A new test for the Poisson distribution

Lawrence D. Brown *

Linda H. Zhao †

Department of Statistics

University of Pennsylvania

Philadelphia, PA 19104

November 7, 2001

Abstract

We consider the problem of testing whether a sample of observations comes from a single Poisson distribution. Of particular interest is the alternative that the observations come from Poisson distributions with different parameters. Such a situation would correspond to the frequently discussed situation of overdispersion.

We propose a new test for this problem that is based on Anscombe's variance stabilizing transformation. There are number of tests commonly proposed, and we compare the performance of these tests under the null hypothesis with that of our new test. We find that the performance of our test is competitive with the two best of these. The asymptotic distribution of the new test is derived and discussed.

We also describe how to compute Minimum Bayes Factors for our test and the various alternative tests.

Use of these tests is illustrated through two examples of analysis of call-arrival times from a telephone call center. The example facilitates careful discussion of the performance of the tests for small λ and moderately large n.

Key words: Poisson variables, Anscombe's transformation, likelihood ratio test, Chi-squared test, overdispersion

*Research supported in part by NSF grant DMS-9971751.

†Research supported in part by NSF grant DMS-9971848.

1 Introduction

A variety of tests is available for testing whether a sample of observations comes from a Poisson distribution. This article proposes an additional test based on Anscombe's (1948) variance stabilizing transformation. We examine the performance of this test and compare it with four other tests in current use. We find this new test to be competitive in performance with the best of these alternatives. We recommend it on this basis, and also because the heuristic idea underlying it easily adapts for a variety of related applications. In this connection, see Brown, Mandelbaum, Sakov, Shen, Zeltyn and Zhao (2001) and Brown, Zhang and Zhao (2001).

In this article we use call-arrival data gathered at an Israeli call center as motivation and illustration of the various problems and methodologies we discuss. We provide a very brief discussion in Section 2 of this data application.

We investigate both the conventional P-values for these test settings and the minimum Bayes factors as described by Berger (2001). We feel that such factors are more appropriate than P-values in contexts like the principle examples treated here, and hence feel it is important to provide a methodology that enables computation of these factors.

The four additional types of test statistics we examine are the likelihood ratio statistic, the corresponding chi-squared statistic sometimes called the "dispersion test", a putatively normal version of this statistic sometimes attributed to Neyman and Scott and the multinomial Pearson chi-squared statistic. The multinomial chi-squared statistic is not well suited for the range of applications we address. The performance of the Neyman-Scott test is shown to be inferior to those built from the remaining three statistics. Among those three we favor the new test based on its ease of use, diagnostic ability and breadth of application.

Suppose the null hypothesis is true, that the data come from a Poisson(λ) distribution. When λ is not small all three recommended tests (the new test, the dispersion χ^2, and likelihood ratio) appear fully satisfactory for practical applications. When λ is small the nominal null distribution for the likelihood ratio test is quite inaccurate. The test should not then be used in the usual form as presented here.

In Section 5 we derive the asymptotic distribution of our new test statistic as $n \to \infty$ and λ fixed. It is shown that this implies that the heuristic nominal null distribution is not fully accurate when λ is small, even if $n \to \infty$. Thus, when λ is very small (say $\lambda \leqslant 5$), the new test we propose is slightly inaccurate. The source of that inaccuracy is explained in Section 4, and an easily implemented correction is proposed that is satisfactory for moderately large sample sizes (say 50 or more, depending partly on how small is λ).

In Section 2 we describe the call center data we will use as an example of an application of our methodology. The various tests are described in Section 3, including the new test we propose based on Anscombe's variance stabilizing transformation. Section 4 presents some simulation results comparing our test and the various other tests. The asymptotic distribution of the new test is discussed in Section 5. Section 6 contains a derivation of the Minimum Bayes factors for the respective statistics. Section 7 concludes with some empirical results for the call center data and some further empirical results about the situation with small λ.

2 Call Center Arrival Data

The data accompanying our study was gathered at a relatively small Israeli bank telephone call center in 1999. The portion of data of interest to us here involves records of the arrival time of

service-request calls to the center. There are calls in which the caller requests service from a call center representative. It is reasonable to conjecture that these arrival times are well modeled by an inhomogeneous Poisson process. The arrival rate for this process should depend only on time of day, and perhaps other calendar related covariates such as month or day of the week. There are different categories of service that may be requested, and preliminary analysis clearly shows that this factor should also be considered since the arrival rate patterns differ considerably. For more information about various aspects of this data see Brown, Gans, Mandelbaum, Sakov , Shen, Zeltyn and Zhao (2001). Other features of the call arrival process are investigated in Brown, Mandelbaum, Sakov, Shen, Zeltyn and Zhao (2001).

If the arrival process for a given call category is as above then the number of arrivals each day within any given interval of time should be independent Poisson variables with a parameter that depends only on the given time interval. If other covariates are involved (such as day of the week) then the Poisson parameter may also depend on these.

The histograms in Figure 1 and 2 show the results from two typical samples. Figure 1 shows the number of standard calls arriving on each regular workday in Nov. and Dec., between 4:30pm and 4:45pm. Figure 2 is a similar histogram for the special category of calls requesting internet assistance arriving between 4:30pm and 4:45pm from Aug. through Dec. In each case it is of interest to test the null hypothesis that these data arise from Poisson populations with their own respective means. Note the different levels of calls/day in these two samples, as well as the different sample sizes.

One reason for considering standard calls only for Nov. and Dec. is that there is some evidence of an increased rate of standard calls in Nov. and Dec. Fig 3 presents the data that provides

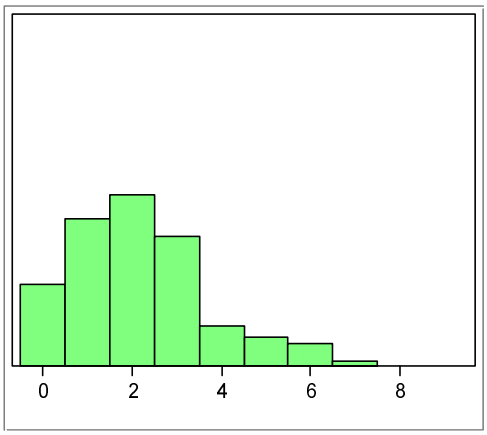

Figure 1: No. of daily calls for Internet Service arriving between 4:30pm and 4:45pm, Regular weekdays, Aug. – Dec. n = 107, $\bar{x} = 2.18$, $s^2 = 2.47$

