# Communicating quantitative information: tables vs graphs

Klein, Torsten L.

PAS Research Unit

11 December 2014

December 11, 2014

# COMMUNICATING QUANTITATIVE INFORMATION: TABLES VS GRAPHS

KLEIN, TORSTEN L.[*]

*In applied statistics and computational econometrics a key task for researchers is to bring the sizable but unstructured body of numeric evidence, for example from Monte Carlo simulation, in a form ready for introducing to scientific dialog. At their disposal they find established means of arrangement: narrative text, tables, graphs. Employing classical principles of communication to evaluate their suitability graphical devices seem optimal. They absorb large quantities of data, and organize content into a productive tool. Graphs confirm the advantage when put to work in a standard simulation exercise. However, theory and application contrast with the norm observed in peer-reviewed journals – by a wide margin and with considerable persistency researchers prefer tables.*

JEL CODE: *A14, C10, C15, C35, C52, Y10.*

KEYWORDS: *econometric and statistical methods, Monte Carlo, bivariate probit model, exogeneity testing; modes of communication, data visualization; economics of science.*

## CONTENTS

[*] PAS Research Unit, B7,17, 68159 Mannheim, Germany, `tl.klein@gmx.de`. I would like to thank Melanie Rübsamen for excellent research assistance. Rainer Fath, Gerald Meier and Jack Wu provided helpful comments. All errors are mine.

# 1 SENTENCE, TABLE, GRAPH: WEIGHING THE MEANS OF COMMUNICATION

When presenting numeric evidence researchers use, apart from standard word-for-word discourse, tabular or graphical devices. The two examples below come from a Monte Carlo study on statistical procedures and their characteristics. In table 1 the spreadsheet-style list groups quantities of interest according to simulation parameters in one dimension and the procedures under scrutiny in the other. Here, it assigns rejection frequencies of seven different test statistics to specific values the researchers chose for their degrees of freedom in the exercise: data generating process, number of observations, nominal significance level. A comparison between columns indicates, contingent on simulation parameters, the relative performance of the procedures as to empirical size, and so offers a basis to recommend one and dismiss the other. Instead of comparing rejection frequencies the graph in figure 1 employs *P* values of test statistics and their empirical distribution function (EDF). It is calculated for data generating process 2, samples of 500 observations and then displayed at a number of pre-chosen significance levels.[1]

Even though for communicating quantities sentence, table and graph are to a certain degree complementary in nature, mutually enhancing their abilities if used in combination, there are advantages to some and shortcomings to others.[2] Which one to choose? The classical principles of information exchange – precision and concision – bring the three means into order, with sentence and graphical devices at the poles of the spectrum and tables occupying the middle ground.

The sentence excels when conveying selected quantities from Monte Carlo simulations. A numerical value may not only be penned explicitly, but also commented on without delay for better understanding by the reader. Yet this runs into difficulty given the need to communicate a lot of data or interrelation among them that is multivariate and hence ill-suited for the linear nature of the sentence. Here resides the graph's main strength. It absorbs quantitative information in large amounts, consumes space to a minimum still,

[1] The main concepts and definitions used are assembled in appendix A.1. The graphical framework is introduced in detail by Davidson and MacKinnon (1998, section 2).

[2] The discussion follows Tufte (1983) closely, with special focus on section 9.

compressing data while making way for appraisal from various directions. Endowed with scales its axes connect data plots to numerical values, a technique more likely than direct statement to trigger inaccuracy, confusion or mistake. Small is its room for commenting, and thus the tolerance for overcharging quantitative displays with textual information. From these extremes the table occupies an intermediate position, combining the ability of the sentence to state precise quantities as well as supplementary information within a concise and relational format similar to the graph.

So, if presenting simulation data, informing readers of test properties, empirical size or other, which principle should one follow? Precision is irrelevant, concision indispensable.

Of course, any format chosen for display must avoid obscuring numerical results. But numbers given to the fourth decimal, scattered between chunks of words perhaps, perhaps shaped into litanies of coded record, offer few benefits for the task at hand: transfer information meaningful to the reader. Quite the opposite, they instill reassurance which may prove deceptive.

For one, each Monte Carlo experiment relies on a data generating process (DGP), a necessary abstraction from economic data encountered in the real world. Although recreating a shortlist of features deemed important by the researcher, every DGP neglects some characteristics and even rules out others explicitly – one choice only among innumerable possibilities. Hence, there will always remain some arbitrary element, read: imprecision, created by its design. A different kind of arbitrariness stems from the method which is applied to study a procedure's qualities. The researcher simulates tests of the null hypothesis drawing data samples repeatedly, each time calculating the statistic in question. He then estimates empirical size by taking an average, the rejection frequency, the outcome of a random variable. Sampling variability will lead to different estimates in different simulations, although conducted on the same DGP, number of observations and number of repetitions; it will make reproducing values identical to those in table 1 improbable.[3] The accuracy suggested by stating point estimates of empirical size cannot solve the randomness inherent to simulation. A third element in the trade off

[3] Take for instance 0.1022, the empirical rejection frequency resulting from DGP1 and samples of 500 observations, an estimate for the true size of procedure LR at the critical value that corresponds to a nominal level of 10 %. Replicating the simulation yields the rejection frequency 0.1114 with a 95 % confidence interval, calculated according to Davidson and MacKinnon (1998, 18), of [0.1054, 0.1233].

between precision and concision concerns which numerical values to report. True, focusing on the empirical companions of "conventional" suspects as in table 1 may first come to mind. Apart from econometric custom however, ten, five, one per cent levels command no special merit over others.[4] *P* value plots also make a choice in selecting the nodes to be displayed. Yet it is based on a criterion of information processing rather than space restriction: distribute nominal significance covering an interval wide enough for the reader to infer the properties of tests.

