A GLOSSARY OF SELECTED STATISTICAL TERMS

Harrison B. Prosper[1], *James T. Linnemann* [2], *and Wolfgang A. Rolke*[3]
[1] Department of Physics, Florida State University, Tallahassee, Florida 32306, USA
[2] Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan, USA
[3] Department of Mathematics, University of Puerto Rico at Mayaguez, Mayaguez, Puerto Rico 00681

This glossary brings together some statistical concepts that physicists may happen upon in the course of their work. The aim is not absolute mathematical precision—few physicists would tolerate such a burden. Instead, (one hopes) there is just enough precision to be clear. We begin with an introduction and a list of notations. We hope this will make the glossary, which is in alphabetical order, somewhat easier to read.

The mistakes that remain can be entirely attributed to stubbornness! (H.B.P.)

Probability, Cumulative Distribution Function and Density

1) A sample space S is the set of all possible outcomes of an experiment.
2) Probability cannot be defined on every subset of the sample space but only those sets in a σ-algebra. (A σ-algebra is the set of subsets of S, that contains the empty set, $\bar{A}$ if it contains A, $\cup A_j$ if it contains the sequence A_j.) Don't panic—this isn't used elsewhere in this glossary.
3) Statisticians define a random variable (r.v.) X as a map from the sample space into the real numbers, that is, $X : S \rightarrow R$ (at least in one dimension).
4) The (cumulative) distribution function is defined by $F(x) = P(X \leq x)$. Here x is a real number, X is a random variable, and $P(X \leq x)$ is interpreted as $P(A)$ where $A = \{X \leq x\} = \{\omega \in S : X(\omega) \leq x\}$. We talk about the probability of an event (set). Note that this definition does not distinguish between discrete and continuous random variables. The distribution may depend on parameters.
5) In the discrete case, we define the probability mass function (pmf) by

$$f(x) = P(X = x).$$

Note that if X is a continuous r.v., all these probabilities are 0. In the continuous case, we define the probability density function (pdf) by

$$F(x) = \int_{-\infty}^{x} p(z)dz,$$

or equivalently $p(x) = \frac{d}{dx}F(x)$.

Notation

$P(A)$	Probability of A
$P(AB)$	Probability of A and B
$P(A+B)$	Probability of A or B
$P(A\|B)$	Probability of A given B
$\cup A_j$	Union of sets A_j
$\cap A_j$	Intersection of sets A_j
X	Random variable
x	Particular instance of a random variable
$t(X)$	Statistic
$d(X)$	Estimator

θ	Parameters of some model M
$\hat{\theta}$	Estimate of parameter θ
$p(X\|\theta)$	Probability density of X given θ
$P(X\|\theta) = p(X\|\theta)dX$	Probability of X given θ
$D(p\|\|q)$	Kullback-Liebler divergence between densities p and q
$X \sim p(X\|\theta)$	X is distributed according to the probability density $p(X\|\theta)$
iid	identically and independently distributed
$P(\theta\|x) = p(\theta\|x)d\theta$	Posterior probability of θ given data x
$P(M)$	Prior probability of model M
$P(x\|M)$	Evidence for model M (probability of data x given model M)
$P(M\|x)$	Posterior probability of model M given data x
$b(\theta)$	Bias
$\mathcal{E}_{\mathcal{X}}[*]$	Expectation operator
$\mathcal{V}_{\mathcal{X}}[*]$	Variance operator
$L(\theta)$	Likelihood function
$\mathcal{L}(d, \theta)$	Loss function
$\mathcal{R}(\theta)$	Risk function
$\mathcal{R}_{emp}(\theta)$	Empirical risk function

GLOSSARY

Ancillary Statistic

Consider a probability density function $p(X|\theta)$. If the distribution of the statistic $t(X)$ is independent of θ and the statistic is also independent of θ, then the function $t(X)$ is said to be an ancillary statistic for θ. The name comes from the fact that such a statistic carries no information about θ itself, but may carry subsidiary (ancillary) information, such as information about the uncertainty of the estimate. *Example* In a series of n observations x_i, n is an ancillary statistic for θ. The independence of the distribution on the ancillary statistic suggests the possibility of inference conditional on the value an ancillary statistic. See also Conditioning, Distribution Free, Pivotal Quantity, Sufficient Statistic.

Bayes Factor

See Posterior Odds.

Bayes' Theorem

$$P(B|A) = P(A|B)P(B)/P(A),$$

is a direct consequence of the definition of conditional probability $P(A|B)P(B) = P(AB) = P(B|A)P(A)$. A similar algebraic structure applies to densities

$$p(z|w) = p(w|z)p(z)/p(w),$$

where $p(w) = \int p(w|z)p(z)dz$. Frequentists and Bayesians are happy to use this theorem when *both* the variables z and w are related to frequency data. Bayesians, however, are quite happy to use the theorem when w is *not* a random variable; in particular, when w may be an hypothesis or one or more parameters. See also Probability.

Bayesian

The school of statistics that is based on the *degree of belief* interpretation of probability, whose advocates included Bernoulli, Bayes, Laplace, Gauss and Jeffreys. For these thinkers, probability and frequency are considered logically distinct concepts. The dominant sub-group among Bayesians is known as the Subjective school, interpreting probability as a personal degree of belief; for these, use in a scientific setting depends on accepting conclusions shown to be robust against differing specifications of prior knowledge. Sufficient data (and the consequent peaking of the likelihood function) makes such robustness more likely. The injunction of distinguished American probabilist Mark Twain against overly-informative prior densities is apposite: “It ain’t what people don’t know that hurts them, it’s what they do know that ain’t so.” See also Prior Density, Default Prior, Posterior Density, and Exchangeability.

Bias

Let $d(X)$ be an estimator of the unknown parameter θ. The bias is defined by

$$b(\theta) \equiv \mathcal{E}_X[d(X)] - \theta,$$

where the expectation $\mathcal{E}_X[d]$ is with respect to an ensemble of random variables $\{X\}$. The bias is just the difference between the expectation value, *with respect to a specified ensemble*, of the estimator $d(X)$ and the value of the parameter being estimated. If the ensemble is not given, the bias is undefined. If an estimator is such that $b(\theta) = 0 \;\; \forall\, \theta$ then the estimator is said to be *unbiased*; otherwise, it is biased. Bias, in general, is a function of the *unknown* parameter θ and can, therefore, only be estimated. Further, bias is a property of a particular choice of metric. In high energy physics, much effort is expended to reduce bias. However, it should be noted that this is usually at the cost of increasing the variance and being further away, in the root-mean-square sense, from the true value of the parameter. See also Ensemble, Quadratic Loss Function, Re-parameterization Invariance.

