Geometric distribution

In probability theory and statistics, the **geometric distribution** is either of two discrete probability distributions:

- The probability distribution of the number *X* of Bernoulli trials needed to get one success, supported on the set { 1, 2, 3, ... }
- The probability distribution of the number *Y* = *X* − 1 of failures before the first success, supported on the set { 0, 1, 2, 3, ... }

Which of these one calls "the" geometric distribution is a matter of convention and convenience.

These two different geometric distributions should not be confused with each other. Often, the name *shifted* geometric distribution is adopted for the former one (distribution of the number *X*); however, to avoid ambiguity, it is considered wise to indicate which is intended, by mentioning the support explicitly.

The geometric distribution gives the probability that the first occurrence of success requires *k* independent trials, each with success probability *p*. If the probability of success on each trial is *p*, then the probability that the *k*th trial (out of *k* trials) is the first success is

$$\Pr(X=k)=(1-p)^{k-1}p$$

for *k* = 1, 2, 3, ...

The above form of the geometric distribution is used for modeling the number of trials up to and including the first success. By contrast, the following form of the geometric distribution is used for modeling the number of failures until the first success:

$$\Pr(Y=k)=(1-p)^{k}p$$

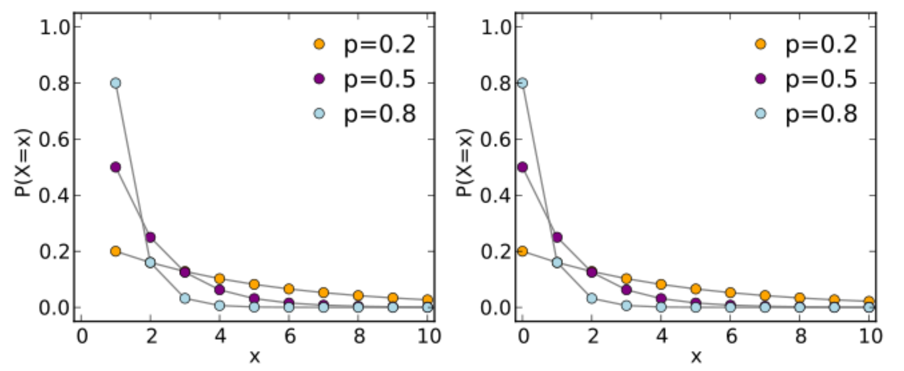
for *k* = 0, 1, 2, 3, ...

In either case, the sequence of probabilities is a geometric sequence.

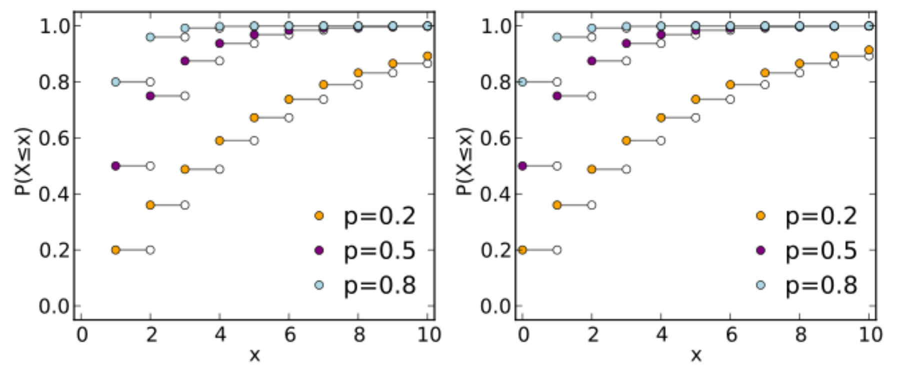
For example, suppose an ordinary die is thrown repeatedly until the first time a "1" appears. The probability distribution of the number of times it is thrown is supported on the infinite set { 1, 2, 3, ... } and is a geometric distribution with *p* = 1/6.

