

# Models for Counts

## Political Science 150C/350C

Simon Jackman

Revision 132

April 30, 2007

### 1 Mathematical and Statistical Background

A random variable  $X$  is said to have a *Poisson* distribution with parameter  $\lambda$  if

$$f(x; \lambda) = \Pr[X = x; \lambda] = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots, \text{ and } \lambda > 0 \quad (1)$$

This distribution takes its name from the famous French mathematician Denis Poisson. Poisson (1837) considered the limit of a sequence of binomial distributions with

$$p_{x,n} = \Pr[X = x] = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \dots, n, \\ 0 & \text{for } x > n \end{cases}$$

where  $n \rightarrow \infty$  and  $p \rightarrow 0$ , while  $np = \lambda < \infty$ . The Poisson distribution in equation (1) follows from noting that as  $n \rightarrow \infty$ ,

$$\frac{n!}{(n-x)!} \approx n^x,$$

and

$$(1-p)^{n-x} = \left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}.$$

With these two asymptotically-valid approximations, the binomial probability density given above becomes

$$\begin{aligned} p_{x,n} &\approx \frac{n^x}{x!} p^x e^{-\lambda} \\ &\approx \frac{n^x}{x!} \left(\frac{\lambda}{n}\right)^x e^{-\lambda} \\ &\approx \frac{\lambda^x}{x!} e^{-\lambda}. \end{aligned}$$

which gives us the Poisson density

$$\lim_{\substack{n \rightarrow \infty \\ np = \lambda}} \sum_w p_{x,n} = \sum_w \frac{e^{-\lambda} \lambda^x}{x!}, \quad (2)$$

where  $\sum_w$  denotes summation over any subset  $w$  of the nonnegative integers  $0, 1, 2, \dots$

Some further insight into this result is gained by recalling that in a binomial experiment the number of successes has an upper-bound equal to the number of trials. This constraint is removed in the Poisson case. Note that

$$1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \quad (3)$$

converges to  $e^\lambda, \forall \lambda$ . Now given the Poisson density defined above, we can see that the  $e^{-\lambda}$  argument plays the role of a normalizing constant, ensuring that

$$\sum_x f(x; \lambda) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^\lambda = 1$$

and so  $f(x)$  behaves like a proper discrete probability distribution, summing to one over its support points, and being non-negative everywhere (since  $\lambda > 0 \Rightarrow f(x) \geq 0$ ); see [Berry and Lindgren \(1996, § 4.6\)](#). Without the normalizing constant each term in the series in (3) can be conceptualized as a *frequency* (instead of a probability), approximating the number of times we observe  $x$  successes from  $n$  trials where the expected number of successes is  $np = \lambda$ , but remembering we consider what happens as  $n \rightarrow \infty$  and  $p \rightarrow 0$ .

[Johnson, Kotz and Kemp \(1992, 152\)](#) note that this result had been derived earlier by [de Moivre \(1711\)](#); they also note that [Bortkiewicz \(1898\)](#) considered situations in which Poisson's distribution might arise (e.g., the number of deaths from kicks by horses, per annum, in the Prussian Army Corps). The critical features for the applicability of Poisson's distribution were that  $N$  be large but that  $p$  be small, *in addition* to the requirements that each trial be independent and that  $p$  be constant over trials.

A number of other derivations followed in the late 19th century and early 20th century. For instance, "Student" (W.S. Gosset, who derived the  $t$  distribution) ([1907](#)) used the Poisson distribution to represent (as a first approximation) the number of particles falling a small area  $A$  when a large numbers of such areas are spread at random over a surface large in comparison with  $A$ . [Rutherford and Geiger \(1910\)](#) provided some data on the emission of radioactive particles that was fitted well by a Poisson distribution; this remains one of the classic areas of application of Poisson models (e.g., [Feller 1968, 160](#)). Other examples include models of arrival times (e.g., customers at stores), when phone calls are made (e.g., [Erlang 1909](#)), electrical demand, accidents in factories or traffic accidents; models of these processes sometimes go under the label of "queueing theory". Other examples include models of mutation in cells, or waiting times between spontaneous impulses in nerve cells.

