

DISCUSSION PAPER

M.C. McCracken
Informetrica Limited

Re: Demand Curves, Cost Functions, and
Tradeoffs in Statistical Data

Although it is common to apply economic concepts of demand, supply, and costs to the production of goods, it is less so with services. And with services provided as a public good, such as statistical data, it is unusual. Yet we expect government departments to establish priorities, work within budgets, and select optimal combinations of resources. As a first step, this paper will examine some dimensions of the supply and demand for statistical data.

Demand for Statistical Data

In the simple world of an economics textbook it is normal to consider two elements - the price per unit and quantity consumed per unit of time. For our purposes the "quantity consumed" will be the number of statistical series or data points used by an individual or "the public" in a year. It is likely that the demand looks something like Figure A. If "price" is interpreted to be not only the sales price but also the cost to the consumer of using the information, then the consumer will consume a quantity Q consistent with a price P . The position of the demand curve is fixed, given that the education, tools, and requirements of the consumer remain unchanged, and that other characteristics of the series remain constant.

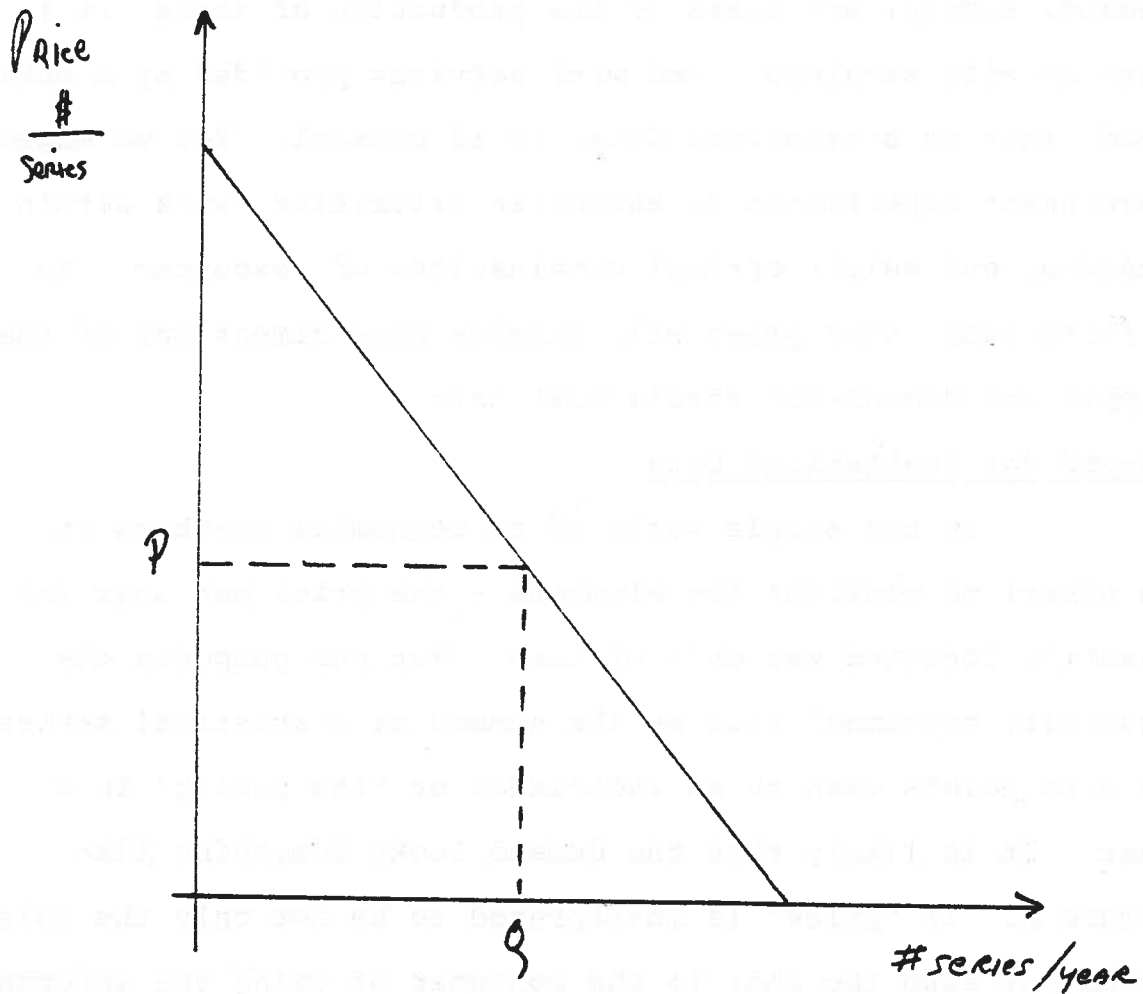


Figure A - Demand of an Individual for Statistical Data