Despite their inaccuracy, stemming from the degrees of freedom granted to users and the randomness introduced by method, Monte Carlo experiments are in ample use. They offer a quick escape route from the prison of asymptotic inference, determining test properties in samples of finite length. Insights come from three avenues to discovery that strain all modes of display and their ability to carry the associated information. First, in a given simulation environment the test under scrutiny may be related to a theoretical or asymptotic absolute: is the true null rejected too often or not often enough, is there a high risk of not rejecting a false null? The evidence may then be set against competing procedures. Finally, changing the environment may offer answers on how previous results vary with simulation parameters. Framing test properties from multiple angles induces a challenge to sentence, table and graph alike since each is tied to the medium of communication the researcher employs. In case of the printed page or the inanimate computer screen this leaves two dimensions at their disposal to absorb multivariate data produced by the simulation. For the graph, though, there is a way out. The small multiple loosens the constraint of rectangularity without sacrificing clarity. It also adds to the strength of conveying large amounts of information using minimal space.[5]

[4] Cf. Davidson and MacKinnon (1998, 14) and Fiorio et al. (2010, 283, footnote 10).

[5] For the "curse of dimensionality" in communications (rather than econometrics) cf. Tufte (1990, 12–15). Tufte (1983, 170–174) and Tufte (1990, 67–79) introduce the concept of small multiples and cite numerous applications dating back as early as the Seventeenth century. In econometrics small multiples fit well with the topics studied by vector autoregressions, and thus already find their way into pioneering contributions such as Sargent and Sims (1977) or Sims (1982); for an example from a simulation study, cf. Davidson and MacKinnon (1999). An alternative to deal with dimensionality issues are surface objects, cf. Davidson and MacKinnon (1998, 5) and Arribas-Bel et al. (2011). Leaving the confinement of bi-dimensional media numerous possibilities arise for the display of quantitative information, cf. Chen et al. (2008).

Figure 2 suggests a small multiple that reports on the same Monte Carlo experiment as table 1.[6] It organizes simulated evidence on procedures' empirical size from all parameter pairings: switching DGP for a given sample length while moving down a column, or increasing sample length for a given DGP while moving through a row. Now compare. The table's spreadsheet anatomy leaves each procedure three slots to state the values of empirical size associated with simulation settings, settings that its stub indicates and orders linearly, one variation after the other. The small-multiple's matrix concedes 108 slots which correspond to the node count the researcher picks. Instead of citing values it merely labels a necessary few coordinate pairs for orientation; instead of sequencing variation on the line it spans a senior plane of Cartesian tuples – of DGP and sample length. The number of size results stored to the graph is far superior thanks to its aptitude for compressing information, yet in reality this seems only a secondary virtue. For why would providing more evidence imply to facilitate its appraisal? Tipping the balance is not the mere excess amount of data displayed. It is a skill of the nuclear *P* value plot to craft them into one unit thereby revealing information absent from its constituents, and a skill of the small multiple to show this unit evolve through each simulation stage using an arrangement that facilitates inquiry and furthers understanding.

Concision counts. Precision is neither feasible nor necessary nor desired. Therefore, to share quantitative information, to get across the point from a Monte Carlo experiment in economy and clarity, the format is the graph. Nevertheless, closing here would mean closing early. The fit of graphical devices reaches beyond the publication stage: to exploratory data analysis. Patterns and regularities are easily discovered, their dependencies on simulation design made transparent.[7] This is shown in the next section by employing a small multiple of *P* value plots to reassess the simulation that governs table 1 and setting off discoveries against those from its spreadsheet competitor.

[6] With figure 1 as a starting point the companion paper Klein (2014a) lists the steps taken to arrive at the small multiple.

[7] A vintage example of using graphical methods in data exploration is Anscombe's quartet, cf. Anscombe (1973) and Tufte (1983, 13–14).

# 2 A PRACTICAL ASSESSMENT FROM TESTS OF EXOGENEITY IN THE BIVARIATE PROBIT MODEL

Table 1 and figure 2 display results from the Monte Carlo experiment proposed in Monfardini and Radice (2008). They investigate size characteristics of seven procedures testing exogeneity in a bivariate probit model. Exogeneity requires that across the two relations determining the latent variables stochastic disturbances are uncorrelated. All procedures test the null hypothesis of zero correlation, and differ by the a priori restrictions they impose as well as the quantity of information they require.[8]

Although Figure 2 concurs with the evidence in table 1 multiple *P* value plots offer some refinement to analysis.

> LR systematically outperforms the other tests for all values of *N* and different nominal levels [i.e. at 10%, 5%, 1%].
>
> – Monfardini and Radice (2008, 276)

Indeed, the likelihood ratio test retains the smallest distance between empirical and nominal size at the nodes presented by the small multiple. Sorting through its rows makes the distance as well as the size advantage over other procedures decline at a diminishing rate with sample information available. This monotonicity is stalled in case of DGP2 and DGP3. Here graphical observation detects a marked pattern of under-rejection for small samples in the vicinity of conventional significance levels. If information is sufficient LR is also the only procedure to remain within 0.05 critical values of the Kolmogorov-Smirnov test, delimiting the shaded area: for DGP1 and 1,000 or 2,000 observations the null, differences of the *P* value plot from the *P* values' theoretical cumulative distribution are due entirely to experimental randomness, can no longer be rejected at the 5 per cent significance level. Furthermore, having to evaluate the bivariate probit model from non-standard samples does not always

[8] The study of empirical size retraced here constitutes only a minor part of the paper. Instead, the authors focus on model identification and misspecification, cf. Monfardini and Radice (2008, 277–280). Details on the bivariate probit model, testing procedures, and simulation environment are relegated to appendix A.2. Results for the power of tests in a small multiple version of the Davidson-MacKinnon framework do not alter those obtained from size characteristics. They are available from the author upon request.

entail a clear advantage for LR which considers both specifications, restricted and unrestricted. This becomes obvious from scanning the small multiple column-wise. DGP1 seems easiest to estimate, yet provides LR with a larger edge over others than the more challenging DGP2.