[Johnson, Kotz and Kemp \(1992, 154--156\)](#) summarize a number of other derivations of the Poisson distribution. See also the book-length treatment by [Haight \(1967\)](#) or the treatment in [Taylor and Karlin \(1998\)](#). These contemporary derivations usually start with a *Poisson point process*. Here I consider a special case of a Poisson point process, a *homogeneous Poisson point process in one dimension*.<sup>1</sup> Following [Tuckwell \(1988, 38ff\)](#), let  $t$  represent a

---

<sup>1</sup>Two and higher dimensional processes are used as the basis of models of ecological patterns, and in geology, geography, and urban studies; for instance, ecologists refer to "Poisson forests" (e.g., [Tuckwell 1988, 45-46](#)). More generally, [Ripley \(1981\)](#) is an excellent reference for models of spatial processes.

time variable (although this could also be a spatial interval). Let events occur randomly over time, starting with  $t = 0$ , and with the first event occurring at time  $T_1$ , the second at  $T_2$ , and so on. Let  $(s_1, s_2]$  be a subinterval of the interval  $[0, s]$ , where  $s < \infty$ . If  $N(s_1, s_2)$  is the number of events in  $(s_1, s_2]$ , then the count variables  $N(s_1, s_2)$  is a random variable, and the collection of all such random variables (defined with respect to any subsets of  $[0, s]$ ), abbreviated to  $N$ , is called a *point process* on  $[0, s]$ . Specifically,

$N$  is a *homogeneous Poisson point process* with rate  $\lambda$  if:

1. for any  $0 \leq s_1 < s_2 \leq s$ ,  $N(s_1, s_2)$  is a Poisson random variable with parameter  $\lambda(s_2 - s_1)$ ;
2. for any collection of times  $0 \leq s_0 < s_1 < s_2 < \dots < s_n \leq s$ , where  $n \geq 2$ , the random variables  $\{N(s_{x-1}, s_x), x = 1, 2, \dots, n\}$  are mutually independent.

Or in words: the number of events in the time interval  $(0, t]$ , which we can denote as  $N(t)$ , is a Poisson random variable with parameter  $\lambda t$ . And the number of points falling in disjoint time intervals are independent.

The expected value of  $N(t)$  is  $\lambda t$  and so the expected number of events in a time period of unit length is  $\lambda$ . For this reason the  $\lambda$  parameter is often referred to as the *intensity* of the Poisson point process. Finally, the Poisson point process is said to be *homogeneous* because the expected number of events depends only on the length of the time interval, not on its location of the interval on the time line (i.e., the intensity of the process is constant over the interval  $(0, t]$ ).

The way we move from a Poisson point process to a Poisson model is relatively straightforward, mirroring the limiting behavior of the binomial distributions noted by Poisson and de Moivre. We assumed that points (times of events) are “sprinkled randomly” (Tuckwell 1988, 41) on the interval  $[0, s]$  under the assumptions:

1. the numbers of points in disjoint subintervals are independent
2. the probability of finding a point in a very small subinterval is proportional to its length, whereas the probability of finding more than one point is negligible

Now divide  $[0, s]$  into  $n$  small subintervals each of length  $\Delta s = s/n$ . Then the probability  $p$  that a given subinterval contains a point is  $\lambda s/n$  where  $\lambda$  is a positive constant. The probability that there are  $x$  occupied subintervals is

$$\begin{aligned} \Pr(x \text{ points in } [0, s]) &= \text{Binomial}(x|n, p) \\ &= \text{Binomial}\left(x|n, \frac{\lambda s}{n}\right) \end{aligned}$$

As  $n \rightarrow \infty$ ,  $\lambda s/n \rightarrow 0$  and the Poisson distribution becomes an appropriate characterization for the data, using the results of Poisson and de Moivre noted earlier. Formally, we have

$$\lim_{n \rightarrow \infty} \text{Binomial}(x|n, p) \xrightarrow{d} \frac{\exp(-np)(np)^x}{x!}.$$

But  $np = n(\lambda s)/n = \lambda s$ , and so as  $n \rightarrow \infty$ ,

$$\Pr(x \text{ points in } [0, s]) = \frac{\exp(-\lambda s)(\lambda s)^x}{x!}$$

which is the Poisson distribution derived above.