But what happens when there are changes to these factors? At this point we should consider some other characteristics of the statistical data that may be relevant to the consumer. Three aspects often raised are timeliness, frequency of publication, and accuracy. More timeliness, frequency or accuracy are expected to shift the demand curve outward. That is, at a given price, the consumer will consume more series, or for the same number of series will be willing to pay a higher price.

I suspect that the relationships look something like those shown on Figure B for a given quantity of series. Below some level of timeliness, say T, there is no demand for the data. Similarly, there is some minimum level of accuracy (A) required. The maximum value for timeliness and accuracy has been assigned the value of 100. The underlying definitions might be:

$$\text{Accuracy} = 100 - \% \text{ error}$$

$$\text{Timeliness} = 100 (1 - \text{Elapsed Time in years} \\ \times \text{Frequency/Parameter})$$

For timeliness the parameter may depend on the variability of the data series, or the nature of the decision it supports. For discussion purposes I would suggest that the parameter is about 5. That is for a monthly frequency there should be less than 5 months elapsed time, for quarterly less than 5 quarters, and for annual data less than 5 years.

The units of frequency may be expressed as the number of times per year that the series is published. There is some minimum frequency, F, and some maximum for a given series.

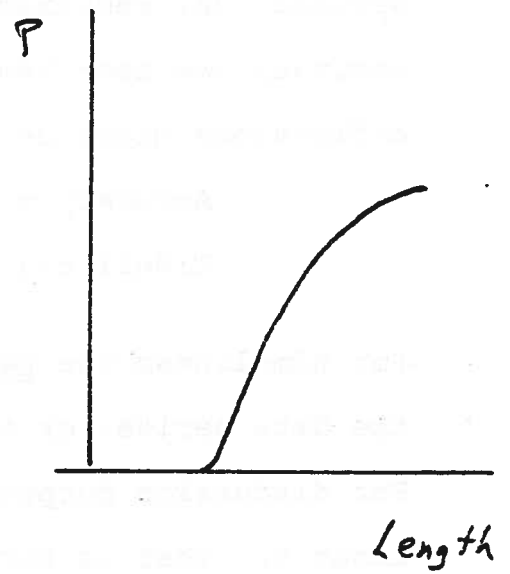
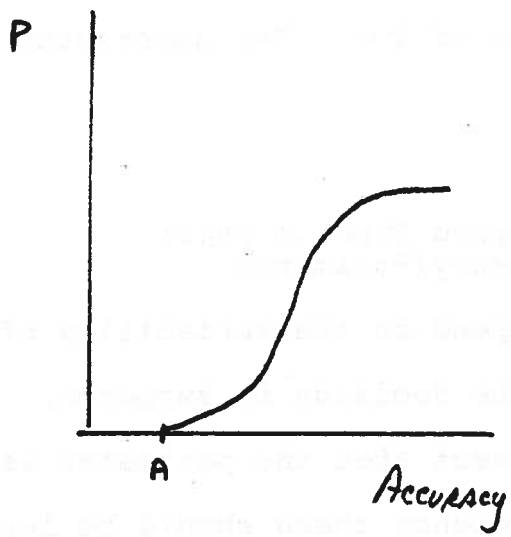
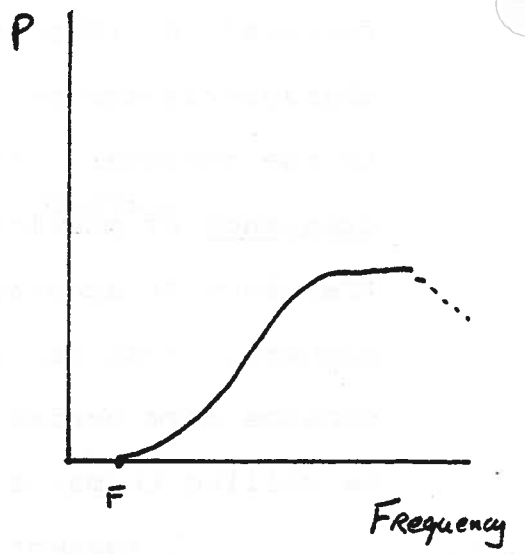
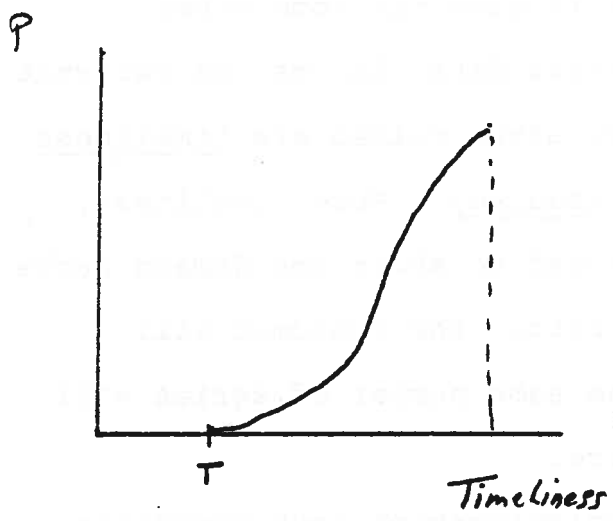