Central Credible Interval

In Bayesian inference, a credible interval defined by apportioning the probability content outside the region equally to the left and right of the interval boundaries. This definition is invariant under change of variable, but may be unsatisfactory if the posterior density is heavily peaked near boundaries. See also Central Interval, Highest Posterior Density Region, Re-parameterization Invariance.

Central Interval

An interval estimate where the probability in question is intended to be in some sense centered, with the probability of being outside the interval equally disposed above and below the the interval boundaries. See Confidence Interval, Central Credible Interval.

Compound Hypothesis

See simple hypothesis.

Conditional Probability

This is defined by

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

The symbol $P(A|B)$ is read as the “probability of A given B.” The idea is very intuitive. Say we want to guess the probability whether an experiment will result in an outcome in the set A. Without any additional information this is given by $P(A)$, where this probability is computed using the sample space S. Now say we do have some more info, namely that the outcome is in the set B. Then one way to proceed is to change from the sample space S to the sample space B, and this is indicated by the notation $P(A|B)$. The formula $P(A|B) = \frac{P(A\cap B)}{P(B)}$ says that instead of computing one probability using the sample space B we can find two probabilities using the old sample space S, which is often easier. See Probability.

Conditioning

Making an inference contingent on some statistic $t(x)$ of the observed data x. Conditioning amounts to selecting that sub-ensemble, of the ensemble of all possible data X, consistent with the observed data. Bayesian inference entails the most extreme conditioning, namely, conditioning on the data observed and nothing else.

Confidence Interval

A set of random intervals $[l(X), u(X)]$, defined over an ensemble of random variables X, is said to be a set of confidence intervals if the following is true

$$\text{Prob}\{\theta \in [l(X), u(X)]\} \geq \beta \;\; \forall\, \theta \text{ and } \forall\, \lambda,$$

where θ is the parameter of interest and λ represent all nuisance parameters. The values $l(X)$ and $u(X)$ are called *confidence limits*. Confidence intervals are a frequentist concept; therefore, the probability "Prob" is interpreted as a relative frequency.

For *fixed* values of θ and λ there is an associated set $C(X, \theta, \lambda) = \{[l(X), u(X)] : \theta, \lambda = \text{constant}\}$ of intervals, of which some fraction bracket the (true) value θ. That fraction "Prob" is called the *coverage probability*. In general, as we move about the parameter space $\Omega(\theta, \lambda)$, the set $C(X, \theta, \lambda)$ changes, as does its associated coverage probability "Prob." Neyman introduced the theory of confidence intervals in 1937, requiring that their coverage probabilities *never* fall below a pre-defined value called the *confidence level* (CL), whatever the true values of all the parameters of the problem. We refer to this criterion as the *Neyman Criterion*. A set of intervals $C(X)$ is said to have *coverage*, or cover, if they satisfy the Neyman criterion. Exact coverage can be achieved for continuous variables, but for discrete variables, the interval over-covers for most true values.

To fully specify a confidence interval, the CL alone does not suffice: it merely specifies the probability content of the interval, not how it is situated. Adding a specification that during construction of an interval of size CL, the remaining probability $(1 - CL)$ is apportioned equally defines Central Intervals; other procedures provide upper ($l(X) = 0$) or lower limits ($U(X) = \infty$),or move smoothly from limits to central intervals.

Confidence Interval (CI) Estimation and Hypothesis Testing are like two sides of the same coin: if you have one you have the other. Technically, if $[l(X), u(X)]$ is a $(1 - \alpha)\,100\%$ CI for a parameter θ, and if you define $C(\theta_0) = \{X : \theta_0 \in [l(X), u(X)]\}$ then this is a critical region (see Hypothesis Testing) for testing $H_0 : \theta = \theta_0$ with level of significance α. The most important use of this duality is to find a CI: First find a test (which is often easier because we have a lot of methods for doing this) and then "invert the test" to find the corresponding CI.

It is important to note the probability statements for Confidence Intervals concern the probability that the limits calculated from data in a series of experiments surround the true with the specified probability. This is a statement about probability of data, given theory. It is not a statement about the probability that the true value lies within the limits calculated for this particular experiment (a statement of probability of theory, given data). To make such a statement, one needs the Bayesian definition of probability as a degree of belief, and a statement of one's degree of belief in the theory (or parameter values) prior to the measurement.

See also Neyman Construction, Hypothesis Testing, Re-parameterization Invariance. Contrast with Credible Region.

Confidence Level

See Confidence Interval.

Confidence Limit

See Confidence Interval.

Consistency

An estimator $d(X)$ is consistent for a parameter θ if $d(X)$ converges in probability to θ as n (the number of samples) goes to infinity, that is $P(|d(X) - \theta| > \epsilon) \to 0$ for all $\epsilon > 0$. That means both the bias and the variance also have to go to 0. Estimators obtained using Bayesian methods or maximum likelihood are usually consistent.

Coverage

See Confidence Interval.

Coverage Probability

See Confidence Interval.

Cramér-Rao Bound

See Minimum Variance Bound.

Credible Interval

See Credible Region.

Credible Region

In Bayesian inference, this is any sub-set ω of the parameter space Ω of a posterior probability $P(\theta|x)$ having a given probability content β, that is, *degree of belief*. A credible region ω is defined by

$$\beta = \int_\omega P(\theta|x) = \int_\omega p(\theta|x) d\theta.$$

If θ is one-dimensional, one speaks of a *credible interval*. The latter is the Bayesian analog of a *confidence interval*, a frequentist concept introduced by Neyman in 1937.

The above specification of probability content is insufficient to fully define the region, even in the case of a single parameter; one must further specify how to choose among the class of intervals with the correct probability content. See also Highest Posterior Density Region, Central Credible Interval.

Default Prior

Default, reference, conventional, non-informative etc. are names given to priors that try to capture the notion of indifference with respect to entertained hypotheses. Although such priors are usually improper (see Improper Priors), they are often useful in practice and practically unavoidable in complex multi-parameter problems for which subjective elicitation of prior densities is well-nigh impossible.

