STATISTICAL METHODS IN RELIABILITY

Alexander A. Yefremov
alexyefremov@tpu.ru

Tomsk Polytechnic University, 2022

BASIC CONCEPTS OF STATISTICS

Basic Concepts of Statistics

In statistics, a *population* is a set of similar items or events which is of interest for some question or experiment.

A statistical population can be a group of existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker).

A common goal of statistical analysis is to produce information about some chosen population.

A *data sample (dataset, sample)* is a set of data collected from a statistical population by a defined procedure.

The elements of a sample are known as *sample points, sampling units or observations*.

The data sample may be drawn from a population

- without replacement (i.e. no element can be selected more than once in the same sample), in which case it is a subset of a population;
- with replacement (i.e. an element may appear multiple times in the one sample), in which case it is a *multisubset*.

Basic Concepts of Statistics

A *statistic* (singular) or *sample statistic* is a single measure of some attribute of a sample. It is calculated by applying a function (statistical algorithm) to the values of the items of the dataset.

More formally, statistical theory defines a statistic as a function of a sample where the function itself is independent of the unknown *estimands*; that is, the function is strictly a function of the data.

The term statistic is used both for the function and for the value of the function on a given sample.

A statistic is distinct from a *statistical parameter*, which is not computable in cases where the population is infinite, and therefore impossible to examine and measure all its items.

However, a statistic, when used to estimate a population parameter, is called an *estimator*.

Samples are collected and statistics are calculated from the samples, so that one can make *inferences* or *extrapolations* from the sample to the population.

For example, the arithmetic mean of a sample is called a statistic; its value is frequently used as an estimate of the mean value of all items comprising the population from which the sample is drawn.

However, the population mean is not called a statistic, because it is not obtained from a sample; instead it is called a population parameter, because it is obtained from the whole population.

Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample variance (unbiased): $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - \bar{X}^2\right)$

where *n* is a sample size, x_i – elements of a sample.

A ***histogram*** is an estimate of the probability distribution of a continuous variable.

To construct a histogram, the first step is to "bin" the range of values—that is, to divide the entire range of values into a series of intervals—and then to count how many values fall into each interval.

The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins must be adjacent, and are often (but are not required to be) of equal size.

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency—the number of cases in each bin.

A histogram may also be *normalized* to display "relative" frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

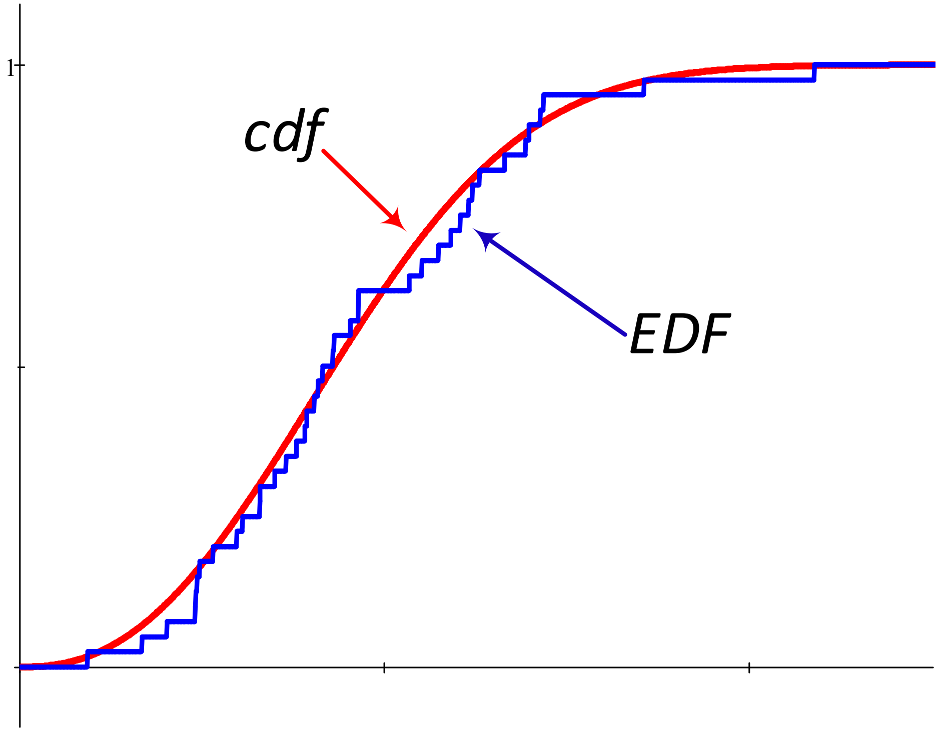
Histograms give a rough sense of the density of the underlying distribution. The total area of a histogram used for probability density is always normalized to 1.

