Paper 187-2009

Bubble, Bubble: Less Toil, No Trouble

Wendy Dickinson, Ringling College of Art and Design, Sarasota, FL
Constance Hines, University of South Florida, Tampa, FL
Bruce Hall, Professor Emeritus, University of South Florida, Palm Harbor, FL

ABSTRACT

In 1801, the idea of using a circular image to portray frequency and amount was formalized by Playfair. These first "statistical charts" illustrated the number of inhabitants in major European cities by a set of hand-drawn circles, with the diameter of the circle indicating population thus enabling effective comparison by city. Today's bubble charts enable the same effective comparison of frequency of phenomena of interest but with two pronounced improvements. The first contemporary improvement is the addition of a location matrix- Cartesian coordinate system- to enable display of the bubble plots in relation to a second variable.

More importantly, the second improvement is the ability to use computers to draw and replicate the bubble plots quickly and inexpensively. Utilizing SAS/GRAPH and PROC GPLOT to construct bubble plots saves both time and money. The resultant bubble plots effectively communicate contemporary research results with both simplicity and elegance, reaching a diverse research audience with information needed for vital decision-making processes.

INTRODUCTION

In 1801, the idea of using a circular image to portray frequency and amount was formalized by Playfair. These first "statistical charts" illustrated the number of inhabitants in major European cities by a set of hand-drawn circles, with the diameter of the circle indicating population thus enabling effective comparison by city. Today's bubble charts enable the same effective comparison of frequency of phenomena of interest but with two pronounced improvements. The first contemporary improvement is the addition of a location matrix- Cartesian coordinate system- to enable display of the bubble plots in relation to a second variable.

More importantly, the second improvement is the ability to use computers to draw and replicate the bubble plots quickly and inexpensively. Utilizing SAS/GRAPH and PROC GPLOT to construct bubble plots saves both time and money. The resultant bubble plots effectively communicate contemporary research results with both simplicity and elegance, reaching a diverse research audience with information needed for vital decision-making.

By design, bubble plots are especially effective for categorical data display due to their visual size differential. In this presentation we provide examples, using data on births and deaths, to show how this visual size differential is linked to corresponding data values, resulting in accessible and readily understandable graphical outputs. By utilizing bubble plots to display categorical data, we create a more effective visual means of transmission for our research results, thus enhancing decision-making and the interpretation of research outcomes.

CONTEXT

PROPERTIES OF NOMINAL OR CATEGORICAL DATA

Researchers in the social, behavioral and health sciences often need to contend with data representing such variables as race, ethnicity, gender, marital status, occupation, state of residence, academic major, blood type, religious affiliation, political affiliation, and presence or absence of some characteristic. These are all, in fact, nominal or categorical variables (Defays, 1988, p. 316), and the process of assigning numbers to elements of these variables is the process of *nominal measurement*.

Nominal measurement may be looked upon as the process of grouping units (objects, persons, responses, etc.) into classes or categories so that all of those in a single class are equivalent, or nearly so, with respect to some property or attribute. The classes are then assigned numbers for identification. With nominal (categorical) scales, the assigned numbers define each distinct grouping of the attribute and serve merely as a substitute for labels or names. When measurement is nominal, we use only the uniqueness property of numbers; e.g., "1" is distinct from "2," so if object or response A is coded using a "1" and object or response B is coded using a "2", then A and B are different with respect to the attribute. The numbers serve to make categorical distinctions only; each distinct number represents a different category. The magnitude of the numbers does not reflect any inherent ordering of the objects, or distance among the objects, to which the numbers are assigned (Glass & Hopkins, 1996; Hopkins, Hopkins, & Glass, 1996).

According to McDonald (1999), with categorical variables the only properties of the categorical scheme in correspondence to those of numbers are the following:

Either A = B or it is not.
If A = B, then B = A
If A = B and B = C, then A = C

Within the above mapping rules, we may assign the same number (e.g., "1") to objects or responses A and B if and only if they are in the same equivalence class. While the interpretations placed upon nominal or categorical scale values are clearly limited, it does not follow that such numerical labeling is of little practical significance. Many of our most critical variables in the social, behavioral and health sciences, and in education, are categorical in nature. Often our data collection consists exclusively of categorical attributes. In studying one such attribute, we may be seeking information for purely descriptive purposes, or we might want to test a hypothesis about the expected frequency of each category value. In a multivariate situation—where each object or person is measured on two or more categorical attributes—we may want to determine whether or not the attributes are related to one another in some way (Kachigan, 1986).

