Distribution Basics

All statistical methods are based on random sampling. There are 4 functions used to characterize the behaviour of random variables:

- The probability density function,
- The cumulative distribution function,
- The reliability function, and
- The hazard function.

In addition to these distributions, the expected values of distribution parameters are used to describe data sets.

Statistical methods are used to describe populations by using samples. If the population of interest is small enough, statistical methods are not required; every member of the population can be measured. If every member of the population has been measured, a confidence interval for the mean is not necessary, because the mean is known with certainty (ignoring measurement error).

For statistical methods to be valid, all samples must be chosen randomly. Suppose the time to fail for 60-watt lightbulbs is required to be greater than 100 hours, and the customer verifies this on a monthly basis by testing 20 lightbulbs. If the first 20 lightbulbs produced each month are tested, any inferences about the reliability of the lightbulbs produced during the entire month would be invalid, because the 20 light-bulbs tested do not represent the entire month of production. For a sample to be random, every member of the population must have an equal chance of being selected for testing.

Now suppose that when 20 light-bulbs, randomly selected from an entire month's production, are placed on a test stand, the test is ended after 10 of the lightbulbs fail. An example of what this test data might look like is given in the table below.

23 hours	63 hours	90 hours
39 hours	72 hours	96 hours
41 hours	79 hours	Ten units survive for 96 hours without failing
58 hours	83 hours	

What is the average time to fail for the light bulbs? It is obviously not the average time to fail of the 10 failed light bulbs, because if testing was continued until all 20 lightbulbs failed, ten data points would be added to the data set that are all greater than any of the ten previous data points. When testing is ended before all items fail, the randomness of the sample has been destroyed. Since only ten of the lightbulbs failed, these ten items are NOT a random sample that is representative of the population. The initial sample of 20 items is a random sample representative of the population. By ending testing after 10 items have failed, the randomness has been destroyed by systematically selecting the 10 items with the smallest times to fail.

The situation described above is called ***censoring***. Statistical inferences can be made using censored data, but special techniques are required. The situation described above is *right* censoring; the time to fail for a portion of the data is not known, but it is known that the time to fail is greater than a given value. Right censored data may be either *time* censored of *failure* censored. If testing is ended after a predetermined amount of time, it is time censored. If testing is ended after a predetermined number of failures, the data is *failure* censored. The data in the table above are failure censored. Time censoring is also known as Type I censoring, and failure censoring is also known as Type II censoring.

The opposite of right censoring is *left* censoring; For a portion of the data the absolute value is not known, but it is known that the absolute value is *less* than a given value. An example of this is an instrument used to measure the concentration of chemicals in solution. If the instrument cannot detect the chemical below certain concentrations, it does not mean there is no chemical present, but that the level of chemical is below the detectable level of the instrument.

Another type of right censoring is *multiple* censoring. Multiple censoring occurs when items are removed from testing at more than one point in time. Field data is often multiply censored. An example of multiple censoring is given in the table below. A "+" next to a value indicates the item was removed from testing at that time without failing.

112 hours	172 hours	220 hours
145 + hours	183 hours	225 + hours
151 hours	184 + hours	225 + hours
160 hours	191 hours	225 + hours
166 + hours	199 hours	225 + hours

There are two types of data commonly used in reliability engineering: ***continuous*** and ***discrete***. Continuous variables are unlimited in their degree of precision. For example, a rod may be 5 inches long, or 5.01 inches long, or 5.001 inches long. It is impossible to state that a rod is exactly 5 inches long, only that the length of the rod falls within a specific interval. Discrete variables are limited to specific values. For example, if a die is rolled, the result is either 1, 2, 3, 4, 5, or 6. There is no possibility of obtaining any value other than these six values.