In statistics, an *empirical distribution function (EDF)* is the distribution function associated with the empirical measure of a sample.

This cumulative distribution function is a step function that jumps up by $\frac{1}{n}$ at each of the *n* data points.

Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value.

The empirical distribution function is an *estimate* of the *cdf* that generated the points in the sample.



Let $(x_1, x_2, \dots, x_n)$ be independent, identically distributed real random variables with the common *cdf* $F(t)$. Then the *EDF* is defined as

$$\widehat{F}_n(t) = \frac{number\ of\ elements\ in\ the\ sample \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq t}$$

where $\mathbf{1}_A$ is the indicator of event *A*.

$\widehat{F}_n(t)$ is an unbiased estimator for $F(t)$. The mean of the empirical distribution is an unbiased estimator of the population's mean. The variance of the empirical distribution times $n/(n-1)$ is an unbiased estimator of the population's variance.

However, since the real life datasets have limited number of elements and often are prone to fluctuations, determining the appropriate *y* plotting positions requires some corrections when estimating distribution parameters.

The most widely used method of determining these values is the method of obtaining the ***median rank*** for each data point, as discussed next.

MEDIAN RANKS

Regarding reliability theory, the Median Ranks method is used to obtain an estimate of the unreliability for each failure.

The median rank is the value that the true probability of failure, $F\left(t_j\right)$, should have at the j-th failure out of a sample of n units at the 50% confidence level.

Median Ranks

The exact median ranks are obtained by solving the following equation for MR_j:

$$\sum_{k=j}^{n} \binom{n}{k} MR_j^k \left(1 - MR_j\right)^{n-k} = \frac{1}{2}$$

where n is the sample size, j is the order number, and MR_j is the median rank for the j-th failure.

Median Ranks

A more straightforward and easier method of estimating median ranks is by applying the following formula:

$$MR_j = \frac{1}{1 + \frac{n-j+1}{j} \cdot F^{-1}(0.5, m, n)}$$

where $m = 2 \cdot (n - j + 1)$, $n = 2j$, and $F^{-1}\ (0.5, m, n)$ - is the *quantile function of F-distribution* with *m* and *n* degrees of freedom, evaluated at the 0.5 point.

Median Ranks

Another quick, though less accurate, approximation of the median ranks is also given by:

$$MR_j = \begin{cases} 1 - 0.5^{1/n} & if\ j = 1; \\ \dfrac{j - 0.3175}{n + 0.365} & if\ j = 2,3,\dots,n-1; \\ 0.5^{1/n} & if\ j = n. \end{cases}$$

This formula is known as *Filliben's estimate*.

Benard's approximation:

$$MR_j = \frac{j - 0.3}{n + 0.4}$$

Median Ranks

Ex.: Assume that six identical units are being reliability tested at the same operation stress levels.

All of these units fail during the test after operating the following number of hours: 93, 34, 16, 120, 53 and 75.

Calculate median ranks for the data points using different methods discussed in this sections. Compare the results.

Median Ranks

First, we should rank the times-to-failure in ascending order as shown next.