A slightly more formal derivation<sup>2</sup> proceeds by letting  $g(x, w)$  denote the probability of  $x$  events in each time interval of length  $w$ . Once again, assume that

1.  $g(1, h) = \lambda h + o(h)$ , where  $\lambda$  is a positive constant and  $h > 0$  and  $o(h)$  is a function such that  $\lim_{h \rightarrow 0} [o(h)/h] = 0$
2.  $\sum_{x=2}^{\infty} g(x, h) = o(h)$ ; i.e., the probability of two or more events in the interval  $h$  is arbitrarily close to zero.
3. The numbers of events in disjoint intervals are independent

These assumptions repeat the conditions stated less formally, earlier. Assume also that  $g(0, 0) = 1$ . Subject to these assumptions, the probability of at least one event in an interval of length  $h$  is the probability of one event plus the probability of two events or  $\lambda h + o(h) + o(h) = \lambda h + o(h)$ . Since probabilities sum to one, the probability of zero events in an time interval of length  $h$  is  $1 - \lambda h + o(h)$ .

Consider the probability of zero events in the time interval  $w + h$ . By the assumption of independence (assumption 3 above), this joint probability is equal to the product of the two individual probabilities

$$\begin{aligned} \Pr[\text{zero events in the interval } w + h] &= g(0, w + h) \\ &= g(0, w)g(0, h) \\ &= g(0, w)[1 - \lambda h - o(h)] \\ &= g(0, w) - \lambda h g(0, w) - o(h)g(0, w) \end{aligned}$$

Re-arranging yields

$$\frac{g(0, w + h) - g(0, w)}{h} = -\lambda g(0, w) - \frac{o(h)g(0, w)}{h}.$$

If we let the incremental time interval  $h$  shrink arbitrarily close to zero, we obtain the differential equation

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{g(0, w + h) - g(0, w)}{h} &= \frac{\partial g(0, w)}{\partial w} \\ &= -\lambda g(0, w) \end{aligned}$$

Solving this differential equation for  $g(0, w)$  requires that we find a function of  $w$  whose derivative is the function times  $-\lambda$ . The only function equal to its own derivative is the exponential function, and so

$$g(0, w) = ce^{-\lambda w}.$$

---

<sup>2</sup>This derivation follows that in (Hogg and Craig 1978, 99-101), and subject to changes of notation is the proof in the Appendix of King (1988).

But the condition  $g(0, 0) = 1$  constrains  $c = 1$ , since we require that  $ce^{-\lambda \times 0} = 1$ .

Now consider the more general case  $g(x, w + h)$ , for  $x = 1, 2, \dots$ . In the time interval  $h$ , two things can happen: we observe no events and so  $x$  remains unchanged from time interval  $w$ , or we observe (at least) one event and so we had  $x - 1$  events at the end of time interval  $w$ . Thus

$$g(x, w + h) = g(x, w)[1 - \lambda h - o(h)] + g(x - 1, w)[\lambda h + o(h)] + o(h)$$

We can re-arrange as before and obtain

$$\frac{g(x, w + h) - g(x, w)}{h} = -\lambda g(x, w) + \lambda g(x - 1, w) + \frac{o(h)}{h}$$

and letting  $h \rightarrow 0$

$$\frac{\partial g(x, w)}{\partial w} = -\lambda g(x, w) + \lambda g(x - 1, w).$$

Consider the case of  $x = 1$ . Then using the result for  $g(0, w)$  from above,

$$\frac{\partial g(1, w)}{\partial w} = -\lambda g(1, w) + \lambda e^{-\lambda w}.$$

Now we are looking for a function  $g(1, w)$  such that its derivative with respect to  $w$  is the function itself times negative  $\lambda$ , plus  $\lambda e^{-\lambda w}$ . This function is

$$g(1, w) = \lambda w e^{-\lambda w}$$

which can be verified using the product rule of differentiation. We can use this result to consider the case of  $x = 2$ , which yields

$$\frac{\partial g(2, w)}{\partial w} = -\lambda g(2, w) + \lambda g(1, w)$$

which after substituting for  $g(1, w)$  yields

$$\frac{\partial g(2, w)}{\partial w} = -\lambda g(2, w) + \lambda^2 w e^{-\lambda w}.$$

The solution of this differential equation for  $g(2, w)$  is

$$g(2, w) = \frac{(\lambda w)^2 e^{-\lambda w}}{2}$$

By induction it can be shown that

$$g(x, w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}$$

for  $x = 1, 2, 3, \dots$ . This is the Poisson distribution given above, and when we normalize the time interval to have unit length ( $w = 1$ ), we obtain the more familiar version of the Poisson density

