Measurement, Statistical Consulting, and Computing Issues

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 800717 Charlottesville VA 22908 USA
fharrell@virginia.edu
hesweb1.med.virginia.edu/biostat

May 18, 2001

Abstract

This paper addresses a number of issues that arise in the course of statistical consulting and statistical computing in biopharmaceutical research. For example, there are many basic problems in how lab measurements are quantified and analyzed, many of which involve subconscious decisions by analysts that should be critically examined. Some suggestions and personal observations are made.

1 Measurement Issues in Basic Lab Data

Standards often exist for metrics on which laboratory measurements are encoded, but it is not necessary and often is not preferable to analyze the measurements on this same metric. For example, log and reciprocal transformations are often used during basic analyses. When effects or changes are of interest, the analysis becomes more dependent on the choice of metric. Even if the optimum metric for measurements obtained under a single experimental condition is already known, the optimum measure of change in such measurements remains highly problematic.

*Presented in part at the SmithKline Beecham Biometrics Advisory Board Meeting 30 November – 1 December 2000

Researchers seem especially eager to use ratios as measures of change or effect, but ratios have a number of problems[4,6,13]. Unlike differences or log ratios, ratios are asymmetric, that is, one obtains different results depending on which measurement is placed in the denominator. Then there is the problem of data origin. Should something be subtracted from the denominator? Most researchers act as if zero is the only choice. Also, ratios have strange distributions. Kronmal[8] cited many problems with using ratios in statistical modeling:

1. There will be spurious correlations in using ratio variables even if all component variables of ratios are uncorrelated.

2. Division of only the dependent variable by an independent variable can result in regression coefficient estimates for the other independent variables that lead to inappropriate conclusions.

3. Use of a ratio as an independent variable can result in inadequate adjustment for component variables of the ratio.

4. Results of regression analyses incorporating ratios are not readily comparable across studies with different distributions.

Kronmal found that ratio variables should only be used in a full model containing all the component variables. However his paper assumes throughout that a linear model and not a multiplicative one is the correct model. Kaiser[6] states that whatever effect measure is chosen (ratio, difference, etc.) should be demonstrated to be uncorrelated with the base value

2 Special Measurement Issues in Assay and Microarray Data

Assays frequently result in several observations being below the lower limit of detectability (LLD). When computing mean values and other statistics, researchers often replace such values with a value that is lower than all "real" values. This is better than eliminating samples, but is arbitrary, especially if using parametric statistical methods (rank-based analyses are less affected). For parametric analysis it may be preferable to treat values below the LLD as being left censored.

Microarray data has its own special problems. For example, Wikman *et al.*[15] found that the Affymetrix p53 Genechip's 1464 positions have their own unique noise and threshold characteristics.

Note that the practice of using geometric means to reduce the influence of extremely high values causes as many problems as it solves. Very low values become outliers once logs are taken, and an origin of zero is implicitly assumed by geometric means.

3 Problems with Safety Assessments using Clinical Lab Data

It is common in safety analyses to report the proportion of patients with a lab value $> 3\times$ upper limit of normal. This results in a loss of information from continuous variables. Also, patients who were "almost abnormal" at baseline will have an easy time moving to the abnormal category.

If one wishes to analyze lab values as continuous variables, a different issue must be tackled. Many lab parameters have non-monotonic relationships with patient risk. For example, the "normal range" for a parameter may be in the middle of its distribution. There is a need to treat lab values as continuous variables without allowing abnormally low values to cancel abnormally high values. In some cases it will be powerful and more interpretable to transform measurements to a scale for which "abnormality points" can be added, e.g., predicted log odds of short-term mortality, log hazard rate, or predicted log survival time. To obtain these clinical outcome–based transformations extensive databases are required. An example of risk scoring of physiologic variables is shown in Figure 1.

4 Strategies That are "Correct Enough"

A question that arises frequently in statistical consulting is "I know that this analytic method is imperfect, but is it good enough?" For example, one may suspect that there is heteroscedasticity in a regression analysis in which the response variable Y is transformed in some traditional way. Should weighted least squares be used, or is it not worth the trouble? I prefer to change the question to recognize that there is often no obvious metric to choose for the response variable, unless there is a single predictor variable that uses the same metric as the response. Solving for an "optimal" transformation of Y may be made a formal part of the modeling process.

