Probability Theory

Luc Demortier

The Rockefeller University

INFN School of Statistics, Vietri sul Mare, June 3-7, 2013

Goals of this Lecture

This lecture is a brief overview of the results and techniques of probability theory that are most relevant for statistical inference as practiced in high energy physics today.

There will be lots of dry definitions and a few exciting theorems.

Results will be stated without proof, but with attention to their conditions of validity.

I will also attempt to describe various contexts in which each of these results is typically applied.

Useful References

1. Alan F. Karr, "Probability," Springer-Verlag New York, Inc., 1993, 282pp.
   A lucid, graduate-level introduction to the subject.
2. George Casella and Roger L. Berger, "Statistical Inference," 2nd ed., Duxbury, 2002, 660pp.
   Covers probability as an introduction to statistical inference, has good examples and clear explanations.
3. David Pollard, "A User's Guide to Measure Theoretic Probability," Cambridge University Press, 2002, 351pp.
   A modern, abstract treatment.
4. Bruno de Finetti, "Theory of Probability: A critical introductory treatment," translated by Antonio Machi and Adrian Smith, John Wiley & Sons Ltd., 1974, in two volumes (300pp. and 375pp.)
   "One of the great books of the world", "This book is about life: about a way of thinking that embraces all human activities" (D. V. Lindley in the Foreword).

Useful References

Most HEP introductions to statistics contain material about probability theory, see for example:

5. Glen Cowan, “Statistical Data Analysis,” Clarendon Press, Oxford, 1998, 197pp.
6. Frederick James, “Statistical Methods in Experimental Physics,” 2nd Ed., World Scientific, 2006, 345pp.

And then there is always Wikipedia of course...

Outline

What Is Probability?

For the purposes of theory, it doesn't really matter what probability "is", or whether it even exists in the real world. All we need is a few definitions and axioms:

1. The set S of all possible outcomes of a particular experiment is called the sample space for the experiment.
2. An event is any collection of possible outcomes of an experiment, that is, any subset of S (including S itself).

To each event A in sample space we would like to associate a number between zero and one, that will be called the probability of A, or $\mathbb{P}(A)$. For technical reasons one cannot simply define the domain of $\mathbb{P}$ as "all subsets of the sample space S". Care is required...

3. A collection of subsets of S is called a sigma algebra, denoted by $\mathcal{B}$, if it has the following properties:
   - $\emptyset \in \mathcal{B}$ (the empty set is an element of $\mathcal{B}$).
   - If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation).
   - If $A_1, A_2, \cdots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

What Is Probability?

4. Given a sample space S and an associated sigma algebra $\mathcal{B}$, a probability function is a function $\mathbb{P}$ with domain $\mathcal{B}$ that satisfies:
   - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{B}$.
   - $\mathbb{P}(S) = 1$.
   - If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.
     This is known as the axiom of *countable additivity*. Some statisticians find it more plausible to work with *finite additivity*: If $A \in \mathcal{B}$ and $B \in \mathcal{B}$ are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. Countable additivity implies finite additivity, but accepting only the latter makes statistical theory more complicated.

These three properties are usually referred to as the axioms of probability, or the Kolmogorov axioms. They define probability but do not specify how it should be interpreted or chosen.

What Is Probability?

There are two main philosophies for the interpretation of probability:

1. **Frequentism**
   If the same experiment is performed a number of times, different outcomes may occur, each with its own relative rate or frequency. This "frequency of occurrence" of an outcome can be thought of as a probability. In the frequentist school of statistics, the only valid interpretation of probability is as the long-run frequency of an event. Thus, measurements and observations have probabilities insofar as they are repeatable, but constants of nature do not. The Big Bang does not have a probability.
2. **Bayesianism**
   Here probability is equated with uncertainty. Since uncertainty is always *someone*'s uncertainty about *something*, probability is a property of someone's relationship to an event, not an objective property of that event itself. Probability is someone's informed degree of belief about something. Measurements, constants of nature and other parameters can all be assigned probabilities in this paradigm.

Frequentism has a more "objective" flavor to it than Bayesianism, and is the main paradigm used in HEP. On the other hand astrophysics deals with unique cosmic events and tends to use the Bayesian methodology.

What Is Probability?

From Bruno de Finetti (p. x): “My thesis, paradoxically, and a little provocatively, but nonetheless genuinely, is simply this:

PROBABILITY DOES NOT EXIST.

The abandonment of superstitious beliefs about the existence of Phlogiston, the Cosmic Ether, Absolute Space and Time, ... , or Fairies and Witches, was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs.”

“In investigating the reasonableness of our own modes of thought and behaviour under uncertainty, all we require, and all that we are reasonably entitled to, is consistency among these beliefs, and their reasonable relation to any kind of relevant objective data (‘relevant’ in as much as subjectively deemed to be so). This is Probability Theory. In its mathematical formulation we have the Calculus of Probability, with all its important off-shoots and related theories like Statistics, Decision Theory, Games Theory, Operations Research and so on.”

Random Variables

1. A random variable X is a mapping from the sample space S into the real numbers. Given a probability function on S, it is straightforward to define a probability function on the range of X. For any set A in that range:

$$\mathbb{P}_X(X \in A) = \mathbb{P}(\{s \in S : X(s) \in A\}).$$

   The induced probability $\mathbb{P}_X$ satisfies the probability axioms.

2. A random variable N is discrete if there exists a countable set C such that $\mathbb{P}_N(C) = 1$.

3. A random variable X is absolutely continuous if there exists a positive function f_X on $\mathbb{R}$, the probability density function of X, such that for every interval $(a, b]$,

$$\mathbb{P}((a, b]) = \int_a^b f_X(t)\, dt.$$

4. With every random variable X is associated a cumulative distribution function:

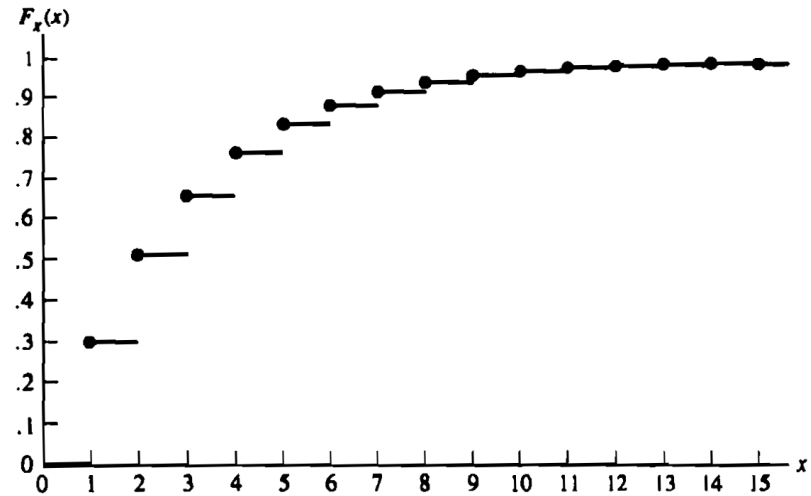
$$F_X(x) \equiv \mathbb{P}_X(X \leq x)$$

Cumulative Distribution Functions

Example: In a coin-tossing experiment, let p be the probability of heads, and define the random variable X to be the number of tosses required to get a head. This yields a geometric distribution: $\mathbb{P}(X = x) = (1-p)^{x-1}p$, and

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i=1}^{x} \mathbb{P}(X = i) = \sum_{i=1}^{x} (1-p)^{i-1}p = 1 - (1-p)^x.$$

Example for $p = 0.3$:



Note the three properties of a cdf:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
2. $F(x)$ is a non-decreasing function of x.
3. $F(x)$ is right-continuous: $\lim_{x \downarrow x_0} F(x) = F(x_0)$ for every x_0.

Quantiles

The inverse of a cumulative distribution function F is called quantile function:

$$F^{-1}(\alpha) \equiv \inf\{x : F(x) \geq \alpha\}, \text{ for } 0 < \alpha < 1.$$

In words, the quantile of order α of a continuous cdf F, or its α-quantile, is the value of x to the left of which lies a fraction α of the total probability under F. Instead of an α-quantile, one sometimes speaks of a 100α-percentile.

Some related concepts include:

- The median of a distribution, which is its 50th percentile.
- The lower quartile (25th percentile), and the upper quartile (75th percentile).
- A measure of dispersion that is sometimes used is the interquartile range, the distance between the lower and upper quartiles.
- A random variable with distribution function F can be constructed by applying F^{-1} to a random variable uniformly distributed on $[0, 1]$, a process known as the quantile transformation.

Probability Densities and Probability Mass Functions

We have already seen that for a continuous random variable one can write probabilities as integrals of a probability density function (pdf):

$$\mathbb{P}((a, b]) = \int_a^b f_X(t)\, dt.$$

The cdf is a special case of this equation:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

The equivalent equation for a discrete random variable involves a probability mass function (pmf) and a summation instead of an integration:

$$F_N(n) = \mathbb{P}(N \leq n) = \sum_{i=0}^{n} f_N(i),$$

where $f_N(i) \equiv \mathbb{P}(N = i)$.

