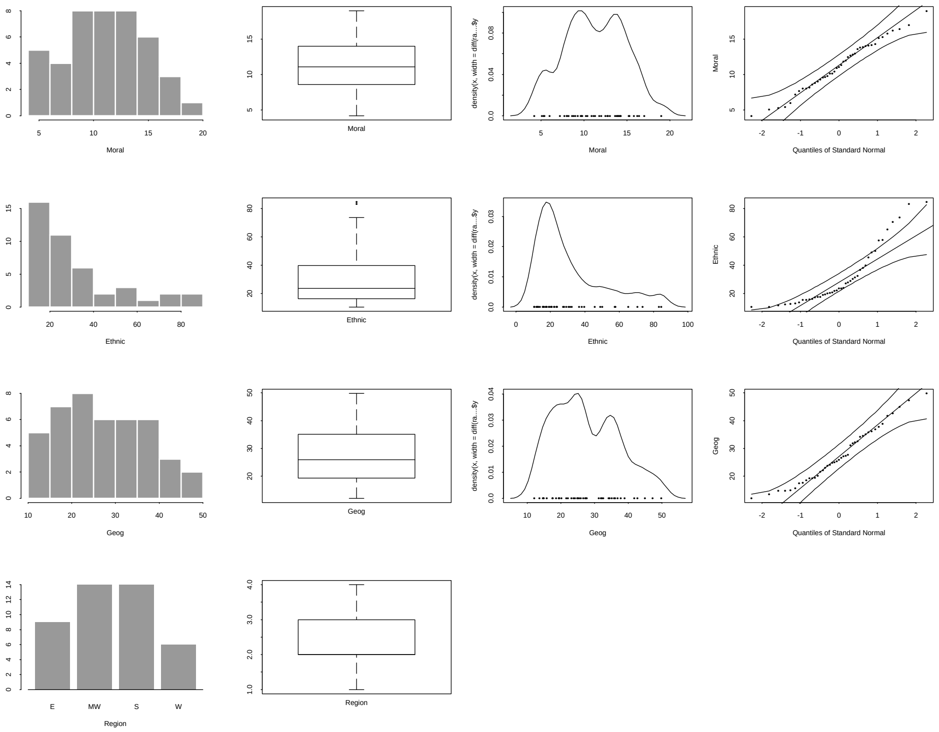
36-707 Homework 1
Solutions

Note: As a rule, hw solutions won't be as detailed as this first one. –BJ

3. Ex. 3.1(a), p. 49

   Here are the plots.



From them, we can see that

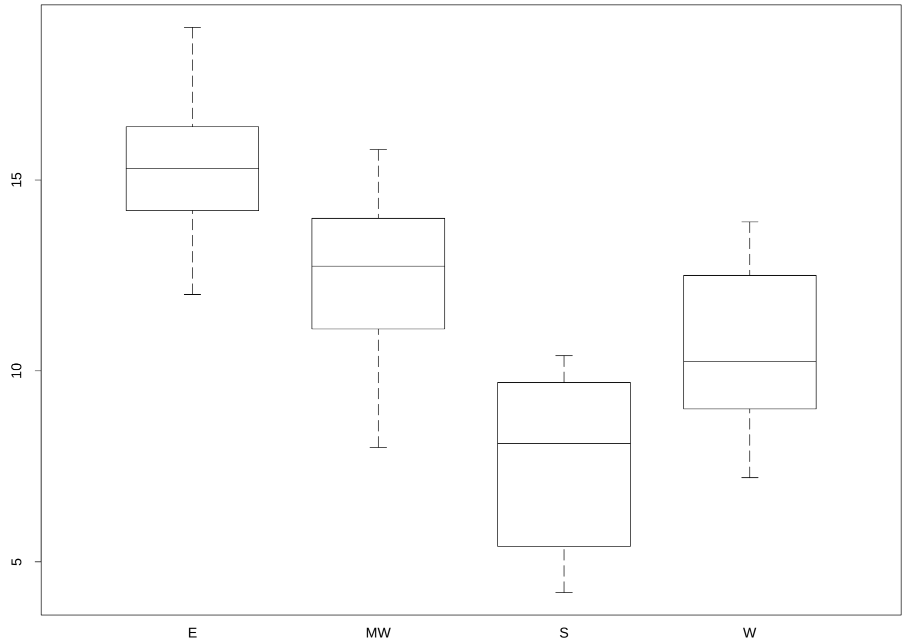
- `Moral` appears to be a continuous variable, relatively symmetric (boxplot), with a possible secondary mode near the value 5 (histogram and density plot). The density plot also suggests a possible gap or low-density area in the distribution near the value 12. The normal probability plot (a.k.a. quanitile-normal plot) gives no evidence against normality for the data.
- `Ethnic` also appears to be continuous, but is severly skewed to the right (long right tail; histogram, boxplot, density plot), with a single dominant mode near the value 18. The qqplot also shows a departure from normality, bending upward where the high quantiles of `Ethnic` are farther from the middle of the data than they would if the data were more nearly normal. There are also two high outliers (both can be seen if the plot is enlarged); they are Memphis and Birmingham (see below).
- `Geog` appears to be continous, with some very mild right skew (histogram, boxplot, density plot). There is a clearer gap in the middle of the distribution, around 30 (density plot). But there is no strong evidence against normality for this variable.
- `Region` is a discrete, categorical variable with values E (eastern USA), MW (Midwest), S (South) and W (West). Issues of symmetry, etc., do not arise since the values are not ordered in any way (in particular the boxplot is not helpful). The histogram shows that there were 14 cities in each of the MW and S, nine cities in the East, and 6 in the West, in this data set.

