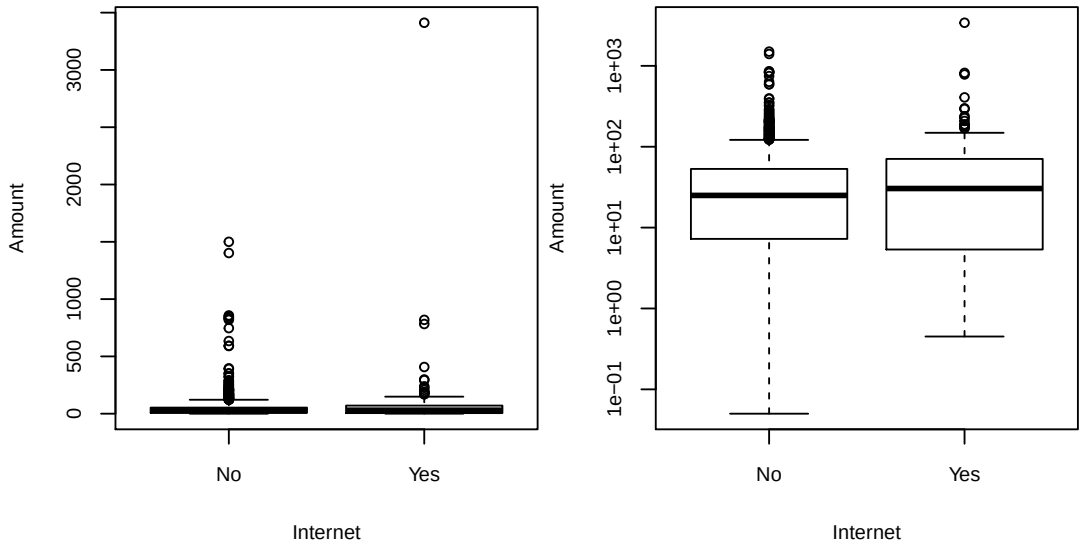


Left - do not suspect association since the means/medians are about the same.
Right/Middle - suspect an association since means/medians look different.

Reminder: consider logarithmic plots

If the distribution is highly skewed, the boxplots may not be very informative. Try plotting $\log_{10} y$ for each group instead. In the plots below, the potential association between the amount of money a customer spends and whether it is an internet purchase is examined. We do not suspect an association.

```
CUST <- read.csv("Customer37783.dat")
plot(Amount~Internet,data=CUST) #or associate(Amount~Internet,data=CUST)
plot(Amount~Internet,log="y",data=CUST) #or associate(log10(Amount)~Internet,data=CUST)
```



Philosophy

When we look at side-by-side boxplots, we never expect means or medians to match up *exactly*, even if the distributions of y for each level of x are fundamentally the same. Since the data represent a sample from a much larger population, we expect there to be some variation.

When both variables are categorical, we quantified the difference in the distribution of y between levels of x with the discrepancy between observed and expected counts. Here, we will use the discrepancy in the averages or medians between the levels of x.

When comparing averages is appropriate, we compare the *observed variability* in averages between levels of x to the *expected variability* had there been no association (each level of x has the same fundamental underlying average value of y). The test is called an **ANOVA** (analysis of variance).

Test for means: F-statistic

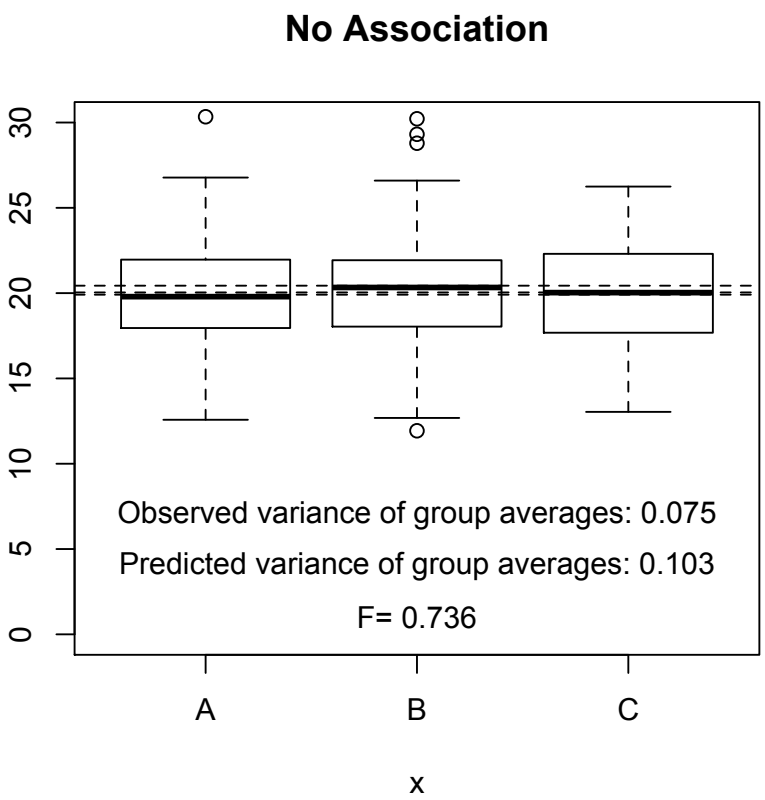
The discrepancy in the average value of y between levels of x is called the F statistic and is given by