Figure B - Demand Curves for Characteristics

Frequency = number of times per year.

In the case of five-year census data the frequency would be 0.2. For exchange rates it could be 365 or greater.

Another characteristic that may be important to the consumer is the "length" of the series. In the case of time series this would mean the number of previous periods for which the data source was available without a break in methodology so that comparability was present. Different consumers will have varying requirements. Someone interested in recent changes may be satisfied with 12 months of history; for analytical purposes five or more years on monthly data may be needed for seasonal adjustment, modelling, etc. Again the "typical" curve is likely to apply as indicated on Figure B.

In the case of cross-sectional data (census, consumer survey) the analogous concept to "length" may be the availability of the data across dimensions (e.g., province, income class, age, etc.) that allows for contextual interpretation of a data point.

The measurement of length would be the number of observations in the data series, either over time or across some other dimension.

These characteristics (and probably others) are often referred to as the quality of the data; more properly the qualities of the data. Other qualities might include relevance, documentation, and discriminating power. Changes in these characteristics or qualities can shift the demand curve for statistical series and, along with the price, can be thought of as determining the demand curve.

Cost Curves

In the most simple terms, the production of statistical data is likely to be a decreasing-cost industry. Given that a database exists, a number of additional series could be produced at low marginal cost. Thus an average cost curve would appear as indicated on Figure C. But this cost curve is based on a given set of qualities for each series (e.g., timeliness, frequency, accuracy, length, etc.).

For a given concept the qualities will generally be increasing cost activities. The posited curves are shown in Figure D.

The posited shapes of the cost curves for a given number of series are hypothetical, but perhaps realistic. For timeliness, there is a range within which there is little increase in cost and then a range within which the costs can increase quite rapidly.

For frequency, there will be additional costs of observation at a more frequent interval. The dotted line would apply if the series was being observed monthly, but additional series were being generated at higher levels of