Note: the Splus code used for this problem follows.

```
> # To read the data and generate the plots:
> angell _ read.table("Angell.dat")
> names(angell)<-c("Moral","Ethnic","Geog","Region")
> source("qqenv.q")
> dim(angell)
[1] 43  4
> angell[1:4,]
          Moral Ethnic Geog Region
Rochester  19.0   20.6 15.0      E
 Syracuse  17.0   15.6 20.2      E
Worcester  16.4   22.1 13.6      E
     Erie  16.2   14.0 14.8      E
> par(mfrow=c(4,4))
> for (i in names(angell)) {
+ x _ angell[,i]
+ hist(x,xlab=i)
+ boxplot(x,names=i,style.bxp="old")
+ plot(density(x,width=diff(range(angell[,i]))/4),type="l",xlab=i)
+ points(x,rep(0,length(x)))
+ qqenv(x,ylab=i)
+ }
> # To find the high outliers on the ``Ethnic'' variable:
> angell[order(-angell$Ethnic),][1:5,]
             Moral Ethnic Geog Region
     Memphis   5.4   84.5 26.7      S
  Birmingham   8.2   83.1 25.9      S
Jacksonville   6.0   73.7 27.7      S
     Atlanta   4.2   70.6 32.6      S
    Richmond  10.4   65.3 24.9      S
```

4. Ex. 3.3(a), p. 53.

Here are the boxplots.



We see from the boxplots that the median level of moral integration is highest in the East, next highest in the Midwest, next in the West, and lowest in the South. Moreover the distributions are fairly well separated, with no overlap between the South and the East for example, and the middle 50% of each distribution exhibiting strong overlap only between the East and the Midwest. The overall range of moral integration scores seems about the same in each region, though the IQR (range of the middle 50% of the data) is largest in the South, followed by the West, the Midwest and the East. The data is fairly symmetric in each region, with some light skewing to the left (long left tail) in the South adn possibly the Midwest, and a little evidence of right skewing (especially in the middle 50% in the West. There are no outliers in any of the regions.

Here is the Splus code that generated the plots.