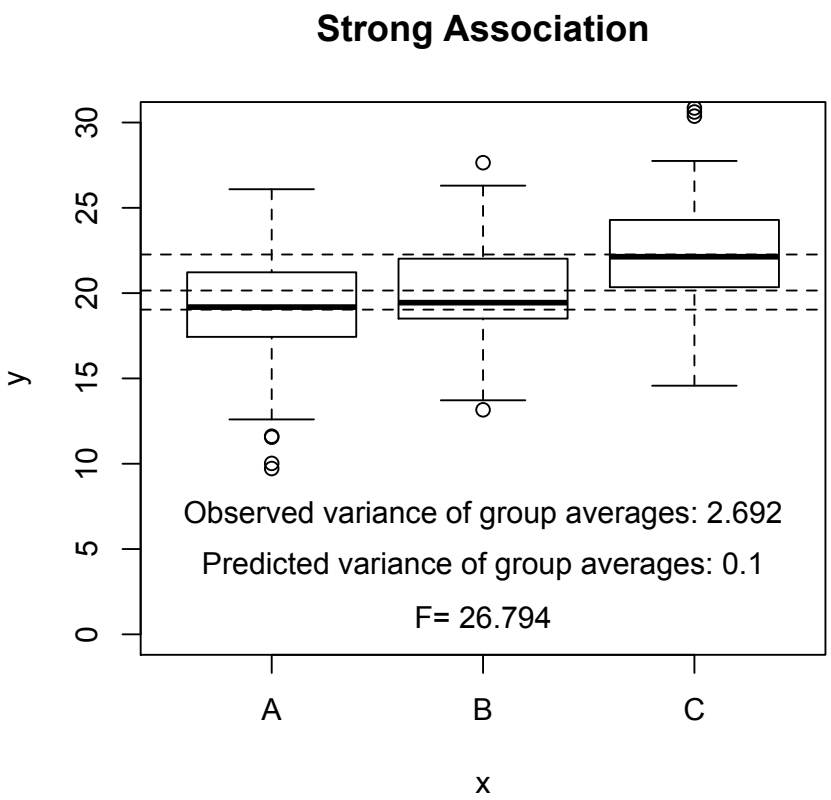
$$F = \frac{\text{Observed variance of group averages}}{\text{Expected variance of group averages if there was no association}}$$

The formula for the "predicted" variance assumes that, in reality, each group comes from identical Normal distributions. You can look up the exact formula for how F is computed online, but software will always output F so that you don't have to.

The formulas are a bit involved, but if there is no association then $F \approx 1$. When there is an association, then the value of F will be "large".

Illustration of ANOVA concepts





Interpreting the value of F

The value of F tells us something interesting about the association.

- If $F = 1$, then the observed variability in the average values of y between levels of x is exactly what you would expect had all levels had the same fundamental underlying average value of y.
- If $F = 5$, then the observed variability in the average values of y between levels of x is *five times higher* than what you would expect had all levels had the same fundamental underlying average value of y.

The question is: how large does F need to be in order to convince that an association exists between y and x, i.e., not all levels of x have the same fundamental underlying average?

Permutation procedure for finding "chance" values of F

Sampling variability alone will give values of $F > 1$ even when all levels have the same fundamental underlying average. We can find how large F can get by chance using the permutation procedure.

