What Can 'CSI' Teach Us about Statistical Literacy?

Jane E. Miller, Ph.D.
Institute for Health, Health Care Policy and Aging Research, and the Edward J. Bloustein School of Planning and Public Policy at Rutgers University, 30 College Avenue, New Brunswick NJ 08901.

Abstract: What do the hit television show "CSI: Crime Scene Investigation" and statistical literacy have in common? Both involve using technical information as evidence in an inquiry, asking whether the scientific or numeric facts support a particular hypothesis. In this paper, I show how to apply standard expository writing techniques to quantitative writing: introducing the topic or question, describing individual facts in context, and finally relating the entire body of evidence back to the original question. I explain how to step back from the procedural details of conducting forensic chemistry tests or calculating inferential statistics to reveal how those results answer the question under study. I demonstrate several basic principles for effective quantitative communication, integrating numeric information with good expository writing. Using these guidelines, statisticians and other quantitative writers will learn to conduct and present a coherent statistical inquiry.

KEYWORDS: Expository writing; inferential statistics; statistical literacy.

What do the hit television show "CSI: Crime Scene Investigation" and good quantitative writing have in common? Both involve using technical information as evidence in an inquiry, asking whether the scientific or numeric facts support or refute a hypothesis. On 'CSI,' the criminalists (as the investigators are called) use a range of laboratory tests to conduct a coherent scientific investigation. Tying together evidence from forensic chemistry, ballistics, DNA analysis, and trace chemical analysis, the investigators use established scientific approaches to put together a convincing case that the suspect committed the crime of which he is accused. They use technical language when speaking with one another, but must be able to translate their findings into plain English to explain them to juries of nonscientists.

Likewise, people who write about statistics should seek to conduct a coherent statistical inquiry, using bivariate and multivariate inferential statistics to test a hypothesis about the relationship between the concepts under study. When conducting the tests and talking with other statisticians, they too use a specialized vocabulary. Unfortunately, few courses teach how to write a clear narrative linking numeric evidence to substantive questions, or how to explain the answers in words that non-statisticians can comprehend easily. As a consequence, many people present statistical results in ways that leave their audience struggling to understand what questions those numbers are intended to answer or what conclusions they support.

In this paper, I show how to apply expository writing techniques to enhance statistical literacy. Tracing a single research question, I show how to introduce the question, describe individual facts in context, and finally relate the entire body of evidence back to the original question. I demonstrate several basic principles for describing numeric patterns: introducing your topic; reporting and interpreting numbers; specifying direction and magnitude of an association, summarizing patterns; writing a conclusion; and explaining a chart to an audience". I illustrate these principles with examples of "poor" and "better" descriptions – samples of ineffective writing annotated to point out weaknesses, followed by concrete examples and explanations of improved presentation. Using these guidelines, statisticians and other quantitative writers will learn to conduct and present a coherent statistical inquiry.

SIX PRINCIPLES FOR WRITING A CLEAR STATISTICAL INQUIRY

Introducing your topic

As with other types of expository writing, start by introducing the topic of your work and the questions you seek to answer with the numbers that follow. If a criminalist starts talking about a specific fingerprint or piece of DNA without having first outlined the basic facts of the crime and its context, a jury will have a hard time understanding where that evidence fits in the overall case. Likewise, if you jump directly to presenting a chi-square statistic or comparing specific numbers without orienting your audience to your topic and objectives, it will be difficult for them to see what those numbers mean.

Begin with a good topic sentence, introducing the characters (variables, in a statistical paper) and setting the context with the W's (when, where, what). Word your introduction either as a statement:

> "Consider how birth weight varies by race and socioeconomic status in the United States around 1990."

or as a rhetorical question:

> "Do differences in socioeconomic status explain observed racial differences in birth weight in the U.S. around 1990?"

Notice that there are no numbers yet, just a statement that mentions the purpose and concepts behind the numbers to be described later in the paragraph.