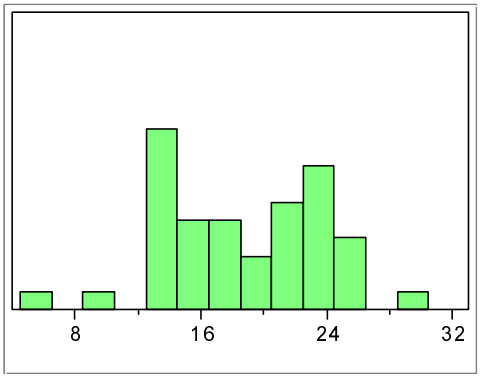

Figure 2: No. of daily calls for Standard Service arriving between 4:30pm and 4:45pm, Regular weekdays, Nov. – Dec. n = 44, $\bar{x} = 18.66$, $s^2 = 25.95$

this evidence. The hypothesis of interest here would be that the daily arrival rate is the same in each month. An extensive discussion related to testing such a hypothesis can be found in Brown, Mandelbaum, Sakov, Shen, Zeltyn and Zhao (2001).

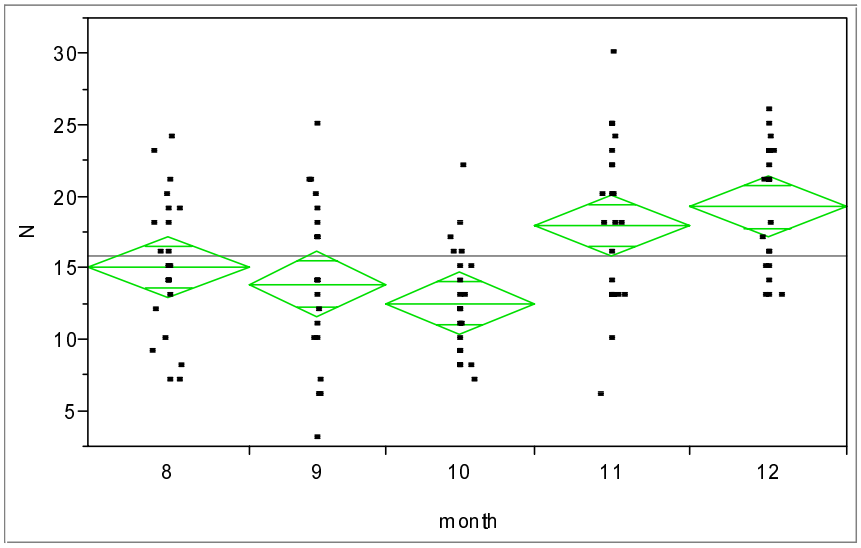

Figure 3: Conventional side-by-side dot plot for Standard Service calls from 4:30pm – 4:45pm (Means diamonds are from the conventional one-way ANOVA analysis. For description of a more appropriate type of analysis see Brown, Mandelbaum, Sakov, Shen, Zeltyn and Zhao (2001).)

3 Tests for the Poisson distribution

Let $X_1, \ldots, X_n$ be independent non-negative integer valued random variables with $P(X = x) = f(x)$. The basic null hypothesis of interest is that

$$H_0 : X_i \sim \text{Poiss}(\lambda_i), \quad \lambda_1 = \ldots = \lambda_n. \tag{1}$$

In a context such as ours the alternative hypothesis is not always delineated precisely. In general one usually wishes to focus on alternatives that are "over-dispersed" in the sense that

$$\frac{E(S^2)}{E(\overline{X})} > 1 \tag{2}$$

where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

For this reason we consider the alternative hypothesis that

$$H_a : X_i \sim \text{Poisson}(\lambda_i), \sum(\lambda_i - \overline{\lambda})^2 > 0. \tag{3}$$

We propose a new test for this problem. We also briefly describe four tests in common use for H_0. We will later focus our attention on properties of the new test in relation to the others.

3.1 A new test based on Anscombe's statistic

Anscombe (1948) derived the second order variance stabilizing transformation for a Poisson variable. If $N \sim \text{Poiss}(\lambda)$ he showed that

$$\text{Var}_\lambda \left(\sqrt{N + \frac{3}{8}} \right) = \frac{1}{4} + O\left(\frac{1}{\lambda} \right). \tag{4}$$

On this basis it is natural to define $Y_i = \sqrt{X_i + 3/8}$ and use the statistic

$$T_{new} = 4 \sum (Y_i - \overline{Y})^2$$

to provide a test for H_0.

Formula (4) suggests that Y_i is approximately normal with variance $1/4$ and mean

$$\nu(\lambda_i) = E_{\lambda_i}(Y_i) = E_{\lambda_i}(\sqrt{N + 3/8}). \tag{5}$$

Under this approximation it would follow that when H_0 is true T_{new} has approximately a Chi-squared distribution with $(n-1)$df. We thus reject H_0 if $T_{new} > \chi^2_{n-1;1-\alpha}$. Further one may conclude that under H_a T_{new} has approximately a noncentral χ^2_{n-1} distribution. In summary it is

reasonable to act as if

$$T_{new} \sim \chi^2_{n-1}(4\sum(\nu(\lambda_i) - \overline{\nu}_n)^2) \tag{6}$$

where

$$\overline{\nu}_n = \frac{1}{n}\sum_{i=1}^{n}\nu(\lambda_i).$$

The empirical results in Section 4 indicate that this approximation is reasonably accurate under H_0 even for fairly small λ and n. Further simulations we have carried out (not reported here) suggest that this approximation is also fairly good for a variety of choices of $\{\lambda_i\}$ in H_a, even for moderate n so long as all λ_i are not small.

Section 5 presents some asymptotic theory concerning the distribution of T_{new}. This theory helps explain why (6) provides numerically satisfactory results even though it is not quite asymptotically valid as $n \to \infty$, even under H_0.

In the context of nonparametric density estimation Brown, Zhang and Zhao (2001) has suggested using the transformation $\sqrt{N + 1/4}$ instead of $\sqrt{N + 3/8}$. This is because

$$E_\lambda(\sqrt{N + 1/4}) = \sqrt{\lambda} + O(1/\lambda).$$

In the context of Brown, Zhang and Zhao (2001) accuracy in estimation of $\sqrt{\lambda}$ is of prime importance, rather than stability of the variance. However for the Poisson tests under investigation here validity of (4) is more important, and the transformation $\sqrt{X_i + 3/8}$ performs slightly better than would $\sqrt{X_i + 1/4}$.

