Basic Statistics Reference

Understanding basic statistics is essential for conducting proper analysis of physical data. This document provides a reference on the concepts of probability, mean, variance, and curve-fitting.

Mean and Variance of a Probability Distribution

A probability distribution can be described by a probability function which assigns a probability $p(x)$ to each outcome x of some event. Probability distributions can either be continuous or discrete. A continuous distribution has outcomes x with probablities $p(x)$, while a discrete distribution has outcomes x_i with probabilities p_i (i is an integer label). Formulas are given here for the discrete case; to convert to the continuous case, replace

$$x_i \to x, \qquad p_i \to p(x), \qquad \sum_i \to \int dx.$$

Expected Value Let $f(x)$ be a function of the outcome of some probabilistic event. The *expected value* $E(f)$ of $f(x)$ is the average value of the quantity $f(x)$ expected over many iterations. The formula is

$$E(f) = \sum_i f(x_i)\, p_i.$$

Mean The *mean* μ of a probability distribution is the expected value of the outcome, which is simply its average value. The formula is

$$\mu = E(x) = \sum_i x_i\, p_i$$

Variance The *variance* σ^2 of a probability distribution is a measure of its "spread-out-ness". The formula is

$$\sigma^2 = E((x-\mu)^2) = \sum_i (x_i - \mu)^2\, p_i = E(x^2) - E(x)^2.$$

Standard Deviation *Standard deviation* σ is the square root of the variance, so $\sigma = \sqrt{\sigma^2}$.

Population Distribution and Sample Distribution

Suppose we know that an experiment has a set of measurement outcomes x and a probability distribution $p(x)$ for obtaining those outcomes. The underlying, idealized, probability distribution $p(x)$ is called the *population distribution*. The mean and variance associated with $p(x)$ are called the *population mean* and population variance.

Now suppose we perform the experiment N times, obtaining outcomes $x_1, x_2, \ldots, x_N$. We say that the data points x_i are "sampled from" the population distribution. The distribution of these values is called the *sample distribution*. When N is large, the sample distribution should closely approximate the population. When N is small, the sample outcomes can be less reliable. However, the sample values can always be used to approximate the population mean and population variance. Formulas are given in the following section.

Mean and Variance of a Data Set

A data set is described by a sequence $x_k = x_1, x_2, \ldots, x_N$ of N outcome values for some measurement. The proper way to analyze a data set is usually to assume it was sampled from some idealized underlying population distribution $p(x)$. The sample values can then be used to reconstruct information about the underlying distribution.

Sample Average Value Let $f(x)$ be a function of the outcome of the measurement. The *sample average* $\bar{f}$ of $f(x)$ is the average value of $f(x)$ over the measurement trials. The formula is

$$\bar{f} = \frac{1}{N}\sum_{k=1}^{N} f(x_k).$$

Sample Mean The *sample mean* $\bar{x}$ of a data set is the average value of the outcome over the trials. The formula is

$$\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k.$$

Sample Variance The *sample variance* s^2 of a data set is a measure of its "spread-out-ness". The formula is

$$s^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})^2.$$

Standard Deviation *Standard deviation* s is the square root of the variance, so $s = \sqrt{s^2}$.

Samples and Population Each of the above sample values approximates the corresponding population value when many trials are taken. That is

$$\bar{f} \approx E(f), \qquad \bar{x} \approx \mu, \qquad s^2 \approx \sigma^2,$$

with the approximation converging for large N.

Variance of the Mean Taking the average of many values increases certainty that the value is accurate. If the sample mean from N trials is calculated repeatedly, it will have a smaller variance than the samples themselves. The formula is

$$\sigma_{\bar{x}}^2 = \frac{1}{N}\sigma^2 \approx \frac{1}{N}s^2.$$

Normal Distribution

A *normal distribution* is a probability distribution such that $p(x)$ is a normalized Gaussian. The normal distribution with standard deviation σ and mean μ is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

For a normal distribution, there is approximately a 68%, 95%, and 99.7% probability of a value lying within 1, 2, and 3 standard deviations of the mean, respectively. In physics, unknown distributions are often assumed to be approximately normal. Normal dists also arise in math by the central limit theorem.

Curve Fitting

Consider a set of N data points (x_k, y_k), and a function $f(x)$. How well does the curve $f(x)$ fit the data?

Chi-Squared with Sigmas If the standard deviation σ_{y_k} of each data point is known, then one should quantify the goodness-of-fit by the quantity

$$\chi^2 = \sum_{k=1}^{N} \frac{(f(x_i) - y_i)^2}{\sigma_{y_k}^2}.$$

In this case the value χ^2 has an absolute meaning.

Chi-Squared without Sigmas If the standard deviations of the data points are not known, then one can quantify the goodness-of-fit by

$$\chi^2 = \sum_{k=1}^{N} (f(x_i) - y_i)^2.$$

In this case χ^2 has only a relative meaning. Sometimes, using an alternative denominator could restore the absolute meaning to the value, but more details about the measurement would have to be known.

Reduced Chi-Squared In the either of the above cases, the quantity

$$\bar{\chi}^2 = \frac{\chi^2}{N_{eff}}$$

is called the *reduced chi-squared*, where the effective number of degrees of freedom N_{eff} is N minus the number of parameters that were used to fit $f(x)$ to the data. A smaller value of $\bar{\chi}^2$ indicates a closer fit to the data. When χ^2 is calculated with the standard deviation values, $\bar{\chi}^2$ has an expected value of 1.

Fitting by Chi-Squared Minimization Consider a data set (x_k, y_k), and a function $f(\vec{p}, x)$ which depends on a set of parameters $\vec{p}$. This function is called a *model* for the data. The value $\chi^2(\vec{p})$ represents the goodness of fit as a function of the model parameters. Various algorithms can be used to find a set of parameters $\vec{p}_0$ which exactly or approximately minimizes χ^2 for the model. Often $f(\vec{p}_0, x)$ is then called a *best fit curve* for the model. This name is slightly misleading to students. Once a suitably well-fitting set of parameters has been found, it's not particularly important whether they are "best" or not; what matters is how well the resulting curve fits the data.

Uncertainty in Fit Parameters Suppose you find best fit parameters $\vec{p}_0$ for a model by a process of chi-squared minimization. What is the uncertainty in these parameters? Sometimes, the fitting algorithm used will output an uncertainty along with the value. This can be very useful, but make sure to know where that number comes from and what it means, as context can vary. Alternatively, you can get a quick estimate of the uncertainty by hand: change each parameter by hand until you observe a substantial change in the fit (you can do this by χ^2 or by eye) — the amount of parameter change needed to have an effect gives you a rough estimate of the uncertainty in the parameter.