Gumbel Based P-Value Approximations for Spatial Scan Statistics

Allyson M. Abrams[§], Ken Kleinman, Martin Kulldorff

Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, USA

[§]Corresponding author

Email addresses:

Allyson Abrams: allyson_abrams@harvardpilgrim.org
Ken Kleinman: ken_kleinman@hms.harvard.edu
Martin Kulldorff: martin_kulldorff@hms.harvard.edu

Abstract

Background

The spatial scan statistic is commonly applied for the detection and evaluation of geographical disease clusters. Monte Carlo hypothesis testing is typically used to test whether the geographical clusters are statistically significant as there is no known way to calculate the null distribution analytically. In Monte Carlo hypothesis testing, simulated random data are generated multiple times under the null hypothesis, and the p-value is $r/(R+1)$, where R is the number of simulated random replicates of the data and r is the rank of the test statistic from the real data compared to the same test statistics calculated from each of the random data sets. A drawback to this powerful technique is that to obtain each additional digit of p-value precision, ten times as many simulated replicates are required, which can lead to excessive computing requirements when the data set is large.

Results

We propose a new method for obtaining more precise p-values with a given number of replicates. The collection of test statistics from the random replicate is used to estimate the true distribution of the test statistic under the null hypothesis, by fitting a continuous distribution to these observations. The choice of distribution is critical, and for the spatial scan statistic, the extreme value Gumbel distribution performs very well while the gamma, normal and lognormal distributions perform poorly. From the fitted Gumbel distribution, we show that it is possible to estimate the analytical p-value with great precision even when the test statistic is far out in the tail beyond any of the test statistics observed in the simulated replicates. In addition, Gumbel-based rejection probabilities have smaller variability than Monte Carlo-based rejection probabilities, suggesting that the proposed approach may result in greater power than the true Monte Carlo hypothesis test for a given number of replicates.

Conclusions

For large data sets, it is often advantageous to replace computer intensive Monte Carlo hypothesis testing with this new method of fitting a Gumbel distribution to random data sets generated under the null, in order to obtain much more precise p-values and slightly higher statistical power.

Background

Introduction

Geographic cluster detection and evaluation is important in disease surveillance. One frequently-used method for cluster detection is the spatial scan statistic [1-3]. This method has been used to study the geography of infectious diseases such as malaria [4], vector borne diseases such as West Nile Virus [5], many different forms of cancer [6-10], low birth weight [11], syndromic surveillance [12-16], and bovine spongiform encephalopathy [17], among many other diseases.

The spatial scan statistic is found by moving a scanning window across the geographical region of interest, generating a large collection of window locations and sizes that meet pre-defined criteria. For each location and size, the likelihood ratio is calculated, and the spatial scan statistic is the maximum of these likelihood ratios. The window location and size with the maximum likelihood ratio is the most likely cluster; that is, the cluster that is least likely to have occurred by chance [1, 2]. Except for the simplest scenarios, there is no known closed-form theoretical distribution for the spatial scan statistic. Therefore, a p-value for the spatial scan statistic is usually obtained using Monte Carlo hypothesis testing [18].

In Monte Carlo hypothesis testing, a large number of random replicates of the observed data are generated under the null hypothesis. Monte Carlo p-values are asymptotically equivalent to p-values from exact permutation tests as the number of random replicates increases, but the key property is that it has been shown that Monte Carlo hypothesis testing p-values maintain the correct alpha level, exactly, as long as the number of replicates plus one is a multiple of $1/\alpha$ [18]. Monte Carlo hypothesis testing can therefore be useful when theoretical distributions are unknown and the number of permutations prohibits a full enumeration. One major drawback to the approach is that small p-values can only be obtained through a very large number of Monte Carlo replicates, which may be computer intensive and time consuming. For the spatial scan statistic, Monte Carlo hypothesis testing requires the calculation of the likelihood ratio for each location and size of the scanning window, for each replicated data set. Thus, the approach can be computer intensive for very large data sets.

In disease surveillance, the spatial scan statistic is sometimes calculated on a daily basis, to continuously monitor a disease in near real-time [12, 19]. These clusters may then be reported to local, state, or federal public health officials for potential investigation. Using a conventional 0.05 α-level would on average result in one false rejection of the null hypothesis every 20 days. Because of limited resources, health officials are not able to investigate a lot of false alarms [12, 19]. To control the number of false rejections at a more tenable level, one might instead use an α-level of 1/365 or 1/3650, corresponding to one expected false positive every year or every ten years, respectively, for daily analyses. If multiple diseases are under surveillance, the desired α-level may be even smaller to adjust for the multiple testing inherent in the multiple diseases being evaluated. If Monte Carlo hypothesis testing is used, such α-levels may require hundreds of thousands of random replicates to be simulated each day. Due to the large number of computations associated with each replicate, the computational burden associated with small p-values can be large.

In this article, we propose a way to generate approximate p-values for the spatial scan statistic with fewer calculations. The approach we take is to find a distribution which closely approximates the distribution of the test statistics that were generated under the null hypothesis, which themselves reflect the distribution of the spatial scan statistic under the null. To do this we generate a relatively small number of random simulated replicates under the null hypothesis. We then use them to estimate parameters for a distribution with a well-characterized functional form. If this distribution fits the sample distribution well, we can use it as an estimate of the distribution of the spatial scan statistic under the null and use it to generate arbitrarily small p-

values. Because we are interested in small p-values, it is particularly important that the estimate is good in the tail of the distribution.

We note that although this paper is focused on the spatial scan statistic, the general methodology that we propose in this article can easily be applied to other test statistics that rely on Monte Carlo hypothesis testing.

Spatial scan statistics

The spatial scan statistic is used to determine whether there is any clustering of events on a map. Events may, for example, be cases of disease incidence, prevalence or mortality. We look at all unique subsets of events that lie within a collection of scanning windows to detect clusters. Although any shape scanning window may be used, we use circles throughout this paper.