As Friendly states, "categorical data means different things in different contexts", including "types of categorical variables, data in case form versus frequency form, frequency data vs. count data, and the distinction between explanatory (predictor) and response (criterion) variables "(p. 2, 2000). By design, bubble plots are especially effective for categorical data display due to their visual size differential. This visual size differential is linked to the corresponding data values, resulting in explanatory graphical output.

METHODS

DATA SOURCE

To create the bubble plots, two extensive national datasets were utilized: natality (births), and mortality (deaths). The data sources were the Centers for Disease Control and Prevention (CDC), and the National Center for Health Statistics (NCHS). Using the natality dataset from the National Vital Statistics Reports (NVSS), bubble plots were generated using combinations of variables, including census region, state of residence, Hispanic origin of mother, gender of baby, age of mother, ethnicity of mother, and frequency counts of births. Using the mortality data set, bubble plots were generated using combinations of variables including age, gender, cause of death, year, ethnicity, and census region. Only two samples of the bubble graph outputs are presented in this paper due to space constraints. The complete set of graphical outputs are displayed via the poster presentation.

Table 1 shows the natality dataset variables used within the graphing algorithm for natality bubble plots, and the mortality dataset variables used within the graphing algorithm for mortality bubble plots.

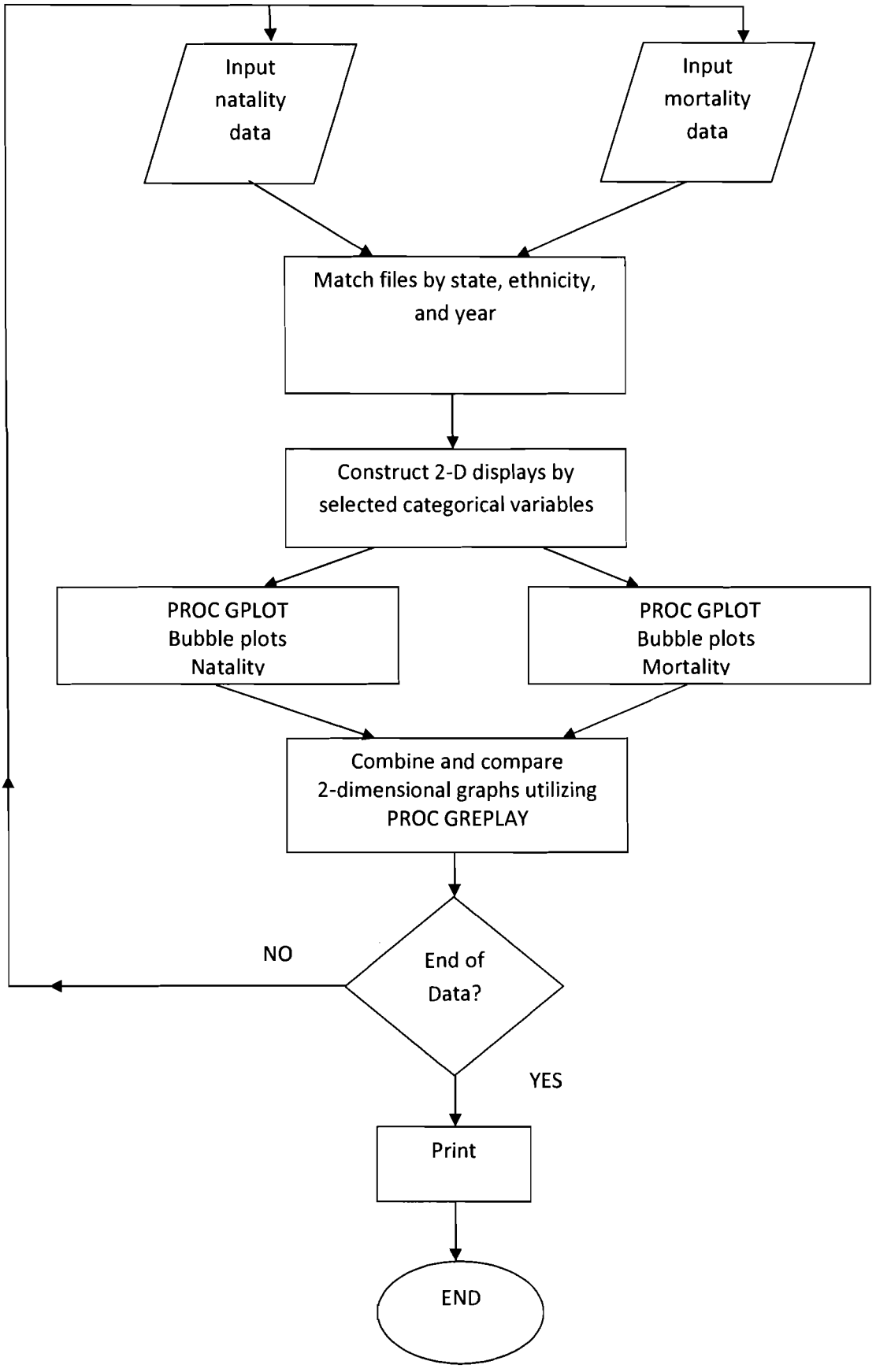
Table 1. SAS/Graph algorithm variables for natality and mortality bubble plots

