Contents List

THE IMPORTANCE, EMPOWERMENT, AND TRANSFORMATION OF STATISTICAL ANALYSIS

Patricia B. Cerrito

Information is empowering. With electronic transmission of information, it is now possible for the general public to have ready access to knowledge previously limited to those designated as "expert" in the field. However, the privilege of access must be accompanied by the responsibility to use knowledge properly. Thus, empowerment requires the development of basic skills for finding, interpreting, and using information.

One of the essential basic skills for information processing is statistical literacy. Statistics has transformed virtually every discipline in the physical, medical, and social sciences as well as made significant inroads into the humanities (Jaeger, 1990). Statistics were originally developed to legitimize the social sciences as science. It has been so successful as a discipline that most research is not regarded as legitimate without the collection of data.

Statistical analysis is not without detractors. The physical sciences have been more resistant to the invasion of statistical techniques. Mathematicians, in particular, disregard statistics because of a lack of rigor in its theory. Because of the rift between mathematics and statistics, Tukey (1962) wrote a paper concerning the philosophy of statistics. He claimed that statisticians cannot serve two masters and must choose between the needs of data analysis and the needs of mathematical deduction. This was clearly demonstrated in Davis (1994, p132):

It is only to be expected that mathematics, the field where applications can be forgotten, should turn out to be the field where they most often are forgotten. So we are seeing more than a mere rejection of applications. The contemporary "pure" mathematician does not say only, I am too noble to get my hands dirty on mechanical problems like you mere engineers," but something even more hostile in its defensiveness: something like, "You are destroying my true science if you entangle me with your reality.

There are those in women’s studies who are also resistant to statistical analysis (Rryse, 1998; Harding, 1987; Mies, 1991). There is much emphasis upon qualitative analysis. As practiced, however, qualitative analysis is extremely intertwined with quantitative analysis. For example, Sacks, Wolffe, and Tierney, 1998) used a qualitative analysis to validate the results of their quantitative analysis, and still used quantitative measures such as Likert scales in its qualitative analysis. Similarly, the qualitative analysis of Polce-Lynch, Myers, Kilmartin, Forssmann-Falck, and Kliewer (1998) also depended heavily upon statistical methods.

Statistics as a discipline has been transformed by the needs for development in other areas. Certain techniques were needed in psychology (reliability, validity; Keren and Lewis, 1993) while others were needed in medicine (survival analysis, logistic regression; Bailar III and Mosteller, 1992). Some statistical methods generally cross all disciplines (analysis of variance; Geisser, 1992) while initially developed because of needs in one area (agriculture; MacNeill and Umphrey, 1987).

The computer explosion has required that statistics as a discipline develop techniques which can deal with the enormous amounts of information routinely collected and stored in electronic databases. One of the more exciting developments in recent years is that of data mining. High speed computers now permit the use of very complex algorithms to automate the process of hypothesis generation and pattern recognition. The techniques optimize the amount of information which can be gathered from a dataset, enabling researchers investigate individual rather than group characteristics.

In 1962, John W. Tukey published a paper in the Annals of Mathematical Statistics. Without stating a single theorem, it was the longest paper ever published in that journal. The purpose of the paper was to define the steps of data:

- Recognition of the problem
- Application of multiple analysis techniques
- Comparisons of efficacy
- Optimization

The process of data analysis requires a statistical judgment based upon

- experience of the particular field of subject matter from which the data come.
- a broad experience with how particular techniques of data analysis have worked out in a variety of fields of application.
- abstract results about the properties of particular techniques.

The judgment of the statistician is always subject to challenge and should be defensible. There are many confounding factors in observational studies which can explain the results obtained from the statistics; it is the responsibility of the researchers to put meaning to the statistical outcome. Data must be collected and analyzed to make a convincing argument. Yet a statistical argument is rarely convincing because of a denial of the statistical conclusions. Why are statistics mistrusted? Consider the following quotations:

"Figures won’t lie but liars will figure." (Charles H. Grosvenor)

"There are three kinds of lies: Lies, damned lies, and statistics." (Benjamin Disraeli)

"In God we trust; all others must have data." (W. Edwards Demming)

The first two indicate a real suspicion of statistics; the third indicates the way in which statistical analysis permeates all aspects of society. The main reason for this ambivalence is a lack of true understanding of statistical concepts. This tends to generate a belief that statistics can prove or disprove whatever the analyst wants to find. It also leads to a misuse of statistical parameters. Students often avoid statistics courses because of the mathematical content. "A little knowledge is a dangerous thing" and most people have only a little knowledge of statistics. Since statistical software is readily available, statistical values can easily be computed by anyone with little understanding for what they mean. This can cause severe misunderstandings. It is important that every individual become statistically literate and to understand when statistics are used correctly-or misused badly.