Suppose that there are p geographical coordinates marked on a map, each representing a region, and consider all circles, $C_{i,r}$, where $i = 1,\ldots, p$ indicates the coordinates around which the circle is centered, and r indicates its radius, which ranges from 0 to some pre-specified maximum. Based on the observed and expected number of events in the circle, calculate the likelihood ratio for each distinct circle [1, 2]. The circle with the maximum likelihood ratio is the most likely cluster, that is, the cluster that is least likely to have occurred by chance. For computational simplicity, the logarithm of the likelihood ratio is typically used instead of the ratio itself, and the log likelihood ratio of this circle is defined as the *scan statistic*.

For this study, we used the SaTScan™ [20] statistical software program, which calculates the scan statistic and implements Monte Carlo hypothesis testing to calculate a p-value. SaTScan™ allows the user to vary many parameters including the maximum cluster size, the probability model, and the number of Monte Carlo replicates.

Monte Carlo hypothesis testing

Proposed by Dwass [18], Monte Carlo hypothesis testing is useful for determining how unusual an observed statistic is when the underlying distribution for the test statistic is unknown but when it is possible to generate data under the null hypothesis. To do this, one first calculates the test statistic from the real data. Then, a large number of random data sets are generated according to the null hypothesis, and the test statistic is calculated for each of these data sets. If one creates R random replicates of the data and r-1 of those replicates have a test statistic which is greater than or equal to the test statistic from the real data, so that r is the rank of the test statistics from the real data, then the p-value of the observed test statistic is $\frac{r}{1+R}$.

If the test statistic from the real data set is among the highest 5 percent from the random data sets, then we can reject the null hypothesis at the α=0.05 level of statistical significance. A nice feature of Monte Carlo hypothesis testing is that the correct 0.05 α-level is maintained exactly even when there are only, for example, 19, 99 or 999 random replicates, although fewer replicates means lower statistical power [21].

In many applications, it matters little whether a p-value is 0.01 or a lot less, as they all reject the null hypothesis, and 99 Monte Carlo replicates is then sufficient . However, in the context of hourly, daily or weekly analyses in real-time disease surveillance, the goal is to detect clusters of disease that are very unusual, and only the most unusual clusters will be investigated. P-values on the order of 0.0001 or even smaller may be required before an investigation is launched. These p-values can only be calculated with 9999 or more Monte Carlo replicates. The number of Monte Carlo replicates required is determined only by the desired precision of the p-value, and each additional decimal place requires 10 times the number of Monte Carlo replicates and hence about 10 times the computing time.

The Gumbel distribution

The spatial scan statistic described above is the maximum value taken over many circle locations and sizes, so the collection of these statistics generated from the Monte Carlo replicates is a distribution of maximum values. Since our technique involves finding a distribution which closely matches the distribution of the replicated statistics, it is natural to consider one of the extreme value distributions as one possible candidate to approximate the desired distribution. The Gumbel distribution is a distribution of extreme values, either maxima or minima. Here, we limit ourselves to distributions of maxima since the scan statistic is a maximum. The cdf for the Gumbel distribution of maxima is $G(x) = e^{-e^{-\frac{x-\mu}{\beta}}}$ and the pdf is $g(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} e^{-e^{-\frac{x-\mu}{\beta}}}$, where μ is a scale parameter and β is a shape parameter. Both μ and β can be estimated from a sample of observations using method of moments estimators as follows:

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi}$$

$$\hat{\mu} = \bar{X} - 0.5772\hat{\beta}$$

where $\bar{X}$ is the sample mean and s is the sample standard deviation [22, 23].

Methods

To evaluate whether it is possible to obtain approximate small p-values with only a limited number of Monte Carlo replicates, we performed computer simulations fitting different probability distributions to the sample test statistics from the random data sets generated under the null. For our baseline set-up, we use a map of 245 counties and county equivalents in the Northeast United States, with each county represented by its census-defined centroid [20]. Under the null hypothesis, the number of cases in each county is Poisson distributed. Conditioning on a total of 600 cases, the cases were randomly and independently assigned to a county with probability proportional to the 1994 female population in that county [24]. The maximum circle size of the scan statistic was set to 50% of the population.

First, we generated 100,000,000 Monte Carlo replicates of the data under the null hypothesis. The maximum log likelihood ratio among all distinct circles is the statistic reported from each replicate. These 100,000,000 statistics generated our "gold standard" distribution of log likelihood ratios, which we treat as if it were the actual distribution of the statistic under the null. Using this distribution, we find the 'true' log likelihood ratio corresponding to a given α-level by

finding the log-likelihood ratio for which the rank divided by 100,000,000 gives the desired α-level. For example, the log likelihood ratio with a rank of 1,000,000 corresponds to an α-level of 0.01, since 1,000,000/100,000,000=0.01.

Using the same parameter settings, we also generated 999 Monte Carlo replicates of the data. We used the 999 maximum log-likelihood ratios obtained from the Monte Carlo replicates to fit normal, gamma, lognormal and Gumbel distributions to the data to see if any of them would approximate the true distribution of these log-likelihood ratios. To fit the normal, gamma, and lognormal distributions we used maximum likelihood estimates, and we used the method of moments estimators for the Gumbel distribution.

