CHAPTER 1

EXPLORATORY DATA ANALYSIS

> We were together learning how to use the analysis of variance, and perhaps it is worth while stating an impression that I have formed-that the analysis of variance, which may perhaps be called a statistical method, because that term is a very ambiguous one — is not a mathematical theorem, but rather a convenient method of arranging the arithmetic. Just as in arithmetical textbooks — if we can recall their contents — we were given rules for arranging how to find the greatest common measure, and how to work out a sum in practice, and were drilled in the arrangement and order in which we were to put the figures down, so with the analysis of variance; its one claim to attention lies in its convenience.
>
> *The Future of Data Analysis*
> J. W. TUKEY

IN 1986, the Vietnamese government began a policy of textitdoi moi (renovation), and decided to move from a centrally planned command economy to a "market economy with socialist direction". As a result, Vietnam was able to evolve from near famine conditions in 1986 to a position as the world's third larger exporter of rice in the mid nineties. Between 1992 and 1997 Vietnam's GDP rose by 8.9% annually (World Bank, 1999).

The first Vietnam Living Standards Survey (VLSS) was conducted in 1992-93 by the State Planning Committee (SPC) (now Ministry of Planning and Investment) along with the General Statistical Office (GSO). The second VLSS was

conducted by the GSO in 1997-98. The survey was part of the Living Standards Measurement Study (LSMS) household surveys conducted in a number of developing countries with technical assistance from the World Bank.

The second VLSS was designed to provide an up-to-date source of data on households to be used in policy design, monitoring of living standards and evaluation of policies and programs. One part of the evaluation was whether the policies and programs that were currently available were age appropriate for the population. For example, if a country has a higher proportion of older people, then there needs to be programs available that appeal to that sector of the population. Another concern was whether the living standards for different sections of the country were equitable. We will use data from the second VLSS (available in the *VLSSage.dat* and *VLSSperCapita.txt* data sets) to examine the following research questions,

1. What is the age distribution for the Vietnamese population?
2. Are there differences in the annual household per capita expenditures between the rural and urban populations in Vietnam?
3. Are there differences in the annual household per capita expenditures between the seven Vietnamese regions?

1.1 Research Question #1: What is the age distribution for the Vietnamese population?

The variable age from the *VLSSage.dat* contains the ages of 28,633 individuals (in years ranging from 0 to 99) living in the 5999 sampled households. To examine this research question, we need to read in the data and look at some graphical and numerical summaries.

```
> ages <- read.table("VLSSage.dat", header=T)
> attach(ages)
> head(ages)

  age
1  68
2  70
3  31
4  28
5  22
6   7


> hist(age, xlab="Age", main="")
```

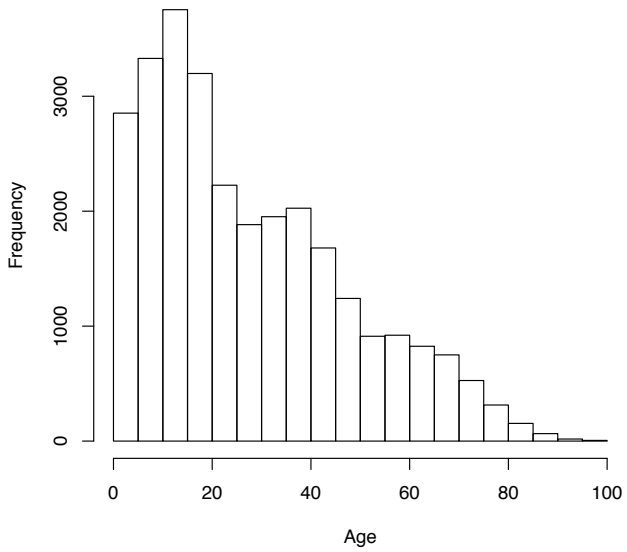

Figure 1.1: Distribution of 28,633 ages from the VLSS data.

The `hist()` function in R can be used to create a histogram. The function takes a required argument of the variable name (e.g., `age`). The optional arguments `xlab=` and `main=` can be used to add a label to the x-axis and a title to the histogram respectively. The histogram of ages is displayed in Figure 1.1.

The Number of Bins in a Histogram

Statistical packages such as R, use a default number of bins Ð or equivalently a default bin width, but this default can often be an arbitrary choice. The matter of selecting the number of bins or classes to be used in the histogram is not trivial. Quite often, the bin width is a subjective choice for the methodologist, using contextual knowledge about the data. In R, the The `breaks=` argument can be used in the The `hist()` function to specify the number of breakpoints between histogram bins.

