Discovery or fluke: statistics in particle physics

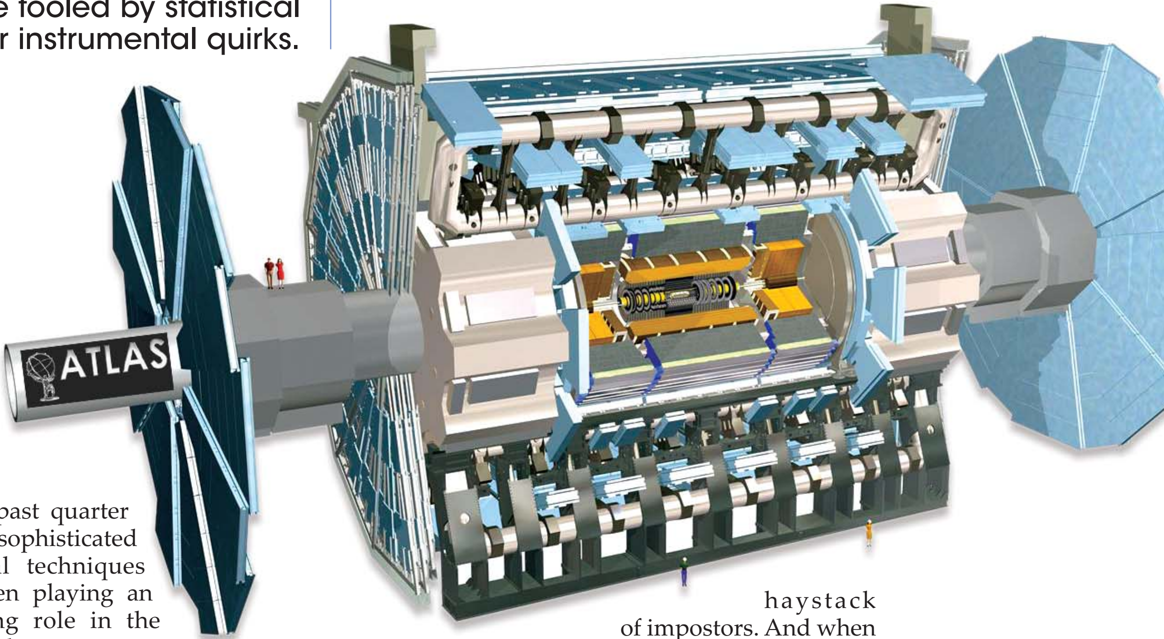
Louis Lyons

Citation: Phys. Today **65**(7), 45 (2012); doi: 10.1063/PT.3.1642
View online: http://dx.doi.org/10.1063/PT.3.1642
View Table of Contents: http://www.physicstoday.org/resource/1/PHTOAD/v65/i7
Published by the American Institute of Physics.

Discovery or fluke: STATISTICS IN PARTICLE PHYSICS

When you're searching for elusive manifestations of new physics, it's easy to be fooled by statistical fluctuations or instrumental quirks.

Louis Lyons



Over the past quarter century, sophisticated statistical techniques have been playing an increasing role in the analysis of particle-physics experiments. For one thing, the experiments have become much more elaborate and difficult, typically producing enormous volumes of data in search of very small effects.

In the decades after World War II, discoveries of new particles—for example, many of the early strange-quark-bearing mesons and hyperons—were often based on one or a few bubble chamber photographs. In the oft-told case of the J/ψ meson, the first known particle harboring charmed quarks, the cross section for its formation in an electron–positron collider rose so dramatically at the resonant energy that the discovery was clear within hours of the first hint of a signal.

Such discoveries were deemed obvious; there was no need to calculate the probability that statistical fluctuations had produced a spurious effect. Contrast that with today's search for the Higgs boson, the only remaining undiscovered fundamental particle required by particle theory's standard model (see PHYSICS TODAY, February 2012, page 16). Because the Higgs search involves a signal-to-background ratio of order 10^{-10}, sophisticated multivariate techniques such as artificial neural networks are needed for finding needle candidates in the haystack of impostors. And when a possible signal does appear, assessing its statistical significance nowadays requires great care.

Gargantuan instruments

The standard model, which took shape in the late 1970s, does not predict a specific mass M_H for the Higgs, but it does predict, as functions of M_H, all of its couplings to other particles and therefore its production and decay rates. So non-observations of anticipated decay modes are used to disfavor the existence of the standard-model Higgs in various M_H ranges.

The Higgs searches, as well as searches for obscure manifestations of new physics beyond the spectacularly successful but manifestly incomplete standard model, are usually performed at large particle accelerators such as the Large Hadron Collider at CERN, shown in figure 1. In the LHC, collisions between beams of multi-TeV protons produce enormous debris showers in which experimenters seek evidence of new particle species and new properties of particles already known.

Louis Lyons, a particle physicist retired from the University of Oxford, is now based at the Blackett Laboratory, Imperial College, London.