> RHO tends to display high over-rejection patterns, especially in DGP2 and DGP3 (the most difficult to estimate), where over-rejection remains serious even when $N = 2{,}000$.
>
> – Monfardini and Radice (2008, 276)

Similar to other tests the size distortion of RHO, a Wald-type statistic, becomes more problematic if the dichotomous variable suspected for endogeneity is observed with outcomes favoring either 0 or 1, and if the number of observations is small. For these combinations of DGP and sample length figure 2 plots discrepancies between the empirical distribution function and its uniformly distributed theoretical counterpart well above zero. Parting with likelihood ratio, score or conditional moment procedures however, the common hump occurs at low values of nominal significance only. Hence, over-rejection tends to be gravest at conventional levels of ten, five and one per cent. In addition, the small multiple suggests that over-rejection declines with sample size slightly faster under DGP3 than under DGP2.

> CM and LM1 [. . .] give highly similar results [and] display empirical sizes quite close to the nominal ones [. . . They constitute] a reliable inference tool for testing exogeneity without simultaneous estimation.
>
> – Monfardini and Radice (2008, 276)

From every combination of DGP and sample length the conditional moment test gives a smaller size distortion than the Lagrange multiplier procedure, except for a few singular nodes where performance is identical. As seen before, results vary systematically with simulation parameters. The advantage of CM is more pronounced for DGP2 and DGP3 as well as in small samples; with evidence on $y_{1t}$ scarce and leaning towards one outcome the gap between empirical and nominal size widens sharply too, above all for LM1. Both effects serve as a forceful warning that all simulation evidence remains conditional, tied to the particularities of a setting chosen. Minor deviations like shrinking samples a little further may alter results – and verdict – to large extent.

> [T]he set of LM tests (LM2, LM3, LM4) originated by the Kiefer formulation of the test statistic variance displays very unsatisfactory size properties in finite samples, with zero rejection frequencies for the sample dimensions analyzed.
>
> – Monfardini and Radice (2008, 276)

*P* value plots confirm the unsatisfactory performance of the three tests, yet figure 2 omits them.[9] The little knowledge gained from inclusion is outweighed by littering space with redundant facts which will inevitably mar comprehension. Table or figure, both methods of display cannot apply their skill to condense information since each datum gives the same. The quoted sentence carries all content.

Even though irrelevant for the communication of size results the graphical tool helps when exploring their link to simulation characteristics, and along the way creates fresh insight. Because the empirical distribution function of LM2, LM3 and LM4 is skewed less compared to other procedures, empirical and nominal size attain their maximum distance at higher nominal significance values; lower skewness means not only equally poor results at conventional levels, but also negligible effects if simulation parameters change, as reported in table 1. Taking into account the whole domain of EDF, size distortion tends to be higher for DGP2 than for DGP3, and for DGP3 higher than for DGP1. Similar to other tests rejection frequency decreases with the number of observations, albeit at a diminishing rate. In consequence, under-rejection of the three statistics actually deteriorates from an increase in sample length. LM2 and LM4 display similar size properties while LM3 fares somewhat "better" under DGP1 and DGP3, but even worse for DGP2.

## 3 MAKING A SENSIBLE CHOICE IN THE REALITY OF ACADEMIC PUBLISHING?

The clean split between means of communicating quantitative information in section 1 is effective for introducing available possibilities, necessary when discussing their benefit or flaw, and, of course, wrong.

[9] The *P* value plots are reproduced in Klein (2014a).

Sentence, table and graph can hardly replace each other: divergent features turn them into complements. How to organize simulation results in prose without tabular or graphical devices converting those of interest in a readily digestible, concise format? How to access a graph's or table's contents without textual guidance on the way to read them, without comment on the bottom line of the exercise? Thus, the point in this note is not to advocate one over the other, but to suggest that choosing the means of communicating quantitative information – just like choosing simulation design – belongs to the researcher's degrees of freedom. It deserves conscious effort. Assigning contents to sentence, table or graph he decides on which elements of the Monte Carlo experiment to disclose and which liberties to grant to readers.[10] If the researcher aims for providing comprehensive quantities of data while allowing his audience to explore simulation results for themselves he will supply a graph. If he prefers to regulate the flow of information more tightly and leave readers less room to draw their own conclusions he will offer numeric evidence in a table or resort to textual presentation only.

Researchers choose the second option. This is the upshot from a cursory look at published practice during the past quarter of a century in table 2. When reporting quantitative information tables rule. Graphs do feature in articles on Monte Carlo simulation, although they mostly take a subsidiary role. While exclusive use has somewhat expanded since the late 1990s they fall clearly short of tabular devices accompanying narrative displays consistently in well over half of the published work in the sample. Explaining preference for tables over graphs does not preoccupy this brief note – the economics of academic publishing has already established a collection of factors that will operate on the choice between the two as well.[11] To discover empirically which of these factors in the end account for such distinct yet stable bias sets the task in Klein (2014b).

[10] Cf. Tufte (1983, 192).

[11] Zamora Bonilla (2012) reviews the main hypotheses while some of their adverse consequences in empirical research are presented by Ioannidis and Doucouliagos (2013).