The idea now is to use the fitted empirical distribution function to obtain a p-value. The p-value is calculated by finding the area under this distribution that is to the right of the observed test statistic. For this to work, it is important that the right tail of this function is similar to the right tail of the true distribution that is represented by the gold standard distribution from the 100,000,000 replicates. In order to check this, we used the cumulative distribution function (cdf) of each fitted empirical distribution function to find the critical value of the log likelihood ratio corresponding to the nominal α-level. We then ranked each critical value among the 100,000,000 log likelihood ratios in the gold standard distribution to find the true probability of rejecting the null at that critical value. We call this the rejection probability, and for the test to be unbiased, the expected value of this rejection probability must be equal to the nominal (desired) α-level. For each type of distribution, we did this 1000 times which resulted in 1000 critical values and, therefore, 1000 rejection probabilities. The average of these rejection probabilities is an estimate of the true (actual) α-level, which is then compared with the nominal α-level .

Here we present a formal description of this process; a schematic diagram is shown in Figure 1. Let $\phi_d(x)$ be the probability density function (pdf) of the distribution, *d*, obtained by using the log likelihood ratios generated from the Monte Carlo replicates to estimate the parameters for *d*. Here we use $d =$ normal, lognormal, gamma, and Gumbel. Let $\Phi_d(x)$ be the associated cdf. For the nominal α-level, α_n, and each distribution, we first find the critical value ω_{d,α_n} for which $1-\Phi_d\left(\omega_{d,\alpha_n}\right)=\int_{\omega_{d,\alpha_n}}^{\infty}\phi_d\left(x\right)dx=\alpha_n$; that is, we find $\omega_{d,\alpha_n}=\Phi_d^{-1}\left(1-\alpha_n\right)$. Note that ω_{d,α_n} is the value for which the area under $\phi_d(x)$ and to the right of ω_{d,α_n} is α_n. Now, let $\gamma\left(x\right)$ be the true pdf of the log-likelihood ratios and let $\Gamma\left(x\right)$ be the corresponding cdf. Then $1-\Gamma\left(\omega_{d,\alpha_n}\right)=\int_{\omega_{d,\alpha_n}}^{\infty}\gamma\left(x\right)dx=r_{d,\alpha_n}$ is the rejection probability from distribution *d* and associated with α_n. Of course, $\Gamma\left(x\right)$ is unknown, and here we use the observed 100,000,000 replicates as a proxy. Effectively, $1-\Gamma\left(\omega_{d,\alpha_n}\right)=\frac{1}{100,000,000}\sum_{i=1}^{100,000,000}I\left(llr_i>\omega_{d,\alpha_n}\right)=r_{d,\alpha_n}$ where llr_i is the i^{th} log-likelihood ratio in the gold-standard distribution. The average of the 1000 r_{d,α_n}'s is our

estimate of the true α-level, when the nominal α-level is α_n. Note that the α-level found using Monte Carlo hypothesis testing, which we denote α_{MC,α_n} is proven theoretically to be correct, so that $\alpha_{MC,\alpha_n} \equiv \alpha_n$.

In addition to the baseline set-up, we repeated the same experiment with different numbers of cases, different maps, different probability models generating the data, different maximum circle sizes, and different numbers of Monte Carlo replicates. The combinations that we used are summarized in Table 1. United States 3-digit zip code populations were obtained from the 1990 United States Census [25]. For each set of parameters in Table 1, we performed the experiment using 99, 999, and 9999 Monte Carlo replicates to generate the empirical/asymptotic distribution. We chose 5 nominal α-levels ($\alpha_n = 0.05, 0.01, 0.001, 0.0001, 0.00001$). All calculations were done using the SaTScan™ software.

Results

α-levels

The most important evaluation criteria is to ensure that the α-levels (type I error) are correct. For the baseline experimental set-up, the histograms in Figure 2 show 1000 rejection probabilities for each empirically fitted distribution. The mean of these rejection probabilities is an estimate of the true α-level achieved through this process. Note that in order to maintain the correct α-level, it is enough for the mean of the distribution to equal α, while the variance around α is irrelevant. Therefore, to maintain the correct α-level, it is sufficient that the rejection probabilities are centered around the desired α-level. Among the distributions assessed here, the rejection probabilities from the Gumbel approximation are centered around the nominal α-levels; for the other distributions the rejection probabilities are centered to the right of the nominal α-level. Thus, using any distribution other than the Gumbel distribution to approximate the underlying distribution of the spatial scan statistic results in anti-conservatively biased α-levels, or in other words, p-vales that are too small. This can also be seen in Figure 3, where we show a plot of the ratio of the estimation of the true α-level to the nominal α-level for each distribution. Figure 3 shows that the Gumbel approximation has very little bias. When 999 or 9999 Monte Carlo replicates were used the slight bias is conservative, whereas the bias from all other distributions is large and anti-conservative.

The above results are all based on the baseline experimental setting. For the other settings, the true α-levels are presented in Table 2 for the Gumbel distribution, showing the robustness of this method. Overall, the Gumbel approximation works quite well. The α-levels are slightly conservative whenever 999 or 9999 Monte Carlo replicates were used, and slightly anti-conservative with 99 replicates. The bias becomes increasingly conservative with an increasing number of replicates. The worst results using the Gumbel distribution were for the very extreme scenarios when there are only 6 disease cases in all of the northeastern United States, when there is a non-negligible conservative bias, and when the maximum scanning window size was limited to contain at most 1 county, when there is a non-negligible anti-conservative bias. We also evaluated the other three distributions using all of the settings, and the bias was similar to, and as bad as the results shown in Figures 2 and 3 (data not shown).