Geometric

Probability mass function



Cumulative distribution function



Parameters	$0<p<1$ success probability (real)	$0<p\leq 1$ success probability (real)
Support	*k* trials where $k\in\{1,2,3,\dots\}$	*k* failures where $k\in\{0,1,2,3,\dots\}$
Probability mass function (pmf)	$(1-p)^{k-1}p$	$(1-p)^{k}p$
CDF	$1-(1-p)^{k}$	$1-(1-p)^{k+1}$
Mean	$\frac{1}{p}$	$\frac{1-p}{p}$
Median	$\left\lceil\frac{-1}{\log_2(1-p)}\right\rceil$ (not unique if $-1/\log_2(1-p)$ is an integer)	$\left\lceil\frac{-1}{\log_2(1-p)}\right\rceil-1$ (not unique if $-1/\log_2(1-p)$ is an integer)
Mode	1	0
Variance	$\frac{1-p}{p^2}$	$\frac{1-p}{p^2}$
Skewness	$\frac{2-p}{\sqrt{1-p}}$	$\frac{2-p}{\sqrt{1-p}}$
Excess kurtosis	$6+\frac{p^2}{1-p}$	$6+\frac{p^2}{1-p}$
Entropy	$\frac{-(1-p)\log_2(1-p)-p\log_2 p}{p}$	$\frac{-(1-p)\log_2(1-p)-p\log_2 p}{p}$
MGF	$\frac{pe^t}{1-(1-p)e^t}$, for $t<-\ln(1-p)$	$\frac{p}{1-(1-p)e^t}$
CF	$\frac{pe^{it}}{1-(1-p)e^{it}}$	$\frac{p}{1-(1-p)e^{it}}$

Contents

Introduction to the geometric distribution

Consider a sequence of trials, where each trial has only two possible outcomes (designated failure and success). The probability of success is assumed to be the same for each trial. In such a sequence of trials, the geometric distribution is useful to model the number of failures before the first success. The distribution gives the probability that there are zero failures before the first success, one failure before the first success, two failures before the first success, and so on.

Examples

A newlywed couple plans to have children, and will continue until the first girl. What is the probability that there are zero boys before the first girl, one boy before the first girl, two boys before the first girl, and so on?

A doctor is seeking an anti-depressant for a newly diagnosed patient. Suppose that, of the available anti-depressant drugs, the probability that any particular drug will be effective for a particular patient is p=0.6. What is the probability that the first drug found to be effective for this patient is the first drug tried, the second drug tried, and so on? What is the expected number of drugs that will be tried to find one that is effective?

A patient is waiting for a suitable matching kidney donor for a transplant. If the probability that a randomly selected donor is a suitable match is p=0.1, what is the expected number of donors who will be tested before a matching donor is found?

Assumptions: When is the geometric distribution an appropriate model?

The geometric distribution is an appropriate model if the following assumptions are true.

- The phenomenon being modelled is a sequence of independent trials.
- There are only two possible outcomes for each trial, often designated success or failure.
- The probability of success, p, is the same for every trial.

If these conditions are true, then the geometric random variable is the count of the number of failures before the first success. The possible number of failures before the first success is 0, 1, 2, 3, and so on. The geometric random variable Y is the number of failures before the first success. In the graphs above, this formulation is shown on the right.

An alternative formulation is that the geometric random variable X is the total number of trials up to and including the first success, and the number of failures is X-1. In the graphs above, this formulation is shown on the left.

Probability of outcomes

Consider the anti-depressant example above. The probability that any given drug is effective (success) is *p* = 0.6. The probability that a drug will not be effective (fail) is *q* = 1 – *p* = 1 – 0.6 = 0.4. Here are probabilities of some possible outcomes.