# A APPENDIX

## A.1 ELEMENTS OF COMMUNICATING QUANTITATIVE INFORMATION: REJECTION FREQUENCIES, THE $p$ VALUE AND ITS EMPIRICAL DISTRIBUTION FUNCTION

For the purpose of analyzing test statistic $\tau$ a Monte Carlo simulation is conducted based on a data-generating process (DGP) to obtain $j = 1, \ldots, M$ realizations $\tau_j$.[12] With the aid of these realizations rejection frequencies like those in table 1 may be obtained. They are calculated for a chosen significance level $\alpha$ as the ratio of two figures: the number of times $\tau_j$ exceeds a corresponding critical value $c_\alpha$ from the statistic's cumulative distribution function $F(\tau)$, and the number of simulations:

$$\frac{1}{M}\sum_{j=1}^{M} 1[\tau_j > c_\alpha] \tag{1}$$

where $1[\cdot]$ is the indicator function.[13]

The $P$ value, or marginal significance level, of statistic $\tau$ assigns to each realization $\tau_j$ the probability of observing a value of at least $\tau_j$:

$$p_j \equiv p(\tau_j) = 1 - F(\tau_j).$$

Its empirical distribution function (EDF) is an estimate of the cumulative distribution function of $p(\tau)$:

$$\hat{F}(x_i) \equiv \frac{1}{M}\sum_{j=1}^{M} 1[p_j \leqslant x_i], \tag{2}$$

where $x_i \in (0, 1)$.

The $P$ value plot displays $\hat{F}$, generated from a DGP under the null hypothesis and evaluated at a series of nodes.[14] Using the correct distribution $F(\tau)$, $p_j$ is the realization of a random variable distributed

[12] This appendix merely restates the presentation in Davidson and MacKinnon (1998, 2–3).

[13] The distribution under the null hypothesis $F(\tau)$ is known and may originate for example from asymptotic theory or bootstrap simulation. A critical value $c_\alpha$ is defined by $1 - F(c_\alpha) = \alpha$. The frequency given assumes a one-sided test, rejecting in the upper tail of the distribution. This is consistent with statistics reported in table 1.

[14] To sketch their diagnostic tools Davidson and MacKinnon (1998, 3) recommend two series of significance levels $x_i$, $i = 1, \ldots, n$: either $x_i$ = 0.002, 0.004, . . ., 0.01, 0.02, . . ., 0.99, 0.992, . . ., 0.998 ($n = 107$); or $x_i$ = 0.001, 0.002, . . ., 0.010, 0.015, . . ., 0.990, 0.991, . . ., 0.999 ($n = 215$).

uniformly in the interval [0, 1]. This will result in a graph close to the 45° line, and signal that $\tau$ maintains correct size.

## A.2 THE BIVARIATE PROBIT MODEL – TESTING EXOGENEITY AND SIMULATING SIZE

Monfardini and Radice (2008) consider a recursive model for two latent variables, the potentially endogenous variable $y_{1i}^*$ and the variable of interest $y_{2i}^*$:[15]

$$\begin{aligned} y_{1i}^* &= \boldsymbol{\beta}_1' \boldsymbol{x}_{1i} + u_{1i} && (3) \\ y_{2i}^* &= \boldsymbol{\beta}_2' \boldsymbol{x}_{2i} + u_{2i} = \delta_1 y_{1i} + \boldsymbol{\delta}_2' \boldsymbol{z}_{2i} + u_{2i}. && (4) \end{aligned}$$

Both are tied to observable dichotomous variables $y_{1i}$ and $y_{2i}$ with outcome:

$$y_{ji} = \begin{cases} 1, & \text{if } y_{ji}^* > 0 \\ 0, & \text{if } y_{ji}^* \leqslant 0 \end{cases} ; \; j = 1, 2. \qquad (5)$$

Exogenous variables of the model include $\boldsymbol{x}_{1i}$ and $\boldsymbol{z}_{2i}$. They affect latent variables according to location parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\delta}_2$. The relation is disturbed by error terms $u_{1i}$ and $u_{2i}$ which follow a bivariate normal distribution with unit variance and correlation $\rho$. Across observational units $i$ there is no dependence. If errors are uncorrelated the dummy regressor $y_{1i}$ in the equation for $y_{2i}^*$ will be exogenous. The model's $K$ parameters $(\boldsymbol{\beta}_1' \; \boldsymbol{\beta}_2' \; \rho)$ are estimated from a sample of $N$ observations in compliance with the maximum likelihood principle.

Testing $\mathrm{H}_0$: $\rho = 0$ Monfardini and Radice examine standard MLE approaches to asymptotic inference: a likelihood ratio test requiring estimation of both the general and the restricted model, a Wald-type test adopting the former only, Lagrange multiplier as well as conditional moment procedures exclusively the latter. Details on each approach and its constituent elements are grouped by table 3. The four Lagrange multiplier tests differ in the method they employ to calculate the observed information matrix $-\mathcal{H}_N$. While LM1 uses the outer product of gradient, all others rely on the asymptotic block-diagonality of the scaled Hessian under $\mathrm{H}_0$.[16] Therefore, the $(K, K)$ element of $(-\mathcal{H}_N)^{-1}$ selected by the score vector in the statistics'

[15] The exposition of the model and simulation environment is adapted from Monfardini and Radice (2008, 272–276).

[16] Cf. Kiefer (1982)

quadratic form equals the inverted $(K, K)$ element of $-\mathcal{H}_N$. To arrive at this item LM2 computes the outer product of gradient, LM3 its probability limit, and LM4 the corresponding entry in the Hessian. With a single parameter restriction imposed all test statistics have limiting $\chi^2(1)$ distributions if $\rho = 0$.[17]

For their Monte Carlo experiment Monfardini and Radice make the bivariate probit model of equations (3) to (5) operational. They formulate three different DGPs, described in table 4, and use them to draw samples of $N = 500$, 1,000, or 2,000 observations. They estimate the model, calculate the seven statistics given by table 3, and compare them to critical values associated with nominal significance level $\alpha$. They repeat each step until $M = 5,000$ realizations are obtained for determining the rejection frequencies cited in table 1. Replicated simulation differs from the original in two ways. It combines DGPs with every sample length conceived, and includes two shorter versions of $N = 50, 100$. In order to smooth $P$ value plots it increases the number of repetitions to $M = 100,000$.