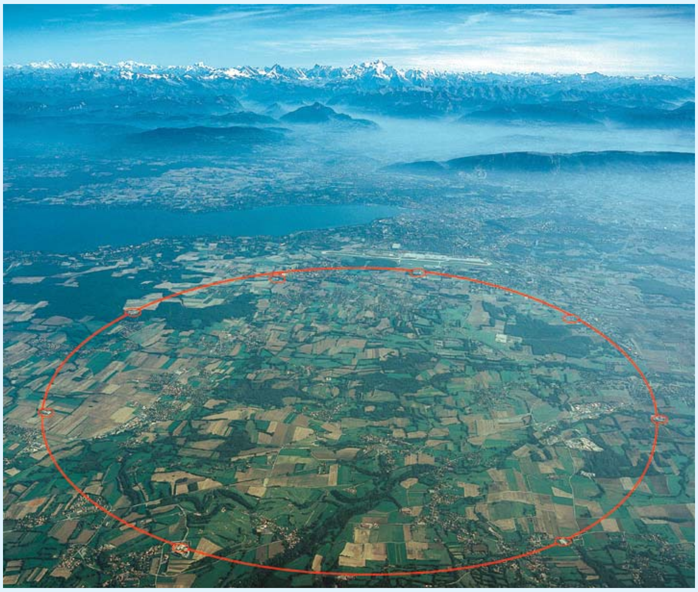


Figure 1. The CERN laboratory straddles the Swiss–French border near Lake Geneva. The 27-km-circumference red circle traces the underground tunnel of the lab's Large Hadron Collider. Mont Blanc, the highest peak in the Alps, appears on the horizon.

Ironically, the very success of the standard model means that statistically robust manifestations of really new physics beyond its purview have thus far eluded experimenters—though there have been tantalizing false alarms. Many experiments now employ "blind analyses," lest some desired outcome subconsciously bias the choice of criteria by which data are accepted or discarded.

The detectors that surround the beam-crossing points and record collision products are gargantuan and intricately multifaceted. The ATLAS detector at the LHC, for example, is as big as a seven-story building, and it incorporates 10^8 channels of electronics. Nowadays, large experimental collaborations with thousands of members from scores of institutions establish committees of experts to advise on statistical issues. Their websites provide useful tutorial articles on such matters, as do the proceedings of the PHYSTAT series of conferences and workshops.[1]

Statistical precision improves with running time only like its square root, and running time at the big accelerators and detectors is costly. Therefore, it's particularly worthwhile to invest effort in developing statistical techniques that optimize the accuracy with which a physics parameter is determined from data taken during a given running time.

More and more searches for hypothesized new phenomena don't actually find them but rather set upper limits on their strength. It has become clear that different methods of setting such upper limits can yield different valid answers. In that regard, the difference between the so-called Bayesian and frequentist approaches is now better appreciated. Another statistical issue nowadays is combining measurements of the same quantity in different experiments, where possible correlations between the different measurements need to be taken into account. And some analyses seek to determine the parameters of a putative theory such as supersymmetry from many different measurements.

Although the issues and techniques discussed in this article have been developed largely for analyses of particle-physics data, many are also applicable elsewhere. For example, it's possible to look for low-grade biological attacks by terrorists in data on the number of people checking in each day at hospitals across the country. Finding a statistically significant enhancement of patients in a specific spacetime region requires statistical tools similar to those with which one seeks the Higgs as an enhancement in the number of interactions in the space of M_H and other relevant variables.

Bayesians versus frequentists

There are two fundamental but very different approaches to statistical analysis: the Bayesian and frequentist techniques. The former is named after the 18th-century English theologian and mathematician Thomas Bayes (see the book review on page 54 of this issue). The two approaches differ in their interpretation of "probability." For a frequentist, probability is defined in terms of the outcomes for a large number of essentially identical trials. Thus the probability of some particular number coming up in a throw of dice could be estimated from the fraction of times it actually happens in many throws.

The need for many trials, however, means that frequentist probability doesn't exist for one-off situations. For example, one can't speak of the frequentist probability that the first team of astronauts to Mars will return safely to Earth. Similarly, a statement like "dark matter, at present, constitutes less than 25% of the cosmic mass–energy budget" cannot be assigned a frequentist probability. It's either true or false.

Bayesians claim that the frequentist view of probability is too narrow. Probability, they say, should be thought of as a measure of presupposition, which can vary from person to person. Thus the probability a Bayesian would assign to whether it rained yesterday in Eilat would depend on whether he knew Eilat's location and seasonal rainfall pattern, and whether he had communicated with anyone there yesterday or today. A frequentist, on the other hand, would refuse to assign any probability to the correctness of such a statement about a past event.

Despite that very personal definition, it is possible to determine numerical values for Bayesian probabilities. To do that, one invokes the concept of a fair bet. A Bayesian asserting that the chance of something being the case is 20% is, in essence, prepared to offer 4 to 1 odds either way.