In many cases there will be a transformation that results in homoscedasticity and in an improved R^2 using unweighted least squares. A good procedure is Tibshirani's[12] AVAS (additive model with variance stabilization) algorithm. This is a nonparametric transform–both–sides procedure whose

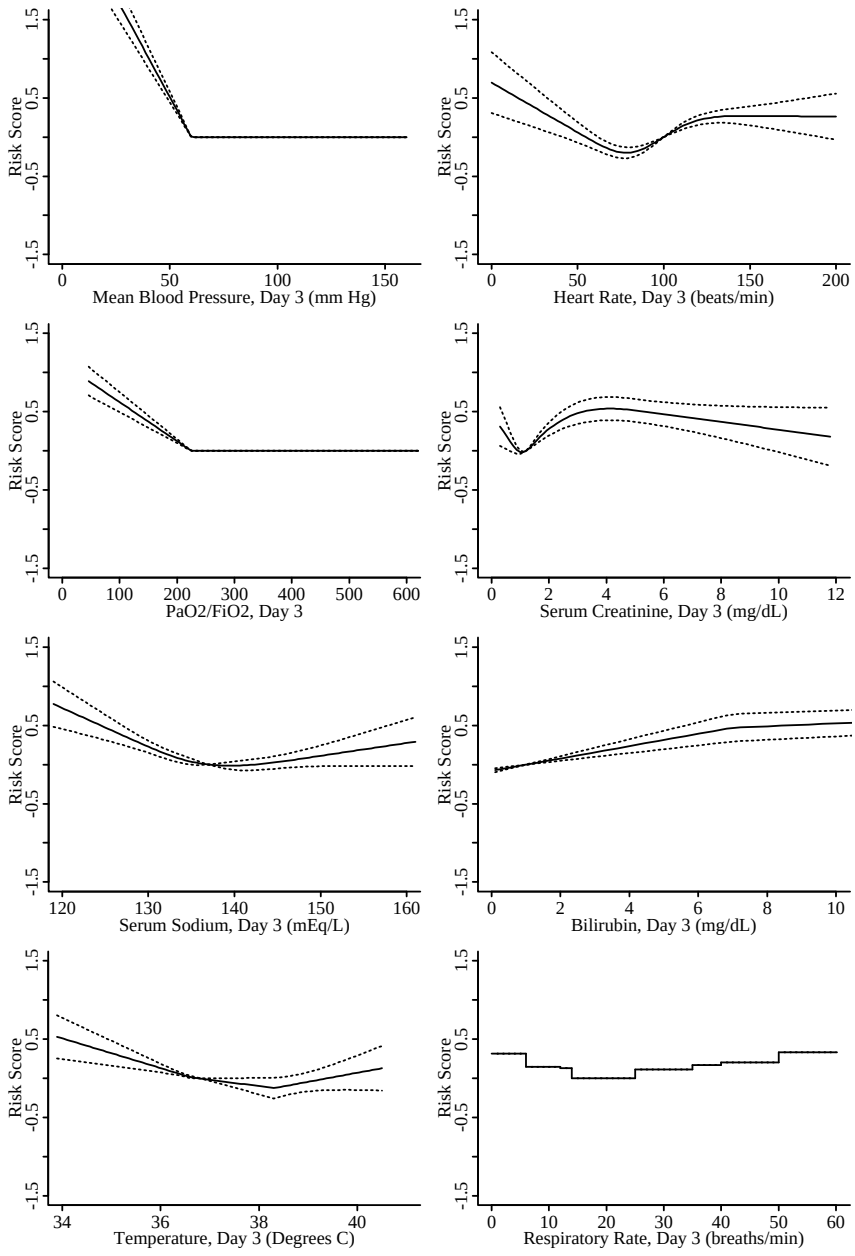


Figure 1: *Risk scoring of physiologic variables in the SUPPORT study*[7]*. Y-axis values represent predicted log hazard of death from a Cox model in which regression effects were modeled flexibly using restricted cubic spline functions.*

goal in transforming Y is to make the variances of residuals independent of $\hat{Y}$ and whose goal in transforming the Xs is to maximize R^2. Bootstrapping can be used to obtain standard errors of estimates and confidence bands, taking into account the uncertainty in all the transformations. This solves a problem that is seldom admitted by statisticians: when multiple Y transforms are tried, there is a large inflation in regression estimates once variances are computed correctly[3] (e.g., by bootstrapping). Most statisticians compute variances that are conditional on the Y-transform. Ordinary variance estimates are biased low because of unaccounted for model uncertainty.

To complete the process of semiparametric estimation using flexible nonparametric transformations, Duan's smearing estimator[2] can be used to obtain predicted means on the original untransformed scale. All aspects of these calculations including bootstrap confidence bands have been automated in the S-PLUS and R function `areg.boot` written by the author (see `http://hesweb1.med.virginia.edu/biostat/presentations/feh/ichpr99/slide.pdf` for details).

When only two variables are being analyzed at a time and only a P-value is needed, the most robust approach is to use rank tests and rank correlation coefficients. When multiple Xs are present, robust rank-based regression (e.g., Cox and proportional odds models) is worth considering. Methods based on ranks are efficient for assessing associations,[1] and they are robust to outliers and strange data distributions. They are somewhat robust to heteroscedasticity.