In Brown, Cai and DasGupta (2001) we investigated confidence intervals for a Poisson mean. This is a related problem but techniques for best confidence intervals do not necessarily extend

to best tests of H_0, and vice-versa. Some results about the confidence interval problem are also reported in Brown, Zhang and Zhao (2001).

The test statistic T_{new} appears to us a natural proposal given Anscombe's well known variance stabilizing transformation. We expect it has been used in the form (6) by some practitioners. But the only reference we have found is Huffman (1984) that presents a sample size two ($n = 2$) version of this test, and also discusses testing a generalization of H_0 when $n = 2$.

3.2 Likelihood ratio statistic

The likelihood ratio statistic for testing H_0 versus H_a is

$$T_{LR} = 2 \sum_{i=1}^{n} X_i \ln \left(\frac{X_i}{\overline{X}} \right).$$

Under the null hypothesis this statistic is asymptotically distributed as a Chi-squared variable with $n - 1$ df. (asymptotically as $n \to \infty$ for fixed λ). Hence this test rejects H_0 when $T_{LR} > \chi^2_{n-1;1-\alpha}$.

Under alternatives in H_a this statistic has approximately a non-central Chi-squared distribution with $(n-1)$df and non-central parameter $\psi^2 = \sum_{i=1}^{n} (\lambda_i - \overline{\lambda})^2 / \overline{\lambda}$ where $\overline{\lambda} = \sum_{i=1}^{n} \lambda_i / n$. We write, $T_{LR} \sim \chi^2_{n-1}(\psi^2)$. This approximation is asymptotically valid as $\lambda \to \infty$ for fixed n with $\lambda_1, \ldots, \lambda_n$ chosen to depend on n in such a way that ψ^2 remains constant, or as $n \to \infty$ with $\lambda_1, \ldots, \lambda_n$ chosen so that $\liminf \overline{\lambda} > 0$ and $\psi^2 = O(\sqrt{n})$.

3.3 Conditional chi-squared statistic; also called the Poisson dispersion test

Under the null hypothesis the conditional distribution of $X_1, \ldots, X_n$ given $\sum X_i = n\overline{X}$ is multinomial $(n\overline{X}, (1/n, \ldots, 1/n))$. This motivates as a test statistic,

$$T_{CC} = \sum \frac{(X_i - \overline{X})^2}{\overline{X}} = \frac{(n-1)S^2}{\overline{X}}$$

where under H_0 has an (asymptotic) Chi-squared distribution with $(n-1)$df. (Hence reject H_0 if $T_{CC} > \chi^2_{n-1;1-\alpha}$.) This statistic can also be motivated as the asymptotic chi-squared approximation to the likelihood ratio test of Section 3.2. Some authors (e.g., Rice(1995)) call this the Poisson dispersion test or the variance test (Cochran (1954)). See also Agresti (1990, p. 479).

Under H_a $T_{CC} \sim \chi_{n-1}(\psi^2)$; with this approximation being asymptotically valid under the same conditions as described for T_{LR}.

3.4 Neyman-Scott statistic

This statistic is directly motived by the expression (2). It is often proposed as test of H_0. See for example Lindsay (1995); and see Joengbloed and Koole (2000) for application of this test to telephone call-center data. The statistic is

$$T_{NS} = \sqrt{\frac{n-1}{2}} \left(\frac{S^2}{\overline{X}} - 1 \right).$$

This statistic is normalized so that asymptotically $T_{NS} \sim N(\psi^2/\sqrt{2n}, 1)$. (Hence this test rejects if $T_{NS} > \Phi^{-1}(1-\alpha)$.) The asymptotic assertion here is valid as $n \to \infty$ with $\lambda_1, \ldots, \lambda_n$ chosen so that $\psi^2 = O(\sqrt{n})$ and $\liminf \overline{\lambda} > 0$.

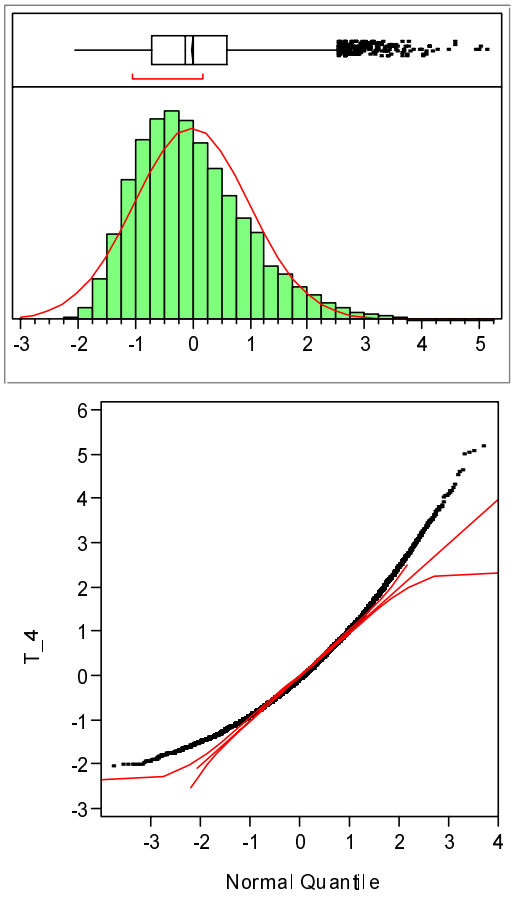

Figure 4: Histogram (with best fitting normal curve) and Normal Quantile plot for T_{NS}; $\lambda = 12$, $n = 12$, 10,000 Monte Carlo samples

It can be seen that under H_0 T_{NS} is the standard normal approximation to the chi-squared statistic T_{CC}. It should therefore not be surprising that the null distribution of T_{NS} is not close to its limiting normal distribution until n is moderately large. Fig 4 shows this non-normality in the case $n = 12$. The fact that the true null distribution of T_{NS} is not close to its nominal limiting distribution means that tests constructed using critical values from this will not have close to their nominal significance level. Correspondingly their nominal P-values based on the limiting distribution will also be considerably in error. For this reason we recommend against use of T_{NS}. (For comparative purposes we have nevertheless included T_{NS} in the numerical results in Section

4.)

3.5 Multinomial Chi-squared statistic

Let $M_j = \#\{X_i : X_i = j\}$, $j = 0, 1, \ldots$. It is possible to view $\{M_j\}$ as a multinomial variable and construct a Chi-squared test of H_0 on this basis. This can be reasonably satisfactory if n is fairly large and λ is rather small, as for the data in Fig 2. Otherwise we do not recommend it. We include this test in this subsection for the sake of completeness, but will not discuss it in the remainder of the paper.