In a recent revision (1997) of the guidelines for promotion and tenure published by the College of Arts and Sciences at the University of Louisville, any faculty member who received below average student teaching evaluations was defined as deficient in teaching. Members of the faculty with quantitative skills pointed out that regardless of actual teaching ability, approximately half of the faculty would defined as regardless of performance and would remain deficient even if scores went up. The criterion was shortly rescinded. However, without statistically literate faculty members, the criterion would have been enforced without regard to the consequences.

Yet, when this same use of the average is translated to the medical field, it receives a general acceptance. The value of 200 is defined in the medical community as the upper level of normal blood cholesterol level. Studies are being conducted to determine if people with below average cholesterol levels should be treated (Rubins, 1995), and to determine just how aggressively children and adolescents should be treated for cholesterol levels (Benuck, Gidding, Donovan, 1995). When this is done, virtually everyone will have a cholesterol "disease". One report did issue a dissenting opinion (Newman, Garber, Holtzman, Hulley, 1995):

The expert Panel’s recommendations do not address important gender differences. Girls have higher average cholesterol levels than boys. They will therefore qualify for more dietary and drug treatment despite their lower age-adjusted risk of heart disease and the lack of association between cholesterol levels and cardiovascular mortality in women.

In other words, most of the cholesterol studies have been conducted on men and the conclusions based on the results involving men. The aggressive treatment of cholesterol has yet to have a demonstrated benefit in women. General consensus is not proof. However, without an ability to access the medical literature, an individual must rely on the recommendations provided by the "experts", i.e. the physician in clinical practice making recommendations for patients. A knowledge and understanding of statistical concepts coupled with an ability to access information which is now readily available empowers the individual to weigh the evidence to make decisions.

The book, The Bell Curve (Herrnstein and Murray, 1994) would not have been so widely discussed if people had a better understanding of the bell curve. Many wrote to challenge (Fischer, 1996; Kincheloe, Steinberg, Gresson III, 1996; Fraser, 1995; Wallace and Graves, 1995). However, the basic assumption of the book, that the general population of the United States can be represented be a bell-shaped curve, is false. It is only valid when examining one, homogeneous population. As soon as multiple heterogeneous populations are considered, the underlying distribution of the population is multi-peaked. Those involved with diversity issues need a solid understanding of this statistical issue.

Consider a population which can be divided into two subpopulations. For example, consider the parameter of height and a population of adult men and adult women. If the two subpopulations are graphed separately, the curves are represented as in Figure 1. When the data are merged and graphed, the result has two peaks (Figure 2). Men and women occur in roughly equal proportions in the population so the peaks are of the same general height. However, when a subpopulation is a minority, the second peak can be hidden by the overwhelming numbers of the larger subpopulation (Figure 3).

This assumption of homogeneity causes some well-known problems. Recommended doses of medications assume that all adults (and children) have average weight. The obvious result here is that women, with a lower average weight than men actually get a higher dose of medication for their body weight, with a corresponding increase in adverse reactions.

Statistically, this problem of heterogeneity versus homogeneity is accounted for by examining possible factors which contribute to heterogeneity. The population is divided into small enough segments so that each one remaining is reasonably homogeneous. The researcher must explain why the heterogeneity exists, and how it must examined. That is not a statistical issue.

Therefore, when data are collected on one subpopulation, the results cannot be generalized to the entire population, although this is commonly done particularly in medicine. To reduce the time of study, the subjects are restricted to those identified as high-risk. In observational studies, subjects enrolled in a clinical trial tend to be self -selected and a number of possible confounding factors are introduced which may account for the results.

For example, Premarin has become the number one prescribed drug in the United States. Its widespread use is based totally upon observational studies indicating that it might reduce the incidence of heart disease in women. The medical profession has reached a general consensus based upon observation only (Birkhauser, 1997):

In spite of the benefits induced by HRT such as the treatment of menopausal symptoms, the prevention of osteoporosis and cardiovascular diseases and the reduction of the incidence of M. Alzheimer, the percentage of substituted postmenopausal women varies between approximately 25% in the best and 1-2% in the worst situation in Europe.