Random Vectors

Random d-vectors generalize random variables: They are mappings from sample space into $\mathbb{R}^d$. Distributional concepts describing random vectors include the following:

1. The distribution of $X = (X_1, \ldots, X_d)$ is the probability $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$ on $\mathbb{R}^d$.
2. The joint distribution function of $X_1, \ldots, X_d$ is the function $F_X : \mathbb{R}^d \to [0, 1]$ given by:

$$F_X(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

3. Let X be a random d-vector. Then for each i the distribution function of component i can be recovered as follows:

$$F_{X_i}(t) = \lim_{t_j \to \infty, j \neq i} F_X(t_1, \ldots, t_{i-1}, t, t_{i+1}, \ldots, t_d).$$

4. A random vector X is discrete if there is a countable subset C of $\mathbb{R}^d$ such that $\mathbb{P}(X \in C) = 1$, and absolutely continuous if there is a function $f_X : \mathbb{R}^d \to \mathbb{R}_+$, the joint density of $X_1, \ldots, X_d$, such that:

$$\mathbb{P}(X_1 \leq t_1, \ldots, X_d \leq t_d) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_d} f_X(y_1, \ldots, y_d)\, dy_1 \ldots dy_d.$$

5. If $X = (X_1, \ldots, X_d)$ is absolutely continuous, then for each i, X_i is absolutely continuous, and

$$f_{X_i}(x) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(y_1, \ldots, y_{i-1}, x, y_{i+1}, \ldots, y_d) \times dy_1 \ldots dy_{i-1} dy_{i+1} \ldots dy_d.$$

Integrating out the variables of the joint density, other than that for X_i, yields the marginal density function of X_i. The procedure can be generalized to any subvector of X.

Expectation, Covariance, Correlation, and Independence

Let X and Y be two continuous random variables with joint pdf $f_{XY}(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$, respectively. We define:

1. The expectation of X: $\mu_X = \mathbb{E}(X) \equiv \int x\, f_X(x)\, dx$
2. The variance of X:
   $\sigma_X^2 = \mathbb{V}\mathrm{ar}(X) \equiv \mathbb{E}[(X - \mathbb{E}(X))^2] = \int (x - \mu_X)^2\, f_X(x)\, dx.$
3. The covariance of X and Y:
   $\mathbb{C}\mathrm{ov}(X, Y) \equiv \mathbb{E}[(X - \mu_X)\,(Y - \mu_Y)] = \int (x - \mu_X)\,(y - \mu_Y)\, f_{XY}(x, y)\, dx\, dy.$
4. The correlation of X and Y: $\rho_{XY} \equiv \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\sigma_X\, \sigma_Y}$. This is a number between -1 and $+1$.

In addition, we say that X and Y are independent random variables, if for every x and y,

$$f_{XY}(x, y) = f_X(x)\, f_Y(y).$$

Note that if X and Y are independent random variables, then $\mathbb{C}\mathrm{ov}(X, Y) = 0$ and $\rho_{XY} = 0$. However the converse is not true. It is possible to find uncorrelated, dependent random variables. This is because covariance and correlation only measure a particular kind of linear relationship.

The above definitions can be adapted to discrete random variables by replacing the integrals with sums.

Transformation Theory

If X is a random variable, then any function of X, say $g(X)$, is also a random variable. Often $g(X)$ itself is of interest and we write $Y = g(X)$. The probability distribution of Y is then given by

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(x \in \mathcal{X} : g(x) \in A).$$

This is all that is needed in practice. For example, if X and Y are discrete random variables, the formula can be applied directly to the probability mass functions:

$$f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x),$$

keeping in mind that $g^{-1}(y)$ is a set that may contain more than one point.

For the case that X and Y are continuous, an example will illustrate the issues. Suppose $g(X) = X^2$. We can write:

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{(y)}) \\ &= \mathbb{P}(X \leq \sqrt{y}) - \mathbb{P}(X \leq -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

The pdf of Y can now be obtained from the cdf by differentiation:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}).$$

This example can be generalized as follows. Suppose the sample space of X can be partitioned into k sets, such that $g(x)$ is a one-to-one transformation from each set onto the sample space of Y. Then:

$$f_Y(y) = \sum_{i=1}^{k} f_X\left(g_i^{-1}(y)\right) \left|\frac{d}{dy} g_i^{-1}(y)\right|,$$

where g_i is the restriction of g to the i^{th} set.

Characterization of Probability Distributions

There are several ways to characterize the probability distribution of a random variable X.

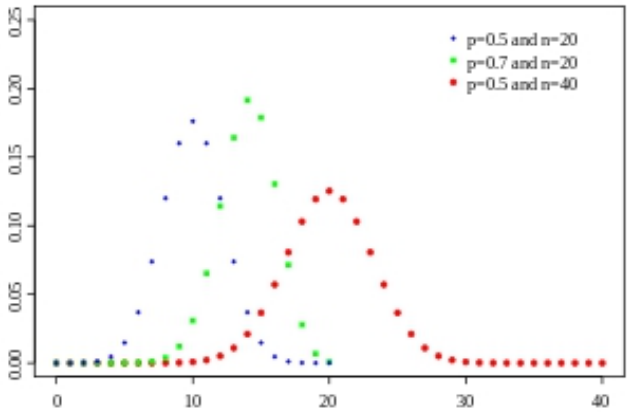
1. Functional characterizations:
   - The probability density function (pdf) for a continuous random variable, or the probability mass function (pmf) for a discrete random variable.
   - The cumulative distribution function (cdf).
   - The characteristic function, $\phi_X(t) = \mathbb{E}(e^{itX})$ (a Fourier transform).
2. Measures of location:
   - The mean: the expectation value of X, $\mathbb{E}(X)$.
   - The median: the point at which the cdf reaches 50%.
   - The mode: the location of the maximum of the pdf or pmf, when unique.
3. Measures of dispersion:
   - The variance: the expectation of $(X - \mu)^2$, where μ is the mean.
   - The standard deviation, or the square root of the variance.
4. Measures of shape:
   - The skewness, defined by $\mu_3/\mu_2^{3/2}$, where $\mu_n \equiv \mathbb{E}[(X - \mathbb{E}(X))^n]$ is the n^{th} central moment.
   - The excess kurtosis, $\mu^4/\mu_2^2 - 3$

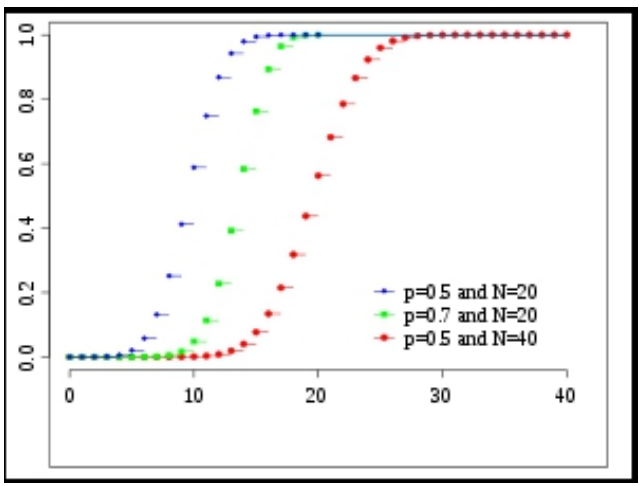
Examples of Probability Distributions

In HEP we encounter the following distributions (among others!):

1. Binomial;
2. Multinomial;
3. Poisson;
4. Uniform;
5. Exponential;
6. Gaussian;
7. Log-Normal;
8. Gamma;
9. Chi-squared;
10. Cauchy (Breit-Wigner);
11. Student's t;
12. Fisher-Snedecor.

The Binomial Distribution





Parameters:	Number of trials n, Success probability p
Support:	$k \in \{0, \ldots, n\}$
pmf:	$\binom{n}{k} p^k (1-p)^{n-k}$
cdf:	$I_{1-p}(n-k, 1+k)$
Mean:	np
Median:	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode:	$\lfloor (n+1)p \rfloor$ or $\lfloor (n+1)p \rfloor - 1$
Variance:	$np(1-p)$
Skewness:	$\frac{1-2p}{\sqrt{np(1-p)}}$
Ex. Kurtosis:	$\frac{1-6p(1-p)}{np(1-p)}$

(`http://en.wikipedia.org/wiki/Binomial_distribution`)

The Binomial Distribution

The binomial distribution comes up mostly when calculating the efficiency of an event selection. With the total number of events fixed, one counts the number of events that passed the selection and draws inferences about the "probability of success".

