A Poissonness Plot

David C. Hoaglin

The American Statistician, Vol. 34, No. 3 (Aug., 1980), 146-149.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198008%2934%3A3%3C146%3AAPP%3E2.0.CO%3B2-X

The American Statistician is currently published by American Statistical Association.

A Poissonness Plot

DAVID C. HOAGLIN*

A graphical technique, similar in spirit to probability plotting, can be used to judge whether a Poisson model is appropriate for an observed frequency distribution. This "Poissonness plot" can equally be applied to truncated Poisson situations. It provides a type of robustness for detecting isolated discrepancies in otherwise well-behaved frequency distributions.

KEY WORDS: Binomial distribution; Data analysis; Household size; Poisson distribution; Probability plotting; Robustness.

1. INTRODUCTION

In analyzing data possibly describable as Poisson, we need a simple graphical device to determine whether the Poisson distribution is an appropriate model. Except for an isolated large count or a substantial discrepancy in one cell, a Poisson model might give a good fit to the data. The data analyst needs to detect these and other simple departures quickly, without actually having to fit the Poisson model, and then to estimate the Poisson parameter in a way that is not sensitive to them. This article derives a simple "Poissonness plot," illustrates its application to Poisson and truncated Poisson data, and discusses robustness and some related plotting techniques.

2. DERIVATION

The most common form of probability plot, the quantile-quantile plot (Wilk and Gnanadesikan 1968), compares data quantiles and the corresponding standard values. It yields a straight-line pattern when the data come from the assumed location-scale family of continuous distributions (e.g., the normal distributions). Because the Poisson distributions form a one-parameter family of discrete distributions instead of a two-parameter family of continuous distributions, the situation is not the same, but it is still possible to construct a plot in which the standard of comparison is a straight line.

Let $x_0, x_1, x_2, \ldots$, denote the counts of the observed frequency distribution (i.e., x_0 of the observations are 0, x_1 are 1, and so on), and let $N = x_0 + x_1 + x_2 + \ldots$. We recall that the Poisson probability function is

$$P_\lambda\{X = k\} = p_\lambda(k) = e^{-\lambda}\lambda^k/k!, \quad k = 0, 1, 2, \ldots. \tag{2.1}$$

For a sample of N, the expected frequencies are

$$m_k = Np_\lambda(k) = Ne^{-\lambda}\lambda^k/k!, \quad k = 0, 1, 2, \ldots. \tag{2.2}$$

To derive the plot, we assume that, for some fixed value of λ, each observed frequency, x_k, equals the expected frequency, m_k. Then taking natural logarithms on both sides of (2.2) gives

$$\log(x_k) = \log(N) - \lambda + k\log(\lambda) - \log(k!), \tag{2.3}$$

and it is evident that plotting $\log(x_k) + \log(k!)$ against k will yield a straight line with slope equal to $\log(\lambda)$ and intercept equal to $\log(N) - \lambda$. (If $x_k = 0$, no point is calculated or plotted.)

If this "Poissonness plot" for a set of data is satisfactorily straight, we may use the maximum likelihood estimator of λ

$$\hat{\lambda} = \Sigma k x_k/N. \tag{2.4}$$

Otherwise, we may estimate λ from the slope of a line fitted to the plot. If the plot is straight, except for a very few discrepant points, we can downweight those points in fitting a line by eye, or we can use the "resistant line" (Tukey 1977; Velleman and Hoaglin 1980) or a suitable technique for simple robust regression (e.g., Andrews 1974 or Mosteller and Tukey 1977, Ch. 14). How we fit a line will depend primarily on what procedure is available. The aim is to get a convenient estimate of λ from the plot (as one would estimate μ or σ from a normal probability plot) and then perhaps look at residuals from the fitted line, rather than to estimate λ in any formal sense. Of course, wherever possible, one would try to discover an explanation for any discrepant points. One benefit of the Poissonness plot is that it makes such points less likely to escape detection.