Expectation

Several terms are used to describe distributions, among the most common are the mean, variance, skewness, and kurtosis. These descriptors are derived from moment generating functions. Readers with an engineering background may recall that the center of gravity of a shape is:

$$cog = \frac{\int_{-\infty}^{\infty} x f(x)\,dx}{\int_{-\infty}^{\infty} f(x)\,dx}$$

The mean, or average, of a distribution is its center of gravity. In the equation above, the denominator is equal to the area below *f*(*x*), which is equal to one for valid probability distributions. The numerator in the equation above is the first moment generating function about the origin. Thus, the mean of a distribution can be determined from the expression:

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

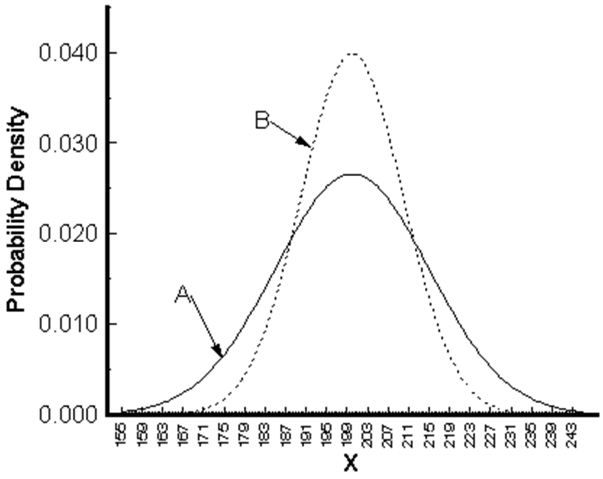
The second moment generating function about the origin is:

$$\mu = \int_{-\infty}^{\infty} x^2 f(x) dx$$

The variance of a distribution is equal to the second moment generating function about the *mean*, which is:

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

The variance is a measure of the dispersion in a distribution. In the figure below, the variance of distribution *A* is greater than the variance of distribution *B*.



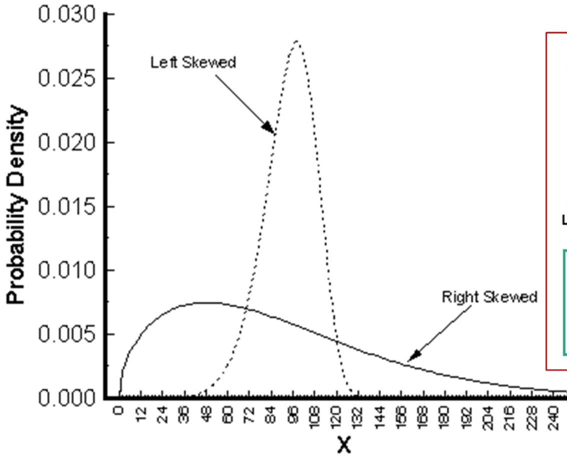
Another common measure of dispersion is the standard deviation. The standard deviation is equal to the square root of the variance. The distribution standard deviation is estimated from the *sample* standard deviation.

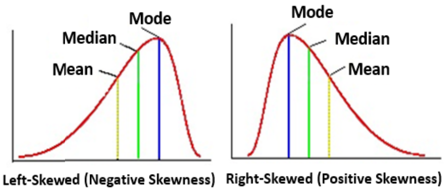
$$\hat{\sigma} = s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}}$$

Note: The sample standard deviation can be found using the Excel function "=STDEV()".

Interactive Example http://www.engineeredsoftware.com/iqf/variance.asp?LessThan=70&B1=Update+Graph

The skewness of a distribution is equal to the third moment generating function about the *mean.* If the skewness is positive, the distribution is *right* skewed. If the skewness is negative, the distribution is *left* skewed. Right and left skewness are demonstrated in the figure below.





Left-Skewed (Negative Skewness) Right-Skewed (Positive Skewness)

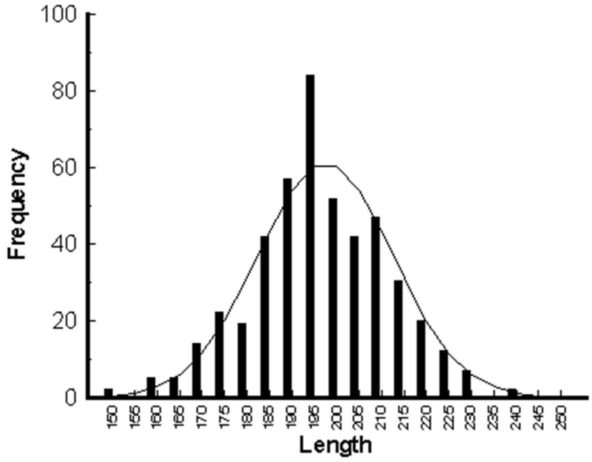
- A *symmetrical distribution* has a skew of zero. A positive result means that your data is positively skewed. A negative results means that your data is negatively skewed.

