Graphical and numerical descriptive analysis: Exploratory tools applied to Vietnamese data

Dominique Haughton
Bentley College

Nguyớn Phong
General Statistical Office, Vietnam

Journal of Statistics Education Volume 12, Number 1 (2004),
ww2.amstat.org/publications/jse/v12n1/haughton.html

Copyright © 2004 by Dominique Haughton and Nguyớn Phong, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: MISSING KEY WORDS.

Abstract

This case study covers several exploratory data analysis ideas, the histogram and boxplot, kernel density estimates, the recently introduced bagplot - a two-dimensional extension of the boxplot - as well as the violin plot, which combines a boxplot with a density shape plot. We apply these ideas and demonstrate how to interpret the output from these tools in the context of data on living standards in Vietnam. The level of the presentation is suitable for an upper-level undergraduate or beginning graduate course in applied statistics. We use data from the Vietnam Living Standards Survey of 1998 (VLSS98) and from the 2000 Vietnam statistical yearbook, the statistical package Stata, and special programs provided by the authors who introduced the bagplot and the violin plot.

1. Intended audience and contents of the paper

The exploratory tools presented in this paper are suitable for students in an upper-level undergraduate or beginning graduate course in applied statistics. The article can be used as reading material with the objective of understanding the tools presented and their interpretation. In addition, the article can be used in a workshop format where participants actually manipulate the data and create their own plots. The data are provided in the form of MS Excel files and references to the necessary software tools are given.

Section 2 gives an introduction to Vietnam and the Vietnam Living Standard Surveys. Section 3 discusses the issue of how many classes to use in a histogram. Kernel density estimators are presented in Section 4. Section 5 is devoted to boxplots and their two-dimensional extension - bagplots. Section 6
2. Introduction to Vietnam and the Vietnam Living Standards Surveys

With China to the North, Laos and Cambodia to the West and the South China Sea to the East, Vietnam is a country about the size of Italy, or New Mexico in the United States.

Figure 1 goes here

The country is often referred to as a bamboo pole with rice baskets at each end: the Red River Delta in the North, and the Mekong Delta in the South. These deltas are among the most densely populated parts of the world. Hanoi is the capital of Vietnam, and Ho Chi Minh City (formerly Saigon) the largest city in the country.

Vietnam (officially the Socialist Republic of Vietnam) consists of 61 provinces (see Figure 1); in this paper we will refer to the seven regions of the country, 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).

In 1986, the government began a policy of doi 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 largest 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 1993) used a multistage cluster sampling procedure on two strata (rural and urban areas) to select a nationally representative sample of 4,800 households. The second of these surveys, VLSS 1998, kept as many households as possible from VLSS 1993 (4,305 households) and used a multistage cluster sampling procedure on 10 strata to augment the sample to 5,999 households. The 10 strata consist of Hanoi and Ho Chi Minh City, medium size towns, smaller towns, and rural parts of the seven regions of Vietnam mentioned above. For further details on the sampling methodology and the surveys see for example Haughton, Haughton and Phong (2001). The data are generally considered to be of high quality and have been used in a number of studies and publications. A Google search on “Vietnam Living Standards Surveys” gives a sense of the range of uses of the data.

3. The number of classes in a histogram

The familiar histogram can be thought of as a filter which filters out noise to display the underlying distribution of the data. The amount of filtering is determined by the class width: at one extreme a minimal class width implies no grouping and thus no filtering; at the other extreme a maximal class width (equal to the range of the data) filters out everything, so the histogram consists of a single block, clearly of no use.
The matter of selecting the number of classes to be used in the histogram can be delicate and in fact has an element of art in it. Quite often, the class width is chosen subjectively, using knowledge about the context of the data. Many statistical packages use a default number of classes (equivalently a default class width), but this default is fairly arbitrary and often has limitations. Some introductory statistics textbooks recommend using about classes where \( n \) is the sample size, within a minimum of 5 and a maximum of 15 classes. However, these rules are meant to be useful only for samples sizes between 25 and 200.

