### Discrete Statistical Distributions

October 17, 2007

Discrete random variables take on only a countable number of values. The commonly used distributions are included in SciPy and described in this document. Each discrete distribution can take one extra integer parameter: L. The relationship between the general distribution and the standard one is

$$p\left(x\right) = p_0\left(x - L\right)$$

which allows for shifting of the input. When a distribution generator is initialized, the discrete distribution can either specify the beginning and ending (integer) values a and b which must be such that

$$p_0(x) = 0$$
  $x < a \text{ or } x > b$ 

in which case, it is assumed that the pdf function is specified on the integers  $a + mk \leq b$  where k is a non-negative integer (0, 1, 2, ...) and m is a positive integer multiplier. Alternatively, the two lists  $x_k$  and  $p(x_k)$  can be provided directly in which case a dictionary is set up internally to evaluate probabilities and generate random variates.

### 0.1 Probability Mass Function (PMF)

The probability mass function of a random variable X is defined as the probability that the random variable takes on a particular value.

$$p\left(x_{k}\right) = P\left[X = x_{k}\right]$$

This is also sometimes called the probability density function, although technically

$$f(x) = \sum_{k} p(x_k) \, \delta(x - x_k)$$

is the probability density function for a discrete distribution<sup>1</sup>.

### 0.2 Cumulative Distribution Function (CDF)

The cumulative distribution function is

$$F(x) = P[X \le x] = \sum_{x_k \le x} p(x_k)$$

and is also useful to be able to compute. Note that

$$F(x_k) - F(x_{k-1}) = p(x_k)$$

#### 0.3 Survival Function

The survival function is just

$$S(x) = 1 - F(x) = P[X > k]$$

the probability that the random variable is strictly larger than k.

 $<sup>^{1}</sup>$ Note that we will be using p to represent the probability mass function and a parameter (a probability). The usage should be obvious from context.

### 0.4 Percent Point Function (Inverse CDF)

The percent point function is the inverse of the cumulative distribution function and is

$$G\left(q\right) = F^{-1}\left(q\right)$$

for discrete distributions, this must be modified for cases where there is no  $x_k$  such that  $F(x_k) = q$ . In these cases we choose G(q) to be the smallest value  $x_k = G(q)$  for which  $F(x_k) \ge q$ . If q = 0 then we define G(0) = a - 1. This definition allows random variates to be defined in the same way as with continuous rv's using the inverse cdf on a uniform distribution to generate random variates.

### 0.5 Inverse survival function

The inverse survival function is the inverse of the survival function

$$Z(\alpha) = S^{-1}(\alpha) = G(1 - \alpha)$$

and is thus the smallest non-negative integer k for which  $F(k) \ge 1 - \alpha$  or the smallest non-negative integer k for which  $S(k) \le \alpha$ .

#### 0.6 Hazard functions

If desired, the hazard function and the cumulative hazard function could be defined as

$$h\left(x_{k}\right) = \frac{p\left(x_{k}\right)}{1 - F\left(x_{k}\right)}$$

and

$$H(x) = \sum_{x_k \le x} h(x_k) = \sum_{x_k \le x} \frac{F(x_k) - F(x_{k-1})}{1 - F(x_k)}.$$

#### 0.7 Moments

Non-central moments are defined using the PDF

$$\mu'_{m} = E\left[X^{m}\right] = \sum_{k} x_{k}^{m} p\left(x_{k}\right).$$

Central moments are computed similarly  $\mu = \mu'_1$ 

$$\mu_{m} = E\left[ (X - \mu)^{2} \right] = \sum_{k} (x_{k} - \mu)^{m} p(x_{k})$$
$$= \sum_{k=0}^{m} (-1)^{m-k} {m \choose k} \mu^{m-k} \mu'_{k}$$

The mean is the first moment

$$\mu = \mu_1' = E[X] = \sum_k x_k p(x_k)$$

the variance is the second central moment

$$\mu_2 = E\left[ (X - \mu)^2 \right] = \sum_{x_k} x_k^2 p(x_k) - \mu^2.$$

Skewness is defined as

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

while (Fisher) kurtosis is

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3,$$

so that a normal distribution has a kurtosis of zero.