Interactive Example http://www.engineeredsoftware.com/iqf/skewness.asp

Kurtosis is the fourth moment generating function about the mean and is a measure of the peaked ness of the distribution.

Probability Density Function

The probability density function, $f(x)$, describes the behaviour of a random variable. Typically, the probability density function is viewed as the shape of the distribution. Consider the histogram of the length of fish shown in the figure below.



The *probability density function* is like the overlaid model in the figure above. The area below the probability density function to the left of a given value, x, is equal to the probability of the random variable represented on the x-axis being less than the given value x. Since the probability density function represents the entire sample space, the area under the probability density function must equal one. Since negative probabilities are impossible, the probability density function, $f(x)$, must be positive for all values of x. Stating these two requirements mathematically,

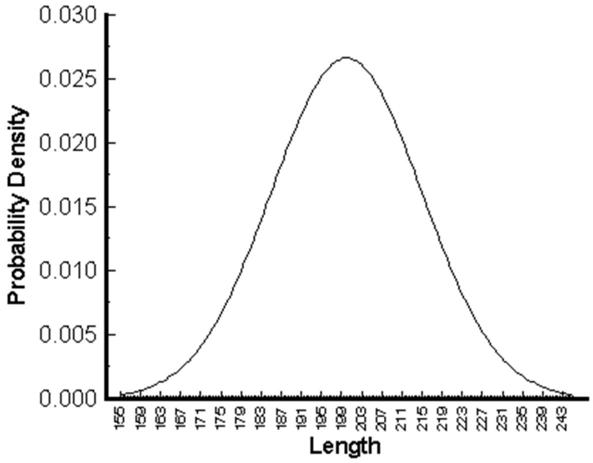
$$\int_{-\infty}^{\infty} f(x) = 1$$

and $f(x)> 0$ for continuous distributions. For discrete distributions

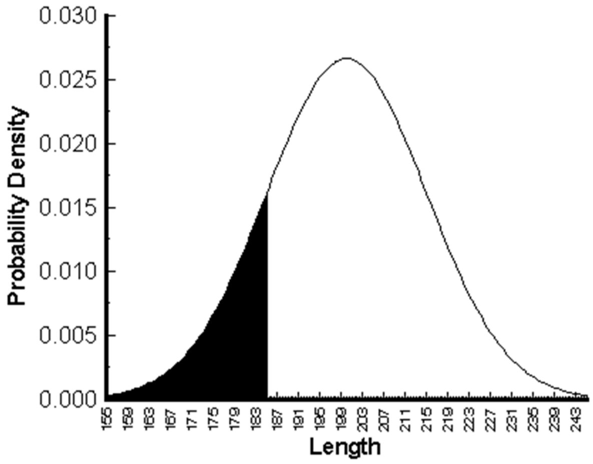
$$\sum^{n} f(x) = 1$$

for all values of n, and $f(x)> 0$.

The area below the smooth curve in the figure above is greater than one, thus this curve is not a valid probability density function. The density function representing the data in this figure is shown below.