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for  $x = 0, 1, 2, \dots$

## 2 Regression Models

In analyzing count data, a frequent question of interest is how to relate the observed counts  $\mathbf{y}$  (non-negative integers) to some predictors,  $\mathbf{X}$ . This is usually done by normalizing the observational time unit to have unit length, setting  $s = 1$  in the derivation above. The next step is to find a way to embed the effects of predictors into the Poisson model. This is almost always done by positing a log-linear model for the mean of the Poisson distribution,

$$\ln \lambda_i = \mathbf{x}_i \boldsymbol{\beta}, \quad (4)$$

or

$$\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta}) \quad (5)$$

Alternative functional forms were once considered; for instance [Cochrane \(1940\)](#) considered a linear model and a square-root model  $\lambda_i = (\mathbf{x}_i \boldsymbol{\beta})^2 \iff \sqrt{\lambda_i} = \mathbf{x}_i \boldsymbol{\beta}$ . However, “model appropriateness was emphasized as an important issue; and in this spirit, the product (or log-linear) model was suggested as being of potential interest” ([Koch, Atkinson and Stokes 1986](#), 33).

Why is the log-linear specification somehow more “appropriate” than other specifications? [King \(1988, 841--843\)](#) offers a fairly detailed justification for the “exponential Poisson regression model”. For one thing, any model for the mean of Poisson variable must map onto the non-negative half-line. There is no guarantee that  $\lambda_i > 0$  if we use the familiar additive, linear specification  $\lambda_i = \mathbf{x}_i \boldsymbol{\beta}$ . A non-linear specification is also preferred, since for many applications it seems plausible that “the ‘effort’ (in terms of a change in  $x$ ) that it would take to move  $y_i$  from 0 to 1 should be proportionally greater than the effort required to move  $y_i$  from, say, 20 to 21” ([King 1988, 842-843](#)). If the idea is that the marginal effect of  $x$  should increase with the mean of the Poisson process  $\lambda$ , then the log-linear functional form is well-suited to the task; note that if  $E(y_i | \mathbf{x}_i) = \lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$  then

$$\begin{aligned} \frac{\partial y_i}{\partial \mathbf{x}_i} &= \frac{\partial \exp(\mathbf{x}_i \boldsymbol{\beta})}{\partial \mathbf{x}_i} \\ &= \lambda_i \boldsymbol{\beta}, \end{aligned}$$

since the exponential function is the only function that is its own derivative.

Whatever model we use for the mean, Poisson regression models are easily fit via maximum likelihood. Assuming (conditional) independence, the log-likelihood for the Poisson model has the relatively simple form:

$$\log L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^n [y_i \log \lambda_i - \lambda_i - \log(y_i!)]$$

where the usual model for the conditional mean is  $\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$ .

## 3 Negative-Binomial

What happens if the Poisson assumption of mean equals variance doesn't hold?  
Consequences:

- Poisson consistent, but inefficient
- Poisson standard errors too small.

Assumption usually doesn't hold due via *overdispersion*: i.e., conditional variance is greater than the conditional mean. This can arise in a number of different ways: unobserved heterogeneity (predictors omitted from the model) or contagion (see Long 1997, 236) *within individuals*, such that the occurrence of an event for individual  $i$  increases the probability of further events for individual  $i$ , even after conditioning on covariates  $\mathbf{x}_i$ . Long (1997, 236) points out that in a cross-section (one observation per individual  $i$ ), there is no way to distinguish unobserved heterogeneity (so-called "spurious contagion") from actual or "true contagion").

See Long (1997, 231) for development of the *negative binomial* model as an alternative to the Poisson. Let  $\varepsilon_i$  be an unobserved source of heterogeneity such that

$$\begin{aligned} E(y_i) &= \tilde{\lambda}_i \\ &= \exp(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i) \\ &= \exp(\mathbf{x}_i\boldsymbol{\beta})\exp(\varepsilon_i) \\ &= \lambda_i\delta_i \end{aligned}$$

where  $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$  is the usual conditional mean from the Poisson model and  $\delta_i = \exp(\varepsilon_i)$ . To identify this expanded conditional mean specification, assume  $E(\delta_i) = 1$ , since the intercept in the  $\mathbf{x}_i\boldsymbol{\beta}$  part of the model could pick up any unknown  $E(\delta_i)$  if it were allowed to vary. This means that the Poisson model and the negative binomial model have the same model for the conditional mean, or formally,

$$E(y_i) = E(\tilde{\lambda}_i) = \lambda_i E(\delta_i) = \lambda_i.$$

And conditional on  $\mathbf{x}_i$ ,  $\boldsymbol{\beta}$  and  $\delta_i$ ,  $y_i$  still has a Poisson distribution: i.e.,