We have restricted ourselves in this paper to histograms with a common class width. There are however cases where variable widths are useful, as for example in the case of a skewed distribution where narrow classes are suitable for the peak and wide classes are suitable for the tail.

---

**Figure 2 goes here**

Figure 2 shows on an example the effect on the shape of the histogram of choosing 50, 40, 30, 20 and 10 classes. The variable we use is the age of individuals (in years) from the 1998 Vietnam Living Standards Survey (28,633 observations, ages in years ranging from 0 to 99).

From Figure 2, we can see that 50 classes give a clear picture of three generations, the young, the middle-aged and the older individuals. A choice of 40 classes creates some artificial peaks, for example the third and the fifth classes. This is because since the width of each class is 2.475 (99/40), the second class, ranging from 2.475 to 4.95, contains only children aged 3 and 4 while the third class, ranging from 4.95 to 7.425, contains more children since children aged 5, 6 and 7 fall in this class. With 50 classes, each class covers two years of age, so we do not see artificial peaks such as appeared with 40 classes. With 10 classes, the three groups are not clearly visible.

Two interesting features emerge out of the histogram with 50 classes. The second bump 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. This decrease in fertility rates has been documented elsewhere (see Haughton et al., 2001).

The question arises as to whether three subpopulations might occur in the distribution of ages. The population pyramid based on the Vietnam 1999 census does seem to indicate the presence of three subpopulations (see General Statistical Office, 2001), but this is of course based on an arbitrary class width of 5 years used in the pyramid. Wand (1997) proposes a series of “plug-in” rules for selecting the width (and therefore the number) of the classes in a histogram, which are guidelines to help choose an appropriate amount of filtering. The justification for the rules lies in the fact that the resulting histogram provides an estimator for the density which is closest to the true density in the mean-integrated-square-error sense. The rules are relatively complicated to compute, but we can get a first estimate by using the zero-stage rule which defines the width of the classes as \( \sqrt{\frac{7S}{IQR}} \), where \( S \) is the sample standard deviation and IQR is the inter-quartile range. The zero-stage rule coincides with the normal scale class width selection proposed by Scott (1979). Compared to the other rules proposed by Wand, the zero-stage rule probably overestimates the class width slightly, but is still a significant improvement over default rules proposed by most statistical packages (Wand 1997, p. 63). In our case, the zero-stage rule yields a width of about 2.314, thus about 43 classes. This choice of width yields three bumps in the histogram.
A good choice of a number of classes in this example is probably 20 or 50. Note that appropriate choices of class widths may very well depend on the position of the classes (equivalently on the choice of the lower limit of the first class).

4. An extension of the histogram: kernel density estimators

A kernel density estimator is a natural development of the histogram which gives an estimation of the density of a random variable.

Given a data set X1, X2, ..., Xn, a naïve estimator of the density can be written as falling in , which equals , where if /2, 0 otherwise. By extension, a kernel density estimator is defined as , where K is a function, called kernel, which integrates to one on the real line, often a symmetric probability density such as the normal density.

Note that a naïve estimator differs from a histogram. For example, consider the data set

1, 3, 3, 5, 6, 8, 9, 13, 14, 17, 18

and let the class width h be 4. For x = 1, 2 ..., 20, the frequencies for classes centered at x are given by:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 3 2 3 2 2 2 2 2 1 0 1 2 2 1 1 2 2 1 0

while the class frequencies for a histogram with class width 4 are given by:

[0 , 4) 3 [4 , 8) 3 [8 , 12) 1 [12 , 16) 2 [16 , 20) 1

so that the density - as estimated by the histogram - is constant over each interval such as [0 , 4), etc.

Note that the kernel must be chosen as well as the width h so the kernel density has a limitation similar to that of a histogram in that respect. The smaller h is, the more details are shown on the graph of the kernel density; in other words, the wider the classes, the more the data are filtered. The default width chosen for example by the Stata kdensity command is a function of the sample size, as well as of the spread and variability of the data. It is given by: , where , S is the sample standard deviation and IQR the inter-quartile range. Silverman (1986, p. 48) states that this choice of h, for a Gaussian kernel, will yield a mean integrated square error (between the true and estimated densities) within 10% of the optimum choice for a wide range of true densities. The Epanechnikov kernel (described below) yields a kernel density which is very close to that arising from a Gaussian kernel, so the default choice of h above is likely to be suitable in most cases. Note however that in long tailed distributions, it can be difficult to select a width which gives enough detail in the main part of the distribution but not too much detail in the tail.

