SI140 Discussion 05

Li Zeng, Tao Huang, Xinyi Liu

ShanghaiTech University, China
{zengli,huangtao1,liuxy10}@shanghaitech.edu.cn

1 Bernoulli & Binomial Random Variables

1.1 Bernoulli Random Variable

A Bernoulli random variable is the simplest kind of random variable. It can take on two values, 1 and 0. It takes on a 1 if an experiment with probability p resulted in success and a 0 otherwise. Some example uses include a coin flip, a random binary digit, whether a disk drive crashed, and whether someone likes a Netflix movie.
If X is a Bernoulli random variable, denoted $X \sim \operatorname{Ber}(p)$:

$$P(X=1)=p,\ P(X=0)=1-p$$

$$E[X]=p$$

$$\operatorname{Var}(X)=p(1-p)$$

1.2 Binomial random variable

A Binomial random variable is random variable that represents the number of successes in n successive independent trials of a Bernoulli experiment. Some example uses include the number of heads in n coin flips, the number of disk drives that crashed in a cluster of 1000 computers, and the number of advertisements that are clicked when 40,000 are served.
If X is a Binomial random variable, we denote this $X \sim \operatorname{Bin}(n,p)$, where p is the probability of success in a given trial. A binomial random variable has the following properties:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \quad \text{if } k \in N, 0 \leq k \leq n (0 \text{ otherwise})$$

$$E[X]=np$$

$$\operatorname{Var}(X)=np(1-p)$$

There is a strong relationship between the Binomial random variable and the Bernoulli random variable- in that a Binomial RV is the sum of n independent Bernoulli RVs. Next week we'll talk more about what independence in the context of random variables means, but for now:
Let $X_i \sim \operatorname{Ber}(p)$, for $i=1,\ldots,n$. Let $Y=\sum_{i=1}^{n} X_i$. If the X_i are independent, then $Y \sim \operatorname{Bin}(n,p)$ Another way to think about this is that for the Binomial RV Y to take on value k, it must be true that there are exactly k of the X_i 's take on value 1, and all other X_i 's must take on value 0. There are $\binom{n}{k}$ ways to pick which X_i 's will have value 1 . (after which we set the rest to 0). When we sum up these k ones and $n-k$ zeros, we get $Y=k$.

Example 1. **Hamming Code.** When sending messages over a network, there is a chance that the bits will be corrupted. A Hamming code allows for a 4 bit code to be encoded as 7 bits, with the advantage that if 0 or 1 bit(s) are corrupted, then the message can be perfectly reconstructed. You are working on the Voyager space mission and the probability of any bit being lost in space is 0.1. How does reliability change when using a Hamming code?

Solution 1. Image we use error correcting codes. Let $X \sim \mathrm{Bin}(7, 0.1)$

$$P(X = 0) = \binom{7}{0}(0.1)^0(0.9)^7 \approx 0.468$$

$$P(X = 1) = \binom{7}{1}(0.1)^1(0.9)^6 = 0.372$$

$$P(\text{correctable}) = P(X = 0) + P(X = 1) = 0.850$$

What if we didn't use error correcting codes? Let $X \sim \mathrm{Bin}(4, 0.1)$

$$P(X = 0) = \binom{4}{0}(0.1)^0(0.9)^4 \approx 0.656$$

Using Hamming Codes improves reliability by about 30%!

2 Expectation & Variance

2.1 Definition of Expectation

A useful piece of information about a random variable is the average value of the random variable over many repetitions of the experiment it represents. This average is called the **expectation.**

Definition 1 (Expectation). *The* **expectation** *of a discrete random variable X is defined as:*

$$E[X] = \sum_{x:P(x)>0} xP(X = x)$$

It goes by many other names: mean, expected value, weighted average, center of mass, 1st moment.

2.2 Properties of Expectation

- Linearity: $E[aX + bY + c] = aE[x] + bE[Y] + c$.
- Law Of The Unconscious Statistician(LOTUS): $E[g(x)] = \sum_x g(x)P(X = x)$.
- Square: $E[X^2] = \sum_x x^2 P(X = x)$.

Note that by using LOTUS, we do not have to explicitly compute a distribution (i.e., a PMF) on Y; we can simply use the known distribution of X. This property of expectation is especially useful when we must compute the expectation of a complex function of X. See the St. Petersburg Paradox example.