From Figure 1.2, we can see that the choice of 55 bins gives a clear picture of three distinct generations, the young, the middle-aged and the older individuals. With ten bins, the three groups are not clearly visible. The argument `breaks=n` added to the `histogram()` function changes the number of bins. The

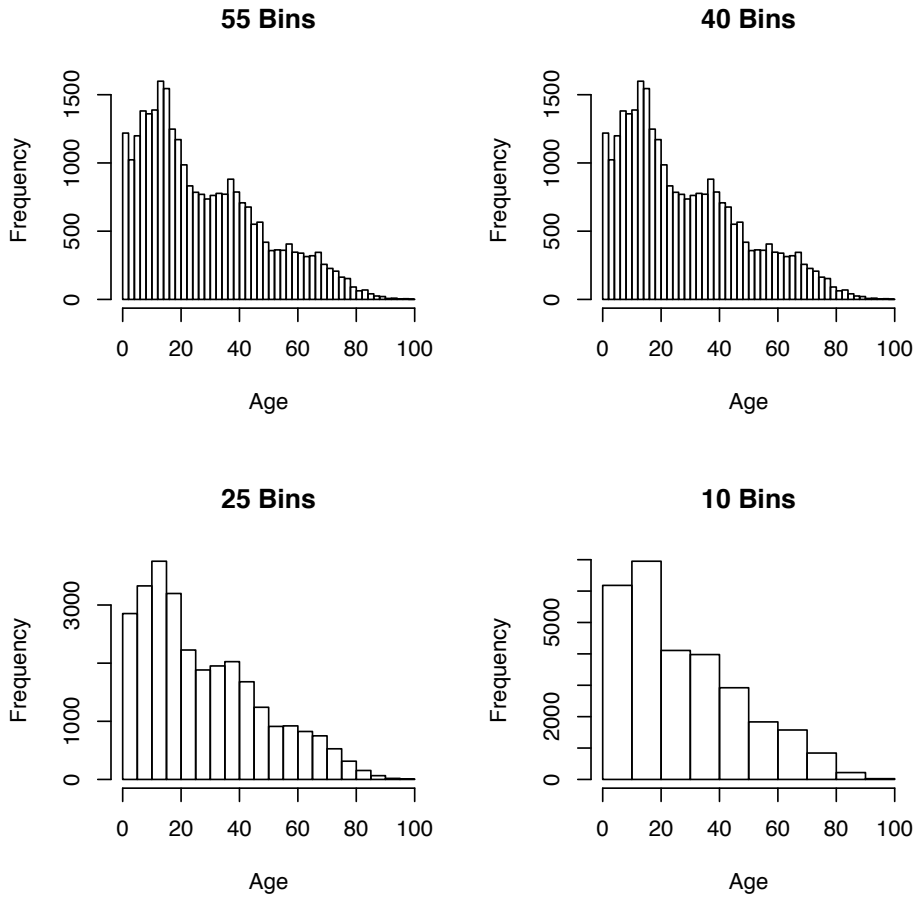

Figure 1.2: Distribution of 28,633 ages from the VLSS data with 55, 40, 25, and 10 bins.

code to produce Figure 1.2 is shown below. The `par(mfrow=c(2,2))` command sets the layout of the graphical output to display the four graphs in a 2x2 array.

```
> par(mfrow=c(2,2))
> hist(age, xlab="Age", main="55 Bins", breaks=55)
> hist(age, xlab="Age", main="40 Bins", breaks=40)
> hist(age, xlab="Age", main="25 Bins", breaks=25)
> hist(age, xlab="Age", main="10 Bins", breaks=10)
> par(mfrow=c(1,1))
```

Several statisticians have provided formulae for choosing the number of bins (e.g., Freedman & Diaconis, 1981; Scott, 1979; Sturges, 1926). All of these methods have been employed in the The `hist()` function. The argument `breaks="sturges"`,

`breaks="fd"`, or `breaks="scott"` will use the appropriate methodology to compute the number of bins. If the argument `breaks=` is omitted, the Sturges method is used.

Recently, Wand (1997) proposed a series of "plug-in" rules for selecting the width (and therefore the number) of the bins in a histogram. The justification for the rules lies in the fact that the resulting histogram provides a good estimate for the density. The rules are relatively complicated to compute, but the `dpih()` function from the `KernSmooth` library can be used to compute the appropriate width for each bin. In our case, the zero-stage rule yields a width of 1.33.

```
> library(KernSmooth)
> dpih(age)

[1] 1.329865
```

Using this bin width in the range of data — from 0 to 100 — produces about 68 bins (100/1.33 = 68). A histogram with 68 bins is shown in Figure 1.3.