Another example is a study of forward-backward asymmetry. One collects N events from a given process and is looking for an asymmetry between the numbers F and B of events in the forward and backward hemispheres, respectively. The distribution of F is binomial:

$$\binom{N}{F} p^F (1-p)^B$$

with mean pN and standard deviation $\sqrt{Np(1-p)} \approx \sqrt{F(1-p)}$. The forward-backward asymmetry is usually defined as

$$R \equiv \frac{F-B}{F+B} = \frac{2F}{N} - 1$$

and has variance

$$\mathbb{V}\mathrm{ar}(R) = \frac{4p(1-p)}{N} \approx \frac{4FB}{N^3}.$$

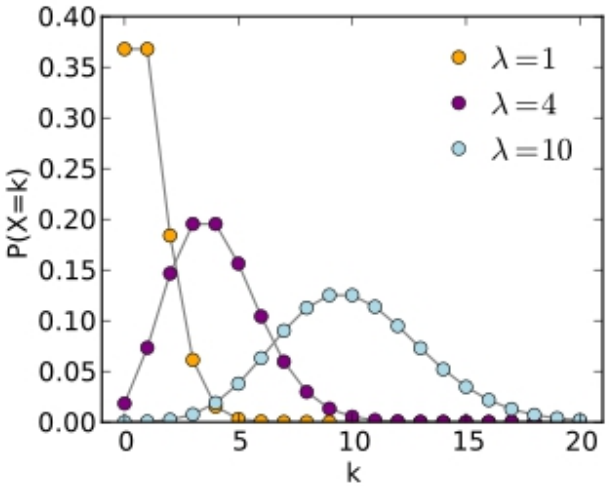
The Multinomial Distribution

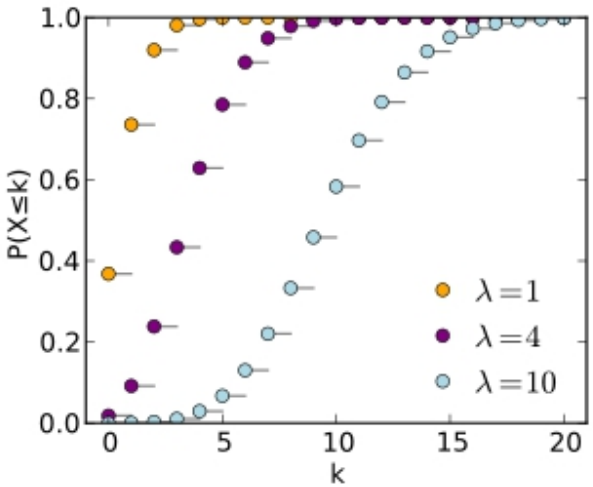
Parameters:	Number of trials n, Event probabilities $p_1, \ldots, p_k$, $\sum_{i=1}^{k} p_i = 1$
Support:	$x_i \in \{0, \ldots, n\}$, $\sum_{i=1}^{k} x_i = n$
pmf:	$\frac{n!}{x_1! \ldots x_k!} \, p_1^{x_1} \ldots p_k^{x_k}$
Mean:	$\mathbb{E}(X_i) = np_i$
Variance:	$\mathbb{V}\mathrm{ar}(X_i) = np_i(1 - p_i)$
Covariance:	$\mathbb{C}\mathrm{ov}(X_i, X_j) = -np_ip_j \ (i \neq j)$

This is the distribution of the contents of the bins of a histogram, when the total number of events n is fixed.

(`http://en.wikipedia.org/wiki/Multinomial_distribution`)

The Poisson Distribution





Parameters:	$\lambda > 0$
Support:	$k \in \{0, 1, 2, 3, \ldots\}$
pmf:	$\frac{\lambda^k}{k!}\, e^{-\lambda}$
cdf:	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$
Mean:	λ
Median:	$\approx \lfloor \lambda + \frac{1}{3} - \frac{0.02}{\lambda} \rfloor$
Mode:	$\lfloor \lambda \rfloor, \lceil \lambda \rceil - 1$
Variance:	λ
Skewness:	$\lambda^{-1/2}$
Ex. Kurtosis:	λ^{-1}

(`http://en.wikipedia.org/wiki/Poisson_distribution`)

The Poisson Distribution

The Poisson distribution is ubiquitous in HEP. It can be derived from the so-called Poisson postulates:

For each $t \geq 0$, let N_t be an integer-valued random variable with the following properties (think of N_t as denoting the no. of arrivals from time 0 to time t):

1. Start with no arrivals: $N_0 = 0$.
2. Arrivals in disjoint time periods are independent: $s < t \Rightarrow N_s \text{ and } N_t - N_s \text{ are independent.}$
3. Number of arrivals depends only on period length: N_s and $N_{t+s} - N_t$ are identically distributed.
4. Arrival probability is proportional to period length, if length is small: $\lim_{t \to 0} \frac{\mathbb{P}(N_t = 1)}{t} = \lambda$.
5. No simultaneous arrivals: $\lim_{t \to 0} \frac{\mathbb{P}(N_t > 1)}{t} = 0$.

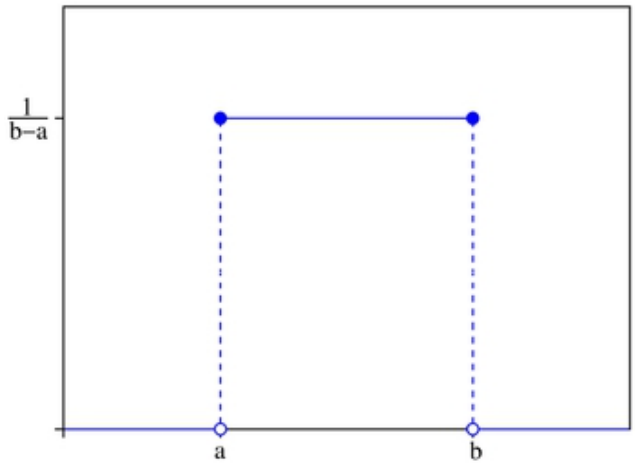
Then, for any integer n:

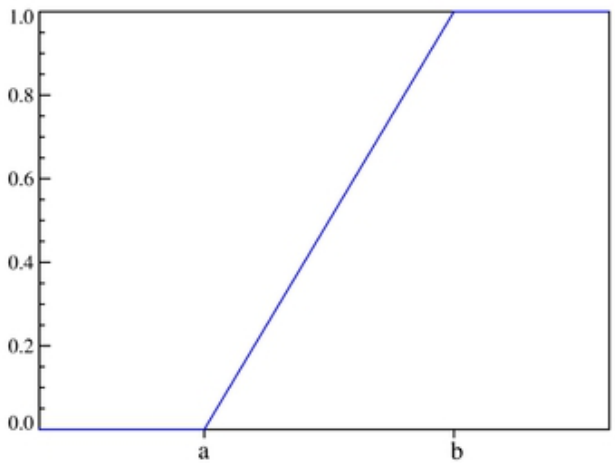
$$\mathbb{P}(N_t = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!},$$

that is, $N_t \sim \text{Poisson}(\lambda t)$.

Example: The number of particles emitted in a fixed time interval t from a radioactive source.

The Uniform Distribution

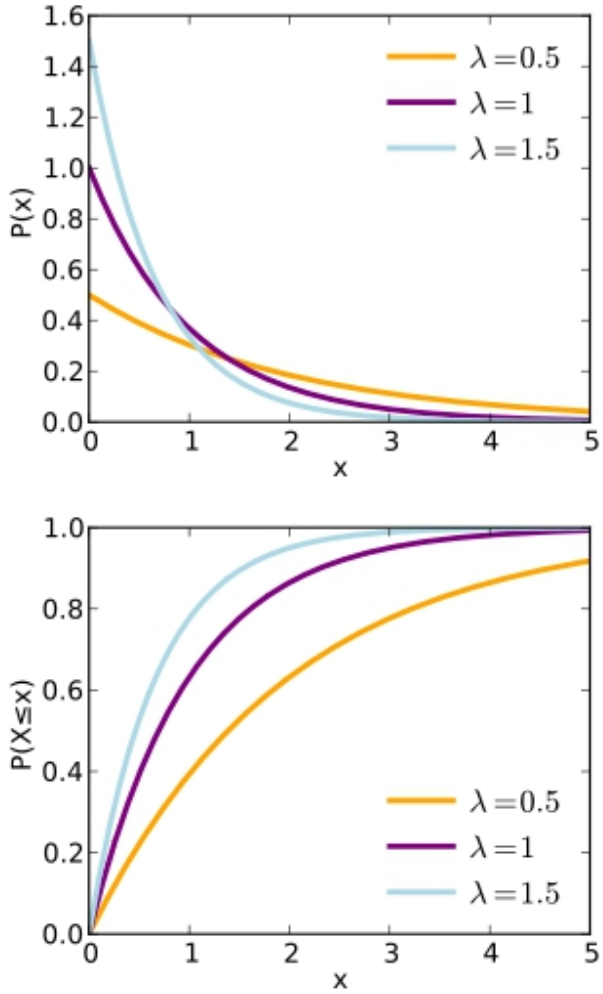




Parameters:	$a, b \in \mathbb{R}$
Support:	$x \in [a, b]$
pdf:	$\frac{1}{b-a}$ for $x \in [a, b]$, 0 otherwise
cdf:	0 for $x < a$, $\frac{x-a}{b-a}$ for $a \leq x < b$, 1 for $x \geq b$
Mean:	$\frac{1}{2}(a + b)$
Median:	$\frac{1}{2}(a + b)$
Mode:	Any value in $[a, b]$
Variance:	$\frac{1}{12}(b - a)^2$
Skewness:	0
Ex. Kurtosis:	$-\frac{6}{5}$