### 0.8 Moment generating function

The moment generating funtion is defined as

$$M_X(t) = E\left[e^{Xt}\right] = \sum_{x_k} e^{x_k t} p\left(x_k\right)$$

Moments are found as the derivatives of the moment generating function evaluated at 0.

### 0.9 Fitting data

To fit data to a distribution, maximizing the likelihood function is common. Alternatively, some distributions have well-known minimum variance unbiased estimators. These will be chosen by default, but the likelihood function will always be available for minimizing.

If  $f_i(k; \boldsymbol{\theta})$  is the PDF of a random-variable where  $\boldsymbol{\theta}$  is a vector of parameters (e.g. L and S), then for a collection of N independent samples from this distribution, the joint distribution the random vector  $\mathbf{k}$  is

$$f(\mathbf{k}; \boldsymbol{\theta}) = \prod_{i=1}^{N} f_i(k_i; \boldsymbol{\theta}).$$

The maximum likelihood estimate of the parameters  $\theta$  are the parameters which maximize this function with  $\mathbf{x}$  fixed and given by the data:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} f(\mathbf{k}; \boldsymbol{\theta})$$
$$= \arg \min_{\boldsymbol{\theta}} l_{\mathbf{k}}(\boldsymbol{\theta}).$$

Where

$$l_{\mathbf{k}}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \log f(k_i; \boldsymbol{\theta})$$
$$= -N \overline{\log f(k_i; \boldsymbol{\theta})}$$

#### 0.10 Standard notation for mean

We will use

$$\overline{y(\mathbf{x})} = \frac{1}{N} \sum_{i=1}^{N} y(x_i)$$

where N should be clear from context.

#### 0.11 Combinations

Note that

$$k! = k \cdot (k-1) \cdot (k-2) \cdot \dots \cdot 1 = \Gamma(k+1)$$

and has special cases of

$$\begin{array}{rcl}
0! & \equiv & 1 \\
k! & \equiv & 0 & k < 0
\end{array}$$

and

$$\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{(n-k)!k!}.$$

If 
$$n < 0$$
 or  $k < 0$  or  $k > n$  we define  $\binom{n}{k} = 0$ 

## 1 Bernoulli

A Bernoulli random variable of parameter p takes one of only two values X=0 or X=1. The probability of success (X=1) is p, and the probability of failure (X=0) is 1-p. It can be thought of as a binomial random variable with n=1. The PMF is p(k)=0 for  $k\neq 0,1$  and

$$p(k;p) = \begin{cases} 1-p & k=0 \\ p & k=1 \end{cases}$$

$$F(x;p) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$$

$$G(q;p) = \begin{cases} 0 & 0 \le q < 1-p \\ 1 & 1-p \le q \le 1 \end{cases}$$

$$\mu = p$$

$$\mu_2 = p(1-p)$$

$$\gamma_3 = \frac{1-2p}{\sqrt{p(1-p)}}$$

$$\gamma_4 = \frac{1-6p(1-p)}{p(1-p)}$$

$$M(t) = 1-p(1-e^t)$$

$$\mu'_m = p$$

$$h[X] = p \log p + (1-p) \log (1-p)$$

## 2 Binomial

A binomial random variable with parameters (n, p) can be described as the sum of n independent Bernoulli random variables of parameter p;

$$Y = \sum_{i=1}^{n} X_i.$$

Therefore, this random variable counts the number of successes in n independent trials of a random experiment where the probability of success is p.

$$p(k; n, p) = \binom{n}{k} p^{k} (1-p)^{n-k} \quad k \in \{0, 1, \dots n\},$$

$$F(x; n, p) = \sum_{k \le x} \binom{n}{k} p^{k} (1-p)^{n-k} = I_{1-p} (n - \lfloor x \rfloor, \lfloor x \rfloor + 1) \quad x \ge 0$$

where the incomplete beta integral is

$$I_{x}\left(a,b\right) = \frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} \int_{0}^{x} t^{a-1} \left(1-t\right)^{b-1} dt.$$