5 Bayesian Methods in Drug Discovery and Dose Response Assessment

Various multiplicity adjustment techniques have been developed for analysis of microarray and other large-scale screening data, as well as for more standard analyses of multiple treatments and multiple doses. Bayesian prior distributions are usually a better way to deal with multiplicity. One can incorporate prior distributions for the chances that a biomarker is an efficacy marker or for probability of monotonicity of dose-response. Formal Bayesian decision analysis that incorporates costs of false positives and false negatives is also an area worth pursuing. Bayesian methods have small-sample exactness without conditioning on only part of the data. It takes more time to

[1] One of the best kept secrets in statistics seems to be the 0.96 relative efficiency of the Wilcoxon test compared with the t-test when all assumptions of the t-test are satisfied.

be a Bayesian than be a traditional statistician, so once researchers see the advantages of the Bayesian paradigm, a shift to this paradigm will result in more close involvement of statisticians with researchers.

6 Dealing with Discrete Data

Statisticians have a variety of tools for dealing with continuous response data. When truly discrete data arise, such as event times in *current status* data in which assessments of the presence of a condition are made only, say, monthly, it is best to use statistical models that were specially developed to handle discrete responses. For event time data, the Prentice-Gloeckner[11] model will handle heavily tied data. Occasionally it is also a good approach to use a continuous method, such as the ordinary Cox partial likelihood approach, with random breaking of ties in the data.

7 Competition from Software / Statistical Knowledge Dissemination

A big problem facing every data-rich research operation is that there are not enough statisticians to meet current demands for data analysis. As a result, researchers are choosing (poorly) analytic software such as Excel, teaching themselves how to use it, and are analyzing their own data. Researchers are not aware of severe computational errors in software such as Excel (see `http://hesweb1.med.virginia.edu/biostat/teaching/shortcourse/excel.hazards.txt`).

Statistics sections in companies and academia should consider putting out a newletter containing guidelines for choosing statistical software. An ongoing short course series (e.g., Statistical Thinking in Biomedical Research—see `http://hesweb1.med.virginia.edu/biostat/teaching/handouts.html`) emphasizing study design, bias, measurement, precision, power, graphics, and demonstrations ("what a statistician does with data") can also have a great benefit on statistician / biological researcher interaction. Clients should know almost as much about statistics as we know about biology.

8 Web-Based Computing as a Statistician Extender

Pikounis, Gunter, *et al.* of Merck Research Labs[10] have had to face the formidable problem of supporting 3000 scientists with 10 statisticians. They are starting to solve some of the problem by developing web-based statistical applications tailored to the needs of these researchers. Much of their applications involve drug discovery. Pikounis *et al.* propose having web-based statistician extendors practice "safe statistics". They use the S-PLUS StatServer to implement strategies that

- "Produce useful answers 'most' of the time
- Indicate where answers may not be useful
- Have 'adequate' performance
- Handle missing values and other data problems
- Are tuned to user skill level

In practice this means

- Graphics
- Well designed user interface
- Resistant methods
- Fewest assumptions possible (nonparametric procedures)
- Use of subject matter knowledge whenever possible"

A big advantage to this approach as compared with the epidemic of use of Excel and other packages by non-statistician researchers is that statisticians can control which methods are distributed or emphasized to non-statisticians, assuming they spend significant up-front efforts in gathering input from potential users.

The Merck S-PLUS High Throughput Screening (HTS) StatServer is Web based and requires no special client software. It is set up to handle 96–3456 well plates for HTS assays in drug discovery. It takes into account positional effects within plates (esp. edge effects), changing background response and assay sensitivity, trends, cycles, shifts, and missing values. The software allows drilling down after potential problems are seen (e.g., analysis by rows

or by columns). It is heavy on graphics and nonparametric trends. An important feature of the software is that error messages to users are also E-mailed to statisticians. Detailed usage accounting data are stored and analyzed.

9 S-PLUS vs. SAS

After SAS 8 replaces the default of SAS 6 at most companies, and after a few more bugs are solved (e.g., fully implementing downward compatible export of SAS V8 datasets), SAS version 8 will provide a number of advantages to users. Before SAS Version 8 began being distributed in production mode, I estimated that S-PLUS was about eight years ahead of SAS in its analytic capabilities. The gap has now been narrowed to about five years. My opinion is that S-PLUS (and R—see the next section) will continue to provide much greater analytic capabilities to the modern statistician than SAS will provide. SAS excels in providing solid procedures for very frequently used methods but it cannot provide the breadth of statistical computing tools to handle the great variety of problems seen in biomedical research. More importantly, the SAS batch procedure–oriented model is cumbersome and inflexible.

