CPH 636 – Spring 2009 – Dr. Charnigo

Lecture 3

Exploring continuous variables

Motivation. Examining numerical and graphical summaries for the continuous variables in your data set can alert you to outlying observations and pronounced departures from normality, which may signal problems that need to be addressed before data mining. [[We also include here those cardinal variables that are not continuous but that are so finely discretized that they may be treated as if they were continuous.]]

Measures of location or central tendency. Let $x_1, x_2, \ldots, x_n$ be the sample values for the variable we are looking at, where n is the sample size. Measures of location or central tendency describe what value is "typical" for the variable we are looking at. You have already seen some of these in your introductory statistics course; others will be new.

Arithmetic mean. The arithmetic mean (or, more simply, "the mean") is denoted $\bar{x}$ and defined by

$$\bar{x} := \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \cdots + x_n}{n}.$$

The arithmetic mean can be influenced by extreme values, sometimes so much so that it no longer describes what value is typical.

Median. The median is a number such that about 50% of the sample values are smaller and about 50% are larger. If n is even, the median is the average of the $(n/2)^{th}$ and $(n/2+1)^{th}$ ordered observations. If n is odd, the median is the $(n/2+1/2)^{th}$ ordered observation.

Mode. The mode is the most frequently occurring sample value. If no sample value occurs twice, then we can either say that the mode does not exist or that every value is a mode.

Geometric mean. The geometric mean is defined as

$$(x_1 \times \cdots \times x_n)^{1/n},$$

which is equivalent to

$$\exp\left[\frac{\log x_1 + \cdots + \log x_n}{n}\right].$$

In effect, you apply a log transformation to the sample values, compute the arithmetic mean, and then undo the log transformation. Of course, for this to make sense, the sample values should be positive.

Harmonic mean. The harmonic mean is defined as

$$\frac{n}{1/x_1 + \cdots + 1/x_n}.$$

In effect, you apply a reciprocal transformation to the sample values, compute the arithmetic mean, and then undo the reciprocal transformation. Again, the sample values should be positive.

Winsorized mean. The Winsorized mean is like the arithmetic mean except that the most extreme sample values are truncated. For example, any sample value above the 97.5^{th} percentile is replaced by the 97.5^{th} percentile, while any sample value below the 2.5^{th} percentile is replaced by the 2.5^{th} percentile. The truncation prevents the Winsorized mean from being unduly influenced by extreme sample values (unless there are a lot of them). [[The p^{th} percentile is a number such that about $p\%$ of the sample values are smaller while about $(100 - p)\%$ of the sample values are larger. Not all people agree on how to compute percentiles. Various rules have been proposed, one of which is as follows: Order the sample values from smallest to largest, letting k be the largest integer less than or equal to $(np/100)$, and take the $(k + 1)^{th}$ value if $(np/100)$ is not an integer. Take the average of the k^{th} and $(k + 1)^{th}$ values if $(np/100)$ is an integer.]]

Trimmed mean. The trimmed mean is like the arithmetic mean except that the extreme sample values are thrown out.

Discussion questions. Suppose that seven Olympic judges assign ratings of $8.3, 9.4, 9.5, 9.5, 9.5, 9.6, 9.7$ and that the athlete's score is calculated as the average of the middle five ratings, $(9.4 + 9.5 + 9.5 + 9.5 + 9.6)/5 = 9.5$. What measure of central tendency is being used here? Why is it (arguably) better than the arithmetic mean?

Suppose that a person invests $\$5,000$ in the stock market and gets annual returns of 2%, 1%, and 21%. In three years, the investment will be worth

$$\$5,000 \times 1.02 \times 1.01 \times 1.21 = \$6,233.$$

Clearly, the arithmetic mean of the annual returns is 8%, yet

$$\$5,000 \times 1.08 \times 1.08 \times 1.08 = \$6,299 > \$6,233.$$

What measure of central tendency would be less misleading as a summary of the annual returns? What is its value? [[Rewrite $1.08 \times 1.08 \times 1.08 > 1.02 \times 1.01 \times 1.21$ as $\log 1.08 > (1/3) \log 1.02 + (1/3) \log 1.01 + (1/3) \log 1.21$. If you have taken an advanced course in probability, can you prove this inequality without using a calculator?]]

