Noise is drawn from a distribution and the nature of that distribution is dependent on the type of noise you are measuring. In probability theory and statistics, a probability distribution identifies either the probability of each value of an unidentified random variable (when the variable is discrete), or the probability of the value falling within a particular interval (when the variable is continuous). The probability function describes the range of possible values that a random variable can attain and the probability that the value of the random variable is within any (measurable) subset of that range.

Introduction

When the random variable takes values in the set of real numbers, the probability distribution is completely described by the cumulative distribution function, whose value at each real x is the probability that the random variable is smaller than or equal to x. The concept of the probability distribution and the random variables which they describe underlies the mathematical discipline of probability theory, and the science of statistics. There is spread or variability in almost any value that can be measured in a population (e.g. height of people, durability of a metal, etc.); almost all measurements are made with some intrinsic error; in physics many processes are described probabilistically, from the kinetic properties of gases to the quantum mechanical description of fundamental particles. For these and many other reasons, simple numbers are often inadequate for describing a quantity, while probability distributions are often more appropriate.

There are various probability distributions that show up in various different applications. One of the more important ones is the normal distribution, which is also known as the Gaussian distribution or the bell curve, and approximates many different naturally occurring distributions. The toss of a fair coin yields another familiar distribution, where the possible values are heads or tails, each with probability 1/2.

Several examples of distributions:

1. Normal
2. Random (or uniform)
3. Poisson
4. Laplacian
Example 1

To see the uniform distribution (aka completely random) type in matlab:

- \texttt{R=rand(1000,1)};
- \texttt{plot(R)}
- \texttt{hist(R,2000)};

EXAMPLE 2: To see the normal distribution (aka Gaussian) type (in matlab):

- \texttt{N=randn(1000,1)};
- \texttt{plot(N)}
- \texttt{hist(N,2000)};

2. Standard deviations

In probability and statistics, the standard deviation is a measure of the dispersion of a collection of values. It can apply to a probability distribution, a random variable, a population or a data set. The standard deviation is usually denoted with the letter \( \sigma \) (lowercase sigma). It is defined as the root-mean-square (RMS) deviation of the values from their mean, or as the square root of the variance. Formulated by Galton in the late 1860s, the standard deviation remains the most common measure of statistical dispersion, measuring how widely spread the values in a data set are. If many data points are close to the mean, then the standard deviation is small; if many data points are far from the mean, then the standard deviation is large. If all data values are equal, then the standard deviation is zero. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.

When only a sample of data from a population is available, the population standard deviation can be estimated by a modified standard deviation of the sample, explained below. The standard deviation of a probability distribution is the same as that of a random variable having that distribution. The standard deviation \( \sigma \) of a real-valued random variable \( X \) is defined as:

\[
\begin{array}{l}
\sigma & = & \sqrt{\operatorname{E}((X - \operatorname{E}(X))^2)} \\
& = & \sqrt{\operatorname{E}(X^2) - (\operatorname{E}(X))^2}.
\end{array}
\]

where \( \operatorname{E}(X) \) is the expected value of \( X \) (another word for the mean), often indicated with the Greek letter \( \mu \).

For a real data set:

File:C:/DOCUME~1/user/LOCALS~1/Temp/msohtmlclip1/01/clip_image003.jpg
Example 2

Suppose we wished to find the standard deviation of the data set consisting of the values 3, 7, 7, and 19.

Step 1: find the arithmetic mean (average) of 3, 7, 7, and 19,

- \((3 + 7 + 7 + 19) / 4 = 9\).

Step 2: find the deviation of each number from the mean,

- \(3 - 9 = -6\)
- \(7 - 9 = -2\)
- \(7 - 9 = -2\)
- \(19 - 9 = 10\).

Step 3: square each of the deviations, which amplifies large deviations and makes negative values positive,

- \((-6)^2 = 36\)
- \((-2)^2 = 4\)
- \((-2)^2 = 4\)
- \(10^2 = 100\).

Step 4: find the mean of those squared deviations,

- \((36 + 4 + 4 + 100) / 4 = 36\).

Step 5: take the non-negative square root of the quotient (converting squared units back to regular units),

- \(\sqrt{36} = 6\).