Some of the fundamental advantages of S-PLUS include the following.

1. There is no distinction between `DATA` and `PROC` steps.
2. S-PLUS has no macro language; all commands are "live". For example, the following S language command will compute frequency tables for discrete variables or quantiles for continuous ones without being told anything about the variable by the user:

   ```
   if(is.category(x) | is.character(x) | length(unique(x)) < 20)
     table(x) else quantile(x)
   ```

3. S-PLUS has many more data types than SAS, and users can add their own attributes to data (e.g., flag strange or imputed values).
4. S-PLUS is truly interactive, not batch oriented.
5. S-PLUS has vastly superior graphics.
6. S-PLUS 2000 comes with 2900 functions.

7. The S language is extendible and relatively simple to program; user-written functions are written in the same language used by the developers; statisticians world-wide are writing S functions. SAS modules written by users are typically written in the SAS macro language, a preprocessing language seldom used by developers at SAS Institute.

8. Modern statistical methods may be implemented quite quickly in S (see Statlib (`lib.stat.cmu.edu`) for many examples).

9. S-PLUS has many methods for modeling, exploratory data analysis, missing data, graphics after model fitting, bootstrapping, advanced table making, etc.[14]

10. SAS has an Output Delivery System (ODS) for saving and customized formatting of the results of statistical (and other) procedures. S-PLUS has no need for an ODS because:

    - All entities are objects, allowing all functions to communicate their results directly.
    - It is easy to write special methods for formatting output, e.g.:

```
# create LaTeX table (can also use HTML)
latex(summary(marker ~ age+sex))
# logistic regression model with regression splines, interactions
f ← lrm(y ~ rcs(age,5)*sex +
          rcs(pressure,4))
f               # ordinary printout
plot(f)         # show fitted shapes
Function(f)     # create S+ function to compute y-hat
sascode(Function(f))      # SAS code for y-hat

# typeset fit in algebraic form
w ← latex(f)
html(w)         # convert LaTeX to HTML
# future: convert to XML with embedded MathML:
xml(f)
```

The examples above show how results of analyses can be converted to various representations under complete control of the user. In the future special XML methods for S objects will become important (see `http://www.omegahat.org`).

The following example shows how more powerful user control can result in advanced, clearly formatted tables. A typesetting language such as LaTeX allows fine control of fonts, superscripting, subscripting, etc. The `summary`

method for S formulas, in Harrell's `Hmisc` library, will create three types of tables using appropriate stratifications. The example that follows shows how a "baseline characteristics" table stratified by treatment is created for a clinical trial (Using the Mayo Clinic PBC dataset). The object `s` contains the data summaries. It can be printed using ordinary output, plotted, or typeset using the `latex` function in `Hmisc` as was done below. Note that the statistics being emphasized (medians and percents) are in a larger bold font, and subsidiary information such as outer percentiles, numerators, and denominators appear in smaller nonbold fonts.

```
s ← summary(drug ∼ bili + albumin +
             stage + protime + sex + age +
             spiders,  method='reverse')
latex(s, npct='both')
```

	N	D-penicillamine ($N = 154$)	placebo ($N = 158$)
Serum Bilirubin (mg/dl)	418	0.725 **1.300** 3.600	0.800 **1.400** 3.200
Albumin (gm/dl)	418	3.34 **3.54** 3.78	3.21 **3.56** 3.83
Histologic Stage, Ludwig Criteria : 1	412	**3%** $\frac{4}{154}$	**8%** $\frac{12}{158}$
2		**21%** $\frac{32}{154}$	**22%** $\frac{35}{158}$
3		**42%** $\frac{64}{154}$	**35%** $\frac{56}{158}$
4		**35%** $\frac{54}{154}$	**35%** $\frac{55}{158}$
Prothrombin Time (sec.)	416	10.0 **10.6** 11.4	10.0 **10.6** 11.0
Sex : female	418	**90%** $\frac{139}{154}$	**87%** $\frac{137}{158}$
Age	418	41.4 **48.1** 55.8	43.0 **51.9** 58.9
Spiders	312	**29%** $\frac{45}{154}$	**28%** $\frac{45}{158}$

$_a\ b\ _c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.

S-PLUS, augmented by the `Hmisc` and `Design` libraries, has many easy to use graphics capabilities that are very difficult to implement in SAS. Seven examples follow.