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \delta_i) = \frac{e^{-\lambda_i\delta_i}(\lambda_i\delta_i)^{y_i}}{y_i!},$$

where  $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$ , but this is non-operational since  $\delta_i$  is an unknown random quantity. If we knew the pdf of  $\delta_i$ , say  $g(\delta_i)$ , we could integrate to obtain

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \int_0^\infty f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \delta_i)g(\delta_i)d\delta_i$$

where the lower limit of integration is zero since  $\delta_i = \exp(\varepsilon_i)$ . The usual assumption in the literature is that  $\delta_i$  has a gamma distribution with parameter  $\theta_i$ :

$$g(\delta_i) = \frac{\theta_i^{\theta_i}}{\Gamma(\theta_i)}\delta_i^{\theta_i-1}\exp(-\delta_i\theta_i)$$

where  $\theta_i > 0$  and the gamma function  $\Gamma(\theta) = \int_0^\infty t^{\theta-1}e^{-t}dt = (\theta-1)!$ , the last equality holding for integer values of  $\theta$ . The gamma distribution has the properties: (1)  $E(\delta_i) = 1$ ; (2)  $V(\delta_i) = 1/\theta_i$ .

The negative binomial density arises from performing the integration given above, with a gamma density for  $\delta_i$  and the Poisson density for  $f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \delta_i)$ . In this sense the negative binomial is a gamma mixture of Poisson densities, and is

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \theta_i) = \frac{\Gamma(y_i + \theta_i)}{y_i! \Gamma(\theta_i)} \left( \frac{\theta_i}{\theta_i + \lambda_i} \right)^{\theta_i} \left( \frac{\lambda_i}{\theta_i + \lambda_i} \right)^{y_i}$$

and has expected value identical to that from the Poisson,

$$E(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

but no longer has the Poisson feature that the conditional mean necessarily equals the conditional variance, i.e.,

$$V(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \theta_i) = \lambda_i \left( 1 + \frac{\lambda_i}{\theta_i} \right) = \exp(\mathbf{x}_i\boldsymbol{\beta}) \left( 1 + \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{\theta_i} \right)$$

with the conditional variance equal to the conditional mean if and only if  $\theta_i = \infty$ . The flexibility in the variance lets the model be quite flexible, able to shift more probability mass onto high or low counts conditional on a mean count; see [Long \(1997, Fig 8.6\)](#) for examples.

The variance parameter,  $\theta_i$  (or  $\delta_i$ ) can be modeled as a function of covariates (i.e., if we had some predictors that we thought tapped heterogeneity or contagion, but not the mean count), but is usually held constant and  $\theta$  is an extra parameter to be estimated jointly with the parameters for the mean function,  $\boldsymbol{\beta}$ .

The likelihood function for the negative binomial model is

$$\begin{aligned} L(\boldsymbol{\beta}, \theta|\mathbf{y}, \mathbf{X}) &= \prod_{i=1}^n f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \theta) \\ &= \prod_{i=1}^n \frac{\Gamma(y_i + \theta)}{y_i! \Gamma(\theta)} \left( \frac{\theta}{\theta + \lambda_i} \right)^{\theta} \left( \frac{\lambda_i}{\theta + \lambda_i} \right)^{y_i} \end{aligned}$$

See function `glm.nb` in `library(MASS)` in R for implementation.

As a matter of practice, it is commonplace to test for overdispersion when estimating count models. The negative binomial model is always a “live competitor” against the Poisson model, is easy to estimate and compare with the Poisson model. The difference in the log-likelihoods of a Poisson model and a (less restrictive) negative binomial model can be used to test the two models (but the critical value of the  $\chi^2$  test should be that for twice the nominal significance level, since  $\theta$  is bounded from below by zero).

