

A Note on Testing for Agglomeration and Dispersion*

Marc Rysman
Boston University

Shane Greenstein
Kellogg School of Management, Northwestern University

June 11, 2003

Abstract

This article proposes a statistical test for determining whether agents in discrete locations are more agglomerated or disperse than predicted by independent random choice. The article provides comparisons to similar tests in the literature.

1 Introduction

Economic theory is often concerned with forces that lead to agglomeration and dispersion. For instance, network effects or socially transmitted neighborhood effects lead interacting agents to *agglomerate* by making similar decisions. Similarly, standard models of market competition predict that firms differentiate from one another such that they would *disperse* in product space. This paper presents a simple methodology for testing for agglomeration and dispersion, even when there are a small number of decision-making agents. We test whether an allocation of agents is more agglomerated or more dispersed than the agents would be if they made independent random choices. Our test is based on the

*We wish to thank the seminar audience and the Antitrust Division of the Department of Justice, Victor Aguirregabiria, Kim Sau Chung, Kevin Lang and Chuck Romeo for advice and encouragement. Martino De Stefano provided excellent research assistance. Rysman was supported by NSF Grant SES-0112527.

likelihood function from a multinomial distribution, and we term it the multinomial test for agglomeration and dispersion (MTAD).

The most similar existing methodology is the dartboard index of Ellison and Glaeser (1997).¹ Ellison and Glaeser (1997) use the index to measure whether manufacturing plants are more concentrated across states than would be predicted by state populations and the size distribution of plants. Whereas the dartboard index is based on a method of moments of approach, MTAD is based on likelihood. Also, their test is derived from a model of agent choices in the context of agglomeration. MTAD has a non-parametric motivation that is easier to apply to cases of dispersion. Despite these differences, we show in Monte Carlo experiments that the two tests perform remarkably similarly although we highlight some circumstances in which they find contradictory results. Both MTAD and the dartboard index treat locations as discrete, although an alternative literature incorporates geographic distance as a continuous measure into indices of concentration. See for example, see Busch and Reinhardt (1999) and Marcon and Puech (2003).

All of these approaches compare the distribution of outcomes to the expected distribution under independent random choice and in this sense are similar to the Pearson Chi-squared test of goodness of fit. However, the Pearson test does not indicate whether the rejection of independent random choice occurred

¹We take the popularity of the dartboard index as motivation for our paper and further work in this area. For example, all of the following papers employ the dartboard index. Gautier and Teulings (2002) use the dartboard index to study labor market density. Klimek and Merrell (1999) analyze concentration in the retail sector. Head, Mayer and Ries (2002) test for a home market bias in international trade. Karlan (2003) measures social capital in rural Peru. Rosenthal and Strange (2001) use the dartboard index derived for U.S. manufacturing industry as a dependent variable in regression analysis. Holmes (1999) uses the dartboard index in an intermediate step for studying vertical relationships between firms. Ellison (2002) measures concentration in the journal publication process. Black and Henderson (1999) use the dartboard index to measure the concentration of population and industry in cities. This list is not exhaustive.

because of agglomeration or dispersion. While a weakness for our purposes, this feature gives the Pearson test the power to reject independent random choice if the data comes from some mix of agglomeration and dispersion, in which case MTAD would fail to reject. We provide further analysis below.

We present MTAD and discuss its asymptotic properties in Section 2. In Section 3, we present an example based on the work of Augereau, Greenstein and Rysman (2003), henceforth AGR, in the context of firms adopting competing standards that exhibit network effects. In Section 4, we present a comparison of MTAD and the dartboard index and in Section 5, we compare our method to the Pearson Chi-squared test of goodness of fit.