These studies do not take into consideration the fact that women may have been denied access to hormone replacement therapy because they had heart disease. Populations do differ (Luoto, Mannisto, Vartiainen, 1998):

Hormone replacement therapy users (28%, n = 463) had higher education, were more often from the capital area, had a significantly higher healthy diet factor score, and were leaner than nonusers.

This is enough to show a difference in outcomes. The "gold standard" of a prospective, randomized controlled trial has not yet been performed. Without it, the results must be considered tentative at best.

Nevertheless, standard statistical methods still deal with groups rather than individuals. This is because the standard methods can only handle a limited number of variables. This has a tremendous impact in society as a whole since individuals tend to be identified in terms of their groups. A more recent method, data mining, allows analysis that can be tailored to individuals. The methods have been used in advertising, particularly on the internet. As written in Elder and Pregibon (1996):

The experienced statistician, perhaps the most capable of guiding the development of automated tools for data analysis, may also be the most acutely aware of all the difficulties that can arise when dealing with real data. This hesitation has bred skepticism of what automated procedures can offer and has contributed to the strong focus by the statistical community on model estimation to the neglect of the logical predecessor to this step, namely model identification.

The statistician’s tendency to avoid complete automation out of respect for the challenges of the data, and the historical emphasis on models with interpretable structure, has led that community to focus on problems with a more manageable number of variables (a dozen, say) and cases (several hundred typically) than may be encountered in KDD problems, which can be orders of magnitude larger at the outset. With increasingly huge and amorphous databases, it is clear that methods for automatically hunting down possible patterns worthy of fuller, interactive attention, are required. The existence of such tools can free one up to, for instance, posit a wider range of candidate data features and basis functions (building blocks) than one would wish to deal with, if one were specifying a model structure "by hand."

Data mining has become a distinct discipline with objectives different from the emphasis of each field (Matheus, Piatetsky-Shapiro, and McNeill, 1992; Fayyad, 1996). Its purpose is to find patterns and regularities in the data, to find connections and links which are not immediately obvious. An outline of the process is as follows:

Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end user	What are the goals? What performance criteria are important? How will the final product be used? Is understandability an issue? What is the trade-off between simplicity and accuracy?
Selecting a dataset.	Consider the homogeneity of the data, change over time, sampling strategy, etc.
Data cleaning and preprocessing	Remove noise and outliers, decide on strategies for handling missing data fields, etc.
Data reduction and transformation	Find useful features to represent the data; use dimensionality reduction or transformation methods to reduce the effective number of variables under consideration
Choosing the data mining task	Deciding whether the goal is classification, regression, clustering, summarization, modeling, etc.
Data mining	The actual analysis, usually done with software.
Evaluating output	An essential step and often overlooked. Determining what is actually knowledge and what is "fool’s gold." Knowledge must be filtered from other outputs.
Use the discovered knowledge	Checking and resolving potential conflicts with previously extracted knowledge.

Data mining can be used to find patterns and relationships in a large, complex database. Then, instead of attempting to reduce an individual to a handful of categories, decisions can be based on the entire base of knowledge. However, any conclusion must be validated or the results will be at best tentative, and at worst, misleading.

Statistical analysis is routinely used in a variety of disciplines. It can provide useful information, but only if it is used correctly. It is important for any student or researcher to have sufficient understanding of statistical concepts to know what can and cannot be proven with statistical tests. Data analysis involves not only computation but judgment based upon experience. That judgment requires not only statistical knowledge but knowledge of the discipline under study.

REFERENCES

Bailar III John C, Mosteller Frederick. (1992). Medical uses of statistics, 2nd Ed. Waltham, MA: NEJM Books.

Benuck I. Gidding SS. Donovan M. Year-to-year variability of cholesterol levels in a pediatric practice. Archives of Pediatrics & Adolescent Medicine. 149(3):292-6, 1995.

Birkhäuser MH. (1997).The Problem of Menopause in Europe. European Menopause Journal. 4(2):68-72.

Chae CU. Ridker PM. Manson JE. Postmenopausal hormone replacement therapy and cardiovascular disease. Thrombosis & Haemostasis. 78(1):770-80, 1997.