Statistical power

Statistical power is another important evaluation criterion. While the variance in the probabilities depicted in Figure 2 do not influence the α-level, a larger variance will slightly reduce the statistical power of the test [21]. We informally compared the power for the Gumbel approximation to the power obtained from Monte Carlo hypothesis testing by looking at the variance of the rejection probabilities used to calculate a range of α-levels. Using the parameter set from the baseline experimental settings, Figure 4 shows histograms of the rejection probabilities from the Gumbel approximation and from Monte Carlo hypothesis testing using different numbers of Monte Carlo replicates. The figure suggests that the variance in the rejection probabilities from the Gumbel approximation is smaller than the variance in the rejection probabilities from Monte Carlo hypothesis testing. Numerically, the ratio of the standard deviation of the rejection probabilities from the Gumbel approximation to the standard deviation of the rejection probabilities from Monte Carlo hypothesis testing is less than 1 when the same number of replicates are used for both methods, indicating that the scan statistic has greater power when the Gumbel approximation is used than with the traditional Monte Carlo hypothesis testing. When we use 9999 replicates for Monte Carlo hypothesis testing and only 999 replicates for the Gumbel approximation the ratio of the standard deviation of the rejection probabilities from the Gumbel approximation to the standard deviation of the rejection probabilities from Monte Carlo hypothesis testing is about 1; the same is true if 999 replicates are used for Monte Carlo hypothesis testing and 99 replicates are used for the Gumbel approximation. This means that in this example, 10 times as many replicates are required in order to get about the same power with Monte Carlo hypothesis testing as with the Gumbel approximation.

Discussion

We have shown that the Gumbel distribution can be used to obtain approximate p-values for the spatial scan statistic with great accuracy in the far tail of the distribution. This can be done using far less computation than required by the traditional method based on Monte Carlo hypothesis testing. A key question is then when to use Monte Carlo hypothesis testing versus Gumbel based p-values.

If the primary interest is in 0.05 and 0.01 alpha levels, or if the data set is small so that it is easy to generate and calculate the test statistic for hundreds of thousands of simulated replicas, then traditional Monte Carlo hypothesis testing works well, and the benefit of Gumbel based p-values is at most marginal. There are several instances in which the Gumbel approximations offer a clear advantage though.

If the same number of replicates is used, then the Gumbel approximation has higher power than Monte Carlo hypothesis testing. When the number of replicates divided by the desired alpha level is large, the difference in power is marginal, but when it is small, there is a clear advantage of the Gumbel approximation. More specifically, the Gumbel approximation with one-tenth the number of replicates used by Monte Carlo hypothesis testing provides approximately the same

statistical power, while using one-tenth of the computing time. Although there is some bias with the Gumbel approximation, the bias is small and, in most cases, conservative.

The most important benefit of the Gumbel approximation is its ability to calculate very small p-values with a very modest number of simulated replicates. For example, p-values on the order of 0.00001 can be accurately calculated with only 999 random replicates by using the Gumbel approximation, while it would require more than 999,999 replicates to get the same power and precision from Monte Carlo hypothesis testing.

The attempts to calculate p-values with the help of the normal, lognormal and gamma distributions all resulted in anti-conservatively biased α-levels. The bias from these approximations was so large that we do not recommend their use to approximate p-values for spatial scan statistics.

The circular purely spatial scan statistic for count data is only one of many types of scan statistics. Other types include the space-time scan statistics [2, 26], elliptical shaped spatial scan statistics [27], non-parametric irregular shaped spatial scan statistics [28-30], as well as spatial and space-time scan statistics for ordinal [31] and exponential data [32, 33]. While we have not tested the Gumbel approximation for other types of scan statistics, these statistics are all maxima and generating p-values for any of them relies on Monte Carlo hypothesis testing. It would be reasonable, then, to evaluate whether p-values for these other scan statistics could also be approximated with the Gumbel distribution.

The method used here of fitting a distribution to the statistics obtained from the Monte Carlo replicates can be applied to many other applications in which Monte Carlo hypothesis testing is used and where very small p-values are required or where computing time is limited. There is no reason to expect the Gumbel distribution to work well in all situations. In this particular example it makes sense intuitively because the scan statistic generated in each replicate is a maximum over many circles and the Gumbel distribution is a distribution of maxima. Other applications may lend themselves naturally to a different choice of distribution.

To summarize, in applications in which the precision of small p-values is not important, we suggest using Monte Carlo hypothesis testing to obtain the p-values for the spatial scan statistic. In applications in which the precision of p-values is important or where each replicate takes a long time to complete, the Gumbel based p-values are often advantageous for reasons of both computational speed and statistical power. To facilitate its use, Gumbel based p-values have been added to version 9 of the freely available SaTScan software, which can be downloaded from www.satscan.org.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AA programmed the simulations, analyzed the simulation results, and drafted the manuscript. KK and MKconceived of the idea, provided guidance, and helped draft the manuscript.

Acknowledgements

This research was funded by grant #RO1CA095979 from the National Cancer Institute and by Models of Infectious Disease Agent Study (MIDAS) grant #U01GM076672 from the National Institute of General Medical Sciences.