Such introductions are especially important if you are presenting a series of charts and tables, each of which addresses one part of an overview of your topic. For instance, a lecture on birth weight might include information on trends across time, consequences of low birth weight for infant mortality, differences in birth weight by race and other characteristics, and finally, a multivariate analysis of birth weight. As you move from one topic to another, introduce it before presenting the associated numeric evidence.

Reporting and interpreting numbers

Reporting the numbers you work with is an important first step toward writing effective numeric descriptions. By including the numbers in the text, table or chart, you give your audience the raw materials with which to perform additional calculations or to compare your data with information for other times, places, or groups.

After reporting the raw numbers, interpret them - showing how they answer the question you have posed. An isolated number or scientific fact that has not been introduced or explained leaves its explication entirely to your audience. Those who are not familiar with your topic are unlikely to know which comparisons to make or to have the information for those comparisons immediately at hand. To help them grasp the meaning of the numbers you report, provide the relevant data and explain the comparisons.

> *Poor:* "In the late 1980's, mean birth weight for non-Hispanic black infants was 3,181 grams (about 7 lbs.; Miller 2005)."
>
> > *Comment:* From this sentence, readers can't assess whether mean birth weight for black infants was high or low, changing or stable. If they knew the rates for other racial groups or time periods, they could compare the figures, but you will make the point more directly if you do the calculation for them.
>
> *Better*: "In the late 1980's, mean birth weight for non-Hispanic black infants was 3,181 grams (about 7 lbs.), nearly 250 grams (half a pound) lower than mean birth weight for non-Hispanic whites (Miller 2005). This difference has been stable for more than two decades (Martin et al. 2002)."
>
> > *Comment*: This version reports the mean value for black infants and compares it against that for whites. The second sentence places those values in recent historical context.

Although it is important to interpret quantitative information, it is also essential to report the numbers. If you *only* describe a relative difference or percentage change, for example, you will have painted an incomplete picture. Suppose that a report states that the incidence of low birth weight (LBW: < 2,500 grams or 5.5 pounds) in some country is 30% greater than it was five years ago but does not report the rate of LBW for either year. A 30% difference is consistent with many possible combinations: 1.0 and 1.3 LBW infants per 1,000, or 10 and 13 per 1,000, or 500 and 650 per 1,000, for example. The first pair of numbers suggests a very low incidence of LBW, the last pair an extremely high rate. Unless the rates themselves are mentioned, you can't determine whether that nation has nearly eradicated low birth weight or faces a huge infant health problem. Furthermore, you can't compare LBW figures from other times or places.

Explaining direction and magnitude

Writing about numbers often involves portraying associations between two or more variables. To describe an association, explain both its shape and size rather than simply stating whether the variables are correlated. For instance, to compare birth weight by race, report *which* racial group weighs more as well as *how much* more. Well-chosen adjectives ("minuscule difference"), adverbs ("increased markedly"), and analogies to familiar shapes ("bell-shaped" or "J-shaped") can enhance a description of a pattern considerably. For statistically-oriented audiences, also report results of inferential statistical tests.

Direction of association. Variables can have a positive or direct association – as the value of one variable increases, the value of the other variable also increases; or a negative or inverse association – as one variable increases, the other decreases. In the U.S., as family income increases, so does mean birth weight; hence family income and birth weight are positively related. Conversely, as maternal cigarette smoking increases, infant birth weight decreases, so smoking and birth weight are negatively related.

For nominal variables such as gender, race, or religion that are classified into categories that have no inherent order, describe direction of association by specifying which category has the highest or lowest value. "Race is negatively associated with birth weight" cannot be interpreted. Instead, write "Non-Hispanic blacks had the lowest mean birth weight," and then mention how other racial groups compare.

Size of association. An association can be large – a given change in one variable is associated with a big change in the other variable – or small – a given change in one variable is associated with a small change in the other. A $5,000 increase in family income might increase birth weight by 50 grams or only 10 grams, depending on maternal smoking and other factors. If several factors each affect birth weight, knowing which make the biggest difference can help reduce the incidence of low birth weight.