Conditional probability $P(A|B)$ is the probability of A, given the fact that B has happened or is the case. For example, the probability of obtaining a 4 on a throw of a die is 1/6; but if we accept only even results, the conditional probability for a 4 becomes 1/3. One shouldn't wrongly equate $P(A|B)$ with $P(B|A)$. The probability of being pregnant, assuming you are female, is much smaller than the probability of being female, assuming you are pregnant. Similarly, if the probability of obtaining the observed data set under the assumption that the Higgs boson does not exist in a certain mass range is only 1%, it is incorrect to deduce from those data that the probability that there is no Higgs in that mass range is only 1%.

Given two correlated circumstances A and B, the probability that both happen or are true is given in terms of the conditional probabilities by

$$P(A \text{ and } B) = P(A|B) \times P(B) = P(B|A) \times P(A),$$

which can be rewritten as

$$P(A|B) = P(B|A) \times P(A)/P(B). \quad (1)$$

Equation 1 is known as Bayes's theorem. (See the Quick Study by Glen Cowan in PHYSICS TODAY, April 2007, page 82.) For example, if you consider a random day in the past few years, Bayes's theorem would relate the probability that the day was rainy, given the fact that it was, say, December 25th, to the probability that the day was December 25th, given the fact that it was rainy. They're not the same.

Estimating parameters

Bayesians and frequentists differ in how they estimate parameter values from data, and in what they mean by their characterization of an estimate. For concreteness, consider a simple counting experiment to determine the flux μ of cosmic rays passing through a detector. In one hour it records n events.

Though most experimenters would then estimate μ as n per hour, they could differ on the range of μ values that are acceptable. Usually one aims to find a μ range that corresponds to a specified "confidence level"—for example 68%, the fraction of the area under a Gaussian curve within one standard deviation of its peak. The individual cosmic rays are believed to appear at random and independently of each other and therefore follow a Poisson distribution. That is, the probability of finding n events if, on average, you expect μ is

$$P_\mu(n) = e^{-\mu} \times \mu^n/n!\ . \quad (2)$$

Frequentists don't dispute Bayes's theorem. But they insist that its constituent probabilities be genuine frequentist probabilities. So Bayesians and frequentists differ fundamentally when probabilities are assigned to parameter values.

The frequentist method is to define a confidence band in the (μ,n) plane such that for every possible value of μ, the band contains those values of n that are "likely" in the sense that the Poisson probabilities $P_\mu(n)$ add up to at least 68%. Then the observed number n_{obs} is used to find the range of μ values for which n_{obs} is within the 68% probability band. The recipe is embodied in the Neyman construction shown in figure 2, named after mathematician Jerzy Neyman (1894–1981).

One is thus making a statement about the range of values for which n_{obs} was likely. For values of μ outside that range, n_{obs} would have been too small or too large to be typical.

The Bayesian approach is quite different. In equation 1, Bayesians replace outcomes A and B, respectively, by "parameter value" and "observed data." So Bayes's theorem then gives

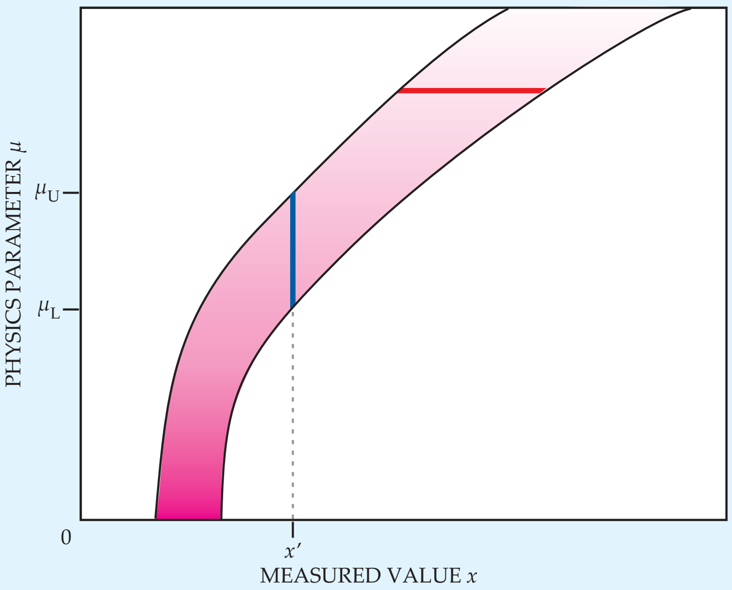


Figure 2. Setting a confidence range for a physics parameter μ by means of a Neyman construction. For any putative value of μ, one presumes to know the probability density $P_\mu(x)$ for obtaining a measured experimental result x. For each μ, the pink-shaded confidence band indicates a range in x that encloses, say, 68% of the probability (the red line). Then, from a particular measurement result x', the 68% confidence interval from μ_L to μ_U (lower to upper, the blue line) gives the range of μ for which the measured x' is deemed probable. An x' smaller than the one shown here might yield only an upper limit or perhaps even no plausible μ value at all.

$$P(param\,|\,data) \propto P(data\,|\,param) \times P(param), \quad (3)$$

where $P(param)$ is called the Bayesian prior. It's a probability density that quantifies what was believed about the parameter's value before the current measurement. The so-called likelihood function $P(param\,|\,data)$ is, in fact, simply the probability of seeing the observed data if a specific parameter value is the true one.