Davis Chandler. (1994). Where did twentieth-century mathematics go wrong? In The Intersection of History and Mathematics. Saski, Sigura, Dauben eds. Boston: Birkhauser-Verlag.

Elder John F. Pregibon Daryl. A statistical perspective on knowledge discovery in databases in Advances in Knowledge Discovery and Data Mining. Fayyad WM. Piatetsky-Shapiro G. Smyth P. Uthurusamy R. eds. 1996. Cambridge, Mass: MIT Press.

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. (knowledge discovery in databases) Communications of the ACM. v39 n11 p27(8). 1996.

Fischer Claude S. (1996). Inequality by design: cracking the bell curve myth. Princeton, NJ: Princeton University Press.

Fraser Steven. (1995). The bell curve wars: race intelligence, and the future of America. New York: Basic Books.

Geisser S. Introduction to On the Mathematical Foundations of Theoretical Statistics. In Breakthroughs in Statistics, Vol. 1. Kotz S and Johnson NL, eds. Springer-Verlag: Berlin.

Harding, Sandra. Feminism & methodology. Bloomington, IN: Indiana University Press. 1987.

Herrnstein Richard J, Murray, Charles CA (1994). The bell curve: intelligence and class structure in American life. New York: Free Press.

Jaeger Richard M. (1990). Statistics: A Spectator Sport, 2nd Ed. Newbury Park, CA: Sage Publications

Keren Gideon, Lewis Charles. (1993). A handbook for data analysis in the behavioral sciences: statistical issues. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kincheloe Joe L, Steinberg Shirley R, Gresson III Aaron D. (1996). The bell curve examined. New York: St. Martin’s Press.

LaRosa JC. [Is there a relationship between cholesterol reduction, low levels of cholesterol and mortality?]. Revista Espanola de Cardiologia. 48 Suppl. 2:14-7, 1995.

Luoto R. Mannisto S. Vartiainen E. Hormone replacement therapy and body size: how much does lifestyle explain?. American Journal of Obstetrics & Gynecology. 178(1 Pt 1):66-73, 1998.

MacNeill Ian B, Umphrey GJ. (1987). Foundations of statistical inference. Norwell, MA: Kluwer Academic Publishers.

Matheus, CJ. Piatetsky-Shapiro G. McNeill Mark D. Selecting and reporting what is interesting, in Advances in Knowledge Discovery and Data Mining. Fayyad WM. Piatetsky-Shapiro G. Smyth P. Uthurusamy R. eds. 1996. Cambridge, Mass: MIT Press.

Mies, Maria. Women’s research or feminist research? In Mary Margaret Fonow & Judith A. Cook (Eds.)Beyond methodology: Feminist scholarship as lived research: 60-84. Bloomington, IN: Indiana University Press.

Newman TB. Garber AM. Holtzman NA. Hulley SB. Problems with the report of the Expert Panel on blood cholesterol levels in children and adolescents. Archives of Pediatrics & Adolescent Medicine. 149(3):241-7, 1995.

Polce-Lynch, Mary; Myers, Barbara J; Kilmartin, Christopher T; Forssmann-Falck, Renate; Kliewer, Wendy. Gender and age patterns in emotional expression, body image, and self-esteem: a qualitative analysis. 38(11-12): 1025-1048. 1998.

Rryse, Marjorie. Critical interdisciplinarity, women's studies, and cross-cultural insight. NWSA Journal, 10(1):1-22. 1998.

Rubins HB. Cholesterol in patients with coronary heart disease: how low should we go?. Journal of General Internal Medicine. 10(8):464-71, 1995.

Sacks, Sharon Zell, Wolffe, Karen E., Tierney, Deborah. Lifestyles of students with visual impairments: preliminary studies of social networks. 64(4): 463-478.

Tukey, John W. (1962). The future of data analysis. Annals of Mathematical Statistics. 33:1-67.

Wallace Betty, Graves William. (1995) Poisoned Apple: How Our Schools' Reliance on the "Bell-Curve" Creates Frustration, Mediocrity, and Failure.. New York: St. Martin's Press.