Measures of dispersion or variability. These describe how spread out the sample values are.

Range. The range is simply the maximum sample value minus the minimum sample value. The range can be influenced by extreme values and thus is not too useful as a measure of variability, although the range may be useful as a check for errors in data entry.

Discussion question. How can the range help us find errors in the data?

Interquartile range. The interquartile range is denoted IQR and defined as the difference between the 75^{th} percentile (also called the third quartile, Q_3) and the 25^{th} percentile (also called the first quartile, Q_1). The interquartile range is often reported in tandem with the median.

Variance. The variance is denoted s^2 and defined by

$$s^2 := \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Apart from the division by $(n-1)$ instead of by n, the variance is the average of the sample values' squared deviations from the mean. The variance can be inflated by extreme values.

Standard deviation. The standard deviation is denoted s and defined as the (positive) square root of the variance. Whereas the mean identifies a typical sample value, the standard deviation identifies what size deviation from the mean is typical.

Interpreting measures of dispersion. If the distribution of sample values is approximately normal, then about 68% of the sample values will fall within one standard deviation of the mean and about 95% will fall within two standard deviations of the mean. Moreover, the interquartile range will be roughly 1.35 times the standard deviation.

Skewness and kurtosis. Skewness, a measure of asymmetry in the distribution of sample values, is defined as

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3/n}{\hat{\sigma}^3},$$

where $\hat{\sigma}^2$ is like s^2 but has n instead of $(n-1)$ in the denominator,

$$\hat{\sigma}^2 := \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}.$$

[[Thus, if the data arise from a normal distribution, $\hat{\sigma}^2$ is the maximum likelihood estimate of the population variance.]] When values to the right of the median are spread out more than values to the left, then the skewness is positive. When values to the left of the median are spread out more than values to the right, then the skewness is negative. In practice, positive skewness is much more common than negative skewness. When the distribution of sample values is approximately normal, then the skewness will be close to zero.

Kurtosis is defined as

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4/n}{\hat{\sigma}^4}.$$

A large kurtosis indicates that the variation in sample values comes mainly from having a few extreme deviations from $\bar{x}$, while a small kurtosis indicates that the variation in sample values comes mainly from having lots of moderate deviations. When the distribution of sample values is approximately normal, then the kurtosis will be close to 3.

Discussion questions. Name some variables for which you anticipate positively skewed data. Name some variables for which you anticipate negatively skewed data. [[If you have taken an advanced course in probability, what can you say about the kurtosis for data with a near-uniform distribution?]]

Introducing our illustrative example. In what follows, I will refer to {RCLIB.saheart1.rtf}. I obtained it by applying Fernandez's UNIVAR macro to the data set in {saheart.sas7bdat}. A few tips for running the macro are in {SASMacroInstr2.txt}. [[See also pages 60-63 of the textbook.]]

Let me give some background on {saheart.sas7bdat}. [[See also Hastie et al, page 100, and {http://www-stat.stanford.edu/~tibs/ElemStatLearn}.]] The data come from a retrospective study conducted in a part of South Africa with an elevated incidence of heart disease. [[A retrospective study is one in which a dichotomous response variable represents disease and is observable at the time the study is conducted. The investigators take a random sample of people with the disease and a random sample of people without the disease. Then they try to relate the disease to explanatory variables such as demographics, behavioral risk factors, and environmental risk factors. A retrospective study may be contrasted with a prospective study, in which the investigators take a random sample of currently-healthy people with an exposure and a random sample of currently-healthy people without the exposure. Then they follow the people forward in time and monitor in which sample the disease develops more frequently.]] The dichotomous response variable is "chd" (coronary heart disease), coded as 1 for those with coronary heart disease and 0 for those without. Other variables in the data set include "sbp" (systolic blood pressure), "tobacco", "ldl" (low-density lipoprotein cholesterol), "adiposity", "famhist" (family history), "typea" (type A behavior), "obesity", "alcohol", and "age".