The $P(param\,|\,data)$, called the Bayesian posterior, can be thought of as an update of the Bayesian prior after the new measurement. It can then be used to extract a new median value of the parameter, a 68%-confidence central interval, a 90%-confidence upper limit, or whatever.

A big issue in Bayesian analysis is what functional form to use for the prior, especially in situations where little is known in advance about the parameter being sought. Take, for example, the mass M_1 of the lightest neutrino. The argument that the prior function should simply be a constant so as not to favor any particular range of values is flawed. That's because it's not at all obvious whether one should assume a constant prior probability density in M_1 or, for example, in M_1^2 or ln M_1. For a given experimental result, all those options yield different conclusions.

The Bayesian priors may be straightforward for parameterizing real prior knowledge, but they are

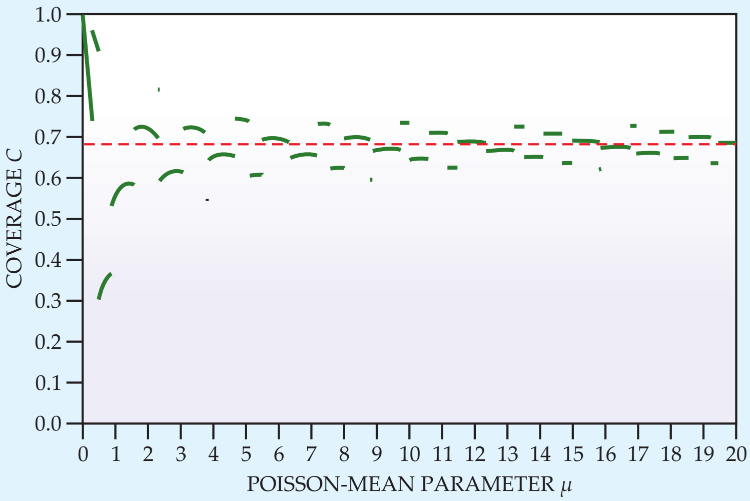


Figure 3. For an idealized counting experiment whose distribution of counts *n* is Poissonian (equation 2 in the text), the coverage *C* is plotted against the Poisson-mean parameter *μ*. Coverage is defined as the fraction of all trials in which the *μ* range, here given by the maximum-likelihood prescription (equation 6), actually does include the true *μ*. The curve jumps around because *n* must be an integer, while *μ* is a continuous parameter. Only when *μ* is large does *C* approach the expected 68% (dashed red line).

problematic for dealing with prior ignorance. Therefore, if you do Bayesian analysis, it's important to perform a test of its sensitivity—the extent to which different choices of priors affect the final result.[2]

Interpreting parameter ranges

Both Bayesian and frequentist analyses typically end with statements of the form

$$\mu_L \le \mu \le \mu_U \qquad (4)$$

at 68% confidence level. And indeed, in some simple problems, the numerical values of the lower and upper limits μ_L and μ_U can agree for the two methods. But they mean different things. The frequentists assert that μ is a physics parameter with a fixed but unknown value about which no direct probability statement can be made. For them, equation 4 is a statement about the interval μ_L to μ_U. Their claim is that if the measurement were to be repeated many times, statistical fluctuations would vary that range from measurement to measurement. But 68% of those quoted ranges should contain the true μ. The actual fraction of ranges that contain the true μ is called the statistical method's coverage C.

By contrast, Bayesians say that μ_L and μ_U have been determined by the measurement without regard to what would happen in hypothetical repetitions. They regard equation 4 as a statement of what fraction of the posterior probability distribution lies within that range.

One shouldn't think of either viewpoint as better than the other. Current LHC analyses employ both. It is true that particle physicists tend to favor frequentist methods more than most other scientists do. But they often employ Bayesian methods for dealing with nuisance parameters associated with systematic uncertainties.

Maximum likelihood

A very useful approach for determining parameters or comparing different theories to measurements is known as the likelihood method. In the Poisson counting example, the probability for observing n counts in a unit of time, when μ is the expected rate, is given by the Poisson distribution $P_\mu(n)$ of equation 2. It's a function of the discrete observable n for a given Poisson mean μ. The likelihood function

$$L_n(\mu) \equiv P_\mu(n) = e^{-\mu} \times \mu^n/n! \qquad (5)$$

simply reverses those roles. It's the same Poisson distribution, but now regarded as a function of the sought physics result μ, with n fixed at the observed measurement. The best estimate of μ is then the value that maximizes $L_n(\mu)$. When there are several independent observations n_i, the overall likelihood function $L(\mu)$ is the product of the individual likelihood functions.