References

1. **Kulldorff, M., *A Spatial Scan Statistic.* Commun. Statist. - Theory Meth., 1997. 26(6): p. 1481-1496.**
2. **Kulldorff, M., *Prospective time periodic geographical disease surveillance using a scan statistic.* J. R. Statist. Soc. A, 2001. 164(1): p. 61-72.**
3. **Naus, J.I., *Clustering of Points in Two Dimensions.* Biometrika, 1965. 52: p. 263-267.**
4. **Gaudart, J., et al., *Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk.* BMC Medical Research Methodology, 2005. 5(1): p. 22.**
5. **Gosselin, P., et al., *The Integrated System for Public Health Monitoring of West Nile Virus (ISPHM-WNV): a real-time GIS for surveillance and decision-making.* International Journal of Health Geographics, 2005. 4(1): p. 21.**
6. **Fukuda, Y., et al., *Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan.* International Journal of Health Geographics, 2005. 4(1): p. 16.**
7. **Ozonoff, A., et al., *Cluster detection methods applied to the Upper Cape Cod cancer data.* Environmental Health: A Global Access Science Source, 2005. 4(1): p. 19.**
8. **Sheehan, T.J. and L. DeChello, *A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997.* International Journal of Health Geographics, 2005. 4(1): p. 15.**
9. **DeChello, L. and T.J. Sheehan, *Spatial analysis of colorectal cancer incidence and proportion of late-stage in Massachusetts residents: 1995–1998.* International Journal of Health Geographics, 2007. 6(1): p. 20.**
10. **Klassen, A., M. Kulldorff, and F. Curriero, *Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors.* International Journal of Health Geographics, 2005. 4(1): p. 1.**
11. **Ozdenerol, E., et al., *Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters.* International Journal of Health Geographics, 2005. 4(1): p. 19.**
12. **Kleinman, K.P., et al., *A model-adjusted space-time scan statistic with an application to syndromic surveillance.* Epidemiol Infect, 2005. 133: p. 409-419.**
13. **Nordin, J.D., et al., *Simulated Anthrax Attacks and Syndromic Surveillance.* Emerging Infectious Diseases, 2005. 11(9): p. 1394-1398.**
14. **Yih, W.K., et al., *Ambulatory-Care Diagnoses as Potential Indicators of Outbreaks of Gastrointestinal Illness --- Minnesota.* MMWR, 2005. 54(Supplement): p. 157-162.**
15. **Besculides, M., et al., *Evaluation of School Absenteeism Data for Early Outbreak Detection -- New York City, 2001 -- 2002.* MMWR, 2004. 53(Supplement): p. 230.**
16. **Heffernan, R., et al., *Syndromic Surveillance in Public Health Practice, New York City.* Emerging Infectious Diseases, 2004. 10(5): p. 858-864.**
17. **Sheridan, H.A., et al., *A temporal-spatial analysis of bovine spongiform encephalopathy in Irish cattle herds, from 1996 to 2000.* Canadian Journal of Veterinary Research, 2005. 69(1): p. 19-25.**
18. **Dwass, M., *Modified randomization tests for nonparametric hypothesis.* Ann. Math. Statist., 1957. 28: p. 181-187.**

19. **Kleinman, K., R. Lazarus, and R. Platt,** ***A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism.*** **Am. J. Epidemiol., 2004. 159(3): p. 217-224.**
20. **Kulldorff, M., and Information Management Services, Inc.,** ***SaTScan™ v5.1: Software for the spatial and space-time scan statistics.*** **2005.**
21. **Jöckel, K.-H.,** ***Finite Sample Properties and Asymptotic Efficiency of Monte Carlo Tests.*** **The annals of Statistics, 1986. 14(1): p. 336-347.**
22. **Gumbel, E.J.,** ***Statistics of Extremes*****. 2004, Mineola, NY: Dover. 375.**
23. **Coles, S.,** ***An Introduction to Statistical Modeling of Extreme Values*****. Springer Series in Statistics. 2001, London: Springer-Verlag. 208.**
24. **Kulldorff, M., et al.,** ***Breast Cancer Clusters in the Northeast United States: A Geographic Analysis.*** **Am. J. Epidemiol., 1997. 146(2): p. 161-170.**
25. **. p. U.S. Census Bureau.**
26. **Kulldorff, M., et al.,** ***Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico.*** **Am J Public Health, 1998. 88(9): p. 1377-1380.**
27. **Kulldorff, M., et al.,** ***An elliptic spatial scan statistic.*** **Statistics in Medicine, 2006. 25(22): p. 3929-3943.**
28. **Duczmal, L. and R. Assunção,** ***A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters.*** **Computational Statistics & Data Analysis, 2004. 45(2): p. 269-286.**
29. **Tango, T. and K. Takahashi,** ***A flexibly shaped spatial scan statistic for detecting clusters.*** **International Journal of Health Geographics, 2005. 4(1): p. 11.**
30. **Assunção, R., et al.,** ***Fast detection of arbitrarily shaped disease clusters.*** **Statistics in Medicine, 2006. 25(5): p. 723-742.**
31. **Jung, I., M. Kulldorff, and A.C. Klassen,** ***A spatial scan statistic for ordinal data.*** **Statistics in Medicine, 2007. 26(7): p. 1594-1607.**
32. **Huang, L., M. Kulldorff, and D. Gregario,** ***A Spatial Scan Statistic for Survival Data.*** **Biometrics, 2007. 63: p. 109-118.**
33. **Cook, A.J., D.R. Gold, and Y. Li,** ***Spatial Cluster Detection for Censored Outcome Data.*** **Biometrics, 2007. 63: p. 540-549.**

Figure legends

Figure 1: Schematic for finding the rejection probability. First find the critical value ω_{d,α_n} on the fitted empirical distribution, then use that value to find r_{d,α_n}, the 'true' probability of rejecting the null hypothesis given the highly accurate 'gold standrad' pdf obtained from the 100,000,000 replicates under the null.

Figure 2: Histograms of the rejection probabilities obtained from each of the fitted empirical distributions. The vertical gray lines indicate the nominal α-level. For α = 0.05, α = 0.01, and α = 0.001, the scales are the same within those columns with the scale marked at the bottom of each column. For α = 0.0001 and α = 0.00001, the scale is different for the normal distribution than for the other 3 distributions, as is indicated at the bottom of the histograms.

Figure 3: Ratio of estimated α-levels to nominal α-levels for 4 distributions and different numbers of Monte Carlo replicates used to estimate the parameters for each distribution.

Figure 4: Histograms of the rejection probabilities obtained using the Gumbel approximation and from Monte Carlo hypothesis testing based on 99, 999, or 9,999 Monte Carlo replicates.

Table 1: Combinations of settings used; bold indicates baseline settings.