Patricia B. Cerrito

Department of Mathematics

University of Louisville
Louisville, Kentucky 40292

502-852-6826

502-852-7132 (fax)

pcerrito@louisville.edu

NOTE All unreferenced quotations are taken from the web site: http://www.cyber-nation.com/victory/quotations/quotelibrary_page1.html

---

THE IMPORTANCE, EMPOWERMENT, AND TRANSFORMATION OF STATISTICAL ANALYSIS

Patricia B. Cerrito

Information is empowering. With electronic transmission of information, it is now possible for the general public to have ready access to knowledge previously limited to those designated as "expert" in the field. However, the privilege of access must be accompanied by the responsibility to use knowledge properly. Thus, empowerment requires the development of basic skills for finding, interpreting, and using information.

One of the essential basic skills for information processing is statistical literacy. Statistics has transformed virtually every discipline in the physical, medical, and social sciences as well as made significant inroads into the humanities (Jaeger, 1990). Statistics were originally developed to legitimize the social sciences as science. It has been so successful as a discipline that most research is not regarded as legitimate without the collection of data.

Statistical analysis is not without detractors. The physical sciences have been more resistant to the invasion of statistical techniques. Mathematicians, in particular, disregard statistics because of a lack of rigor in its theory. Because of the rift between mathematics and statistics, Tukey (1962) wrote a paper concerning the philosophy of statistics. He claimed that statisticians cannot serve two masters and must choose between the needs of data analysis and the needs of mathematical deduction. This was clearly demonstrated in Davis (1994, p132):

It is only to be expected that mathematics, the field where applications can be forgotten, should turn out to be the field where they most often are forgotten. So we are seeing more than a mere rejection of applications. The contemporary "pure" mathematician does not say only, I am too noble to get my hands dirty on mechanical problems like you mere engineers," but something even more hostile in its defensiveness: something like, "You are destroying my true science if you entangle me with your reality.

There are those in women's studies who are also resistant to statistical analysis (Rryse, 1998; Harding, 1987; Mies, 1991). There is much emphasis upon qualitative analysis. As practiced, however, qualitative analysis is extremely intertwined with quantitative analysis. For example, Sacks, Wolffe, and Tierney, 1998) used a qualitative analysis to validate the results of their quantitative analysis, and still used quantitative measures such as Likert scales in its qualitative analysis. Similarly, the qualitative analysis of Polce-Lynch, Myers, Kilmartin, Forssmann-Falck, and Kliewer (1998) also depended heavily upon statistical methods.

Statistics as a discipline has been transformed by the needs for development in other areas. Certain techniques were needed in psychology (reliability, validity; Keren and Lewis, 1993) while others were needed in medicine (survival analysis, logistic regression; Bailar III and Mosteller, 1992). Some statistical methods generally cross all disciplines (analysis of variance; Geisser, 1992) while initially developed because of needs in one area (agriculture; MacNeill and Umphrey, 1987).

The computer explosion has required that statistics as a discipline develop techniques which can deal with the enormous amounts of information routinely collected and stored in electronic databases. One of the more exciting developments in recent years is that of data mining. High speed computers now permit the use of very complex algorithms to automate the process of hypothesis generation and pattern recognition. The techniques optimize the amount of information which can be gathered from a dataset, enabling researchers investigate individual rather than group characteristics.

In 1962, John W. Tukey published a paper in the Annals of Mathematical Statistics. Without stating a single theorem, it was the longest paper ever published in that journal. The purpose of the paper was to define the steps of data:

- Recognition of the problem
- Application of multiple analysis techniques
- Comparisons of efficacy
- Optimization

The process of data analysis requires a statistical judgment based upon

- experience of the particular field of subject matter from which the data come.
- a broad experience with how particular techniques of data analysis have worked out in a variety of fields of application.
- abstract results about the properties of particular techniques.

The judgment of the statistician is always subject to challenge and should be defensible. There are many confounding factors in observational studies which can explain the results obtained from the statistics; it is the responsibility of the researchers to put meaning to the statistical outcome. Data must be collected and analyzed to make a convincing argument. Yet a statistical argument is rarely convincing because of a denial of the statistical conclusions. Why are statistics mistrusted? Consider the following quotations:

"Figures won’t lie but liars will figure." (Charles H. Grosvenor)

"There are three kinds of lies: Lies, damned lies, and statistics." (Benjamin Disraeli)

"In God we trust; all others must have data." (W. Edwards Demming)

The first two indicate a real suspicion of statistics; the third indicates the way in which statistical analysis permeates all aspects of society. The main reason for this ambivalence is a lack of true understanding of statistical concepts. This tends to generate a belief that statistics can prove or disprove whatever the analyst wants to find. It also leads to a misuse of statistical parameters. Students often avoid statistics courses because of the mathematical content. "A little knowledge is a dangerous thing" and most people have only a little knowledge of statistics. Since statistical software is readily available, statistical values can easily be computed by anyone with little understanding for what they mean. This can cause severe misunderstandings. It is important that every individual become statistically literate and to understand when statistics are used correctly-or misused badly.