An attractive feature of the likelihood method is that it can use individual data observations without first having to bin them in a histogram. That makes the unbinned likelihood approach a powerful method for analyzing sparse data samples.

The best estimate μ_{max} is the value for which the probability of finding the observed data set would be greatest. Conversely, values of μ for which that probability is very small are excluded. The range of acceptable values for μ is related to the width of the likelihood function. Usually, the range is defined by

$$\ln L(\mu_L) = \ln L(\mu_U) = \ln(L_{\text{max}}) - ½. \qquad (6)$$

When the likelihood function is a Gaussian centered at μ_{max} with standard deviation σ, that prescription yields the conventional range $\mu_{\text{max}} \pm \sigma$.

Two important features of likelihoods are commonly misunderstood:

▶ The likelihood method does not automatically satisfy the coverage requirement that 68% of all intervals derived via equation 6 contain the true value of the parameter. The complexity of the situation is illustrated in figure 3, which shows the actual coverage C as a function of μ for the Poisson counting experiment.[3] For small μ, the deviation of C from a constant 68% is dramatic.

▶ For parameter determination, one maximizes L as a function of the parameter being sought, using the fixed data set. So it's often thought that the larger L_{max} is, the better is the agreement between theory and data. That's not so; L_{max} is not a measure of goodness-of-fit. A simple example makes the point: Suppose one seeks to determine the parameter β for a particle's presumed decay angular distribution

$$(1 + \beta \cos^2 \theta)/(2 + 2\beta/3)$$

from a set of decay observations $\cos\theta_i$, which can range from −1 to +1. The presumed distributions are all symmetric about $\cos\theta = 0$, so a data set with all $\cos\theta_i$ negative should give a very bad fit to the theory that yielded the symmetric parameterization. But the likelihood function contains only $\cos^2\theta$. So it's completely insensitive to the signs of the $\cos\theta$ values.

For two precisely defined alternative hypotheses, however, the ratio of likelihood values can be used to maximize the power of an experimental test between those hypotheses.[4]

Discovery, exclusion, and limits

Consider two idealized scenarios in which we're looking for new particle-physics phenomena: In one case, we count the number of observations n. They may be produced just by background sources (with expected Poisson mean b), or by a combination of background sources and new physics (expectation $b + s$). When we find n larger than b, how significant is the evidence that s is greater than zero?

In the second case, we're looking at the so-called invariant-mass distributions of pairs of particles produced in high-energy collisions. The invariant mass m of such a pair is the sum of their energies in their joint center of mass. If enough of those pairs are, in fact, the decay products of some hypothesized new particle, the m distribution will exhibit a narrow peak around the putative parent's mass, on top of a smooth and predictable distribution due to background processes. Figure 4 offers an illustrative but cautionary tale (see PHYSICS TODAY, September 2003, page 19).

In both cases, we're seeking to decide whether the data are more consistent with the null hypothesis H_0 (background only) or with the alternative hypothesis H_1 (background plus signal).[5] We might conclude that the data are convincingly inconsistent with H_0, and so claim a discovery. Alternatively, we might set an upper limit on the possible signal strength or even conclude that the data exclude H_1. Finally, it may be that our experiment is not sensitive enough to distinguish between H_0 and H_1.

The technique consists in choosing a test variable t whose observed distribution should be sensitive to the difference between the two hypotheses. In the first example above, t could simply be n, the number of counts observed in a fixed time interval. In the second case, it could be the likelihood ratio of the two hypotheses.

To illustrate, figure 5 shows the expected distributions of t, assuming either H_0 or H_1, for three different imagined experiments. In one, shown in figure 5a, the expected distributions for the two hypotheses overlap so much that it's hard to distinguish between them. At the other extreme, figure 5b, the larger separation makes the choice easy.

The intermediate case, figure 5c, needs discussion. For a given observed value t', we define probability values p_0 and p_1 for the two hypotheses. As shown in the figure, each is the fractional area under the relevant curve for finding a value of t at least that extreme. Usually, one claims a discovery when p_0 is below some predefined level α. Alternatively, one excludes H_1 if p_1 is smaller than another predefined level γ. And if the observed t' satisfies neither of those conditions, the experiment has yielded no decision.

In particle physics, the usual choice for α is 3×10^{-7}, corresponding to the 5σ tail of a Gaussian H_0 distribution. That requirement is shown by t_{crit} in figure 5c. Why so stringent? For one thing, recent history offers many cautionary examples of exciting 3σ and 4σ signals that went away when more data arrived (figure 4c, for example).