> *Poor*: "Family income and birth weight are correlated."
>
> > *Comment*: This sentence doesn't say whether family income and birth weight are positively or negatively related, or how much weight differs by income.
>
> *Better*: "As family income increases, mean birth weight increases."
>
> > *Comment*: Although this version specifies the direction of the association, the size of the income difference in birth weight is still unclear.
>
> *Best*: "As family income increases, mean birth weight also increases but at a decreasing rate. For example, infants born into a family with income at twice the poverty level weigh about 50 grams more than an infant born to a family at the poverty level, whereas the difference between infants born into families at three and four times the poverty level is only about 10 grams."
>
> > *Comment*: This version explains both the direction, shape, and size of the income/birth weight pattern.

The size of a difference between two values can be calculated in any of several ways, including absolute difference (subtracting one value from the other), relative difference or ratio (dividing one value by the other), or percentage difference or change. To decide which of these alternatives to use as you write, read similar comparisons in the literature for your field. Miller (2004) provides guidelines for how to choose among, calculate, and write about each of these types of quantitative comparisons.

Summarizing patterns from tables or charts

If you only need to compare a few pairs of numbers to answer the question you are analyzing, the above principles will go a long way to improving the description of those contrasts. Often, however, answering a question requires describing a pattern involving many numbers, such as trends in prices for each of four products over several decades, or mean birth weight values for each of three racial groups over the typical range of family incomes in the United States.

The numbers you present, whether in text, tables or charts, are meant to provide evidence about some issue or question. However, if you only provide a table or chart, you leave it to your audience to figure out for themselves what that evidence says. Instead, digest the patterns to help them see the general relationship in the table or chart. When asked to summarize a table or chart, inexperienced writers often make one of two opposite mistakes: (1) they report every single number from the table or chart in their description, or (2) they pick a few arbitrary numbers to contrast in sentence form without considering whether those numbers represent an underlying general pattern. Neither approach adds much to the information presented in the table or chart, and both can confuse or mislead the audience. Paint the big picture, rather than reiterating all of the little details. Describe the forest, not each individual tree. If someone is interested in specific values within the pattern you describe, they can look them up in the accompanying table or chart.

As you summarize, relate the evidence back to the substantive topic: Are trends in birth weight by income-to-poverty (IPR – a measure of socioeconomic status) consistent across racial/ethnic groups (Figure 1)?[1] Are there appreciable differences in mean birth weight across racial groups at a given level of IPR? Summarize broad patterns with a few simple statements instead of writing piecemeal about individual numbers, comparing many pairs of numbers, or describing each of several trend lines separately. For example, answering a question such as "does mean birth weight rise, fall, or remain stable as IPR increases?" is much more instructive than responding to "what was mean birth weight at IPR = 0.0, 1.0, …, 4.0 among non-Hispanic white infants?" or "how much does mean birth weight among whites change between IPR = 0.0 and 1.0? Between IPR = 1.0 and 2.0?..."

Here is a mantra I devised to guide you through the steps of writing an effective description of a pattern involving three or more numbers or facts: "generalization, example, exceptions," or "GEE" for short.[2] The idea is to identify and describe a pattern in general terms, give a representative numeric example to illustrate that pattern, and then explain and illustrate any exceptions. This approach can also be used to summarize findings of previous studies, identifying consensus and pointing out discrepancies.

"Generalization" For a generalization, come up with a description that characterizes a relationship among most, if not all, of the numbers. In Figure 1, is the gen-

[1] The income-to-poverty ratio (IPR) is defined as family income in dollars divided by the Federal Poverty Level (threshold) for a family of given size and age composition (Proctor and Dalaker 2003).

[2] Not to be confused with GEE = generalized estimating equation. After all, there are only so many acronyms to go around!

eral birth weight trend in most racial groups upward, downward, or level? Does one racial group consistently have the highest mean birth weight over the whole income range? Start by describing one such pattern (e.g., trends in mean birth weight among non-Hispanic whites) then consider whether that pattern applies to the other racial groups as well. Or, determine which racial group had the highest mean birth weight at IPR = 0 and see whether it also had the highest value at IPR = 2.0 and 4.0. If the pattern fits most of the groups most of the time, it is a generalization. For the few situations it doesn't fit, you have an exception (see below).