Now

$$\mu = np$$

$$\mu_2 = np (1-p)$$

$$\gamma_1 = \frac{1-2p}{\sqrt{np (1-p)}}$$

$$\gamma_2 = \frac{1-6p (1-p)}{np (1-p)}$$

$$M(t) = \left[1-p (1-e^t)\right]^n$$

## 3 Boltzmann (truncated Planck)

$$p(k; N, \lambda) = \frac{1 - e^{-\lambda}}{1 - e^{-\lambda N}} \exp(-\lambda k) \quad k \in \{0, 1, \dots, N - 1\}$$

$$F(x; N, \lambda) = \begin{cases} 0 & x < 0 \\ \frac{1 - \exp[-\lambda(\lfloor x \rfloor + 1)]}{1 - \exp(-\lambda N)} & 0 \le x \le N - 1 \\ 1 & x \ge N - 1 \end{cases}$$

$$G(q, \lambda) = \left[ -\frac{1}{\lambda} \log\left[1 - q\left(1 - e^{-\lambda N}\right)\right] - 1 \right]$$

Define  $z = e^{-\lambda}$ 

$$\mu = \frac{z}{1-z} - \frac{Nz^N}{1-z^N}$$

$$\mu_2 = \frac{z}{(1-z)^2} - \frac{N^2 z^N}{(1-z^N)^2}$$

$$\gamma_1 = \frac{z(1+z)\left(\frac{1-z^N}{1-z}\right)^3 - N^3 z^N \left(1+z^N\right)}{\left[z\left(\frac{1-z^N}{1-z}\right)^2 - N^2 z^N\right]^{3/2}}$$

$$\gamma_2 = \frac{z\left(1+4z+z^2\right)\left(\frac{1-z^N}{1-z}\right)^4 - N^4 z^N \left(1+4z^N+z^{2N}\right)}{\left[z\left(\frac{1-z^N}{1-z}\right)^2 - N^2 z^N\right]^2}$$

$$M(t) = \frac{1-e^{N(t-\lambda)}}{1-e^{t-\lambda}} \frac{1-e^{-\lambda}}{1-e^{-\lambda N}}$$

# 4 Planck (discrete exponential)

Named Planck because of its relationship to the black-body problem he solved.

$$p(k;\lambda) = (1 - e^{-\lambda}) e^{-\lambda k} \quad k\lambda \ge 0$$

$$F(x;\lambda) = 1 - e^{-\lambda(\lfloor x \rfloor + 1)} \quad x\lambda \ge 0$$

$$G(q;\lambda) = \left[ -\frac{1}{\lambda} \log [1 - q] - 1 \right].$$

$$\mu = \frac{1}{e^{\lambda} - 1}$$

$$\mu_2 = \frac{e^{-\lambda}}{(1 - e^{-\lambda})^2}$$

$$\gamma_1 = 2\cosh\left(\frac{\lambda}{2}\right)$$

$$\gamma_2 = 4 + 2\cosh(\lambda)$$

$$M(t) = \frac{1 - e^{-\lambda}}{1 - e^{t - \lambda}}$$

$$h[X] = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} - \log(1 - e^{-\lambda})$$

### 5 Poisson

The Poisson random variable counts the number of successes in n independent Bernoulli trials in the limit as  $n \to \infty$  and  $p \to 0$  where the probability of success in each trial is p and  $np = \lambda \ge 0$  is a constant. It can be used to approximate the Binomial random variable or in it's own right to count the number of events that occur in the interval [0,t] for a process satisfying certain "sparsity" constraints. The functions are

$$p(k;\lambda) = e^{-\lambda} \frac{\lambda^k}{k!} \quad k \ge 0,$$

$$F(x;\lambda) = \sum_{n=0}^{\lfloor x \rfloor} e^{-\lambda} \frac{\lambda^n}{n!} = \frac{1}{\Gamma(\lfloor x \rfloor + 1)} \int_{\lambda}^{\infty} t^{\lfloor x \rfloor} e^{-t} dt,$$