(i) The first drug works. There are zero failures before the first success. Y = 0 failures. The probability P(zero failures before first success) is simply the probability that the first drug works.

$$\Pr(Y=0)=q^0\,p\ =0.4^0\times 0.6=1\times 0.6=0.6.$$

(ii) The first drug fails, but the second drug works. There is one failure before the first success. Y= 1 failure. The probability for this sequence of events is p(first drug fails) × p(second drug is success) which is given by

$$\Pr(Y=1)=q^1\,p\ =0.4^1\times 0.6=0.4\times 0.6=0.24.$$

(iii) The first drug fails, the second drug fails, but the third drug works. There are two failures before the first success. Y= 2 failures. The probability for this sequence of events is p(first drug fails) × p(second drug fails) × p(third drug is success)

$$\Pr(Y=2)=q^2\,p,=0.4^2\times 0.6=0.096.$$

The general formula to calculate the probability of *k* failures before the first success, where the probability of success is *p* and the probability of failure is *q* = 1 − *p*, is

$$\Pr(Y=k)=q^k\,p.$$

for *k* = 0, 1, 2, 3, ...

For the newlyweds awaiting their first girl, the probability of no boys before the first girl is

$$\Pr(Y=0)=q^0\,p\ =0.5^0\times 0.5=1\times 0.5=0.5.$$

The probability of one boy before the first girl is

$$\Pr(Y=1)=q^1\,p\ =0.5^1\times 0.5=0.5\times 0.5=0.25.$$

The probability of two boys before the first girl is

$$\Pr(Y=2)=q^2\,p\ =0.5^2\times 0.5=0.125.$$

and so on.

Expected number of failures before the first success

For the geometric distribution, the expected (mean) number of failures before the first success is E(*Y*) = (1 − *p*)/*p*.

For the anti-depressant example, with *p* = 0.6, the mean number of failures before the first success is E(*Y*) = (1 − *p*)/*p* = (1 − 0.6)/0.6 = 0.67.

For the kidney-donor example, with *p* = 0.1, the mean number of failures before the first success is E(*Y*) = (1 − 0.1)/0.1 = 9.

For the alternative formulation, where *X* is the number of trials up to and including the first success, the expected value is E(*X*) = 1/*p*.

Moments and cumulants

The expected value of a geometrically distributed random variable *X* is 1/*p* and the variance is (1 − *p*)/*p*²:

$$\mathrm{E}(X)=\frac{1}{p},\qquad \mathrm{var}(X)=\frac{1-p}{p^2}.$$

Similarly, the expected value of the geometrically distributed random variable *Y* = *X* − 1 (where *Y* corresponds to the pmf listed in the right column) is *q*/*p* = (1 − *p*)/*p*, and its variance is (1 − *p*)/*p*²:

$$\mathrm{E}(Y)=\frac{1-p}{p},\qquad \mathrm{var}(Y)=\frac{1-p}{p^2}.$$

Let $\mu = (1 - p)/p$ be the expected value of *Y*. Then the cumulants κ_n of the probability distribution of *Y* satisfy the recursion

$$\kappa_{n+1} = \mu(\mu+1)\frac{d\kappa_n}{d\mu}.$$

Outline of proof: That the expected value is $(1 - p)/p$ can be shown in the following way. Let *Y* be as above. Then

$$\begin{aligned}\mathrm{E}(Y) &= \sum_{k=0}^{\infty}(1-p)^k p \cdot k \\ &= p\sum_{k=0}^{\infty}(1-p)^k k \\ &= p(1-p)\sum_{k=0}^{\infty}(1-p)^{k-1}\cdot k \\ &= p(1-p)\left[\frac{d}{dp}\left(-\sum_{k=0}^{\infty}(1-p)^k\right)\right] \\ &= p(1-p)\frac{d}{dp}\left(-\frac{1}{p}\right) = \frac{1-p}{p}.\end{aligned}$$

(The interchange of summation and differentiation is justified by the fact that convergent power series converge uniformly on compact subsets of the set of points where they converge.)

Parameter estimation

For both variants of the geometric distribution, the parameter *p* can be estimated by equating the expected value with the sample mean. This is the method of moments, which in this case happens to yield maximum likelihood estimates of *p*.