(http://en.wikipedia.org/wiki/Uniform_distribution_(continuous))

The Exponential Distribution



Parameters:	$\lambda > 0$
Support:	$x \geq 0$
pdf:	$\lambda e^{-\lambda x}$
cdf:	$1 - e^{-\lambda x}$
Mean:	$\frac{1}{\lambda}$
Median:	$\frac{\ln 2}{\lambda}$
Mode:	0
Variance:	$\frac{1}{\lambda^2}$
Skewness:	2
Ex. Kurtosis:	6

(`http://en.wikipedia.org/wiki/Exponential_distribution`)

The Exponential Distribution

In a Poisson process with a mean of λ events per unit time:

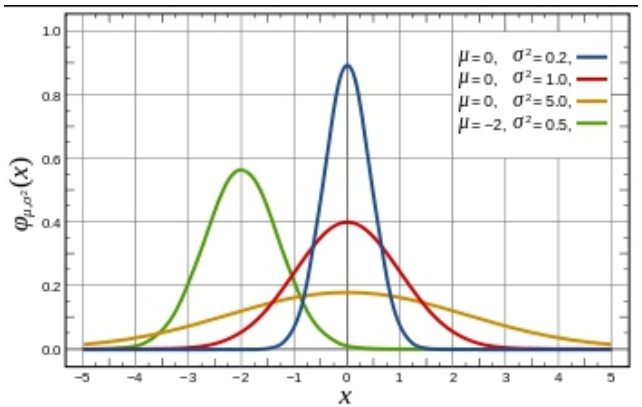
$$\mathbb{P}(N) = e^{-\lambda t} \frac{(\lambda t)^N}{N!},$$

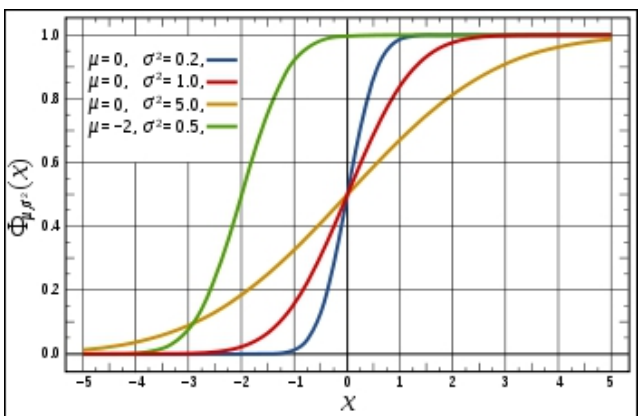
the probability of no events in time t is the exponential distribution $e^{-\lambda t}$.

For the time interval Z between two successive Poisson events, one has:

$$\mathbb{P}(Z > t) = e^{-\lambda t}.$$

The Gaussian (or Normal) Distribution





Parameters:	Mean μ, Variance $\sigma^2 > 0$
Support:	$x \in \mathbb{R}$
pdf:	$\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$
cdf:	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\,\sigma} \right)\right]$
Mean:	μ
Median:	μ
Mode:	μ
Variance:	σ^2
Skewness:	0
Ex. Kurtosis:	0

(`http://en.wikipedia.org/wiki/Normal_distribution`)

The Gaussian (or Normal) Distribution

This is the most important distribution in all of statistics, in large part due to the limit theorems, see later.

The probability content of some commonly used intervals:

$$\mathbb{P}(-1.00 \leq \frac{X-\mu}{\sigma} \leq 1.00) = 0.68 \qquad \mathbb{P}(-1.64 \leq \frac{X-\mu}{\sigma} \leq 1.64) = 0.90$$

$$\mathbb{P}(-1.96 \leq \frac{X-\mu}{\sigma} \leq 1.96) = 0.95 \qquad \mathbb{P}(-2.58 \leq \frac{X-\mu}{\sigma} \leq 2.58) = 0.99$$

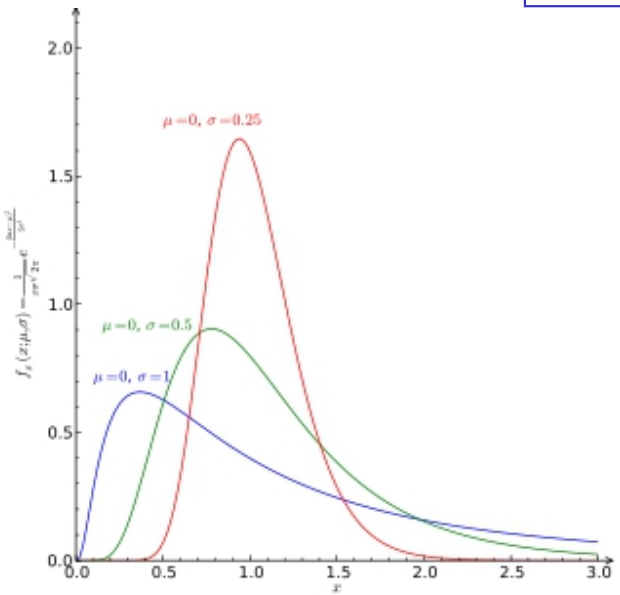
$$\mathbb{P}(-3.29 \leq \frac{X-\mu}{\sigma} \leq 3.29) = 0.999$$

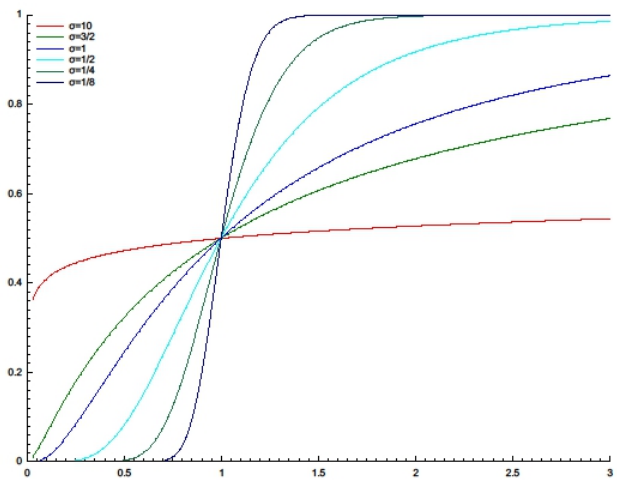
The bivariate Normal distribution has pdf

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}.$$

The Log-Normal Distribution



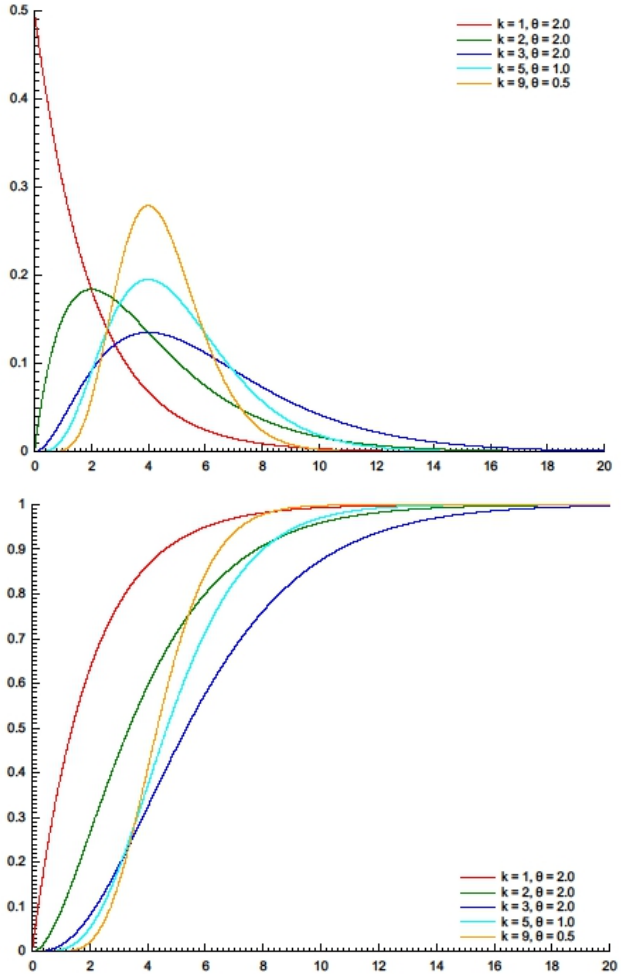


Parameters:	Log-scale μ, Shape $\sigma^2 > 0$
Support:	$x > 0$
pdf:	$\frac{1}{x\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$
cdf:	$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\ln x-\mu}{\sqrt{2}\,\sigma}\right)\right]$
Mean:	$e^{\mu+\sigma^2/2}$
Median:	e^{μ}
Mode:	$e^{\mu-\sigma^2}$
Variance:	$(e^{\sigma^2}-1)\,e^{2\mu+\sigma^2}$
Skewness:	$(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$
Ex. Kurtosis:	$e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-6$

The logarithm of a Lognormal random variable follows a Normal distribution.