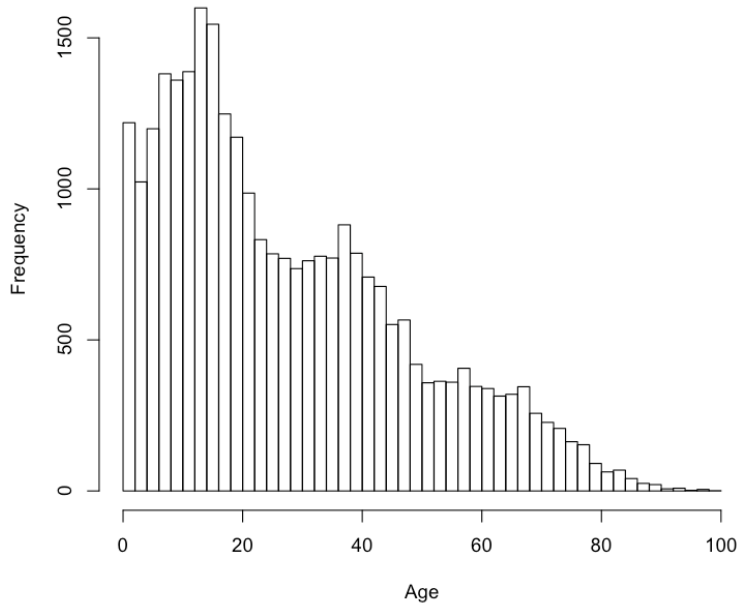

Figure 1.3: Distribution of the VLSS age data with 68 bins.

This choice of 68 bins yields three modal clumps in the histogram from which emerge two interesting features. The second mode in the age distribution occurs near the age of 40, and has lower frequency, in part because of lives lost

during the war, which ended in 1975 with the reunification of North and South Vietnam. One can also see the effect of decreasing fertility rates in the fact that frequencies decrease as ages decrease from about 16 to zero, which has been documented in the literature (e.g., Haughton, Haughton, & Phong, 2001). The question arises as to whether three subpopulations might occur in the distribution of ages. The population pyramid based on the 1999 Vietnam census does seem to indicate the presence of three subpopulations (see General Statistical Office, 2001).

Kernel Density Estimators

Although a histogram can be very useful for examining the distribution of a variable, the graph can differ dramatically depending on the number of bins used. This problem can be overcome (partially) using nonparametric density estimation. Density estimation is an attempt to estimate the probability density function of a variable based on the sample, but less formally it can be thought of as a way of averaging and smoothing the histogram.

Since a density function encloses an area of 1, we first must rescale the histogram so that the total area under the smoothed line (the area within the bins) equals one. In other words, we examine the proportion of cases at specific points in the histogram rather than the frequency counts.

Kernel density estimation is essentially a sophisticated form of locally weighted averaging of the distribution. It uses a weight function (i.e., kernel) that ensures the enclosed area of the curve equals one. The kernal density estimator creates an estimate of the density by placing a "bump" at each data point and then it sums the "bumps" up using,

$$p(x) = \frac{1}{nh} \sum K\left(\frac{x - X_i}{h}\right) \tag{1.1}$$

where K is the kernel density function (the "bump" function); x is the point where the density is estimated; X_i is the center of the interval; and h is the bandwidth (i.e., window half-width). Note that both the kernel function and the bandwidth must be specified by the methodologist.

Selecting the Kernel Function and Bandwidth

A choice of kernel density function recommended in the literature is the Epanechnikov kernel[1]. This kernel is the most efficient in minimizing the error when approximating the true density by the kernel density (see Silverman 1986).

Selecting the bandwidth h, is primarily a matter of trial and error. The smaller h is, the more details are shown on the graph of the kernel density. As the bandwidth increases, the density curve becomes smoother. In Figure 1.4 we give four Epanechnikov kernel density estimators for the age in years of each individual in the 1998 VLSS. The second kernel density with a bandwidth of 2 clearly shows three age groups: the young, middle and old generation.

Ideally a bandwidth is chosen that is small enough to reveal detail in the graph, but large enough to produce random noise. Statistical theory provides some guidance by suggesting,

$$h = .9\sigma n^{-1/5} \tag{1.2}$$

Notice that as the sample size increases, the bandwidth is narrower which permits the graph to show a finer degree of detail. The population standard deviation σ is generally unknown, so we replace it with an adaptive estimator of spread, namely,

$$A = \min\left\{\hat{\sigma}, \frac{\text{IQR}}{1.349}\right\} \tag{1.3}$$

This is a precaution since if the population is not normally distributed, the sample standard deviation tends to be over-inflated which would then produce a poor estimate of the population density. An even further caveat is that if the underlying density distribution is substantially non-normal — as in our example, the bandwidth produces a window width $2h$ that is oftentimes too wide (i.e., the line is too rough). However it is good value to use as a starting point and the value can be adjusted downward until the resulting plot becomes too rough.

In our example, the starting bandwidth would be,

$$h = .9(20.284)(.128) = 2.344$$

Using R to Plot a Kernel Density Estimate

The `density()` function can be used to obtain a kernel density estimate of a distribution in R. Within this function, we need to specify the kernel and

[1] There are other choices that are often used including the familiar standard normal density function.