In {RCLIB.saheart1.rtf} are results for "sbp" stratified by "chd". That is, systolic blood pressure values are examined both for those with coronary heart disease and for those without. The immense number of pages in {RCLIB.saheart1.rtf} is somewhat misleading, as there is some wasted space.

Box plots. Page 1 of {RCLIB.saheart1.rtf} shows side-by-side box plots for systolic blood pressure. The first box plot is for those without coronary heart disease; the second is for those with coronary heart disease. In a box plot, the edges of the box are the 25^{th} and 75^{th} percentiles. The median is marked with a horizontal line segment, and the mean is marked with a plus sign. Whiskers extend from the box to the smallest and largest values that are not deemed to be outliers (i.e., the smallest value not less than $Q_1 - 1.5IQR$ and the largest value not greater than $Q_3 + 1.5IQR$). Values deemed to be outliers are marked with red dots and labeled (albeit not too legibly) by subject number.

Histograms. Page 2 shows histograms with normal density curves superimposed. Each curve represents a normal distribution with the same mean and variance as the sample values. Hence, if the distribution of sample values is approximately normal, the superimposed curve will provide an excellent fit to the histogram.

Control charts. Pages 3 through 12 provide “control charts”. A control chart is a favorite graphical tool for companies that manufacture items. A company identifies some important characteristic of the items (e.g., length) and then monitors that characteristic with a control chart. The manufacturing process is suspected to be out of control if there are values outside the control limits or if there are too many consecutive values below or above the mean. However, in public health and medicine, sample values often pertain to distinct and unrelated subjects. Hence, even if there is a temporal structure to the sample values, there is not necessarily any underlying process that may be out of control. Because of this, I think that you can ignore the control charts.

Numerical output from the UNIVAR macro. Page 13 provides measures of location for those without coronary heart disease, and page 14 provides the “five-number summary”: minimum value, Q_1, median, Q_3, maximum value. Page 15 provides measures of variability. You can ignore pages 16 through 18. Page 19 identifies potential outliers based on a “Student” statistic that Fernandez does not define in his book and that I have not heard of before. As best as I can tell, the Student statistic is proportional to $(x_i - \bar{x})$ but not exactly equal to $(x_i - \bar{x})/s$. Fernandez wrote the macro to flag any observation for which the Student statistic is greater than 2.5 in absolute value.

Normal probability plot. Page 20 provides a normal probability plot in which the Student statistics are plotted against theoretical percentiles of a standard normal distribution in such a way that, if the distribution of sample values is approximately normal, the plot will resemble a straight line. Westward curvature at the bottom and northward curvature at the top indicate positive skewness. Southward curvature at the bottom and eastward curvature at the top indicate negative skewness. Southward curvature at the bottom and northward curvature at the top indicate greater than normal kurtosis. Westward curvature at the bottom and eastward curvature at the top indicate less than normal kurtosis.

Tests for departures from normality. Page 21 presents results of tests for skewness different from 0, kurtosis different from 3, and nonnormality. While making such inferences is not our focus, these tests are useful because they can warn us when there are pronounced departures from normality. That said, these tests can be very sensitive: very small departures from normality may be identified when n is large. Getting a perspective from the normal probability plot is essential.

Identifying outliers. Page 22 identifies outliers based on cutoffs of $Q_3 + 1.5IQR$ and $Q_1 - 1.5IQR$. Pages 23 through 26 present numerical summaries if these outliers are removed, while pages 27 through 31 present numerical summaries if the largest 2.5% and smallest 2.5% of the sample values are removed. Page 32 presents the Winsorized mean.