2 A Multinomial Test of Agglomeration and Dispersion

We begin by pointing out that combinatorics statistics can be a useful way to capture agglomeration and dispersion. For instance, suppose we observe 4 firms choose between 2 locations. A *dispersed* arrangement would have 2 firms in each location whereas an agglomerated arrangement would have all 4 firms in either location. Consider the combinatoric expression $\binom{4}{x}$ where x is the number of firms in the first location. Assuming each firm selected randomly between locations with equal probability, this expression would have an expected value of 4.375. The dispersed arrangement generates $\binom{4}{2} = 6$ whereas the agglomerated arrangements generate $\binom{4}{0} = \binom{4}{4} = 1$. That is, whether the combinatoric statistic is above or below its expectation captures the notion of whether the data is more or less agglomerated than would be expected under independent random choice. Now, we construct a statistic based on likelihood that exploits this intuition.

Suppose we observe M markets each populated by n_m agents $m = 1, \dots, M$,

where n_m is bounded by $\underline{n} > 0$ and $\bar{n} < \infty$. The variable n_m is distributed according to the discrete distribution $f(n_m)$. The agents choose between C options, available in each market. The unconditional probability of observing option c is p_c , $c = 1, \dots, C$. The observed number of agents choosing option c in market m is x_m^c . Let \mathbf{x}_m be the vector of elements x_m^1, \dots, x_m^C and \mathbf{p} be the vector of probabilities p_1, \dots, p_C . If the agents make choices independently, the likelihood of observing the outcome x_m^1, \dots, x_m^C in market m is the multinomial pdf:

$$L(\mathbf{x}_m, n_m, \mathbf{p}) = \binom{n_m}{x_m^1, \dots, x_m^C} p_1^{x_m^1} \dots p_C^{x_m^C}$$

Letting \mathbf{X} be the $M \times C$ matrix of choices for all of the markets and \mathbf{n} be the $M \times 1$ vector of the number of agents in each market, we can define the average log-likelihood of the data set to be:

$$l(\mathbf{X}, \mathbf{n}, \mathbf{p}) = \frac{1}{M} \sum_{m=1}^M \ln \left(\binom{n_m}{x_m^1, \dots, x_m^C} \right) + x_m^1 \ln(p_1) + \dots + x_m^C \ln(p_C)$$

Consider the likelihood value if the data were actually generated by independent random choice. Let the random variable $l(f, \mathbf{p})$ be distributed according to the distribution $l(\mathbf{X}, \mathbf{n}, \mathbf{p})$ if \mathbf{X} was actually drawn from a multinomial distribution and n_m was drawn from f . Then we have that:

$$E[l(f, \mathbf{p})] = \sum_{n_m=\underline{n}}^{\bar{n}} \sum_{\mathbf{z} \in \Xi(n_m)} \left(\ln \left(\binom{n_m}{z^1, \dots, z^C} \right) + z^1 \ln(p_1) + \dots + z^C \ln(p_C) \right) L(\mathbf{z}, n_m, \mathbf{p}) f(n_m)$$

where $\Xi(n_m)$ is the set of all possible configurations of n_m . We are interested in:

$$t(\mathbf{X}, \mathbf{n}, \mathbf{p}) = l(\mathbf{X}, \mathbf{n}, \mathbf{p}) - E[l(f, \mathbf{p})]$$

Now we show that $t(\mathbf{X}, \mathbf{n}, \mathbf{p})$ is distributed asymptotically normal. Let $\sigma^2 = V[l(f, \mathbf{p})]$. Then we have the following theorem:

Theorem 1 *Under the null hypothesis that \mathbf{X} is distributed according to the multinomial distribution with probabilities \mathbf{p} , we have:*

$$\lim_{M \rightarrow \infty} \sqrt{M} t(\mathbf{X}, \mathbf{n}, \mathbf{p}) \sim \mathbf{N}(0, \sigma^2)$$

Proof. Under the null, $\text{plim}_{M \rightarrow \infty} (l(\mathbf{X}, \mathbf{n}, \mathbf{p}) - E[l(f, \mathbf{p})]) = 0$ by the law of large numbers. Then, we have that $l(\mathbf{x}_m, n_m, \mathbf{p}) = \ln(L(\mathbf{x}_m, n_m, \mathbf{p}))$ is a random sample with mean $E[l(f, \mathbf{p})]$ and variance σ^2 . So the result follows by the Central Limit Theorem. ■

It is possible to derive an analytic expression for σ . However, we expect that in many circumstances it will be easier to compute σ (and possibly $E[l(f, \mathbf{p})]$) through Monte Carlo procedures.