## REFERENCES

Anscombe, Francis J., 1973, "Graphs in statistical analysis", *The American Statistician* 27(1), 17–21.

Arribas-Bel, Daniel, Julia Koschinsky and Pedro V. Amaral, 2011, "Improving the multi-dimensional comparison of simulation results: a spatial visualization approach", *Letters in Spatial and Resource Sciences* 5(2), 55–63.

Chen, Chun-houh, Wolfgang Härdle und Antony Unwin (eds.), 2008, *Handbook of data visualization*, Berlin: Springer.

Davidson, Russell and James G. MacKinnon, 1998, "Graphical methods for investigating the size and power of hypothesis tests", *Manchester School* 66(1), 1–26.

Davidson, Russell and James G. MacKinnon, 1999, "Bootstrap testing in nonlinear models", *International Economic Review* 40(2), 487–508.

[17] Instead of the Wald statistic, given in table 3 for notational concision, Monfardini and Radice (2008, 273) study its square root which asymptotically follows the standard normal distribution.

Fiorio, Carlo V., Vassilis A. Hajivassiliou, Peter C.B. Phillips, 2010, "Bimodal $t$-ratios: the impact of thick tails on inference", *Econometrics Journal* 13(2), 271–289.

Ioannidis, John and Chris Doucouliagos, 2013, "What's to know about the credibility of empirical economics?", *Journal of Economic Surveys* 27(5), 997–1004.

Kiefer, Nicholas M., 1982, "Testing for dependence in multivariate probit models", *Biometrika* 69(1), 161–166.

Klein, Torsten L., 2014a, "The small multiple in econometrics – a redesign", mimeo, PAS Research Unit.

Klein, Torsten L., 2014b, "Communicating quantitative evidence: what determines the preference for tables over graphs?", mimeo, PAS Research Unit.

Monfardini, Chiara and Rosalba Radice, 2008, "Testing exogeneity in the bivariate probit model: a Monte Carlo study", *Oxford Bulletin of Economics and Statistics* 70(2), 271–282.

Sargent, Thomas J. and Christopher A. Sims, 1977, "Business cycle modelling without pretending to have too much a priori economic theory", in *New methods in business cycle research: proceedings from a conference*, Federal Reserve Bank of Minneapolis, 45–109.

Sims, Christopher A., 1982,"Policy analysis with econometric models", *Brookings Papers on Economic Activity*, 1982(1), 107–152.

Tufte, Edward R., 1983, *The visual display of quantitative information*, Cheshire CN: Graphics Press.

Tufte, Edward R., 1990, *Envisioning information*, Cheshire CN: Graphics Press.

Wooldridge, Jeffrey M., 2010, *Econometric analysis of cross section and panel data*, 2nd edition, Cambridge MA: MIT Press.

Zamora Bonilla, Jesús, 2012, "The economics of scientific knowledge", in Uskali Mäki (ed.), *Philosophy of economics*, Amsterdam: North-Holland, 823–862.

Table 1: Empirical test size – rejection frequencies from an exemplary Monte Carlo simulation.

| DESIGN | | | TESTING PROCEDURE | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | $N$ | $\alpha$ | CM | LM1 | LM2 | LM3 | LM4 | LR | RHO |
| 1 | 500 | 10% | 0.1154 | 0.1172 | 0.0028 | 0.0038 | 0.0028 | 0.1022 | 0.1240 |
| | | 5% | 0.0632 | 0.0644 | 0.0004 | 0.0006 | 0.0002 | 0.0536 | 0.0740 |
| | | 1% | 0.0192 | 0.0192 | 0.0000 | 0.0000 | 0.0000 | 0.0148 | 0.0274 |
| | 1,000 | 10% | 0.1090 | 0.1100 | 0.0034 | 0.0044 | 0.0036 | 0.1054 | 0.1156 |
| | | 5% | 0.0586 | 0.0586 | 0.0004 | 0.0008 | 0.0006 | 0.0552 | 0.0654 |
| | | 1% | 0.0112 | 0.0114 | 0.0000 | 0.0000 | 0.0000 | 0.0102 | 0.0156 |
| 2 | 1,000 | 10% | 0.1224 | 0.1228 | 0.0000 | 0.0000 | 0.0000 | 0.1164 | 0.1848 |
| | | 5% | 0.0720 | 0.0724 | 0.0000 | 0.0000 | 0.0000 | 0.0632 | 0.1356 |
| | | 1% | 0.0178 | 0.0180 | 0.0000 | 0.0000 | 0.0000 | 0.0132 | 0.0784 |
| | 2,000 | 10% | 0.1152 | 0.1156 | 0.0000 | 0.0000 | 0.0000 | 0.1084 | 0.1444 |
| | | 5% | 0.0576 | 0.0576 | 0.0000 | 0.0000 | 0.0000 | 0.0548 | 0.0866 |
| | | 1% | 0.0132 | 0.0134 | 0.0000 | 0.0000 | 0.0000 | 0.0120 | 0.0340 |
| 3 | 1,000 | 10% | 0.1170 | 0.1174 | 0.0000 | 0.0004 | 0.0000 | 0.1038 | 0.1464 |
| | | 5% | 0.0596 | 0.0600 | 0.0000 | 0.0000 | 0.0000 | 0.0508 | 0.0946 |
| | | 1% | 0.0148 | 0.0150 | 0.0000 | 0.0000 | 0.0000 | 0.0112 | 0.0414 |
| | 2,000 | 10% | 0.1076 | 0.1078 | 0.0000 | 0.0000 | 0.0000 | 0.1004 | 0.1192 |
| | | 5% | 0.0548 | 0.0552 | 0.0000 | 0.0000 | 0.0000 | 0.0480 | 0.0700 |
| | | 1% | 0.0138 | 0.0138 | 0.0000 | 0.0000 | 0.0000 | 0.0104 | 0.0232 |

*Source*: Monfardini and Radice (2008, 277: table 1)
*Notes*: The table gives rejection frequencies of seven procedures assessing in a bivariate probit model the exogeneity of a dichotomous regressor: conditional moment test CM, Lagrange multiplier tests LM1, LM2, LM3 and LM4, likelihood ratio test LR, and Wald-type test RHO. The frequencies are obtained from simulating data generating processes DGP1, DGP2 and DGP3 to create samples of $N$ observations, evaluating test statistics at asymptotic critical values that correspond to significance levels $\alpha$.