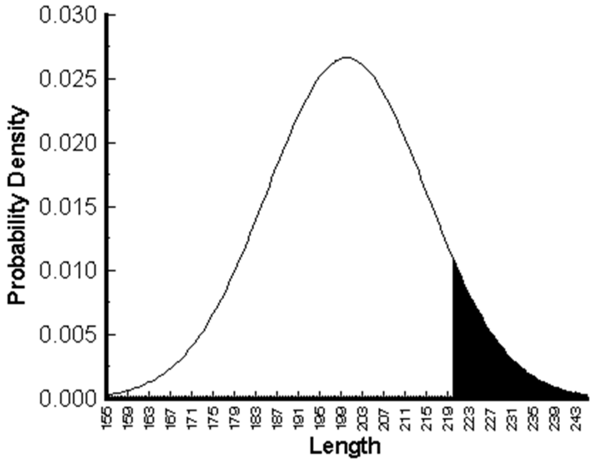


The figure below demonstrates how the probability density function is used to compute probabilities. The area of the shaded region represents the probability of a single fish, drawn randomly from the population having a length less than 185. This probability is 15.9%.



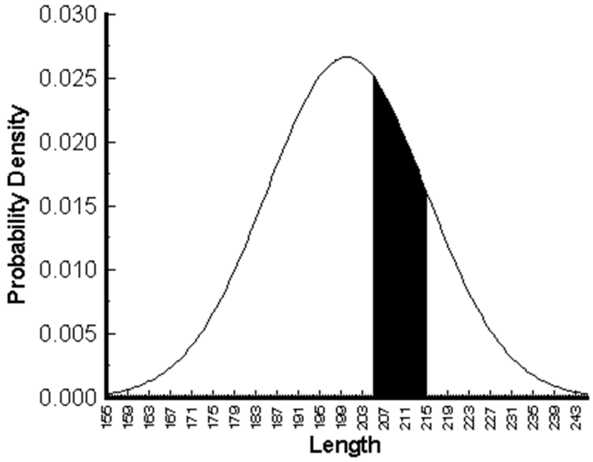
Interactive Example http://www.engineeredsoftware.com/iqf/probabil.asp

The figure below demonstrates the probability of the length of one randomly selected fish having a length greater than 220.



Interactive Example **http://www.engineeredsoftware.com/iqf/probabil2.asp**

The area of the shaded region in the figure below demonstrates the probability of the length of one randomly selected fish having a length greater than 205 and less than 215. Note that as the width of the interval decreases, the area, and thus the probability of the length falling in the interval decreases. This also implies that the probability of the length of one randomly selected fish having a length exactly equal to a specific value is zero. This is because the area of a line is zero.



Interactive Example **http://www.engineeredsoftware.com/iqf/probabil3.asp**

Example: A probability density function is defined as

f(*x*) = a/*x* , 1<*x*<10

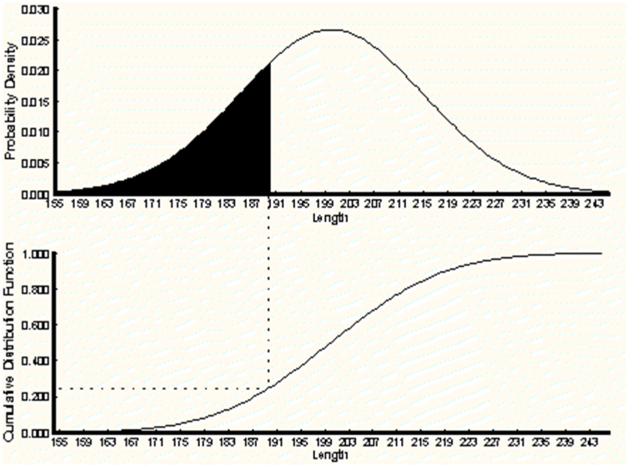
For *f*(*x*) to be a valid probability density function, what is the value of *a*?

Solution: To be a valid probability density function, all values of *f*(*x*) must be positive, and the area beneath *f*(*x*) must equal one. The first condition is met by restricting *a* and *x* to positive numbers. To meet the second condition, the integral of *f*(*x*) from one to ten must equal 1.

$$\int_1^{10} \frac{a}{x} dx = 1$$
$$a\ln(10) - a\ln(1) = 1$$
$$a = \frac{1}{\ln(10)}$$

Cumulative Distribution Function

The cumulative distribution function, *F*(*x*), denotes the area beneath the probability density function to the left of *x*. This is demonstrated in the figure below.