The above derivations assume that we know \mathbf{p} exactly, as if we know \mathbf{p} for the population. A potential problem is that \mathbf{p} will often be estimated. Accounting for estimation error in \mathbf{p} can introduce serious problems into the asymptotics. However, we do not expect this issue to be an important empirical problem. For instance, in AGR, \mathbf{p} is the mean of a dummy variable computed across more than 2000 observations. In Ellison and Glaeser (1997), \mathbf{p} is the 48×1 vector of population shares of states, which is computed over more than 200 million people. Estimation error in \mathbf{p} would seem to be so small in these circumstances that it can be safely ignored.² Similarly, we take f to be the empirical distribution of n_m . In practice, it may be estimated with error although we do not address this issue here.

In this case, the test statistic $\sqrt{M}t(\mathbf{X}, \mathbf{n}, \mathbf{p})/\sigma$ is distributed normally, and provides a test for the null hypothesis that the data was generated from a multinomial model. Rejection can be taken as a rejection of the hypothesis of independent random choice by the agents. Furthermore, whether $t(\mathbf{X}, \mathbf{n}, \mathbf{p})$ is positive or negative is informative as to whether rejection was a result of data that was agglomerated or dispersed. In $t(\mathbf{X}, \mathbf{n}, \mathbf{p})$, we expect the terms associated with the probabilities \mathbf{p} to drop out as M rises, regardless of whether the data was generated by independent random choice or not (assuming that \mathbf{p} is chosen to match the observed choices \mathbf{X}). Asymptotically, differences between $l(\mathbf{X}, \mathbf{n}, \mathbf{p})$ and $E[l(f, \mathbf{p})]$ are driven by the combinatoric terms. If choices are

²If this issue is a concern, it is straightforward to compute $V[l(\mathbf{n}, \hat{\mathbf{p}})]$ via the bootstrap.

agglomerated, then $\binom{n_m}{x_m^1, \dots, x_m^C}$ will be lower than its expected value. If choices are dispersed, then $\binom{n_m}{x_m^1, \dots, x_m^C}$ will be higher than expected. This feature allows us to test between agglomeration and dispersion.

3 An Example

AGR analyze the adoption of 56K modems by Internet Service Providers in the United States upon introduction in 1997. At this point, there were two competing incompatible standards. AGR point out that within each market, network effects would drive ISP's to agglomerate on a single standard (possibly different across markets) whereas incentives to differentiate would cause ISP's to disperse across the two standards. AGR use local calling plans to identify separate markets, as consumers almost never sign with an ISP that requires a long distance call. AGR are interested in the role of these incentives in the ultimate failure of the market to coordinate on a single standard.

AGR observe 2,233 ISP's in 2,298 markets. We ignore the fact that some ISP's appear in multiple markets.³ The median number of ISP's in a market is 3 although the mean is 15.06. We focus on October 1997, when 389 ISP's had adopted the standard called X2, 523 had adopted Flex and 185 had adopted both standards. Let $\{x_m^0, x_m^A, x_m^B, x_m^{AB}\}$ be the number of ISP's in market m (a local calling area) that do not adopt, that adopt X2, that adopt Flex and that adopt both respectively. We start with a graphical presentation of the data. We calculate the number of adopters of only X2 as a percentage of the number of ISP's that adopt only one standard, ignoring markets with only one such firm. That is, we compute $x_m^A / (x_m^A + x_m^B)$ in each market where $x_m^A + x_m^B > 1$. For now, we ignore firms that adopt both or neither standard (x_m^0 and x_m^{AB}). The national adoption rate computed in this way is 58%.

³AGR address this issue in detail. See that paper for more on industry details and construction of the data.