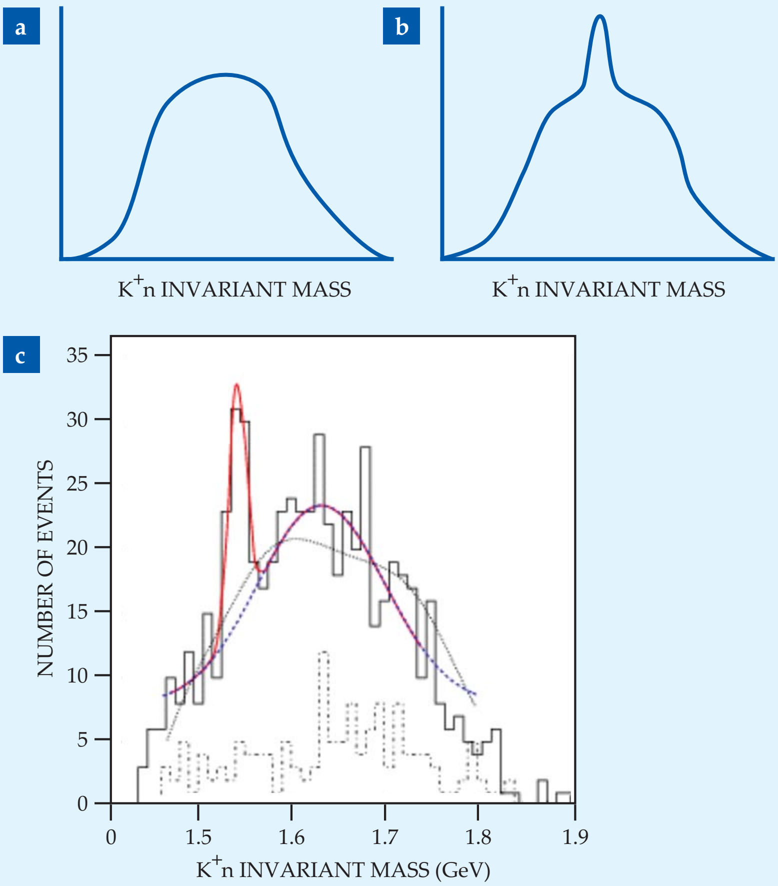


Figure 4. Invariant-mass distributions of K$^+$–neutron pairs produced in the bombardment of deuterons by high-energy photons. The standard assumption that all baryons are three-quark states forbids the existence of a particle species X that decays to K$^+$n. **(a)** Assuming there is no X, one expects a smooth distribution due to background processes. **(b)** The appearance of a peak above background at some invariant mass might signal the existence of X with that mass. **(c)** The actual data from a 2003 experiment[8] exhibit an apparently significant peak (red) at 1.54 GeV. But the evidence for an exotic "pentaquark" baryon at that mass or any other has not survived the accumulation of more data.

Trying to decide between two hypotheses, even when formally employing a frequentist approach, one may subconsciously be using a sort of Bayesian reasoning. The essential question is: What does the experimenter believe about the competing hypotheses after seeing the data?

Though frequentists are loath to assign probabilities P to hypotheses about nature, they're informally using the likelihoods L and the Bayesian priors in the following recipe:

$$\frac{P(H_1|data)}{P(H_0|data)} = \frac{L(data|H_1)}{L(data|H_0)} \times \frac{Prior(H_1)}{Prior(H_0)}. \qquad (7)$$

While the likelihood ratio might favor the new-physics hypothesis H_1—for example, a particular extra-dimensions theory with specified values of the theory's free parameters—the experimenters could well deem H_1 intrinsically much less likely than the