Example 2 (St.Petersburg Paradox). Consider a game played with a fair coin which comes up heads with $p = 0.5$. Let N be the number of coin flips before the first "tails". In this game you win $\$2N$. How many dollars do you expect to win? Let W be a random variable which represents your winnings.

Solution 2. Note that $W = 2^N$ and $P(N = n) = (\frac{1}{2})^{n+1}$. We do not explicitly compute the PMF of W. Instead, using LOTUS,

$$E[W] = E[2^N] = (\frac{1}{2})^1 2^0 + (\frac{1}{2})^2 2^1 + \cdots = \sum_{i=0}^{\infty}(\frac{1}{2})^{i+1}2^i = \sum_{i=0}^{\infty}\frac{1}{2} = \infty$$

This example is nicknamed a paradox because consider the more realistic scenario, where the game dealer has a maximum amount of money (say, \$65,536), and you cannot win more than this amount. If you are projected to win more, then the dealer goes home, you get kicked out of the hall, and you win nothing. In this case, when $N = 16$, you win $W = 2^{16} = 65,536$, and if $N \geq 17$ you win nothing ($W = 0$). Using LOTUS, where $k = \log_2(65,536) = 16$:

$$E[W] = (\frac{1}{2})^1 2^0 + (\frac{1}{2})^2 2^1 + \cdots + (\frac{1}{2})^k 2^{k-1} = \sum_{0}^{k}(\frac{1}{2}) = 8.5$$

Example 3. Players A and B take turns in answering trivia questions, starting with player A answering the first question. Each time A answers a question, she has probability p_1 of getting it right. Each time B plays, he has probability p_2 of getting it right.

(a) If A answers m questions, what is the PMF of the number of questions she gets right?
(b) If A answers m times and B answers n times, what is the PMF of the total number of questions they get right (you can leave your answer as a sum)? Describe exactly when/whether this is a Binomial distribution.
(c) Suppose that the first player to answer correctly wins the game (with no predetermined maximum number of questions that can be asked). Find the probability that A wins the game.

Solution 3. (a) The r.v. is $Bin(m, p_1)$. So the PMF is $p_i = \binom{m}{i} p_1^i (1-p_1)^{m-i}$.
(b) Let T be the total number of questions they get right. To get a total of k questions right, it must be that A got j and B got kj, for $j = 0, ..., k$. These are disjoint events so the PMF is

$$P(T=k) = \sum_{j=0}^{k} \binom{m}{j} p_1^j (1-p_1)^{m-j} \binom{n}{k-j} p_2^{k-j} (1-p_2)^{n-(k-j)}$$

If $p_1 = p_2 = p$, recall the identity

$$\sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j} = \binom{m+n}{k}$$

which implies

$$P(T=k) = \binom{m+n}{k} p^k (1-p)^{m+n-k}$$

Hence, when $p_1 = p_2$, it is Binomial; otherwise, it's not.
(c) Let $r = P(A \text{ wins})$. Conditioning on the results of the first question for each player, we have

$$r = p_1 + (1-p_1)p_2 \cdot 0 + (1-p_1)(1-p_2)r$$

which gives

$$r = \frac{p_1}{1-(1-p_1)(1-p_2)} = \frac{p_1}{p_1 + p_2 - p_1 p_2}$$

2.3 Definition of Variance

Expectation is a useful statistic, but it does not give a detailed view of the probability mass function. **Variance** is a formal quantification of "spread". There is more than one way to quantify spread; variance uses the average square distance from the mean.

Definition 2 (Variance). *The* **variance** *of a discrete random variable X with expected value μ is:*

$$Var(X) = E[(X-\mu)^2]$$

2.4 Properties of Variance

- For any $r.v.X$, $Var(X) = E[X^2] - (E[X])^2$.
- $Var(aX+b) = a^2 Var(X)$. **Note that this implies that variance is nonlinear**.
- $Var(X+Y) = Var(X) + Var(Y)$ if X and Y are independent.

3 Indicator Variable's Definition and Property

Indicator variables are important in probability and statistics, because it bridges expectation with probability of a event, which serves a important role in:

- Dividing up complex random variable into easy dummy variables.
- In statistics, indicator variables are used to deal with data in different categories.