Time-to-failure t_j, hours	Failure Order Number out of Sample Size of 6
16	1
34	2
53	3
75	4
93	5
120	6

Median Ranks

Let's start with the simplest method – Benard's approximation:

$$MR_j = \frac{j - 0.3}{n + 0.4}$$

where $n = 6$ and j runs from 1 to n. Then

t_j, hours	Failure Order Number	Benard's Median Rank
16	1	0,10937
34	2	0,26563
53	3	0,42188
75	4	0,57813
93	5	0,73438
120	6	0,89063

Median Ranks

Now we proceed with Filliben's estimate:

$$MR_j = \begin{cases} 1 - 0.5^{1/n} & if\ j = 1; \\ \dfrac{j - 0.3175}{n + 0.365} & if\ j = 2,3, \dots, n-1; \\ 0.5^{1/n} & if\ j = n. \end{cases}$$

t_j, hours	Failure Order Number	Filliben's Median Rank
16	1	0,1091
34	2	0,26434
53	3	0,42145
75	4	0,57855
93	5	0,73566
120	6	0,8909

Median Ranks

Next, we compute median ranks using the formula with F-distribution.

$$MR_j = \frac{1}{1 + \frac{n-j+1}{j} \cdot F^{-1}(0.5, m, n)} \qquad \begin{aligned} m &= 2(n-j+1) \\ n &= 2j \end{aligned}$$

The values of quantile function of F-distribution can be determined with Mathcad function *qF*.

Order Number	m	n	qF(0.5,m,n)
1	12	2	1.36097
2	10	4	1.11257
3	8	6	1.02975
4	6	8	0.97111
5	4	10	0.89882
6	2	12	0.73477

Median Ranks

Then the median ranks are:

t_j, hours	Failure Order Number	Median Rank
16	1	0.1091
34	2	0.26445
53	3	0.42141
75	4	0.57859
93	5	0.73555
120	6	0.8909

Finally, we obtain exact median ranks by solving the cumulative binomial equation:

$$\sum_{k=j}^{n} \binom{n}{k} MR_j^k \left(1 - MR_j\right)^{n-k} = \frac{1}{2}$$

For $j = 1$ the left-hand side of the equation can be obtained with Mathcad:

$$\sum_{k=1}^{6} \left[\mathrm{combin}(6,k)\cdot MR^k\cdot(1-MR)^{6-k}\right] \Bigg|\begin{matrix}\text{simplify}\\ \text{collect}, MR\end{matrix} \rightarrow 6\cdot MR^5 - MR^6 - 15\cdot MR^4 + 20\cdot MR^3 - 15\cdot MR^2 + 6\cdot MR$$

However, the equation itself is of no particular interest to us. We can obtain its solution as follows:

$$\sum_{k=1}^{6}\left[\mathrm{combin}(6,k)\cdot Z^{k}\cdot(1-Z)^{6-k}\right]-\frac{1}{2}\ \left|\begin{matrix}\mathrm{solve},Z\\ \mathrm{float},5\end{matrix}\right.\rightarrow\begin{pmatrix}1.8909\\ \boxed{0.1091}\\ 1.4454+0.77154i\\ 0.55455-0.77154i\\ 0.55455+0.77154i\\ 1.4454-0.77154i\end{pmatrix}$$

As you see, only one root satisfies the conditions for median ranks: real number between 0 and 1.

Median Ranks

If we continue with this algorithm for all *j*'s, we get:

t_j, hours	Failure Order Number	Median Rank
16	1	0.1091
34	2	0.26445
53	3	0.42141
75	4	0.57859
93	5	0.73555
120	6	0.8909

Median Ranks

Now we compare the values of the median ranks obtained by different methods:

Benard	Filliben	F-Distr.	Binom.
0.10937	0.1091	0.1091	0.1091
0.26563	0.26434	0.26445	0.26445
0.42188	0.42145	0.42141	0.42141
0.57813	0.57855	0.57859	0.57859
0.73438	0.73566	0.73555	0.73555
0.89063	0.8909	0.8909	0.8909

As you can see, obtaining median ranks with F-distribution gives you the same result as with the exact median ranks.

Filliben's and Benard's formulae provide results close to the exact one, with Benard's being slightly less accurate.

Median Ranks

Note that median ranks *are not the replacement for EDF* since they serve different purposes.

Time-to-failure t_j, hours	EDF	Median Rank
16	0.16667	0.1091
34	0.33333	0.26445
53	0.5	0.42141
75	0.66667	0.57859
93	0.83333	0.73555
120	1	0.8909