In a recent revision (1997) of the guidelines for promotion and tenure published by the College of Arts and Sciences at the University of Louisville, any faculty member who received below average student teaching evaluations was defined as deficient in teaching. Members of the faculty with quantitative skills pointed out that regardless of actual teaching ability, approximately half of the faculty would defined as regardless of performance and would remain deficient even if scores went up. The criterion was shortly rescinded. However, without statistically literate faculty members, the criterion would have been enforced without regard to the consequences.

Yet, when this same use of the average is translated to the medical field, it receives a general acceptance. The value of 200 is defined in the medical community as the upper level of normal blood cholesterol level. Studies are being conducted to determine if people with below average cholesterol levels should be treated (Rubins, 1995), and to determine just how aggressively children and adolescents should be treated for cholesterol levels (Benuck, Gidding, Donovan, 1995). When this is done, virtually everyone will have a cholesterol "disease". One report did issue a dissenting opinion (Newman, Garber, Holtzman, Hulley, 1995):

The expert Panel’s recommendations do not address important gender differences. Girls have higher average cholesterol levels than boys. They will therefore qualify for more dietary and drug treatment despite their lower age-adjusted risk of heart disease and the lack of association between cholesterol levels and cardiovascular mortality in women.

In other words, most of the cholesterol studies have been conducted on men and the conclusions based on the results involving men. The aggressive treatment of cholesterol has yet to have a demonstrated benefit in women. General consensus is not proof. However, without an ability to access the medical literature, an individual must rely on the recommendations provided by the "experts", i.e. the physician in clinical practice making recommendations for patients. A knowledge and understanding of statistical concepts coupled with an ability to access information which is now readily available empowers the individual to weigh the evidence to make decisions.

The book, The Bell Curve (Herrnstein and Murray, 1994) would not have been so widely discussed if people had a better understanding of the bell curve. Many wrote to challenge (Fischer, 1996; Kincheloe, Steinberg, Gresson III, 1996; Fraser, 1995; Wallace and Graves, 1995). However, the basic assumption of the book, that the general population of the United States can be represented be a bell-shaped curve, is false. It is only valid when examining one, homogeneous population. As soon as multiple heterogeneous populations are considered, the underlying distribution of the population is multi-peaked. Those involved with diversity issues need a solid understanding of this statistical issue.

Consider a population which can be divided into two subpopulations. For example, consider the parameter of height and a population of adult men and adult women. If the two subpopulations are graphed separately, the curves are represented as in Figure 1. When the data are merged and graphed, the result has two peaks (Figure 2). Men and women occur in roughly equal proportions in the population so the peaks are of the same general height. However, when a subpopulation is a minority, the second peak can be hidden by the overwhelming numbers of the larger subpopulation (Figure 3).

This assumption of homogeneity causes some well-known problems. Recommended doses of medications assume that all adults (and children) have average weight. The obvious result here is that women, with a lower average weight than men actually get a higher dose of medication for their body weight, with a corresponding increase in adverse reactions.

Statistically, this problem of heterogeneity versus homogeneity is accounted for by examining possible factors which contribute to heterogeneity. The population is divided into small enough segments so that each one remaining is reasonably homogeneous. The researcher must explain why the heterogeneity exists, and how it must examined. That is not a statistical issue.

Therefore, when data are collected on one subpopulation, the results cannot be generalized to the entire population, although this is commonly done particularly in medicine. To reduce the time of study, the subjects are restricted to those identified as high-risk. In observational studies, subjects enrolled in a clinical trial tend to be self -selected and a number of possible confounding factors are introduced which may account for the results.

For example, Premarin has become the number one prescribed drug in the United States. Its widespread use is based totally upon observational studies indicating that it might reduce the incidence of heart disease in women. The medical profession has reached a general consensus based upon observation only (Birkhauser, 1997):

> In spite of the benefits induced by HRT such as the treatment of menopausal symptoms, the prevention of osteoporosis and cardiovascular diseases and the reduction of the incidence of M. Alzheimer, the percentage of substituted postmenopausal women varies between approximately 25% in the best and 1-2% in the worst situation in Europe.