Definition 3 (Indicator Variable). *The indicator random variable of an event A is the r.v. which equals 1 if A occurs and 0 otherwise. We will denote the indicator r.v. of A by I_A or $I(A)$.*

3.1 Properties of Indicator Variables

- $P(A) = E\left(I_A\right)$ (bridging expectation with probability)
- Binary:
  - $\left(I_A\right)^k = I_A$ for any positive integer k.
  - $I_{A^c} = 1 - I_A$.
  - $I_{A \cap B} = I_A I_B$.
  - $I_{A \cup B} = I_A + I_B - I_A I_B$
- Moments of indicator: Given n events $A_1, \ldots, A_n$ and indicators $I_j, j = 1, \ldots, n$.
  - Sum of IVs: $X = \sum_{j=1}^n l_j$ (# of events that occur)
  - Pair performance of events: $Y = \binom{X}{2} = \sum_{i<j} I_i I_j$ (# of pairs of distinct events that occur)
  - Relationship between i.v.s & Expectation & Variance:
    - $E\left(\binom{X}{2}\right) = \sum_{i<j} P\left(A_i \cap A_j\right) = \frac{1}{2} \sum_{i \neq j} P\left(A_i \cap A_j\right)$
    - $E\left(X^2\right) = \sum_{i \neq j} E(I_i I_j) + E(X) = 2E\left(\binom{X}{2}\right) + E(X)$
    - $\operatorname{Var}(X) = 2E\left(\binom{X}{2}\right) + E(X) - (E(X))^2$

Problem 1 (Almost Fixed Points in a Circle).

Let Ω be the set of all permutations of the numbers $1, 2, \ldots, n$. Let an almost fixed point be defined as follows: If we put the numbers $i \in 1, 2, \ldots, n$ around a circle in clockwise order (such that 1 and n are next to each other) and then assign another number $\omega(i) \in 1, 2, \ldots, n$ to it, if the number $\omega(i)$ is next to i (or is equal to i), we will say that i is almost a fixed point. So, for the permutation $\omega(1) = 5, \omega(2) = 3, \omega(3) = 1, \omega(4) = 4, \omega(5) = 2$, we have that $1, 2$, and 4 are almost fixed points. Now, let $X(\omega)$ denote the number of almost fixed points in $\omega \in \Omega$.

(a) Find $E[X]$.
(b) Find $\operatorname{var}(X)$.

Solution 4. (a) Suppose X_i is the indicator variable of that number i is a almost fixed point. The total number of almost fixed point is X. Therefore,

$$E\left[X_i\right] = P\left(X_i = 1\right) = \frac{3}{n}$$

(b) since

$$X = \sum_{i=1}^n X_i$$

So, if $n \geq 3$

$$E[X] = \sum_{i=1}^n E\left[X_i\right] = 3$$

since all permutations are equally possible, then for $n \geq 4$, we have

$$E\left[X_1 X_2\right] = E\left[X_1 X_n\right] = \frac{1}{n} \frac{3}{n-1} + \frac{2}{n} \frac{2}{n-1} = \frac{7}{n(n-1)}$$

if $n \geq 5$, we have

$$E\left[X_1 X_3\right] = E\left[X_1 X_{n-1}\right] = \frac{2}{n} \frac{3}{n-1} + \frac{1}{n} \frac{2}{n-1} = \frac{8}{n(n-1)}$$

if $n > 5$,for $4 \leq i \leq n-2$, we have

$$E\left[X_1 X_i\right] = \frac{3}{n} \frac{3}{n-1} = \frac{9}{n(n-1)}$$

Now we are interested in $E\left(X^2\right)$, since X_i is identical for all i, then if $n \geq 5$

$$\begin{aligned}
E\left[X^2\right] &= E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] \\
&= nE\left[X_1^2\right] + n\left(E\left[X_1X_2\right] + E\left[X_1X_3\right] + \ldots + E\left[X_1X_{n-1}\right] + E\left[X_1X_n\right]\right) \\
&= nE\left[X_1\right] + n\left(\frac{2\times 7}{n(n-1)} + \frac{2\times 8}{n(n-1)} + \frac{(n-1-4)\times 9}{n(n-1)}\right) \\
&= 3 + \frac{9n-15}{n-1}
\end{aligned}$$