$\frac{C}{N}$
Series

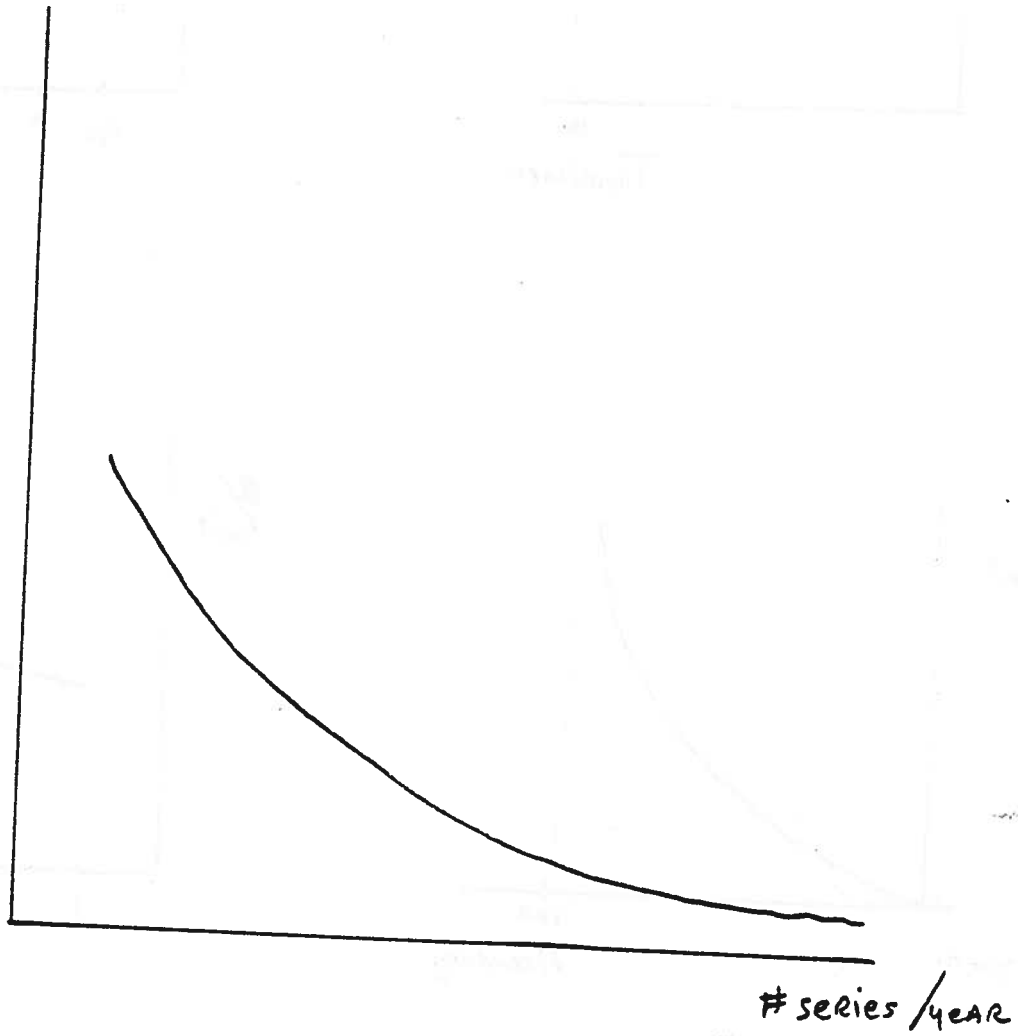


Figure C - Average Cost Curve for Statistical Data

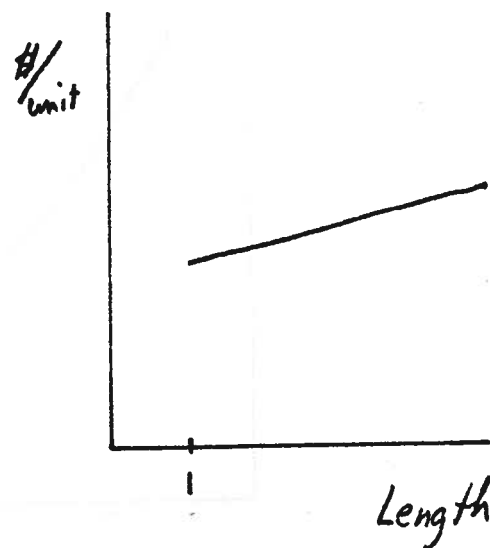
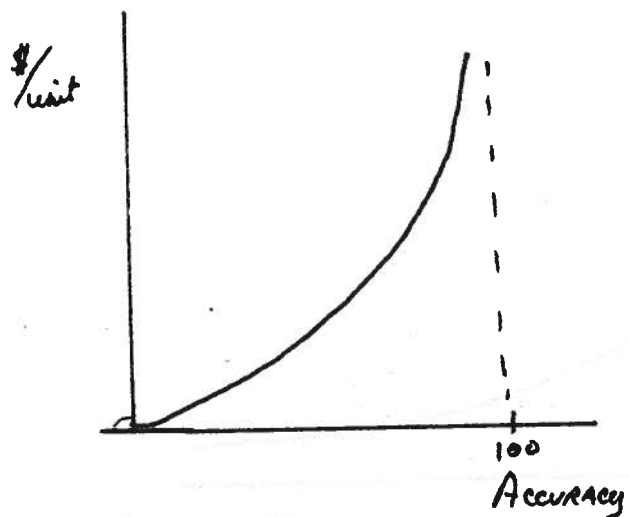
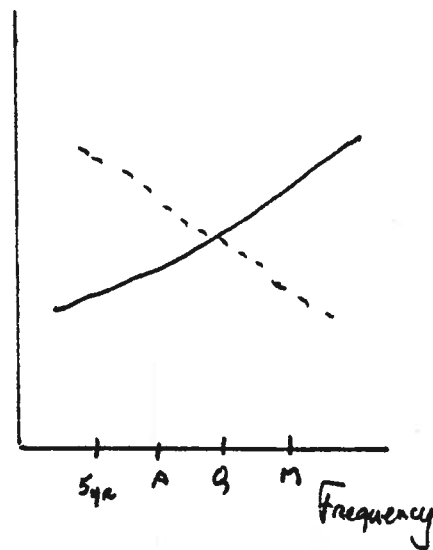
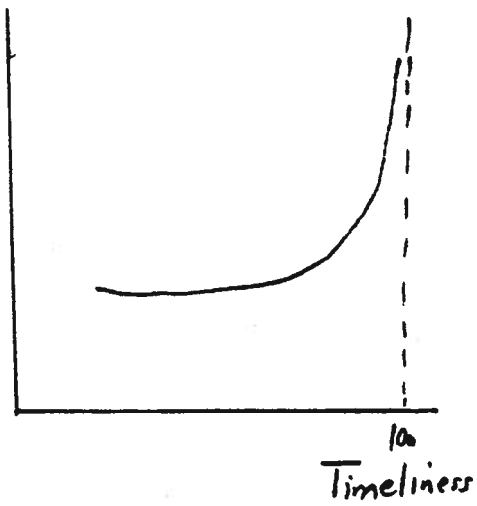


Figure D - Cost Curves for Characteristics

aggregation and published separately. (An expanded notion might distinguish between frequency of observation and frequency of publication.)

Accuracy is thought to be a continuously increasing cost activity at increasing marginal cost. In some cases, complete accuracy may be attainable at finite cost; in other cases there may be inherent limits of measurement.

The length of the data series incurs some increasing costs as a result of storage, but more importantly the cost of maintaining comparability over time increase with length.

Tradeoffs

Nor are these characteristics independent of each other. That is, for any cost curve it is necessary to assume that other characteristics are held constant. For the four characteristics treated above there are six "tradeoff" curves, reflecting the pair-wise linkage of the different characteristics. These are illustrated in Figure E.

Perhaps the most frequently discussed tradeoff is that between timeliness and accuracy, where trying to decrease the publication lag leads to inherent inaccuracy because of the delays in reporting, lack of time to check inputs, etc. One approach to reducing this tradeoff is to publish data points that are "timely" but "preliminary", and then subsequently revise them.

I conjecture that length and frequency are positively related, since we are measuring length by the number of observations and frequency by the number of times per year