Number of cases	Map	Model type	Maximum circle size
600	**NE counties**	**Poisson**	**50%**
600	NE counties	Bernoulli	50%
600	NE counties	Poisson	1 county
600	US 3-digit zip codes	Poisson	50%
6	NE counties	Poisson	50%
6000	NE counties	Poisson	50%
60000	NE counties	Poisson	50%

Number of cases	Maximum circle size	Map	Model type	Number of Monte Carlo replicates	Nominal alpha				
					0.00001	0.0001	0.001	0.01	0.05
6	50%	NE counties	Poisson	99	0.000003	0.00004	0.0006	0.008	0.051
				999	0.000001	0.00002	0.0004	0.007	0.048
				9999	0.000001	0.00002	0.0004	0.007	0.047
600	50%	NE counties	Poisson	99	0.000013	0.00012	0.0012	0.011	0.054
				999	0.000006	0.00008	0.0009	0.010	0.051
				9999	0.000006	0.00007	0.0008	0.010	0.050
			Bernoulli	99	0.000014	0.00013	0.0012	0.011	0.053
				999	0.000007	0.00008	0.0009	0.010	0.050
				9999	0.000007	0.00008	0.0008	0.010	0.047
		US 3 digit zip codes	Poisson	99	0.000014	0.00013	0.0012	0.011	0.054
				999	0.000007	0.00008	0.0009	0.010	0.052
				9999	0.000006	0.00008	0.0009	0.010	0.051
	1 county	NE counties	Poisson	99	0.000033	0.00022	0.0016	0.012	0.053
				999	0.000020	0.00016	0.0012	0.011	0.051
				9999	0.000018	0.00015	0.0018	0.011	0.050
6000	50%	NE counties	Poisson	99	0.000013	0.00012	0.0011	0.011	0.053
				999	0.000007	0.00008	0.0009	0.010	0.051
				9999	0.000006	0.00007	0.0008	0.010	0.050
60000	50%	NE counties	Poisson	99	0.000013	0.00012	0.0011	0.011	0.054
				999	0.000006	0.00007	0.0009	0.010	0.051
				9999	0.000006	0.00007	0.0008	0.009	0.050

Table 2: Estimated α-levels for the Gumbel approximation for different parameters, corresponding to five nominal α-levels.

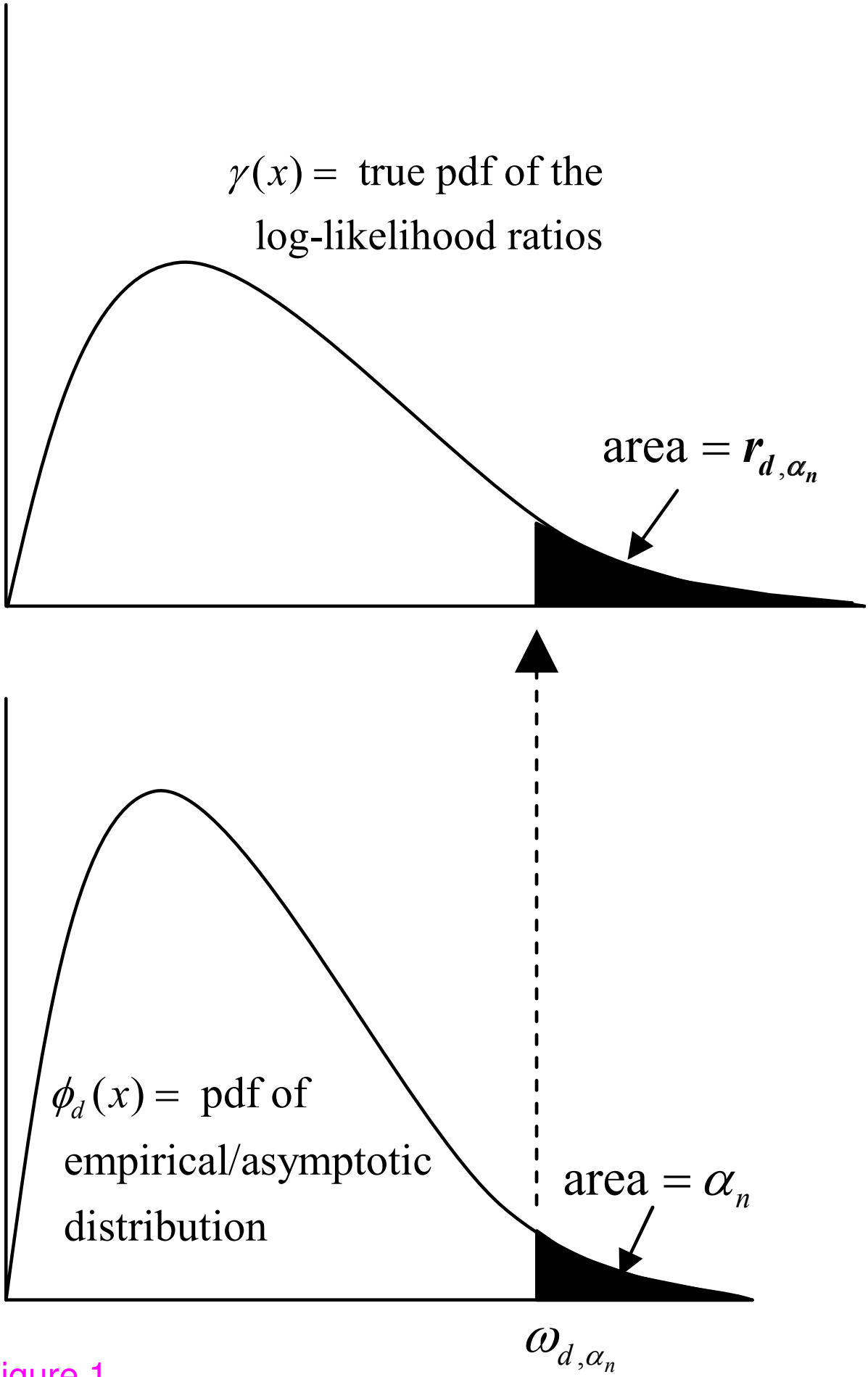
$\gamma(x) =$ true pdf of the
log-likelihood ratios
area $= r_{d,\alpha_n}$
$\phi_d(x) =$ pdf of
empirical/asymptotic
distribution
area $= \alpha_n$
ω_{d,α_n}