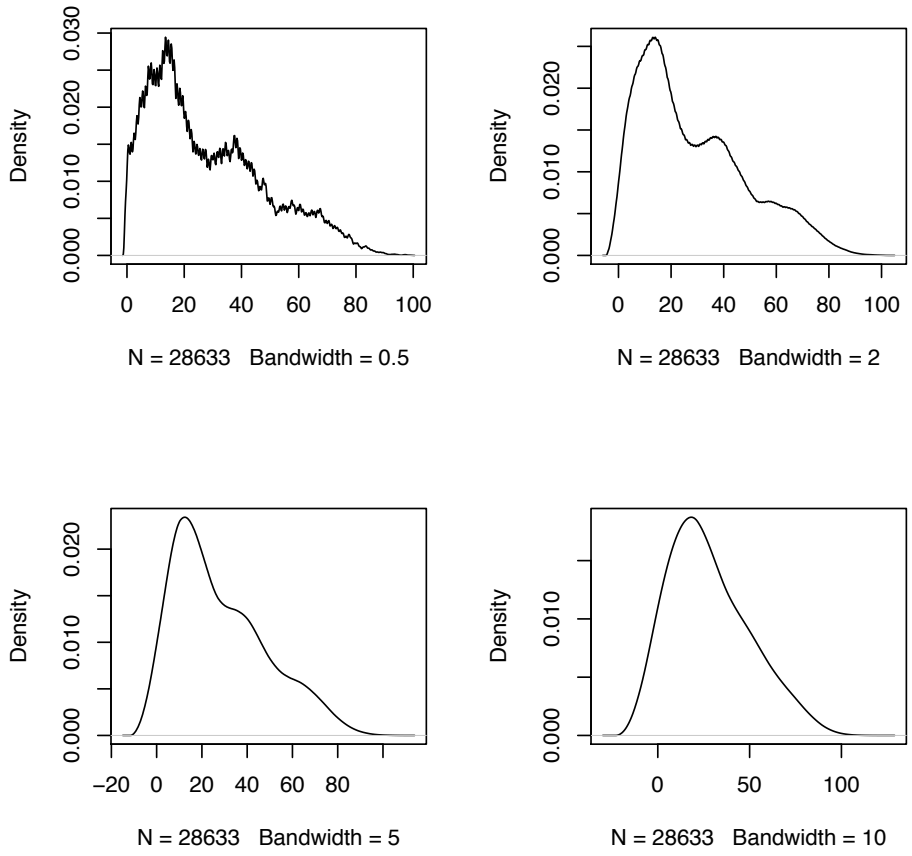

Figure 1.4: Kernel density estimates of the population distribution of Vietnamese ages with bandwidths of .5, 2.344, 5, and 10.

bandwidth using `kernel="epanechnikov"` and `bw="2.344"`. Since this is the default bandwidth selected by R in the `density()` function, the `bw=` argument can be omitted. Lastly, we need to call the `plot()` function on our `density()` function[2] Using trial and error to adjust the bandwidth downward from our initial estimate, it can be found that a good bandwidth might be around 2.1. The R code used to produce a plot of the kernel density (see Figure 1.5) estimate is as follows.

```
> plot(density(age, kernel="epanechnikov"), main="")
> plot(density(age, kernel="epanechnikov", bw=2.1), main="")
```

[2]Note that all of the arguments (e.g., `main=`, `xlab=`) can be used in the `plot()` function.

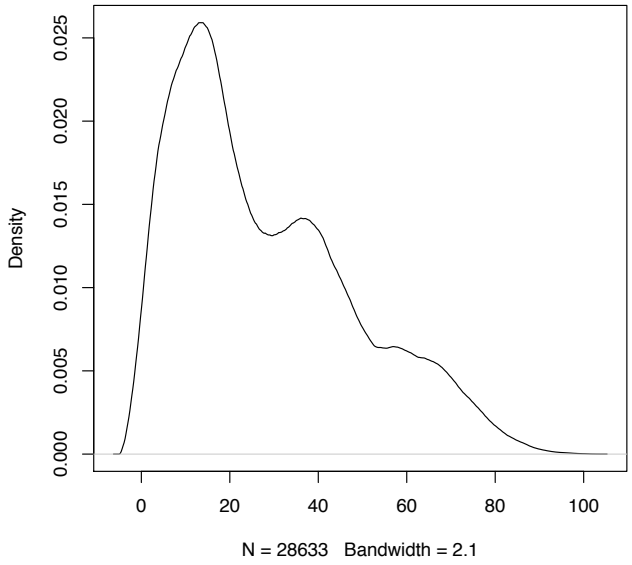

Figure 1.5: Kernel density estimate of the population distribution of Vietnamese ages with bandwidth of 2.1.