There are two generalizations of interest in Figure 1: the relationship of **each** independent variable (race or income) with the dependent variable (birth weight). So, we start with a verbal generalization about each of those patterns, which will serve as the topic sentences for separate paragraphs – one about racial differences in birth weight, the other about how birth weight changes with income.

> [Generalization #1]: "As shown in Figure 1, in every racial group, birth weight increased with family income.
>
> [Generalization #2]: At every income level, blacks weighed considerably less than whites."
>
>> *Comment*: Notice that although these sentences each convey direction of association, they don't include any numbers to assess size. That comes in the second step. Readers are referred to the accompanying chart, which depicts the relationships.

"Example" Having described your generalizable pattern in intuitive language, illustrate it with numbers from your table or chart. This step anchors your generalization to the specific numbers upon which it is based. It ties the prose and table or chart together. By reporting a few illustrative numbers, you implicitly show your audience where in the table or chart those numbers came from as well as the comparison involved. They can then test whether the pattern applies to other times, groups, or places using other data from the table or chart.

To illustrate the above generalizations about Figure 1, include sentences that incorporate examples from the chart into the description. For the racial difference generalization, pick one value of the income-to-poverty ratio and compared birth weight across racial groups holding IPR constant at that value.

> "For example, at IPR =1, black infants weighed roughly 200 grams less than their white peers."

For the income pattern generalization, choose one racial group and present the difference in birth weight across the IPR range.

> "Among whites, mean birth weight rose from about 3,000 grams at IPR = 0 to 3,200 grams at IPR = 4."

"Exceptions" Sometimes you will be lucky enough that the generalizations you have made capture all the relevant variation in your data. If you are working with data from the real world, however, often there will be important exceptions to the general pattern you have sketched. Tiny blips can usually be ignored, but if some parts of a table or chart depart substantially from your generalization, describe those departures.

When portraying an exception, explain its overall shape and how it differs from the generalization you have described and illustrated in your preceding sentences. Is it higher or lower? By how much? If a trend, is it moving toward or away from the pattern you are contrasting it against? In other words, describe both direction and magnitude of the difference between the generalization and the exception. Finally, provide numeric examples from the table or chart to illustrate the exception. In the case of Figure 1, the above generalizations about income and race fail to capture the changing rank order and mean difference in birth weight between whites and Mexican Americans as income increases.

> [To follow the above generalization]: "However, gains were much smaller among Mexican Americans than among blacks or whites. Although Mexican Americans weighed 70 grams more than whites at the low end of the income range, by the top of that range, the pattern was reversed, with whites weighing roughly 100 grams more than Mexican Americans."
>
>> *Comment*: The first sentence describes the exception and identifies the racial group to which it applies. The second sentence reports specific numeric examples to illustrate the consequences of the different rates of increase in birth weight by income for the relation between race and birth weight.

Other types of exceptions include a rising trend in each group but at a slower rate in some groups, or a sustained rise in some groups but an appreciable decline in others. In other words, an exception can occur in terms of magnitude (e.g., small versus large) as well as in direction (e.g., rising versus falling, or higher versus lower).