(`http://en.wikipedia.org/wiki/Log-normal_distribution`)

The Gamma Distribution



Parameters:	Shape $\alpha > 0$, Rate $\beta > 0$
	or Shape $k \equiv \alpha$, Scale $\theta \equiv \frac{1}{\beta}$
Support:	$x > 0$
pdf:	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
cdf:	$\frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)}$
Mean:	$\frac{\alpha}{\beta}$
Median:	No simple closed form
Mode:	$\frac{\alpha-1}{\beta}$ for $\alpha > 1$
Variance:	$\frac{\alpha}{\beta^2}$
Skewness:	$\frac{2}{\sqrt{\alpha}}$
Ex. Kurtosis:	$\frac{6}{\alpha}$

(`http://en.wikipedia.org/wiki/Gamma_distribution`)

The Gamma Distribution

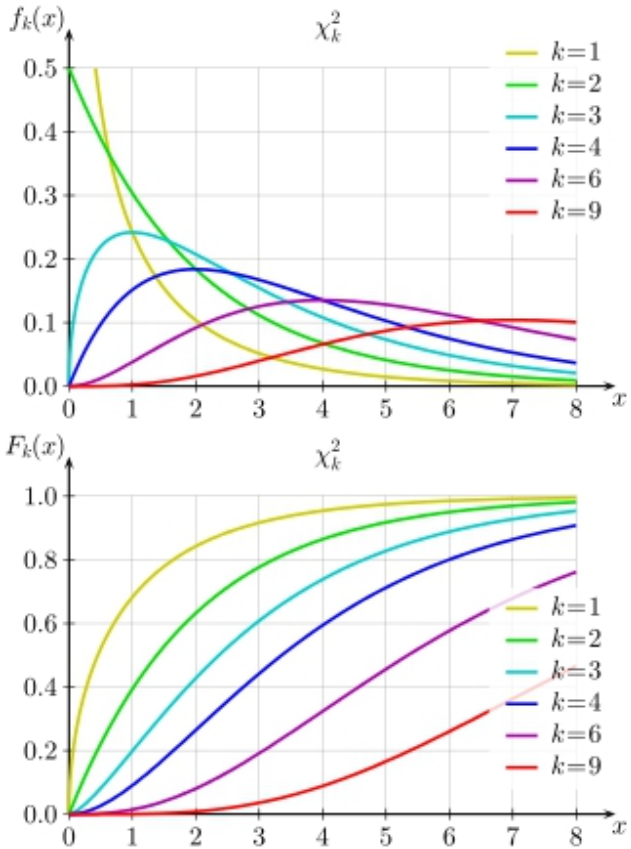
The exponential and chi-squared distributions are special cases of the Gamma distribution.

There is a special relationship between the Gamma and Poisson distributions. If X is a Gamma(α, β) random variable where α is integer, then

$$\mathbb{P}(X \leq x) = \mathbb{P}(Y \geq \alpha),$$

where Y is a Poisson$(x\beta)$ random variable.

The Chi-Squared Distribution



Parameters:	Degrees of freedom $k \in \mathbb{N}$
Support:	$x \geq 0$
pdf:	$\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
cdf:	$\frac{1}{\Gamma\left(\frac{k}{2}\right)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean:	k
Median:	$\approx k\left(1-\frac{2}{9k}\right)^3$
Mode:	$\max(k-2, 0)$
Variance:	$2k$
Skewness:	$\sqrt{8/k}$
Ex. Kurtosis:	$12/k$

(`http://en.wikipedia.org/wiki/Chi-squared_distribution`)

The Chi-Squared Distribution

If $X_1, \ldots, X_n$ are independent, standard Normal random variables ($N(0,1)$), then the sum of their squares,

$$X^2_{(n)} \equiv \sum_{i=1}^{n} X_i^2,$$

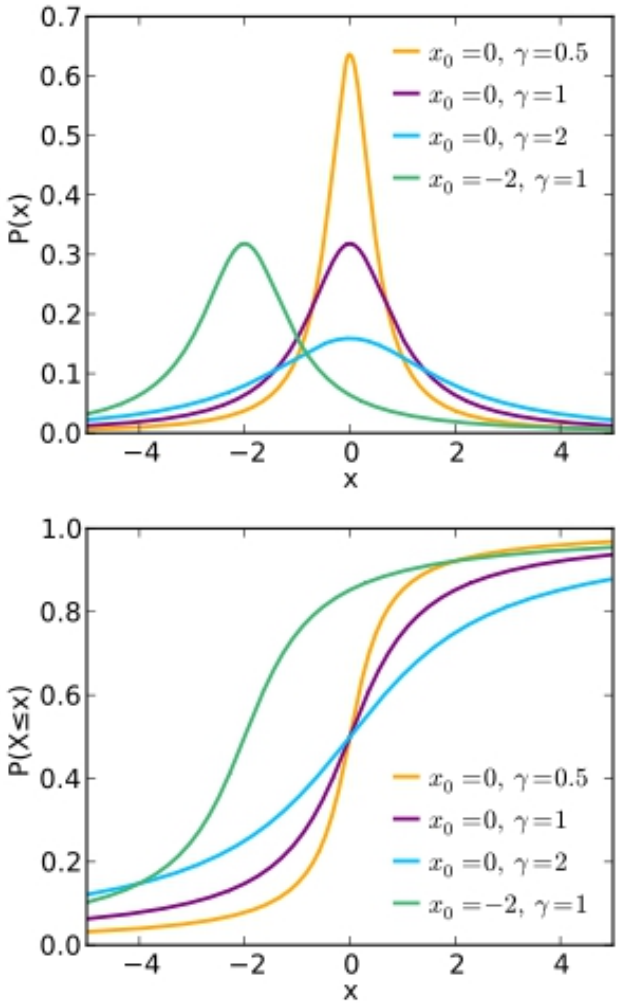
follows a chi-squared distribution for n degrees of freedom.

At large n, the quantities

$$Z_n = \frac{X^2_{(n)} - n}{\sqrt{2n}},$$
$$Z'_n = \sqrt{2X^2_{(n)}} - \sqrt{2n-1},$$

are approximately standard Normal.

The Cauchy (or Lorentz, or Breit-Wigner) Distribution



Parameters:	Location x_0, Scale $\gamma > 0$
Support:	$x \in \mathbb{R}$
pdf:	$\frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$
cdf:	$\frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right)+\frac{1}{2}$
Mean:	undefined
Median:	x_0
Mode:	x_0
Variance:	undefined
Skewness:	undefined
Ex. Kurtosis:	undefined

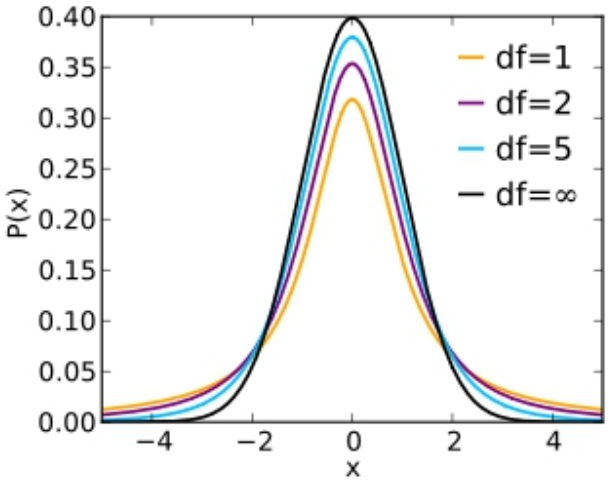
(`http://en.wikipedia.org/wiki/Cauchy_distribution`)

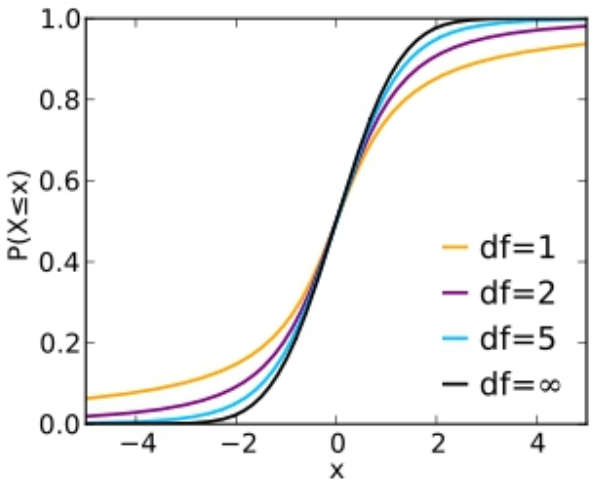
The Cauchy (or Lorentz, or Breit-Wigner) Distribution

Although the Cauchy distribution is pretty pathological, since none of its moments exist, it “has a way of turning up when you least expect it” (Casella & Berger).

For example, the ratio of two standard normal random variables has a Cauchy distribution.