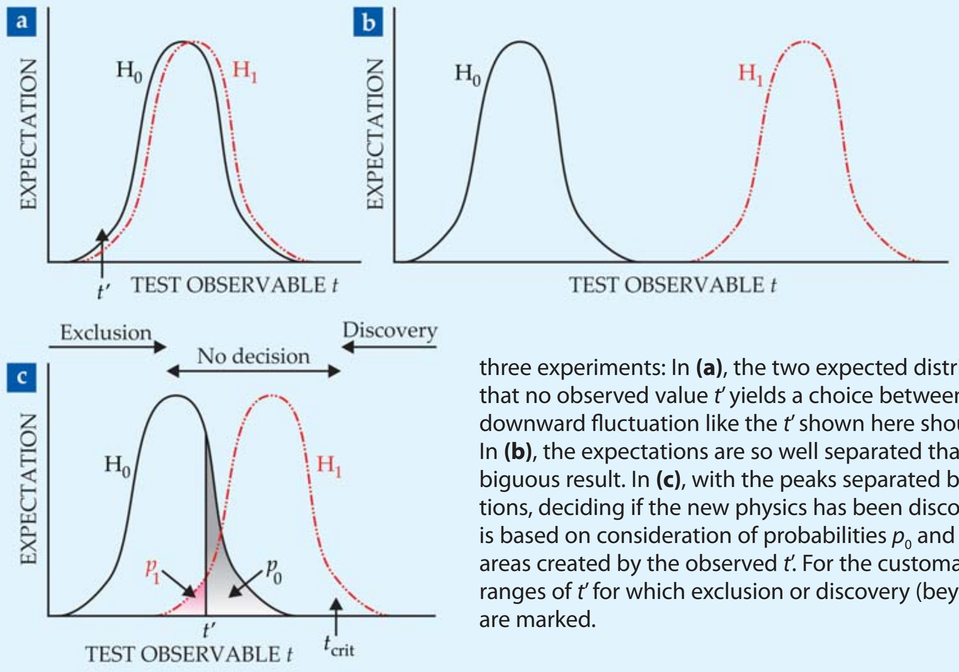


Figure 5. Two different hypotheses, H_0 and H_1, yield different expected distributions (black and red curves) of an experimental observable t meant to distinguish between them. Both hypotheses assume that the t distribution reflects known background processes, but H_1 assumes that it also reflects putative new physics. Consider three experiments: In **(a)**, the two expected distributions have so much overlap that no observed value t' yields a choice between hypotheses. In particular, a downward fluctuation like the t' shown here should not be invoked to exclude H_1. In **(b)**, the expectations are so well separated that there's little chance of an ambiguous result. In **(c)**, with the peaks separated by about three standard deviations, deciding if the new physics has been discovered, excluded, or left in doubt is based on consideration of probabilities p_0 and p_1 given by the shaded fractional areas created by the observed t'. For the customary criteria discussed in the text, ranges of t' for which exclusion or discovery (beyond t_{crit}) of H_1 can be claimed are marked.

standard model and thus assign it a smaller prior. That's a way of invoking the philosophers' maxim that extraordinary claims require extraordinary evidence. It also explains why most physicists were inclined to believe that the recently reported evidence for faster-than-light neutrinos was more likely to be explained in terms of some overlooked experimental feature (see PHYSICS TODAY, December 2011, page 8).

For the exclusion of a discovery hypothesis, the conventional choice for γ is 5%. Because that's so much less stringent than the conventional discovery threshold, "disfavored" might be a better term than "excluded."

With the exclusion threshold so lax, cases like the t' shown in figure 5a present a special problem. Even though such an experiment has no sensitivity to the proposed new physics, there is a 5% chance that a downward fluctuation of t will result in H_1 being wrongly excluded.

A widely used ad hoc prescription for avoiding that pitfall is to base exclusion not on the value of p_1 by itself but instead on the quotient $p_1/(1 - p_0)$. That ratio of the two tails to the left of the observed t' in figure 5a provides some protection against false exclusion of H_1 when the experiment lacks the relevant sensitivity.

The "look elsewhere" effect

Usually the new-physics hypothesis H_1 is not completely specified; it often contains unknown parameters such as the mass of a new particle. In that case, an experimenter looking at a mass distribution like those shown in figure 4 would be excited to see a prominent peak anywhere in the spectrum. If, in truth, there's nothing but background (H_0), the chance of a large random fluctuation occurring *somewhere* in the spectrum is clearly larger than the probability of its occurring in a specifically chosen bin. That consideration, called the look-elsewhere effect, must be allowed for in calculating the probability of a random fluctuation somewhere in the data. Looking in many bins for a possible enhancement is like performing many repeated experiments.[6]

The effect is relevant for other fields too. For example, if a search for evidence of paranormal phenomena is based on a large number of trials of different types on many subjects, the chance of a fluctuation being mistaken for a signal is enhanced if the possible signature is poorly specified in advance and then looked for in many data subsets. That's not even to mention how small a physicist's Bayesian prior ought to be.

Statistical fluctuations must be taken into account, but so must systematic effects. Indeed, systematics nowadays usually require more thought, effort, and time than does the evaluation of statistical uncertainties.[7] That's true not only of precision measurements of parameters but also of search experiments that might yield estimates of the significance of an observed enhancement or the exclusion limit on some proposed new physics.

Concerning discovery claims, it might well be that an observed effect whose statistical significance appears to be greater than 5σ becomes much less convincing when realistic systematic effects are considered. For example, the sought-after signature of dark-matter interactions in an underground detector is the accumulation of significantly more events than the number b expected from uninteresting background processes. (See PHYSICS TODAY, February 2010, page 11.) The excess would be significant, for example, at the 5σ level if we observed 16 events when our estimate of b, based on random fluctuations, was 3.1 events. But if, because of systematic uncertainties such as the poorly known concentrations of radioactive contaminants, b might be as large as 4.4, the probability of observing 16 or more background events goes up by a factor of 100, and the observed excess becomes much less promising.

The searches for new physics continue. Cosmologists tell us that most of the matter in the cosmos consists of particle species whose acquaintance we have not yet made. And the standard model that describes the known species all too well has many more free parameters than one wants in a truly comprehensive theory. But discovery claims of new phenomena will come under close statistical scrutiny before the community accepts them.