An extreme form of discrepancy in a single cell arises when data values equal to zero cannot be observed. For example, a Poisson model may be appropriate for some data on sizes of households or sizes of social groups, but all such groups will necessarily have at least one member. For such data, the "positive Poisson distribution," whose probability function is

$$p_\lambda^*(k) = p_\lambda(k)/(1 - e^{-\lambda}), \quad k = 1, 2, \ldots, \tag{2.5}$$

is used. Since the renormalization affects only the intercept in the Poissonness plot, no change in the plotting technique is involved. Similarly, the plot also serves for more general truncated Poisson distributions, which omit other values on the left besides zero (or on the right beyond a certain point). The simple plot may be all the more valuable in these situations because iterative numerical calculation is required to estimate λ (see Johnson and Kotz 1969, Sec. 4.10). If the plot reveals poor agreement between the data and the Poisson

* David C. Hoaglin is Senior Scientist, Abt Associates Inc., 55 Wheeler Street, Cambridge, MA 02138, and Research Associate, Department of Statistics, Harvard University, Cambridge, MA 02138. The author is grateful to Stephen E. Fienberg for pointing out the related work by Gart, to Peter Fortini for discussions that led to the binomial plot, and to William G. Cochran, Miriam Gasko-Green, Paul W. Holland, Frederick Mosteller, Anita Parunak, Michael A. Stoto, and John W. Tukey for helpful comments and suggestions. This work was supported in part by NSF Grant SOC75-15702 to Harvard University. The analysis of the data from the Housing Allowance Demand Experiment was carried out primarily under Contract H-2040R between the U.S. Department of Housing and Urban Development and Abt Associates Inc.

1. *Scintillations From Radioactive Decay of Polonium, With Calculations for Poissonness Plot*

k	x_k	log (x_k)	log ($k!$)	log (x_k) + log ($k!$)
0	57	4.04	0	4.04
1	203	5.31	0	5.31
2	383	5.95	0.69	6.64
3	525	6.26	1.79	8.06
4	532	6.28	3.18	9.46
5	408	6.01	4.79	10.80
6	273	5.61	6.58	12.19
7	139	4.93	8.52	13.46
8	45	3.81	10.60	14.41
9	27	3.30	12.80	16.10
10	10	2.30	15.10	17.41
11	4	1.39	17.50	18.89
12	0	—	19.99	—
13	1	0	22.55	22.55
14	1	0	25.19	25.19
N = 2608				

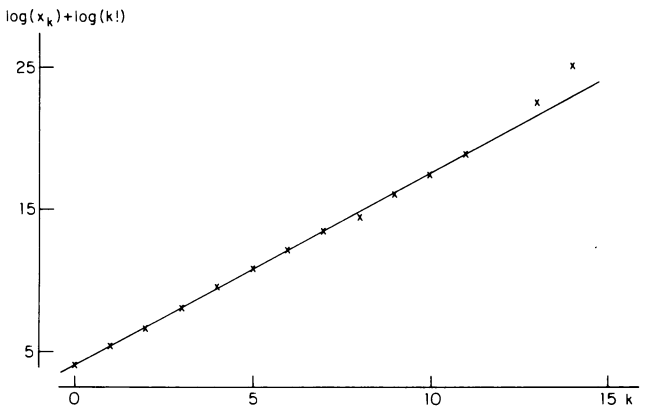
model, one can bypass the iterative calculations and turn immediately to other models.

We now illustrate the Poissonness plot with three examples.

3. EXAMPLES

Rutherford and Geiger (1910) presented data on the number of scintillations in 1/8-minute intervals caused by radioactive decay of polonium. Table 1 shows their data and the calculations for the Poissonness plot, which appears in Figure A. It is clear that this set of data follows a Poisson distribution closely: Only the rightmost point departs substantially from a straight line, and even then the discrepancy is not great, especially since it comes from $x_{14} = 1$. The point for $k = 8$ is somewhat low, and it may deserve further attention.