So, the standard deviation of the set is 6. This example also shows that, in general, the standard deviation is different from the mean absolute deviation (which is 5 in this example).

Standard Error

The standard error of a method of measurement or estimation is the estimated standard deviation of the error in that method. Specifically, it estimates the standard deviation of the difference between the measured or estimated values and the true values. Notice that the true value of the standard deviation is usually unknown and the use of the term standard error carries with it the idea that an estimate of this unknown quantity is being used. It also carries with it the idea that it measures not the standard deviation of the estimate itself but the standard deviation of the error in the estimate, and
these can be very different.

In applications where a standard error is used, it would be good to be able to take proper account of the fact that the standard error is only an estimate. Unfortunately this is not often possible and it may then be better to use an approach that avoids using a standard error, for example by using maximum likelihood or a more formal approach to deriving confidence intervals. One well-known case where a proper allowance can be made arises where the Student's t-distribution is used to provide a confidence interval for an estimated mean or difference of means. In other cases, the standard error may usefully be used to provide an indication of the size of the uncertainty, but its formal or semi-formal use to provide confidence intervals or tests should be avoided unless the sample size is at least moderately large. Here "large enough" would depend on the particular quantities being analysed.

### Standard Error of the Mean

The standard error of the mean (SEM), an unbiased estimate of expected error in the sample estimate of a population mean, is the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size (assuming statistical independence of the values in the sample):

\[
SE_{\bar{x}} = \frac{s}{\sqrt{n}}
\]

where

- \(s\) is the sample standard deviation (i.e. the sample based estimate of the standard deviation of the population), and
- \(n\) is the size (number of items) of the sample.

A practical result: Decreasing the uncertainty in your mean value estimate by a factor of two requires that you acquire four times as many samples. Worse, decreasing standard error by a factor of ten requires a hundred times as many samples.

This estimate may be compared with the formula for the true standard deviation of the mean:

\[
SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}
\]

where

- \(\sigma\) is the standard deviation of the population.

### Example 3

To see the normal distribution (aka Gaussian) type (in matlab):

- \(N = \text{randn}(1000, 1)\);
- \(\text{hist}(N, 2000)\);
- \(\text{std}(N) \) %To Calculate the Standard Deviation of the sample
Rules for normally distributed data

The central limit theorem says that the distribution of a sum of many independent, identically distributed random variables tends towards the normal distribution. If a data distribution is approximately normal then about 68% of the values are within 1 standard deviation of the mean, about 95% of the values are within two standard deviations and about 99.7% lie within 3 standard deviations. This is known as the 68-95-99.7 rule, or the empirical rule.

For various values of \( z \), the percentage of values expected to lie in the symmetric confidence interval \( (-z\sigma, z\sigma) \) are as follows:

<table>
<thead>
<tr>
<th>( z\sigma )</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1( \sigma )</td>
<td>68.27</td>
</tr>
<tr>
<td>1.645( \sigma )</td>
<td>90</td>
</tr>
<tr>
<td>1.960( \sigma )</td>
<td>95</td>
</tr>
<tr>
<td>2( \sigma )</td>
<td>95.450</td>
</tr>
</tbody>
</table>

Dark color is less than one standard deviation from the mean. For the normal distribution, this accounts for 68.27% of the set; while two standard deviations from the mean (medium and dark blue) account for 95.45%; three standard deviations (light, medium, and dark blue) account for 99.73%; and four standard deviations account for 99.994%. The two points of the curve which are one standard deviation from the mean are also the inflection points.
<table>
<thead>
<tr>
<th>$2.576\sigma$</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3\sigma$</td>
<td>99.7300</td>
</tr>
<tr>
<td>$3.2906\sigma$</td>
<td>99.9</td>
</tr>
<tr>
<td>$4\sigma$</td>
<td>99.993666</td>
</tr>
<tr>
<td>$5\sigma$</td>
<td>99.9994267</td>
</tr>
<tr>
<td>$6\sigma$</td>
<td>99.999998027</td>
</tr>
<tr>
<td>$7\sigma$</td>
<td>99.99999997440</td>
</tr>
</tbody>
</table>

**Reference**

- Wikipedia