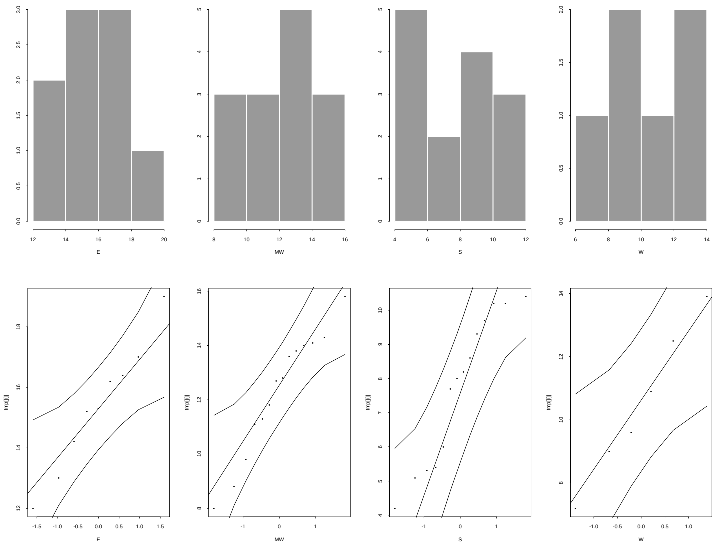
```
> attach(angell)
> boxplot(split(Moral,Region),style.bxp="old")
```

In addition to the boxplots, numerical summaries of the data, histograms and qqplots can help with seeing what is going on.

```
> # ``lapply(a,b)'' applies the function b to each element of the list a:
> lapply(split(Moral,Region),summary)
$E:
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 "12.00" "14.20" "15.30" "15.37" "16.40" "19.00"

$MW:
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
 " 8.00" "11.15" "12.75" "12.28" "13.95" "15.80"

$S:
     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 " 4.200" " 5.550" " 8.100" " 7.736" " 9.600" "10.400"

$W:
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 " 7.20" " 9.15" "10.25" "10.52" "12.10" "13.90"
> tmp _ split(Moral,Region)
> for (i in 1:4) hist(tmp[[i]],xlab=names(tmp)[i])
> for (i in 1:4) qqenv(tmp[[i]],ylab=names(tmp)[i])
```

This generates the following supplementary plots (which are not all that helpful since the sample sizes are so small):



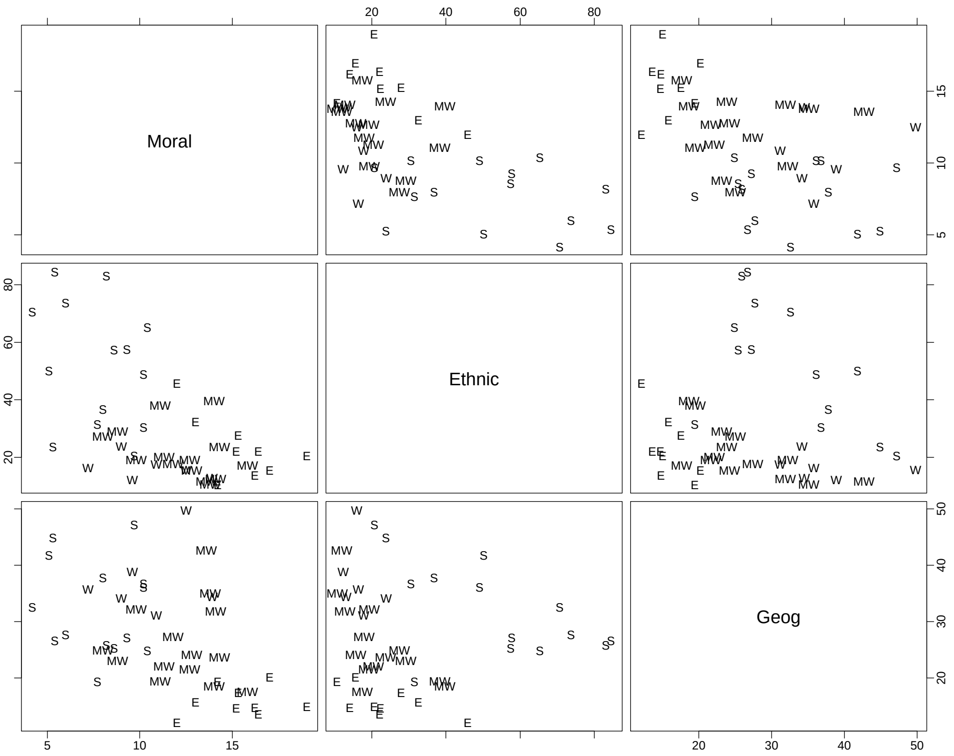
5. Ex. 3.4(a), p. 56.