SAS Variable name	Type of variable	Variable measure	Data base
AVGBIRTH	continuous	average number of births	natality
LIVBIRT	continuous	live births	natality
BIRTHRAT	continuous	birth rate per 1,000	natality
RACE	categorical	race as defined by OMB	natality, mortality
GENDER	categorical	gender	natality, mortality
AGE	continuous	age in years	natality, mortality
STATE	categorical	state of birth, death	natality, mortality
WEEKDAY	categorical	day of week	natality, mortality
AVGDEATH	continuous	average number of deaths	mortality
DEATHCAUS	categorical	cause of death	mortality
MORTRATE	continuous	mortality rate per 1,000	mortality

Bubble plots were generated within the PROC GPLOT procedure. The GPLOT procedure "plots the values of two or more variables on a set of coordinate axes (X and Y), and also "generates bubble plots in which circles of varying proportions representing the value of a third variable are drawn at the data points" (SAS Institute, p.801, 1999).

The resultant bubble plots code was created by using the BUBBLE statement and incorporating BUBBLE statement options, including the AXIS and FORMAT statements. By further defining axis characteristics, bubble appearance, plot appearance, and specifying custom titles and footnotes, the bubble plot output graphs are optimized for display and communication. The algorithm structure to construct the bubble plots is displayed in Figure 1, Natality and Mortality: PROC GPLOT algorithm flowchart.