Distribution Free

A distribution, of a statistic $t(X)$, is said to be distribution free if it does not depend on the parameters of the underlying probability distribution of X. The classic example of such a distribution is that of the statistic $t(X) = [(X - \mu)/\sigma]^2$, where $X \sim \text{Gaussian}(\mu, \sigma)$ with σ known and μ unknown. Although the distribution of X depends on the two parameters μ and σ the distribution of $t(X)$, a χ^2 distribution, depends on neither. This is a useful feature in frequentist statistics because it allows probabilistic statements about $t(X)$ to be transformed into exact probabilistic statements about μ. See also Ancillary Statistic.

Empirical Risk Function

In many analyses, the risk function, obtained by averaging the loss function over all possible data, is usually unknown. (See Loss Function, Risk Function.) Instead, one must make do with only a *sample* $\{x_i\}$ of data, usually obtained by Monte Carlo methods. Therefore, in lieu of the risk function $\mathcal{R}(\theta)$ one is forced to use the *empirical risk function*

$$\mathcal{R}_{emp}(\theta) \equiv \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(d_i, \theta),$$

where $d_i \equiv d(x_i)$ are estimates of θ and $\mathcal{L}(d_i, \theta)$ is the loss function. *Empirical risk minimization* is the basis of many methods used in data analysis, ranging from simple χ^2 based fits to the training of neural networks.

Ensemble

One would be hard-pressed to find this term used by statisticians. But one would be even harder-pressed to excise it from the physics literature! In the context of statistics, an ensemble is the set of repeated trials or experiments, or their outcomes. To define an ensemble one must decide what aspects of an experiment are variable and what aspects are fixed. If experiments are actually repeated no difficulty arises because the actual experiments constitute the ensemble. A difficulty arises, however, if one performs only a single experiment: In that case, because the ensemble is now an abstraction, the embedding of the experiment in an ensemble becomes a matter of *debate*.

The ensemble definition is necessary, for example, to write simulations that evaluate uncertainties in frequentist error calculations, and as such typically requires definition of the relevant models leading to probability densities $p(x|\theta)$ for the contributing processes, and the Stopping Rule for data taking which defines how each experiment in the ensemble ends. See also Stopping Rule, Likelihood Principle.

Estimate

See estimator.

Estimator

Any procedure that provides estimates of the value of an unknown quantity θ. In simple cases, estimators are well-defined functions $d(X)$ of random variables X. In high energy physics, they are often complicated computer programs whose behaviour, in general, cannot be summarized algebraically. When a specific set of data x are entered into the function $d(X)$ one obtains an estimate $\hat{\theta} = d(x)$ of the value of the unknown quantity θ.

Evidence

Given prior $P(\theta|M) = p(\theta|M)d\theta$ for model M, characterized by parameters θ and the likelihood $L(\theta) \equiv p(x|\theta, M)$, the evidence $p(x|M)$ for the model M is given by

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)d\theta.$$

This is a very important quantity in Bayesian inference. See also Model Comparison.

Exchangeability

Exchangeable events are those whose probability is independent of the order in which they occur. The corresponding concept in frequentist statistics is that of independently and identically distributed events. In Bayesian statistics, de Finetti's theorem makes a connection between degree of belief probabilities and classical frequency probabilities. If the number of successful outcomes is S and the number of trials T, the theorem states that under rather general conditions, the limit

$$\lim_{T\to\infty} S/T \to r$$

exists with probability 1 for any exchangeable sequence of events. For a fuller discussion, see for example O'Hagan, Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference, Edward Arnold (1994). See also iid, Law of Large Numbers.

Expectation

$\mathcal{E}_z[*]$ is the *expectation* or *averaging operator* with respect to the variable z. Given a function $f(z)$ its expectation is

$$\mathcal{E}_z[f] = \int f(z)p(z)dz,$$

where $p(z)$ is the probability density of z. In frequentist inference z is a random variable X; in Bayesian analysis it can be a parameter.

Fisher Information

Consider the probability density $p(X|\theta)$, where both X and θ may be multi-dimensional. Define the random vector $\Theta(X) = \frac{\partial}{\partial\theta} \ln p(X|\theta)$. The Fisher information (matrix) is defined by

$$J(\Theta) = \mathcal{E}_X[\Theta(X)\Theta(X)^T].$$

Example: Let $X \sim \mathrm{Poisson}(\theta) = \exp(-\theta)\,\theta^X/X!$; then $\Theta(X) = \frac{X-\theta}{\theta}$. Therefore, $J(\Theta) = \mathcal{E}_X[(\frac{X-\theta}{\theta})^2] = 1/\theta$; J^{-1} is the variance of the Poisson distribution.

See also Likelihood, Jeffreys' Prior, Minimum Variance Bound, and Quadratic Loss Function.

Flat Prior

An attempt at specifying a default prior with minimal structure by specifying that the prior probability is independent of the parameter value. It has the seductive appeal of mathematical simplicity, but hides some pitfalls, encapsulated in the question: "flat in what variable?" If the flat prior represents your actual subjective prior knowledge of values of possible parameters, You should be ready to answer "why flat in mass, rather than cross section, ln(tanβ), or Poisson mean?" If you are consistent, you should re-express (transform) your flat prior to other variables by multiplying by a Jacobian; flat priors in different variables express *inconsistent* states of prior knowledge. Some try to justify this choice, in a particular variable, by finding numerical similarities of credible intervals or limits to confidence intervals, despite the different interpretations of these intervals. Any attempt to estimate probabilities or calculate limits by "integrating the likelihood function" has implicitly assumed a flat prior in a particular variable, and almost certainly is then vulnerable to making inconsistent (but Bayesian degree-of belief) probability statements, depending on which parameterization of the likelihood function is chosen.

See also, Bayesian, Default Prior, Improper Prior, Prior Density, Re-parameterization Invariance.

Frequentist

The school of statistics that is based on the *relative frequency* interpretation of probability, whose advocates included Boole, Venn, Fisher, Neyman and Pearson. This school sees no logical distinction between probability and frequency.

Goodness Of Fit

See P-value.

Highest Posterior Density (HPD) Region

The smallest credible region with a given probability content β. (See Credible Region.) In one dimension this region is found by minimizing the length of the interval $[a, b]$ defined by

$$\beta = \int_a^b p(\theta|x)d\theta.$$

According to this definition, disjoint regions are possible if the posterior probability is multi-modal. HPD regions are *not* invariant under a change of variable: a probability integral transform of the posterior density would render the posterior flat, rather hindering choice of a HPD region. See also Central Credible Interval, Probability Integral Transform, Re-parameterization Invariance.