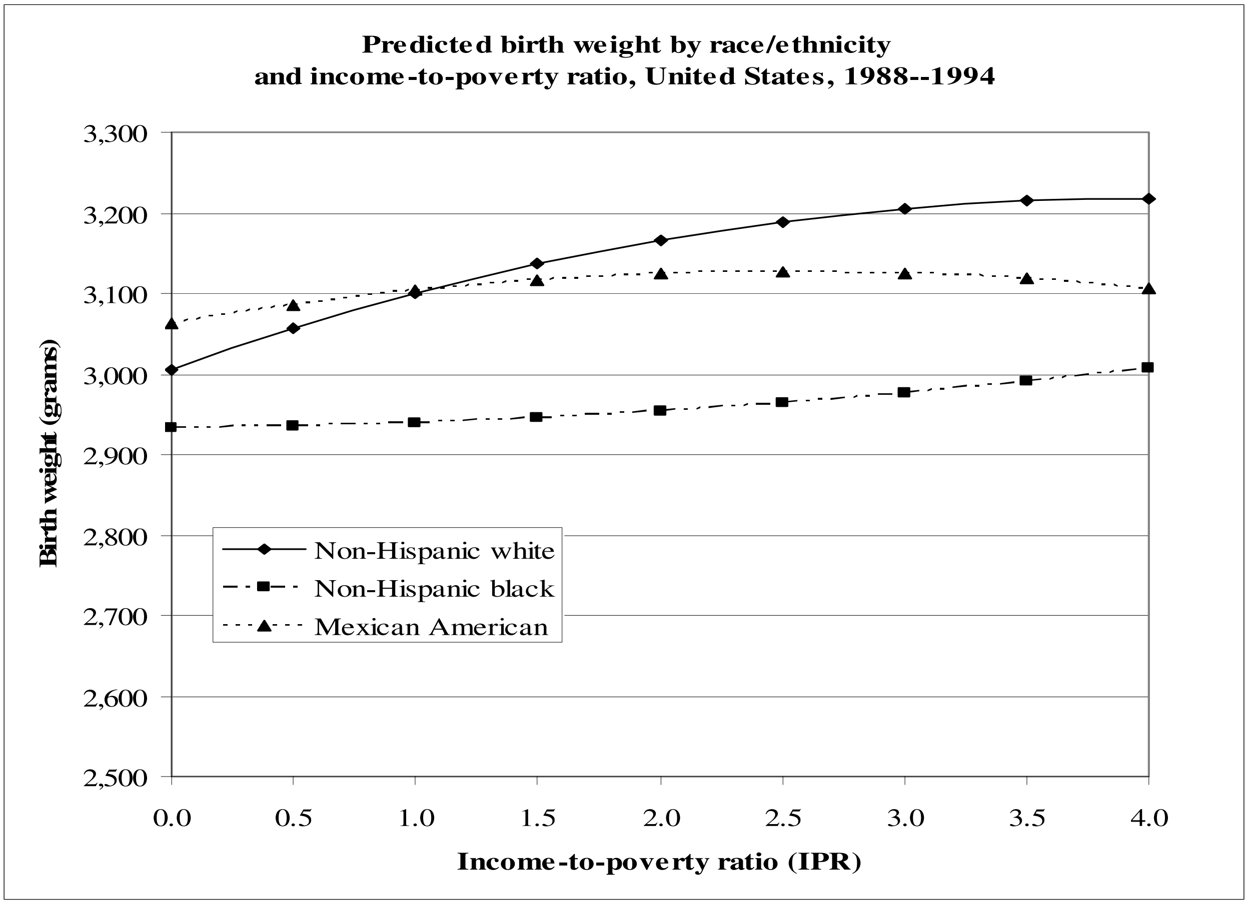

Figure 1.
Source: Miller 2005 with data from U.S. Department of Health and Human Services 1997.
Notes: (1) Predicted birth weight is based on an ordinary least square regression of birth weight with controls for gender, race, family income-to-poverty ratio, mother's age, educational attainment, and cigarette smoking. (2) The income-to-poverty ratio is defined as family income divided by the Federal Poverty Level for a family of comparable size and age composition.

The GEE approach can also be used to compare results across models that test similar hypotheses for different subgroups, time periods, dependent variables, or statistical specifications, or to synthesize findings or theories in previous literature. For instance, rather than slogging through a description of each of a dozen coefficients for a model of birth weight for males and again for an identically-specified model for females, point out which variables had similar associations with birth weight for both genders and which variables had different patterns for males than for females. For a systematic step-by-step approach to recognizing patterns and organizing the ideas for a GEE into paragraphs, see (Miller 2004).

Writing the conclusion

Having presented the individual pieces of evidence, write a summary to explain how that evidence answers the question you posed at the beginning of the paper, just as in the standard expository writing approach to writing an analytic essay.

> "Blacks have lower birth weight at all levels of income, hence SES differences alone do not explain observed racial differences in birth weight."