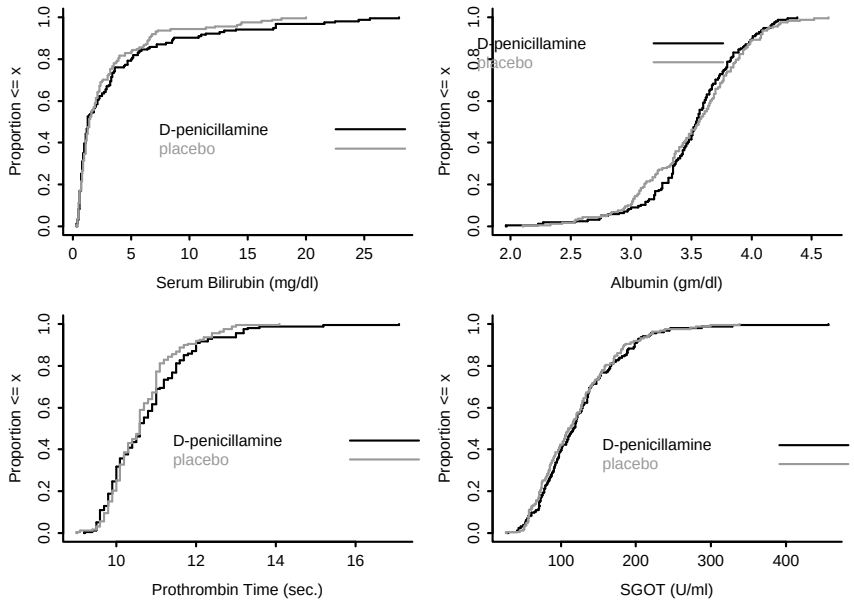


Figure 2: *Empirical cumulative distribution plots for continuous variables stratified by treatment. Produced by the* `ecdf` *function.*

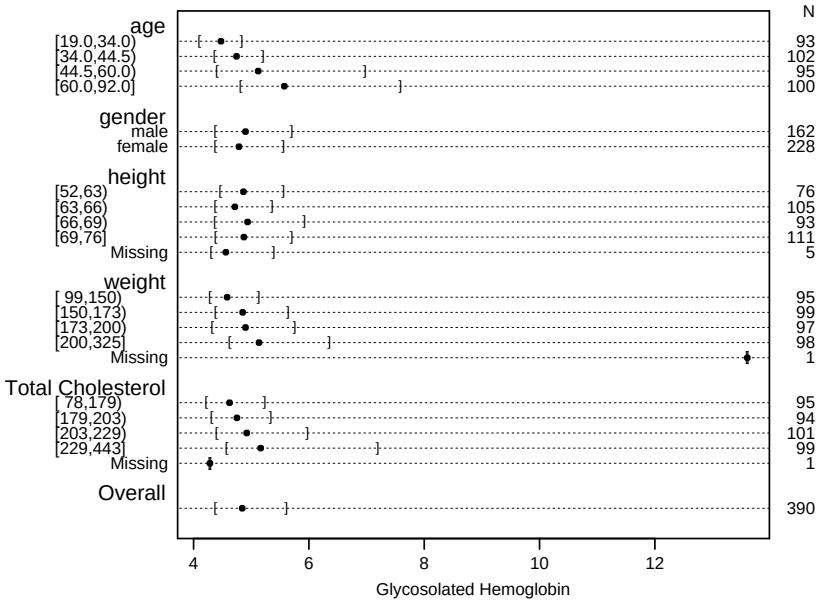


Figure 3: *Three quartiles of glycosolated hemoglobin, stratified separately by a categorical variable (gender) and automatically by quartiles of continuous baseline variables. Produced by the* `summary` *function.*