A choice of kernel recommended in the literature is the above mentioned Epanechnikov kernel, which is a concave quadratic function with zeros at . This kernel is the most efficient in minimizing the error when approximating the true density by the kernel density (see Silverman, 1986, Stata Corporation, 1999).

In Figure 3 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 width of 2.344 (default width) clearly shows three age groups: the young, middle and old generation. The width of 0.5 gives us probably more peaks than the true density of the age, but still shows three groups. The third kernel density with a width of 5 still
gives us a shape of three groups, but not as clearly as the default width does. And with a width of 10 these three groups almost disappear.

---

**Figure 3 goes here**

---

**Figure 4 goes here**

---

Figure 4 gives kernel density estimators separately for the three age groups defined as: up to 28, from 29 to 52, and 53 and above.

On the matter of the selection of the bandwidth $h$, we note a paper by Chiu (1991) which considers the least squares cross-validation estimator of the optimal bandwidth and finds that this estimator tends to be too small, and therefore tends to show fictitious details in the density (Chiu, 1991, p. 1884). Chiu proposes a stabilized bandwidth selector which performs much better than the cross-validation estimates but is quite complicated to compute. Such techniques should be considered when the detailed shape of the density is in doubt and of significant importance.

We also note that the data are always discrete so that plots of kernel densities are - technically speaking - bar graphs, just as histograms are. In our example, the points are joined by drawing straight line segments between consecutive points; the overall result (except when the width equals 5) has the appearance of a smooth curve.

---

**Figure 5 goes here**

---

Kernel densities are particularly useful when comparing two or more groups: in Figure 5, we plot on the same graph the kernel densities for the real per capita expenditure of urban and rural households from the 1998 VLSS, using the Stata default widths (424.45 for urban households, 161.60 for rural households). It is quite clear on the graph that the distribution of real 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.

We added to the graph a vertical line positioned at the poverty line (established at 1,789.871 thousand VND for 1998 by the General Statistical Office, where VND refers to Vietnamese Dong, the Vietnamese currency; one US $ equals approximately 15,000 VND at the time of writing). 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 quite a lot of households to a position above the poverty line. This explains the recent relatively 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
5. Boxplots and bagplots

5.1 Boxplots

In addition to histograms which attempt to uncover the shape of the density function which generated the data, other plots can be drawn to give different views on the distribution. The most familiar example is the boxplot or box-and-whisker plot.

Figure 6 goes here

Boxplots are particularly useful when comparing several subgroups. In Figure 6 we show boxplots of real per capita expenditure from the 1998 VLSS for households in each of the seven regions of Vietnam, from left to right on the plot, Northern Uplands, Red River Delta, North Central Coast, Central Coast, Central Highlands, Southeast, and Mekong Delta. Among other features, it is quite clear that the distribution of real per capita in the Southeast region (which includes Ho Chi Minh City and hinterland) is shifted upwards and more variable (the box is wider), compared with other regions. The distributions in other regions are fairly similar; the Southeast region is the only region which clearly stands out.

5.2 Two-dimensional boxplots: bagplots

Bagplots, introduced by Rousseeuw, Ruts and Tukey (1999) are two-dimensional versions of the boxplot, and make it possible to visualize the location, spread, skewness and outliers of two variables together. In Figure 7a, we show a scatter plot with unidimensional boxplots of the real per capita expenditure (on the y axis) versus annual crop land (in square meters, on the x axis) for 4,230 rural households with non missing values for annual crop land; in order to better display the main part of the distribution for annual crop land and real per capita expenditures, we have excluded households with annual crop land greater than 60,000 square meters or real per capita expenditure greater than 15,000,000 VND.