$$\mu = \lambda$$

$$\mu_2 = \lambda$$

$$\gamma_1 = \frac{1}{\sqrt{\lambda}}$$

$$\gamma_2 = \frac{1}{\lambda}.$$

$$M(t) = \exp\left[\lambda \left(e^t - 1\right)\right].$$

### 6 Geometric

The geometric random variable with parameter  $p \in (0,1)$  can be defined as the number of trials required to obtain a success where the probability of success on each trial is p. Thus,

$$p(k;p) = (1-p)^{k-1} p \quad k \ge 1$$

$$F(x;p) = 1 - (1-p)^{\lfloor x \rfloor} \quad x \ge 1$$

$$G(q;p) = \left[\frac{\log(1-q)}{\log(1-p)}\right]$$

$$\mu = \frac{1}{p}$$

$$\mu_2 = \frac{1-p}{p^2}$$

$$\gamma_1 = \frac{2-p}{\sqrt{1-p}}$$

$$\gamma_2 = \frac{p^2 - 6p + 6}{1-p}.$$

$$M(t) = \frac{p}{e^{-t} - (1-p)}$$

## 7 Negative Binomial

The negative binomial random variable with parameters n and  $p \in (0,1)$  can be defined as the number of extra independent trials (beyond n) required to accumulate a total of n successes where the probability of a success on each trial is p. Equivalently, this random variable is the number of failures encoutered while accumulating n successes during independent trials of an experiment that succeeds with probability p. Thus,

$$p(k; n, p) = {k+n-1 \choose n-1} p^n (1-p)^k \quad k \ge 0$$

$$F(x; n, p) = \sum_{i=0}^{\lfloor x \rfloor} {i+n-1 \choose i} p^n (1-p)^i \quad x \ge 0$$

$$= I_p (n, \lfloor x \rfloor + 1) \quad x \ge 0$$

$$\mu = n \frac{1-p}{p}$$

$$\mu_2 = n \frac{1-p}{p^2}$$

$$\gamma_1 = \frac{2-p}{\sqrt{n(1-p)}}$$

$$\gamma_2 = \frac{p^2 + 6(1-p)}{n(1-p)}.$$

Recall that  $I_p(a,b)$  is the incomplete beta integral.

# 8 Hypergeometric

The hypergeometric random variable with parameters (M, n, N) counts the number of "good" objects in a sample of size N chosen without replacement from a population of M objects where n is the number of "good" objects in the total population.

$$p(k; N, n, M) = \frac{\binom{n}{k} \binom{M-n}{N-k}}{\binom{M}{N}} N - (M-n) \le k \le \min(n, N)$$

$$F(x; N, n, M) = \sum_{k=0}^{\lfloor x \rfloor} \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}},$$

$$\mu = \frac{nN}{M}$$

$$\mu_2 = \frac{nN(M-n)(M-N)}{M^2(M-1)}$$

$$\gamma_1 = \frac{(M-2n)(M-2N)}{M-2} \sqrt{\frac{M-1}{nN(M-m)(M-n)}}$$

$$\gamma_2 = \frac{g(N, n, M)}{nN(M-n)(M-3)(M-2)(N-M)}$$

where (defining m = M - n)

$$\begin{array}{ll} g\left(N,n,M\right) & = & m^3 - m^5 + 3m^2n - 6m^3n + m^4n + 3mn^2 \\ & -12m^2n^2 + 8m^3n^2 + n^3 - 6mn^3 + 8m^2n^3 \\ & + mn^4 - n^5 - 6m^3N + 6m^4N + 18m^2nN \\ & -6m^3nN + 18mn^2N - 24m^2n^2N - 6n^3N \\ & -6mn^3N + 6n^4N + 6m^2N^2 - 6m^3N^2 - 24mnN^2 \\ & + 12m^2nN^2 + 6n^2N^2 + 12mn^2N^2 - 6n^3N^2. \end{array}$$