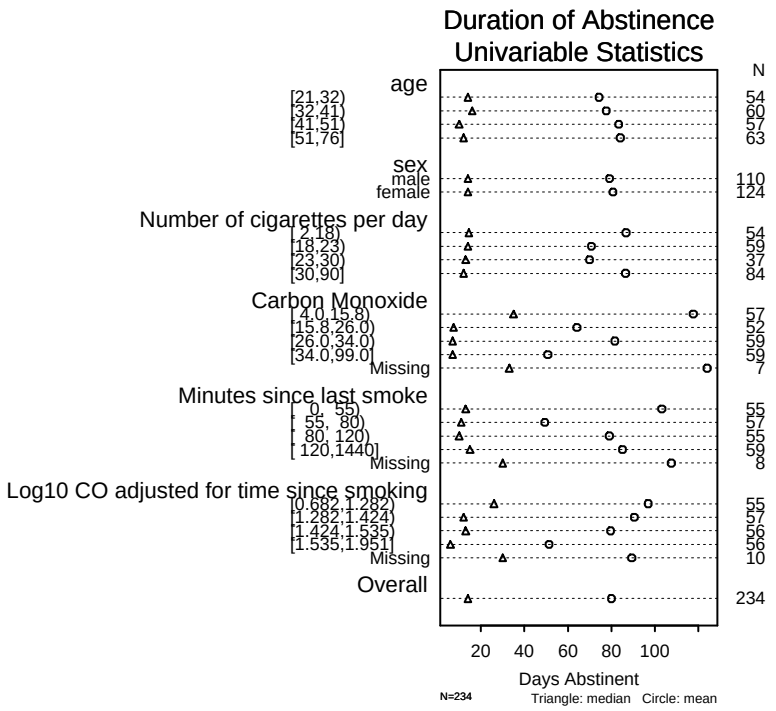


Figure 4: *Mean and median duration of abstinence from cigarette smoking automatically stratified by quartiles of continuous descriptors and by sex. Produced using* `summary`.

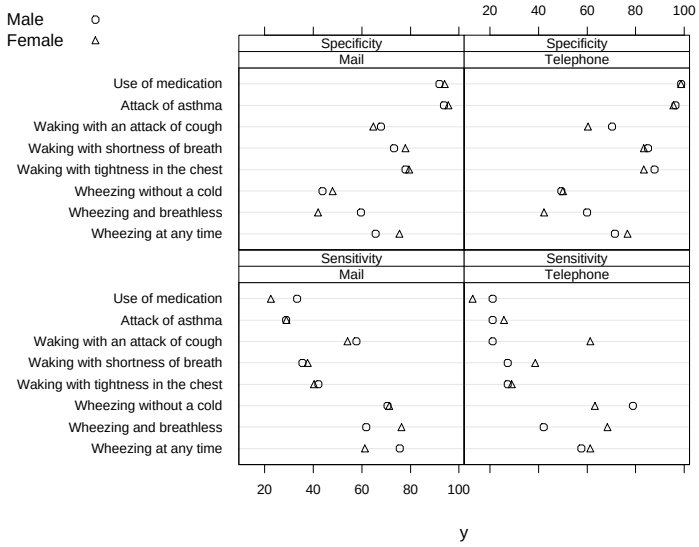


Figure 5: `Trellis` *multipanel display of sensitivity and specificity of various questions, further stratified by sex of respondent. Produced by* `Dotplot`.

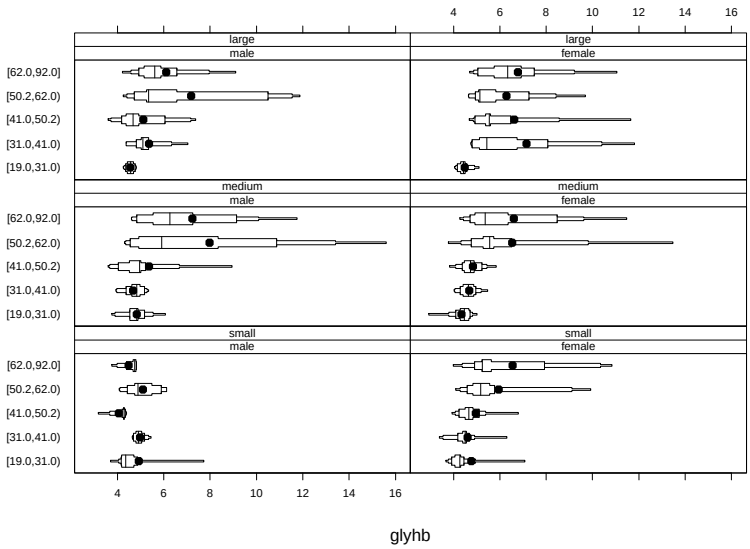


Figure 6: *Extended box plots showing 25%, 50%, 75%, and 90% intervals in addition to the mean (dot) and median (vertical line in the middle). Boxed region depicts the interquartile range. Produced using* `bwplot(..., panel=panel.bpplot)`.