In an appendix to the online version of this article, the author has provided a quiz.

References

1. See, for example, CDF statistics committee, http://www-cdf.fnal.gov/physics/statistics/; CMS statistics committee, http://cms.web.cern.ch/news/importance-statistics-high-energy-physics; Proceedings of the PHYSTAT 2011 Workshop, CERN, http://cdsweb.cern.ch/record/1306523/files/CERN-2011-006.pdf.
2. R. Cousins, *Amer. J. Phys.* **63**, 398 (1995).
3. J. G. Heinrich, http://www-cdf.fnal.gov/physics/statistics/notes/cdf6438_coverage.pdf.
4. J. Neyman, E. S. Pearson, *Philos. Trans. R. Soc. London A* **231**, 289 (1933).
5. L. Lyons, http://www-cdf.fnal.gov/physics/statistics/notes/H0H1.pdf.
6. Y. Benjamini, Y. Hochberg, *J. R. Stat. Soc. B* **57**, 289 (1995).
7. J. Heinrich, L. Lyons, *Annu. Rev. Nucl. Part. Sci.* **57**, 145 (2007).
8. S. Stepanyan et al. (CLAS collaboration), *Phys. Rev. Lett.* **91**, 252001 (2003). ■

Appendix: Quiz for your contemplation

1. The scruffy young man

You see a young man with untidy long hair, dirty jeans, and a sweater with holes at the elbows. You assess the probability that he works in a bank as being 0.1%. Then someone asks you what you think is the probability that he works in a bank and also plays the saxophone at a jazz club at night, and you reply 15%.

Are you being inconsistent?

2. Is this theory acceptable?

You have a histogram of 100 bins containing some data, and you use it to determine the best value p_0 of a parameter p by the chi-square (χ^2) method. It turns out that χ_{min} = 87, which is reasonable because the expected value for χ^2 with 99 degrees of freedom is 99 ± 14. A theorist asks whether his predicted value p_{th} is consistent with your data. So you calculate $\chi^2(p_{\mathrm{th}}) = 112$. The theorist is happy because that's within the expected range. But you point out that the uncertainty in p is calculated by finding where χ^2 increases by 1 unit from its minimum. Because 112 is 25 units larger than 87, it's equivalent to a 5-standard-deviation discrepancy, and so you rule out the theorist's value of p.

Which of you is right?

3. The peasant and his dog

A peasant p is out on a path in the country with his dog d. The path is crossed by two strong streams, situated at $x = 0$ km and at $x = 1$ km. Because the peasant doesn't want to get his feet wet, he's confined to the region between $x = 0$ and $x = 1$; d can be at any x. But because d loves his master, he has a 50% chance of being within 0.1 km of p. That obviously implies that p has a 50% chance of being within 0.1 km of d. But because d has a 50% chance of being farther than 0.1 km from p, d could be at $x = -0.15$ km, in which case there is zero chance of p being within 0.1 km of d. That conclusion appears to be in conflict with the expected equality of the probabilities.

What is the relevance of the story to the analysis of particle-physics data?

Solutions

1. The scruffy young man

In this example, probability means the fraction of people like the scruffy young man who fulfill the stipulated conditions. Obviously there cannot be more men who work in banks and also play the saxophone than there are men who work in banks. Therefore the two probability estimates are inconsistent.

2. Is the theory acceptable?

For our given data, χ^2 has a minimum of 87 when the parameter $p = p_0$ and the shape of the $\chi^2(p)$ curve around the minimum will be approximately parabolic. The uncertainty σ on p is given by

$$\chi^2(p_0 \pm \sigma_p) = \chi^2(p_0) + 1.$$

If we were to repeat the experiment many times, fluctuations would cause the parabolae to be in slightly different positions but have the same curvature. One expects the minimum values of χ^2 to vary around 99 by ±14 and the p positions of the minima to move horizontally around the true value by $\pm\sigma_p$. So if $p_{\rm th}$ were the true value, the minimum χ^2 for a different data set could well be 112. But the minimum should be in the vicinity of $p_{\rm th}$, with $\chi^2(p_{\rm th})$ not much more than 1 larger than the minimum value. In our case, the difference is 25, which corresponds to an extremely unlikely 5σ fluctuation. So it can almost certainly be ruled out. In comparing two possible parameter values, the useful chi-square indicator is the difference in χ^2s rather than their individual numerical values.

3. The peasant and his dog

The probabilities are different because we are comparing the probability of separation distances, given the position of the peasant, with that of separation distances, given the position of the dog. The different definitions of condition in the conditional probabilities allow them to be different.

The particle-physics analog of the peasant is a parameter confined to lie in a restricted region—for example, a branching fraction, or the squared sine of an angle. The dog represents the data used to estimate the parameter. One expects the data to yield a result that's close to the true value of the parameter, but imperfect experimental resolution sometimes gives a result outside the parameter's physically allowed range. So the symbols *p* and *d* could stand for parameter and data.