Figure 1 Natality and Mortality: PROC GPLOT algorithm flowchart



RESULTS

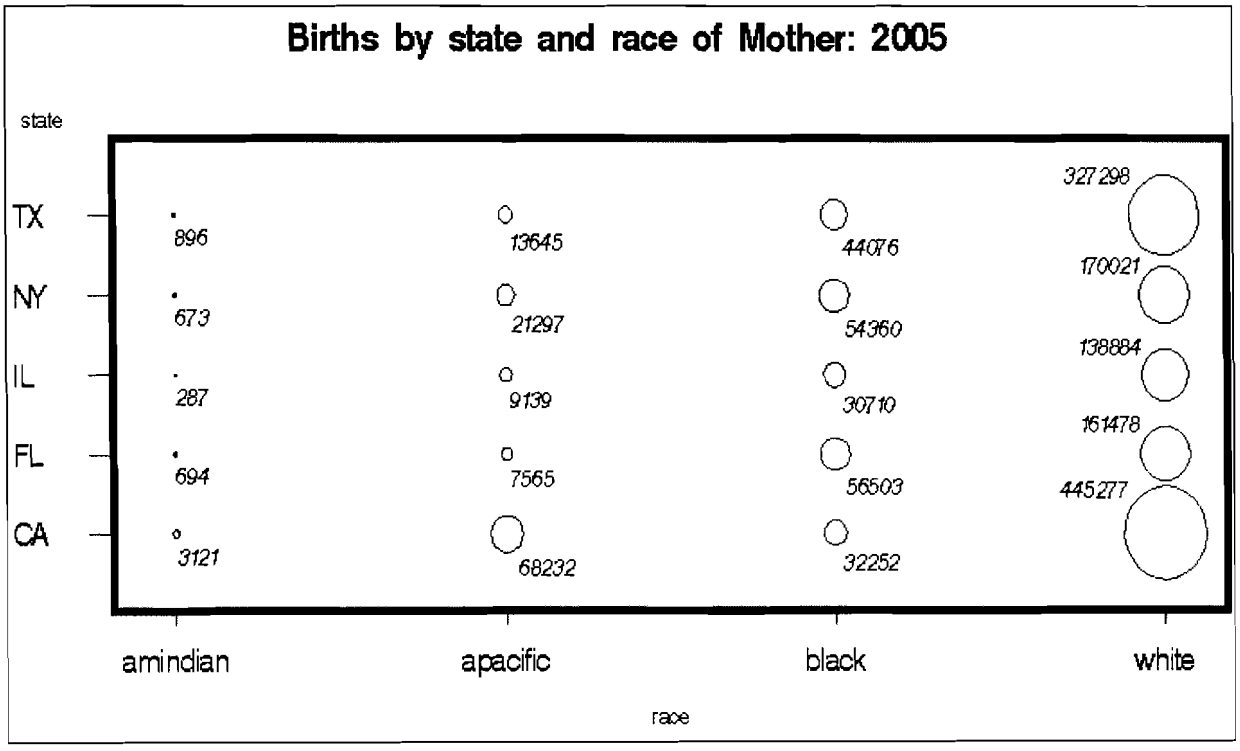
EXAMPLES OF BUBBLE PLOTS: NATALITY

The SAS code for Figure 2, Births by state and race of mother, is shown below. The resultant graph uses the four categories of race as defined by the 1977 Office and Management Budget (OMB) standards: American Indian, Asian-Pacific Islander, Black, and White. The five states with the highest birth frequencies were selected for inclusion.

```
**Natality by state and race, SAS 9.1 **
**Bubble plots for CDC/NCHS mortality data**
**SGF 2009 **
**Dickinson, Hines, Hall**;
************************************************
**reset the graphics environment to new specs **
************************************************;
        goptions reset = all gunit= pct border cback = white
        colors = (black blue green red orange brown)
        ftitle= swissb ftext= swiss htitle = 4 htext = 3;

**Data source: National Vital Statistics Reports, V.56, No.6, 12-05-2007';
**Table 12, Live births by race of mother: United States, 2005';
**graphing top 5 states for births, 2005**;

        data cdc;
                input state $ race $ births;
                datalines;
                ...
                ;
        title1 'Births by state and race of Mother: 2005';
                axis1 offset = (5,5)
                        width = 5
                        value = (height= 4);
                axis2 offset = (10,10)
                        major = (height = 2)    minor = (height = 3)
                        width  = 5
                        value = (height = 4);

        proc gplot data = cdc;
                        bubble state*race = births/haxis = axis1
                        vaxis = axis2   vminor = 1
                        bcolor= blue
                        blabel
                        bsize = 12
                        bfont = swissi
                        caxis = red;
                        run;
                        quit;
```

Note that in Figure 2, the **blabel** option causes each bubble to be labeled with the number of births associated with that particular observation. Additionally, the use of the **bsize = 12** option provides a scaling factor of 12 to be utilized when drawing the bubbles. The diameter of the bubbles will be multiplied by a constant factor to increase the size of each bubble within the graphical output, yet maintain the correct proportion (numeric relationship) of data-to-bubble.

Figure 2. Births by state and race of mother



EXAMPLES OF BUBBLE PLOTS: MORTALITY

The SAS 9.1 code for Figure 3, Top 10 Causes of Death: Percent by Gender, is shown below.