This raises the question of variability. When the plot is close enough to a straight line, how far should log (x_k) be allowed to depart before being considered discrepant? Naturally, these limits vary with λ and N. Each x_k can be regarded as an observation from a binomial distribution with parameters N and $p_k = p_\lambda(k)$. For large N, log (x_k) can be compared to a normal



A. Poissonness Plot for Polonium Scintillation Data

2. *Occurrences of the Word* May *in Papers by Madison in* The Federalist

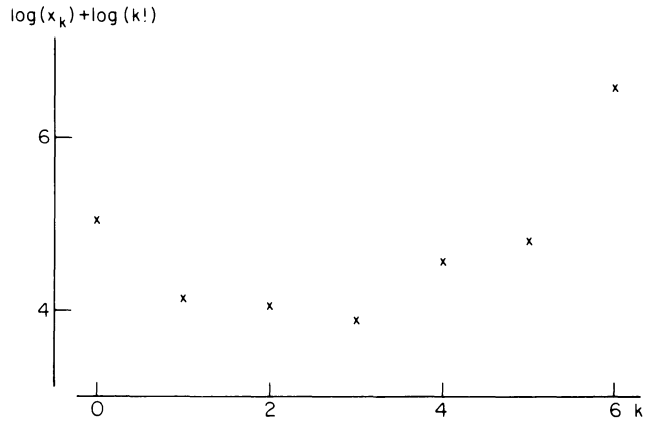
k	x_k	log (x_k)	log (x_k) + log ($k!$)
0	156	5.05	5.05
1	63	4.14	4.14
2	29	3.37	4.06
3	8	2.08	3.87
4	4	1.39	4.56
5	1	0	4.79
6	1	0	6.58
N = 262			

distribution with mean log (Np_k) and variance $(1 - p_k)/(Np_k)$. When we take variability into account in this way and estimate λ as 3.877 from the slope of a resistant line[1] fitted to Figure A, we find that the standard deviation for log (x_{14}) is 3.05. Thus that point lies only .73 standard deviations above the fitted line. For log (x_8), on the other hand, the standard deviation is .119, and this point is low by 3.51 standard deviations, a substantial amount even when we recognize that we have focused on x_8 because it seemed discrepant.

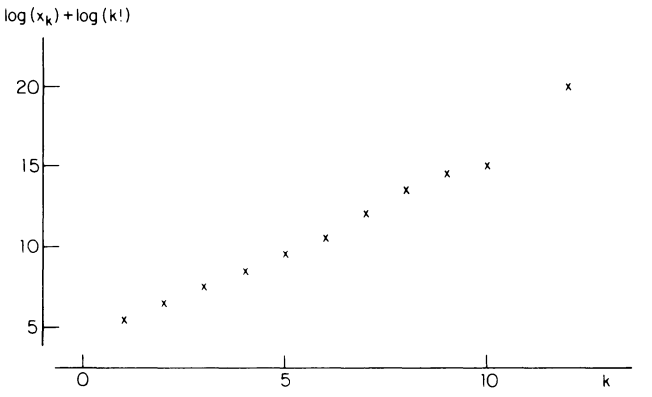
For a second example we turn to the number of occurrences of the word *may* in 262 blocks of text (each approximately 200 words in length) in those papers in *The Federalist* that were written by James Madison. The data come from Mosteller and Wallace (1964, p. 33). Table 2 shows the calculations, and Figure B is the plot. The clear curvature suggests that a Poisson distribution cannot provide a good fit to these data. (Mosteller and Wallace obtained a much better fit by using a negative binomial distribution.)

The third example involves a truncated Poisson distribution. The Housing Allowance Demand Experiment (Bakeman, Kennedy, and Wallace 1977), part of the Experimental Housing Allowance Program established by the U.S. Department of Housing and Urban Development, examined the effects of providing housing subsidy payments directly to low-income households as an alternative to subsidized housing. Table 3 shows the frequency distribution of household size (i.e., the number of persons in the household) at enrollment for the 1,239 households in Allegheny county, Pennsylvania, that were still participating in the program at the end of two years. In Figure C we see that the leftmost six points lie close to a straight line, but the points for household sizes from 7 to 10 are somewhat above that line, and the one household with 12 members clearly does not fit well with the rest of the sample. On the whole, the evidence in Figure C is adequate to suggest that a positive Poisson model would be a satisfactory starting point in analyzing these

[1] The resistant line is based on separating a set of (x,y) data into three equal-sized groups according to the x values, summarizing each group by the point (median x, median y), calculating the slope from the summary points for the first and third groups, and using all three summary points to determine the intercept. Velleman and Hoaglin (1980) give full details of the procedure. The line fitted to Figure A is $y = 1.355x + 3.99$.