Agglomeration would imply that within each market, firms are on one standard or the other but there are many markets of both types. Dispersion would imply that within each market, adoption is close to the national number. Figure 1 provides a histogram of the adoption percentage across markets. The dark bars represent the observed data. Note that most markets are in the 0.5, 0.6 and 0.7 bins, with relatively few on the ends. For comparison purposes, we show what the histogram would have looked like if the firms had really made independent random choices, that is if $x_m^A + x_m^B$ firm in each market chose independently with probability $p_1=0.58$. The gray bars put less weight in the middle of the graph and more weight on the ends. This result suggests that the data is more disperse then would be expected by independent random choice.⁴

MTAD provides a method for rejecting independent random choice in a statistical sense. For these purposes, $M = 1,595$ and \mathbf{X} is the $M \times 2$ matrix identifying the number of adoptions of each standard in each market. Results appear in Table 1. We have computed $l(\mathbf{X}, \mathbf{n}, \mathbf{p})$ for the observed data, as well as $E[l(f, \mathbf{p})]$ and a standard deviation from $V[l(f, \mathbf{p})]$. For comparison, we also compute the minimum and maximum that $l(\mathbf{X}, \mathbf{n}, \mathbf{p})$ could take on. The maximum is $l(\overline{\mathbf{X}}, \mathbf{n}, \mathbf{p})$, where $\overline{x}_m^A = p_1/n_m$, or the closest integer, and $\overline{x}_m^B = n_m - \overline{x}_m^A$. Similarly, the minimum is $l(\underline{\mathbf{X}}, \mathbf{n}, \mathbf{p})$ where $\underline{x}_m^A = n_m$ with probability p_1 and $\underline{x}_m^B = 0$ with probability p_2 . Again, $\underline{x}_m^B = n_m - \underline{x}_m^A$.

Results appear in row 1. One would have expected a likelihood value of $E[l(f, \mathbf{p})] = -1.631$, and allowed for a standard deviation of 0.016.⁵ However, the observed likelihood value of -1.472 is well out of the 95% confidence range. Not only does this result reject independent random choice, it tells us that the reason is because the ISP's are spread more evenly across the two standards

⁴The "bumpiness" in Figure 1 is due to the distribution of the number of firms in a market, which is also uneven.

⁵The standard errors were generated by drawing new matrices \mathbf{X}^s 500 times, and computing $l(\mathbf{X}^s, \mathbf{n}, \mathbf{p})$ for each sample, and then computing the standard deviation based on this sample. Note that \mathbf{p} is not re-estimated for each draw s .

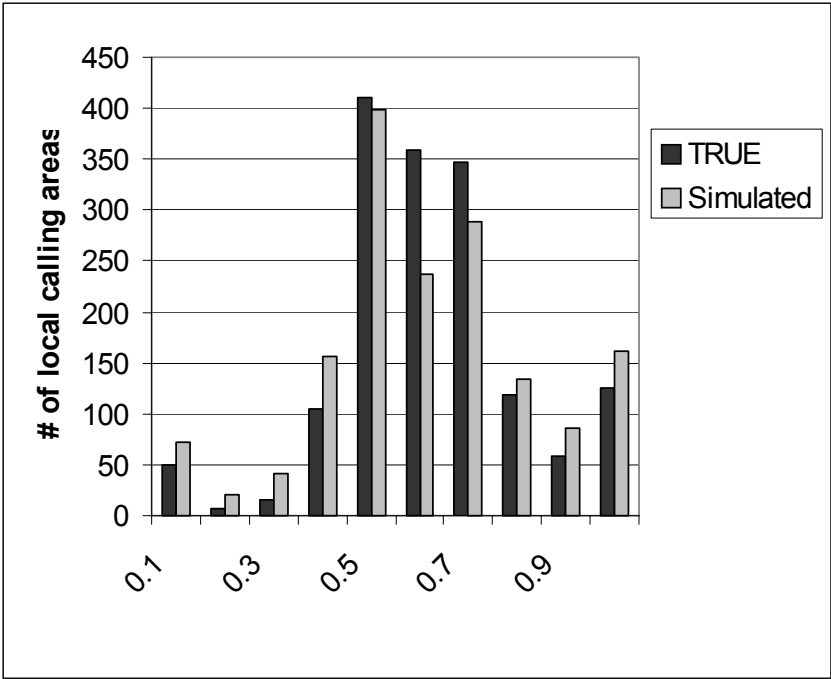


Figure 1: Percentage of ISP's adopting X2.