Figure 1

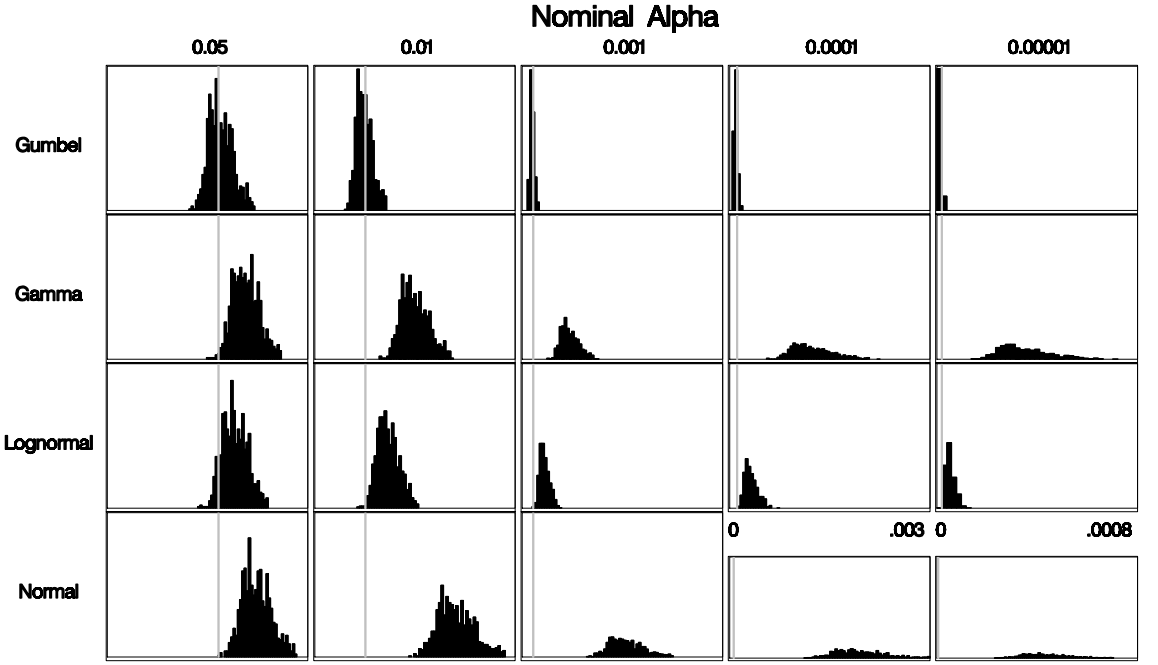
Nominal Alpha
0.05
0.01
0.001
0.0001
0.00001
Gumbel
Gamma
Lognormal
Normal
0
.09
0
.04
0
.02
0
.007
0
.004
0
.003
0
.0008

Figure 2

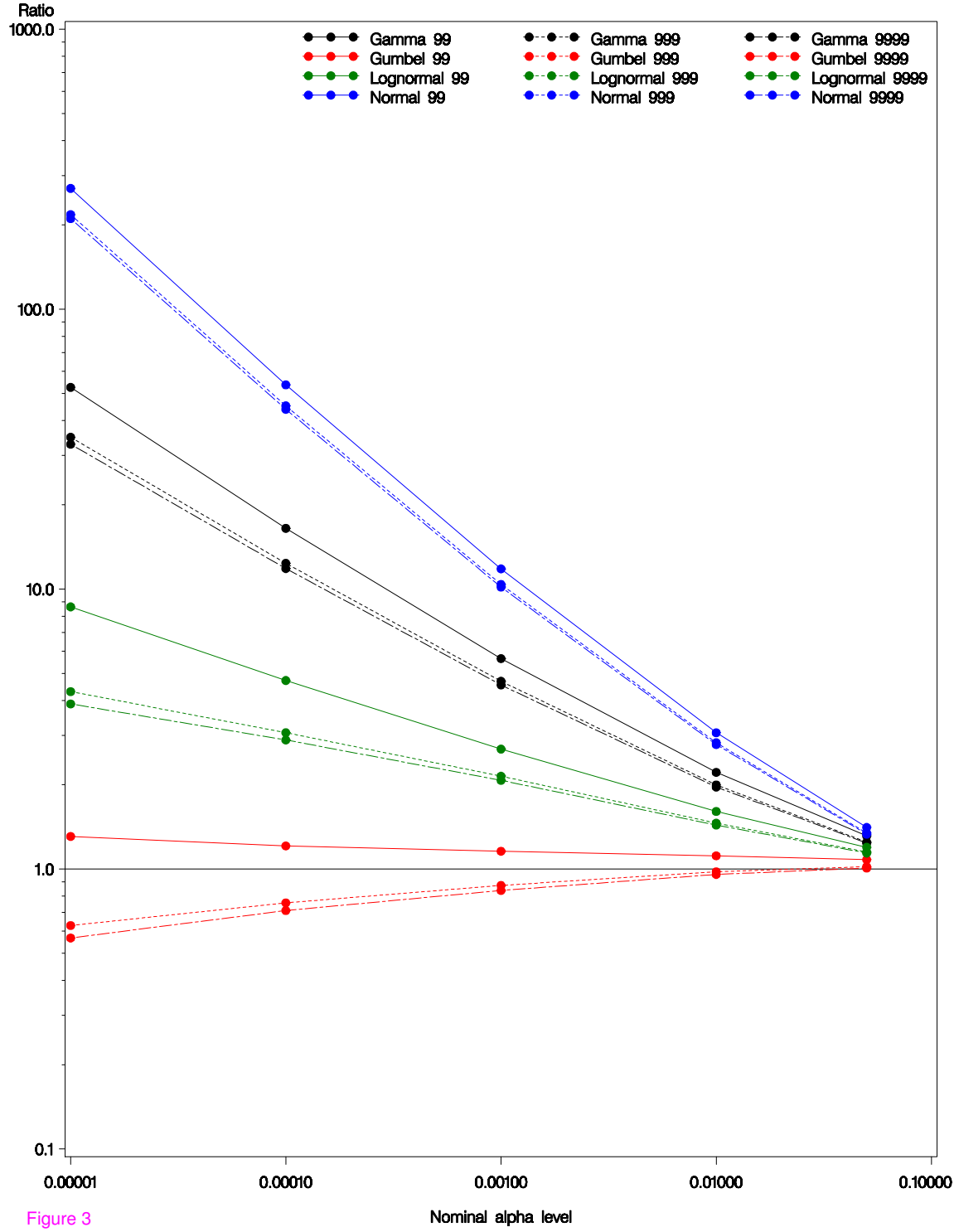
Ratio
1000.0
100.0
10.0
1.0
0.1
Gamma 99
Gumbel 99
Lognormal 99
Normal 99
Gamma 999
Gumbel 999
Lognormal 999
Normal 999
Gamma 9999
Gumbel 9999
Lognormal 9999
Normal 9999
0.00001
0.00010
0.00100
0.01000
0.10000
Nominal alpha level