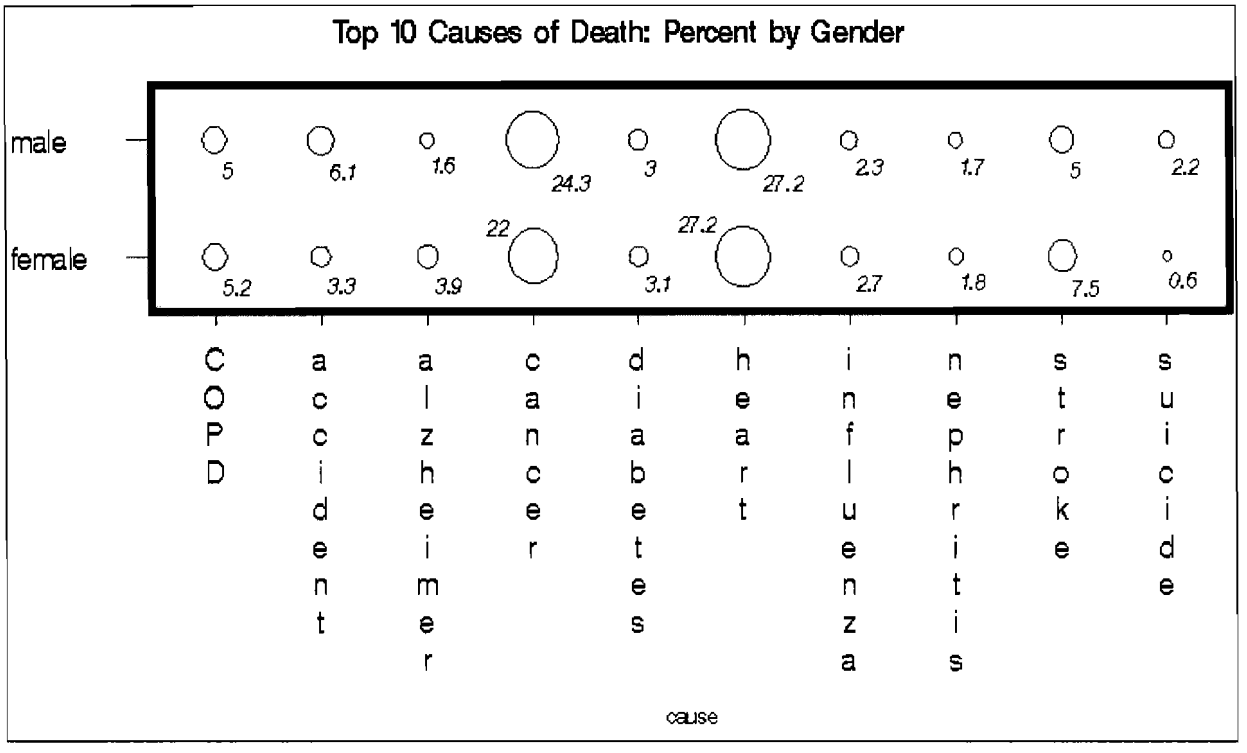
```
**Causes of Death, SAS 9.1 **
**Bubble plots for CDC/NCHS mortality data**
**SGF 2009 **
**Dickinson, Hines, Hall**;
************************************************
**reset the graphics environment to new specs **
************************************************;
goptions reset = all gunit= pct border cback = white
        colors = (black blue green red orange brown)
        ftitle= swissb ftext= swiss htitle = 4 htext = 3;

**Data source: Top 10 causes of death by gender: 2004"**
** National Vital Statistics Reports, V. 56 No. 5, 11-20-2007**
** Table D, page 9**
** Cause of death codes from ICD-10: International Causes of Disease, Tenth Revision**;

        data cdc;
                length cause $9;
                input gender $ cause $ deaths percent;
                datalines;
                ...
                ;
        title1 'Top 10 Causes of Death: Percent by Gender';
        *footnote1 h= 2 j = 1 'Source: National Vital Statistics Report, 2007';
                axis1 offset = (5,5)
                        width = 5       value = (height= 4);
                axis2 offset = (7,7)
                        label = none
                        major = (height =2)   minor = (height = 3)
                        width  = 5      value = (height = 4);

                proc gplot data = cdc;
                        bubble  gender*cause = percent/haxis = axis1
                        vaxis = axis2    vminor = 1
                        bcolor= blue
                        blabel
                        bfont = swissi
                        bsize = 10
                        caxis = red;
                        run;
                        quit;
```