Specifically, for the first variant let $k = k_1, ..., k_n$ be a sample where $k_i \geq 1$ for $i = 1, ..., n$. Then *p* can be estimated as

$$\hat{p} = \left(\frac{1}{n}\sum_{i=1}^{n}k_i\right)^{-1} = \frac{n}{\sum_{i=1}^{n}k_i}.$$

In Bayesian inference, the Beta distribution is the conjugate prior distribution for the parameter *p*. If this parameter is given a Beta(α, β) prior, then the posterior distribution is

$$p \sim \mathrm{Beta}\left(\alpha+n,\ \beta+\sum_{i=1}^{n}(k_i-1)\right).$$

The posterior mean E[*p*] approaches the maximum likelihood estimate $\hat{p}$ as α and β approach zero.

In the alternative case, let $k_1, ..., k_n$ be a sample where $k_i \geq 0$ for $i = 1, ..., n$. Then *p* can be estimated as

$$\hat{p} = \left(1+\frac{1}{n}\sum_{i=1}^{n}k_i\right)^{-1} = \frac{n}{\sum_{i=1}^{n}k_i+n}.$$

The posterior distribution of *p* given a Beta(α, β) prior is

$$p \sim \mathrm{Beta}\left(\alpha+n,\ \beta+\sum_{i=1}^{n}k_i\right).$$

Again the posterior mean E[*p*] approaches the maximum likelihood estimate $\hat{p}$ as α and β approach zero.

Other properties

- The probability-generating functions of *X* and *Y* are, respectively,

$$G_X(s) = \frac{s\,p}{1-s\,(1-p)},$$

$$G_Y(s) = \frac{p}{1-s\,(1-p)}, \quad |s| < (1-p)^{-1}.$$

- Like its continuous analogue (the exponential distribution), the geometric distribution is memoryless. That means that if you intend to repeat an experiment until the first success, then, given that the first success has not yet occurred, the conditional probability distribution of the number of additional trials does not depend on how many failures have been observed. The die one throws or the coin one tosses does not have a "memory" of these failures. The geometric distribution is the only memoryless discrete distribution.
- Among all discrete probability distributions supported on {1, 2, 3, ... } with given expected value μ, the geometric distribution *X* with parameter $p = 1/\mu$ is the one with the largest entropy.
- The geometric distribution of the number *Y* of failures before the first success is infinitely divisible, i.e., for any positive integer *n*, there exist independent identically distributed random variables $Y_1, ..., Y_n$ whose sum has the same distribution that *Y* has. These will not be geometrically distributed unless *n* = 1; they follow a negative binomial distribution.

- The decimal digits of the geometrically distributed random variable *Y* are a sequence of independent (and *not* identically distributed) random variables. For example, the hundreds digit *D* has this probability distribution:

  $$\Pr(D=d) = \frac{q^{100d}}{1+q^{100}+q^{200}+\cdots+q^{900}},$$

  where $q = 1 - p$, and similarly for the other digits, and, more generally, similarly for numeral systems with other bases than 10. When the base is 2, this shows that a geometrically distributed random variable can be written as a sum of independent random variables whose probability distributions are indecomposable.

- Golomb coding is the optimal prefix code for the geometric discrete distribution.

Related distributions

- The geometric distribution *Y* is a special case of the negative binomial distribution, with $r = 1$. More generally, if $Y_1, \ldots, Y_r$ are independent geometrically distributed variables with parameter *p*, then the sum

  $$Z = \sum_{m=1}^{r} Y_m$$

  follows a negative binomial distribution with parameters *r* and *p*.[1]

- The geometric distribution is a special case of discrete compound Poisson distribution.
- If $Y_1, \ldots, Y_r$ are independent geometrically distributed variables (with possibly different success parameters p_m), then their minimum

  $$W = \min_{m \in 1,\ldots,r} Y_m$$

  is also geometrically distributed, with parameter $p = 1 - \prod_m (1 - p_m)$.