This sentence brings the analysis full circle, relating that evidence back to the original question. The conclusion could then be fleshed out with possible explanations for this pattern, a discussion of study strengths and limitations, and suggestions for future research on the topic.

Explaining a chart or table to a live audience

Tables, charts, maps, and other diagrams offer real advantages for presenting evidence, whether results of forensic tests at a trial, or results of statistical tests in a course lecture or conference presentation. Unfortunately, many speakers devote far too little time to describing such exhibits. They put up a slide with the table or chart, state "as you can see, . . ." and then describe the pattern in a few seconds before moving on to the next slide. As the slide disappears, many listeners are still trying to locate the numbers or pattern in question and have not had time to digest the meaning of the statistics.

Although it may appear to save time, failing to orient your audience to your tables or charts reduces the effectiveness of your talk. A criminalist who works every day with certain types of diagrams and tabular output from forensic tests knows exactly where to look and how to interpret the information shown on their exhibits. Likewise, if you designed a chart and wrote the corresponding lecture, you know it well enough to home in quickly on the exact number or table cell or trend line you wish to discuss. Give your viewers the same advantage by showing them where to find your numbers and what questions they address before you report and interpret patterns.

Introduce the topic. First, state the topic or purpose of the table, chart, or other diagram, just as you would in the introductory sentence of a written paragraph. Rather than reading the title from the slide, paraphrase it into a full sentence or rephrase it as a rhetorical question.

> "Figure 1 examines predicted birth weight by race and income level among U.S. women who gave birth in the late 1980's and early 1990's. In other words, 'do differences in family income explain observed racial differences in birth weight in the U.S.?'"

Explain the layout. Second, explain the layout of the table or chart. Don't discuss any numbers, patterns, or contrasts yet. Just give your audience a chance to digest what is where. For a table, name what is in the columns and rows. For a chart, identify the concepts and units on the different axes and in the legend, mentioning the color or shading of bars or line styles that correspond to each major group you will discuss. Also explain the purpose of features such as reference lines or regions, colors, symbols, or other annotations. (If you don't have time to mention such features, omit them to avoid distracting or confusing your viewers.)

Use a "Vanna White"[3] approach as you explain the layout, literally pointing out the applicable portion of the table or chart as you mention it. Point with a laser pointer, pen, or finger – it doesn't matter. The important thing is to lead your viewers' eyes across the key features of the exhibit before reporting or interpreting the information found there. At first this may seem silly or awkward, but most audiences follow and retain the subsequent description much more easily than if you omit the guided tour. Below, I use bracketed comments to describe the Vanna White motions that accompany the surrounding script; they are there to guide you, not to be spoken as part of the presentation. For Figure 1:

> "Across the bottom [wave left to right along *x* axis] is the income-to-poverty ratio or IPR, which is calculated as family income divided by the Federal Poverty Level for a family of comparable size and age composition. On the *y* axis is predicted birth weight in grams [gesture vertically along *y* axis]. Results are based on a multivariate model with controls for the variables listed at the bottom of the chart [point to note]. Non-Hispanic whites are shown with the solid line connecting circles. Non-Hispanic blacks are shown with the dashed line and squares, while Mexican Americans are shown with the dotted line and triangles [point to each line in turn. Note: for a color slide mention the line colors in lieu of the line styles].

Describe the patterns. Finally, having introduced your audience to the purpose and layout of the table or chart, describe the patterns it embodies. Use the GEE approach, starting with a general descriptive sentence followed by specific numeric examples and exceptions (where pertinent), as in the above description of Figure 1 under "Generalization, Example, Exception."

Again, gesture to show comparisons and point to identify specific values on the chart as you mention them. For example, when reporting numeric values to illustrate the trend generalization, coordinate the bracketed gestures with the associated script:

> "Among whites, mean birth weight rose from about 3,000 grams at IPR = 0 to 3,200 grams at IPR = 4" [wave along the solid line for non-Hispanic whites, pointing out the *y* values corresponding to *x* =0 and *x* = 4.0.]