1.2 Research Question #2: Are there differences in the annual household per capita expenditures between the rural and urban populations in Vietnam?

The data set *VLSSperCapita.txt* contains data on the household per capita expenditures for 5999 households along with two demographic variables. We will read in the data and examine it using the following code in R.

```
> household <- read.table("VLSSperCapita.txt", header=T)
> attach(household)
> head(household)

  household      dong region urban    dollar       regionName
1     12225 2764.8999      1     0 184.32666 Northern Uplands
2     13314  940.9336      1     0  62.72891 Northern Uplands
3     12101 1786.9258      1     0 119.12839 Northern Uplands
4      7505 2130.5046      1     1 142.03364 Northern Uplands
5     11709 1149.1510      1     0  76.61007 Northern Uplands
6     15309 1461.8467      1     0  97.45645 Northern Uplands
```

Plots that Allow Comparisons

Plots of the kernel density estimates (KDE) are particularly useful when comparing two or more groups. For example, we could compare the distribution of the household per capita expenditures for the urban and rural populations in Vietnam. In R, we will need to plot the KDE for the rural population, and then plot the KDE for the urban population on the same graph. We will need to differentitate between the two plots by specifying arguments in the `plot()` function and also by adding a legend to our plot. Before we do anything, we need to split up our data based on whether the household is urban or rural.

Subsetting Data Using R

Since we have a factor that specifies whether each household in our sample is urban (urban = 1) or rural (urban = 0), we can use the `subset()` function to only plot a subset of our data, namely just the rural households. This function takes two arguments: the variable that you would like subsetted, and a logical expression that indicates how to subset. For example to obtain a subset of the household per capita expenditures (in dollars) for only the rural households (the urban variable is equal to 0), we would use,

```
> subset(dollar, urban==0)
```

This can then be used in the `density()` function as

```
> plot(density(subset(dollar, urban==0), kernel="epanechnikov"), main="")
```

To make the plot easier to read, we will set the limit on the y-axis. To differentiate between the two plots, we can also change the type of line (e.g., dotted). Both of these are arguments in the `plot()` function. Namely, `ylim=c(min,`

`max)` and `lty=value`. Finally, we will add a label to the x-axis (`xlab="label"`. The complete code for all of this is,

```
> plot(density(subset(dollar, urban==0), kernel="epanechnikov"), main="", lty=2,
       ylim=c(0,.008), xlab="Expenditures per capita")
```

Adding a Plot Using R

We can draw on top of a plot using the `lines()` function. We will add a density plot of the household per capita expenditures for the urban households to our existing plot of the rural households. To help further to tell the two plots apart, we will plot this KDE using a solid line. The code is,

```
> lines(density(subset(dollar, urban==1), kernel="epanechnikov"), main="", lty=1)
```

Adding Other Helpful Pieces to the Plot

It would be helpful to readers if we added a legend to our plot. This can be done using the `legend()` function. The following code adds a legend to the top-righthand corner of our plot.

```
> legend("topright", legend=c("Rural","Urban"), lty=c(2,1))
```

The last piece we will add to the plot is a vertical line that depicts the poverty line in Vietnam. This can be done using the `abline()` function and specifying the argument `v=value`. We add a red, solid line at the poverty line of $119.32 using the code,

```
> abline(v=119.32, lty=1, col="red")
```

In Figure 1.6, the kernel densities estimates for the per capita expenditures of urban (solid) and rural (dotted) households are plotted on the same graph. A vertical line positioned at the poverty line[3] has also been added to the graph. It is quite clear on the graph that the distribution of per capita expenditure is shifted to the right, and is more variable, for urban households. Rural areas are more homogeneous, with shared poverty except for a rather small number of wealthier households.

[3]The poverty line was established at 1,789.871 thousand VND for 1998 by the General Statistical Office. This is equivalent to $119.32.

It is also worth noting that the poverty line is close to the mode of the rural expenditure per capita distribution, which means that a small increase in expenditure per capita is enough to shift many of the rural households to a position above the poverty line. This is one likely explanation for recent dramatic reductions in poverty rates in Vietnam. As the poverty line moves further to the right, further reductions in poverty rates are likely to be smaller in magnitude.