   One can get a nice scatterplot matrix (with no point labels) using `pairs(angell[,1:3])`; and one can use different plotting symbols for the different regions using `brush(angell)`. However, as pointed out in class, you can't easily print the "brush" window, and it's hard to get "pairs" to plot different characters for each region.

   After spelunking in the guts of `pairs`, `pairs.default`, `text`, and `text.default` for a while (using `help` and `invisible(vi())`), I came up with the following:

```
> myplot _ function(x,y) { text(x,y,labels=as.character(angell[,4])) }
> myplot
function(x, y)
{
        text(x, y, labels = as.character(angell[,4]))
}
> pairs(angell[,1:3],panel=myplot)
```

   Note that `myplot` only works for the `angell` data frame! Anyway, here's the result:



What stands out here is that cities in the South are separated from cities in the other four regions, in the moral and ethnic integration variables.

`brush()` is also a bit frustrating, in that it is difficult to get the region labels to appear on each point in the plots. As a compromise, I tried

```
> brush(angell[,1:3],rowlab=as.character(angell[,4]))
```

and then highlighted the cities in each region by hand.

6. **Writing exercise**

   Here is my first draft.

   Transformations are a common tool in applied statistics, and in regression analysis in particular. They are used to make the data better comply with the assumptions underlying the models we wish to use to analyze the data. There are three general reasons for transforming data:

   - *To reduce skewness*, or equivalently enhance symmetry, in the distribution of individual variables. This is done because many basic statistical procedures, up through multiple regression, assume that the underlying variables follow a normal distribution, and the procedures still work reasonably well if the variables are at least symmetrically distributed with no severe outliers (extreme values).
   - *To correct nonconstant spread* in one variable when plotted against another variable. This is done again because many basic statistical procedures (and those based on the assumption that the variables jointly follow the normal distribution, in particular) assume that the spread in one variable does not depend on the value of another variable in the problem.
   - *To promote a linear relationship* between two or more variables. Again, this is because basic regression analysis assumes a linear relationship between the variables.

   The most common transformations for continuous, unbounded variables are power transformations, $y = x^p$. For example, $p < 1$ corrects positive skew (long right tails), and $p > 1$ corrects negative skew (long left tails). The log and exponential functions can be viewed as the extremes of these two families of power functions. For continuous bounded variables (say, proportions bounded between 0 and 1), $\text{logit}(x) = log(x/(1-x))$, or another S-shaped transformation, is a good first step. Generally a transformation should be chosen both to improve the appropriateness of the assumptions of the statistical model, and to facilitate—or at least not hinder—an interpretation of the variables. Thus, powers such as 2, 1/2, and log and exponential functions are to be preferred over powers such as 0.324534 or 4.35254.

Here is my second draft. The first parag and the bullets were heavily edited, the second parag hardly at all.

Transformations are a common tool in applied statistics, and in regression analysis in particular, because they can help the data to better comply with the assumptions underlying statistical models. Transformations are used: (a) to reduce skewness and enhance symmetry in the distribution of a single variable, since many basic statistical models assume that the variables are normally distributed, or at least symmetric; (b) to correct nonconstant spread in one variable as a function of another, since the multivariate normal distribution on which many basic procedures are based assumes this; and (c) to promote linearity in the relationship between variables, since basic regression models assume this.

The most common transformations for continuous, unbounded variables are power transformations, $y = x^p$. For example, $p < 1$ corrects positive skew (long right tails), and $p > 1$ corrects negative skew (long left tails). The log and exponential functions can be viewed as the extremes of these two families of power functions. For continuous bounded variables (say, proportions bounded between 0 and 1), $\text{logit}(x) = log(x/(1-x))$, or another S-shaped transformation, is a good first step. Generally a transformation should be chosen both to improve the appropriateness of the assumptions of the statistical model, and to facilitate—or at least not hinder—an interpretation of the variables. Thus, powers such as 2, 1/2, and log and exponential functions are to be preferred over powers such as 0.324534 or 4.35254.