Table 1: Binomial Test for Differentiation

Choices	Obs.	Coord	Expected	Std Dev	True	Diff
X2, Flex	1,595	-5.785	-1.631	0.016	-1.460	-1.220
X2, Flex, Both	1,698	-11.366	-3.130	0.020	-2.959	-2.367
None, X2, Flex, Both	2,200	-23.496	-4.429	0.022	-4.270	-3.379

then independent random choice would allow. For comparison purposes, if the ISP's had been perfectly differentiated, the likelihood value would be -1.223, whereas perfect coordination generates -6.01. The second row shows results when allowing firms to adopt both standards is a third option (adding x_m^{AB} to the analysis), and row 4 includes firms that do not adopt (x_m^0 as a fourth choice). For each case, we include markets for with at least two firms choosing. In all cases, we can reject that the data was generated from independent random choice and we can always do so in favor of dispersion.

4 A Comparison to the Dartboard Index

In this section, we compare MTAD to the dartboard index in their ability to correctly determine when data comes from a distribution that is agglomerated or dispersed. We show that the two tests perform remarkably similarly, although we find some circumstances where one is preferred to the other.

Suppose we observe M markets made up of n firms deciding to between 2 choices. We assume firms within each market choose sequentially according to a pre-specified ordering. Firm j draws profit $\pi_{jc} = \beta n_{jc} + \varepsilon_{jc}$ from option $c = 1, 2$, where n_{jc} is the number of firms that have chosen c previous to j and ε_{jc} is distributed standard normal.⁶ The parameter β determines the extent of agglomeration or dispersion. For each set of parameters M , n and β , we

⁶For this exercise, we assume firms choose based on previous choosers and do not attempt to predict choices of firms that follow.

Table 2: Comparison of MTAD and the dartboard index (DI)

Row	M	n	β	Percent Rejection	
				MTAD	DI
1	10	4	-0.5	13.5%	13.5%
2	10	4	-1	50.9	50.9
3	10	4	0.5	71.1	71.1
4	10	4	1	91.1	91.1
5	10	4	0	0.5	0.5
6	10	10	0	0.05	0.19
7	10	10	-0.5	26.1	46.4
8	10	10	0.15	64.6	64.3
9	1	100	0.1	78.2	81.4

draw 10,000 samples and compute both tests for each sample. We report the parameters and the percent of times that independent random choice is rejected by each test.

Results appear in Table 2. We see that when there are 4 firms in the market and 10 markets, results coincide exactly regardless of β . When there are 10 firms in a market, results differ somewhat. MTAD performs slightly better in the case where $\beta = 0$, in which case neither test should reject. MTAD rejects 0.05% of the time where the dartboard index rejects 0.19% of the time. On the other hand, the dartboard index performs significantly better in the case of dispersion ($\beta = 0.5$). This is surprising given that the dartboard index was designed with agglomeration in mind. In the case of agglomeration ($\beta = 0.15$), MTAD performs better, but only slightly. The last row simulates a case closer to the type of data used in Ellison and Glaeser (1997). There is only one market (an industry), 100 firms and 48 possible choices (states), and some agglomeration. Both tests perform well, although the dartboard index performs slightly better.