Hypothesis Testing

A hypothesis is a statement about the state of nature, often about a parameter. Hypothesis testing compares twohypotheses, called the null hypothesis H_0 and the alternative hypothesis H_1. A hypothesis may be simple, such as $H_0 : p = 0.5$, or compound, such as $H_0 : p \leq 0.5$. In practice H_0 is usually a simple hypothesis whereas H_1 is often compound. The null hypothesis is the boring, bland explanation. The alternative hypothesis is why you did the research: the more interesting and exciting possibility, for which evidence must be offered. Nature, alas, is not constrained to be described by either of the hypotheses under consideration.

A hypothesis test is a procedure that decides whether H_0 or H_1 is true. The subspace of the sample space for which H_0 is rejected is called the critical region (or the rejection region). When performing a hypothesis test we face the following possibilities:

	True State of Nature	
Decision made by hypothesis test	H_0 is true	H_0 is false
Fail to reject H_0 ("Accept" H_0)	correct decision	Type II error (false negative), with probability β
Reject H_0 ("Accept" H_1)	Type I error (false positive), with probability α	correct decision

The probability α is called the significance of the test; $1 - \beta$ is called the power of the test. If H_1 is compound, the power is a function of the true parameter. More later on why one can only "Accept", not simply Accept, hypotheses.

The logic of hypothesis testing is as follows: Choose H_0, H_1, α and the test procedure. Ideally these should be chosen before looking at the data, although this is often not practical. The choice of α should be made by considering the consequences of committing the type I error (such as claiming to have discovered a new decay mode that really does not exist) or the type II error (such as not publishing a discovery and then have the glory go to somebody else). The crucial point here is that making α smaller usually means getting a larger β.

One very important, and often overlooked, point in hypothesis testing is the role of the sample size. What can we conclude after we performed a hypothesis test, and failed to reject H_0? There are two possibilities: either H_0 is actually true (and we should accept it), or we did not have the sample size necessary to reject it. As an illustration of the role of the sample size consider this mini MC: generate n observations from a normal distribution with mean 0.1 and variance 1. Then do the standard test for $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$, using $\alpha = 0.05$. Repeat this 500 times and check how often H_0, which we know is wrong, is actually rejected. The result is as follows:

sample size n	10	20	30	40	50	60	70	80
% that correctly reject H_0	29	60	79	90	93	97	98.6	98.8

Clearly whether we reject the null hypothesis depends very much on the sample size. In real life, we never know why we failed to reject H_0 and so the terminology "failed to reject H_0" really is more correct than "accept H_0".

See also Neyman-Pearson Test, Simple Hypothesis.

iid A set of measurements are iid (identically and independently distributed) if they are independent, and all governed by the same probability distribution.

Improper Prior

A prior density function that cannot be normalized to unity. A flat prior density over an infinite domain is an example of an improper prior.

Indicator Function

Any function $I(x, ..)$, of one or more variables, that assumes only two values, 0 or 1, depending on the values of the variables. An example is the Kronecker $\delta(i, j)$, which is equal to 1 if $i = j$ and 0 otherwise. Another is the Heaviside step function $h(x)$, which is 1 if $x > 0$, 0 otherwise.

Invariance

See Re-parameterization Invariance.

Jeffreys' Prior

Jeffreys suggested the following general default prior density

$$\pi(\theta) = \sqrt{J(\theta)},$$

based on the Fisher information $J(\theta)$. (See Fisher Information.) It is re-parameterization invariant in the sense that if one transforms from θ to the new set of variables ϕ the Jeffreys priors $\pi(\phi)$ and $\pi(\theta)$ are related by the Jacobian of the transformation. Many different arguments yield this prior. (See, for example, Kullback-Liebler Divergence.) However, while it works extremely well for one-dimensional problems, typically, it is less than satisfactory in higher dimensions.

Use of this prior may violate the Likelihood Principle, as the form taken by Jeffreys' Prior can depend on the stopping rule. For example, the binomial and negative binomial distributions produce different Jeffreys' priors, even though they produce likelihoods which are proportional to each other.

Jeffreys also had made other suggestions for priors in specific cases (location parameters, for example). Confusingly, these other specific suggestions (which may conflict with the general rule above) are also sometimes referred to as Jeffreys' prior or Jeffreys' rule.

See also Stopping Rule, Re-parameterization Invariance, Likelihood Principle.

Kullback-Liebler Divergence

This is a measure of the "dissimilarity", or divergence, between two densities with the property that it is zero if and only if the two densities are identical. Given two densities $p(X|\theta)$ and $q(X|\phi)$, the Kullback-Liebler divergence is given by

$$D(p||q) = \int p(X|\Theta)\ \ln[p(X|\Theta)/q(X|\Phi)]dX.$$

Because $D(p||q)$ is not symmetric in its arguments it cannot be interpreted as a "distance" in the usual sense. However, if the densities p and q are not too different, that is, $q(X|\phi) \approx p(X|\theta + \Delta\theta)$, it is possible to write $D(p||q) \approx \frac{1}{2}\Delta\theta^T J(\theta)\Delta\theta$, which may be interpreted as the invariant distance in the parameter space between the densities (Vijay Balasubramanian, adap-org/9601001). The metric turns out to be the Fisher information matrix $J(\theta)$. Consequently, it follows from differential geometry that the invariant volume in the parameter space is just $\sqrt{J(\theta)}\, d\theta$, which we recognize as none other than the Jeffreys prior. See also Jeffreys Prior.

Law of Large Numbers

There are several versions of the weak and strong laws of large numbers. We shall consider one version of the weak law. The weak law of large numbers is the statement, first proved by Jakob Bernoulli, about the *probability* that the ratio of two numbers, namely the number of successful outcomes S over the number of independent trials T, converges to the *probability* p of a success, assuming that the probability of success is the same for each trial. The statement is

$$\forall\, \epsilon > 0,\ \text{Prob}(|p - S/T| > \epsilon) \to 0 \text{ as } T \to \infty.$$

In words: The probability that S/T deviates from p goes to zero as the number of trials goes to infinity. A sequence that converges in this manner is said to *converge in probability*. This theorem provides the connection between relative frequencies S/T and probabilities p in repeated experiments in which the probability of success does not change. While the theorem provides an operational definition of the probability p, in terms of S/T, it leaves *un-interpreted* the probability "Prob." Note that it is not satisfactory to interpret "Prob" in the same way as p because that would entail an interpretation of p that is infinitely recursive. For this reason, Bayesians argue that "Prob" is to be interpreted as some sort of *degree of belief* about the statement $S/T \to p$ as $T \to \infty$.