Student's t-Distribution





Parameters:	Degrees of freedom $\nu > 0$
Support:	$x \in \mathbb{R}$
pdf:	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
cdf:	$1 - \frac{1}{2} I_{\frac{\nu}{x^2+\nu}}\left(\frac{\nu}{2}, \frac{1}{2}\right)$, for $x > 0$
Mean:	0 for $\nu > 1$
Median:	0
Mode:	0
Variance:	$\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \leq 2$
Skewness:	0 for $\nu > 3$
Ex. Kurtosis:	$\frac{6}{\nu-4}$ for $\nu > 4$, ∞ for $2 < \nu \leq 4$

(`http://en.wikipedia.org/wiki/Student's_t-distribution`)

Student's t-Distribution

Let $X_1, X_2, \ldots, X_n$ be Normal random variables with mean μ and standard deviation σ. Then the quantities

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

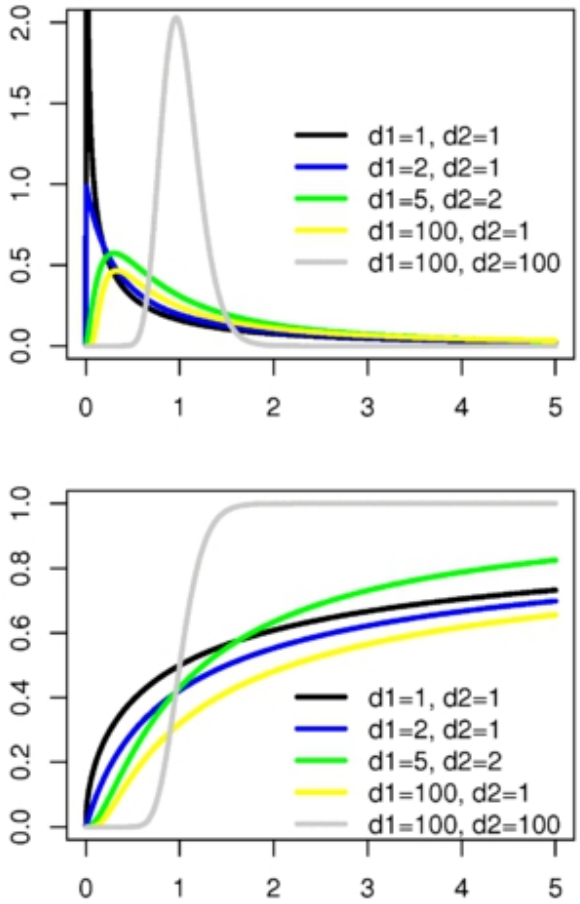
are independently distributed, $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ as standard normal, and $(n-1)S^2/\sigma^2$ as chi-squared for $(n-1)$ degrees of freedom.
To test whether $\bar{X}$ is statistically consistent with μ, when σ is unknown, the correct test statistic to use is the ratio

$$t = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{S/\sigma} = \frac{\sqrt{n}\,(\bar{X} - \mu)}{S},$$

which has a Student's t-distribution with $n - 1$ degrees of freedom.

Note that for $\nu = 1$ the t-distribution becomes the Cauchy distribution.

The Fisher-Snedecor (or F-) Distribution



Parameters:	Degrees of freedom $d_1, d_2 > 0$
Support:	$x \geq 0$
pdf:	$\sqrt{\frac{(d_1 x)^{d_1}\ d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}\ \frac{1}{x\ B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$
cdf:	$I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$
Mean:	$\frac{d_2}{d_2 - 2}$ for $d_2 > 2$
Mode:	$\frac{d_1 - 2}{d_1}\ \frac{d_2}{d_2 + 2}$ for $d_1 > 2$
Variance:	$\frac{2 d_2^2 (d_1 + d_2 - 2)}{d_1 (d_2 - 2)^2 (d_2 - 4)}$ for $d_2 > 4$
Skewness:	$\frac{(2 d_1 + d_2 - 2)\sqrt{8 (d_2 - 4)}}{(d_2 - 6)\sqrt{d_1 (d_1 + d_2 - 2)}}$ for $d_2 > 6$

(`http://en.wikipedia.org/wiki/F-distribution`)

The Fisher-Snedecor (or F-) Distribution

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu_X, \sigma_X^2)$ distribution, and $Y_1, \ldots, Y_m$ a random sample from a $N(\mu_Y, \sigma_Y^2)$ distribution. We may be interested in comparing the variabilities of the X and Y populations. A quantity of interest in this case is the ratio σ_X^2/σ_Y^2, information about which is contained in the ratio of sample variances S_X^2/S_Y^2, where

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

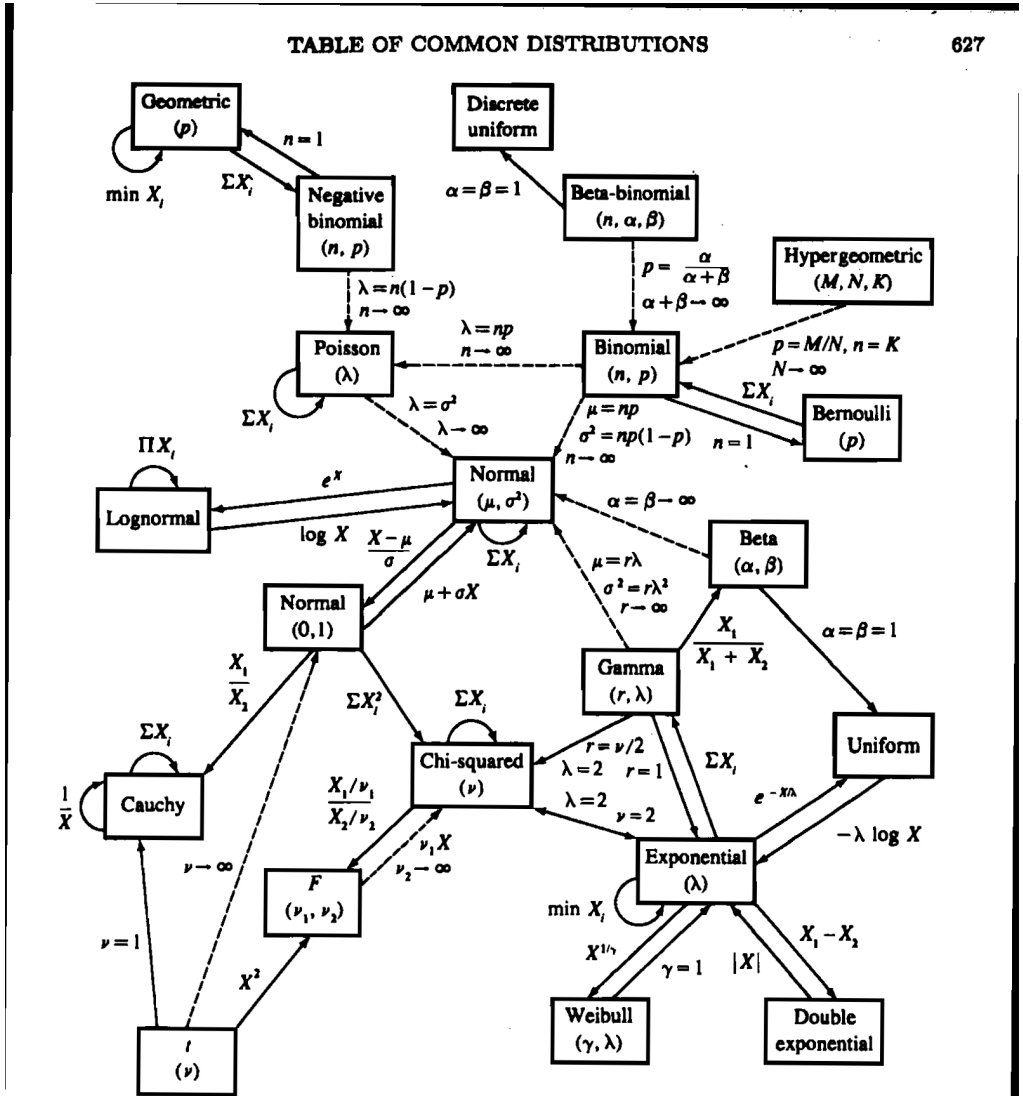
$$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

The F-distribution with $n-1$ and $m-1$ degrees of freedom provides the distribution of the ratio

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2},$$

which is a ratio of scaled chi-squared variates that are independent.

Relationships Among Distributions



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

(From Casella & Berger)

Conditional Probability

It is sometimes desirable to revise probabilities to account for the knowledge that an event has occurred. This can be done using the concept of conditional probability:

Let A and B be events. Provided that $\mathbb{P}(A) > 0$, the conditional probability of B given A is

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

If $\mathbb{P}(A) = 0$, one sets $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ by convention.

What happens in the conditional probability calculation is that B becomes the sample space: $\mathbb{P}(B \mid B) = 1$. The original sample space S has been updated to B.

It follows from the definition that $\mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)$, and by symmetry, $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$. Equating the right-hand sides of both equations yields:

$$\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\,\frac{\mathbb{P}(A)}{\mathbb{P}(B)},$$

which is known as Bayes' rule.

Conditional Probability

A more general version of Bayes' rule is obtained by considering a partition $A_1, A_2, \ldots$ of the sample space S, that is, the A_i are pairwise disjoint subsets of S, and $\cup_{i=1}^{\infty} A_i = S$.

Let B be any set. By the Law of Total Probability we have

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(B \mid A_i)\, \mathbb{P}(A_i).$$

Substituting this result in the basic version of Bayes' rule shows that, for each $i = 1, 2, \ldots$:

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\, \mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B \mid A_j)\, \mathbb{P}(A_j)}.$$

Conditional Probability for Random Variables

So far we defined conditional probability for subsets of sample space ("events"). The extension to random vectors is straightforward.