Cumulative Distribution Function

The area of the shaded region of the probability density function in the figure is 0.2525. This is the corresponding value of the cumulative distribution function at *x* = 190. Mathematically, the cumulative distribution function is equal to the integral of the probability density function to the left of *x*.

$$F(x) = \int_{-\infty}^{x} f(\tau)\,d\tau$$

Interactive Example http://www.engineeredsoftware.com/iqf/cumulati1.asp

Example: A random variable has the probability density function *f*(*x*) = 0.125*x,* where *x* is valid from 0 to 4. What is the probability of *x* being less than or equal to 2?

Solution:

$$F(2) = \int_{0}^{2} 0.125x \; dx = \frac{0.125x^2}{2}\Big|_{0}^{2} = 0.0625x^2\Big|_{0}^{2} = 0.25$$

Example: The time to fail for a transistor has the following probability density function. What is the probability of failure before *t* = 200?

$$f(t) = 0.01e^{-0.01t}$$

Solution:

The probability of failure before *t* = 200 is

The integral of *f*(*t*) is:

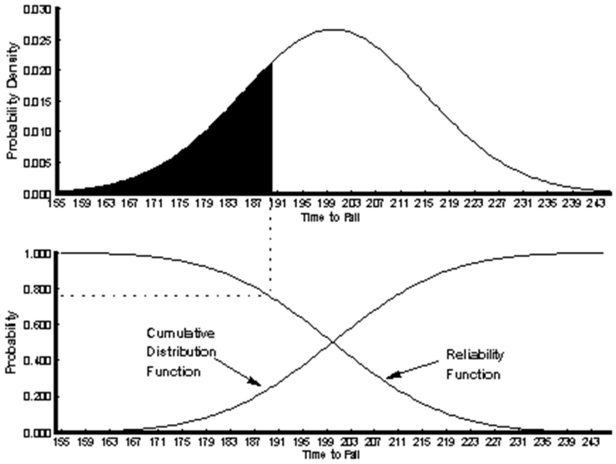
-exp(-0.01*t*)

evaluating this expression from 0 to 200 is

P(*t*<200) = -exp[-0.01(200)] - {-exp[-0.01(0)]} = -0.135 - (1) = 0.865

Reliability Function

The reliability function is the complement of the cumulative distribution function. If modelling the time to fail, the cumulative distribution function represents the probability of failure and the reliability function represents the probability of survival. Thus, the cumulative distribution function increases from zero to one as the value of *x* increases, and the reliability function decreases from one to zero as the value of *x* increases. This is shown in the figure below.



As seen from the figure above the probability that the time to fail is greater than 190 is 0.7475 which is the reliability at time = 190. The probability that the time to fail is less than 190, the cumulative distribution function, is 1 - 0.7475 = 0.2525. Mathematically, the reliability function is the integral of the probability density function from *x* to infinity.

$$R(x) = \int_x^\infty f(\tau)d\tau$$

Interactive Example http://www.engineeredsoftware.com/IQF/reliabil1.asp

Hazard Function

The hazard function is a measure of the tendency to fail, the greater the value of the hazard function, the greater the probability of impending failure. Technically, the hazard function is the probability of failure in the very small-time interval, x_0 to $x_0 + d_x$ given survival until x_0. The hazard function is also known as the instantaneous failure rate. Mathematically, the hazard function is defined as

Commented [AC1]: Impending means about to happen or forthcoming.

$$h(x) = \frac{f(x)}{R(x)}$$

Using the expression above, and the two expressions below, if either the hazard function, reliability function, or probability density function is known, the remaining two functions can be derived.

$$R(x) = e^{-\int_{-\infty}^{x} h(\tau)d\tau}$$

$$f(x) = h(x)e^{-\int_{-\infty}^{x} h(\tau)d\tau}$$

If the hazard function is increasing, failures are caused by wear out. If the hazard function is decreasing infant mortality failures are occurring. Some causes of infant mortality failures are:

- improper use,
- improper installation,
- improper setup,
- inadequate training,
- poor quality control,
- defective materials,
- power surges,
- inadequate testing, and
- damage during storage or shipping

Example: Given the hazard function, $h(x) = 2x$, derive the reliability function and the probability density function.