Figure 3

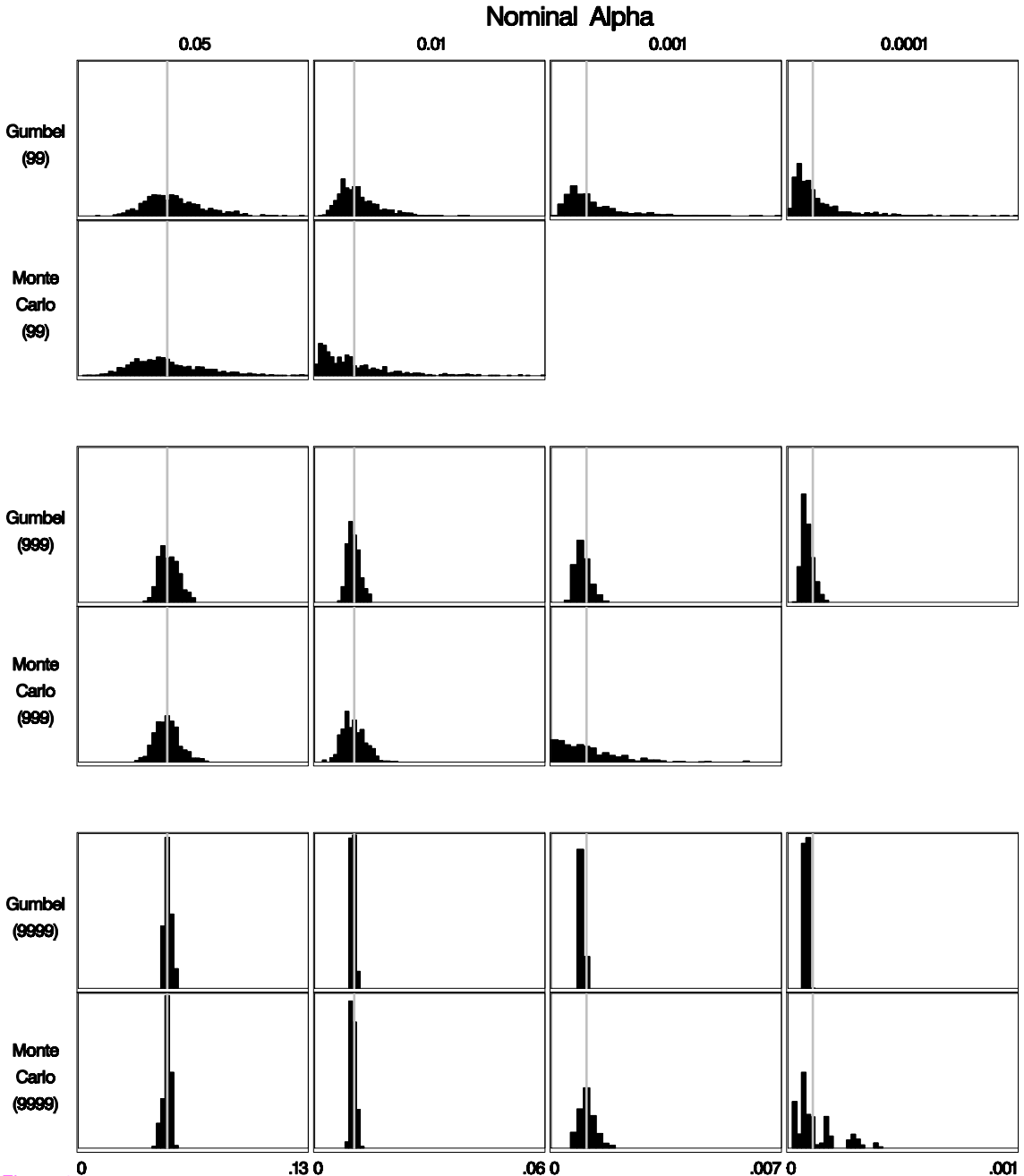
Nominal Alpha
0.05
0.01
0.001
0.0001
Gumbel (99)
Monte Carlo (99)
Gumbel (999)
Monte Carlo (999)
Gumbel (9999)
Monte Carlo (9999)
0
.13 0
.06 0
.007 0
.001

Figure 4