Therefore, for $n \geq 5$

$$\operatorname{var}(X) = E\left[X^2\right] - E[X]^2 = \frac{3n-9}{n-1}$$

Notice that this also hold when n = 3 and n = 4 Therefore, in summary, for $n \geq 3$,

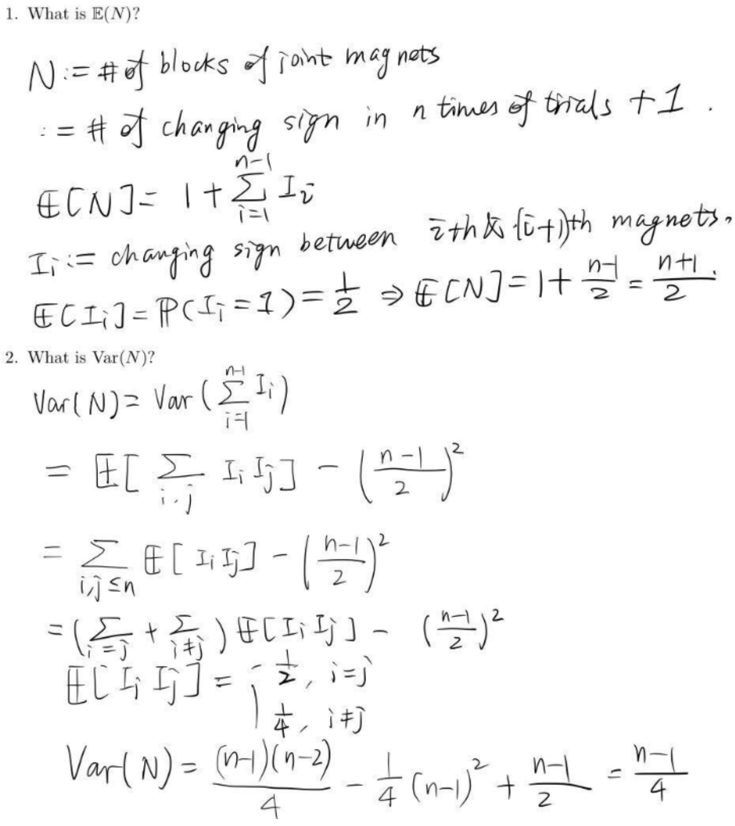
$$E[X] = 3$$

$$\operatorname{var}(X) = \frac{3n-9}{n-1}$$

Problem 2 (Magnets). There are n bar magnets, $n > 1$, placed in a line end to end. Assume that each magnet takes one of the two possible orientations, say (NS) or (SN), with equal probability, and magnets have independent orientations. Adjacent magnets with like poles repel, while those with opposite poles join and form blocks. For instance, if $n = 5$, and the orientation of magnets is $(NS)(SN)(SN)(NS)(NS)$, they form 3 blocks of the form $(NS)|(SN)(SN)|(NS)(NS)$. Let N be the number of blocks of joint magnets.

(a) What is $E(N)$?
(b) What is $\operatorname{Var}(N)$?

Solution 5. See soln in the following piture:



1. What is $\mathbb{E}(N)$?

$N :=$ # of blocks of joint magnets

$:=$ # of changing sign in n times of trials $+1$.

$\mathbb{E}[N] = 1 + \sum_{i=1}^{n-1} I_i$

$I_i :=$ changing sign between ith 及 $(i+1)$th magnets.

$\mathbb{E}[I_i] = \mathbb{P}(I_i = 1) = \frac{1}{2} \Rightarrow \mathbb{E}[N] = 1 + \frac{n-1}{2} = \frac{n+1}{2}.$

2. What is $\operatorname{Var}(N)$?

$$\begin{aligned}
\operatorname{Var}(N) &= \operatorname{Var}\left(\sum_{i=1}^{n-1} I_i\right) \\
&= \mathbb{E}\left[\sum_{i,j} I_i I_j\right] - \left(\frac{n-1}{2}\right)^2 \\
&= \sum_{i,j\leq n} \mathbb{E}\left[I_i I_j\right] - \left(\frac{n-1}{2}\right)^2 \\
&= \left(\sum_{i=j} + \sum_{i\neq j}\right) \mathbb{E}\left[I_i I_j\right] - \left(\frac{n-1}{2}\right)^2
\end{aligned}$$

$$\mathbb{E}\left[I_i I_j\right] = \begin{cases} \frac{1}{2}, & i=j \\ \frac{1}{4}, & i\neq j \end{cases}$$

$$\operatorname{Var}(N) = \frac{(n-1)(n-2)}{4} - \frac{1}{4}(n-1)^2 + \frac{n-1}{2} = \frac{n-1}{4}$$