- Create an artificial **permutation** dataset where the observed values of y and x are *randomly* paired together. By design, there is no association between these variables and all levels of x fundamentally have the same underlying average.
- Calculate F for this permutation sample and record it.
- Repeat this process a "lot" of times and make a histogram of the values of F to see what happens "by chance" when x and y have no association.

Permutation Dataset

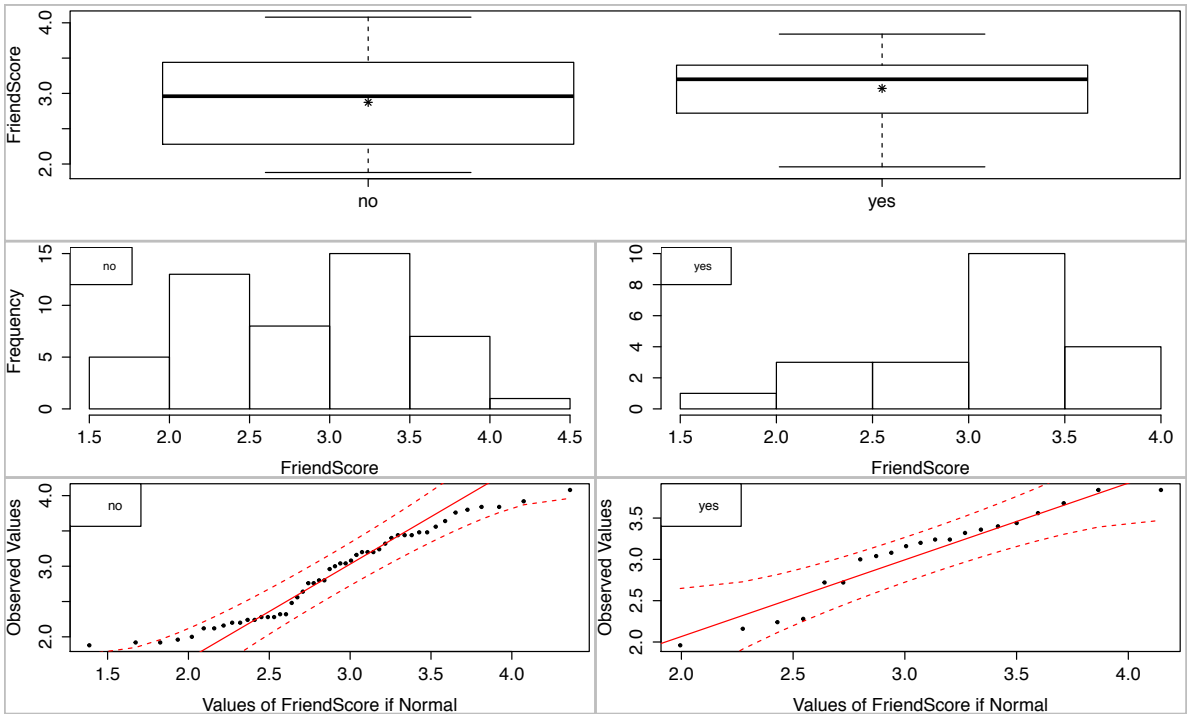
The **permutation procedure** is most easily illustrated when x and y are both numerical as we have seen. Below is a table of the original observed values and of three permutation datasets where the values of x and y have been randomly paired together (more specifically, the values of y are shuffled up and assigned at random to each individual).

Individual	Observed Data		Permutation 1		Permutation 2		Permutation 3	
	Rent	*Spend*	*Rent*	*Spend*	*Rent*	*Spend*	*Rent*	*Spend*
1	Yes	0.2	Yes	3.4	Yes	5.6	Yes	3.4
2	Yes	3.4	Yes	6.0	Yes	3.4	Yes	0.2
3	No	3.4	No	5.6	No	6.0	No	5.6
4	No	5.6	No	3.4	Yes	3.4	Yes	3.4
5	Yes	6.0	Yes	0.2	No	0.2	No	6.0

Comparing means example

Is there an association between smiling and friendship potential?

```
associate(FriendScore~Smile,data=FFRIEND)
```



