Regression Modeling with Actuarial and Financial Applications

Edward W. Frees

Contents

20

Report Writing: Communicating Data Analysis Results

Chapter Preview. Statistical reports should be accessible to different types of readers. These reports inform managers who desire broad overviews in nontechnical language as well as analysts who require technical details in order to replicate the study. This chapter summarizes methods of writing and organizing statistical reports. To illustrate, we will consider a report of claims from third party automobile insurance.

20.1 Overview

The last relationship has been explored, the last parameter has been estimated, the last forecast has been made, and now you are ready to share the results of your statistical analysis with the world. The medium of communication can come in many forms: you may simply recommend to a client to "buy low, sell high" or give an oral presentation to your peers. Most likely, however, you will need to summarize your findings in a written report.

Communicating technical information is difficult for a variety of reasons. First, in most data analyses there is no one "right" answer that the author is trying to communicate to the reader. To establish a "right" answer, one only need position the pros and cons of an issue and weigh their relative merits. In statistical reports, the author is trying to communicate data features and the relationship of the data to more general patterns, a much more complex task. Second, most reports written are directed to a primary client, or audience. In contrast, statistical reports are often read by many different readers whose knowledge of statistical concepts varies extensively; it is important to take into consideration the characteristics of this heterogeneous readership when judging the pace and order in which the material is presented. This is particularly difficult when a writer can only guess whom the secondary audience may be. Third, authors of statistical reports need to have a broad and deep knowledge base, including a good understanding of underlying substantive issues, knowledge of statistical concepts and language skills. Drawing on these different skill sets can be challenging. Even for a generally effective writer, any confusion in the analysis is inevitably reflected in the report.

Provide enough details of the study so that the analysis could be independently replicated with access to the original data.

Communication of data analysis results can be a brief oral recommendation to a client or a 500-page Ph.D. dissertation. However, a 10- to 20-page report summarizing the main conclusions and outlining the details of the analysis suffices for most business purposes. One key aspect of such a report is to provide the reader with an understanding of the salient features of the data. Enough details of the study should be provided so that the analysis could be independently replicated with access to the original data.

20.2 Methods for Communicating Data

To allow readers to interpret numerical information effectively, data should be presented using a combination of words, numbers and graphs that reveal its complexity. Thus, the creators of data presentations must draw on background skills from several areas including:

- an understanding of the underlying substantive area,
- a knowledge of the related statistical concepts,
- an appreciation of design attributes of data presentations and
- an understanding of the characteristics of the intended audience.

This balanced background is vital if the purpose of the data presentation is to inform. If the purpose is to enliven the data ("because data are inherently boring") or to attract attention, then the design attributes may take on a more prominent role. Conversely, some creators with strong quantitative skills take great pains to simplify data presentations in order to reach a broad audience. By not using the appropriate design attributes, they reveal only part of the numerical information and hide the true story of their data. To quote Albert Einstein, "You should make your models as simple as possible, but no simpler."

This section presents the basic elements and rules for constructing successful data presentations. To this end, we discuss three modes of presenting numerical information: (i) within text data, (ii) tabular data and (iii) data graphics. These three modes are ordered roughly in the complexity of data that they are designed to present; from the within text data mode that is most useful for portraying the simplest types of data, up to the data graphics mode that is capable of conveying numerical information from extremely large sets of data.

Within Text Data

Within text data simply means numerical quantities that are cited within the usual sentence structure. For example:

The price of Vigoro stock today is $36.50 per share, a record high.

When presenting data within text, you will have to decide whether to use figures or spell out a particular number. There are several guidelines for choosing between figures and words, although generally for business writing you will use words if this choice results in a concise statement. Some of the important guidelines include:

1. Spell out whole numbers from one to ninety-nine.
2. Use figures for fractional numbers.
3. Spell out round numbers that are approximations.
4. Spell out numbers that begin a sentence.
5. Use figures in sentences that contain several numbers.

For example:

There are forty-three students in my class.
With 0.2267 U.S. dollars, I can buy one Swedish kroner.
There are about forty-three thousand students at this university.
Three thousand, four hundred and fifty-six people voted for me.
Those boys are 3, 4, 6 and 7 years old.

Text flows linearly; this makes it difficult for the reader to make comparisons of data within a sentence. When lists of numbers become long or important comparisons are to be made, a useful device for presenting data is the *within text table*, also called the *semitabular* form. For example:

For 2005, net premiums by major line of business written by property and casualty insurers in billions of US dollars, were:

Private passenger auto — 159.57
Homeowners multiple peril — 53.01
Workers' compensation — 39.73
Other lines — 175.09.

(*Source: The Insurance Information Institute Fact Book 2007.*)

Tables

When the list of numbers is longer, the tabular form, or *table*, is the preferred choice for presenting data. The basic elements of a table are identified surrounding Table 20.1.

Title

Table 20.1. *Summary Statistics of Stock Liquidity Variables*

Column Headings

	Mean	Median	Standard deviation	Minimum	Maximum
VOLUME	13.423	11.556	10.632	0.658	64.572
AVGT	5.441	4.284	3.853	0.590	20.772
NTRAN	6436	5071	5310	999	36420
PRICE	38.80	34.37	21.37	9.12	122.37
SHARE	94.7	53.8	115.1	6.7	783.1
VALUE	4.116	2.065	8.157	0.115	75.437
DEB_EQ	2.697	1.105	6.509	0.185	53.628

Body

Stub

Source: Francis Emory Fitch, Inc., Standard & Poor's Compustat, and University of Chicago's Center for Research on Security Prices.

Rule

These are:

1. *Title.* A short description of the data, placed above or to the side of the table. For longer documents, provide a table number for easy reference within the main body of the text. The title may be supplemented by additional remarks, thus forming a *caption*.
2. *Column Headings.* Brief indications of the material in the columns.
3. *Stub.* The left hand vertical column. It often provides identifying information for individual row items.

4. *Body.* The other vertical columns of the table.
5. *Rules.* Lines that separate the table into its various components.
6. *Source.* Provides the origin of the data.

As with the semitabular form, tables can be designed to enhance comparisons between numbers. Unlike the semitabular form, tables are separate from the main body of the text. Because they are separate, tables should be self-contained so that the reader can draw information from the table with little reference to the text. The title should draw attention to the important features of the table. The layout should guide the reader's eye and facilitate comparisons. Table 20.1 illustrates the application of some basic rules for constructing "user friendly" tables. These rules include:

1. For titles and other headers, STRINGS OF CAPITALS ARE DIFFICULT TO READ, keep these to a minimum.
2. Reduce the physical size of a table so that the eye does not have to travel as far as it might otherwise; use single spacing and reduce the type size.
3. Use columns for figures to be compared rather than rows; columns are easier to compare, although this makes documents longer.
4. Use row and column averages and totals to provide focus. This allows readers to make comparisons.
5. When possible, order rows and/or columns by size in order to facilitate comparisons. Generally, ordering by alphabetical listing of categories does little for understanding complex data sets.
6. Use combinations of spacing and horizontal and vertical rules to facilitate comparisons. Horizontal rules are useful for separating major categories; vertical rules should be used sparingly. White space between columns serves to separate categories; closely spaced pairs of columns encourage comparison.
7. Use tinting and different type size and attributes to draw attention to figures. Use of tint is also effective for breaking up the monotonous appearance of a large table.
8. The first time that the data are displayed, provide the source.

Graphs

For portraying large, complex data sets, or data where the actual numerical values are less important than the relations to be established, graphical representations of data are useful. Figure 20.1 describes some of the basic elements of a *graph*, also known as a *chart*, *illustration* or *figure*. These include:

1. *Title* and *Caption.* As with a table, these provide short descriptions of the main features of the figure. Long captions may be used to describe everything that is being graphed, draw attention to the important features and describe the conclusions to be drawn from the data. Include the source of the data here or on a separate line immediately below the graph.
2. *Scale Lines (Axes)* and *Scale Labels.* Choose the scales so that the data fill up as much of the data region as possible. Do not insist that zero be included; assume that the viewer will look at the range of the scales and understand them.
3. *Tick Marks* and *Tick Mark Labels.* Choose the range of the tick marks to include

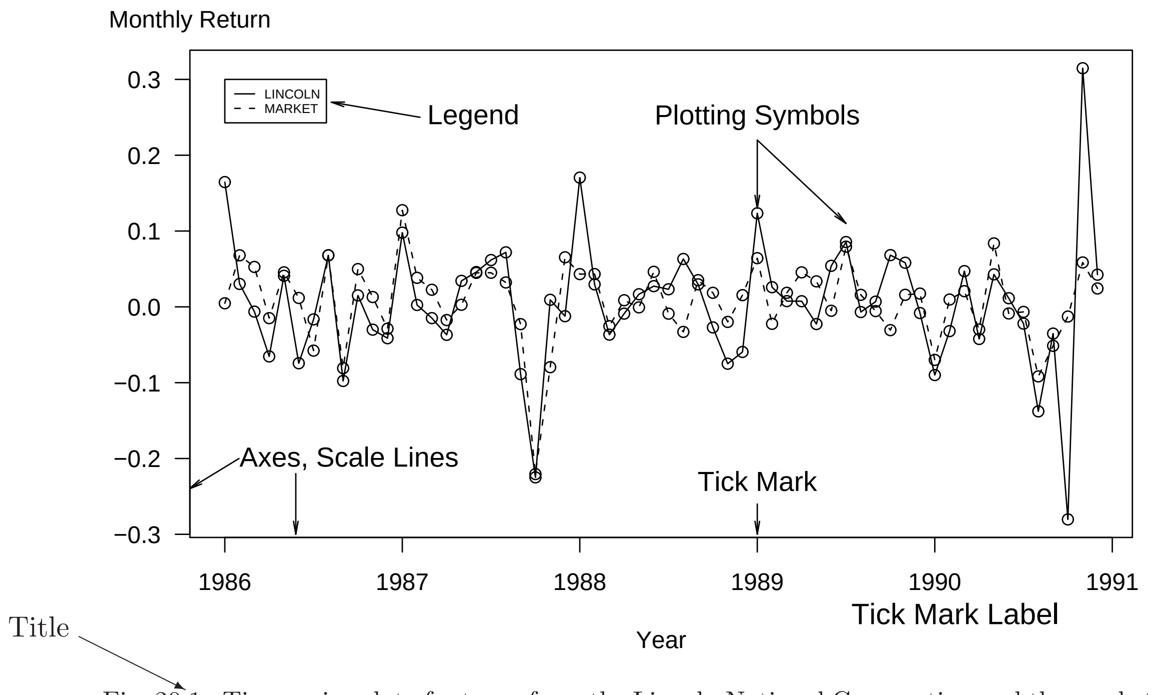


Fig. 20.1. Time series plot of returns from the Lincoln National Corporation and the market. There are 60 monthly returns over the period January, 1986 through December, 1990.

almost all of the data. Three to ten tick marks are generally sufficient. When possible put the tick outside of the data region, so that they do not interfere with the data.

4. *Plotting Symbols.* Use different plotting symbols to encode different levels of a variable. Plotting symbols should be chosen so that they are easy to identify, for example, "O" for one and "T" for two. However, be sure that plotting symbols are easy to distinguish; for example, it can be difficult to distinguish "F" and "E".
5. *Legend (Keys).* These are small textual displays that help to identify certain aspects of the data. Do not let these displays interfere with the data or clutter the graph.

As with tables, graphs are separate from the main body of the text and thus should be self-contained. Especially with long documents, tables and graphs may contain a separate story line, providing a look at the main message of the document in a different way than the main body of the text. Cleveland (1994) and Tufte (1990) provide several tips to make graphs more "user-friendly."

1. Make lines as thin as possible. Thin lines distract the eye less from the data when compared to thicker lines. However, make the lines thick enough so that the image will not degrade under reproduction.
2. Try to use as few lines as possible. Again, several lines distract the eye from the data, which carries the information. Try to avoid "grid" lines, if possible. If you must use grid lines, a light ink, such as a gray or half tone, is the preferred choice.
3. Spell out words and avoid abbreviations. Rarely is the space saved worth the potential confusion that the shortened version may cause the viewer.

4. Use a type that includes both capital and small letters.
5. Place graphs on the same page as the text that discusses the graph.
6. Make words run from left to right, not vertically.
7. Use the substance of the data to suggest the shape and size of the graph. For time series graphs, make the graph twice as wide as tall. For scatter plots, make the graph equally wide as tall. If a graph displays an important message, make the graph large.

Of course, for most graphs it will be impossible to follow all these pieces of advice simultaneously. To illustrate, if we spell out the scale label on a left hand vertical axis and make it run from left to right, then we cut into the vertical scale. This forces us to reduce the size of the graph, perhaps at the expense of reducing the message.

A graph is a powerful tool for summarizing and presenting numerical information. Graphs can be used to break up long documents; they can provoke and maintain reader interest. Further, graphs can reveal aspects of the data that other methods cannot.

20.3 How to Organize

Writing experts agree that results should be reported in an organized fashion with some logical flow, although there is no consensus as to how this goal should be achieved. Every story has a beginning and an end, usually with an interesting path connecting the two endpoints. There are many types of paths, or methods of development, that connect the beginning and the end. For general technical writing, the method of development may be organized chronologically, spatially, by order of importance, general-to-specific or specific-to-general, by cause-and-effect or any other logical development of the issues. This section presents one method of organization for statistical report writing that has achieved desirable results in a number of different circumstances, including the 10- to 20-page report described previously. This format, although not appropriate for all situations, serves as a workable framework on which to base your first statistical report.

The broad outline of the recommended format is:

1. Title and Abstract
2. Introduction
3. Data Characteristics
4. Model Selection and Interpretation
5. Summary and Concluding Remarks
6. References and Appendix

Sections (1) and (2) serve as the preparatory material, designed to orient the reader. Sections (3) and (4) form the main body of the report while Sections (5) and (6) are parts of the ending.

Title and Abstract

If your report is disseminated widely (as you hope), here is some disappointing news. A vast majority of your intended audience gets no further than the title and the abstract. Even for readers who carefully read your report, they will usually carry in their memory the impressions left by the title and abstract unless they are experts in the subject that you are reporting on (which most readers will not be). Choose the title of your report carefully. It should be concise and to the point. Do not include deadwood (phrases like *The Study of, An Analysis of*) but do not be too brief, for example, by using only one word titles. In addition to being concise, the title should be comprehensible, complete and correct.

The language of the abstract should be nontechnical.

The abstract is a one- to two-paragraph summary of your investigation; 75 to 200 words are reasonable rules of thumb. The language should be nontechnical as you are trying to reach as broad an audience as possible. This section should summarize the main findings of your report. Be sure to respond to such questions as: What problem was studied? How was it studied? What were the findings? Because you are summarizing not only your results but also your report, it is generally most efficient to write this section last.

Introduction

As with the general report, the introduction should be partitioned into three sections: orientation material, key aspects of the report and a plan of the paper.