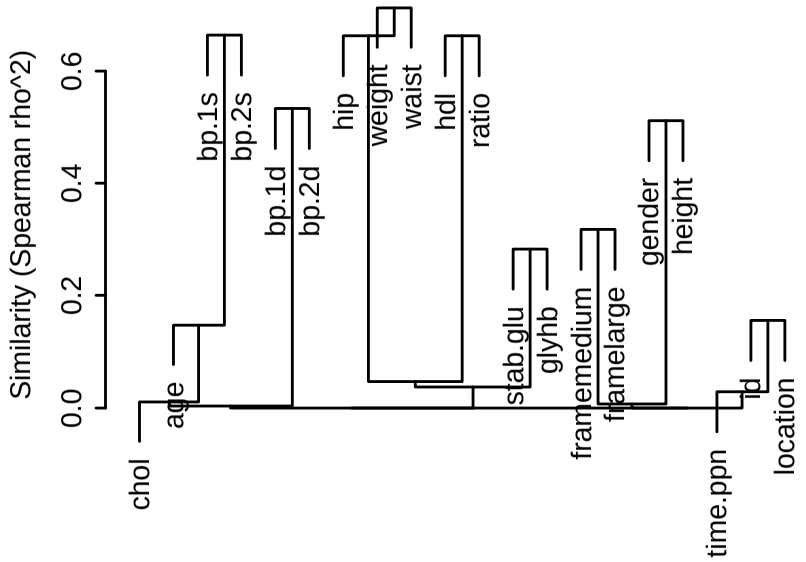


Figure 7: *Hierarchical clustering of variables using Spearman* ρ^2 *as similarity measures. Produced by* `varclus`.

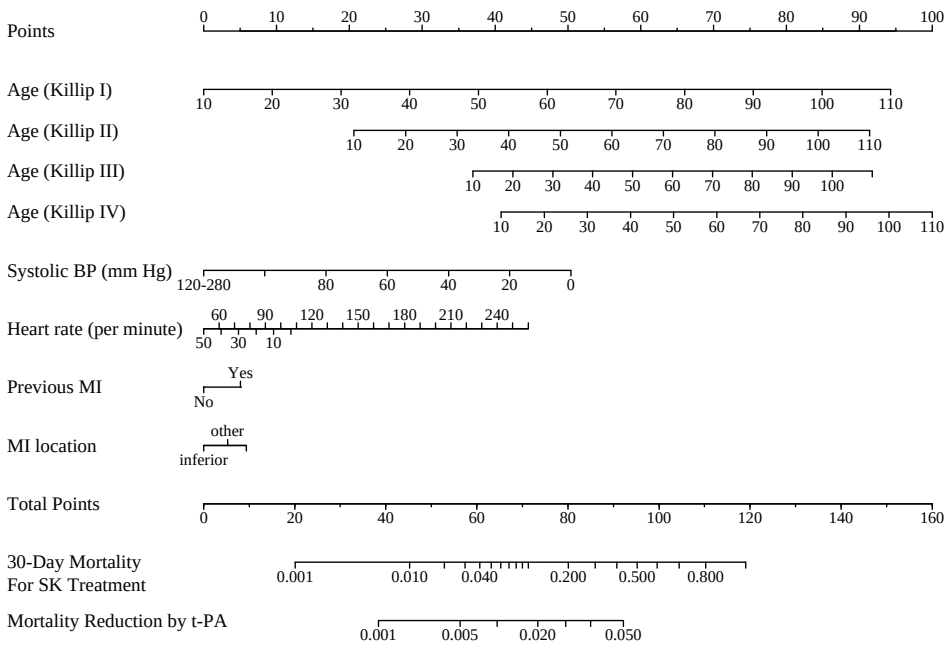
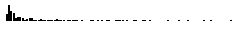


Figure 8: *Nomogram for manually computing the absolute reduction in 30-day mortality by t-PA over streptokinase for patients with acute myocardial infarction in the GUSTO-I trial*[1]*. Produced by the* `Design` *library's* `nomogram` *function.*

pbc
19 Variables 418 Observations

---

bili : Serum Bilirubin (mg/dl)



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
418	0	98	3.221	0.50	0.60	0.80	1.40	3.40	8.03	14.00

lowest : 0.3 0.4 0.5 0.6 0.7, highest: 21.6 22.5 24.5 25.5 28.0

---

albumin : Albumin (gm/dl)

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
418	0	154	3.497	2.750	2.967	3.243	3.530	3.770	4.010	4.141

lowest : 1.96 2.10 2.23 2.27 2.31, highest: 4.30 4.38 4.40 4.52 4.64

---

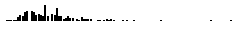
stage : Histologic Stage, Ludwig Criteria

n	missing	unique	Mean
412	6	4	3.024

1 (21, 5%), 2 (92, 22%), 3 (155, 38%), 4 (144, 35%)

---

protime : Prothrombin Time (sec.)



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
416	2	48	10.73	9.60	9.80	10.00	10.60	11.10	12.00	12.45

lowest : 9.0 9.1 9.2 9.3 9.4, highest: 13.8 14.1 15.2 17.1 18.0

---

sex : Sex

n	missing	unique
418	0	2

male (44, 11%), female (374, 89%)

---

fu.days : Time to Death or Liver Transplantation

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
418	0	399	1918	245.1	606.8	1092.8	1730.0	2613.5	3524.2	4040.6

lowest : 41 43 51 71 77, highest: 4500 4509 4523 4556 4795

---

age : Age

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
418	0	345	50.74	33.84	36.37	42.83	51.00	58.24	64.30	67.92

lowest : 26.28 28.88 29.56 30.28 30.57
highest: 74.52 75.00 75.01 76.71 78.44

---

spiders : Spiders

n	missing	unique
312	106	2

absent (222, 71%), present (90, 29%)

Figure 9: *Mixing text and graphics. Descriptive statistics combined with histograms of continuous variables. Produced by* `latex(describe(datasetname))`.