B. Poissonness Plot for The Federalist *Data*



C. Poissonness Plot for Household Size Data

household size data. And if one wanted to fit a truncated Poisson model to the data for $k = 1$ through $k = 6$, the estimate of λ calculated from the slope of that portion of Figure C would be an excellent initial value for the maximum likelihood calculation. A desirable parallel step would be to look closer at the data on the 63 individual households of sizes 7 through 10. Their composition may indicate that some of them were formed through a process different from that for smaller households.

4. A PLOT FOR BINOMIAL DATA

The approach that led to the Poissonness plot in Section 2 can also be used for data from a binomial distribution (and, in fact, for any one-parameter exponential family of discrete distributions). For fixed n and p, the expected frequencies in a binomial sample of N are

$$N\binom{n}{k}p^k(1-p)^{n-k} = N\binom{n}{k}(1-p)^n\left(\frac{p}{1-p}\right)^k,$$

and one plots $\log(x_k) - \log\binom{n}{k}$ against k, looking for a straight line with slope $\log(p/(1-p))$.

Applications of this binomial plot are likely to be rarer than for the Poissonness plot because n must be constant. Meier and Zabell (1980), however, give one interesting data set to which it could be applied. The data, analyzed as binomial by Benjamin Peirce, come from comparing all possible pairs among 42 specimens of a woman's signature and counting (for each pair) the number of matches in downstrokes among 30 possible downstrokes.

3. Household Size at Enrollment

k	x_k	$\log(x_k)$	$\log(x_k) + \log(k!)$
1	210	5.35	5.35
2	315	5.75	6.45
3	292	5.68	7.47
4	176	5.17	8.35
5	125	4.83	9.62
6	57	4.04	10.62
7	38	3.64	12.16
8	18	2.89	13.50
9	6	1.79	14.59
10	1	0	15.10
11	0	—	—
12	1	0	19.99
	$N = 1239$		

5. DISCUSSION

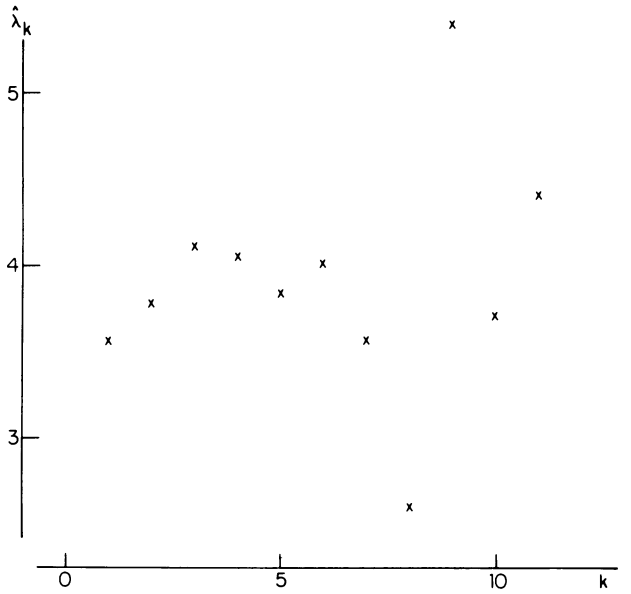
The three examples, together with a number of others not included here, indicate that the Poissonness plot can serve quite well as a diagnostic tool. It indicates whether a Poisson model will fit the data, and it can handle truncated Poisson distributions with no additional effort. Of course, if the sample size is relatively small, the plot may provide rather uncertain guidance, but the same is true for the customary quantile-quantile plots.

If desired, use of the Poissonness plot could be made even simpler (at least for samples that do not include too large a value of k) by designing a special-purpose graph paper.