Solution: The reliability function is

$$R(x) = e^{-\int_{-\infty}^{x} 2x dx}$$

$$R(x) = e^{-x^2}$$

The probability density function is:

$$f(x) = h(x)R(x) = 2xe^{-x^2}$$

Discrete Distributions

The Poisson, Binomial, Hypergeometric and Geometric distributions are used to model discrete data. Time to fail is continuous data, but some situations call for discrete data, such as the number of missiles required to destroy a target or the number of defects in a lot of 1000 items.

- Binomial Distribution
- Hypergeometric Distribution
- Poisson Distribution
- Geometric Distribution

Binomial Distribution

The binomial distribution is used to model situations having only 2 possible outcomes, usually labelled as success or failure. For a random variable to follow a binomial distribution, the number of trials must be fixed, and the probability of success must be equal for all trials. The binomial probability density function is:

$$P(x,n,p)=\binom{n}{x}p^{x}(1-p)^{n-x}$$

where P(*x*, *n*, *p*) is the probability of exactly *x* successes in *n* trials with a probability of success equal to p on each trial. Note that:

$$\binom{n}{x}=\frac{n!}{x!(n-x)!}$$

This notation is referred to as "*n* choose *x*” and is equal to the number of combinations of size *x* made from n possibilities. This function is found on most calculators.

Interactive Example http://www.engineeredsoftware.com/iqf/binomial_ie1.asp

The binomial cumulative distribution function is:

$$P(x,n,p)=\sum_{i=0}^{x}\binom{n}{i}p^{i}(1-p)^{n-i}$$

where P(*x*, *n*, *p*) is the probability of exactly *x* or fewer successes in *n* trials with a probability of success equal to *p* on each trial.

Interactive Example http://www.engineeredsoftware.com/iqf/binomial_ie2.asp

The mean and variance of the binomial distribution are:

$\mu = np$

$\sigma^2 = np(1 - p)$

Example

The probability of a salesman making a successful sales call is 0.2, and 8 sales calls are made in a day. What is the probability of making exactly 2 successful sales calls in a day? What is the probability of making more than 2 successful sales calls in a day?

Solution

The probability of exactly 2 successes in 8 trials is

$$P(2,8,0.2)=\binom{8}{2}0.2^2(1-0.2)^{8-2}=0.2936$$

This solution is found using the Excel function:

=BINOMDIST(2,8,0.2,0)

The probability of more than 2 successes is equal to the one minus the probability of 2 or fewer successes. The binomial probability density function can be used to compute the probability of exactly 0 successes and the probability of exactly 1 success, but it is easier to use the Excel function:

=BINOMDIST(2,8,0.2,1)

The "1" at the end of the function above returns the cumulative binomial distribution. The probability of 2 or fewer successes is 0.7969. The probability of more than 2 successes is 1-0.7969 = 0.2031.

Detailed calculations are available in the Microsoft Excel file "**Chapt-1 Discrete Distributions.xlsx**".

Before electronic spreadsheets were common, the Poisson distribution was used to approximate the binomial distribution because Poisson tables were more accommodating than binomial tables. This approximation is now useless; why approximate a value when you can get an exact answer. The requirement for a valid approximation is *p* must be small and *n* must be large. The approximation is done by using *np* as the mean of the Poisson distribution.

Hypergeometric Distribution

The hypergeometric distribution is like the binomial distribution. Both are used to model the number of successes given:

- a fixed number of trials, and
- two possible outcomes on each trial.

The difference is that the binomial distribution requires the probability of success to be the same for all trials, while the hypergeometric distribution does not. Consider drawing from a deck of cards. If 5 cards are drawn, the probability of getting exactly 2 hearts can be computed using the binomial distribution if after each draw the card is replaced in the deck and the cards are shuffled. By replacing the card and shuffling, the probability of getting a heart on each of the 5 draws remains fixed at 13/52. If the card is not replaced after each draw, the probability of getting a heart on the first draw is 13/52, but the probability of getting a heart on the second draw is dependent on the outcome of the first draw. If the first draw resulted in a heart, the probability of getting a heart on the second draw is 12/51. If the first draw did not result in a heart, the probability of getting a heart on the second draw is 13/51. The hypergeometric distribution is used to model this situation. This is also why the hypergeometric distribution is referred to as the distribution that models sampling without replacement.