To begin the orientation material, re-introduce the problem at the level of technicality that you wish to use in the report. It may or may not be more technical than the statement of the problem in the abstract. The introduction sets the pace, or the speed at which new ideas are introduced, in the report. Throughout the report, be consistent in the pace. To clearly identify the nature of the problem, in some instances a short literature review is appropriate. The literature review cites other reports that provide insights on related aspects of the same problem. This helps to crystallize the new features of your report.

As part of the key aspects of the report, identify the source and nature of the data used in your study. Make sure that the manner in which your data set can address the stated problem is apparent. Give an indication of the class of modeling techniques that you intend to use. Is the purpose behind this model selection clear (for example, understanding versus forecasting)?

At this point, things can get a bit complex for many readers. It is a good idea to provide an outline of the remainder of the report at the close of the introduction. This provides a map to guide the reader through the complex arguments of the report. Further, many readers will be interested only in specific aspects of the report and, with the outline, will be able to "fast-forward" to the sections that interest them most.

It is also useful to describe the data without reference to a specific model.

Data Characteristics

In a data analysis project, the job is to summarize the data and use this summary information to make inferences about the state of the world. Much of this summarization is done through statistics that are used to estimate model parameters. However, it is also useful to describe the data without reference to a specific model for at least two reasons. First, by using basic summary measures of the data, you can appeal to a larger audience than if you had restricted your considerations to a specific statistical model. Indeed, with a carefully constructed graphical summary device, you should be able to reach virtually any reader who is interested in the subject material. Conversely, familiarity with statistical models requires a certain amount of mathematical sophistication and you may or may not wish to restrict your audience at this stage of the report. Second, constructing statistics that are immediately connected to specific models leaves you open to the criticism that your model selection is incorrect. For most reports, the selection of a model is an unavoidable juncture in the process of inference but you need not do it at this relatively early stage of your report.

In the data characteristics section, identify the nature of data. For example, be sure to identify the component variables, and state whether the data are longitudinal versus cross-sectional, observational versus experimental and so forth. Present any basic summary statistics that would help the reader develop an overall understanding of the data. It is a good idea to include about two graphs. Use scatter plots to emphasize primary relationships in cross-sectional data and time series plots to indicate the most important longitudinal trends. The graphs, and concomitant summary statistics, should not only underscore the most important relationships but may also serve to identify unusual points that are worthy of special consideration. Carefully choose the statistics and graphical summaries that you present in this section. Do not overwhelm the reader at this point with a plethora of numbers. The details presented in this section should foreshadow the development of the model in the subsequent section. Other salient features of the data may appear in the appendix.

Model Selection and Interpretation

This is the heart and soul of your report. The results reported in this section generally took the longest to achieve. However, the length of the section need not be in proportion to the time it took you to accomplish the analysis. Remember, you are trying to spare readers the anguish that you went through in arriving at your conclusions. However, at the same time you want to convince readers of the thoughtfulness of your recommendations. Here is an outline for the Model Selection and Interpretation Section that incorporates the key elements that should appear:

1. An outline of the section
2. A statement of the recommended model
3. An interpretation of the model, parameter estimates and any broad implications of the model
4. The basic justifications of the model
5. An outline of a thought process that would lead up to this model
6. A discussion of alternative models.

In this section, develop your ideas by discussing the general issues first and specific details later. Use subsections (1)-(3) to address the broad, general concerns that a nontechnical manager or client may have. Additional details can be provided in subsections (4)-(6) to address the concerns of the technically inclined reader. In this way, the outline is designed to accommodate the needs of these two types of readers. More details of each subsection are described in the following.

You are again confronted with the conflicting goals of wanting as large an audience as possible and yet needing to address the concerns of technical reviewers. Start this all-important section with an outline of things to come. That will enable the reader to pick and choose. Indeed, many readers will wish only to examine your recommended model and the corresponding interpretations and will assume that your justifications are reliable. So, after providing the outline, immediately provide a *statement of the recommended model* in no uncertain terms. Now, it may not be clear at all from the data set that your recommended model is superior to alternative models and, if that is the case, just say so. However, be sure to state, without ambiguity, what you considered the best. Do not let the confusion that arises from several competing models representing the data equally well drift over into your statement of a model.

The statement of a model is often in statistical terminology, a language used to express model ideas precisely. Immediately follow the statement of the recommended model with the concomitant *interpretations*. The interpretations should be done using nontechnical language. In addition to discussing the overall form of the model, the parameter estimates may provide an indication of the strength of any relationships that you have discovered. Often a model is easily digested by the reader when discussed in terms of the resulting implications of a model, such as a confidence or prediction interval. Although only one aspect of the model, a single implication may be important to many readers.

It is a good idea to discuss briefly some of the technical *justifications of the model* in the main body of the report. This is to convince the reader that you know what you are doing. Thus, to defend your selection of a model, cite some of the basic justifications such as t-statistics, coefficient of determination, residual standard deviation, and so forth in the main body and include more detailed arguments in the appendix. To further convince the reader that you have seriously thought about the problem, include a brief description of a *thought process* that would lead one from the data to your proposed model. Do *not* describe to the reader all the pitfalls that you encountered on the way. Describe instead a clean process that ties the model to the data, with as little fuss as possible.

As mentioned, in data analysis there is rarely if ever a "right" answer. To convince the reader that you have thought about the problem deeply, it is a good idea to mention *alternative models*. This will show that you considered the problem from more than one perspective and are aware that careful, thoughtful individuals may arrive at different conclusions. However, in the end, you still need to give your recommended model and stand by your recommendation. You will sharpen your arguments by discussing a close competitor and comparing it with your recommended model.

Summary and Concluding Remarks

This final section may may serve as a springboard for questions and suggestions about future investigations.

This section should rehash the results of the report in a concise fashion, in different words than the abstract. The language may or may not be more technical than the abstract, depending on the tone that you set in the introduction. Refer to the key questions posed when you began the study and tie these to the results. This section may look back over the analysis and may serve as a springboard for questions and suggestions about future investigations. Include ideas that you have about future investigations, keeping in mind costs and other considerations that may be involved in collecting further information.

References and Appendix

The appendix may contain many auxiliary figures and analyses. The reader will not give the appendix the same level of attention as the main body of the report. However, the appendix is a useful place to include many crucial details for the technically inclined reader and important features that are not critical to the main recommendations of your report. Because the level of technical content here is generally higher than in the main body of the report, it is important that each portion of the appendix be clearly identified, especially with respect to its relation to the main body of the report.

20.4 Further Suggestions for Report Writing

1. Be as brief as you can although still include all important details. On one hand, the key aspects of several regression outputs can often be summarized in one table. Often a number of graphs can be summarized in one sentence. On the other hand, recognize the value of a well-constructed graph or table for conveying important information.
2. Keep your readership in mind when writing your report. Explain what you now understand about the problem, with little emphasis on how you happened to get there. Give practical interpretations of results, in language the client will be comfortable with.
3. Outline, outline. Develop your ideas in a logical, step-by-step fashion. It is *vital* that there be a logical flow to the report. Start with a broad outline that specifies the basic layout of the report. Then make a more detailed outline, listing each issue that you wish to discuss in each section. You only retain literary freedom by imposing structure on your reporting.
4. Simplicity, simplicity, simplicity. Emphasize your primary ideas through simple language. Replace complex words by simpler words if the meaning remains the same. Avoid the use of cliches and trite language. Although technical language may be used, avoid the use of technical jargon or slang. Statistical jargon, such as "Let $x_1, x_2, \ldots$ be i.i.d. random variables ..." is rarely necessary. Limit the use of Latin phrases (e.g., i.e.) if an English phrase will suffice (such as, that is).
5. Include important summary tables and graphs in the body of the report. Label all figures and tables so each is understandable when viewed alone.
6. Use one or more appendices to provide supporting details. Graphs of secondary importance, such as residuals plots, and statistical software output, such as regression fits, can be included in an appendix. Include enough detail so that another analyst, with access to the data, could replicate your work. Provide a strong link between the primary ideas that are described in the main body of the report and the supporting material in the appendix.

20.5 Case Study: Swedish Automobile Claims

® **Empirical**
Filename is "SwedishMotorInsurance"

Determinants of Swedish Automobile Claims

Abstract
Automobile ratemaking depends on an actuary's ability to estimate the probability of a claim and, in the event of a claim, the likely amount. This study examines a classic Swedish data set of third party automobile insurance claims. Poisson and gamma regression models were fit to the frequency and severity portions, respectively. Distance driven by a vehicle, geographic area, recent driver claims experience and the type of automobile are shown to be important determinants of claim frequency. Only geographic area and automobile type turn out to be important determinants of claim severity. Although the experience is dated, the techniques used and the importance of these determinants give helpful insights into current experience.

What problem was studied? How was it studied? What were the findings?

Section 1. Introduction
Actuaries seek to establish premiums that are fair to consumers in the sense that each policyholder pays according to his or her own expected claims. These expected claims are based on policyholder characteristics that may include age, gender and driving experience. Motivation for this rating principle is not entirely altruistic; an actuary understands that rate mispricing can lead to serious adverse financial consequences for the insuring company. For example, if rates are too high relative to the marketplace, then the company is unlikely to gain sufficient market share. Conversely, if rates are too low relative to actual experience, then premiums received will be unlikely to cover claims and related expenses.

Begin with some orientation material.

Setting appropriate rates is important in automobile insurance that indemnifies policyholders and other parties in the event of an automobile accident. For a short term coverage like automobile insurance, claims resulting from policies are quickly realized and the actuary can calibrate the rating formula to actual experience.

For many analysts, data on insurance claims can be difficult to access. Insurers wish to protect the privacy of their customers and so do not wish to share data. For some insurers, data are not stored in an electronic format that is convenient for statistical analyses; it can be expensive to access data even though it is available to the insurer. Perhaps most important, insurers are reluctant to release data to the public because they fear disseminating proprietary information that will help their competitors in keen pricing wars.

When describing the key aspects of the report, include sources of data.

Because of this lack of up to date automobile data, this study examines a classic Swedish data set of third party automobile insurance claims that occurred in 1977. Third party claims involve payments to someone other than the policyholder and the insurance company, typically someone injured as a result of an automobile accident. Although the experience is dated, the regression techniques used in this report work equally well with current experience. Further, the determinants of claims investigated, such as vehicle use and driver experience, are likely to be important into today's driving world.

Provide a plan for the remainder of the paper.

The outline of the remainder of this report is as follows. In Section 2, I present the most important characteristics of the data. To summarize these characteristics, in Section 3 is the discussion of a model to represent the data. Concluding remarks can be found in Section 4 and many of the details of the analysis are in the appendix.

Section 2. Data Characteristics

These data were compiled by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance, summarized in Hallin and Ingenbleek (1983) and Andrews and Herzberg (1985). The data are cross-sectional, describing third party automobile insurance claims for the year 1977.

Identify the nature of the data.

The outcomes of interest are the number of claims (the frequency) and sum of payments (the severity), in Swedish kroners. Outcomes are based on 5 categories of distance driven by a vehicle, broken down by 7 geographic zones, 7 categories of recent driver claims experience (captured by the "bonus") and 9 types of automobile. Even though there are 2,205 potential distance, zone, experience and type combinations ($5 \times 7 \times 7 \times 9 = 2,205$), only $n = 2,182$ were realized in the 1977 data set. For each combination, in addition to outcomes of interest, we have available the number of policyholder years as a measure of exposure. A "policyholder year" is the fraction of the year that the policyholder has a contract with the issuing company. More detailed explanations of these variables are available in Appendix A2.

Use selected plots and statistics to emphasize the primary trends. Do not refer to a statistical model in this section.

In this data, there were 113,171 claims from 2,383,170 policyholder years, for a 4.75% claims rate. From these claims, a total of 560,790,681 kroners were paid, for an average of 4,955 per claim. For reference, in June of 1977, a Swedish kroner could be exchanged for 0.2267 U.S. dollars.

Table 20.2 provides more details on the outcomes of interest. This table is organized by the $n = 2,182$ distance, zone, experience and type combinations. For example, the combination with the largest exposure (127,687.27 policyholder years) comes from those driving a minimal amount in rural areas of southern Sweden, having at least six accident free years and driving a car that is not one of the basic eight types (Kilometres=1, Zone=4, Bonus=7 and Make=9, see Appendix A2). This combination had 2,894 claims with payments of 15,540,162 kroners. Further, I note that there were 385 combinations that had zero claims.

Table 20.2. *Swedish Automobile Summary Statistics*

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
Policyholder Years	1,092.20	81.53	5,661.16	0.01	127,687.27
Claims	51.87	5.00	201.71	0.00	3,338.00
Payments	257,008	27,404	1,017,283	0	18,245,026
Average Claim Number (per Policyholder Year)	0.069	0.051	0.086	0.000	1.667
Average Payment (per Claim)	5,206.05	4,375.00	4,524.56	72.00	31,442.00

Note: Distributions are based on $n = 2,182$ distance, zone, experience and type combinations.
Source: Hallin and Ingenbleek (1983)

Table 20.2 also shows the distribution of the average claim number per insured. Not surprisingly, the largest average claim number occurred in a combination where there was only a single claim with a small number (0.6) of policyholder years. Because we will be using policyholder years as a weight in our Section 3 analysis, this type of aberrant behavior will be automatically down-weighted and so no special techniques are required to deal with it. For the largest average payment, it turns out that there are 27 combinations with a single claim of 31,442 (and one combination with two claims of 31,442). This apparently represents some type of policy limit imposed that we do not have documentation on. I will ignore this feature in the analysis.

Figure 20.2 shows the relationships between the outcomes of interest and exposure bases. For the number of claims, we use policyholder years as the exposure basis. It is clear that the number of insurance claims increases with exposure. Further, the payment amounts increase with the claims number in a very linear fashion.

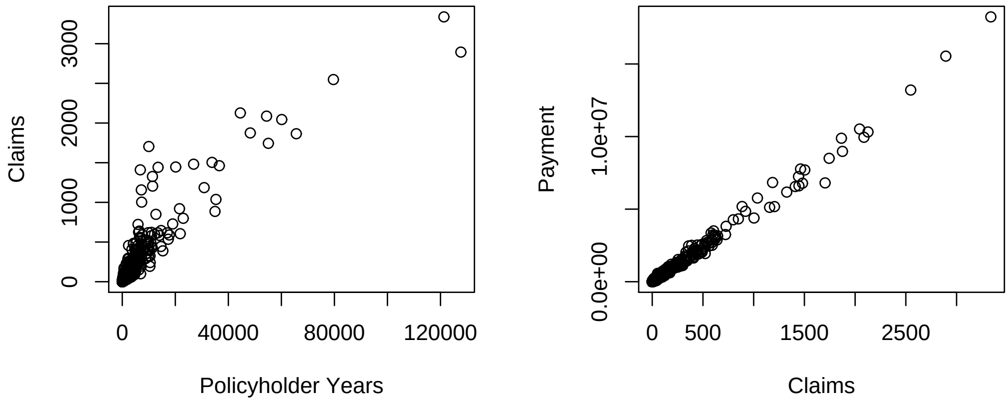


Fig. 20.2. Scatter Plots of Claims versus Policyholder Years and Payments versus Claims.