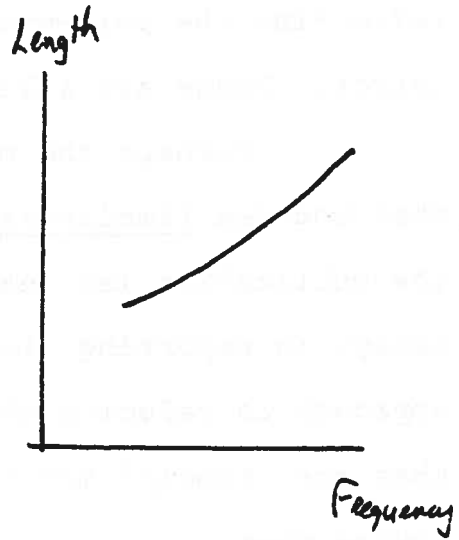
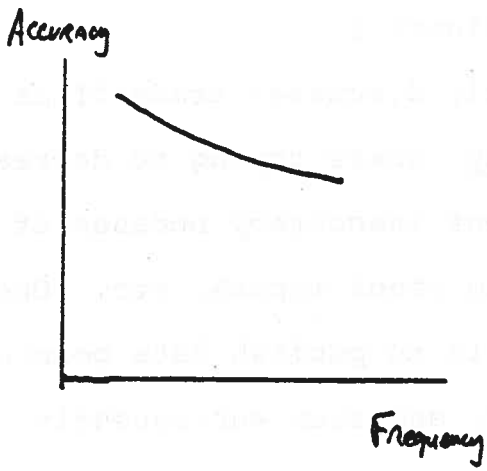
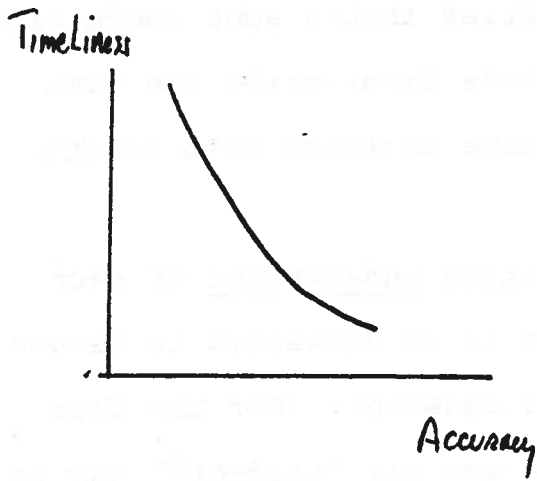
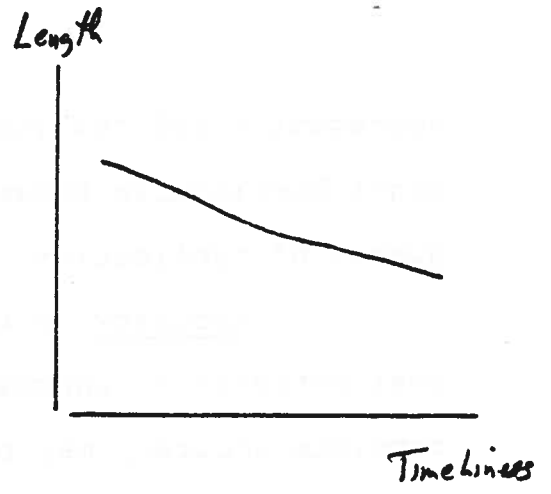
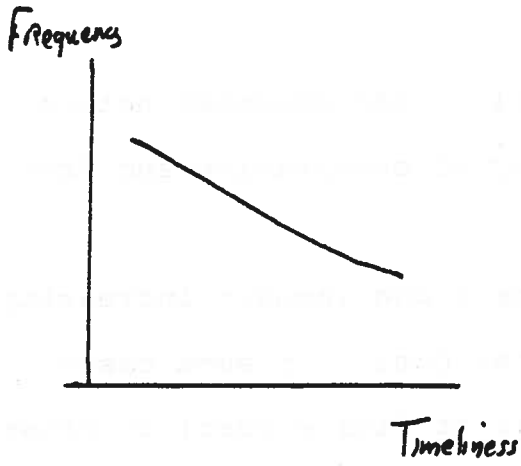


Figure E - Tradeoffs between Characteristics

published and observed. But if we measured length as the number of observations divided by frequency of observation then we may find a different curve because we stressed the comparability of data points in length. One reason for the frequent observation might be the changing nature of the observed phenomenon, which suggests that comparability may be difficult to maintain.

So What?

At this point, not much! A number of curves have been posited, suggesting that the demand and cost functions for statistical data are complex (as are most other realistic demand and cost functions). Before a meaningful distillation of demand and cost functions to identify the key characteristics determining the costs and demands can be done, it would be desirable to put units on the axes and to investigate the aggregation of individual demand functions and other dimensions of the cost structure.

I conjecture that accuracy can be improved by the number of series being collected; that timeliness can be improved by use of econometric models, and that technological change will impact both the supply and demand side ⁱⁿ ~~is~~ positive ways - allowing greater benefits to the user through lower cost of using data and reduced costs in producing a set of data with given qualities.

The reader is invited to feedback reactions to this paper; suggest other characteristics, argue with the shapes of the curves, and try to draw implications for organization, pricing of services, allocation of resources, etc.