10 Open Source Statistical Computing

A major development in computing since 1991 is the rapid release of operating systems (e.g., Linux) and application software by the Internet-based open source community[9]. Highly skilled programmers are developing all types of freely available software that is also open source, meaning that any user can see all of the source code and can modify it for their own purposes as long as they are willing to give back any improvements to the originator of the program. The Internet has allowed clusters of geographically separated programmers to work cooperatively to develop the highest quality code. These "hackers" are motivated by a desire to produce code that is far superior to that written by Microsoft, by pride in writing the most efficient, reliable, and elegant code, and by a desire for fame instead of money. A major result of this revolution is that developers obtain rapid feedback from growing communities of users, and they rapidly correct errors and make enhancements that are popular with actual users. The overall quality of this chaotic system of software development has in the past two years exceeded the quality of expensive commercial software in many areas.[2] One obvious advantage of the open approach is that users need not wait for yearly update cycles of commercial vendors for errors to be corrected. It is not uncommon for open source software to issue two new versions on the same day.

I removed Microsoft 2000 from my office computer in the fall of 2000 after finding that it was slow and did not run several of my applications correctly. I installed RedHat Linux, and estimate that my personal productivity has increased about 10% even after adjusting for the Linux learning curve. Future productivity will be greater. I have gone from rebooting the computer almost every day to rebooting every two months.

It was only a matter of time before serious open source statistical computing systems became viable options for the statistician. The R system[5] (`www.r-project.org`), an open source version of the S language upon which S-Plus is based, began to mature around 2000 and now has a wide following, particular in other countries where S-Plus is very expensive. R lacks the Microsoft Office interface of S-Plus, the ability to import and export some databases and graphics, and it lacks a graphical user interface and trellis multipanel graphics, but otherwise one can do most everything in R that one can do in S-Plus. R's documentation files are more logically designed than those in S-Plus, and R has builtin functions for installing or

[2]For example, the most widely used Web server is now the open source APACHE server.

updating add-on packages from the Internet. Ironically by the developers of R being much more aggressive in making changes to the system (while being responsive to users), R has become as or more reliable as S-PLUS. R is particularly well suited for Web application development as there are no licensing issues.

The open source nature of R is important in a regulatory environment, as statisticians can examine 100% of its source code when in doubt about a calculation.

References

[1] R. M. Califf, L. H. Woodlief, F. E. Harrell, K. L. Lee, H. D. White, A. Guerci, G. I. Barbash, R. Simes, W. D. Weaver, M. L. Simoons, E. J. Topol, and The GUSTO-I Investigators. Selection of thrombolytic therapy for individual patients: Development of a clinical model. *American Heart Journal*, 133:630–639, 1997.

[2] N. Duan. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78:605–610, 1983.

[3] J. J. Faraway. The cost of data analysis. *Journal of Computational and Graphical Statistics*, 1:213–229, 1992.

[4] R. J. Hayes. Methods for assessing whether change depends on initial value. *Statistics in Medicine*, 7:915–927, 1988.

[5] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

[6] L. Kaiser. Adjusting for baseline: Change or percentage change? *Statistics in Medicine*, 8:1183–1190, 1989.

[7] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, and D. P. Wagner. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122:191–203, 1995.

[8] R. A. Kronmal. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society A*, 156:379–392, 1993.

[9] G. Moody. *Rebel Code: The Inside Story of Linux and the Open Source Revolution*. Perseus Publishing, Cambridge MA, 2001.

[10] B. Pikounis, B. Gunter, A. Liaw, and N. Pajni. Automated analysis software for screening using S-PLUS StatServer. S-PLUS Users Conference, October 2000.

[11] R. L. Prentice and L. A. Gloeckler. Regression analysis of grouped survival data with applications to breast cancer data. *Biometrics*, 34:57–67, 1978.

[12] R. Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83:394–405, 1988.

[13] L. Törnqvist, P. Vartia, and Y. O. Vartia. How should relative changes be measured? *American Statistician*, 39:43–46, 1985.

[14] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, third edition, 1999.

[15] F. P. Wilman, M. Lu, T. Thykjaer, S. H. Olesen, L. D. Andersen, C. Cordon-Cardo, and T. F. Orntoft. Evaluation of the performance of a p53 sequencing microarray chip using 140 previously sequenced bladder tumor samples. *Clinical Chemistry*, 46:1555–1561, 2000.