To show your audience where the numbers for the across-race generalization come from, say:

> "For example, at IPR =1, black infants weighed roughly 200 grams less than their white peers." [Point to respective values on the dotted and solid lines above the *x* value 1.0.]

SUMMARY

In conclusion, both crime scene investigators and people who work with statistics face the often challenging task of weaving together complex scientific information into a form that answers the substantive question in a straightforward and clear fashion. Before either occupa-

3. The "Vanna White" moniker is in honor of the long-time hostess of the TV game show *Wheel of Fortune* who gestures at the display to identify each item or feature as it is introduced.

tion can perform their assigned task, indisputably they must master the technical aspects of their craft.

In this day and age, however, technology is widely used both for analysis of forensic evidence and calculation of statistical tests. When training all but the most technically oriented CSIs and statisticians, therefore, emphasize how to choose the appropriate test for the task and explain the results, rather than focusing solely on the technical details of conducting those tests. Just as few criminalists need to know how a particular DNA analysis technique was developed, few statisticians need to know how the log-likelihood function was derived. Just as most criminalists can rely on specialized machines to conduct chemical titrations or match fingerprints to a database, most statisticians will use computers to calculate a chi-square statistic or estimate a regression.

Moreover, technical forensic or statistical skills alone are not sufficient to convey results to an audience, whether a jury weighing forensic evidence or readers of an applied statistical paper. By keeping the focus on the choice of tests and the interpretation of results, both comprehension and communication of statistics can be improved. Aim to write about statistical findings so that readers understand what they mean, crafting a logical narrative with a beginning, middle, and end. In the introduction, ask the question in plain English, mentioning the specific concepts under study. In the body of the results section, systematically review the statistical evidence. Finally, close by answering the original question in everyday language. By applying these approaches, statisticians can learn to tell a clear story with numbers as evidence.

CHECKLIST FOR WRITING AN EFFECTIVE STATISTICAL INQUIRY

- Introduce your topic or question before presenting associated numeric evidence.
- Report and interpret numbers in the text:
  - Interpret the numbers, showing how they answer your main question.
  - Report them to allow readers to perform other calculations or compare to other data.
- Specify both the direction and size of an association:
  - If describing a difference across groups, which has the higher value and by how much?
  - If a trend, is it rising or falling? How rapidly?
- To describe a pattern involving many numbers, summarize the overall pattern rather than repeating all the numbers:
  - Find a generalization that fits most of the data, painting a verbal picture of the pattern's shape.
  - Report a few illustrative numbers from the associated table or chart.
  - Describe exceptions to the general pattern.
- Conclude your paper with a verbal summary of findings, relating the numeric evidence back to the question posed at the outset of the work.
- Use the 'Vanna White' technique to describe charts or tables to a live audience:
  - Paraphrase the purpose of the exhibit.
  - Explain the layout of the table (contents of rows and columns) or chart (contents of axes and legend), pointing to the pertinent elements as you mention them.
  - Describe the pattern, pointing to the illustrative numbers as you speak.

ACKNOWLEDGEMENTS

This paper is adapted from material in *The Chicago Guide to Writing about Numbers*. Chicago: The University of Chicago Press. Jane Miller can be contacted at Voicemail: (732) 932-6730; fax (732) 932-6872. Email: jem@rci.rutgers.edu

REFERENCES

Martin, Joyce A., Brady E. Hamilton, Stephanie J. Ventura, Fay Menacker, and Melissa M. Park (2002). "Births: Final Data for 2000." *National Vital Statistics Reports* 50 (5).

Miller, Jane E. (2004). The Chicago Guide to Writing about Numbers. Chicago: University of Chicago Press.

----. (2005). The Chicago Guide to Writing about Multivariate Analysis. Chicago: University of Chicago Press.

Proctor, Bernadette D., and Joseph Dalaker (2003). "Poverty in the United States: 2002." *Current Population Reports*. P60–222. Washington, DC: U.S. Government Printing Office.

U.S. Department of Health and Human Services (1997). National Health and Nutrition Examination Survey, III, 1988–1994. CD-ROM Series 11, no. 1. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.