The hypergeometric probability density function is:

$$P(x,N,n,m)=\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

where $p(x, N, n, m)$ is the probability of exactly x successes in a sample of n drawn from a population of N containing m successes.

Interactive Example http://www.engineeredsoftware.com/iqf/hyper_ie1.asp

The hypergeometric cumulative distribution function is:

$$P(x,N,n,m)=\sum_{i=0}^{x}\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

Interactive Example http://www.engineeredsoftware.com/iqf/hyper_ie2.asp

The mean and the variance of the hypergeometric distribution are:

$$\mu = \frac{nm}{N}$$

$$\sigma^2 = \left(\frac{nm}{N}\right)\left(1 - \frac{m}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Example

Fifty items are submitted for acceptance. If it is known that there are 4 defective items in the lot, what is the probability of finding exactly 1 defective item in a sample of 5? What is the probability of finding less than 2 defective items in a sample of 5?

Solution

For this example,

$x = 1$
$N = 50$
$n = 5$
$m = 4$

The probability of finding exactly 1 defective item in a sample of 5 is:

$$P(1{,}50{,}5{,}4) = \frac{\binom{4}{1}\binom{50-4}{5-1}}{\binom{50}{5}} = \frac{(4)(163{,}185)}{2{,}118{,}760} = 0.30808$$

This answer can be found using the Excel function:

=HYPGEOMDIST(1,5,4,50)

The probability of finding less than 2 defective items in a sample of 5 is equal to the probability of exactly zero plus the probability of exactly 1. The probability of exactly zero is:

$$p(0{,}50{,}5{,}4) = \frac{\binom{4}{0}\binom{50-4}{5-0}}{\binom{50}{5}} = \frac{(1)(1{,}370{,}754)}{2{,}118{,}760} = 0.64696$$

The probability of finding less than 2 defective items in a sample of 5 is 0.30808 + 0.64696 = 0.95504. Unfortunately, Excel does not have a cumulative form of the hypergeometric distribution.

Detailed calculations are available in the Microsoft Excel file "**Chapt-1 Discrete Distributions.xlsx**".

The binomial distribution can be used to approximate the hypergeometric distribution when the population is large with respect to the sample size. When *N* is much larger than *n*, the change in the probability of success on a single trial is too small to significantly affect the results of the calculations. Again, it is silly to use approximations when exact solutions can be found with electronic spreadsheets. These approximations were useful for the engineer toiling with a slide rule but are of little use now.

Poisson Distribution

The Poisson distribution is used to model rates, such as rabbits per acre, defects per unit, or arrivals per hour. The Poisson distribution is closely related to the exponential distribution. If *x* is a Poisson distributed random variable, the 1/*x* is an exponential random variable. If *x* is an exponential random variable, then 1/*x* is a Poisson random variable. For a random variable to be Poisson distributed, the probability of an occurrence in an interval must be proportional to the length of the interval, and the number of occurrences per interval must be independent.

The Poisson probability density function is:

$$p(x,\mu)=\frac{e^{-\mu}\mu^{x}}{x!}$$

Microsoft Excel contains the Poisson probability distribution function. The format is:

=POISSON(*x*,μ,0)

Interactive Example **http://www.engineeredsoftware.com/iqf/poissonie.asp**

The term p(x,μ) represents the probability of exactly *x* occurrences in an interval having an average of μ occurrences. The mean and variance of the Poisson distribution are both equal to μ. The Poisson cumulative distribution function is simply the sum of the Poisson probability density function from 0 to x.

$$P(x,\mu)=\sum_{i=0}^{x}\frac{e^{-\mu}\mu^{i}}{i!}$$

The cumulative Poisson distribution has been computed in many textbooks to eliminate the need for tedious calculations, but most people prefer to use the Excel formula:

=POISSON(x,μ,1)

Interactive Example **http://www.engineeredsoftware.com/iqf/poissonie2.asp**

Example
A complex software system averages 7 errors per 5,000 lines of code. What is the probability of exactly 2 errors in 5,000 lines of randomly selected lines of code?