These studies do not take into consideration the fact that women may have been denied access to hormone replacement therapy because they had heart disease. Populations do differ (Luoto, Mannisto, Vartiainen, 1998):

> Hormone replacement therapy users (28%, n = 463) had higher education, were more often from the capital area, had a significantly higher healthy diet factor score, and were leaner than nonusers.

This is enough to show a difference in outcomes. The "gold standard" of a prospective, randomized controlled trial has not yet been performed. Without it, the results must be considered tentative at best.

Nevertheless, standard statistical methods still deal with groups rather than individuals. This is because the standard methods can only handle a limited number of variables. This has a tremendous impact in society as a whole since individuals tend to be identified in terms of their groups. A more recent method, data mining, allows analysis that can be tailored to individuals. The methods have been used in advertising, particularly on the internet. As written in Elder and Pregibon (1996):

> The experienced statistician, perhaps the most capable of guiding the development of automated tools for data analysis, may also be the most acutely aware of all the difficulties that can arise when dealing with real data. This hesitation has bred skepticism of what automated procedures can offer and has contributed to the strong focus by the statistical community on model estimation to the neglect of the logical predecessor to this step, namely model identification.
>
> The statistician’s tendency to avoid complete automation out of respect for the challenges of the data, and the historical emphasis on models with interpretable structure, has led that community to focus on problems with a more manageable number of variables (a dozen, say) and cases (several hundred typically) than may be encountered in KDD problems, which can be orders of magnitude larger at the outset. With increasingly huge and amorphous databases, it is clear that methods for automatically hunting down possible patterns worthy of fuller, interactive attention, are required. The existence of such tools can free one up to, for instance, posit a wider range of candidate data features and basis functions (building blocks) than one would wish to deal with, if one were specifying a model structure "by hand."

Data mining has become a distinct discipline with objectives different from the emphasis of each field (Matheus, Piatetsky - Shapiro, and McNeill, 1992; Fayyad, 1996). Its purpose is to find patterns and regularities in the data, to find connections and links which are not immediately obvious. An outline of the process is as follows:

Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end user	What are the goals? What performance criteria are important? How will the final product be used? Is understandability an issue? What is the trade-off between simplicity and accuracy?
Selecting a dataset.	Consider the homogeneity of the data, change over time, sampling strategy, etc.
Data cleaning and preprocessing	Remove noise and outliers, decide on strategies for handling missing data fields, etc.
Data reduction and transformation	Find useful features to represent the data; use dimensionality reduction or transformation methods to reduce the effective number of variables under consideration
Choosing the data mining task	Deciding whether the goal is classification, regression, clustering, summarization, modeling, etc.
Data mining	The actual analysis, usually done with software.
Evaluating output	An essential step and often overlooked. Determining what is actually knowledge and what is "fool’s gold." Knowledge must be filtered from other outputs.
Use the discovered knowledge	Checking and resolving potential conflicts with previously extracted knowledge.

Data mining can be used to find patterns and relationships in a large, complex database. Then, instead of attempting to reduce an individual to a handful of categories, decisions can be based on the entire base of knowledge. However, any conclusion must be validated or the results will be at best tentative, and at worst, misleading.

Statistical analysis is routinely used in a variety of disciplines. It can provide useful information, but only if it is used correctly. It is important for any student or researcher to have sufficient understanding of statistical concepts to know what can and cannot be proven with statistical tests. Data analysis involves not only computation but judgment based upon experience. That judgment requires not only statistical knowledge but knowledge of the discipline under study.

REFERENCES

Bailar III John C, Mosteller Frederick. (1992). Medical uses of statistics, 2nd Ed. Waltham, MA: NEJM Books.

Benuck I. Gidding SS. Donovan M. Year-to-year variability of cholesterol levels in a pediatric practice. Archives of Pediatrics & Adolescent Medicine. 149(3):292-6, 1995.

Birkhäuser MH. (1997).The Problem of Menopause in Europe. European Menopause Journal. 4(2):68-72.

Chae CU. Ridker PM. Manson JE. Postmenopausal hormone replacement therapy and cardiovascular disease. Thrombosis & Haemostasis. 78(1):770-80, 1997.