Remark. The output from the UNIVAR macro is organized differently if you do not stratify, but I think that you will be able to understand the parts that you need based on the descriptions above. If not, please ask.

Pragmatics. If there are several continuous variables in your data set, you don't want to spend an inordinate amount of time examining every piece of output from the UNIVAR macro for each variable. In the absence of specific directions otherwise (e.g., those that you may receive on a Written Assignment in CPH 636), my practical recommendations are as follows:

• If your response variable is continuous, carefully examine output from the UNIVAR macro for your response variable.

• If there are not many other continuous variables (say, no more than three others), carefully examine output from the UNIVAR macro for each.

If there are several other continuous variables (say, between four and eight), run the UNIVAR macro for each, but be brief in your examination of the output. Check the box plots and the normal probability plots, from which you can get an idea about whether the distribution of sample values is approximately normal and whether there are outliers you need to be concerned about.

If there are lots of other continuous variables (say, more than eight), run the UNIVAR macro only on variables for which the minimum and maximum values seem out of line. You can quickly check minimum and maximum values in an Excel spreadsheet with the "MIN" and "MAX" commands in Excel or in SAS with PROC MEANS.

• Extreme outliers that are suspected mistakes should be investigated, if possible, and then corrected, if appropriate. Extreme outliers that cannot be investigated and corrected may be removed from the data set, but I advise against removing mild outliers that are not suspected mistakes.

• If the distribution of sample values is positively skewed for one of your continuous variables, then you may want to consider a transformation to reduce the skewness. Transformations that reduce positive skewness include the logarithm and the square root. [[Functions with positive first derivatives and negative second derivatives are likely to help with positive skewness. Functions with positive first derivatives and positive second derivatives are likely to help with negative skewness.]]

I cannot give a universal rule about when to transform and when not to transform, but factors that may be considered are: the degree of skewness, whether you are looking at the response variable or another variable, and how much a transformation may complicate subject matter interpretations.

• Stratification may be helpful if the distribution of sample values for the continuous variable under examination differs across strata. In that case, failure to stratify may result in some low values being declared outliers that are not really outliers (merely below average for their stratum, for which the center of the distribution is lower) and in some high values being declared outliers that are not really outliers (merely above average for their stratum, for which the center of the distribution is higher).

Of course, the potential benefit of stratification has to be weighed against the increased amount of output that will be produced. I think that the case for stratification is strongest when the continuous variable under examination is the response variable or when the variable by which you want to stratify is the response variable. In any event, you should avoid creating too many strata and avoid creating strata that have very few or no observations.

Exploring categorical variables

Introduction. Exploration of categorical variables is much easier. You do not have to think about normality or outliers. Moreover, in many data sets the number of categorical variables is modest. For the most part, you just want to get a sense of how often certain values occur. Of course, you also want to see if there are any obvious mistakes, which will manifest as categories that should not exist (e.g., a category 99 when there should be only category 0 and category 1).

Illustrative example. Refer to {RCLIB.saheart3.rtf}, which, to our relief, is much shorter than {RCLIB.saheart1.rtf}. I applied the FREQ (or FREQUENCY) macro to the data set in {saheart.sas7bdat}. A few tips for running the macro are given in {SASMacroInstr2.txt}. [[See also pages 53-55 of the textbook.]] I am specifically focusing on the "chd" variable and stratifying by "famhist".

Pages 1 and 2 reveal the proportions of individuals in the sample for whom "famhist" = Absent and for whom "chd" = 0. The confidence intervals and hypothesis tests are not of particular interest for data mining. Page 3 reveals the proportions of individuals in each of $4 = 2 \times 2$ groups determined by the variables "famhist" and "chd". Again, the hypothesis tests are not of particular interest for data mining. Pages 4 through 10 include several graphical displays that are largely self-explanatory. There are some redundancies in the graphical displays.