In general the test statistic is formally as follows. Let $G_k = \{m_{k-1} + 1, \ldots, m_k\}$, $k = 1, \ldots, K$, be mutually disjoint sets of consecutive integers, where $m_0 = 0$, $m_K = \infty$. Let

$$M_k^* = \sum_{j \in G_k} M_j, \ k = 1, \ldots, K. \tag{7}$$

under H_0 $M^* = \{M_k\}$ is multinomial$(n\overline{X}, p^*)$ where $p^* = \{p_k^*\}$ with

$$p_k^*(\lambda) = \sum_{j \in G_k} \frac{\lambda^j \exp(-\lambda)}{j!}. \tag{8}$$

Let $\hat{\lambda}^*$ denote the MLE of λ based on observation of M^* under the multinomial model (8). Define the test statistic to be

$$T_{MC} = \sum \frac{(M_k^* - p_k^*(\hat{\lambda}^*))^2}{p_k^*(\hat{\lambda}^*)}.$$

Reject H_0 if $T_{MC} > \chi^2_{K-2;1-\alpha}$ where $\chi^2_{K-2;1-\alpha}$ denotes the $(1-\alpha)^{th}$ quantile of the Chi-squared distribution with $K - 1$ df.

It is generally asserted that this test has approximately the nominal level of significance, α, so

long as no $p_k^*(\hat{\lambda}^*)$ is small. (Values of $p_k^*(\hat{\lambda}^*) > 3$, or > 5, are often claimed to be satisfactory here; we have not investigated this issue.) The necessity to form suitable groups G_k that depend somewhat on $n\overline{X}$ and λ makes this test awkward to implement in general. Calculation of $\hat{\lambda}^*$, while feasible is also not in general convenient. Further, this test is an omnibus test of H_0 that is not particularly focused on detecting over-dispersion. For these reasons we do not feel this test is desirable for the range of application we are considering, and we do not discuss it in the remainder of the paper.

When λ is small and n is moderately large one may satisfactorily implement this test with $G_k = \{k-1\}$, $k = 1, \ldots, K-1$ and $G_K = \{K-1, \ldots\}$ and one may take $\hat{\lambda}^* = \hat{\lambda} = \overline{X}$. ($K$ should be chosen so that $p_k^*(\hat{\lambda}^*)$ is not small.) This is the case for the data in Fig 2. For that data choosing $K = 7$ yields $T_{MC} = 2.970$ with a P-value of $p = 0.295$.

4 Empirical results under H_0

This section reports selected empirical results about the null distribution of the statistics T_{new}, T_{LR}, T_{CC}, T_{NS}. These results are summarized in Table 1. This table gives information about the empirical type I error rates for tests computed using the nominal null distribution of various statistics. The table also contains an overall measure of how close is the empirical χ^2 or normal null distribution. The table also indirectly provides information about the accuracy of P-values calculated from the nominal distributions since accuracy of type I error rates and of P-values are linked concepts.

The general impression from the table is that the empirical type I error rates using any of T_{new}, T_{LR}, T_{CC} are reasonably accurate when $\lambda \geqslant 12$. Even when $\lambda = 5$ satisfactory accuracy is evident for T_{new} and T_{CC}. The results in Section 5 suggest a modified nominal null distribution be used

when λ is small to calculate critical values for T_{new}. The results in Section 5 also confirm that T_{LR} is a less desirable choice when $\lambda \leqslant 5$. Overall, the empirical type I errors using the T_{LR} are less accurate than those from the other three statistics, as one would also expect from the results reported in Fig 4.

The quantities reported in Table 1 are defined as follows. Let G denote generally the nominal null cumulative distribution of a statistic T. (For T_{NS}, G is standard normal. For the other statistics G is χ^2_{n-1}.) Let κ_α denote the α critical values, $\kappa(\alpha) = G^{-1}(1-\alpha)$. Let H denote the true null distribution of the statistic. Then the true type I error is $1 - H(\kappa(\alpha))$. The table reports Monte-Carlo estimates based on 10,000 samples of these quantities for various statistics and values of n, λ. The standard errors are the theoretical values $\sqrt{\alpha(1-\alpha)/10000}$.

Table 1 also reports a measure of the disparity between the nominal G and the true H as measured via the Kolmogorov-Smirnov distance

$$D^* = \sup_t |H(t) - G(t)|.$$

Again, the values reported derive from 10,000 simulations. To be more precise, each entry in the last column of the table reports the value of

$$\hat{D}^*_N = \sup_t |\hat{H}_N(t) - G(t)| \tag{9}$$

where $\hat{H}$ denotes the sample CDF from the N=10000 simulated values of T.

Simulated values of $\hat{D}^*_N$ have the Kolmogorov-Smirnov limiting distribution. This is not a normal

Table 1: Empirical Type I errors (10,000 repetitions) and $\hat{D}_N^*$ defined in (9)

n	λ	Statistic	$\alpha = .1$ SE. = 0.003	$\alpha = .05$ SE. = 0.002	$\alpha = .01$ SE. = 0.001	$\alpha = .005$ SE. = 0.001	$\hat{D}^* = \sup \lvert \hat{H} - G \rvert$ ESE. = 0.007
20	5	T_{new}	0.1107	0.0585	0.0132	0.0070	0.0130
20	5	T_{LR}	0.1359	0.0724	0.0173	0.0089	0.0588
20	5	T_{CC}	0.0977	0.0495	0.0103	0.0059	0.0105
20	5	T_{NS}	0.1039	0.0620	0.0220	0.0148	0.0457
12	12	T_{new}	0.1050	0.0540	0.0122	0.0065	0.0094
12	12	T_{LR}	0.1102	0.0563	0.0120	0.0062	0.0130
12	12	T_{CC}	0.1007	0.0505	0.0104	0.0054	0.0057
12	12	T_{NS}	0.1082	0.0670	0.0260	0.0179	0.0611
5	25	T_{new}	0.1008	0.0510	0.0103	0.0053	0.0035
5	25	T_{LR}	0.1027	0.0517	0.0101	0.0051	0.0069
5	25	T_{CC}	0.0994	0.0490	0.0095	0.0046	0.0066
5	25	T_{NS}	0.1059	0.0696	0.0312	0.0231	0.0955

distribution. In particular, a 95% confidence region for $H(t)$ is

$$\sup_t |H(t) - \hat{H}_n(t)| \leqslant 2\, ESE$$

where

$$ESE = \frac{1.96 \times 0.5}{1.36\sqrt{10000}} = 0.007.$$

For this reason we have chosen to report the effective standard error, ESE, as the measure of the precision of our Monte-Carlo simulation.