## 4 Hurdle and Zero-Inflated Count Models

See [Long \(1997, §8.5.2\)](#), [Greene \(2003, §21.9.6\)](#), and [Cameron and Trivedi \(1998, §4.7\)](#). We often have too many zeros. Model this via a branching process. Two approaches:

1. Hurdle model: zeros generated by a different process than the ones, for Poisson case

$$\begin{aligned}
 P(y_i = 0) &= F(\mathbf{z}_i; \boldsymbol{\gamma}) \\
 P(y_i = j) &= \frac{1 - F(\mathbf{z}_i; \boldsymbol{\gamma})}{1 - f_2(y_i = 0; \mathbf{x}_i, \boldsymbol{\beta})} f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}), \quad j > 0 \\
 f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}) &= \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \\
 \lambda_i &= \exp(\mathbf{x}_i; \boldsymbol{\beta})
 \end{aligned}$$

2. Zero-Inflated models: put extra probability mass on the zeros, for Poisson case

$$\begin{aligned}
 \psi_i &= F(\mathbf{z}_i; \boldsymbol{\gamma}) \\
 P(y_i = 0) &= \psi_i + (1 - \psi_i) e^{-\lambda_i} \\
 P(y_i = j) &= (1 - \psi_i) \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad j = 1, 2, \dots \\
 \lambda_i &= \exp(\mathbf{x}_i; \boldsymbol{\beta})
 \end{aligned}$$

- ZIP vs Poisson; ZINB versus NB.
- testing ZIP versus Poisson, or ZINB versus NB. Vuong's non-nested test.

## References

- Berry, Donald and Bernard W. Lindgren. 1996. *Statistics: Theory and Methods*. Second ed. Belmont, California: Duxbury.
- Bortkiewicz, L. von. 1898. *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.
- Cameron, A. Colin and Pravin K. Trivedi. 1998. *Regression analysis of count data*. Cambridge, United Kingdom: Cambridge University Press.
- Cochrane, W. G. 1940. "???" *Annals of Mathematical Statistics* 11:335--347.
- de Moivre, A. 1711. "De Mensura Sorits." *Philosophical Transactions of the Royal Society*, No. 329 27:213--264.
- Erlang, A. K. 1909. "Probability calculus and telephone conversations." *Nyt Tidsskr. Mat. B* 20:33--39. (in Danish).
- Feller, W. 1968. *An Introduction to Probability Theory and its Applications*. Third ed. New York: Wiley.

- Greene, William H. 2003. *Econometric Analysis*. Fifth ed. Upper Saddle River, New Jersey: Prentice Hall.
- Haight, F. A. 1967. *Handbook of the Poisson Distribution*. New York: Wiley.
- Hogg, Robert V. and Allen T. Craig. 1978. *Introduction to Mathematical Statistics*. Fourth ed. New York: Collier Macmillan.
- Johnson, Norman L., Samuel Kotz and Adrienne W. Kemp. 1992. *Univariate Discrete Distributions*. Second ed. New York: Wiley.
- King, Gary. 1988. "Statistical models for political science event counts: bias in conventional procedures and evidence for the exponential Poisson regression model." *American Journal of Political Science* 32:838--863.
- Koch, Gary G., Susan S. Atkinson and Maura E. Stokes. 1986. "Poisson Regression." In *Encyclopedia of Statistical Sciences*, ed. Samuel Kotz and Norman L. Johnson. Vol. 7 New York: Wiley pp. 32--41.
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Number 7 in Advanced Quantitative Techniques in the Social Sciences. Thousand Oaks, California: Sage.
- Poisson, S. D. 1837. *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, Précédées des Règles Générales du Calcul des Probabilités*. Paris: Bachelier, Imprimeur-Librairie pour les Mathématiques, la Physique, etc.
- Ripley, B. D. 1981. *Spatial Statistics*. New York: Wiley.
- Rutherford, R. S. and H. Geiger. 1910. "The probability variations in the distribution of  $\alpha$  particles." *Philosophical Magazine, 6th series* 20:698--704.
- Student. 1907. "On the error of counting with a haemocytometer." *Biometrika* 5:351--360.
- Taylor, Howard M. and Samuel Karlin. 1998. *An Introduction to Stochastic Modeling*. Third ed. San Diego: Academic Press.
- Tuckwell, Henry C. 1988. *Elementary Applications of Probability Theory*. London: Chapman and Hall.