The strong law of large numbers is similar to the weak in that it is a statement about the convergence in probability of S/T to p, except that the convergence in probability is to unity rather than zero. (See for example, E. Parzen, *Modern Probability Theory and Its Applications* (Wiley, New York, 1992), Chapter 10.)

Likelihood

The common notation for a likelihood is $L(\theta) \equiv p(x|\theta)$, found by evaluating a probability density function $p(X|\theta)$ at the observed data $X = x$. Note the distinction between the probability density function $p(X|\theta)$, which is a function of the random variable X and the parameter θ, and the likelihood function $L(\theta)$, which, because the data are fixed, is a function of θ only. In practice, the structure of the probability calculus is often clearer using the notation $p(x|\theta)$ rather than $L(\theta)$; contrast $p(\theta|x)p(x) = p(x|\theta)p(\theta)$ with $p(\theta|x)p(x) = L(\theta)p(\theta)$. If x are multiple iid observations, the likelihood may be written as a product of pdf's evaluated at the individual observations x_i. The likelihood concept was championed by Fisher. In the method of Maximum Likelihood, the value of the parameter at which the likelihood has a mode is used as an estimate of the parameter. Because of their good asymptotic properties, frequentists often use maximum likelihood estimators. See Likelihood Ratio, Likelihood Principle, and contrast with Posterior Mode.

Likelihood Principle

The principle that *inferences* ought to depend only on the data observed and relevant prior information. Thus any two choices of pdf (probability model) which produce the same or proportional likelihoods should, according to the Likelihood Principle, produce the same inference.

Acceptance of this principle does *not* imply that one must, of necessity, eschew ensembles. Indeed, ensembles must be considered in the *design* of experiments, typically, to test how well a procedure, be it frequentist or Bayesian, might be expected to perform on the average. But ensembles are not needed to effect an inference in methods, such as standard Bayesian inference, that obey the Likelihood Principle. Frequentist methods such as use of minimum variance unbiased estimators violate the Likelihood Principle. This can be seen by examining the definition of bias, which involves Expectation over all values of x of the values of the statistic. This average includes $p(x|\theta)$ for values of x other than that actually observed, and thus not part of the likelihood. See also Ensemble, Stopping Rule, and Jeffreys' Prior.

Likelihood Ratio

The ratio of two likelihoods: $\frac{L(H_1)}{L(H_2)}$. Likelihood ratios are important in many statistical procedures, such as the Neyman-Pearson test of simple hypotheses, H_1 and H_2. See also Neyman-Pearson Test, Prior and Posterior Odds, Simple Hypothesis.

Loss Function

Any function that quantifies the loss incurred in making a decision, such as deciding, given some data, on a particular value for an unknown parameter. In practice, the loss function $\mathcal{L}(d, \theta)$ is a function of the estimator $d(X)$ and the unknown quantity θ to be estimated. The loss function is a random variable by virtue of its dependence on the random variable $d(X)$. For a specific example, see Quadratic Loss Function.

Marginal Distribution

Given any distribution $p(X, Y)$ the marginal distribution is

$$p(X) = \int p(X, Y) dY.$$

Marginalization

Summation or integration over one or more variables of a probability distribution or density. Such a procedure follows directly from the rules of probability theory. In Bayesian inference it is the basic technique for dealing with nuisance parameters, λ, in a posterior probability density

$$p(\theta|x) = \int p(\theta, \lambda|x) d\lambda,$$

that is, one integrates them out of the problem to arrive at the marginal posterior density $p(\theta|x)$.

Mean

The first moment, about zero, of a distribution $p(X)$

$$\text{Mean} \equiv \int X p(X) dX.$$

Median

For a one-dimensional distribution, the median is the point at which the distribution is partitioned into two equal parts.

Minimum Variance Bound

A lower bound on the variance of an estimator, based on the Fisher Information.

The Fisher Information describes in some sense the information in a (prospective) data set. As such, it provides a bound on the variance of an estimator d for a parameter θ of the form

$$\mathcal{V}_X[d] \geq (1 + \partial b/\partial\theta)^2 / \, J(\theta) \, ,$$

where b is the bias of the estimator. That is, the parameter is better estimated when the Fisher Information is larger (for example if more measurements are made). The Fisher Information, from its definition, is clearly related to the (expected) curvature of the likelihood function, and is thus sensitive to how well-defined is the peak of the likelihood function (particular for a maximum likelihood Estimator). In the multidimensional case, one compares diagonal elements of the covariance matrix and the Fisher Information. See also Fisher Information, Variance, Bias, and Quadratic Loss Function.

Mode

The point at which a distribution assumes its maximum value. The mode depends on the metric chosen. See Re-parameterization Invariance.

Model

A model is the abstract understanding of underlying physical processes generating some or all of data of a measurement. A well specified model can be realized in a calculation leading to a pdf. This might follow directly if the model is simple, such as a process satisfying the assumptions for a Poisson distribution; or indirectly, via a Monte Carlo simulation, for a more complex model such as $p(x|\theta), \; \theta = M_{Higgs}$, for a number of potential Higgs masses. See also Model Comparison.

Model Comparison

The use of posterior probabilities $P(M_i|x)$ to rank a set of models $\{M_i\}$ according to how well each is supported by the available data x. The posterior probabilities are given by

$$P(M_i|x) = \frac{p(x|M_i)P(M_i)}{\sum_i p(x|M_i)P(M_i)},$$

where $p(x|M_i)$ is the evidence for model M_i and $P(M_i)$ is its prior probability. See also, Model, Evidence and Posterior Odds.

Moment

The r^{th} moment $M_r(a)$, about the point a, is defined by

$$M_r(a) = \int (X - a)^r p(X) dX.$$

Neyman Construction

The method by which confidence intervals are constructed. (See, for example, the discussion in G. Feldman and R. Cousins, Phys. Rev. **D57**, 3873 (1998), or the Statistics section of the current Review of Particle Properties, published by the Particle Data Group.) The theory of confidence intervals was introduced by Jerzy Neyman in a classic 1937 paper.