Davis Chandler. (1994). Where did twentieth-century mathematics go wrong? In The Intersection of History and Mathematics. Saski, Sigura, Dauben eds. Boston: Birkhauser-Verlag.

Elder John F. Pregibon Daryl. A statistical perspective on knowledge discovery in databases in Advances in Knowledge Discovery and Data Mining. Fayyad WM. Piatetsky-Shapiro G. Smyth P. Uthurusamy R. eds. 1996. Cambridge, Mass: MIT Press.

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. (knowledge discovery in databases) Communications of the ACM. v39 n11 p27(8). 1996.

Fischer Claude S. (1996). Inequality by design: cracking the bell curve myth. Princeton, NJ: Princeton University Press.

Fraser Steven. (1995). The bell curve wars: race intelligence, and the future of America. New York: Basic Books.

Geisser S. Introduction to On the Mathematical Foundations of Theoretical Statistics. In Breakthroughs in Statistics, Vol. 1. Kotz S and Johnson NL, eds. Springer-Verlag: Berlin.

Harding, Sandra. Feminism & methodology. Bloomington, IN: Indiana University Press. 1987.

Herrnstein Richard J, Murray, Charles CA (1994). The bell curve: intelligence and class structure in American life. New York: Free Press.

Jaeger Richard M. (1990). Statistics: A Spectator Sport, 2nd Ed. Newbury Park, CA: Sage Publications

Keren Gideon, Lewis Charles. (1993). A handbook for data analysis in the behavioral sciences: statistical issues. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kincheloe Joe L, Steinberg Shirley R, Gresson III Aaron D. (1996). The bell curve examined. New York: St. Martin’s Press.

LaRosa JC. [Is there a relationship between cholesterol reduction, low levels of cholesterol and mortality?]. Revista Espanola de Cardiologia. 48 Suppl. 2:14-7, 1995.

Luoto R. Mannisto S. Vartiainen E. Hormone replacement therapy and body size: how much does lifestyle explain?. American Journal of Obstetrics & Gynecology. 178(1 Pt 1):66-73, 1998.

MacNeill Ian B, Umphrey GJ. (1987). Foundations of statistical inference. Norwell, MA: Kluwer Academic Publishers.

Matheus, CJ. Piatetsky-Shapiro G. McNeill Mark D. Selecting and reporting what is interesting, in Advances in Knowledge Discovery and Data Mining. Fayyad WM. Piatetsky-Shapiro G. Smyth P. Uthurusamy R. eds. 1996. Cambridge, Mass: MIT Press.

Mies, Maria. Women’s research or feminist research? In Mary Margaret Fonow & Judith A. Cook (Eds.)Beyond methodology: Feminist scholarship as lived research: 60-84. Bloomington, IN: Indiana University Press.

Newman TB. Garber AM. Holtzman NA. Hulley SB. Problems with the report of the Expert Panel on blood cholesterol levels in children and adolescents. Archives of Pediatrics & Adolescent Medicine. 149(3):241-7, 1995.

Polce-Lynch, Mary; Myers, Barbara J; Kilmartin, Christopher T; Forssmann-Falck, Renate; Kliewer, Wendy. Gender and age patterns in emotional expression, body image, and self-esteem: a qualitative analysis. 38(11-12): 1025-1048. 1998.

Rryse, Marjorie. Critical interdisciplinarity, women’s studies, and cross-cultural insight. NWSA Journal, 10(1):1-22. 1998.

Rubins HB. Cholesterol in patients with coronary heart disease: how low should we go?. Journal of General Internal Medicine. 10(8):464-71, 1995.

Sacks, Sharon Zell, Wolffe, Karen E., Tierney, Deborah. Lifestyles of students with visual impairments: preliminary studies of social networks. 64(4): 463-478.

Tukey, John W. (1962). The future of data analysis. Annals of Mathematical Statistics. 33:1-67.

Wallace Betty, Graves William. (1995) Poisoned Apple: How Our Schools' Reliance on the "Bell-Curve" Creates Frustration, Mediocrity, and Failure.. New York: St. Martin’s Press.

Patricia B. Cerrito

Department of Mathematics

University of Louisville
Louisville, Kentucky 40292

502-852-6826

502-852-7132 (fax)

pcerrito@louisville.edu

NOTE All unreferenced quotations are taken from the web site: http://www.cyber-nation.com/victory/quotations/quotelibrary_page1.html