Note that for T_{new} and T_{CC} D_N^* is acceptably small. Indeed, it is less than $2 \times ESE$, and hence using this we would not reject at level .05 the null hypothesis that $H = G$. This is also true for T_{LR} when $\lambda = 12$ and 25. But when $\lambda = 5$ the performance in this regard is less satisfactory, as is the performance of T_{NS} for all combinations of n, λ in the table.

5 Asymptotic distribution of T_{new}

We have suggested approximating the null distribution of T_{new} as a Chi-squared with $(n-1)$df. The empirical results in the previous section suggest that this approximation is satisfactory for practical applications. We now explore the asymptotic distribution of T_{new} as $n \to \infty$. We show that the limiting null distribution is not Chi-squared $(n-1)$ but is very close to Chi-squared $(n-1)$ so long as λ is not small. This closeness explains why the Chi-squared approximation is suitable for nearly all practical applications. Finally, we also provide similar results about the distribution under H_a.

Note that $E_\lambda(T_{new}) = 4(n-1)\mathrm{Var}_\lambda(Y)$. As noted at (4), Anscombe (1948) proved by an asymptotic expansion that

$$\xi(\lambda) \stackrel{\triangle}{=} 4\mathrm{Var}_\lambda(Y) = 1 + O(1/\lambda). \tag{10}$$

This expression is not only asymptotically accurate – it is nearly the exact truth so long as $\lambda > 4$. Figs 5 and 6 show plots of $\xi(\lambda) = 4\mathrm{Var}_\lambda(Y) = E_\lambda(T_{new})/(n-1)$ derived via direct calculation. In particular,

$$\xi(\lambda) = (n-1)^{-1}E_\lambda(T_{new}) \leqslant 1.0025. \tag{11}$$

(The maximum value of $E_\lambda(T_{new})$ occurs at approximately $\lambda = 5.5$.) This means that T_{new} is positively biased by at most a very small amount, and so suggests that a test based on T_{new} will not have significance levels much below their nominal value. That is, this suggests while the test based on T_{new} may be conservative, it will not be radical by very much.

The results in Figs 5 and 6 suggest that the distribution of Y may effectively be very close to normal. As further exploration of this possibility, note that if Y were exactly normal then we would