Neyman-Pearson Test

A frequentist test of a simple hypothesis H_0, whose outcome is either rejection or non-rejection of H_0 (for example, that an observation is from a signal with a know pdf). The test is performed against an alternative simple hypothesis H_1(for example, that the observation is due to a background with a known pdf). For two simple hypotheses, the Neyman-Pearson test is optimal in the sense that for a given probability α to commit a type I error, it achieves the smallest possible probability β to commit type II errors. (See Type I and Type II Errors.)

The test statistic is the ratio of probability densities

$$\lambda(X) = \frac{p(X|H_1)}{p(X|H_0)}.$$

The critical region is defined by $C = \{X : \lambda(X) > k\}$ with the *significance* or *size* of the test given by $\alpha = P(X \in C|H_0)$, supposing H_0 to be true. The basis of the test is to include regions of the highest λ (ratio of probability densities) first, adding regions of lower values of λ until the desired size is obtained. Thus, nowhere outside the critical region, where H_0 would fail to be rejected, is the ratio $p(X|H_1)/p(X|H_0)$ greater than in the critical region. The test, based on a likelihood ratio, is not optimal if the hypotheses are not simple. See also Hypothesis Testing, Simple Hypothesis.

Nuisance Parameter

Any parameter whose true value is unknown but which must be excised from the problem in order for an inference on the parameter of interest to be made. For example, in an experiment with imprecisely known background, that latter is a nuisance parameter.

Null Hypothesis

See P-value, Hypothesis Testing, and Neyman-Pearson Test.

Occam Factor

In Bayesian inference, the Occam factor is a quantity that implements Occam's razor: "Plurality shouldn't be posited without necessity" (William of Occam, 1285-1349). Basically, keep it simple! Consider the evidence $p(x|M) = \int p(x|\theta, M)p(\theta|M)d\theta$. Let $p(x|\hat{\theta}, M)$ be the value of the likelihood $L(\theta) \equiv p(x|\theta, M)$ at its mode. Suppose that the likelihood is tightly peaked about its mode $\hat{\theta}$ with a width $\Delta\theta$. We can write the evidence as

$$p(x|M) \approx p(x|\hat{\theta}, M) \times p(\hat{\theta}|M)\Delta\theta.$$

The factor $p(\hat{\theta}|M)\Delta\theta$ is called the *Occam factor*. Complex models tend to have prior densities spread over larger volumes of parameter space and consequently smaller values of $p(\hat{\theta}|M)$. On the other hand, a model that fits the data too well tends to yield smaller values of $\Delta\theta$. The Occam factor is seen to penalize models that are unnecessarily complex or that fit the data too well.

From the form of the Occam factor one may be tempted to conclude that the absolute value of the prior density is important. This is not so. What matters is that prior densities be correctly calibrated across the set of models under consideration. That is, the *ratio* of prior densities across models must be well-defined. See also Model, Model Comparison.

P-value

The probability that a random variable X could assume a value greater than or equal to the observed value x. Consider a hypothesis H, observed data x and a probability density function $p(X|H)$ that is contingent on the hypothesis being *true*. We suppose that $X > x$ represents values of X that are judged more extreme than that observed—usually, those values of X that render the hypothesis H less likely. The p-value is defined by

$$p \equiv \int_x^\infty p(Y|H)dY.$$

P-values are the basis of a frequentist procedure for *rejecting* H: Before an experiment, decide on a *significance* α; perform the experiment; if $p < \alpha$—implying that the data observed are considered too unlikely if H is true—one *rejects* the hypothesis H. A hypothesis that can be rejected according to this protocol is called a *null hypothesis*.

A significance test requires one to decide ahead of time the value x_0 of X, with significance $\alpha = \int_{x_0}^{\infty} p(Y|H)dY$, such that if the observed value $x > x_0$ the hypothesis is to be rejected. Clearly, this is equivalent to rejection when $p < \alpha$.

Note that smaller p-values imply *greater* evidence against the hypothesis being tested. Note also, that a *goodness of fit* test is just a particular application of a significance test. “Goodness of fit” is a misnomer because a significance test provides evidence *against* an entertained hypothesis H. So what to do if $p > \alpha$ or equivalently $x < x_0$? Do another experiment, since one may only conclude that the data and the test have failed to reject the hypothesis. The question is not the natural one “Does the curve fit?” but rather “Does the curve not fit?”!

There is an (unfortunately widespread) incorrect application of the p-value. Using the p-value approach to hypothesis testing, which is what statisticians advocate, requires the analyst to worry about the level of significance α. Indeed, deciding on an α is the very first thing one needs to do, even before any data are analyzed or maybe even before any data are taken. For example, an experiment should decide ahead of time what level of significance is required to claim a certain type of discovery. It is not correct to do the analysis, find $p = 0.0012$, say, and then decide that this is sufficient to reject H_0. What the p-value adds to the process is an idea of how close one got to rejecting H_0 instead of failing to do so, or vice versa. If one decided before hand that an appropriate α is 0.005, and then finds $p = 0.0007$, one can be much more certain of not committing a type I error (rejecting a true hypothesis) than if $p = 0.0041$, but, in either case, one would reject H_0. See also Hypothesis Testing.

Pivotal Quantity

A function of data and parameters whose distribution, given the parameters, does not depend on the value of the parameters of the sampling distribution . *Example* Suppose that $X \sim \text{Gaussian}(\mu, \sigma)$, with mean μ and known standard deviation σ. The distribution of the statistic $t(X) = [(X - \mu)/\sigma]^2$ (a χ^2 variate) is independent of μ. Therefore, $t(X)$ is a pivotal quantity for μ and σ, but not an ancillary statistic, because it includes μ and σ in its definition. Any ancillary statistic is also a pivotal quantity, but not vice versa; ancillary statistics are much rarer. Pivotal quantities may be useful in generating Confidence Limits. See also Ancillary Statistic.