# 9 Zipf (Zeta)

A random variable has the zeta distribution (also called the zipf distribution) with parameter  $\alpha > 1$  if it's probability mass function is given by

$$p(k; \alpha) = \frac{1}{\zeta(\alpha) k^{\alpha}} \quad k \ge 1$$

where

$$\zeta\left(\alpha\right) = \sum_{n=1}^{\infty} \frac{1}{n^{\alpha}}$$

is the Riemann zeta function. Other functions of this distribution are

$$F(x;\alpha) = \frac{1}{\zeta(\alpha)} \sum_{k=1}^{\lfloor x \rfloor} \frac{1}{k^{\alpha}}$$

$$\mu = \frac{\zeta_1}{\zeta_0} \quad \alpha > 2$$

$$\mu_2 = \frac{\zeta_2 \zeta_0 - \zeta_1^2}{\zeta_0^2} \quad \alpha > 3$$

$$\gamma_1 = \frac{\zeta_3 \zeta_0^2 - 3\zeta_0 \zeta_1 \zeta_2 + 2\zeta_1^3}{[\zeta_2 \zeta_0 - \zeta_1^2]^{3/2}} \quad \alpha > 4$$

$$\gamma_2 = \frac{\zeta_4 \zeta_0^3 - 4\zeta_3 \zeta_1 \zeta_0^2 + 12\zeta_2 \zeta_1^2 \zeta_0 - 6\zeta_1^4 - 3\zeta_2^2 \zeta_0^2}{(\zeta_2 \zeta_0 - \zeta_1^2)^2}.$$

$$M(t) = \frac{\operatorname{Li}_{\alpha}(e^{t})}{\zeta(\alpha)}$$

where  $\zeta_i = \zeta\left(\alpha - i\right)$  and  $\operatorname{Li}_n\left(z\right)$  is the  $n^{\operatorname{th}}$  polylogarithm function of z defined as

$$\operatorname{Li}_{n}(z) \equiv \sum_{k=1}^{\infty} \frac{z^{k}}{k^{n}}$$

$$\mu'_{n} = M^{(n)}(t) \Big|_{t=0} = \frac{\operatorname{Li}_{\alpha-n}(e^{t})}{\zeta(a)} \Big|_{t=0} = \frac{\zeta(\alpha-n)}{\zeta(\alpha)}$$

# 10 Logarithmic (Log-Series, Series)

The logarimthic distribution with parameter p has a probability mass function with terms proportional to the Taylor series expansion of  $\log (1-p)$ 

$$p(k;p) = -\frac{p^k}{k\log(1-p)} \quad k \ge 1$$

$$F(x;p) = -\frac{1}{\log(1-p)} \sum_{k=1}^{\lfloor x \rfloor} \frac{p^k}{k} = 1 + \frac{p^{1+\lfloor x \rfloor} \Phi(p, 1, 1+\lfloor x \rfloor)}{\log(1-p)}$$

where

$$\Phi\left(z,s,a\right) = \sum_{k=0}^{\infty} \frac{z^k}{\left(a+k\right)^s}$$

is the Lerch Transcendent. Also define  $r = \log(1 - p)$ 

$$\mu = -\frac{p}{(1-p)r}$$

$$\mu_2 = -\frac{p[p+r]}{(1-p)^2 r^2}$$

$$\gamma_1 = -\frac{2p^2 + 3pr + (1+p)r^2}{r(p+r)\sqrt{-p(p+r)}}r$$

$$\gamma_2 = -\frac{6p^3 + 12p^2r + p(4p+7)r^2 + (p^2 + 4p+1)r^3}{p(p+r)^2}.$$

$$M(t) = -\frac{1}{\log(1-p)} \sum_{k=1}^{\infty} \frac{e^{tk} p^k}{k}$$
$$= \frac{\log(1-pe^t)}{\log(1-p)}$$

Thus,

$$\mu'_{n} = M^{(n)}(t)\Big|_{t=0} = \frac{\operatorname{Li}_{1-n}(pe^{t})}{\log(1-p)}\Big|_{t=0} = -\frac{\operatorname{Li}_{1-n}(p)}{\log(1-p)}.$$