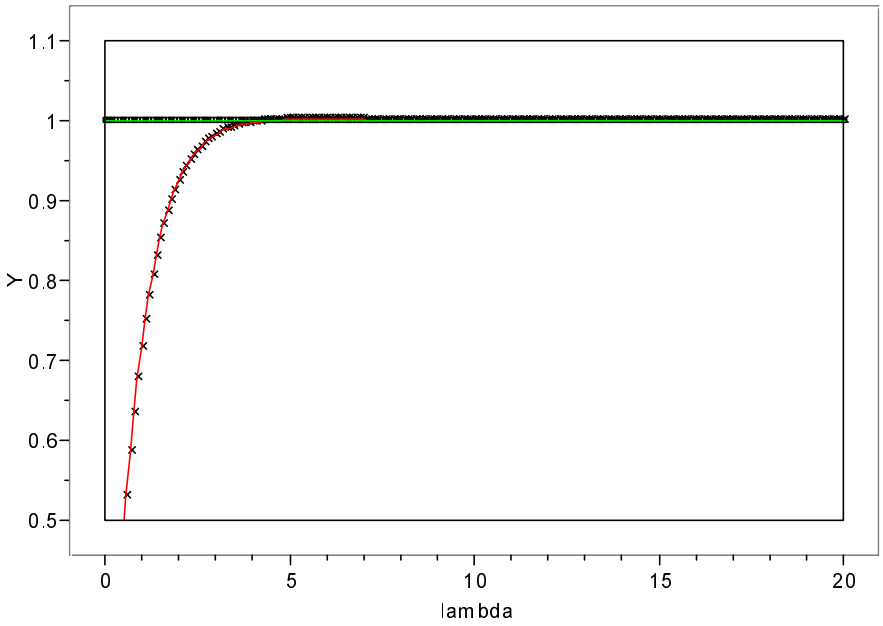

Figure 5: Plot of $E_\lambda(T_{new}/(n-1))$

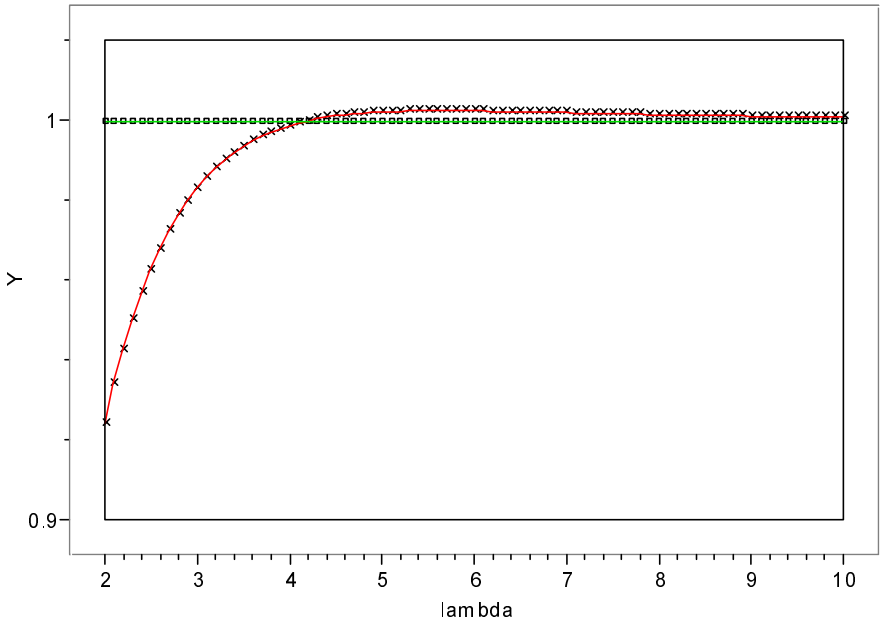

Figure 6: Detail of Fig 5

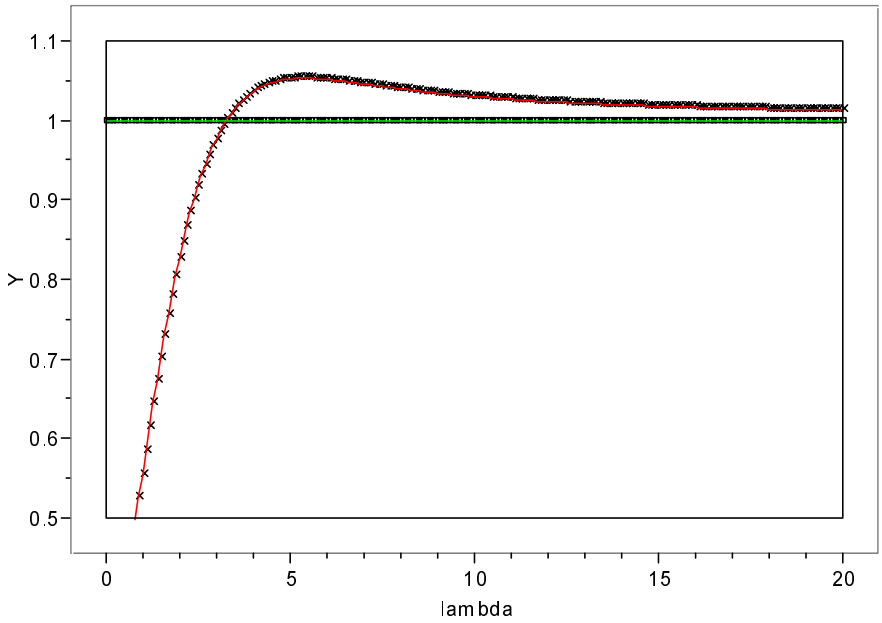

Figure 7: Plot of $\rho(\lambda)$ as defined in (12)

have $\mathrm{Var}((Y-\nu_\lambda)^2) = 2$. Fig 7 is a plot of

$$\rho(\lambda) = \left[\frac{\mathrm{Var}((Y-\nu_\lambda)^2)}{2}\right]^{1/2}. \tag{12}$$

Note that $\rho(\lambda) \approx 1$ whenever $\lambda > 4$. In particular, $\rho(\lambda) \leqslant 1.054$ with the maximum occurring at $\lambda = 5.4$. Again, this suggests that test based on T_{new} will be conservative for very small λ, but will not for any λ be "radical" by very much.

Here is a formal statement of the asymptotic result.

Theorem 5.1 *Assume H_0 is true, λ is fixed and $n \to \infty$. Then*

$$\frac{1}{\rho(\lambda)}\sqrt{\frac{n-1}{2}}\left(\frac{T_{new}}{n-1} - \xi(\lambda)\right) \to N(0,1) \tag{13}$$

in distribution where ξ, ρ are defined in (11) and (12).

Remarks: Recall that if $Z \sim \chi^2_{n-1}$ then

$$\sqrt{\frac{n-1}{2}}\left(\frac{Z}{n-1}-1\right) \to N(0,1).$$

Note that both $\xi(\lambda) \approx 1$ and $\rho(\lambda) \approx 1$. It is thus clear that for large n T_{new} is reasonably closely approximated as a χ^2_{n-1} variable, even though its asymptotic distribution is not exactly χ^2_{n-1} as $n \to \infty$ for fixed λ.

If one were in a situation where n is moderately large and λ is small then (13) suggests that the nominal χ^2 critical values (and P-values) can be slightly improved by calculating critical values (and P-values) from the normal distribution in (13) calculated at $\hat{\lambda} = \overline{X}$.

The formula for the approximate P-value thus becomes

$$P \approx 1 - \Phi^{-1}\left(\frac{1}{\rho(\overline{X})}\sqrt{\frac{n-1}{2}}\left(\frac{T_{new}}{n-1} - \xi(\overline{X})\right)\right). \tag{14}$$

See Table 4 in Section 7 for a data example.

Proof: The result follows from the definition of ξ, ρ, the central limit theorem and Slutsky's theorem. □

Similar reasoning using the central limit theorem for independent non-identically distributed random variables yields

Theorem 5.2 *Let $\lambda_1, \ldots, \lambda_n$ depend on n. Let*

$$\begin{aligned} \overline{\xi}_n &= \frac{1}{n}\sum \xi(\lambda_i) \\ \overline{\rho}_n^2 &= \frac{1}{n}\sum \rho^2(\lambda_i) \\ \psi^2 &= 4\sum(\nu(\lambda_i) - \overline{\nu}_n)^2 \\ \overline{\nu}_n &= \frac{1}{n}\sum \nu(\lambda_i). \end{aligned} \tag{15}$$

Assume

$$\liminf_{n\to\infty} \overline{\rho}_n > 0 \ \ and \ \ \limsup_{n\to\infty} \overline{\rho}_n < \infty. \tag{16}$$

Then

$$\frac{1}{\overline{\rho}_n}\sqrt{\frac{n-1}{2}}\left(\frac{T_{new}}{n-1} - \overline{\xi}_n\right) \to N\left(\frac{4\sum(\nu(\lambda_i) - \overline{\nu}_n)^2}{\overline{\rho}_n\sqrt{2(n-1)}}, 1\right) \tag{17}$$

in distribution as $n \to \infty$.

Proof: The theorem follows from Lindeberg's central limit theorem (Feller (1966, p491)) and Slutsky's theorem. We omit the details, but note that the condition (16) could be considerably weakened. □

It is possible to effectively implement Theorem 5.2 to get values of the power of the test when more accuracy is desired than is provided by (6) and n is quite large. In order to best use (13) and (17) we suggest defining

$$\begin{aligned} \tilde{\overline{\xi}}_n &= \frac{1}{n}\sum_{i=1}^{n} \xi(X_i) \\ \tilde{\overline{\rho}}_n^2 &= \frac{1}{n}\sum_{i=1}^{n} \rho^2(X_i), \end{aligned} \tag{18}$$

since these are the obvious estimates of the corresponding quantities in (15). Then construct the test that rejects when

$$\frac{1}{\tilde{\overline{\rho}}_n}\sqrt{\frac{n-1}{2}}\left(\frac{T_{new}}{n-1}-\tilde{\overline{\xi}}_n\right) > \Phi^{-1}(1-\alpha). \tag{19}$$

Note for later use that under H_0

$$\xi(\overline{X}) \approx \tilde{\overline{\xi}}, \quad \rho^2(\overline{X}) \approx \tilde{\overline{\rho}}^2 \tag{20}$$

with asymptotic equality as $n \to \infty$. (20) should also be approximately valid when the alternative is not far from H_0. In such situations one could use the simpler values $\xi(\overline{X})$, $\rho^2(\overline{X})$ in place of $\tilde{\overline{\xi}}$, $\tilde{\overline{\rho}}^2$. Because of (20) the test in (19) is very similar to that described in (14).