The simplest case is that of discrete random vectors. Let (X, Y) be a discrete bivariate random vector with joint probability mass function (pmf) $f_{X,Y}(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any y such that $f_Y(y) > 0$, the conditional pmf of X given that $Y = y$ is the function of x defined by

$$f_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

This is a properly normalized pmf that is everywhere positive.

For an absolutely continuous random (n+1)-vector $(X, Y_1, \ldots, Y_n)$ with density f, the conditional density of X given $Y_1, \ldots, Y_n$ is the function of x defined by

$$f_{X|Y_1,\ldots,Y_n}(x \mid y_1, \ldots, y_n) = \frac{f(x, y_1, \ldots, y_n)}{\int_{-\infty}^{+\infty} f(z, y_1, \ldots, y_n)\, dz}.$$

The conditional cumulative distribution function of X given $Y_1, \ldots, Y_n$ is obtained by integrating the conditional density:

$$F_{X|Y_1,\ldots,Y_n}(t \mid y_1, \ldots, y_n) = \int_{-\infty}^{t} f_{X|Y_1,\ldots,Y_n}(x \mid y_1, \ldots, y_n)\, dx.$$

Bayes' Theorem and the Predictive Distribution

The Bayesian paradigm makes extensive use of conditional probabilities, starting with Bayes' theorem, which is a reformulation of Bayes' rule for random variables. Suppose we have some data X with a known distribution $p(X \mid \theta)$ depending on some unknown parameter θ. If we have prior information about θ in the form of a prior distribution $\pi(\theta)$, Bayes' theorem tells us how to update that information to take into account the new data. The result is a posterior distribution for θ:

$$p(\theta \mid X) = \frac{p(X \mid \theta)\,\pi(\theta)}{\int p(X \mid \theta')\,\pi(\theta')\,d\theta'}.$$

The result of a Bayesian analysis is this posterior distribution. Usually one tries to summarize it by providing a mean, or quantiles, or intervals with specific probability content.

The posterior distribution also provides the basis for predicting the values of future observations X_{new}, via the predictive distribution:

$$p(X_{new} \mid X) = \int p(X_{new} \mid \theta)\,p(\theta \mid X)\,d\theta.$$

The Borel-Kolmogorov Paradox

This paradox illustrates some of the pitfalls attached to conditional probability calculations. Imagine a uniform distribution of points over the surface of the earth. Intuitively it is obvious that the subset of points along the equator will also be uniformly distributed. That is, their conditional distribution, given a latitude of zero, is uniform. However the equator is only the equator because of our choice of coordinate system. Any other great circle could serve as equator. By invariance one would expect the points to be just as uniformly distributed along meridians, for example.

Unfortunately this cannot be the case. A quarter of a meridian's length lies north of latitude 45°N. Integrating over all meridians would lead to the expectation that the earth's cap above 45°N covers a quarter of the total surface, which is clearly incorrect.

To resolve the paradox, one must realize that conditioning on a set of probability zero is inadmissible. The proper procedure is first to condition on a small but finite set around the value of interest, and then take a limit. However the limit is not uniquely defined. In the case of the equator, it is a band of parallels whose width goes to zero, whereas in the case of the meridian it is a lune of meridians whose opening angle goes to zero...

The Borel-Kolmogorov Paradox

For a mathematical explanation, we introduce a latitude ϕ and a longitude λ, with $-\pi/2 \leq \phi \leq +\pi/2$ and $-\pi < \lambda \leq +\pi$. The probability density function of a uniform distribution on the unit sphere is given by:

$$f(\phi, \lambda) = \frac{1}{4\pi} \cos\phi.$$

The marginal densities are:

$$f(\phi) = \int_{-\pi}^{+\pi} f(\phi, \lambda)\, d\lambda = \frac{1}{2} \cos\phi,$$

$$f(\lambda) = \int_{-\pi/2}^{+\pi/2} f(\phi, \lambda)\, d\phi = \frac{1}{2\pi}.$$

For the conditional densities at zero longitude and zero latitude this gives:

$$f(\phi \mid \lambda = 0) = \frac{f(\phi, \lambda = 0)}{f(\lambda = 0)} = \frac{1}{2} \cos\phi,$$

$$f(\lambda \mid \phi = 0) = \frac{f(\phi = 0, \lambda)}{f(\phi = 0)} = \frac{1}{2\pi},$$

showing that the conditional distributions along the Greenwich meridian and along the equator are indeed different.

The Borel-Kolmogorov Paradox

Let's now change coordinates: Keep the latitude ϕ but replace the longitude:

$$\lambda \to \mu = \frac{\lambda}{g(\phi)},$$

where $g(\phi)$ is a strictly positive function. In terms of ϕ and μ, the density of points on the unit sphere is given by:

$$f(\phi, \mu) = \frac{1}{4\pi} \cos\phi \left| \frac{\partial(\phi, \lambda)}{\partial(\phi, \mu)} \right| = \frac{1}{4\pi} \cos\phi \, g(\phi)$$

For the conditional density of latitudes given $\mu = 0$ we find:

$$f(\phi \mid \mu = 0) \propto \cos\phi \, g(\phi).$$

Observe that $\mu = 0$ is entirely equivalent to $\lambda = 0$; both conditions correspond to the same set of points on the unit sphere. And yet the conditional distributions differ by the factor $g(\phi)$. This is because of the implicit limiting procedure used in conditioning. To condition on $\lambda = 0$ we consider the set $\{\lambda : |\lambda| \leq \epsilon\}$ and let ϵ go to zero. To condition on $\mu = 0$ the set is $\{\lambda : |\lambda| \leq \epsilon \, g(\phi)\}$.

Classical Limit Theorems

The limit theorems describe the behavior of certain sample quantities as the sample size approaches infinity.

The notion of an infinite sample size is somewhat fanciful, but it provides an often useful approximation for the finite sample case due to the simplification it introduces in various formulae.

First we have to clarify what we mean by random variables “approaching infinity”. . .

Modes of Stochastic Convergence

Consider a sequence of random variables $X_1, X_2, \ldots$, *not* necessarily independent or identically distributed. There are many different ways this sequence can converge to a random variable X. Here we only consider three:

1. Convergence in distribution:
   $X_n \xrightarrow{\text{d}} X$ if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all points x where $F_X(x)$ is continuous.

2. Convergence in probability (or weak convergence):
   $X_n \xrightarrow{\text{P}} X$ if, for every $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1$.
   "For any $\epsilon > 0$ and $\delta > 0$ there is an integer N such that the individual deviation probabilities $\mathbb{P}(|X_n - X| > \epsilon)$ are less than δ for all $n > N$."

3. Almost sure convergence (or strong convergence):
   $X_n \xrightarrow{\text{a.s.}} X$ if, for every $\epsilon > 0$, $\mathbb{P}(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1$.
   "For any $\epsilon > 0$ and $\delta > 0$ there is an integer N such that the probability for any deviation $|X_n - X|$, $n > N$, to be greater than ϵ, is less than δ."

Almost sure convergence implies convergence in probability, and convergence in probability implies convergence in distribution.

Classical Limit Theorems: The Statements

Let $X_1, X_2, \ldots$ be i.i.d.; set $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}(X_1)$, and $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1)$.

The weak law of large numbers:

$$\text{If } \mu < \infty, \text{ then } \bar{X}_n \xrightarrow{\mathsf{P}} \mu.$$

The strong law of large numbers:

$$\text{If } \mu < \infty, \text{ then } \bar{X}_n \xrightarrow{\text{a.s.}} \mu.$$

The central limit theorem:

$$\text{If } \sigma^2 < \infty, \text{ then } \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathsf{d}} N(0,1).$$

The law of the iterated logarithm:

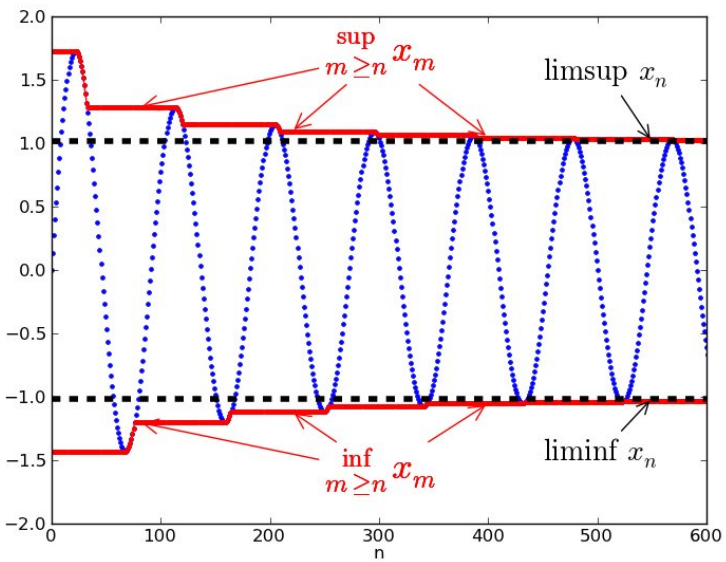
$$\text{If } \sigma^2 < \infty, \text{ then } \limsup_{n\to\infty} \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \frac{1}{\sqrt{2\ln\ln n}} = 1.$$

Short Digression on "Lim Sup"

The limit superior and limit inferior of a sequence of X_n are defined by

$$\limsup_{n\to\infty} X_n \equiv \lim_{n\to\infty}\left(\sup_{m\geq n} X_m\right) = \inf_{n\geq 0}\sup_{m\geq n} X_m$$

$$\liminf_{n\to\infty} X_n \equiv \lim_{n\to\infty}\left(\inf_{m\geq n} X_m\right) = \sup_{n\geq 0}\inf_{m\geq n} X_m$$



The limit superior of X_i is the smallest real number b such that, for any $\epsilon > 0$, there exists an n such that $X_m < b + \epsilon$ for all $m > n$. In other words, any number larger than the limit superior is an eventual upper bound for the sequence. Only a finite number of elements of the sequence are greater than $b + \epsilon$.