To understand the explanatory variable effects on frequency, Figure 20.3 presents box plots of the average claim number per insured versus each rating variable. To visualize the relationships, three combinations where the average claim exceeds 1.0 have been omitted. This figure shows lower frequencies associated with lower driving distances, non-urban locations and higher number of accident free years. The automobile type also appears to have a strong impact on claim frequency.

For severity, Figure 20.4 presents box plots of the average payment per claim versus each rating variable. Here, effects of the explanatory variables are not as pronounced as with frequency. The upper right hand panel shows that the average severity is much smaller for Zone=7. This corresponds to Gotland, a county and municipality of Sweden that occupies the largest island in the Baltic Sea. Figure 20.4 also suggests some variation based on the type of automobile.

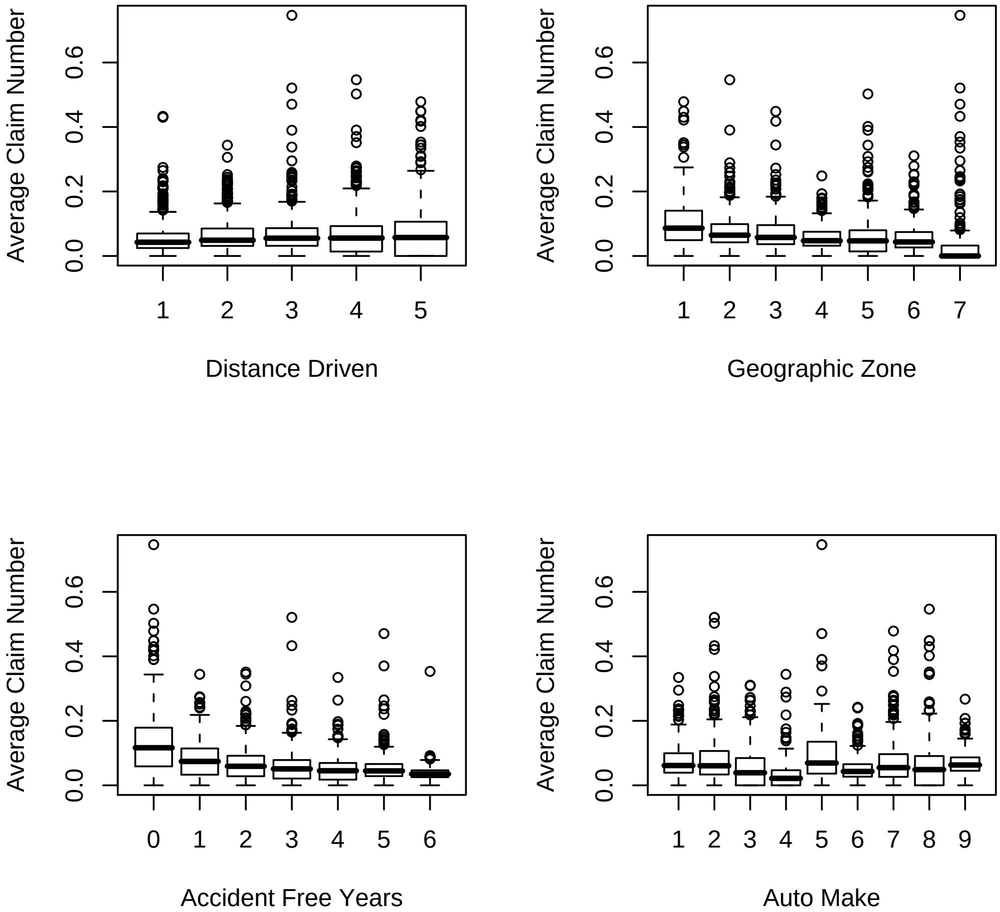


Fig. 20.3. Box Plots of Frequency by Distance Driven, Geographic Zone, Accident Free Years and Make of Automobile.

Section 3. Model Selection and Interpretation

Section 2 established that there are real patterns between claims frequency and severity and the rating variables, despite the great variability in these variables. This section summarizes these patterns using regression modeling. Following the statement of the model and its interpretation, this section describes features of the data that drove the selection of the recommended model.

Start with a statement of your recommended model.

As a result of this study, I recommend a Poisson regression model using a logarithmic link function for the frequency portion. The systematic component includes the rating factors distance, zone, experience and type as additive categorical variables as well as an offset term in logarithmic number of insureds.

Interpret the model; discuss variables, coefficients and broad implications of the model

This model was fit using maximum likelihood, with the coefficients appearing in Table 20.3; more details appear in Appendix A4. Here, the base categories correspond to the first level of each factor. To illustrate, consider a driver living in Stockholm (Zone=1) who drives between one and fifteen thousand kilometers per year (Kilometres=2), has had an accident within the last year (Bonus=1) and

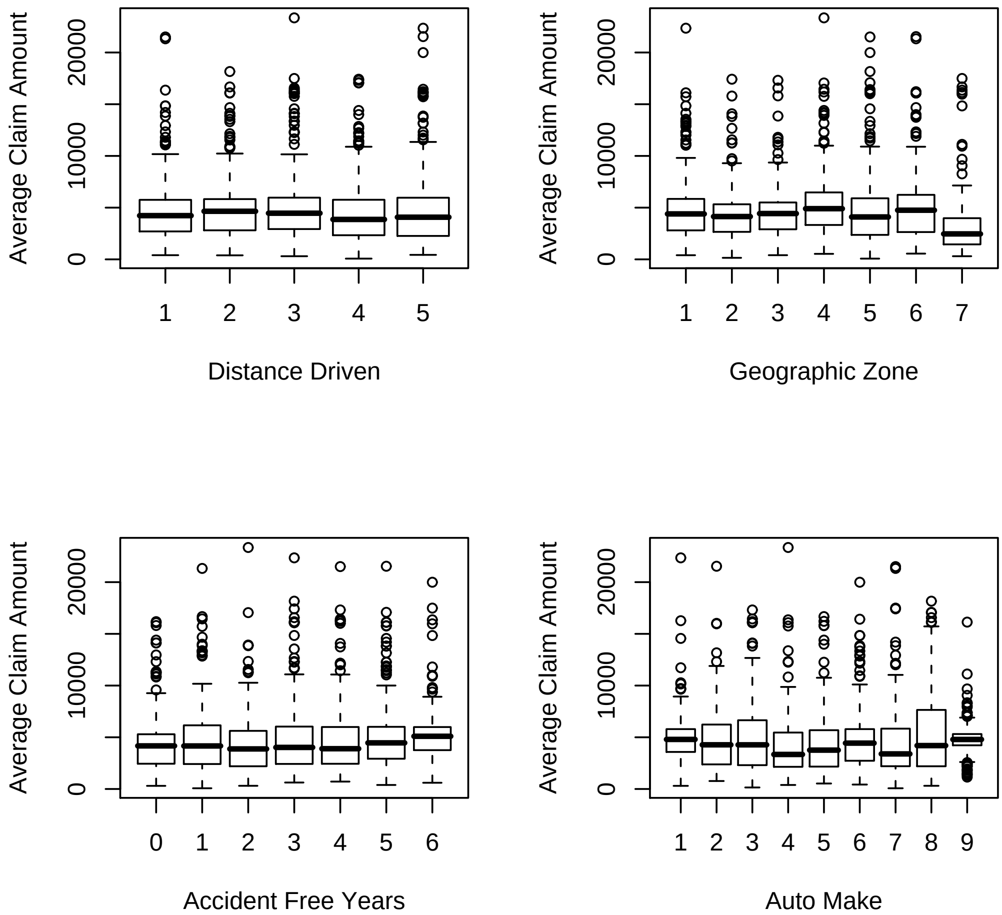


Fig. 20.4. Box Plots of Severity by Distance Driven, Geographic Zone, Accident Free Years and Make of Automobile.

driving car type "Make=6". Then, from Table 20.3, the systematic component is $-1.813 + 0.213 - 0.336 = -1.936$. For a typical policy from this combination, we would estimate a Poisson number of claims with mean $\exp(-1.936) = 0.144$. For example, the probability of no claims within a year is $\exp(-0.144) = 0.866$. In 1977, there were 354.4 policyholder years in this combination, for an expected number of claims of $354.4 \times 0.144 = 51.03$. It turned out that there were only 48 claims in this combination in 1977.

For the severity portion, I recommend a gamma regression model using a logarithmic link function. The systematic component consists of the rating factors zone and type as additive categorical variables as well as an offset term in logarithmic number of claims. Further, the square root of the claims number was used as a weighting variable to give larger weight to those combinations with greater number of claims.

This model was fit using maximum likelihood, with the coefficients appearing in Table 20.4; more details appear in Appendix A6. Consider again our illustrative driver living in Stockholm (Zone=1) who drives between one and fifteen thou-

Table 20.3. *Poisson Regression Model Fit*

Variable	Coefficient	t-ratio	Variable	Coefficient	t-ratio
Intercept	-1.813	-131.78	Bonus=2	-0.479	-39.61
Kilometres=2	0.213	28.25	Bonus=3	-0.693	-51.32
Kilometres=3	0.320	36.97	Bonus=4	-0.827	-56.73
Kilometres=4	0.405	33.57	Bonus=5	-0.926	-66.27
Kilometres=5	0.576	44.89	Bonus=6	-0.993	-85.43
Zone=2	-0.238	-25.08	Bonus=7	-1.327	-152.84
Zone=3	-0.386	-39.96	Make=2	0.076	3.59
Zone=4	-0.582	-67.24	Make=3	-0.247	-9.86
Zone=5	-0.326	-22.45	Make=4	-0.654	-27.02
Zone=6	-0.526	-44.31	Make=5	0.155	7.66
Zone=7	-0.731	-17.96	Make=6	-0.336	-19.31
			Make=7	-0.056	-2.40
			Make=8	-0.044	-1.39
			Make=9	-0.068	-6.84

sand kilometers per year (Kilometres=2), has had an accident within the last year (Bonus=1) and driving car type "Make=6". For this person, the systematic component is $8.388 + 0.108 = 8.496$. Thus, the expected claims under the model are $\exp(8.496) = 4,895$. For comparison, the average 1977 payment was 3,467 for this combination and 4,955 per claim for all combinations.

Table 20.4. *Gamma Regression Model Fit*

Variable	Coefficient	t-ratio	Variable	Coefficient	t-ratio
Intercept	8.388	76.72	Make=2	-0.050	-0.44
Zone=2	-0.061	-0.64	Make=3	0.253	2.22
Zone=3	0.153	1.60	Make=4	0.049	0.43
Zone=4	0.092	0.94	Make=5	0.097	0.85
Zone=5	0.197	2.12	Make=6	0.108	0.92
Zone=6	0.242	2.58	Make=7	-0.020	-0.18
Zone=7	0.106	0.98	Make=8	0.326	2.90
			Make=9	-0.064	-0.42
Dispersion	0.483				

Discussion of the Frequency Model

What are some of the basic justifications of the model?

Both models provided a reasonable fit to the available data. For the frequency portion, the t-ratios in Table 20.3 associated with each coefficient exceed three in absolute value, indicating strong statistical significance. Moreover, Appendix A5 demonstrates that each categorical factor is strongly statistically significant.

Provide strong links between the main body of the report and the appendix.

There were no other major patterns between the residuals from the final fitted model and the explanatory variables. Figure A1 displays a histogram of the deviance residuals, indicating approximate normality, a sign that the data are in congruence with model assumptions.

A number of competing frequency models were considered. Table 20.5 lists two others, a Poisson model without covariates and a negative binomial model with the same covariates as the recommended Poisson model. This table shows that the recommended model is best among these three alternatives, based on the Pearson goodness of fit statistic and a version weighted by exposure. Recall that the Pearson fit statistic is of the form $\sum(O-E)^2/E$, comparing observed (O) to data expected under the model fit (E). The weighted version summarizes $\sum w(O-E)^2/E$, where our weights are policyholder years in units of 100,000. In each case, we prefer models with smaller statistics. Table 20.5 shows that the recommended model is the clear choice among the three competitors.

Table 20.5. *Pearson Goodness of Fit for Three Frequency Models*

Model	Pearson	Weighted Pearson
Poisson without Covariates	44,639	653.49
Final Poisson Model	3,003	6.41
Negative Binomial Model	3,077	9.03

Is there a thought process that leads us to conclude the model is a useful one?

In developing the final model, the first decision made was to use the Poisson distribution for counts. This is in accord with accepted practice and because a histogram of claims numbers (not displayed here) showed a skewed Poisson-like distribution.

Covariates displayed important features that could affect the frequency, as shown in Section 2 and Appendix A3.

A good way to justify your recommended model is to compare it to one or more alternatives.

In addition to the Poisson and negative binomial models, I also fit a quasi-Poisson model with an extra parameter for dispersion. Although this seemed to be useful, ultimately I chose not to recommend this variation because the ratemaking goal is to fit expected values. All rating factors were very statistically significant with and without the extra dispersion factor and so the extra parameter added only complexity to the model. Hence, I elected not to include this term.

Discussion of the Severity Model

For the severity model, the categorical factors zone and make are statistically significant, as shown in Appendix A7. Although not displayed here, residuals from this model were well-behaved. Deviance residuals were approximately normally distributed. Residuals, when rescaled by the square root of the claims number were approximately homoscedastic. There were no apparent relations with explanatory variables.

This complex model was specified after a long examination of the data. Based on the evident relations between payments and number of claims in Figure 20.2, the first step was to examine the distribution of payments per claim. This distribution was skewed and so an attempt to fit logarithmic payments per claim was made. After fitting explanatory variables to this dependent variable, residuals from the model fitting were heteroscedastic. These were weighted by the square root of the claims number and achieved approximate homoscedasticity. Unfortunately, as seen in Appendix Figure A2, the fit is still poor in the lower tails of the distribution.

A similar process was then undertaken using the gamma distribution with a log-link function, with payments as the response and logarithmic claims number as the offset. Again, I established the need for the square root of the claims number as a weighting factor. The process began with all four explanatory variables but distance and accident free years were dropped due to their lack of statistical significance. I also created a binary variable "Safe" to indicate that a driver had six or more accident free years (based on my examination of Figure 20.4). However, this turned out to be not statistically significant and so was not included in the final model specification.

Section 4. Summary and Concluding Remarks

Although insurance claims vary significantly, we have seen that it is possible to establish important determinants of claims number and payments. The recommended regression models conclude that insurance outcomes can be explained in terms of the distance driven by a vehicle, geographic area, recent driver claims experience and type of automobile. Separate models were developed for the frequency and severity of claims. It part, this was motivated by the evidence that fewer variables seem to influence payment amounts compared to claims number.

Rehash the results in a concise fashion. Discuss shortcomings and potential extensions of the work.

This study was based on 113,171 claims from 2,383,170 policyholder years, for a total of 560,790,681 kroners. This is a large data set that allows us to develop complex statistical models. The grouped form of the data allows us to work with only $n = 2,182$ cells, relatively small by today's standards. Ungrouped data would have the advantage of allowing us to consider additional explanatory variables. One might conjecture about any number of additional variables that could be included; age, gender and good student discount are some good candidates. I note that the article by Hallin and Ingenbleek (1983) considered vehicle age - this variable was not included in my database because analysts responsible for the data publication considered to be an insignificant determinant of insurance claims.