Solution
The probability of exactly 2 errors in 5,000 lines of randomly selected lines of code is:

$$p(2,7)=\frac{e^{-7}7^{2}}{2!}=0.022$$

Detailed calculations are available in the Microsoft Excel file "**Chapt-1 Discrete Distributions.xlsx**".

Example
A complex software system averages 7 errors per 5,000 lines of code. What is the probability of exactly 3 errors in 15,000 lines of randomly selected lines of code?

Solution
The average number of errors in 15,000 lines of code is:

$$\mu=\left(\frac{7}{5{,}000}\right)(15{,}000)=21$$

The probability of exactly 3 errors in 15,000 lines of randomly selected lines of code is:

$$p(3,21)=\frac{e^{-21}21^{3}}{3!}=0.00000117$$

Example
A complex software system averages 6 errors per 5,000 lines of code. What is the probability of less than 3 errors in 2,500 lines of randomly selected lines of code? What is the probability of more than 2 errors in 2,500 lines of randomly selected lines of code?

Solution
The average number of errors in 2,500 lines of code is μ= 3. The probability of less than 3 defects is equal to the probability of exactly 0 defects plus the probability of exactly 1 defect plus the probability of exactly 2 defects. Entering a cumulative Poisson table in the with r = 2 and μ= 3 gives the probability of 2 or fewer defect which is 0.4232. This value can also be computed manually. The same solution is found using the Excel formula:

=POISSON(2,3,1)

The "1" at the end of this formula gives the cumulative Poisson.

The probability of more than 2 errors is equal to the probability of exactly 3 plus the probability of exactly 4 plus the probability of exactly 5, etc. A simpler approach is to consider that the probability of more than 2 errors is equal to one minus the probability of 2 or fewer errors. Thus, the probability of more than 2 errors is 1 - 0.4232 = 0.5768.

Geometric Distribution

The geometric distribution is similar to the binomial distribution in that the probability of occurrence is constant from trial to trial and the trials are independent. The binomial distribution models situations where the number of trials is fixed, and the random variable is the number of successes. The geometric distribution requires exactly 1 success, and the random variable is the number of trials required to obtain the first success. The geometric distribution is a special case of the negative binomial distribution. The negative binomial distribution models the number of trials required to obtain *m* successes, and *m* is not required to be equal to one.

The geometric probability density function is:

$$p(x,p) = p(1-p)^{(x-1)}$$

where $p(x, p)$ is the probability that the first success occurs on the x-th trial given a probability of success on a single trial of p.

Interactive Example **http://www.engineeredsoftware.com/iqf/geometric_ie.asp**

The probability that more than n trials is required to obtain the first success is:

$$p(x > n) = (1-p)^n$$

Interactive Example http://www.engineeredsoftware.com/iqf/geometric_ie2.asp

The mean and variance of the geometric distribution are:

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

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

Example

The probability of an enemy aircraft penetrating friendly airspace is 0.01. What is the probability that the first penetration of friendly airspace is accomplished by the 80th aircraft to attempt the penetration of friendly airspace?

Solution

The probability that the first success occurs on the 80th trial with a probability of success of 0.01 on each trial is

$$p(80, 0.01) = 0.01(1-0.01)^{(80-1)} = 0.00452$$

Detailed calculations are available in the Microsoft Excel file "**Chapt-1 Discrete Distributions.xlsx**".

Example

The probability of an enemy aircraft penetrating friendly airspace is 0.01. What is the probability that it will take more than 80 attempts to penetrate friendly airspace?

Solution

The probability that it will take more than 80 attempts to penetrate friendly airspace with a probability of success of 0.01 on each trial is:

$$p(x > 80) = (1-0.01)^{80} = 0.4475$$

Detailed calculations are available in the Microsoft Excel file "**Chapt-1 Discrete Distributions.xlsx**".