A sequence converges when its limit inferior and limit superior are equal.

Classical Limit Theorems: Applications

1. **Consistency**
   The property described by the Weak Law of Large Numbers, a sequence of random variables converging to a constant, is known as *consistency*. This is an important property for the study of *point estimators* in statistics.

2. **Frequentist Statistics**
   As an application of the Weak Law of Large Numbers, let X_i be a Bernoulli random variable, with $X_i = 1$ representing "success", and $X_i = 0$ "failure". Then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the frequency of successes in the first n trials. If p is the probability of success, the Weak Law of Large Numbers states that $\bar{X}_n \xrightarrow{\mathrm{P}} p$.

   This result is often used by frequentist statisticians to justify their identification of probability with frequency. One must be very careful with this however. Logically one cannot assume something as a definition and then prove it as a theorem. Furthermore, there is a contradiction between a definition that would assume as certain something that the theorem only states to be very probable.

3. Monte Carlo Integration:
   Let f be a function on $[a, b]$, with $\int_a^b |f(x)|\,dx < \infty$, and let $U_1, U_2, \ldots$ be independent random variables that are uniformly distributed between a and b. Then:
   $$\frac{1}{n}\sum_{i=1}^{n} f(U_i) \xrightarrow{\text{a.s.}} \int_a^b f(x)\,dx.$$
   This follows directly from the Strong Law of Large Numbers. This Monte Carlo technique of integration is one of the simplest, and it extends to multi-dimensional integrals.

Classical Limit Theorems: Applications

4. **Maximum Likelihood Estimation:**
Suppose that we have observed data $X_1, \ldots, X_n$ from a distribution $f(x \mid \theta)$, where θ is an unknown parameter. The likelihood function is:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta).$$

We can estimate θ by its value at the maximum of the likelihood:

$$\hat{\theta}_n = \arg\max_{\{\theta\}} \mathcal{L}_n(\theta).$$

The estimator $\hat{\theta}_n$ depends on the sample size n. If the likelihood function satisfies some standard regularity conditions, one can prove two limit theorems for $\hat{\theta}_n$, as $n \to \infty$:

- Consistency: $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta$;
- Asymptotic normality: $\sqrt{n}[\hat{\theta}_n - \theta] \xrightarrow{\mathsf{d}} N(0, \sigma^2(\theta))$, where $\sigma^2(\theta) = 1/I(\theta)$ and $I(\theta)$ is the Fisher information:

$$I(\theta) \equiv \mathbb{E}\left[\left(\frac{d}{d\theta} \ln f(X_1 \mid \theta)\right)^2\right]$$

Classical Limit Theorems: Applications

5. **Empirical Distribution Functions:**
Let $X_1, \ldots, X_n$ be i.i.d. with entirely unknown distribution function F. We can estimate F by the *empirical distribution function*:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq t).$$

This estimator is justified by the Strong Law of Large Numbers:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq t) \xrightarrow{\text{a.s.}} \mathbb{E}[\mathbf{1}(X_1 \leq t)] = F(t).$$

The convergence is actually uniform in t, a result known as the Glivenko-Cantelli theorem:

$$\sup_{\{t\}} |\hat{F}_n(t) - F(t)| \xrightarrow{\text{a.s.}} 0.$$

Accompanying this result is a central limit theorem known as the Kolmogorov-Smirnov theorem:

$$\sqrt{n} \sup_{\{t\}} |\hat{F}_n(t) - F(t)| \xrightarrow{\text{d}} Z, \text{ where } \mathbb{P}(Z \leq t) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2 t^2}.$$

Classical Limit Theorems: Applications

6. **The Delta Method:**
   Suppose we make an observation X that we can use to estimate a parameter θ. However we are not interested in θ itself, but in some function g of θ. We could then consider using $g(X)$ as an estimator of $g(\theta)$. What are the properties of this estimator? What is its variance, its sampling distribution?

   A first-order Taylor expansion of $g(X)$ around $X = \theta$ yields:

$$g(X) = g(\theta) + g'(\theta)(X - \theta) + \text{Remainder}$$

   Assuming that the mean of X is θ, this leads to $\mathbb{E}(g(X)) \approx g(\theta)$ (neglecting the remainder in the Taylor expansion), and therefore:

$$\mathbb{V}\text{ar}g(X) \approx \mathbb{E}([g(X) - g(\theta)]^2) \approx \mathbb{E}([g'(\theta)(X - \theta)]^2) = [g'(\theta)]^2 \mathbb{V}\text{ar}X$$

   So far all we have done is rederive the rule of error propagation. Where it gets interesting is that this rule is associated with a generalization of the Central Limit Theorem.

Classical Limit Theorems: Applications

6. The Delta Method (continued):
   - Basic Delta Method:
     Let $Y_1, Y_2, \ldots$ be random with $\sqrt{n}(Y_n - \theta) \xrightarrow{\text{d}} N(0, \sigma^2)$. For given g and θ, suppose that $g'(\theta)$ exists and is not zero. Then
     $$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{\text{d}} N(0, \sigma^2[g'(\theta)]^2).$$
   - Second-Order Delta Method:
     Let $Y_1, Y_2, \ldots$ be random with $\sqrt{n}(Y_n - \theta) \xrightarrow{\text{d}} N(0, \sigma^2)$. For given g and θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not zero. Then
     $$n[g(Y_n) - g(\theta)] \xrightarrow{\text{d}} \frac{1}{2}\sigma^2 g''(\theta)\chi^2_1.$$
   - Multivariate Delta Method:
     Let $\vec{X}_1, \vec{X}_2, \ldots$ be a sequence of random p-vectors such that $\mathbb{E}(X_{ik}) = \mu_i$ and $\mathbb{C}\text{ov}(X_{ik}, X_{jk}) = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\vec{\mu}$, suppose that $\tau^2 \equiv \sum\sum \sigma_{ij} \frac{\partial g}{\partial \mu_i} \frac{\partial g}{\partial \mu_j} > 0$. Then:
     $$\sqrt{n}[g(\bar{X}_1, \ldots, \bar{X}_p) - g(\mu_1, \ldots, \mu_p)] \xrightarrow{\text{d}} N(0, \tau^2).$$

7 Sampling to a Foregone Conclusion

Suppose we are looking for a new physics signal in a specific channel. We have a sample of n events in that channel, and for each event in the sample we measure property X. We plan to claim discovery if the observed significance Z_n exceeds a pre-set threshold c:

$$Z_n \equiv \frac{|\bar{X}_n - \mu|}{\sigma/\sqrt{n}} \geq c,$$

where μ is the expected value of X in the absence of signal and σ is the measurement resolution. The discovery threshold c is typically set to 3 or 5 in HEP.

Suppose that after an initial data collection run we observe some indication of a signal, but not enough to claim discovery. We then proceed to take more data and regularly check our discovery criterion. As the sample size increases, is there any guarantee that the probability for making the correct decision regarding the presence of signal goes to 1?

7. Sampling to a Foregone Conclusion (continued)

The answer is NO! At least not if we keep c constant with n. This is a consequence of the Law of the Iterated Logarithm (LIL), according to which the event $Z_n \geq (1+\epsilon)\sqrt{2 \ln \ln n}$ happens infinitely many times if $\epsilon \leq 0$. Therefore, regardless of the choice of c, Z_n will eventually exceed c for some n, even if there is no signal.

Furthermore, the LIL tells us exactly how to vary the discovery threshold with sample size in order to avoid "sampling to a foregone conclusion".

"A blind use of [significances] allows the statistician to cheat, by claiming at a suitable point in a sequential experiment that he has a train to catch. This must have been known to Khintchine when he proved in 1924 that, in sequential binomial sampling, a "sigmage" of nearly $\sqrt{2 \ln \ln n}$ is reached infinitely often, with probability 1. [...] But note that the iterated logarithm increases with fabulous slowness, so that this particular objection to the use of [significances] is theoretical rather than practical. To be reasonably sure of getting 3σ one would need to go sampling for billions of years, by which time there might not be any trains to catch." (I.J. Good, J. R. Statist. Soc. B 27 (1965) p.197.)