A description of how to calculate rejection frequencies offers appendix A.1. Details on the bivariate probit model, testing procedures as well as the Monte Carlo simulation presents appendix A.2.

Table 2: Methods of display when communicating quantitative information – relative frequencies [%] from a small sample of journal publications.

| DISPLAY METHOD | 1989–1993 | 1994–1998 | 1999–2003 | 2004–2008 | 2009–2013 |
|---|---|---|---|---|---|
| Sentence only | 9.52 | 10.00 | 3.45 | 2.70 | 10.13 |
| Table | 68.25 | 55.00 | 55.17 | 56.76 | 53.16 |
| Graph | 7.94 | 5.00 | 12.64 | 16.22 | 15.19 |
| Graph & table | 14.29 | 30.00 | 28.74 | 24.32 | 21.52 |
| Sum | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Observations | 63 | 80 | 87 | 74 | 79 |

*Source*: various issues of *Econometrica*, *The Journal of Economic Studies*, *The Journal of Economics and Statistics*; own calculations.
*Notes*: the sample comprises papers employing Monte Carlo simulations which were published during the respective five-year intervals in the journals mentioned before. The papers are grouped according to the method of communicating simulation results: those which discuss them in the body of text only, those which additionally use either tabular or graphical displays, and those which use both methods.

Table 3: Testing exogeneity ($\mathrm{H}_0$: $\rho = 0$) – procedures evaluated by Monte Carlo simulation and their quadratic form statistics.

| PROCEDURE | TEST STATISTIC | |
|---|---|---|
| *Distance* (LR) | $2[\sum_{i=1}^{N} \ell(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) - \sum_{i=1}^{N} \ell(\boldsymbol{w}_i, \tilde{\boldsymbol{\theta}})]$ | |
| *Single model* | $\boldsymbol{R}(\boldsymbol{\theta}_N)'\{\mathrm{Avar}_N[\boldsymbol{R}(\boldsymbol{\theta}_N)]\}^{-1}\boldsymbol{R}(\boldsymbol{\theta}_N)$ | |
| | $\boldsymbol{R}(\boldsymbol{\theta}_N)$ | $\mathrm{Avar}_N[\boldsymbol{R}(\boldsymbol{\theta}_N)]$ |
| Wald (RHO²) | $\boldsymbol{r}'\hat{\boldsymbol{\theta}} = [\mathbf{0}'_{K-1}\ 1]'\hat{\boldsymbol{\theta}}$ | $\boldsymbol{r}'(-\boldsymbol{\mathcal{H}}_N)^{-1}\boldsymbol{r}$,<br>$\boldsymbol{\mathcal{H}}_N = \widehat{\boldsymbol{\mathcal{H}}} = \sum_i \frac{\partial^2 \ell_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big\vert_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ |
| Score | $\tilde{\boldsymbol{s}} = [\mathbf{0}'_{K-1}\ \tilde{s}_\rho]'$ | $-\boldsymbol{\mathcal{H}}_N$ |
| LM1 | | $\widetilde{\boldsymbol{\mathcal{H}}}1 = -\tilde{\boldsymbol{S}}'\tilde{\boldsymbol{S}} = -\sum_i \frac{\partial \ell_i}{\partial\boldsymbol{\theta}} \frac{\partial \ell_i}{\partial\boldsymbol{\theta}'}\Big\vert_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ |
| Kiefer (1982) | | $-\operatorname{plim} N^{-1}\boldsymbol{\mathcal{H}}\Big\vert_{\mathrm{H}_0} = -\begin{bmatrix} \overline{\boldsymbol{\mathcal{H}}}_{\beta\beta} & \mathbf{0}_{K-1} \\ \mathbf{0}'_{K-1} & \overline{\mathcal{H}}_{\rho\rho} \end{bmatrix}$ |
| LM2 | | $\widetilde{\widetilde{\overline{\mathcal{H}}}}2_{\rho\rho} = -\tilde{\boldsymbol{S}}'_\rho\tilde{\boldsymbol{S}}_\rho = -\sum_i \frac{\partial \ell_i}{\partial\rho} \frac{\partial \ell_i}{\partial\rho}\Big\vert_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ |
| LM3 | | $\widetilde{\widetilde{\overline{\mathcal{H}}}}3_{\rho\rho} = -\sum_i \frac{\phi_{1i}^2\phi_{2i}^2}{\Phi_{1i}\Phi_{2i}(1-\Phi_1)(1-\Phi_{2i})}\Big\vert_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}$ |
| LM4 | | $\widetilde{\widetilde{\overline{\mathcal{H}}}}4_{\rho\rho} = \sum_i \frac{\partial^2 \ell_i}{\partial\rho\partial\rho}\Big\vert_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ |
| CM | $\mathbf{1}'_N\tilde{\boldsymbol{m}} = \sum_i \tilde{u}_{1i}\tilde{u}_{2i}$ | $N\tilde{\boldsymbol{m}}'[\boldsymbol{I}_{K-1} - \tilde{\boldsymbol{S}}_{\boldsymbol{\beta\beta}}(\tilde{\boldsymbol{S}}'_{\boldsymbol{\beta\beta}}\tilde{\boldsymbol{S}}_{\boldsymbol{\beta\beta}})^{-1}\tilde{\boldsymbol{S}}'_{\boldsymbol{\beta\beta}}]\tilde{\boldsymbol{m}}$ |