Theorem 5.2 implies the power of the test given in (19) is

$$P_{\boldsymbol{\lambda}}(T_{new} \text{ satisfied (19)}) \to 1 - E\left(\Phi\left(\frac{\tilde{\overline{\rho}}_n}{\overline{\rho}_n}\Phi^{-1}(1-\alpha)\right) + \sqrt{\frac{n-1}{2}}\frac{\tilde{\overline{\xi}}_n - \overline{\xi}_n}{\overline{\rho}_n} - \frac{4\sum(\nu(\lambda_i)-\overline{\nu}_n)^2}{\overline{\rho}_n\sqrt{2(n-1)}}\right), \tag{21}$$

where $\boldsymbol{\lambda} = \{\lambda_i\}$.

Now, $\tilde{\overline{\rho}}_n \to \overline{\rho}_n$ in probability. Also, $\overline{\rho}_n \approx 1$ so long as $\min \lambda_i > 4$ as a consequence of the results plotted in Fig 7.

Let

$$\mathrm{Var}(\sqrt{n-1}(\tilde{\overline{\xi}}_n - \overline{\xi}_n)) = \epsilon(\boldsymbol{\lambda}).$$

Recall that $\xi(\lambda)$ is nearly constant for $\lambda > 4$. Hence ϵ is numerically quite small so long as

$\min \lambda_i > 4$. It follows that then

$$P_{\boldsymbol{\lambda}}(T_{new} \text{ satisfies (19)}) = 1 - E\left(\Phi(\Phi^{-1}(1-\alpha) - \frac{\psi^2}{\sqrt{2(n-1)}} + \epsilon(\boldsymbol{\lambda})) + O_p(1)\right) \quad (22)$$
$$\rightarrow 1 - \Phi(\Phi^{-1}(1-\alpha) - \frac{\psi^2}{\sqrt{2(n-1)}} + \epsilon^*)$$

for some numerically small ϵ^*. (ϵ^* is numerically small because of its relation to the random variable $\epsilon(\boldsymbol{\lambda})$ which is also numerically small.)

If T_{new} were exactly noncentral χ^2 as assumed in (6) then we would have

$$P_{\boldsymbol{\lambda}}(\sqrt{\frac{n-1}{2}}(\frac{T_{new}}{n-1} - 1) > \Phi^{-1}(1-\alpha)) \rightarrow 1 - \Phi(\Phi^{-1}(1-\alpha) - \frac{\psi^2}{2\sqrt{n-1}}), \text{ under (6).} \quad (23)$$

Since ϵ^* is numerically small, these facts suggest that so long as all (or most) $\lambda_i > 4$ (6) is a very good approximation even though it is not asymptotically exact as $n \rightarrow \infty$ with $\overline{\lambda} = O(1)$.

6 Minimum Bayes Factors

Sellke, Bayarri and Berger (2001) ("SBB", below) and Berger (2001) note that small P-values are commonly interpreted to imply greater evidentiary evidence against H_0 than is actually warranted. They propose minimum Bayes factors as an alternative measure. We show in our situation that approximately minimum Bayes factors for testing H_0 versus H_a are readily computable. In our context we agree with SBB, and feel that these factors provide a better evidentiary measure than P-values. In Section 7 these factors are implemented on our call-center arrival data with tests T_{LR}, T_{CC} and T_{new}.

Suppose T is a test statistic such that $T \sim \chi^2_m(\psi^2)$, and we wish to test

$$H_0 : \ \psi^2 = 0 \text{ versus } H_a(R) : \ \psi^2 = R \tag{24}$$

for some fixed $R > 0$. Let f_r denote the noncentral $\chi^2_m(r)$ density. Then the Bayes factor is

$$\frac{f_0(T)}{f_R(T)}. \tag{25}$$

If instead we wish to test H_0 versus

$$H_a : \ \psi^2 > 0 \tag{26}$$

then the minimum Bayes factor (MBF) is

$$MBF = \inf_{\psi^2>0} \frac{f_0(T)}{f_{\psi^2}(T)} = \frac{f_0(T)}{f_{\hat{\psi}^2}(T)} \tag{27}$$

where $\hat{\psi}^2$ is the maximum likelihood estimate of ψ^2 based on T. It is possible to calculate $\hat{\psi}^2$ numerically for any given m and T but we have found that the simple estimate

$$\tilde{\psi}^2 = (T - m)_+ \tag{28}$$

gives very good numerical accuracy in place of $\hat{\psi}^2$ in (27). ($\tilde{\psi}^2$ is the truncated UMVUE of ψ^2.) Over a range of values of m, T we found the error in the MBF calculated from using $\tilde{\psi}^2$ to be well less than 1% of the true value as calculated from $\hat{\psi}^2$.

We refer the reader to SBB and articles cited therein for a discussion of various interpretations of the MBF, and for discussion of why the MBF provides a better evidentiary measure than P-values.

Table 2: P-values and MBF's for H_a of (26)

df	T	P-values	MBF	MBF*
5	10	0.0752	0.402	0.5291
5	15	0.0103	0.0775	0.1287
5	20	0.0013	0.0115	0.0227
5	25	0.0001	0.0015	0.0034
10	15	0.1321	0.5885	0.7268
10	20	0.0293	0.1866	0.2808
10	25	0.0053	0.0431	0.0760
10	30	0.0009	0.0082	0.0165
20	30	0.0699	0.3665	0.5053
20	35	0.0201	0.1346	0.2135
20	40	0.0050	0.0402	0.0720
20	45	0.0011	0.0103	0.0204
20	50	0.0002	0.0023	0.0051
50	70	0.1958	0.0324	0.3019
50	80	0.0045	0.0360	0.0659
50	90	0.0004	0.0044	0.0094

Table 2 illustrates the difference in evidence provided by MBF's and P-values in the general Poisson problem. For a selection of values of df and possible observations T it gives the MBF and P-value for testing (26). SBB also suggest a non-parametric (omnibus) MBF which is

$$MBF^* = -eP\log(P) \tag{29}$$

where P is the conventional P-value. For comparison this value is also included in Table 2. We remark that direct comparison of P-values and MBF's may be facilitated by transforming the MBF to MPP=MBF/(1+MBF). (As explained in SBB, this is the Minimum Posterior Probability of H_0 under the prior giving equal mass to H_0 and some $\{\lambda_i\} \in H_a$.) We do not include MPP in Table 2, but note that for small values of MBF one has MBF $\approx$ MBF / (1 + MBF)= MPP.