Figure 7a goes here

In Figure 7b, we show the same scatter plot, but this time with its bagplot. The "+" symbol in the lower left part of the graph is the depth median, the two-dimensional equivalent of the one-dimensional median; it is defined as the point with the highest halfspace location depth (see Rousseeuw, Ruts and Tukey (1999)). The halfspace location depth of a point is the smallest number of data points contained in any closed half-plane with boundary line through the point. If there is no unique point with highest halfspace location depth, the depth median is defined as the center of gravity of the deepest region.
The inner bag (shaded) contains about 50% of the data (we provide more details and an example below) and corresponds to the box in a boxplot. The outer bag (bounded by a dotted curve) is obtained by inflating the inner bag by a ratio of 3 relative to the depth median, and roughly corresponds to the (usually not printed) fences in a boxplot. The value of 3 is relatively arbitrary, although it was decided by Rousseeuw et al. (1999) on the basis of simulations. Observations outside the outer bag are defined as outliers, although they can be quite numerous, as is the case in our example. The bagplot was calculated with Fortran and Matlab programs provided by Rousseeuw et al. (1999); these programs are available on .

Figure 7b goes here

Note that alternatives to the bagplot are discussed in Rousseeuw et al. (1999) section 7. The bagplot has a more general shape than the relplot and quelplot of Goldberg and Iglewicz (1992), which construct elliptic or quarter-elliptic plots. Other bivariate boxplots which rely on depth measures or convex hull peeling are referred to in Rousseeuw et al. (1999). Finally, a reference is made there to plots of highest density regions (Hyndman, 1996). Such plots are considered to be well-suited to distributions with multiple modes, but are not extensions of the one-dimensional boxplot (Rousseeuw et al., 1999).

We can see that the bivariate variable (annual crop land, real per capita expenditure) is right skewed, with 50% of the observations concentrated in the lower left corner of the graph, where the annual crop land and real per capita expenditure variables are low. At the same time, we can see a "funnel shape" in the scatter plot; the relationship between real per capita expenditure and annual crop land is quite heteroskedastic.

It is also clear from Figure 7a that the association between real per capita expenditures and annual crop land is tenuous. This means that there is more to household wealth than land; households without land have to look elsewhere to make a living.

A striking feature in Figure 7b is how small the inner bag is, indicating that there is a lot of equality in real per capita expenditure and land. The image of a "poor household on a tiny plot of land" still holds, a legacy of how land was redistributed in modern Vietnam.

The inner bag is defined as follows. Following Rousseeuw et al. (1999), let # denote the cardinality of a set and let us define the depth region as the set of data points with halfspace location depth greater than or equal to k. One first determines the value of k for which # = [n/2] < # , where [x] denotes the largest integer less than or equal to x. One then interpolates linearly between and relative to the depth median. The exact algorithm, which is intuitively simple, but quite awkward to describe, is given in detail in a technical report which preceded the article by Rousseeuw et al. (1999) and is available at on . To illustrate the process, Figure 8 shows a data set consisting of the vertices of the square and the octagon and of the central point. The sample size is 13; the set consists of the whole dataset (cardinality 13), the set ( = = ) consists of the red square and the middle point (cardinality 5), and the set ( = = ) consists of the middle point (cardinality 1; the depth median). In this case k = 2; the green bagplot lies about half-way between and , and covers about half the data points.
6. Boxplots combined with density shape plots: violin plots

Recent attempts have been made to modify boxplots to also show the shape of the density. The violin plot introduced by Hintze and Nelson in 1998 features a boxplot (with box and whiskers, but no outliers) together with a plot of the “naïve estimator of the density”, also called density trace, defined at the beginning of Section 4. As in the usual boxplot, the whiskers are extended to the farthest points within 1.5 inter-quartile ranges from the 25th and 75th percentiles. Violin plots can be produced with the software NCSS (Number Cruncher Statistical Systems, www.ncss.com). In Figure 9 below, we show a violin plot of the secondary school graduation rates in the 61 provinces of Vietnam from the 2000 Statistical Yearbook (GSO 2001).