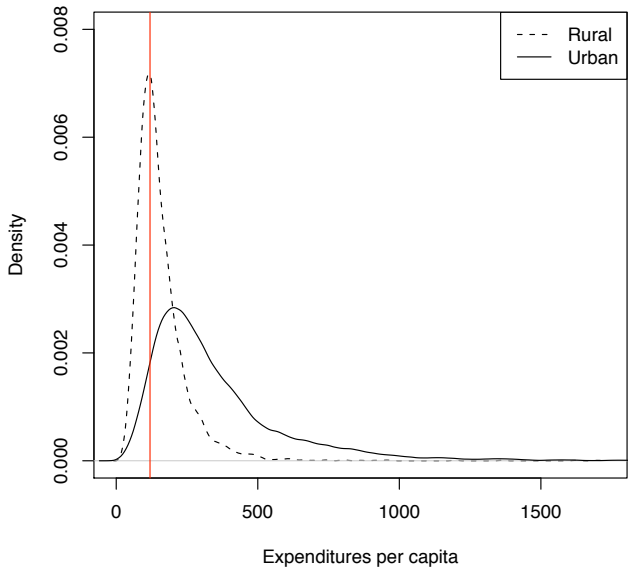

Figure 1.6: Kernel density estimate of the rural and urban population distribution of household per capita income.

Robust Estimators

Now that we have seen that the two populations seem to differ in per capita expenditures, we would like to quantify that difference. Often this is quantification is carried out by providing the sample mean difference between the two groups. Unfortunately, both of these distributions are heavily skewed. In such cases, the sample mean is often not the best estimate of the center of the population distribution. Thus, the difference between the two sample means may offer a poor summary of how the two populations differ and the magnitude of that difference. For that reason we need to use a more robust esti-

mator than the sample mean. Two robust estimates of central tendency are described and illustrated below.

Winsorized Mean

One problem with the mean is that the tails of a distribution can dominate its value. If a measure of location, such as the mean, is intended to reflect what the typical subject — or in our case, household — is like, the mean can fail because its value can be inordinately influenced by a very small proportion of the subjects who fall in the tails of the distribution. One strategy for dealing with this problem is to give less weight to values in the tails of the distribution, and pay more attention to the values near the center. A specific strategy for implementing this idea is to winsorize[4] the distribution.

Let us compute the 20% winsorized mean for the 1730 urban households. In essence, winsorizing the distribution changes the highest x% of the scores to the next smallest score, and changes the x% smallest scores to the next largest score. In our example, the highest and lowest 20% of the urban per capita expenditures (346 households from each end of the distribution) will be changed into the next nearest score. The 346 lowest values will be changed to \$173.01, the 346 highest values will be changed to \$467.59, and the middle 1038 values will remain the same. Thus,

$$\hat{\mu}_{W_{20\%}} = \frac{173.01 + 173.01 + 173.01 + \cdots + 467.59 + 467.59 + 467.59}{1730} = 303.24$$

The winsorized mean for the urban sample is \$303.24. We can use the same process to compute a 20% winsorized mean for the rural sample as well. This turns out to be \$143.23. This would suggest that a robust estimate of the mean difference in per capita expenditures between the rural and urban populations is around \$160.

R does not have a native function to compute the winsorized mean. To obtain the winsorized mean, we have to create our own function. You can copy and paste the code taken from Wilcox (2005) and provided in Table 1.1 into R.

The function can then be used to find the winsorized mean using the following R code. In this code we first create two new variables, *urb* and *rur*, by using

[4] Among the mathematicians recruited by Churchill during the Second World War was one Charles Winsor. For his efforts in removing the effects of defective bombs from the measurement of bombing accuracy, he received a knighthood and we received a new statistical tool — winsorized means.

```
win<-function(x,tr=.2){
#
# Compute the gamma Winsorized mean for the data in the vector x.
# tr is the amount of Winsorization
#
   y<-sort(x)
   n<-length(x)
   ibot<-floor(tr*n)+1
   itop<-length(x)-ibot+1
   xbot<-y[ibot]
   xtop<-y[itop]
   y<-ifelse(y<=xbot,xbot,y)
   y<-ifelse(y>=xtop,xtop,y)
   win<-mean(y)
   win
}
```

Table 1.1: R function to compute winsorized mean.

the `subset()` function. The `tr=.2` argument in the `win()` function specifies the percentage to winsorize from each end of the distribution.

```
> urb <- subset(dollar, urban==1)
> rur <- subset(dollar, urban==0)
> win(urb, tr=.20)

[1] 303.2407

> win(rur, tr=.20)

[1] 143.2265
```

Trimmed Mean

Rather than winsorize, another strategy for reducing the effects of the tails of a distribution is simply to remove them. This is the strategy employed by trimming. To find a trimmed mean, the x% largest and smallest scores are deleted and the mean is computed using the remaining scores. Let us compute the 20% trimmed mean for the 1730 urban households. We would eliminate the 346 (20%) households with the highest and lowest per capita expenditures from the data. The trimmed mean is then found by,

$$\hat{\mu}_{t_{20\%}} = \frac{173.01 + 173.25 + 173.32 + \cdots + 466.97 + 467.50 + 467.59}{1038} = 291.87$$

The 20% trimmed mean for per capita expenditures for households in the sample in rural areas can be found in a similar manner and is equal to $139.47. The difference in per capita expenditures is then found to be around $152.