Further, my analysis of data is based on 1977 experience of Swedish drivers. The lessons learned from this report may or may not transfer to modern drivers that are closer. Nonetheless, the techniques explored in this report should be immediately applicable with the appropriate set of modern experience.

Appendix

A table of contents, or outline, is useful for long appendices.

Appendix Table of Contents

Include references, detailed data analysis and other materials of lesser importance in the appendices.

A1. References

Andrews, D. F. and A. M. Herzberg (1985). Chapter 68 in: *A Collection from Many Fields for the Student and Research Worker*, pp. 413-421. Springer, New York.

Hallin, Marc and Jean-François Ingenbleek (1983). The Swedish automobile portfolio in 1977: A statistical study. *Scandinavian Actuarial Journal* 1983: 49-64.

A2. Variable Definitions

TABLE A.1 Variable Definitions

Name	Description
Kilometres	Kilometers traveled per year 1: <1,000 2: 1,000-15,000 3: 15,000-20,000 4: 20,000-25,000 5: > 25,000
Zone	Geographic zone 1: Stockholm, Göteborg, Malmö with surroundings 2: Other large cities with surroundings 3: Smaller cities with surroundings in southern Sweden 4: Rural areas in southern Sweden 5: Smaller cities with surroundings in northern Sweden 6: Rural areas in northern Sweden 7: Gotland
Bonus	No claims bonus. Equal to the number of years, plus one, since the last claim.
Make	1-8 represent eight different common car models. All other models are combined in class 9.
Exposure	Amount of policyholder years
Claims	Number of claims
Payment	Total value of payments in Swedish kroner

A3. Basic Summary Statistics for Frequency

TABLE A.2. Averages of Claims per Insured by Rating Factor

```
Kilometre
     1      2      3      4      5
0.0561 0.0651 0.0718 0.0705 0.0827

Zone
     1      2      3      4      5      6      7
0.1036 0.0795 0.0722 0.0575 0.0626 0.0569 0.0504

Bonus
     1      2      3      4      5      6      7
0.1291 0.0792 0.0676 0.0659 0.0550 0.0524 0.0364

Make
     1      2      3      4      5      6      7      8      9
0.0761 0.0802 0.0576 0.0333 0.0919 0.0543 0.0838 0.0729 0.0712
```

A4. Final Fitted Frequency Regression Model — R Output

```
Call: glm(formula = Claims ~ factor(Kilometres) + factor(Zone) +
factor(Bonus) +
    factor(Make), family = poisson(link = log), offset = log(Insured))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.985  -0.863  -0.172   0.600   6.401

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.81284    0.01376 -131.78  < 2e-16 ***
factor(Kilometres)2  0.21259    0.00752   28.25  < 2e-16 ***
factor(Kilometres)3  0.32023    0.00866   36.97  < 2e-16 ***
factor(Kilometres)4  0.40466    0.01205   33.57  < 2e-16 ***
factor(Kilometres)5  0.57595    0.01283   44.89  < 2e-16 ***
factor(Zone)2       -0.23817    0.00950  -25.08  < 2e-16 ***
factor(Zone)3       -0.38639    0.00967  -39.96  < 2e-16 ***
factor(Zone)4       -0.58190    0.00865  -67.24  < 2e-16 ***
factor(Zone)5       -0.32613    0.01453  -22.45  < 2e-16 ***
factor(Zone)6       -0.52623    0.01188  -44.31  < 2e-16 ***
factor(Zone)7       -0.73100    0.04070  -17.96  < 2e-16 ***
factor(Bonus)2      -0.47899    0.01209  -39.61  < 2e-16 ***
factor(Bonus)3      -0.69317    0.01351  -51.32  < 2e-16 ***
factor(Bonus)4      -0.82740    0.01458  -56.73  < 2e-16 ***
factor(Bonus)5      -0.92563    0.01397  -66.27  < 2e-16 ***
factor(Bonus)6      -0.99346    0.01163  -85.43  < 2e-16 ***
factor(Bonus)7      -1.32741    0.00868 -152.84  < 2e-16 ***
factor(Make)2        0.07624    0.02124    3.59  0.00033 ***
factor(Make)3       -0.24741    0.02509   -9.86  < 2e-16 ***
factor(Make)4       -0.65352    0.02419  -27.02  < 2e-16 ***
factor(Make)5        0.15492    0.02023    7.66  1.9e-14 ***
factor(Make)6       -0.33558    0.01738  -19.31  < 2e-16 ***
factor(Make)7       -0.05594    0.02334   -2.40  0.01655 *
factor(Make)8       -0.04393    0.03160   -1.39  0.16449
factor(Make)9       -0.06805    0.00996   -6.84  8.2e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 34070.6  on 2181  degrees of freedom
Residual deviance:  2966.1  on 2157  degrees of freedom AIC: 10654
```

A5. Checking Significance of Factors in the Final Fitted Frequency Regression Model — R Output

```
Analysis of Deviance Table

Terms added sequentially (first to last)
                    Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                 2181      34071
factor(Kilometres)   4     1476      2177      32594  2.0e-318
factor(Zone)         6     6097      2171      26498         0
factor(Bonus)        6    22041      2165       4457         0
factor(Make)         8     1491      2157       2966  1.4e-316
```

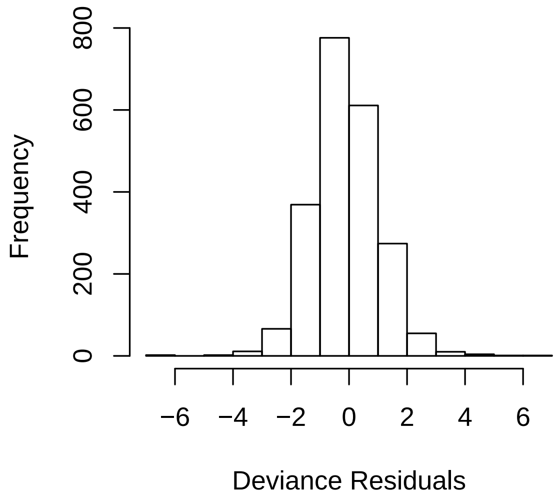


Figure A1. Histogram of deviance residuals from the final frequency model

A6. Final Fitted Severity Regression Model — R Output

```
Call:
glm(formula = Payment ~ factor(Zone) + factor(Make), family = Gamma(link = log),
    weights = Weight, offset = log(Claims))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.56968  -0.39928  -0.06305   0.07179   2.81822

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.38767    0.10933  76.722  < 2e-16 ***
factor(Zone)2 -0.06099    0.09515  -0.641  0.52156
factor(Zone)3  0.15290    0.09573   1.597  0.11041
factor(Zone)4  0.09223    0.09781   0.943  0.34583
factor(Zone)5  0.19729    0.09313   2.119  0.03427 *
factor(Zone)6  0.24205    0.09377   2.581  0.00992 **
factor(Zone)7  0.10566    0.10804   0.978  0.32825
factor(Make)2 -0.04963    0.11306  -0.439  0.66071
factor(Make)3  0.25309    0.11404   2.219  0.02660 *
factor(Make)4  0.04948    0.11634   0.425  0.67067
factor(Make)5  0.09725    0.11419   0.852  0.39454
factor(Make)6  0.10781    0.11658   0.925  0.35517
factor(Make)7 -0.02040    0.11313  -0.180  0.85692
factor(Make)8  0.32623    0.11247   2.900  0.00377 **
factor(Make)9 -0.06377    0.15061  -0.423  0.67205
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 0.4830309)

    Null deviance: 617.32  on 1796  degrees of freedom
Residual deviance: 596.79  on 1782  degrees of freedom
AIC: 16082
```

A7. Checking Significance of Factors in the Final Fitted Severity Regression Model — R Output

```
Analysis of Deviance Table

Terms added sequentially (first to last)

               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            1796     617.32
factor(Zone)    6     8.06      1790     609.26      0.01
factor(Make)    8    12.47      1782     596.79  0.001130
```

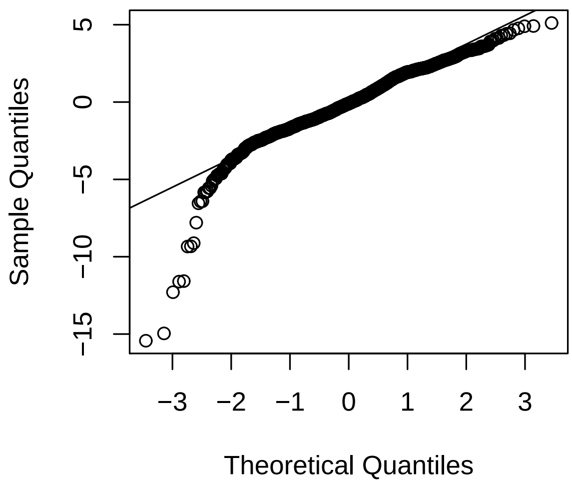


Fig. 20.5. Figure A2. *qq* Plot of Weighted Residuals from a Lognormal Model. The dependent variable is average severity per claim. Weights are the square root of the number of claims. The poor fit in the tails suggests using an alternative to the lognormal model.

20.6 Further Reading and References

You can find further discussion of guidelines for presenting within text data in *The Chicago Manual of Style*, a well-known reference for preparing and editing written copy.

You can find further discussion of guidelines for presenting tabular data in Ehrenberg (1977) and Tufte (1983).

Miller (2005) is a book length introduction to writing statistical reports with an emphasis on regression methods.

Chapter References

The Chicago Manual of Style (1993). The University of Chicago Press, 14th ed. Chicago, Ill.

Cleveland, William S. (1994). *The Elements of Graphing Data.* Monterey, Calif.: Wadsworth.

Ehrenberg, A.S.C. (1977). Rudiments of numeracy. *Journal of the Royal Statistical Society A* 140:27797.

Miller, Jane E. (2005). *The Chicago Guide to Writing about Multivariate Analysis.* The University of Chicago Press, Chicago, Ill.

Tufte, Edward R. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Connecticut.

Tufte, Edward R. (1990). *Envisioning Information.* Graphics Press, Cheshire, Connecticut.

20.7 Exercises

Exercises

20.1 Determinants of CEO Compensation. Chief executive officer (CEO) compensation varies significantly from firm to firm. For this exercise, you will report on a sample of firms from a survey by *Forbes Magazine* to establish important patterns in the

® **Empirical**
Filename is "CeoCompensation"

compensation of CEOs. Specifically, introduce a regression model that explains CEO salaries in terms of the firm's sales and the CEO's length of experience, education level and ownership stake in the firm. Among other things, this model should show that larger firms tend to pay CEOs more and, somewhat surprisingly, that CEOs with a higher educational levels earn less than otherwise comparable CEOs. In addition to establishing important influences on CEO compensation, this model should be used to predict CEO compensation for salary negotiation purposes.

21

Designing Effective Graphs

Chapter Preview.† Actuaries, like other business professionals, communicate quantitative ideas graphically. Because the process of reading, or decoding, graphs is more complex than reading text, graphs are vulnerable to abuse. To underscore this vulnerability, we give several examples of commonly encountered graphs that mislead and hide information. To help creators design more effective graphs and to help viewers recognize misleading graphs, this chapter summarizes guidelines for designing graphs that show important numerical information. When designing graphs, creators should:

(1) Avoid chartjunk
(2) Use small multiples to promote comparisons and assess change
(3) Use complex graphs to portray complex patterns
(4) Relate graph size to information content
(5) Use graphical forms that promote comparisons
(6) Integrate graphs and text
(7) Demonstrate an important message
(8) Know the audience.

Some of these guidelines for designing effective graphs, such as (6), (7) and (8), are drawn directly from principles for effective writing. Others, such as guidelines (3), (4) and (5), come from cognitive psychology, the science of perception. Guidelines (1) and (2) have roots both in effective writing and in graphical perception. For example, the writing principle of brevity demonstrates how eliminating pseudo three-dimensional perspectives and other forms of chartjunk improve graphs. As another example, the writing principle of parallel structure suggests using small multiple variations of a basic graphical form to visualize complex relationships across different groups and over time.

To underscore the scientific aspect of graphical perception, we examine the process of communicating with a graph, beginning with a sender's interpretation of data and ending with a receiver's interpretation of the graph. In keeping with scientific tradition, this chapter discusses several studies in the literature on the effectiveness of graphs.

We conclude that the actuarial profession has many opportunities to improve its practice, making communication more efficient and precise.

21.1 Introduction

Like other business professionals, actuaries communicate ideas orally and in writing, as well as through presentations, which are interactive forms of communication that encompass oral and written messages. Actuaries, as well as other financial analysts, communicate ideas with important quantitative components. Writers express quantitative ideas as (1) numbers within paragraphs, (2) numbers within tabular forms, (3) functional relationships such as equations, and (4) data or equations as graphs.

† This chapter is based on "Designing Effective Graphs," by Edward W. Frees and Robert B. Miller, 1990, *North American Actuarial Journal*, volume 2, number 2, 53-70. Published by the Society of Actuaries - reprinted with permission.

Graphs are a simple yet powerful medium for written communication of quantitative ideas. Graphs can present a large amount of data in a small space, express important relationships between quantities, compare different sets of data, and describe data, thus providing a coherent picture of complex systems. Graphs do more than merely state an idea; they demonstrate it.

Graphs are powerful because they are flexible, but flexibility can be a disadvantage because of the potential for abuse. Well-accepted references dealing with methods of quantitative data presentation mitigate opportunities for abuse. The *Chicago Manual of Style* (1993), a standard reference, discusses presentation of in-text data, and Ehrenberg (1977) and Tufte (1983) discuss presentation of tabular data. In contrast, we focus on data presentation through *graphical* displays.

This chapter seeks to improve actuarial practice as it relates to graphical displays. We intend to: (1) demonstrate the importance of graphical displays, (2) provide guidelines to improve graphical practice, and (3) introduce some of the scientific underpinnings of good graphical practice. The agenda is ambitious, yet the goal of this chapter is to provide practicing actuaries with basic tools that they can use to become critical consumers and effective producers of graphs. We also hope that readers will adopt our enthusiasm and wish to explore the graphical design literature on their own.

An important theme of this chapter is that principles of vigorous writing can and should be applied to the practice of making effective graphs. The *Elements of Style* (Strunk and White 1979, p. xiv) summarizes vigorous writing:

"... sixty-three words that could change the world."

> Vigorous writing is concise. A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, for the same reason that a drawing should have no unnecessary lines and a machine no unnecessary parts. This requires not that the writer make all his sentences short, or that he avoid all detail and treat his subjects only in outline, but that every word tell.

White attributes this quotation to William Strunk. White calls it "a short, valuable essay on the nature and beauty of brevity–sixty-three words that could change the world." We argue that brevity is especially important when making effective graphs. This was also understood by Strunk; as noted above, he said "a drawing should contain no unnecessary lines ... " We use the term *chartjunk*, introduced by Tufte (1983), for any unnecessary appendage in a graph.