4 The Poisson Distribution

4.1 Binomial in the Limit

Recall the example of sending a bit string over a network. In our last class we used a binomial random variable to represent the number of bits corrupted out of 4 with a high corruption probability (each bit had independent probability of corruption $p=0.1$). That example was relevant to sending data to spacecraft, but for earthly applications like HTML data, voice or video, bit streams are much longer (length $\approx 10^4$) and the probability of corruption of a particular bit is very small $\left(p \approx 10^{-6}\right)$. Extreme n and p values arise in many cases: visitors to a website, server crashes in a giant data center.

Unfortunately, $X \sim \operatorname{Bin}\left(10^4, 10^{-6}\right)$ is unwieldy to compute. However, when values get that extreme, we can make approximations that are **accurate and make computation feasible**. Recall that the parameters of the binomial distribution are $n=10^4$ and $p=10^{-6}$. First, define $\lambda=np$. We can rewrite the binomial PMF as follows:

$$
\begin{aligned}
P(X=i) &= \frac{n!}{i!(n-i)!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} \\
&= \frac{n(n-1)\ldots(n-i-1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}
\end{aligned}
$$

This equation can be made simpler using some approximations that hold when n is sufficiently large and p is sufficiently small:

$$
\frac{n(n-1)\ldots(n-i-1)}{n^i} \approx 1
$$

$$
(1-\lambda/n)^n \approx e^{-\lambda}
$$

$$
(1-\lambda/n)^i \approx 1
$$

Using these reduces our original equation to:

$$
P(X=i)=\frac{\lambda^i}{i!} e^{-\lambda}
$$

This simplification, derived by assuming extreme values of n and p, turns out to be so useful that it gets its own random variable type: the **Poisson random variable**.

4.2 Poisson Random Variable

A Poisson random variable approximates Binomial where n is large, p is small, and $\lambda=np$ is "moderate". Interestingly, to calculate the things we care about (PMF, expectation, variance), we no longer need to know n and p. We only need to provide λ, which we call the rate.

There are different interpretations of "moderate". Commonly accepted ranges are $n>20$ and $p<0.05$ or $n>100$ and $p<0.1$ Here are the key formulas you need to know for Poisson. If Y is a Poisson random variable, denoted $Y \sim \operatorname{Poi}(\lambda)$, then

$$
P(Y=i)=\frac{\lambda^i}{i!} e^{-\lambda}
$$

$$
E[Y]=\lambda
$$

$$
\operatorname{Var}(Y)=\lambda
$$

Note here that the expectation = variance = parameter λ.

Example 4. The Poisson distribution is often used to model the number of events that occur independently at any time in an interval of time or space, with a constant average rate. Earthquakes are a good example of this. Suppose there are an average of 2.8 major earthquakes in the world each year. What is the probability of getting more than one major earthquake next year?

Solution 6. Let $X \sim \text{Poi}(2.8)$ be the number of major earthquakes next year. We want to know $P(X>1)$ We can use the complement rule to rewrite this as $1-P(X=0)-P(X=1)$. Using the PMF for Poisson:

$$\begin{aligned} P(X>1) &= 1-P(X=0)-P(X=1) \\ &= 1-e^{-2.8}\frac{2.8^0}{0!}-e^{-2.8}\frac{2.8^1}{1!} \\ &= 1-e^{-2.8}-2.8e^{-2.8} \\ &\approx 1-0.06-0.17 \\ &= 0.77 \end{aligned}$$

4.3 Connection between Poisson and Binomial

Poisson: Binomial in the Limit. See 4.1.