The biggest advantage to the winsorized and trimmed means is that the standard errors associated with them will be smaller than it will be for the untrimmed mean. This is especially true for distributions that are long or heavy tailed. In these distributions a robust measure of central tendency will often provide a better estimate of the population mean. One last note that bears mentioning, is that the median is also sometimes used as a robust estimator, but In fact, the median is just a trimmed mean with the percentage of trim equal to,

$$\frac{1}{2} - \frac{1}{2n}$$

To obtain a trimmed mean using R, we need only add the `trim=` argument to the original `mean()` function. The code to obtain the 20% trimmed means looks like the following using the same urban and rural variables we created earlier.

```
> mean(urb, trim=.2)

[1] 291.8671

> mean(rur, trim=.2)

[1] 152.4
```

Winsorized Variance

It is also possible to compute a robust estimate for the variation in a data set. This is useful especially if one wishes to use inferential methods (e.g., hypothesis tests or confidence intervals). The winsorized variance is computed as,

$$\hat{\sigma}^2_W = \frac{1}{n-1}\sum (w_i - \hat{\mu}_W)^2 \tag{1.4}$$

where w_i are the values in the winsorized distribution, and $\hat{\mu}_W$ is the winsorized mean. The winsorized standard deviation can be found by taking the square root of the winsorized variance. The winsorized variance is used as an accompanying measure of variation for both the trimmed and winsorized means. Wilcox (2005) has once again provided a function for computing this robust measure using R. The code provided in Table 1.2 should be copied into R.

```
winvar<-function(x,tr=.2, na.rm=F){
#
# Compute the gamma Winsorized variance for the data in the vector x.
# tr is the amount of Winsorization which defaults to .2.
#
   if(na.rm)x<-x[!is.na(x)]
   y<-sort(x)
   n<-length(x)
   ibot<-floor(tr*n)+1
   itop<-length(x)-ibot+1
   xbot<-y[ibot]
   xtop<-y[itop]
   y<-ifelse(y<=xbot,xbot,y)
   y<-ifelse(y>=xtop,xtop,y)
   winvar<-var(y)
   winvar
}
```

Table 1.2: R function to compute winsorized variance.

The function can then be used to find the winsorized variance using the following R code.

```
> winvar(urb, tr=.2)

[1] 12831.72
```

Decisions about when to use a robust estimator such as a trimmed or winsorized mean and also about how much to trim or winsorize are not trivial tasks. Trimming or winsorizing 20% of the data is common, while for very heavy tailed distributions, 25% may be more appropriate. The interested student is referred to Rosenberger and Gasko (1983) or Wilcox (2005).

1.3 Research Question #3: Are there differences in the annual household per capita expenditures between the seven Vietnamese regions?

The Socialist Republic of Vietnam consists of 61 provinces (see Figure 1.7); that encompass seven regions of the country. These regions are represented by different colors on the map: from North to South, Northern Uplands (yellow), Red River Delta (light blue), North Central Coast (purple), Central Coast (red), Central Highlands (green), Southeast (brown), and Mekong Delta (blue).

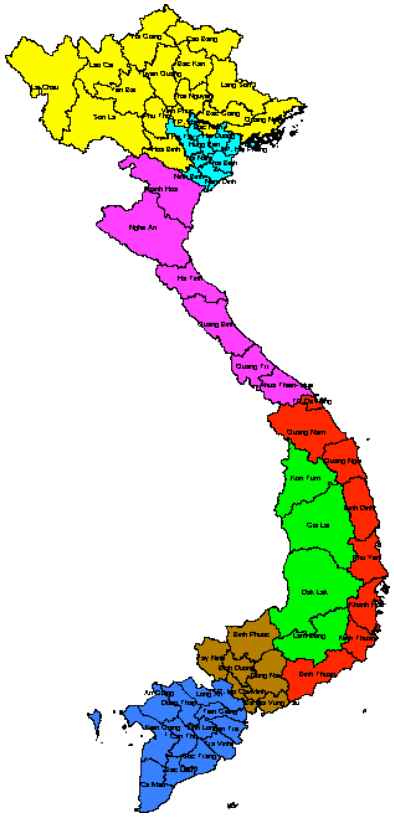

Figure 1.7: Map of Vietnam color coded by region.