Chartjunk is any unnecessary appendage in a graph.

Vigorous writing principles other than brevity also apply to the practice of making effective graphs. Just as with writing, effective graphs are the result of repeated revising and editing. Poorly designed graphs can and do hide information and mislead. Fancy or pretentious graphs are distracting when simpler graphs suffice.

Although the principles of effective writing are valuable, they are not sufficient for producing effective graphs. Writing is processed in a serial manner, word by word, sentence by sentence, with a beginning and an ending. The process of "reading," or *decoding*, a graph is nonlinear and more complex. The additional complexities mean that even authors who follow effective writing practices may produce ineffective graphs. Often the form of written prose is the sole determinant of its value, whereas in graphics the communication process plays the dominant role. We assume that readers are familiar with effective writing forms. Thus, we first review the communication process in which a graph plays a crucial role.

To underscore the importance of effective graphical design, Section 21.2 provides several illustrations of graphs that hide information and are misleading; the defects illustrated are more serious drawbacks than mere chartjunk. The Section 21.2 illustrations motivate the need for additional guidelines and methods for constructing effective graphs.

Section 21.3 introduces eight important guidelines for creating and viewing graphs. Although the guidelines do not provide a panacea for all graphical defects, they do provide business professionals such as actuaries with a key checklist for creating effective graphs. The guidelines are organized so that the first two, on chartjunk and the use of multiples, are based on both effective writing and graphical perception perspectives. Guidelines Three, Four and Five are related primarily to the graphical perception literature, whereas Guidelines Six, Seven and Eight are based primarily on effective writing principles.

As with effective writing, questions of style enter into the discussion of what is and what is not an effective graph. Many style decisions are based upon accepted practices without a firm scientific foundation. However, the process of perceiving graphs has been the subject of inquiry in several scientific disciplines, including psychophysics, cognitive psychology, and computational visions (Cleveland 1995, Chapter 4). Section 21.4 illustrates some types of experimental evidence for determining an effective graphical form based on both the receiver and the graph itself as units of study. Section 21.4 also illustrates how such mainstays of business publications as bar charts and pie charts are poor communicators of numerical information.

Sections 21.5 and S21:References contain concluding remarks and descriptions of some resources for actuaries who wish to learn more about designing effective graphs.

Most readers are removed from the detailed data summarized by a graph. Several difficulties and misconceptions can arise owing to the distance between the original data and a viewer's interpretation of the graph. Figure 21.1 illustrates the challenge of communicating with a graph. The sender (and creator) of the graph has a message derived from an interpretation of data. Although a few graphs communicate raw data, the primary purpose of most graphs is to communicate the sender's interpretation. The message the sender intends is encoded in a graph and passed on to the receiver.

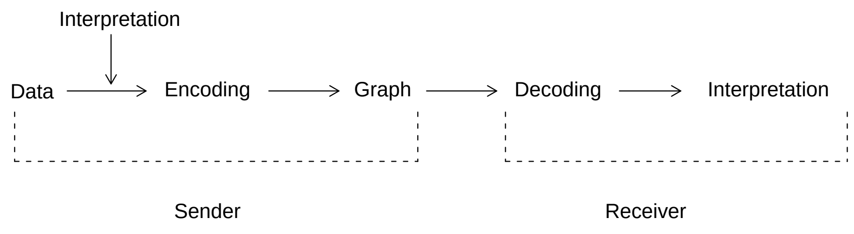


Fig. 21.1. Flow Chart of the Process of Communicating with a Graph. The graph is a crucial intermediary in the process of communicating data interpretation to the receiver.

In general, the receiver is party to neither the exact interpretation intended by the sender nor the raw data. Thus the receiver must decode the graph and develop an interpretation of its message. Two issues arise:

- Whether the interpretation constructed by the receiver is congruent to the interpretation of the sender
- Whether the receiver's interpretation is consistent with and supported by the data.

The first issue depends on the skill with which the sender constructs the graph and the skill with which the receiver decodes it. A poorly constructed graph can hide or distort the sender's message. A graph that is hard to read can discourage the receiver from spending the time necessary to decode the message correctly. The receiver can ignore or misinterpret a graph that is not constructed with care.

The second issue depends not only on the skills mentioned above but also on the skill with which the sender draws meaning from the data. How carefully does the sender document the process of interpretation? Is this communicated to the receiver? Is the receiver capable of assessing the extent to which the graph is a credible summary of the data? Failure at any of these points could result in the receiver ignoring or misinterpreting the graph.

This chapter assumes that the graphs included in business communications are the subject of scrutiny by serious readers. Graphs that appear quickly on the television screen, a flip chart or presentation package are designed to attract attention and to entertain the viewer. Design, rather than information, considerations dominate these media. We focus instead on graphs that are part of professional writing and are designed to inform. As with effective writing, we assume that in creating graphs "...one must believe - in the truth and worth of the scrawl, in the ability of the reader to receive and decode the message" (Strunk and White 1979, p. 84). We now turn to examples of graphs that mislead.

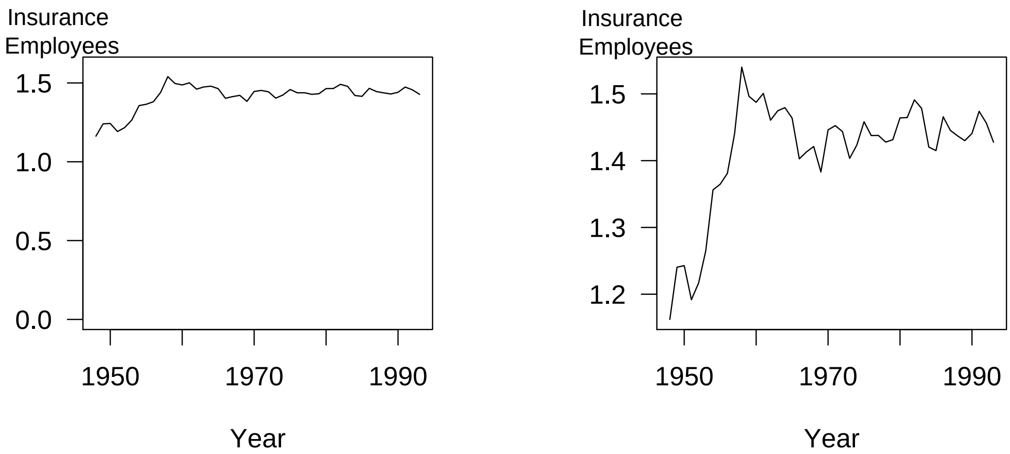
21.2 Graphic Design Choices Make a Difference

As noted by Schmid (1992), the ancient proverb "One picture is worth ten thousand words," when applied to graphs might well read, "One picture *can be* worth ten thousand words *or figures*." Graphic potential is not easily realized. Because of their flexibility, graphs too easily render visual displays of quantitative information that are uninformative, confusing or even misleading.

Examples 21.2.1 through 21.2.5 illustrate five different types of deceptive graphs. In each case, the data were not altered nor were different dimensions of the data portrayed. The common theme of the examples is that, by altering only the data scales, the creator can alter dramatically a viewer's interpretation.

Example 21.2.1: Including Zero To Compress Data. Figure 21.2 shows a time series of the percentage of full-time equivalent workers employed in the insurance industry. The annual data, 1948-1993, are from the National Income and Product Accounts produced by the Bureau of Labor Statistics. The left-hand panel, Figure 21.2(a), provides the impression of a stable employment environment for the insurance industry. Including zero on the vertical axis produces this seeming stability. By doing this, most of the graph is devoted to white space that does not show the variability in the data. In contrast, the right-hand panel, Figure 21.2(b), uses the data to set the range on the axes. This panel clearly shows the large employment increases in the years following the Korean War, circa 1952. It also allows the reader to see the employment declines that the insurance industry has suffered in the most recent three years.

This example is similar to a popular illustration from Huff's well-known *How to Lie with Statistics* (Huff 1954). The point is that motivation external to the data,



(a) **A stable insurance industry**

(b) **The insurance industry work force increased dramatically in the 1950s**

Fig. 21.2. Annual Insurance Employees, 1948-1993. "Insurance employees" is the percentage of full-time-equivalent employees who are working for insurance carriers. Allowing the data to determine the scale ranges reveals interesting aspects of the data.

such as including zero on an axis, can invite us to alter the data scale and change a viewer's interpretation of the data. As Example 21.2.2 shows, creators of graphs can also alter a viewer's interpretation by changing both scales of a two-dimensional graph.

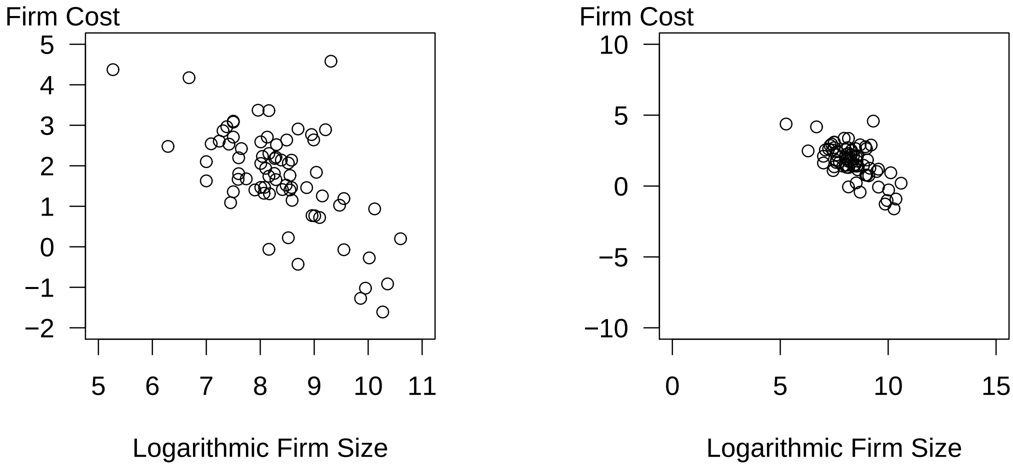
Ⓡ **Empirical** *Filename is "RiskSurvey"*

Example 21.2.2: Perception of Correlation. Figure 21.3 relates risk management cost effectiveness to firm size. These data are from a survey of 73 risk managers of large, U.S.-based, international firms that was originally reported in Schmit and Roth (1990). The data are analyzed in Section 6.5. Here, the measure of risk management cost effectiveness, firm cost, is defined to be the logarithm of the firm's total property and casualty premiums and uninsured losses as a percentage of total assets. The firm size measure is total assets in logarithmic units.

The left-hand panel, Figure 21.3(a), shows a negative relationship between firm costs and firm size, as anticipated by Schmit and Roth. The correlation coefficient between the two variables is -0.64. The data are in a small center portion of Figure 21.3(b) when compared to the left-hand panel, Figure 21.3(a). Figure 21.3(a) uses the data to determine the axes and thus shows more patterns in the data. As Cleveland, Diaconis, and McGill (1982) show, the scaling makes the data in the right-hand panel appear more correlated than in the left-hand panel.

Change of scales can also alter the viewer's perception of trend in time series data, as illustrated in Example 21.2.3.

Example 21.2.3: Transforming to a Logarithmic Scale. Figure 21.4 exhibits a time series of the U.S. credit insurance market over 1950-1989. These data are analyzed in Frees (1996) and are originally from the *Life Insurance Fact Book* (1990). When the amount of insurance is examined on a linear scale in Figure 21.4(a), the credit insurance market appears to be expanding rapidly. However, Fig-

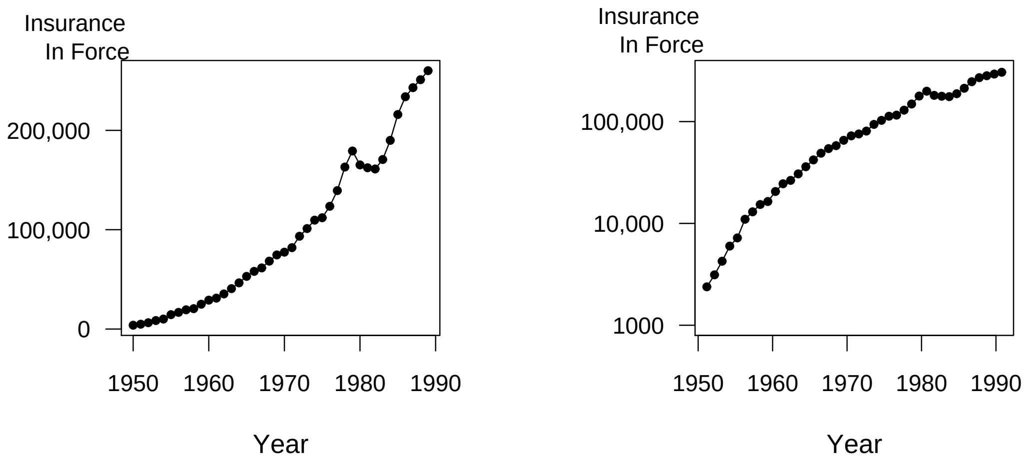


(a) **The data in this figure appear less correlated**

(b) **The data in this figure appear more correlated**

Fig. 21.3. Cost Effectiveness of a Firm's Risk Management Practices Versus Firm Size. The data represented in each figure are the same. However, the wider scales in panel (b) suggest that the data are more highly correlated.

ure 21.4(b) shows that, when examined on a logarithmic scale, the market is leveling off. As discussed in Section 3.2.2, changes on a logarithmic scale can be interpreted as proportional changes. Thus, Figure 21.4(a) shows the market is increasing rapidly, and Figure 21.4(b) shows that the rate of increase is leveling off. These messages are not contradictory, but viewers must interpret each graph critically to understand the intended message.

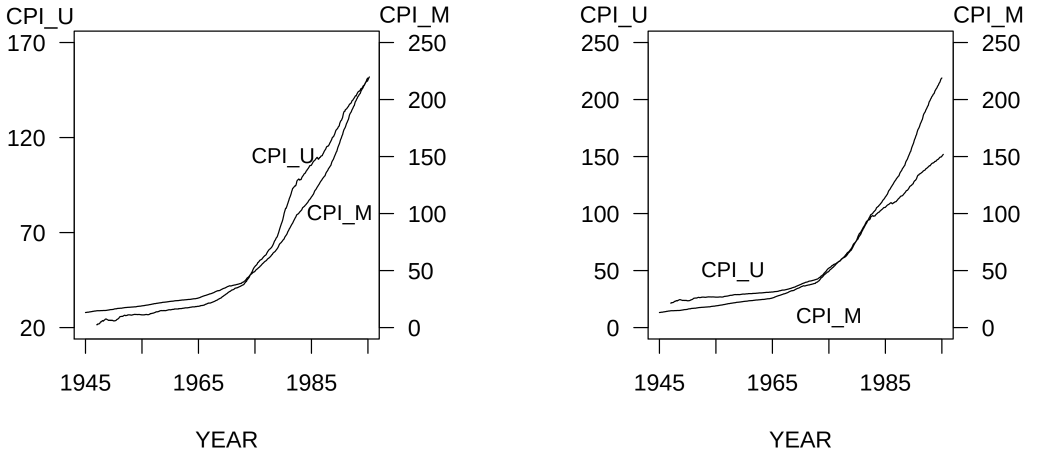


(a) **U.S. credit life insurance market exploding**

(b) **U.S. credit life insurance market leveling off**

Fig. 21.4. Annual U.S. Credit Life Insurance in Force, 1950-1989. Different vertical scales give different impressions of the rate of growth over time.