*Source*: adapted from information given in Monfardini and Radice (2008).
*Notes*: All procedures are based on the model in equations (3) to (5) and a sample of observations $\boldsymbol{w}_i = [y_{1i}\ y_{2i}\ \boldsymbol{x}'_{1i}\ \boldsymbol{z}'_{2i}]'$, $i = 1, 2, \ldots, N$. To estimate the $(K-1)$ location parameters and correlation coefficient, $\boldsymbol{\theta} = [\boldsymbol{\beta}'_1\ \boldsymbol{\beta}'_2\ \rho]' = [\boldsymbol{\beta}'\ \rho]'$, the method of maximum likelihood is used: $\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell = \sum_{i=1}^{N} \ell_i = \sum_{i=1}^{N} \ell(\boldsymbol{w}_i, \boldsymbol{\theta})$, where $\ell$ denotes the log likelihood function. Its score is given by $\boldsymbol{s} = \sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \boldsymbol{\theta}) = \sum_{i=1}^{N} \partial\ell_i/\partial\boldsymbol{\theta}$; scores for individual observations may be stacked in $(N \times K)$ matrix $\boldsymbol{S} = [\boldsymbol{s}(\boldsymbol{w}_1, \boldsymbol{\theta})\ \boldsymbol{s}(\boldsymbol{w}_2, \boldsymbol{\theta})\ \ldots\ \boldsymbol{s}(\boldsymbol{w}_N, \boldsymbol{\theta})]'$. The Hessian contains the second-order derivatives, $\boldsymbol{\mathcal{H}} = \sum_{i=1}^{N} \partial^2\ell_i/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. Both score and Hessian are partitioned in similar fashion to the parameter vector, leading to $\boldsymbol{s} = [\boldsymbol{s}'_{\boldsymbol{\beta}}\ s_\rho]'$, $\boldsymbol{S} = [\boldsymbol{S}_{\boldsymbol{\beta}}\ \boldsymbol{S}_\rho]$, and sub-matrices of the form $\boldsymbol{\mathcal{H}}_{\boldsymbol{\theta}1\boldsymbol{\theta}2} = \sum_{i=1}^{N} \partial^2\ell_i/\partial\boldsymbol{\theta}1\partial\boldsymbol{\theta}2'$ where $\boldsymbol{\theta}1, \boldsymbol{\theta}2 \in \{\boldsymbol{\beta}, \rho\}$. If $\rho = 0$ the model can be separated into two univariate probits, $\ell|_{\mathrm{H}_0} = \sum_{i=1}^{N} \ell_1(y_{1i}|\boldsymbol{x}_{1i}; \boldsymbol{\beta}_1) + \sum_{i=1}^{N} \ell_2(y_{2i}|y_{1i}, \boldsymbol{z}_{2i}; \boldsymbol{\beta}_2)$.

Under regularity conditions $\boldsymbol{\theta}_N$, the parameter estimator based on sample information, asymptotically follows a normal distribution, $\boldsymbol{\theta}_N \overset{a}{\sim} \mathrm{N}(\boldsymbol{\theta}_o, [\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_o)]^{-1})$, with mean equal to the true parameter vector $\boldsymbol{\theta}_o$ and variance equal to the inverse of the information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_o) = -\mathrm{E}[\partial^2\ell/\partial\boldsymbol{\theta}_o\partial\boldsymbol{\theta}'_o]$, cf. Wooldridge (2010, 476–479). Parameters and functions thereof obtained from estimating the unrestricted model are marked with a hat ($\widehat{\ }$), those estimated under $\mathrm{H}_0$ with a tilde ($\widetilde{\ }$). Generalized residuals of the restricted model are defined by $\tilde{u}_{ji} = \phi_{ji}(y_{ji} - \Phi_{ji})/[\Phi_{ji}(1 - \Phi_{ji})]|_{\mathrm{H}_0}$ for $j = 1, 2$, where $\phi_{ji} = \phi(\boldsymbol{\beta}'_j\boldsymbol{x}_{ji})$ and $\Phi_{ji} = \Phi(\boldsymbol{\beta}'_j\boldsymbol{x}_{ji})$ denote the standard normal probability distribution function and cumulative distribution function respectively.

Table 4: Data generating process – parameter sets, regressor and error stochastics.

| | $y_1^* = \beta_{10} + \beta_{11}x + \beta_{12}z + u_1$ | | | $y_2^* = \delta_{10}y_1 + \delta_{11}y_1z + \delta_{20} + \delta_{21}z + u_2$ | | | |
|---|---|---|---|---|---|---|---|
| DGP | $\beta_{10}$ | $\beta_{11}$ | $\beta_{12}$ | $\delta_{10}$ | $\delta_{11}$ | $\delta_{20}$ | $\delta_{21}$ |
| 1 | 0.5 | 1 | 1.5 | 1 | 1 | −0.5 | 0.5 |
| 2 | −1.5 | 0.5 | 0.5 | 1.5 | −1 | −0.5 | −1 |
| 3 | −1.75 | 0.7 | 0.4 | −0.7 | −0.6 | 1.9 | −1 |

*Source*: Monfardini and Radice (2008, 275)

*Notes*: The two latent variables $y_1^*$ and $y_2^*$ form a recursive probit model as suggested in appendix A.2; observations of the related dichotomous variables $y_1$ and $y_2$ are produced according to equation (5). Random variables $x$ and $z$ follow a bivariate standard normal distribution with correlation coefficient 0.5. Error terms $u_1$ and $u_2$ follow a bivariate standard normal distribution with correlation coefficient $\rho$.