Another type of plot can be used to diagnose Poissonness. Gart (1970) uses the fact that $kp_\lambda(k)/p_\lambda(k-1) = \lambda$ as a basis for calculating $\hat{\lambda}_k = kx_k/x_{k-1}$ and plotting $\hat{\lambda}_k$ against k for $k = 1, 2, \ldots$. Rao (1971) includes a

4. Values of $\hat{\lambda}_k$ for the Polonium Data

k	x_k	$\hat{\lambda}_k$
0	57	3.561
1	203	3.773
2	383	4.112
3	525	4.053
4	532	3.835
5	408	4.015
6	273	3.564
7	139	2.590
8	45	5.400
9	27	3.704
10	10	4.400
11	4	(0)
12	0	—
13	1	(14.000)
14	1	



D. Plot of $\hat{\lambda}_k$ for the Polonium Scintillation Data

similar technique, plotting $p_\lambda(k)/p_\lambda(k+1)$ against k. Table 4 shows the values of $\hat{\lambda}_k$ for the polonium scintillation data, and Figure D is the plot. When the data are Poisson, it is useful to have a horizontal line as the reference pattern because deviations show up more clearly; but one discrepant value of x_k will affect both $\hat{\lambda}_k$ and $\hat{\lambda}_{k+1}$. This lack of robustness can give a misleading picture of the location and extent of departures from Poissonness, as Figure D shows. There, $\hat{\lambda}_8$ is unusually low, and $\hat{\lambda}_9$ is unusually high. However, the Poissonness plot in Figure A (supplemented by calculations taking variability into account) indicates that the only substantial discrepancy comes from x_8.

A further difficulty in using the plot of $\hat{\lambda}_k$ arises when $x_{k-1} = 0$ and, to a lesser extent, when $x_k = 0$. On the other hand, if one wants to consider a negative binomial model as an alternative to the Poisson model, it is very useful that the plot of $\hat{\lambda}_k$ against k will indicate this by following a straight line whose slope is between 0 and 1. On balance, an effective strategy would be to use the Poissonness plot of Section 2 initially and then to calculate and plot $\hat{\lambda}_k$ if there is a suggestion of negative binomial behavior (as in Figure B).

[*Received June 1979. Revised March 1980.*]

REFERENCES

Andrews, David F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics*, 16, 523–531.

Bakeman, Helen D., Kennedy, Stephen D., and Wallace, James (1977), *Fourth Annual Report of the Housing Allowance Demand Experiment*, Cambridge, Mass.: Abt Associates Inc.

Gart, John J. (1970), "Some Simple Graphically Oriented Statistical Methods for Discrete Data," in G.P. Patil (ed.), *Random Counts in Scientific Work*, Vol. 1 *Random Counts in Models and Structures*, University Park, Pa.: The Pennsylvania State University Press.

Johnson, Norman L., and Kotz, Samuel (1969), *Distributions in Statistics: Discrete Distributions*, Boston: Houghton Mifflin Co.

Meier, Paul, and Zabell, Sandy (1980), "Benjamin Peirce and the Howland Will," *Journal of the American Statistical Association*, 75, forthcoming.

Mosteller, Frederick, and Tukey, John W. (1977), *Data Analysis and Regression*, Reading, Mass.: Addison-Wesley Publishing Co.

Mosteller, Frederick, and Wallace, David L. (1964), *Inference and Disputed Authorship: The Federalist*, Reading, Mass.: Addison-Wesley Publishing Co.

Rao, G.V. (1971), "A Test for the Fitting of Some Discrete Distributions," *Publications de l'Institut de Statistique de l'Université de Paris*, 20, 121–128.

Rutherford, E., and Geiger, H. (1910), "The Probability Variations in the Distribution of α Particles," *Philosophical Magazine*, Sixth Ser., 20, 698–704.

Tukey, John W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley Publishing Co.

Velleman, Paul F., and Hoaglin, David C. (1980), *Applications, Basics, and Computing of Exploratory Data Analysis*, North Scituate, Mass.: Duxbury Press, forthcoming.

Wilk, M.B., and Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–17.