Example 21.2.4: Double Y-Axes. Figure 21.5 displays two measures of inflation that are produced by the Bureau of Labor Statistics. On the left-hand axes are CPI_U, the consumer price index for urban consumers. On the right-hand axes are CPI_M, the consumer price index for medical components of the overall index. Each series consists of monthly values ranging from January 1947 through April 1995.



(a) **Overall Consumer Price Index (CPI) is similar to the medical component of the CPI**

(b) **Overall Consumer Price Index (CPI) is increasing more slowly than the medical component of the CPI**

Fig. 21.5. Monthly Values of the Overall Consumer Price Index (CPI) and the Medical Component of the CPI, January 1947 through April 1995. Different scale ranges alter the appearances of relative growth of the two series.

The left-hand panel, Figure 21.5(a), suggests that the CPI_U and the CPI_M begin and end in approximately the same position, thus implying that they have increased at about the same rate over the period. The creator could argue that each index measures the value of a standard bundle of goods, thus justifying the argument for using a different scale for each series.

The right-hand panel, Figure 21.5(b), provides a more useful representation of the data by using the same scale for each series. Here, CPI_M begins lower than CPI_U and ends higher. That is, the medical component index has increased more quickly than the index of prices for urban consumers. Other patterns are also evident in Figure 21.5: each series increased at roughly the same rate over 1979-1983 and CPI_M increased much more quickly from 1983 to 1994 when compared to 1948-1979.

Example 21.2.5: Aspect Ratio. Figure 21.6 shows a time series plot of the monthly unemployment rate, April 1953 through December 1992. The unemployment rate is the percentage of unemployed civilian labor force, seasonally adjusted. It is part of the Household Survey produced by the Bureau of Labor Statistics, Department of Labor. This series was analyzed in Frees et al. (1997). The top panel of Figure 21.6 shows that the unemployment rate averaged 5.9% with a peak of 10.8% in the fourth quarter of 1982 and a minimum of 2.7% in the third quarter of 1953.

The two panels in Figure 21.6 differ only in their shape, not in the scaling of either

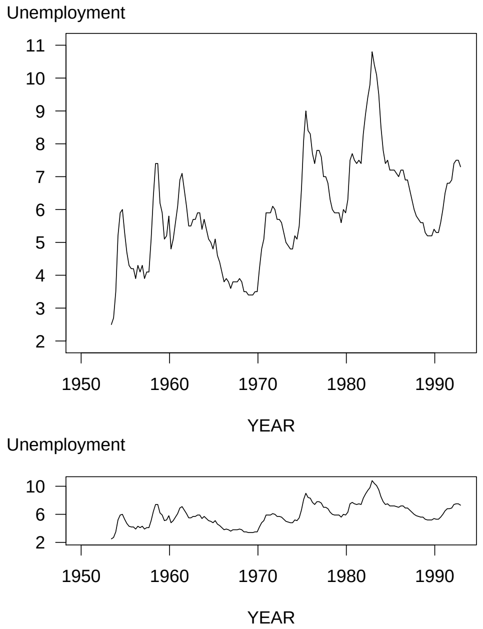


Fig. 21.6. Time Series Plot of Quarterly Values of the U.S. Unemployment Rate, 1953-1992. The lower panel displays a feature that is not evident in the upper panel; unemployment declines more slowly than it rises.

variable or in the relative amount of space that the data take within the figure frame. To differentiate these two shapes, we can use the concept of a figure's *aspect ratio*, defined to be the height of the data frame divided by its width (some sources use the reciprocal of this value for the aspect ratio). The data frame is simply a rectangle whose height and width just allow the graph *of the data* to fit inside. To illustrate, in the upper panel in Figure 21.6, the length of the vertical side is equal to the length of the horizontal side. In the lower panel, the vertical side is only 25% of the horizontal side.

A figure's aspect ratio is defined to be the height of the data frame divided by its width.

Both panels show that the unemployment series oscillated widely over this 39-year period. The lower panel, however, displays a feature that is not apparent in the upper panel; the rise to the peak of an unemployment cycle is steeper than the descent from the peak. Within each unemployment cycle, the percentage of workers unemployed tends to rise quickly to a maximum and then to fall gradually to a minimum. This behavior is surprisingly regular over the almost 39-year period displayed in the plot.

Different aspect ratios can leave substantially different impressions on the eye, as Figure 21.6 illustrates. Thus, the aspect ratio can be chosen to emphasize different features of the data.

21.3 Design Guidelines

Understanding the issues illustrated in Section 21.2 can help actuaries and other business professionals create and interpret graphs. This section presents eight guidelines for designing effective graphs. One of our main points is that current practice is not in accord with these guidelines. Thus, we anticipate that not all of our readers will find the demonstrations of the guidelines visually appealing, but, as stated in Section 21.1, many of the guidelines are based on a scientific foundation outlined in Section 21.4. "Intuition" is something we learn and cultivate; progress in science does not always conform to current intuition. It was widely believed at one time that the earth was flat and that the sun revolved about the earth. The demonstrations of this section may or may not be immediately intuitive, but they are logical conclusions from the design guidelines advocated here.

Guideline One: Avoid Chartjunk

In Section 21.1, we defined chartjunk to be any unnecessary appendage in a graph. Creators of graphs who use chartjunk lower their credibility with serious receivers. Even when senders convey a correct interpretation accompanied by chartjunk, they ask receivers to process and properly ignore the chartjunk. If chartjunk is part of the default, or easily used, options of a software package, then the sender can clutter a graph, or even make a graph misleading, simply by punching a button.

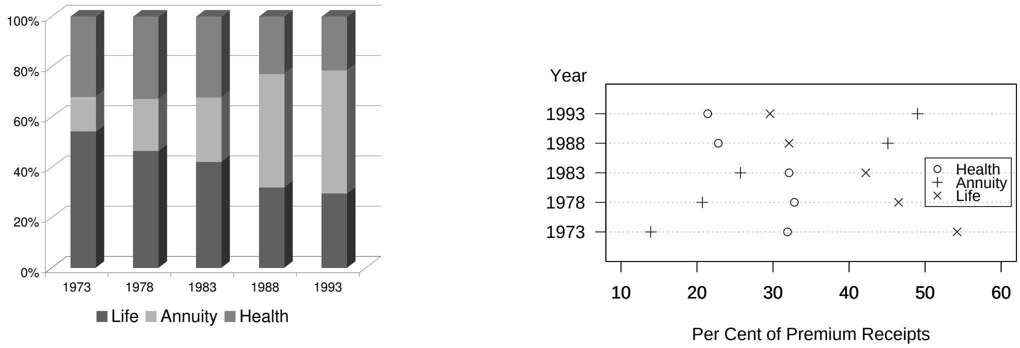
Avoid pictures that are not worth ten thousand words.

Senders who avoid chartjunk raise their credibility. They ask receivers to look only at meaningful characters and marks. Senders may have to spend considerable time with their software to make effective graphs, but the respect and attention of their receivers reward them. Another way to avoid chartjunk is not to use a graph at all if a few words will do. If the message in a graph can be summarized in a few words, then the graph is not needed. Avoid pictures that are not worth ten thousand words!

Avoiding chartjunk is based in part on the concept of brevity in vigorous writing principles. From the graphical perception viewpoint, avoiding chartjunk reduces the noise when communicating between the graph's sender and receiver. Thus, this guideline is important because it has roots in both writing and perception principles.

Example 21.3.1: Premium Receipts of Life Insurance Companies. Figure 21.7(a) is an adaptation of a graph on page 69 of the *Life Insurance Fact Book* (1994). The graph reports 15 bits of information: 5 years and 2 percentages for each year (a third percentage is found by subtraction). A three-dimensional box represents each percentage, and each box displays different shadings to represent the three lines of business: health, annuity and life. These figures could be reported compactly in a small table. However, granting that a graph may help the receiver appreciate trends in the figures, the graph's simplicity should reflect the simplicity of the information available in the figures. In particular, a small plotting symbol suffices to report a percentage. A three-dimensional, shaded box is hardly called for. It is interesting that the three-dimensional box was an "innovation" in 1994. Earlier editions of the *Fact Book* used two-dimensional boxes. The volume of chartjunk took a big jump in 1994.

Figure 21.7(b) is a *dot plot*, discussed by Cleveland (1994). Different plotting symbols show the different lines of business. The tick marks on the lower horizontal axes help us estimate the percentages, and the light, dotted grid lines help us scan across the graph to the plotting symbols of interest. The major shifts, and the approximate magnitudes of the shifts, that happened between 1983 and 1988 are clear here.



(a) **The three-dimensional stacked bar chart is a poor graphical form for making comparisons over time and across lines of business**.

(b) **The dot plot allows for direct comparison over time and across lines of business.**

Fig. 21.7. Distribution of Premium Receipts, 1973-1993. The excessive chartjunk of (a) hides the large change in distribution types between 1983 and 1988.

Guideline Two: Use Small Multiples to Promote Comparisons and Assess Change

Statistical thinking is directed towards comparing measurements of different entities and assessing the change of a measurement over time or some other unit of measurement. Graphical displays are inherently limited when portraying comparisons or assessing changes because they are static, two-dimensional media. Graphs that contain multiple versions of a basic graphical form, each version portraying a variation of the basic theme, promote comparisons and assessments of change. By repeating a basic graphical form, we promote the process of communication.

Tufte (1997) states that using *small multiples* in graphical displays achieves the same desirable effects as using parallel structure in writing. Parallel structure in writing is successful because it allows readers to identify a sentence relationship only once and then focus on the meaning of each individual sentence element, such as a word, phrase or clause. Parallel structure helps achieve economy of expression and draw together related ideas for comparison and contrast. Similarly, small multiples in graphs allow us to visualize complex relationships across different groups and over time.

Small multiples in graphs allow us to visualize complex relationships across different groups and over time.

The Section 21.2 figures illustrated the use of small multiples. In each figure, the two plots portrayed were identical except for the change in scale; this use of parallel structure allowed us to demonstrate the importance of scaling when interpreting graphs. Example 21.3.2 illustrates another application of small multiples in graphical displays, Cleveland's (1993) multiway dot plot.

Example 21.3.2: Relative Importance of Risk Source. Figure 21.8, called a *multiway dot plot*, demonstrates conclusions reached by using a model introduced in Frees (1998) concerning the relative importance of risk sources within a block of short-term insurance contracts. The risk sources are the stochastic interest environment, the frequency of claims (mortality), and the possibility of a catastrophic event (disaster) occurring. The relative importance of these three risk sources is considered by letting two parameters of interest vary. These parameters are the expected year until disaster and, in the event of disaster, the expected proportion (probability) of policyholders that will succumb to disaster.

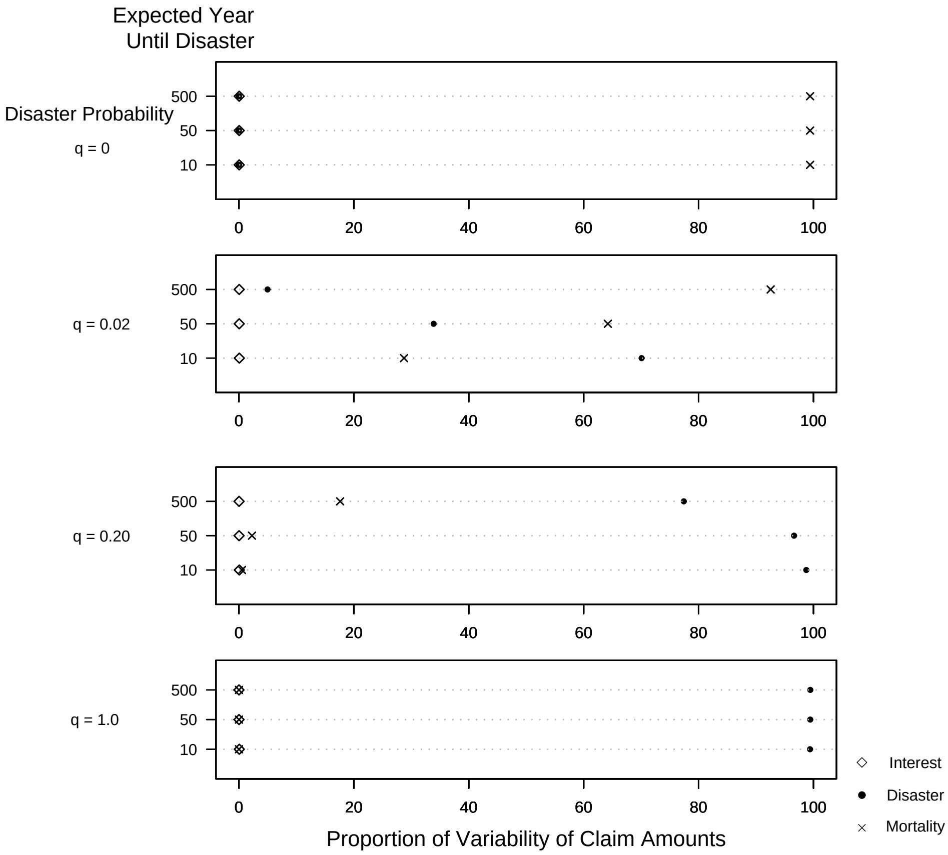


Fig. 21.8. The Relative Importance of Risk Sources. This complex graph allows us to visualize differences over sources of risk (interest, disaster and mortality), expected year until disaster, and probability of disaster. The multiway dot plot demonstrates how quickly the importance of the disaster component increases as the probability of disaster increases.

Figure 21.8 shows that when no policyholders succumb to disaster ($q = 0$), then the frequency component, mortality, dominates the other risk sources. At the opposite extreme, when all policyholders succumb to disaster ($q = 1$), then the disaster component dominates the other risk factors. This is true even when the expected time until disaster is 500 years! For the intermediate cases, when either the expected proportion of policyholders succumbing to disaster increases or the expected year until disaster decreases, the importance of the disaster component increases at the expense of the mortality component. Because of the short-term nature of the contract considered, the interest component does not play an important role in Figure 21.8.

This story of relative importance could not be told using analytic expressions because of the complexity of the underlying models. The story behind Figure 21.8 could be told, however, using tabular displays. The advantage of Figure 21.8 is that it allows the viewer to make comparisons over three different risk sources when two parameters of interest vary. Although such comparisons are possible with tabular displays, graphical displays are more effective devices.

Guideline Three: Use Complex Graphs to Portray Complex Patterns

Many authors believe that a graph should be simple and immediately understood by the viewer. Simple graphs are desirable because they can deliver their message to a broad audience and can be shown quickly and digested immediately. Although this notion may be appropriate for popular writing, for professional writing the concept of instant understanding is limiting in that it precludes the notion that graphs demonstrate complex ideas. Complex patterns should be portrayed as simply as possible, although the patterns themselves should not be unnecessarily simplified.