Binomial: Poisson Conditioning on A Sum of Poissons. Given the sum of two independent Poisson variables, the conditional distribution of either is binomial, i.e. If $X \sim \text{Pois}(\lambda_1), Y \sim \text{Pois}(\lambda_2)$, and X is independent of Y, then the conditional distribution of X given $X+Y=n$ is $\text{Bin}(n, \lambda_1/(\lambda_1+\lambda_2))$. (*Required a proof here.*)

5 Other Discrete Distributions (Checking)

In this part, we briefly give a summary of some other distributions mentioned in this course. You may also need to be able to derive all the expectation and variance formulas on your own.

5.1 Geometric Distribution

X is a geometric random variable ($X \sim \text{Geo}(p)$) if X is number of the independent trials until the first success and p is probability of success on each trial. If $X \sim \text{Geo}(p)$

$$P(X=n)=(1-p)^{n-1}p$$

$$E[X]=1/p$$

$$\text{Var}(X)=(1-p)/p^2$$

The PMF, $P(X=n)$, can be derived using the independence assumption. Let E_i represent the event that the i -th trial succeeds. Then the probability that X is exactly n is the probability that the first $n-1$ trials fail, and the n -th succeeds:

$$\begin{aligned} P(X=n) &= P\left(E_1^C E_2^C \ldots {E_{n-1}}^C E_n\right) \\ &= P\left(E_1^C\right) P\left(E_2^C\right) \ldots P\left({E_{n-1}}^C\right) P\left(E_n\right) \\ &= (1-p)^{n-1} p \end{aligned}$$

A similar argument can be used to derive the CDF, the probability that $X \leq n$. This is equal to $1-P(X>n)$, and $P(X>n)$ is the probability that at least the first n trials fail:

$$\begin{aligned} P(X \leq n) &= 1-P(X>n) \\ &= 1-P\left(E_1^C E_2^C \ldots E_n^C\right) \\ &= 1-P\left(E_1^C\right) P\left(E_2^C\right) \ldots P\left(E_n^C\right) \\ &= 1-(1-p)^n \end{aligned}$$

5.2 Negative Binomial Distribution

X is a negative binomial random variable ($X \sim \text{NBin}(r, p)$) if X is the number of independent trials until r successes and p is probability of success on each trial. If $X \sim \text{NBin}(r, p)$

$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \quad r \leq n$$

$$E[X] = r/p$$

$$\text{Var}(X) = r(1-p)/p^2$$

Example 5. A grad student needs 3 published papers to graduate. (Not how it works in real life!) On average, how many papers will the student need to submit to a conference, if the conference accepts each paper randomly and independently with probability $p = 0.25$? (Also not how it works in real life...though the NIPS Experiment suggests there is a grain of truth in this model!)

Solution 7. Let X be the number of submissions required to get 3 acceptances. $X \sim \text{Neg Bin}(r = 3, p = 0.25)$. So $E[X] = \frac{r}{p} = \frac{3}{0.25} = 12$

5.3 Hypergeometric Distribution

X is a hypergeometric random variable ($X \sim \text{HypG}(n, N, m)$) if X is the number of red balls drawn when n balls are drawn at random, **without replacement**, from an urn with N balls total, m of which are red. If $X \sim \text{HyperG}(p)$

$$P(X = k) = \frac{\binom{m}{k}\binom{n-m}{n-k}}{\binom{N}{n}}, 0 \leq k \leq \min(n, m)$$

$$E[X] = n\frac{m}{N}$$

$$\text{Var}(X) = \frac{nm(N-n)(N-m)}{N^2(N-1)}$$

The Binomial Approximation to the Hypergeometric A Hyper-geometric distribution often occurs in reality, i.e., sampling without replacement to estimate the male-female proportion in population. In such a large N, m setting, it can usually be approximated by a binomial distribution. The reason is that, if the sample size does not exceed 5% of the population size, there is little difference between sampling with and without replacement.

5.4 Zipf Distribution

X is a Zipf random variable ($X \sim \text{Zipf}(s, N)$) if the probability of X obeys an inverse power law:

$$P(X = k) = C \cdot \frac{1}{k^s} \text{ where } 1 \leq k \leq N$$

where C is a normalizing constant (which turns out to be equal to reciprocal of the N th harmonic number).

In human languages, a Zipf distribution is a good model of the frequency rank index of a randomly chosen word, where N is the number of words in the language, and s also depends on various properties of the language (but is often close to 1). Other processes involving rank-ordering quantities also frequently result in a Zipf distribution, such as the rank of populations of large cities.