7 Empirical Examples

Unless λ is small the statistics T_{new}, T_{LR} and T_{CC} all have approximately non-central chi-squared distributions, as discussed in Section 3. Hence their P-values and MBF's can readily be calculated for given data sets.

Table 3 compares these quantities as calculated for the data shown in Fig 2 where λ is estimated by $\overline{X} = 18.66$, which is not small.

Table 3: P-values and MBF's for the data in Fig. 2

Statistic	Value	P-value	MBF
T_{new}	63.79	0.0213	0.1393
T_{LR}	62.93	0.0253	0.1604
T_{CC}	59.80	0.0457	0.2592

Note that among the three tests here T_{new} yields the smallest P-value and MBF, and hence yields the most significant result. (Since λ is clearly not small here we may consider all three test as being valid in the sense that their significance levels are approximately correct under H_0.)

For the data in Fig. 1 the estimate of λ is $\overline{X} = 2.18$. The values of T_{new}, T_{LR} and T_{CC} for this data are given in Table 4. The next column of the table gives the nominal P-value based on the assumption that the null distributions are χ^2_{106}.

The discussion in Section 6 establishes that this nominal null distribution is not accurate for T_{new} with $\lambda \approx 2.18$. Instead, that discussion suggests (14) as a reasonable alternative. The column of Table 4 headed "Asymptotically corrected P-value" gives the result of (14) for T_{new}.

The very different nominal P-values calculated for T_{LR} and T_{CC} suggests that one or both of these do not actually possess the nominal χ^2_{106} distribution under H_0, in spite of the fact that $n = 107$ is fairly large.

Table 4: Chi-squared P-values and corrected P-values for the data in Fig. 1

Statistic	Value	χ^2 P-value	Asymptotically corrected P-value	Empirically corrected P-value
T_{new}	110.86	0.354	0.152	0.19
T_{LR}	134.13	0.034		0.18
T_{CC}	120.15	0.164		0.17

To investigate this possibility we performed a simulation giving 5000 samples of size $n = 107$ from the Poisson ($\lambda = 2.18$) distribution. We compute for each sample the value of T. This gives an empirical estimate of the null distribution of T at $\lambda = 2.18$ and $n = 107$. From this we can determine an empirical P-value corresponding to the actual observed value of T. These empirically corrected P values are given in the last column of Table 4.

Note that the empirically corrected P-value for T_{LR} is very different than the χ^2 P-value. For T_{new} the two corrected P-values are in reasonable (but not perfect) agreement. For T_{CC} the χ^2 P-value appears quite adequate. The process of calculating empirically corrected P-values in the above manner is one standard form of bootstrapped test procedure. See for example Politis, Romano and Wolf (1999, p. 37).

It is interesting to note that the empirical null distributions produced by the simulations for Table 4 turn out to be well approximated by both normal and Gamma distributions.

Table 5: Simulated null distributions under Poiss ($\lambda = 2.18$). (From 5000 repetitions.)

Statistic	Normal	Gamma
T_{new}	$\mu = 99.7$, $\sigma^2 = 163.76$	$r = 60.4$, $s = 1.65$
T_{LR}	$\mu = 120.0$, $\sigma^2 = 251.2$	$r = 57.1$, $s = 2.10$
T_{CC}	$\mu = 106.1$, $\sigma^2 = 216.1$	$r = 52.4$, $s = 2.03$ (approx χ^2_{105})

The entries in Table 5 give the parameters of the normal and gamma distribution that provide

the best fit to the simulated null distribution. In each case in Table 5 the normal approximation is a reasonably good fit to the simulation distribution but the Gamma approximation is an even better fit. (The parameters in this table correspond to a gamma density written in the form $g_{r,s}(x) = C_{r,s}x^{r-1}\exp(-x/s)$.) We can also compute the P values from the gamma distributions in Table 5. They are essentially the same as those reported in Table 4.

Note that the asymptotic normal distribution (13) for T_{new} that was used in Table 4 has mean = 99.8 and variance = 159. This asymptotic approximation therefore agrees quite well with the empirical normal distribution found in Table 5. The empirical gamma distribution for T_{CC} is approximately χ^2_{105} which agrees almost perfectly with the nominal χ^2_{106} distribution for this statistic.

We do not give MBF's in Table 4. There are two reasons for this. From a practical perspective they are not needed; since the P-values are not small, the MBFs will also not be small. Also, since the statistic T_{LR} clearly does not have its nominal limiting distribution here, one would apparently need to calculate its MBFs from a simulation. Perhaps the MBF for T_{CC} could reasonably be calculated from its asymptotic non-central chi-squared distribution. For T_{new} one could use the normal approximation in Section 4 to construct an MBF as described in SBB (2001).

References

[1] Agresti, A. (1990). *Categorical Data Analysis*, Wiley and Sons.

[2] Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35**, 246–254.

[3] Berger, J. (2001). Could Fisher, Jeffreys and Neyman have agreed upon testing. Fisher lecture, Atlanta JSM.

[4] Brown, L.D., Cai, T. and DasGupta, A. (2000). Interval estimation in exponential families. Technical Report.

[5] Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2001). Empirical analysis of a telephone call center. Technical report.

[6] Brown, L. D., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2001). Multifactor Poisson and Gamma-Poisson models for call center arrival times. Technical report.

[7] Brown, L.D., Zhang, R. and Zhao L. (2001). Root un-root methodology for nonparametric density estimation.

[8] Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, **10**, 417-451.

[9] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol II. Wiley and Sons.

[10] Huffman, M.D. (1984). An improved approximate tow-sample Poisson test. *Appl. Statist.*, bf 33, 224-226.

[11] Jongbloed, G. and Koole, G. (2000). Managing uncertainty in call centers using Poisson mixtures. Technical report.

[12] Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS regional conference series in probability and statistics, Vol. 5.

[13] Politis, D. N., Romano, J.P., Wolf, M. (1999). *Subsampling*, Springer.

[14] Rice, J. (1995). *Mathematical Statistics and Data Analysis*, second edition, Duxbury press.

[15] Sellke, T., Bayarri, M.J. and Berger, J. (2001). Calibration of P-values for testing precise null hypotheses. *The American Statistician*, **55**, 62–71.