- Suppose $0 < r < 1$, and for $k = 1, 2, 3, \ldots$ the random variable X_k has a Poisson distribution with expected value r^k/k. Then

  $$\sum_{k=1}^{\infty} k\,X_k$$

  has a geometric distribution taking values in the set {0, 1, 2, ...}, with expected value $r/(1 - r)$.

- The exponential distribution is the continuous analogue of the geometric distribution. If *X* is an exponentially distributed random variable with parameter λ, then

  $$Y = \lfloor X \rfloor,$$

  where $\lfloor\ \rfloor$ is the floor (or greatest integer) function, is a geometrically distributed random variable with parameter $p = 1 - e^{-\lambda}$ (thus $\lambda = -\ln(1 - p)$[2]) and taking values in the set {0, 1, 2, ...}. This can be used to generate geometrically distributed pseudorandom numbers by first generating exponentially distributed pseudorandom numbers from a uniform pseudorandom number generator: then $\lfloor \ln(U)/\ln(1-p) \rfloor$ is geometrically distributed with parameter p, if U is uniformly distributed in [0,1].

- If $p = 1/n$ and *X* is geometrically distributed with parameter *p*, then the distribution of X/n approaches an exponential distribution with expected value 1 as $n \to \infty$, since

  $$\begin{aligned} P(X>a) = (1-p)^a &= \left(1-\frac{1}{n}\right)^{n\frac{1}{n}(a)} = \left[\left(1-\frac{1}{n}\right)^n\right]^{\frac{1}{n}(a)} \\ &\to [e^{-1}]^{\frac{1}{n}(a)} = e^{-\frac{1}{n}a} \text{ as } n \to \infty. \end{aligned}$$

Computer software for the geometric distribution

Geometric distribution using R

The R function `dgeom(k, prob)` calculates the probability that there are k failures before the first success, where the argument "prob" is the probability of success on each trial.

For example,

```
dgeom(0,0.6) = 0.6
```

```
dgeom(1,0.6) = 0.24
```

R uses the convention that k is the number of failures, so that the number of trials up to and including the first success is *k* + 1.

The following R code creates a graph of the geometric distribution from *Y* = 0 to 10, with *p* = 0.6.

```
Y=0:10
```

```
plot(Y, dgeom(Y,0.6), type="h", ylim=c(0,1), main="Geometric distribution for p=0.6", ylab="P(Y=Y)", xlab="Y=Number of failures before first success")
```

Geometric distribution using Excel

The geometric distribution, for the number of failures before the first success, is a special case of the negative binomial distribution, for the number of failures before s successes.

The Excel function `NEGBINOMDIST(number_f, number_s, probability_s)` calculates the probability of k = number_f failures before s = number_s successes where p = probability_s is the probability of success on each trial. For the geometric distribution, let number_s = 1 success.

For example,

`=NEGBINOMDIST(0, 1, 0.6)` = 0.6

`=NEGBINOMDIST(1, 1, 0.6)` = 0.24

Like R, Excel uses the convention that k is the number of failures, so that the number of trials up to and including the first success is k + 1.

See also

- Hypergeometric distribution
- Coupon collector's problem
- Compound Poisson distribution
- Negative binomial distribution

References

1. Pitman, Jim. Probability (1993 edition). Springer Publishers. pp 372.
2. "Wolfram-Alpha: Computational Knowledge Engine" (http://www.wolframalpha.com/input/?i=inverse+p+=+1+-+e%5E-l). *www.wolframalpha.com*.

External links

- "Geometric distribution" (http://planetmath.org/?op=getobj&from=objects&id=3456). *PlanetMath*.
- Geometric distribution (http://mathworld.wolfram.com/GeometricDistribution.html) on MathWorld.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Geometric_distribution&oldid=895500580"

This page was last edited on 4 May 2019, at 18:00 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.