Posterior Density

Given a likelihood $L(\theta) \equiv p(x|\theta)$ and prior density $p(\theta)$, the posterior density, by Bayes theorem, is

$$p(\theta|x) = \frac{L(\theta)p(\theta)}{\int L(\theta)p(\theta)d\theta} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta},$$

where θ represents one or more parameters, one or more of which could be discrete. *Example* Suppose that the likelihood depends on three classes of parameters: θ_j, λ_j and M_j, where θ_j, the parameters of interest and λ_j, the nuisance parameters, pertain to model M_j. The posterior density in this case is given by

$$p(\theta_j, \lambda_j, M_j|x) = \frac{L(\theta_j, \lambda_j, M_j)p(\theta_j, \lambda_j, M_j)}{\sum_j \int\int L(\theta_j, \lambda_j, M_j)p(\theta_j, \lambda_j, M_j)d\lambda_j d\theta_j}.$$

The posterior density can be marginalized to obtain, for example, the probability $P(M_j|x)$ of model M_j, given data x. See Marginalization, Model.

Posterior Mean

The mean of a posterior density. See Mean.

Posterior Median

The median of a posterior density. See Median.

Posterior Mode

The mode of a posterior density; it is near the maximum of the likelihood if the prior density is flat near the peak of the likelihood. See Mode.

Posterior Odds

Given models M_i and M_j, with posterior probabilities $P(M_i|x)$ and $P(M_j|x)$, respectively, the posterior odds is the ratio

$$\frac{P(M_i|x)}{P(M_j|x)} = \frac{p(x|M_i)}{p(x|M_j)}\frac{P(M_i)}{P(M_j)}.$$

The first ratio on the right, that of the evidence $p(x|M_i)$ for model M_i to the evidence $p(x|M_j)$ for model M_j, is called the *Bayes Factor*. The second ratio, $P(M_i)$ to $P(M_j)$, is called the *Prior Odds*. In words

Posterior odds = Bayes factor × Prior odds.

See also, Model, Evidence.

Power

The probability to reject false hypotheses. See also Hypothesis Testing.

Predictive Density

The probability density to observe data y given that one has observed data x. The predictive density is given by

$$p(y|x) = \int p(y|\theta)p(\theta|x)d\theta,$$

where $p(\theta|x)$ is the posterior probability density of the parameters θ. The predictive probability finds application in algorithms, such as the Kalman filter, in which one must predict where next to search for data.

Prior Density or Prior

The prior probability (and prior density for the case of continuous parameters) describe knowledge of the unknown hypothesis or parameters before a measurement. If one chooses to express subjective prior knowledge in a particular variable $\pi(\theta)$, then coherence implies that one would express that same knowledge in a different variable as by multiplying by the Jacobian: $\pi(\phi) = \pi(\theta)\partial\theta/\partial\phi$. Specification of prior knowledge can be difficult, and even controversial, particularly when trying to express weak knowledge or indifference among parameters. See also Bayesian, Default Prior, Flat Prior, Occam Factor, Re-parameterization Invariance.

Prior Odds

See Posterior Odds.

Probability

Probability is commonly defined as a measure $P(A)$ on a set $A \in S$ that satisfies the axioms

1. $P(A) \geq 0 \;\; \forall A$.
2. $P(S) = 1$.
3. $P(\cup A_j) = \sum P(A_j)$, if $A_i \cap A_j = \oslash$ for $i \neq j$.

A measure, roughly speaking, is a real-valued number that assigns a well-defined meaning to the size of a set. Probability is an abstraction of which there are several interpretations, the most common being

1. Degree of belief,
2. Relative frequency.

Statisticians sometimes denote a random variable by an uppercase symbol X and a specific value thereof by its lowercase partner x. For example, $p(X)$ represents a function of the random variable X while $p(x)$ represents its value at the specific point $X = x$.

Abstract notation	*Description*
$P(A)$	Probability of A
$P(AB)$	Probability of A **and** B
$P(A+B)$	Probability of A **or** B
$P(A\|B)$	Probability of A **given** B. This is called a conditional probability: $P(A\|B) = P(AB)/P(B)$ $P(B\|A) = P(AB)/P(A)$ which leads to Bayes' theorem: $P(B\|A) = P(A\|B)P(B)/P(A)$
Concrete notation	*Description*
$P(X\|\theta) = p(X\|\theta)dX$	$p(X\|\theta)$ is a probability density function (pdf); $P(X\|\theta)$ is a probability.
$X \sim p(X\|\theta)$	means that the variable X is distributed according to the pdf $p(X\|\theta)$.

Probability Integral Transform

For a continuous random variable, a (cumulative) distribution function

$$R = F(X|\theta) = \int_{-\infty}^{X} p(Y|\theta)dY,$$

maps X into a number R between 0 and 1 by knowledge of the density $p(X|\theta)$. If the X are distributed according to $p(X|\theta)$, the R are distributed uniformly. So if the pdf is known in one choice of variable, one can use that knowledge to transform (choose a new variable) in which the pdf is flat. A statistic formed by applying this transform to observations satisfies the definition of Pivotal Quantity, and as such can be useful for calculating confidence intervals. Such a transformation also may be used to map data of potentially infinite range into a finite range, which may be convenient during multidimensional analyses. The inversion (often numerical) of this transform is a common technique in Monte Carlo generation of random variates, as the inverse maps a uniform random number r_i into an value x_i with the distribution $p(x|\theta)$. See also Pivotal Quantity, Highest Posterior Density.

Profile Likelihood

Given a likelihood $L(\theta, \lambda)$, which is a function of the parameter θ and one or more parameters λ, the profile likelihood is $L(\theta, \hat{\lambda})$ where $\hat{\lambda}$ is the maximum likelihood estimate of the parameters λ. The profile likelihood is used in circumstances in which the exact elimination of nuisance parameters is problematic as is true, in general, in frequentist inference.

Quadratic Loss Function

This is defined by

$$\mathcal{L}(d, \theta) = (d - \theta)^2,$$

where d is an estimator of θ. The *square root* of the corresponding risk function, $\mathcal{E}_X[(d-\theta)^2]$, obtained by averaging with respect to an ensemble of possible data X, is called the *mean square error* (MSE) or *root mean square* (RMS). It is one measure (the most tractable) of average deviation of an ensemble of estimates from the true value of a parameter. The RMS, bias and variance are related as follows

$$\mathrm{RMS}^2 \equiv \mathcal{E}_X[(d-\theta)^2] = b^2(\theta) + \mathcal{V}_X(d).$$

See also Bias, Variance.