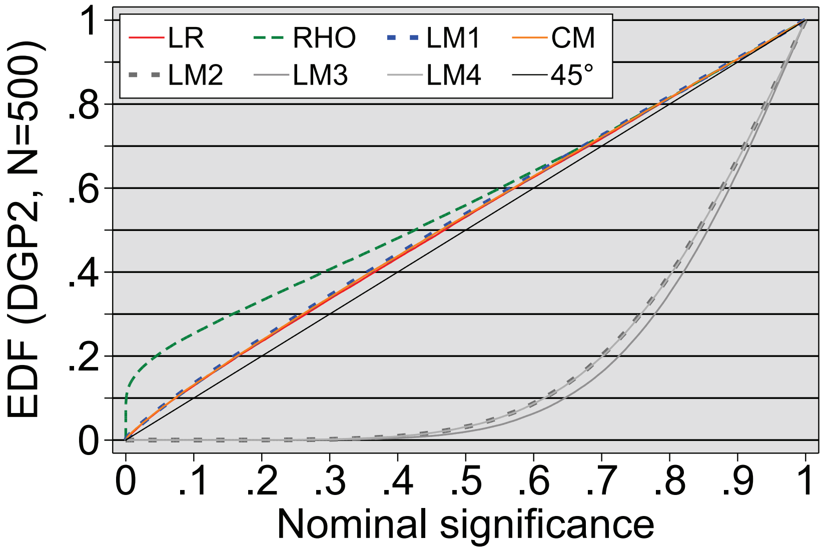

Figure 1: Empirical test size – *P* value plots.

*Notes*: The figure displays the *P* values' empirical distribution function (EDF) of the seven exogeneity tests for data generating process DGP2 and samples of 500 observations, cf. the notes to table 1.

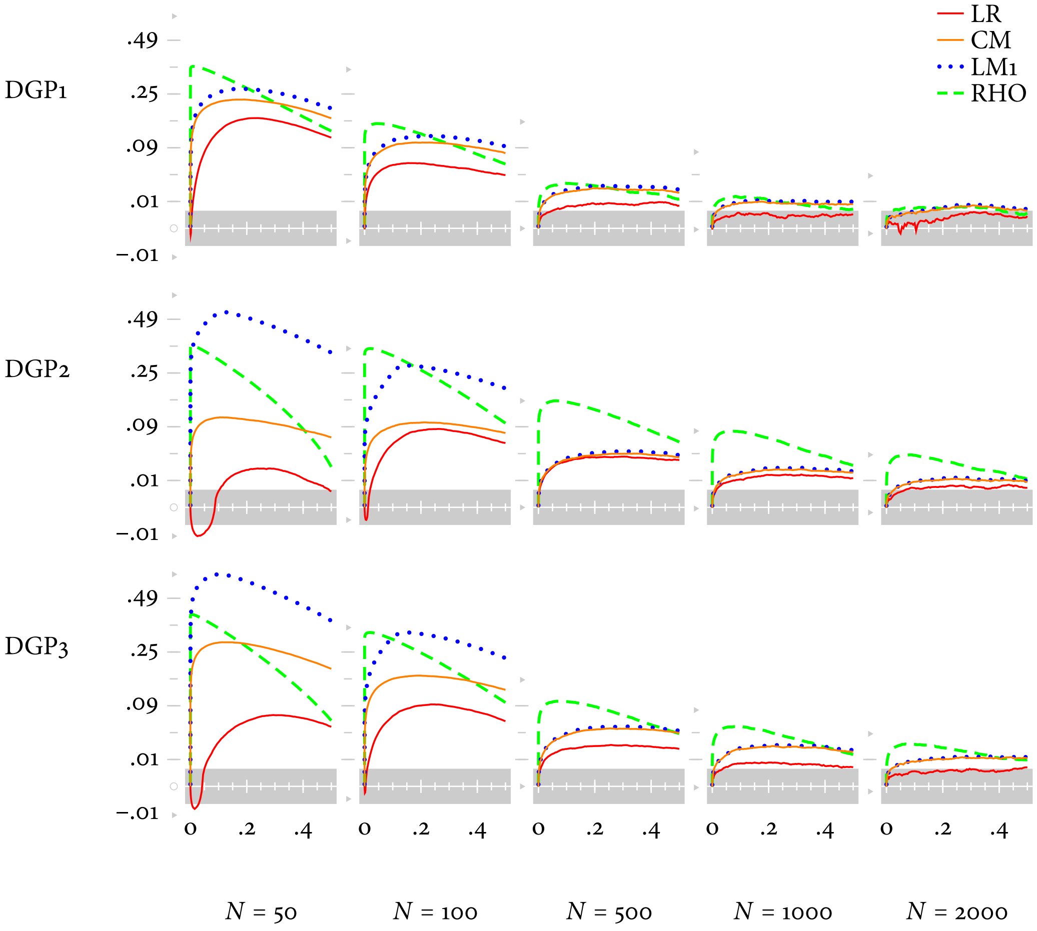

Figure 2: Empirical test size of selected testing procedures – *P* value discrepancy plots in a small multiple.

*Source*: Klein (2014a)

*Notes*: The figure compiles results from replicating the Monte Carlo simulation that feeds table 1, and places them within the framework introduced by Davidson and MacKinnon (1998). It displays for each combination of data generating process and sample length, over the interval (0, 0.5] of nominal significance levels, *P* value discrepancies. They are calculated taking the difference between the *P* values' empirical distribution function $\hat{F}(x_i)$ and nodes $x_i$ inserted for evaluation. Each ordinate is scaled non-linearly according to the power transformation $g\colon y \mapsto \operatorname{sgn}(y) \cdot |y|^{0.5}$. The shaded area enveloping abscissae indicates the non-rejection region of the Kolmogorov-Smirnov test at 0.05 critical values. A description of the graphical objects and their application to the Monte Carlo study can be found in appendix A.1 and the companion paper Klein (2014a).