# 11 Discrete Uniform (randint)

The discrete uniform distribution with parameters (a, b) constructs a random variable that has an equal probability of being any one of the integers in the half-open range [a, b). If a is not given it is assumed to be zero and the only parameter is b. Therefore,

$$p(k; a, b) = \frac{1}{b-a} \quad a \le k < b$$

$$F(x; a, b) = \frac{\lfloor x \rfloor - a}{b-a} \quad a \le x \le b$$

$$G(q; a, b) = \lceil q(b-a) + a \rceil$$

$$\mu = \frac{b+a-1}{2}$$

$$\mu_2 = \frac{(b-a-1)(b-a+1)}{12}$$

$$\gamma_1 = 0$$

$$\gamma_2 = -\frac{6}{5} \frac{(b-a)^2 + 1}{(b-a-1)(b-a+1)}.$$

$$M(t) = \frac{1}{b-a} \sum_{k=a}^{b-1} e^{tk}$$
$$= \frac{e^{bt} - e^{at}}{(b-a)(e^t - 1)}$$

## 12 Discrete Laplacian

Defined over all integers for a > 0

$$\begin{split} p\left(k\right) &= \tanh\left(\frac{a}{2}\right)e^{-a|k|}, \\ F\left(x\right) &= \begin{cases} \frac{e^{a\left(\lfloor x\rfloor+1\right)}}{e^a+1} & \lfloor x\rfloor < 0, \\ 1 - \frac{e^{-a\lfloor x\rfloor}}{e^a+1} & \lfloor x\rfloor \geq 0. \end{cases} \\ G\left(q\right) &= \begin{cases} \left\lceil \frac{1}{a}\log\left[q\left(e^a+1\right)\right] - 1\right\rceil & q < \frac{1}{1+e^{-a}}, \\ \left\lceil -\frac{1}{a}\log\left[\left(1-q\right)\left(1+e^a\right)\right]\right\rceil & q \geq \frac{1}{1+e^{-a}}. \end{cases} \\ M\left(t\right) &= \tanh\left(\frac{a}{2}\right)\sum_{k=-\infty}^{\infty}e^{tk}e^{-a|k|} \end{split}$$

$$M(t) = \tanh\left(\frac{a}{2}\right) \sum_{k=-\infty} e^{tk} e^{-a|k|}$$

$$= C\left(1 + \sum_{k=1}^{\infty} e^{-(t+a)k} + \sum_{1}^{\infty} e^{(t-a)k}\right)$$

$$= \tanh\left(\frac{a}{2}\right) \left(1 + \frac{e^{-(t+a)}}{1 - e^{-(t+a)}} + \frac{e^{t-a}}{1 - e^{t-a}}\right)$$

$$= \frac{\tanh\left(\frac{a}{2}\right) \sinh a}{\cosh a - \cosh t}.$$

Thus,

$$\mu'_{n} = M^{(n)}(0) = [1 + (-1)^{n}] \operatorname{Li}_{-n} (e^{-a})$$

where  $\operatorname{Li}_{-n}(z)$  is the polylogarithm function of order -n evaluated at z.

$$h[X] = -\log\left(\tanh\left(\frac{a}{2}\right)\right) + \frac{a}{\sinh a}$$

### 13 Discrete Gaussian\*

Defined for all  $\mu$  and  $\lambda > 0$  and k

$$p(k; \mu, \lambda) = \frac{1}{Z(\lambda)} \exp \left[-\lambda (k - \mu)^2\right]$$

where

$$Z(\lambda) = \sum_{k=-\infty}^{\infty} \exp\left[-\lambda k^{2}\right]$$

$$\mu = \mu$$

$$\mu_{2} = -\frac{\partial}{\partial \lambda} \log Z(\lambda)$$

$$= G(\lambda) e^{-\lambda}$$

where  $G\left(0\right)\to\infty$  and  $G\left(\infty\right)\to2$  with a minimum less than 2 near  $\lambda=1$ 

$$G(\lambda) = \frac{1}{Z(\lambda)} \sum_{k=-\infty}^{\infty} k^2 \exp\left[-\lambda (k+1) (k-1)\right]$$