One way for a graph to represent complex patterns is for some of its basic elements to serve more than one purpose. Tufte (1983) called such elements *multifunctioning.* For example, we can use plotting symbols to represent not only elements corresponding to the horizontal and vertical scales but also a level of a categorical variable.

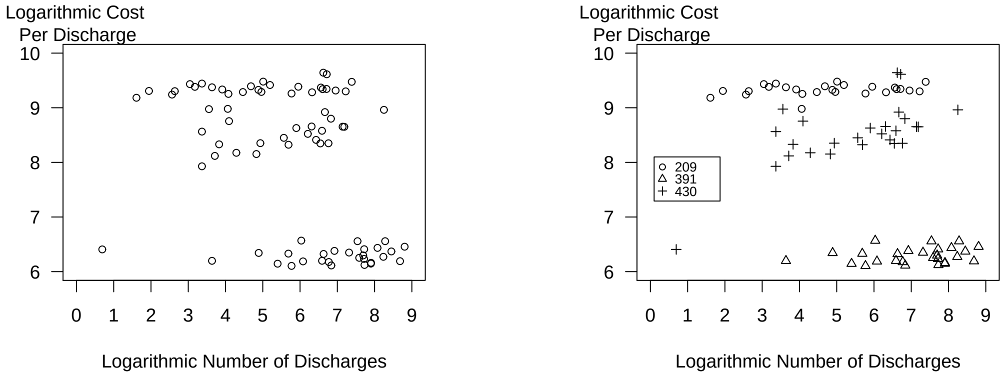
® **Empirical**
Filename is "WiscHospCosts"

Example 21.3.3: Frequency and Severity of Hospital Costs. Figure 21.9 displays the relationship between average hospital costs and frequency of hospital usage. These data for the year 1989 were obtained from the Office of Health Care Information, Wisconsin's Department of Health and Human Services, and are further analyzed in Section 4.4. The data represent averages over the state of Wisconsin, broken down by nine health service areas, three types of providers (fee for service, health maintenance organization, and other) and three types of diagnosis-related groups (DRGs). The three DRGs, numbers 209, 391 and 430, represent major joint and limb reattachment, normal newborns, and psychoses, respectively. Each plotting symbol in Figure 21.9 represents a combination of health service area, type of payer, and type of DRG. The horizontal axis provides the number of patients admitted in 1989 for each combination, in natural logarithmic units. The vertical scale provides the average hospital cost per discharge for each combination, in natural logarithmic units.

The story in the left-hand panel, Figure 21.9(a), is one of increased economies of scale. That is, combinations of health service areas, type of payer, and DRG that have a larger number of patients, measured by discharges, have lower costs. A substantial negative relationship is evident in Figure 21.9(a); the correlation coefficient is -0.43. This is true despite the aberrant point in the lower left-hand region of Figure 21.9(a). The aberrant point is less important economically than the others; it represents a combination with only two discharges. When the point is removed, the correlation becomes -0.50, thus representing an even stronger negative relationship.

Despite its simplicity, Figure 21.9(a) hides an important relationship. The right-hand panel, Figure 21.9(b), is a redrawing of Figure 21.9(a) that includes different plotting symbols for different DRGs. Here, the story is the opposite to the one of increased economies of scale. For combinations representing major joint and limb reattachments and normal newborns, the relationship between frequency and cost is fairly flat. For these DRGs there are few economies of scale. For the psychoses DRG, number 430, Figure 21.9(b) shows a small positive relationship between frequency and cost, even discounting for the combination with only two patients discharged.

The two panels illustrate a phenomenon in statistics referred to as *Simpson's paradox*, or a problem of *aggregation of data.* See Section 4.4 for further discussion. The important point for this chapter is that sometimes simple graphs are mislead-



(a) **With the exception of one outlying observation in the lower left-hand region, there appears to be a significant negative relationship between cost and number of hospital discharges.**

(b) **By introducing the DRG codes, we see a small positive relationship between cost and number of hospital discharges within each group.**

Fig. 21.9. Logarithmic Cost per Discharge Versus the Logarithmic Number of Discharges. By adding a plotting symbol code for the level of DRG, the three distinct groups are evident. The three DRGs, 209, 391, and 430, represent major joint and limb reattachment, normal newborns and psychoses, respectively.

ing. Complex graphs may take more time for viewers to interpret, but they more effectively summarize complex relationships.

Guideline Four: Relate Graph Size to Information Content

"How large should the graph be?" is an important question. The bounds on size are clear. Graphs should not be so small that they are not clearly legible, particularly upon reproduction that degrades an image, nor should they be so large that they exceed a page. With large graphs, it is difficult to compare elements within the graph, thus defeating a primary purpose of graphs.

The data density of a graph is the number of data entries per unit area of the graph.

Within these bounds, a graph should be proportional to the amount of information that it contains. To discuss the proportion of information content, Tufte (1983) introduced the *data density of a graph.* This is defined to be the number of data entries per unit area of the graph. For comparing graph size and information, the data density is a quantity to be maximized, either by increasing the number of data entries or reducing the size of the graph. By examining this density over a number of popular publications, Tufte concluded that most graphs could be effectively shrunk.

For example, Figure 21.7(a) is a chart with a low data density. This chart represents only 15 numbers. With an area of approximately 9 square inches, this graph's data density is roughly 15/9. For comparison, Figure 21.10 shows approximately 600 numbers. Although Figure 21.10's area is about twice as large as that of Figure 21.7(a), the data density is much larger in Figure 21.10 than in Figure 21.7(a).

Example 21.3.4: Inflation Rate Forecasts. Figure 21.10 is a complex graph that contains much information about a complex subject, forecasting the inflation rate (CPI) for projections of Social Security funds (Frees et al. 1997). The graph shows actual experience of quarterly inflation rates up through the first quarter of

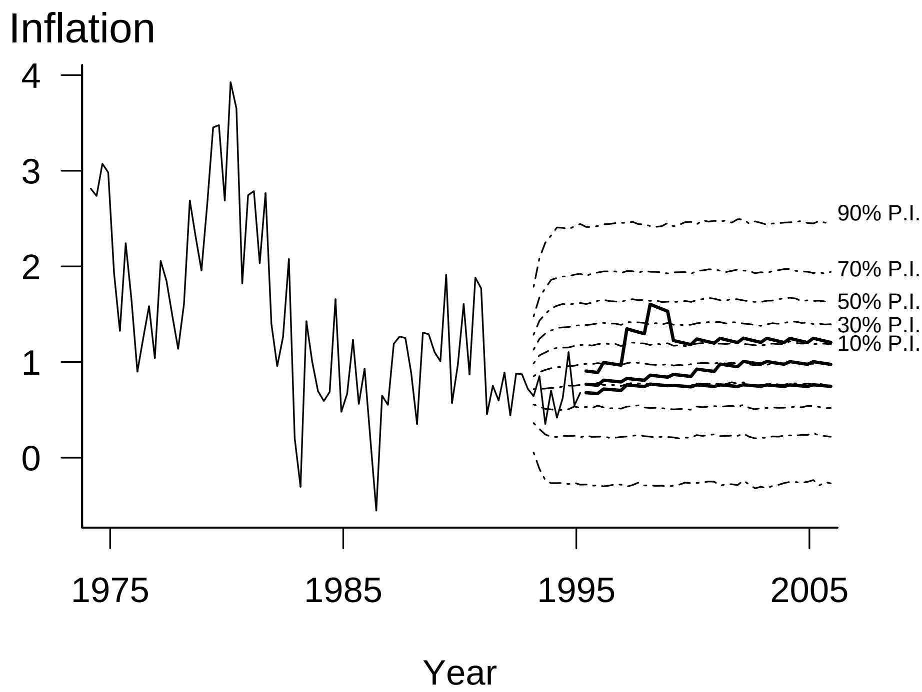


Fig. 21.10. Comparison of Stochastic Prediction Intervals to Held-out Actual Experience and to Social Security's Assumptions. The thin solid lines represent actual inflation rates, and the thick solid lines represent projections by Social Security experts. The dotted lines represent prediction intervals generated by a stochastic time series model. This complex graph allows viewers to make comparisons based on approximately 600 points.

1995. Experience up through 1992 was used to fit a time series model described in Frees et al. (1997), and this model was used to generate prediction intervals (PIs) of the inflation rate. These prediction intervals can be compared to held-out experience that was not used to fit the model (1993-1995) as well as projections of inflation by Social Security experts. The thick lines represent high-, intermediate-, and low-cost inflation projections determined by Social Security experts.

Figure 21.10 is complex and may not be immediately understood by the viewer. However, almost every stroke within the data region represents numerical information. Although complex, Figure 21.10 allows the viewer to compare (1) 20 years of experience to a 10-year forecast, (2) recent held-out experience to forecasts, and (3) expert projections to forecasts generated by a time series model. The graph's complexity reflects the complexity of forecasting inflation rates; this complexity is not due to unnecessary elements that distract viewers and make them more "interested" in the graph.

Guideline Five: Use Graphical Forms That Promote Comparisons

Creators of graphs are often faced with the choice of several graphical forms that could be used to represent a feature of the data. As we describe in Guideline Eight, the receiver's knowledge of graphical forms can influence the choice. Graphical perception is also an important determinant. In Section 21.4, we discuss this issue in detail. We include it here as part of the Guidelines Section for completeness.

Guideline Six: Integrate Graphs and Text

Data graphics should be carefully integrated with text, tables, and other graphs. A legend summarizes the graph and its main message, but the surrounding text develops the theme leading up to the message and discusses its impact. Although "a picture is worth ten thousand words," a graph needs supporting text. Tufte (1983) encourages readers and writers to think of data graphics as paragraphs and to treat them as such.

Data graphics can be complemented by a tabular presentation of data: graphics can highlight relationships among the data, and tables can present precise numerical descriptions of the data. The two modes are complementary. A good writing device is to place a graphical display in the main body of the report and to reinforce the graph with a tabular display in an appendix.

The American Statistical Association, in its *Style Guide* for journal publications, reminds us that a detailed legend is helpful when interpreting graphs. The *Style Guide* recommends that a legend describe a graph, draw attention to the graph's important features, and explain this importance.

Guideline Seven: Demonstrate an Important Message

Detailed legends and graphs should reinforce messages that are developed in the main body of the text. To illustrate, when considering ways of portraying a complex dataset, choose a graphical form that highlights an important message. All too often, creators of graphs display data features that are not part of the theme that is being developed.

Cleveland (1994) recommends that we "put major conclusions in a graphical form." In regression data analysis, major conclusions are about patterns in the data that are summarized using models. Usually major conclusions are best presented graphically. Graphs display a large amount of information that is retained by the viewer because it is visualized. Graphs communicate patterns directly to a viewer, without using an equation to represent the patterns. In this way, a wider audience can be reached than if the presentation relies solely on a model-based interpretation of the data. Further, patterns suggested by a graph reinforce those represented by a model, and vice versa. Thus the two tools, graphs and models, reinforce and strengthen one another.

"The greatest value of a picture is when it forces us to notice what we never expected to see." Tukey (1977)

Tukey (1977) states that "The greatest value of a picture is when it forces us to notice what we never expected to see." Unexpected phenomena are usually memorable events; viewers of graphs remember these results, which makes them powerful. In writing this chapter, we did not expect the results of Figure 21.6. This figure demonstrates that unemployment rises much more quickly than it declines; it is a powerful example of the use of aspect ratios.

Guideline Eight: Know Your Audience

A basic precept of effective writing, familiarity with one's audience, is also valid for designing effective graphs. As stated in the Introduction, our primary motivation in developing guidelines is to encourage the precise and concise communication of quantitative ideas to a scientific audience using a written medium. As discussed in Section 21.4, the graphical form is subservient to the real role of the graphical display, *communicating* quantitative ideas of the creator to the viewer of a graph. If the audience does not have an understanding of the graphical form, then the form will hinder the communication flow rather than aid it. Thus, each of the seven guidelines already discussed can be modified or even ignored upon occasion, depending on the audience for the graph. To illustrate, in Example 21.3.1 we argued that the dot plot was superior to the three-dimensional stacked bar chart. As another example, in Section 21.4 we argue that pie charts are ineffective communicators of information based on the science of cognitive perception. However, for some audiences, creators of graphs will prefer the less effective forms based on the level of audience familiarity. We hope that practice will eventually shift from these ineffective modes of communication. Still, it is important to recognize the background of the audience of the graph. We recommend that creators of graphs not so much swim against the tide of poor graphic design as bend their course towards more effective modes of communication.

21.4 Empirical Foundations For Guidelines

This section consists of two different scientific aspects of graphical studies: science of perception and surveys of graphical practice.

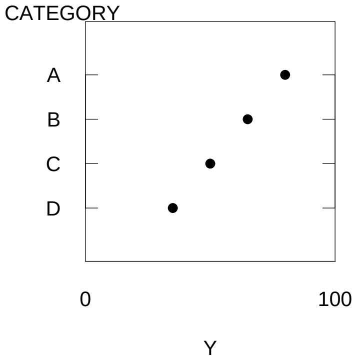
This chapter does not include a number of graphical forms that are mainstays in business publications and the popular press, such as pie charts, pictographs, and stacked bar charts. In fact, we have shown stacked bar charts in Section 21.3 only as an example of how *not* to draw figures. Why are these widely used graphical forms not adopted in an chapter emphasizing data graphics? The reasons lie in how graphical forms communicate information and how we perceive graphical information. We demonstrate that, given how we perceive information, pie and stacked bar charts are poor communicators of numerical information.

As described in Section 21.1, data graphics encode information, and we, as viewers, decode this information when viewing a graph. The efficiency of this transmission can be considered in the context of cognitive psychology, the science of perception. This discipline provides a framework for distinguishing among different types of information processing that we do when decoding graphs. Identifying different types of information processing will help us decide what are effective, and ineffective, graphical forms.

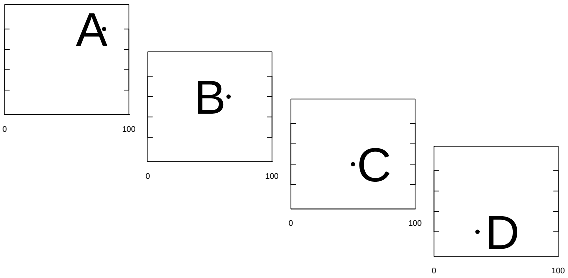
21.4.1 Viewers as Units of Study

Table 21.1 is an ordered list of basic graphical perception tasks, according to Cleveland (1994). Here, the ordering begins with a set of tasks that is least difficult for a viewer to perform and ends with a set that is most difficult. Thus, for example, judging position along a common scale is the least difficult for viewers and judging relative shadings of colors and density (the amount of ink) is the most difficult.