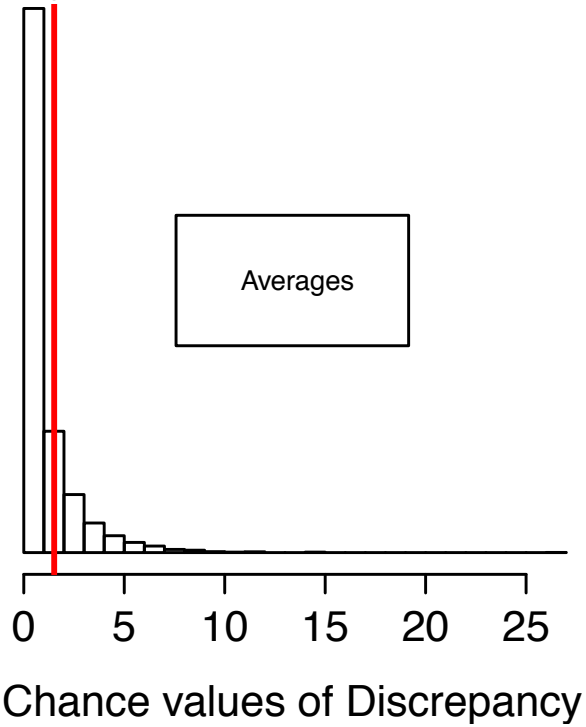
Comparing means example

	no	yes	Discrepancy	Estimated p-value
Averages (ANOVA)	2.873	3.07	1.514	0.2172
Mean Ranks (Kruskal)	39.59	25.95	1.303	0.2526
Medians	2.96	3.2	1.572	0.1926

- The average is appropriate to summarize the distribution (some points outside the bands in the QQ plot at the edges, but this is ok).
- There is a noticeable difference in averages (2.87 vs. 3.07).
- The value of F, the discrepancy in averages, is 1.5. So the variability in averages in levels is 1.5 higher than what we'd expect by chance.
- What values of F appear by chance?

Comparing means example

Creating 500 permutations, we find that such a discrepancy in averages occurs fairly often by chance, so maybe there is not an association after all.



Median test

When the distribution of y is not well-described by a Normal distribution (there may be outliers or skewness), the average does not do a good at summarizing the distribution. Thus, the preceeding ANOVA test to compare averages is not a good idea.

> When at least one of the distributions has extreme outliers or is skewed, we will compare the **medians**.

```
x <- c(3,6,8,9,12)     #Nice symmetric distribution; mean summarizes data
mean(x)
## [1] 7.6
x <- c(3,6,8,9,120)    #Very skewed with an extreme outlier (typo); mean doesn't summarize c
mean(x)
## [1] 29.2
```

Median test

The median test is the most general test and requires no assumptions about the underlying distributions!

The test comes up with a clever way to measure the variability in the median value of y between each level of x, then it uses the permutation procedure to determine whether such variation is explainable by chance or if there is evidence of an association.

Median test mechanics

- When there is no association between y and x, the fraction of individual values above and below the overall median should be about the same.
- If individuals in one group tend to have more values above (or below) the median than expected, this implies that there is an association between y and x (know the group tells you something about the distribution of y).
- Median tests determines if there is an association between x and $y =$ "is value above median?"

Illustration of median test

Original data:

	Values							Larger than median (of 3.1)						
White	3.1	2.1	3.7	1.8	2.3	1.9		no	no	yes	no	no	no	
Black	3.7	3.3	4.1	3.4				yes	yes	yes	yes			
Other	4.1	1.9	3.0	1.7	4.2	3.6	1.6	yes	no	no	no	yes	yes	no

Converted into a contingency table:

	Larger than median	
Race	Yes	No
White	1	5
Black	4	0
Other	3	4

Here, 5/6 of values for "white" are below the median while 0/4 of values for "black" are below the median, indicating a possible association. However, the test comes back with a p-value greater than 5%, indicating no statistically significant association.

p-values

When `associate()` is run, output that calculates and compares the means and medians (you can ignore the middle row about Mean Ranks) is displayed along with the *p*-value of the differences in means/medians (found via the permutation procedure).

> The *p*-value tells us the probability of finding at least as big a discrepancy in averages/medians "by chance" when each level of x has the same fundamental underlying average/median, i.e., when x and y are independent.

The *p*-value is estimated to be the fraction of permutation datasets (where no association exists) that produced a discrepancy in averages/medians at least as large as the discrepancy in the original data.