In order to learn more about the distinctions, we consider row 8 ($\beta = 0.15$) and look at the observations at which the two tests differ. In Figure 2, we graph the distributions of choices from the whole sample of 10,000 observations, and separately the observations at which one test rejects but the other does not. Note that the distribution associated with MTAD has a high proportion

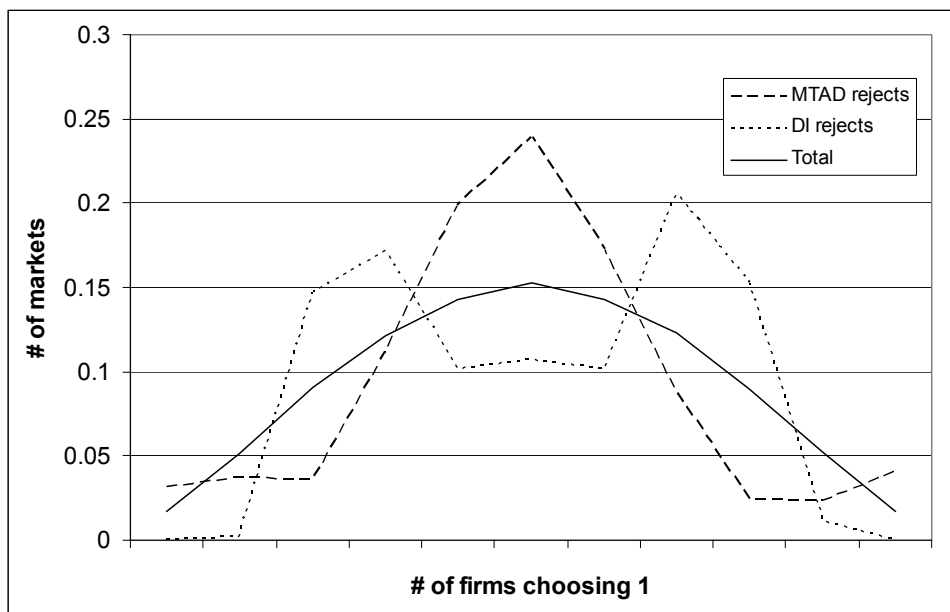


Figure 2: Observations at which one test rejects and the other does not.

of perfectly disperse markets (5,5) and perfectly agglomerated (10,0 or 0,10), whereas the dartboard index has the inverse. MTAD can find agglomeration even where a perfectly disperse market is observed, as long as there is at least one perfectly agglomerated market. Conversely, the dartboard index can find agglomeration without observing a perfectly agglomerated market, but cannot if there is a perfectly disperse market.

5 A Comparison to Pearson's χ^2 test

Another statistic appropriate for testing whether the two distributions in Figure 1 are different is Pearson's χ^2 test.⁷ For instance, suppose we observe M markets with 4 firms in each market choosing between one of two locations (labelled 1 and 2) in their markets. Let q_j be the observed probability of observing a market

⁷A useful reference is Gouriéroux and Monfort (1995), Section 17.5.

with j firms in location 1 and let \hat{q}_j be the probability if firms made independent random choices. Under the null hypothesis of independent random choice, the statistic

$$\xi = M \sum_{j=0}^4 \frac{(q_j - \hat{q}_j)^2}{\hat{q}_j}$$

is distributed χ^2 with 3 degrees of freedom.⁸ A weakness of this statistic is that it does not indicate whether rejection of the null hypothesis occurs because firms are agglomerated or disperse. However, a benefit is that the test can reject when there is some alternative reason to do so.

Consider the following example, designed to show a circumstance in which the Pearson test would reject but MTAD (and the dartboard index) would fail to do so. Suppose we observe M markets in which the probability of observing a given number of firms in location 1 is as described by the “true” bars in Figure 3.⁹ This implies an unconditional probability of each firm choosing location 1 of 0.5. The “simulated” data is what would occur if firms chose randomly between locations with equal probability. This data is more dispersed than expected from the perspective of outcome 2 but is more agglomerated than expected from the perspective of outcomes 0 and 4. We have constructed this data set such that the “true” data generates $l(\mathbf{X}, \mathbf{n}, \mathbf{p}) = E[l(f, \mathbf{p})]$, so MTAD fails to reject independent random choice. On the other hand, the Pearson χ^2 test equals $M \cdot 1.0015$ and rejects at 95% confidence for $M > 8$.

6 Conclusion

Numerous papers are interested in measuring agglomeration and dispersion, as exhibited by the popularity of the dartboard index of Ellison and Glaeser

⁸Writing down the equivalent statistic to apply to Figure 1 is more involved because the markets with different number of firms have different distributions over possible outcomes.