Of course the question of the choice of h arises in the plotting of the density trace; the NCSS software uses a default of 40% of the range of the data for twice h, but following the zero-stage rule mentioned in Section 2 would suggest about 20% of the range for twice h. The latter is the bandwidth we used in Figure 9: at least three groups - and possibly a fourth group - of provinces appear in Figure 9, represented by bumps, as well as a long tail. The top group of provinces has secondary school graduation rates centered at about 95%, while the rates for the other groups are centered at about 87%, 77% and 70% with an outlier at 57%. The kernel density estimate produced with Stata’s defaults (see Figure 10) seems to confirm these features. Note that the density shape is plotted twice (symmetrically) on the violin plot to make it easier to visualize the magnitude of the bumps, but that the violin plot is a univariate plot.

We give in Table 1 the values of the secondary school graduation rates (the data read one column after the other) for all 61 provinces, sorted from the highest to the lowest rate. Overall, graduation rates decrease as one moves from North to South. The lowest rate occurs in the province of Gia Lai, a poor isolated province in the Central Highlands with a relatively high population of ethnic minorities.
7. Conclusion

We have introduced several graphical and numerical exploratory tools in the context of a data set on living standards in Vietnam. We have shown how histograms and plots of kernel density estimators can be used to uncover underlying patterns in distributions. We have discussed how boxplots can be used to show regional differences in the distribution of household real expenditure per capita for the Vietnamese households in our sample. We also demonstrated how two-dimensional boxplots known as bagplots can help uncover interesting features of a bivariate distribution. We presented violin plots which combine information found in boxplots and density shape plots. This article can be used in two ways: reading only, in order to become familiar with the tools presented and their interpretation; or reading together with manipulation of the data used in our results, in order to also obtain some hands-on experience with the tools. The data are provided with the article, in the form of MS Excel files. The file agedata.xls contains the ages in years of 28,633 individuals; the file segrad.xls contains the secondary school graduation rates (in percent) for the 61 provinces of Vietnam, and the file expcap1.xls contains the following variables: a household number (househol), expenditure per capita in ‘000 VND (exppcap), region, with regions coded 1 through 7 (region), urban (1 if urban, 0 if rural), and annual crop land in square meters (croplandpa). The regions are Northern Uplands (1), Red River Delta (2), North Coast (3), Central Coast (4), Central Highlands (5), South East (6) and Mekong Delta (7).

Acknowledgements

The authors would like to thank the Editor, Associate Editor, and three anonymous referees for their very careful reading of the paper, and their many useful comments.

References

Chiu, S.T. (1991), “Bandwidth Selection for Kernel Density Estimation”, Annals of Statistics, 19, 1883-1905.

General Statistical Office (2001), Statistical Yearbook. Hanoi, Vietnam: Statistical Publishing House.

Goldberg, K.M. and Iglewicz, B. (1992), “Bivariate Extensions of the Boxplot”, Technometrics, 34, 307-320.

Haughton, D., Haughton, J. and Phong, N. (editors) (2001), Living Standards during an Economic Boom: the Case of Vietnam, Hanoi, Vietnam: Statistical Publishing House, and United Nations Development Programme.

Hyndman, R.J. (1996), “Computing and Graphing Highest Density Regions”, The American Statistician, 50, 120-126.

Hintze, J. and Nelson, R.D. (1998), “Violin Plots: A Box Plot-Density Trace Synergism”, The American Statistician, 52, 181-184.

Rousseeuw, P., Ruts, I. and Tukey, J. (1999), "The Bagplot: A Bivariate Boxplot", The American
Statistician, 53, 382-387.

Scott, D.W. (1979), “On Optimal and Data-based Histograms”, Biometrika, 66, 605-610.

Silverman, B. (1986), Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

Stata Corporation (1999), "The kdensity command", Stata Manual for Release 6, 2, 144-151.

Tukey, J. (1977), Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Wand, M.P. (1997), “Data-based Choice of Histogram Bin Width”, The American Statistician, 51, 59-64.

World Bank (1999), World Development Indicators 1999 CD-ROM. Washington DC.

Dominique Haughton
Bentley College
USA

Nguyôn Phong
General Statistical Office
Vietnam