To begin to examine our third research question, we would again want to examine a graphical summary of the data. With so many different groups it would be difficult to read a graph including seven plots of the kernel density estimates. It is more common to begin by examining side-by-side boxplots. Using the R code below we obtain the graph depicted in Figure 1.8. We will also add a horizontal line at the poverty line of $119.32.

```
> boxplot(dollar ~ regionName, xlab="Region",
      ylab="Expenditures per capita")
> abline(h=119.32, col="red")
```

Displaying Means along with Standard Errors

Often in statistical work, a graph of the sample means is provided as a visual depiction of the differences between multiple groups. Each of the means should also be accompanied by a measure of the variation inherent in the

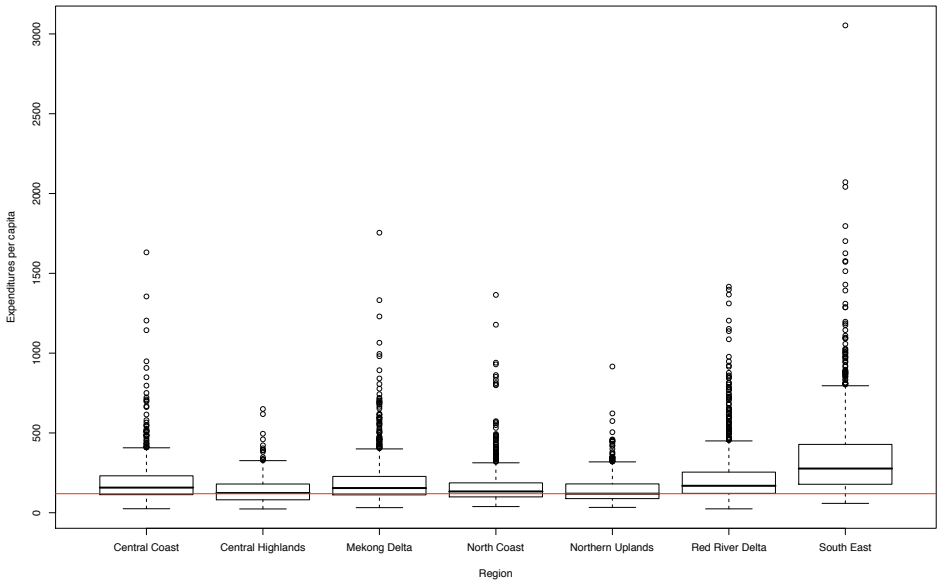

Figure 1.8: Side-by-side boxplots of per capita expenditures by region.

group (e.g., the SEM). The use of error bars helps us facilitate a more accurate interpretation of what is actually happening in the population. It is also in line with the guidelines suggested by the APA appointed Task Force on Statistical Inference, which suggest, "in all figures, include graphical representations of interval estimates whenever possible (Wilkinson, 1999, p. 601)." A plot of the region means and accompanying error bars is shown in Figure 1.9.

To create this graph using R, we can use the `plotmeans()` function found in the `gplots` library. The code used to create the error bar graph can be found below.

```
> library(gplots)
> plotmeans(dollar ~ regionName, data=household, connect=F, p=.95, n.label=F,
    xlab="Region", ylab="Expenditures per capita")
```

From the graphs depicted in Figure 1.9 and Figure 1.8, it appears that there are differences in the per capita expenditures between the regions. The South East region seems to have the highest household per capita expenditure. It is quite clear that the distribution of expenditures per capita in the Southeast region (which includes Ho Chi Minh City and hinterland) is typically higher and more variable, compared with other regions. Also, more than 75% of the people sampled from the South East region are above the poverty line. The

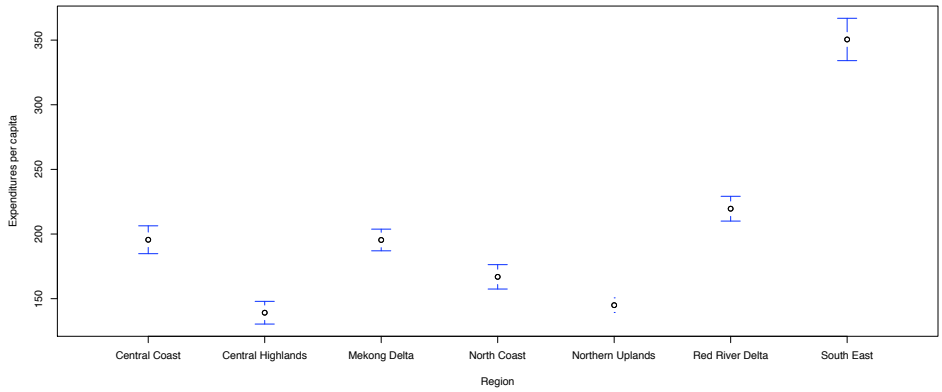

Figure 1.9: Means and error bars of per capita expenditures by region.

distributions in the other six regions seem fairly similar. It is worthwhile to note that in the two most impoverished regions, the Central Highlands and Northern Uplands, have close to 50% of their household below the poverty line. Based on our findings to the second research question, it is perhaps not surprising that both of these regions are primarily rural.