Re-parameterization Invariance

This property holds if a procedure is metric independent, that is, it produces equivalent results, no matter which variable is chosen for the analysis. For example, invariance holds if a procedure for obtaining an estimate $\hat{\theta}$, in a new variable, produces an estimate $\hat{\phi}(\hat{\theta})$, where $\phi(\theta)$ is a new variable which is a (possibly nonlinear) function of θ. More fundamentally, probability densities transform by the Jacobian, $p(z) = p(w)\partial w/\partial z$, so that the probability (integrals over the density) for regions containing values of z are equal to the probabilities for regions contain equivalent values $w(z)$, assuming w and z are related by a one to one, monotonic transformation. However, the values of the densities themselves at equivalent points are *not* the same.

Physicists tend to place more emphasis on this property than do statisticians: we are trying to understand nature, and don't want a particular choice of coordinates to change our conclusions. Maximum Likelihood Estimators have the property: the value of the likelihood itself is unchanged by re-parameterization, $L(\phi) = L(\phi(\theta))$; and since $\partial L/\partial\phi = (\partial L/\partial\theta)\partial\theta/\partial\phi$, zeros of the derivative occur at the corresponding places. By construction, frequentist Confidence Intervals also have this property: the Neyman Construction begins with probability statements for ranges of measured values for x, so with a correct change of variable in the density for x, the same probability will be found for the equivalent region in $y(x)$. For Confidence Intervals, that means, for example, $u(Y) = u(Y(X))$ However, the property of un-biased-ness is not invariant under re-parameterization: for example, the square of the expectation of X does not equal the expectation of X^2.

In the same sense, integrals on Posterior Densities (pdf's for parameters) calculated with Jeffreys' Prior also have this property, since this Prior transforms by a Jacobian, and the Likelihood is unchanged as discussed above. Similarly, subjective prior knowledge described in one variable can be transformed by a Jacobian into a corresponding Prior in another variable. Thus, with these choices of priors (Jeffreys or subjective knowledge), the posterior median, central credible intervals, and any other estimators defined as percentile points of the posterior density are invariant. However, even with these choices of priors, posterior means and modes, and Highest Posterior Density Credible Regions, are not re-parameterization invariant. See also Confidence Interval, Prior Density, Jeffreys' Prior.

Risk Function

The expectation value, $\mathcal{R}(\theta) = \mathcal{E}_X[\mathcal{L}(d,\theta)]$, with respect to an ensemble of possible sets of data X, of the loss function $\mathcal{L}(d,\theta)$. Given a risk function, the optimal estimator is defined to be that which minimizes the risk. In frequentist theory, the risk function is a function of the parameter θ. In Bayesian theory, there is a further averaging with respect to θ, in which each value of θ is weighted by the associated posterior probability $P(\theta|x)$. However, minimizing this risk function with respect to the estimator $d(X)$ is equivalent to minimizing the risk function over an ensemble containing a *single* value of $X = x$. One can summarize the situation as follows: *frequentist risk* is the loss function averaged with respect to all possible data X for *fixed* θ, while *Bayesian risk* is the loss function averaged with respect to all possible θ for *fixed* data $X = x$. This is an illustration of the fact that Bayesian inference typically obeys the *likelihood principle*, whereas frequentist inference typically does not. See Likelihood Principle.

Sample Mean

Given a random sample $x_1, x_2, \ldots, x_n$ of size n, the sample mean is just the average

$$M_n = (x_1 + x_2 + \cdots + x_n)/n,$$

of the sample. Its convergence to the true mean is governed by the law of large numbers. See Law Of Large Numbers.

Sampling Distribution

The sampling distribution is the (cumulative) distribution function of a statistic, that is, of a (possibly vector-valued) function of the data.

Sampling Rule

A rule that specifies how a random sample is to be constructed.

Significance Test

See P-value.

Simple Hypothesis

A completely specified hypothesis. Contrast the simple hypothesis $\theta = 42$ with the non-simple, that is, *compound hypothesis* $\theta > 42$. That an event is due to a signal with a known pdf with no free parameters is a simple hypothesis. That an event is due to one of two backgrounds, each with a known pdf, but whose relative normalization is unknown, is a compound hypothesis. If the relative normalization of the two backgrounds is known, this is again a simple hypothesis. See Neyman-Pearson test, Hypothesis Testing.

Statistic

Any meaningful function of the data, such as those that provide useful summaries, for example, the sample mean. See Sample Mean.

Stopping Rule

A rule that specifies the criterion (or criteria) for stopping an experiment in which data are acquired *sequentially*. It is a matter of debate whether an inference should, or should not, depend upon the stopping rule. This is related to the question of how to embed a finite sequence of experiments into an ensemble.

A classic example arose in connection with the measurement of the top quark mass by the DØ collaboration. The experimental team found 77 events, of which about 30 were estimated to be due to top quark production. To assess systematic effects and validate the methods of analysis required the embedding of the 77 events into an ensemble. The conundrum was this: Should the ensemble be binomial, in which the sample size is fixed at 77? Or should it be a Poisson ensemble with fluctuating sample size? Or, though this was not considered, should it be the ensemble pertaining to experiments that run until 77 events are found, yielding a negative binomial distribution? The answer, of course, is that there is no unique answer. Nonetheless, the choice made has consequences: the Poisson and binomial ensembles produce different likelihoods; the binomial and negative binomial produce equivalent likelihoods, but would produce different confidence intervals.

See also Ensemble, Likelihood Principle, Jeffreys' Prior.

Sufficient Statistic

If a likelihood function can be re-written solely in terms of one or more functions $t(x)$ of the observed data x then the $t(x)$ are said to be sufficient statistics. They are sufficient in the sense that their use does not entail loss of information with respect to the data x. *Example* Consider a sample $x_1, ..., x_n$ with likelihood function $L(\theta) = \prod_{i=1}^{n} \exp(-\theta)\theta^{x_i}$. This can be re-written as $L(\theta) = \exp(-n\theta)\theta^{nt}$, where the statistic t is the sample mean. Since the likelihood can be written solely in terms of n and t, these together are sufficient statistics. See also Ancillary Statistic.

Type I and Type II Errors

One commits a type I error if a true hypothesis is rejected. A type II error is committed if a false hypothesis is accepted.

Variance

The variance of a quantity d, for example an estimator $d(X)$, is defined by

$$\mathcal{V}_X[d] \equiv \mathcal{E}_X[d^2] - \mathcal{E}_X^2[d],$$

where $\mathcal{E}_X[*]$ is the *expectation* or *averaging operator* with respect to an ensemble of values X. See also Quadratic Loss Function.