⁹For this example, observed probabilities of outcomes $\{0, 1, 2, 3, 4\}$ is $\{0.119, 0, 0.762, 0, 0.119\}$.

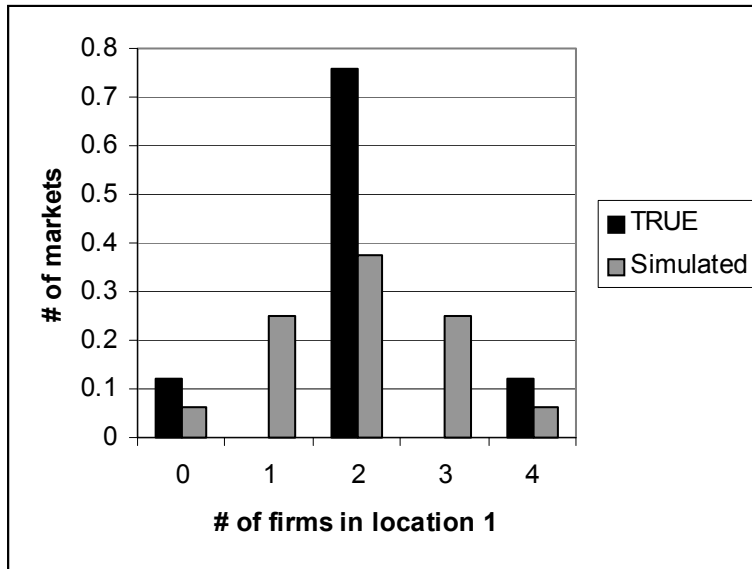


Figure 3: Hypothetical example exhibiting greater power of Pearson’s test

(1997). We present a statistic that provides a test of the hypothesis of independent random assignment based on the multinomial distribution, which we term the multinomial test of agglomeration and dispersion (MTAD). MTAD allows the researcher to test whether rejection took place because the data was agglomerated or dispersed. MTAD is easy to compute and interpret, and performs well in practice.

References

- [1] Augereau, A, Greenstein, S., & Rysman, M. (2003). Coordination vs. Differentiation in a Standards War: The Adoption of 56K Modems. Mimeo, Boston University.
- [2] Black, D., & Henderson, V. (1999). Spatial evolution of population and industry in the United States. *American Economic Review*, 89(2), 321-327.
- [3] Busch, M. L., & Reinhardt, E. (1999). Industrial location and protection: The political and economic geography of US nontariff barriers. *American Journal of Political Science*, 43(4), 1028-1050.

- [4] Head, K., Mayer, T., & Ries, J. (2002). Market Size and Agglomeration, Mimeo, University of British Columbia.
- [5] Ellison, G. (2002). The slowdown of the economics publishing process. *Journal of Political Economy*, 110(5), 947-993.
- [6] Ellison, G., & Glaeser, E. L. (1997). Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105(5), 889-927.
- [7] Gautier, P.A. & Teulings, C.N. (2002) An empirical index for labor market density. Mimeo, University of Tinbergen.
- [8] Gourieroux, C., & Monfort, A. (1995). *Statistics and econometric models* (Vol. 2). Cambridge [England] ; New York, NY, USA: Cambridge University Press.
- [9] Holmes, T. J. (1999). Localization of industry and vertical disintegration. *Review of Economics and Statistics*, 81(2), 314-325.
- [10] Karlan, D.S. (2003). Social Capital in Group Banking. Mimeo, Princeton University.
- [11] Klimek, S.D. & Merrell, D.R. (1999) Geographic Concentration in the U.S. Retail and Wholesale Sectors. Mimeo, Carnegie Mellon University.
- [12] Marcon, E., & Puech, F. (2003). Measures of Geographic Concentration of Industries: Improving Distance Based Methods. Mimeo, TEAM University of Paris.
- [13] Rosenthal, S.S., & Strange, W.C. (2001). The Determinants of Agglomeration, *Journal of Urban Economics*, 50, 191-229.