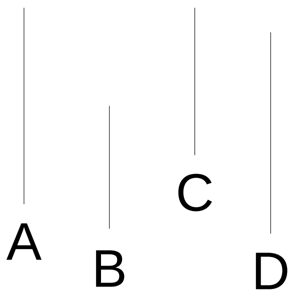
To understand the relative difficulty of the tasks, Cleveland and McGill (1984) performed a series of tests on many experimental subjects. To illustrate, Figures 21.11-**??** presents a series of tests that are analogous to the first five tasks. Cleveland and McGill summarized the performance of the experimental subjects by calculating the accuracy with which the subjects performed each set of tasks. Through these measures of relative accuracy, and arguments from cognitive psychology, Cleveland and McGill developed the ordering presented in Table 21.1.

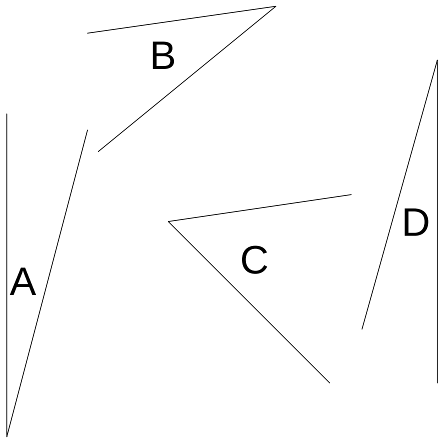


(a) **Experiment to Judge Position along a Common Scale. Assess the relative values of A, B, C and D along this 100-point scale.**



(b) **Experiment to Judge Position along Identical, Nonaligned Scales. Assess the relative values of A, B, C and D on a common 100-point scale.**





(c) **Experiment to Understand Length Judgments. Suppose line A is 100 units long. Assess the relative lengths of lines B, C and D.**

(d) **Experiment to Understand Angle Judgments. Suppose angle A is 100 units. Assess the relative values of angles B, C and D.**

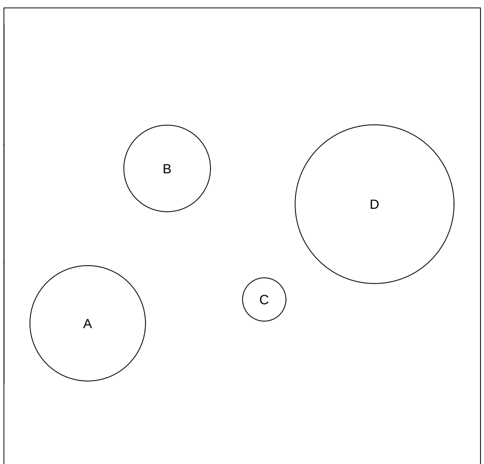


Table 21.1. *Basic Graphical Perception Tasks*

1. Position along a common scale
2. Position along identical, nonaligned scales
3. Length
4. Angles and slopes
5. Area
6. Volume
7. Color and density

This chapter does not discuss the use of color because of the complexities of coding and decoding it effectively. We refer interested readers to Cleveland (1994, Section 3.13) and Tufte (1990, Chapter 5) for further information.

The ordered list of graphical perception tasks can help the creator choose the appropriate graphical form to portray a dataset. When confronted with a choice of two graphical forms, a creator should select the form that is least difficult for the viewer. Other things being equal, a task that can be performed with little difficulty by the viewer means that information can be transmitted more reliably. To illustrate, we discuss two examples in which Table 21.1 can help you decide on the appropriate graphical form for portraying a dataset.

Example 21.4.1: Distribution of Premium Income. The first example demonstrates some shortcomings of the stacked bar chart. For this discussion, we return to Example 21.3.1. Figure 21.7(a) is a three-dimensional stacked bar chart. We have already discussed the substantial amount of chartjunk in this figure. Even without the useless pseudo third dimension, the stacked bar chart requires the viewer to make length judgments to understand, for example, the distribution of annuity receipts over time. In contrast, the dot plot in Figure 21.7(b) requires the viewer to make comparisons only according to positions along a common scale. As described in Table 21.1, the latter is an easier task, resulting in more reliable information for the viewer. Thus, we conclude that the dot plot is preferred to the stacked bar chart.

Example 21.4.2: Distribution of Mortgages. Our second example demonstrates the inadequacy of pie charts. Figure 21.12 is an adaptation of the figure on page 100 of the *Life Insurance Fact Book* (1994). It reports, for the years 1973, 1983 and 1993, commercial, 1- to 4-family, and farm mortgages as percentages of total mortgages. Pie charts make comparisons difficult. For example, the graph makes it difficult to detect whether farm mortgages are more prevalent than 1- to 4-family mortgages in 1983, or whether farm mortgage percentages increased or decreased from 1973 to 1983. The comparison of percentages across years is a linear operation, yet the pie charts require us to decode angles, a difficult task according to the ordering in Table 21.1. As with Example 21.3.1, the charts in Figure 21.12 make things worse by reporting in three dimensions; these figures not only require us to decode volumes but also add substantially to the chartjunk in the graphic. *Only nine numbers are reported in this graphic*, three years and two percentages in each year. (The third percentage can be computed by subtraction.)

If a graphic is needed, then the dot plot in Figure 21.13 is more than sufficient.

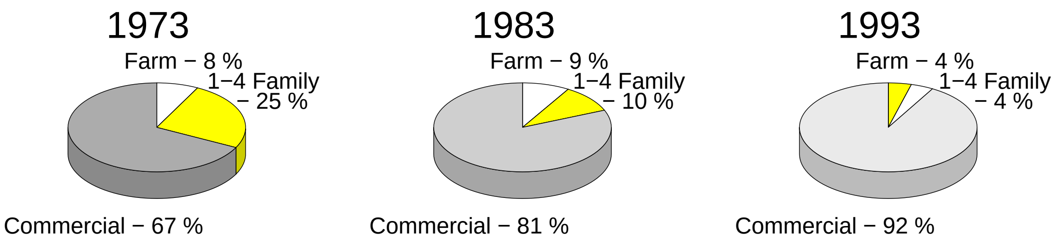


Fig. 21.12. Distribution of Mortgages for the Years 1973, 1983 and 1993. The three-dimensional pie chart is a poor graphical form for making comparisons over time and across types of mortgages.

Here, comparisons are made according to positions along a common scale, a task easier than comparing angles. Pie charts require us to make comparisons using angles, which are more difficult and less reliable than comparisons using other graphical forms.

Fig. 21.13. Commercial, 1- to 4-Family, and Farm Mortgages as Percentages of Total Mortgages for 1973, 1983 and 1993. A negative aspect of this graph is the overlap of the 1- to 4-family and farm plotting symbols in 1983 and 1993.

Although Figure 21.13 is a more effective graph than Figure 21.12, for these data we recommend a tabular display (Table 21.2), which allows for clear comparisons across mortgage types and across years. Further, more detailed information about mortgage percentages is available in Table 21.2 than in Figure 21.12 or 21.13. Of course, we can always superimpose the actual percentages, as is often done with pie charts and as illustrated in Figure 21.12. Our response to this approach is to question the worth of the entire graph. As with writing, each stroke should offer new information; let creators of graphs make each stroke tell!

21.4.2 Graphs as Units of Study

Surveys of graphical practice in professional publications provide an important database with which to assess prevalence of good and bad practice and changes in practice over time. Tufte (1983, pp. 82- 86) discusses a survey of approximately 4,000 graphs randomly selected from 15 news publications for the years 1974 to 1980. The graphs were assessed for "sophistication," defined as presentation of relationship between variables, excluding time series or maps. Cleveland and McGill (1985) report a similar survey of scientific publications, assessing the prevalence of graphical errors.

Table 21.2. *Commercial, 1- to 4-Family, and Farm Mortgages as Percentages of Total Mortgages for 1973, 1983, 1993*

	Year		
Mortgage Type	1973	1983	1993
Commercial	67.5	81.3	91.7
1-4 Family	24.9	10.1	4.1
Farm	7.6	8.6	4.2

Harbert (1995) assessed every graph and table in the 1993 issues of four psychology journals on 34 measures of quality. The measures of quality were gleaned from the current research literature on graphic quality. They were converted into a check sheet, and a check sheet was filled out for each graph and table in the selected psychology journals. Harbert's study yielded data on 439 graphs and tables. We summarize the analysis of the 212 graphs.

Harbert assigned letter grades to the graphics: A, AB, B, BC, C, CD, D, DF and F. These grades reflected her overall evaluation of the graphs as communicators of statistical information. The grades were converted to numerical values: 4.0, 3.5, 3.0, 2.5, 2.0, 1.5, 1.0, 0.5 and 0.0. The numerical values were the dependent variable in a regression. The independent variables were the 34 measures of quality, suitably coded. The purpose of the study was to determine which factors were statistically significant predictors of the grades assigned by an "expert" evaluator of graphics. By trial and error, Harbert selected a multiple linear regression equation in which all the predictors were statistically significant (5% level) and no other predictors achieved this level of significance when added to the equation. Table 21.3 shows the variables included in the regression equation ($R^2 = 0.612$).

Table 21.3. *Factors Affecting Assessment of Graphic Quality, Harbert Study*

Variables with Positive Coefficients	Variables with Negative Coefficients
Data-ink ratio	Proportion of page used by graphic
Comparisons made easy	Vertical labels on Y-axis
Sufficient data to make a rich graphic	Abbreviations used
	Optical art used
	Comparisons using areas or volumes

Data-ink ratio was defined by Tufte (1983, p. 93) as the "proportion of the graphics ink devoted to the nonredundant display of data-information" or equivalently as "1.0 minus the proportion of a graphic that can be erased without loss of data-information." The data-ink ratio is more readily calculated than the data density measure defined in Section 21.3 of this paper. Optical art is decoration that does not tell the viewer anything new.

One variable that had been anticipated as very significant was data density, which is difficult and time-consuming to measure. An important finding of the study was that the easier-to-measure data-ink ratio and proportion of page variables were sufficient to predict the grades. A quotation from Harbert's thesis sums up the finding: "The highest grades were given to those graphics that take up small proportions of the page, have a large data-ink ratio, make comparisons easy, have enough data points, have horizontally printed labels, do not have abbreviations, do not have optical art, and do not use volume or 3-D comparisons" (Harbert 1995, p. 56).

As a small follow-up study to Harbert's work, we examined each of the 19 non-table graphics in the *Life Insurance Fact Book* (1994), assessing them on seven negative factors. Table 21.4 shows the percentage of graphs that displayed each of the negative factors.

Table 21.4. *Percentage of Graphs Displaying Negative Factors in Life Insurance Fact Book 1994*

Negative Factor	Percentage of Graphics
Use of 3-D bars	79 %
Grid lines too dense	79
Making comparison of time series values hard	37
Use of stacked bars	37
Growth displayed poorly	32
Use of lines that are wider than need be	16
Use of pies	5

Our review suggests that every graphic could have been reduced by 50% to 75% without loss of clarity. This observation is in keeping with Harbert's finding about the proportion-of-page variable. In a word, the graphs in the *Life Insurance Fact Book* could be produced much more ably. Doing so would improve the quality of communication and would potentially increase the respect with which knowledgeable professionals in other fields view the insurance industry.

We hope that other investigators will engage in further study of graphic practice in actuarial publications. By using data from such studies, the profession can improve its practice, making communications efficient and precise.

21.5 Concluding Remarks

The Society of Actuaries' motto is a quotation of Ruskin: "The work of science is to substitute facts for appearances and demonstrations for impressions." Armed with the guidelines outlined in this paper and discussed further in the references, actuaries can be leaders in presenting data graphically, thus substituting demonstrations for impressions. Surveys of recent actuarial literature should be the basis for assessing current practice. Editors and referees of professional publications can be especially influential in bringing about a rapid improvement in standards of practice. Moreover, actuaries can recommend and use statistics textbooks that pay attention to graphic quality.

Because actuaries read material that contains graphs, they are consumers. They should become tough customers! All too often the defaults in spreadsheet and statistical graphics software become the norm. Actuaries should not allow the choices made by software programmers to drive graphic quality or standards. Although it is easy to create graphs using defaults in the graphics software, the resulting graphs are seldom fully satisfactory. If a graph is not worth doing well, let's leave it out of our publications.

21.6 Further Reading and References

In addition to the references listed, other resources are available to actuaries interested in improving their graphic design skills. Like the Society of Actuaries, another professional organization, the American Statistical Association (ASA), has special interest sections. In particular, the ASA now has a section on statistical graphics. Interested actuaries can join ASA and that section to get the newsletter *Statistical Computing & Graphics*. This publication has examples of excellent graphical practice in the context of scientific discovery and application.

The technical *Journal of Computational and Graphical Statistics* contains more in-depth information on effective graphs. We also recommend accessing and using the *ASA Style Guide* at http: //www.amstat.org/publications/style-guide.html as an aid to effective communication of quantitative ideas.

Chapter References

American Council of Life Insurance. Various years. *Life Insurance Fact Book*. Washington, D.C.: ACLI.

Cleveland, William S. (1994). *The Elements of Graphing Data*. Monterey, Calif.: Wadsworth.

Cleveland, William S. (1993). *Visualizing Data*. Summit, N.J.: Hobart Press.

Cleveland, William S., Diaconis, P., and McGill. R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science* 216, 1138-1141.

Cleveland, William S., and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* 79, 531-454.

Cleveland, William S., and McGill, R. (1985). Graphical perception and graphical methods for analyzing and presenting scientific data. *Science* 229, 828-833.

Ehrenberg, A.S.C. (1977). Rudiments of Numeracy. *Journal of the Royal Statistical Society A* 140:277-97.

Frees, Edward W. (1996). *Data Analysis Using Regression Models.* Englewood Cliffs, N.J.: Prentice Hall.

Frees, Edward W. (1998). Relative Importance of Risk Sources in Insurance Systems, *North American Actuarial Journal* 2(2), 34-51.

Frees, Edward W., Kung, Yueh C., Rosenberg, Marjorie A., Young, Virginia R., and Lai, Siu-Wai (1997). Forecasting Social Security Assumptions, *North American Actuarial Journal* 1(3), 49-82.

Harbert, D. (1995). The Quality of Graphics in 1993 Psychology Journals, Senior honors thesis, University of Wisconsin-Madison.

Huff, D. (1954). *How To Lie with Statistics.* New York: Norton.

Schmid, C.F. (1992). *Statistical Graphics: Design Principles and Practices* Malabar, Fla.: Krieger Publishing Co.

Schmit, Joan T., and Roth, K. (1990).

Cost Effectiveness of Risk Management Practices, *Journal of Risk and Insurance* 57, 455-470.

Strunk, W., and White, E.B. (1979). *The Elements of Style*. 3rd ed. New York: Macmillan.

Tufte, Edward R. (1983). T*he Visual Display of Quantitative Information.* Cheshire, Conn.: Graphics Press.

Tufte, Edward R. (1990). *Envisioning Information.* Cheshire, Conn.: Graphics Press.

Tufte, Edward R. (1997). *Visual Explanations.* Cheshire, Conn.: Graphics Press.

Tukey, John (1977). *Exploratory Data Analysis.* Reading, Mass.: Addison-Wesley.

University of Chicago Press (1993). *The Chicago Manual of Style.* 14th ed. Chicago, Ill.