Note that the use of the **axis1 offset** and **axis2 offset** statements provides "spacing" for the bubbles to keep them from intersecting the x and y axes. The use of the bubble plot to display data works particularly well with these categorical variables gender and cause of death. It is easy for the viewer to visually compare the numbers and cause of death by looking at the size of the bubble.

In Figure 3, we see the top ten causes of death displayed by cause of death and gender. The ten most frequently reported causes of death, as recorded by the Centers for Disease Control, are heart disease, malignancies (cancer), accidents, cerebrovascular incidents (stroke), COPD and respiratory diseases, diabetes, influenza and pneumonia, suicide, nephritis, and alzheimers disease. Note while the top ten reported causes of death remain the same for both males and females, the actual reported percentages can differ by category of cause.

Examples of gender differences by cause of death include a lower percentage of females (0.6%) committing suicide than males (2.2%); and a higher percentage of males (6.1%) dying as the result of accidents than females (3.3%). A notable exception is the incidence of heart disease, which yielded the same percentage (27.2%) for both genders in 2005.

Figure 3 Top 10 causes of death: Percent by gender



IMPLICATIONS AND IMPORTANCE

As these examples have illustrated, in the public health arena, where timely dissemination of data and treatment strategies is essential for public education, bubble plots can be utilized to quickly and efficiently transmit information to both the trained and the naïve viewer. Bubble plots may also be valuable for data dissemination and interpretation in the social and behavioral sciences where categorical variables within large data sets are common.

By using bubble plots, we can visually compare data values across a variety of variables. This visual comparison provides a way for the viewer to detect patterns and trends over time contained within the data set. Thus, bubble plots can help our data "talk" to us without words, providing a powerful tool for both interpretation and dissemination of research results.

REFERENCES

Defays, D. (1988). Scaling of nominal data. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook*. Oxford, England: Pergamon Press.

Friendly, M. (2000). *Visualizing Categorical Data*. Cary, NC: SAS Institute, Inc.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd Ed.). Needham Heights, MA: Allyn and Bacon.

Heron, Melonie, Centers for Disease Control and Prevention (CDC) (November, 2005). Deaths: Leading causes for 2004, 56(5). http://www.cdc.gov/nchs/data/nvsr/nvsr56/nvsr56_06.pdf

Hopkins, K. D., Hopkins, B. R., & Glass, G. V. (1996). *Basic statistics for the behavioral sciences* (3rd Ed). Needham Heights, MA: Allyn and Bacon.

Kachigan, S. K. (1986). *Statistical analysis: An interdisciplinary introduction to univariate & multivariate methods*. New York, NY: Radius Press.

Kung, H., Hoyert, D., Xu, J. , Murphy, S. (April, 2008). Deaths: Final Data for 2005, 56(10), http://www.cdc.gov/nchs/data/nvsr/nvsr56/nvsr56_06.pdf

Martin, J., Hamilton, B., Sutton, P., Ventura, S., Menacker, F., Kirmeyer, S., Munson, M., Division of Vital Statistics, Centers for Disease Control and Prevention (CDC), December, 2007. Births: Final Data for 2005, 56(5). http://www.cdc.gov/nchs/data/nvsr/nvsr56/nvsr56_06.pdf

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Playfair, W. *Commercial and Political Atlas and Statistical Breviary (1801)*. (2005). Edited by H. Wainer and I. Spence, Cambridge University Press: New York, New York.

SAS /Graph Software: Reference, Version 8 (1999). Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

The Authors greatly appreciate the assistance of Lisa Adkins in the preparation of this document.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the Authors via:

Dr. Wendy Dickinson
Liberal Arts Program
Ringling College of Art and Design
2700 North Tamiami Trail
Sarasota, FL 34239

Email: wdickins@ringling.edu