

Inferring from data

Systematic uncertainties

What is systematics?

I never encountered any precise, rigorous definition. Everybody assesses systematic uncertainties all the time, but when it comes to define them there are only empiric definitions based on examples. Need a precise definition otherwise too much room for subjective arbitrariness.



Any statistical inference is based on $p(x|m)$: observe x to extract information about m , assuming to know the distribution $p(x|m)$, that is “the model”.

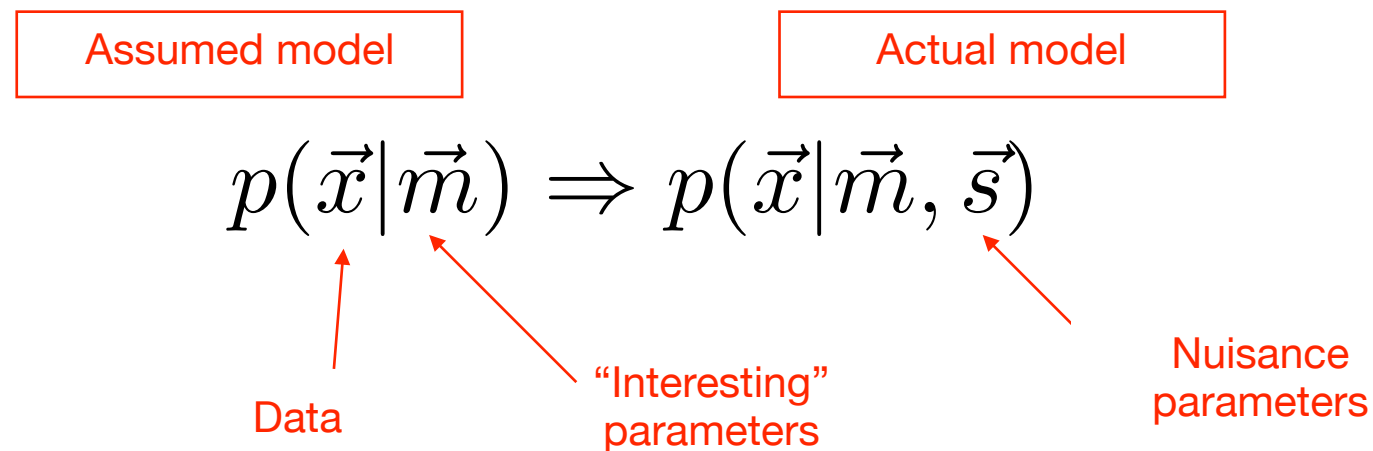
The systematic uncertainty is that component of the uncertainty that is due to the imperfect knowledge about the probability distribution $p(x;m)$.

G. Punzi, 2001

(think a second to all the cases of systematic uncertainties you are familiar with you'll realize they most likely do fit in this definition)

Nuisance parameters

Assume model $p(x|m)$. But the actual true model may differ. The difference is parametrized by the additional dependence on some **unknown nuisance parameters** that are not interesting for the measurement at hand but do influence its outcome.



Width of $p(x|m)$ connects with the statistical uncertainty. The shape, which depends on nuisance parameters, connects with the systematic uncertainty: not only you don't know exactly what value of x would be observed if m had some definite value; you don't even know exactly how probable each possible value of x is.

Cannot define standard deviation for s ; would imply knowing the distribution $p(s)$. But then s wouldn't be any longer a nuisance and would get embedded in the model!

Can only estimate an allowed range for s , and **ensure that any result of the inference hold for any s in that range.**

Experimentally known external inputs

The claim that $p(s)$ is unknown may not apply to systematic effects associated with uncertainties on external inputs whose value is known experimentally from auxiliary measurements like, e.g., input masses or lifetimes of known particles.

Due to the experimental nature of the determination of these parameters, their distributions are likely to be Gaussian (central limit...) and thus can be incorporated in the model using “Gaussian constraints” to the externally known central value μ with width σ

$$p(x|m, s) \approx p(x|m, s) \times G(s|\mu, \sigma)$$

This way, the uncertainty due to external experimental inputs gets encapsulated into the variance of the estimator (the “uncertainty returned by the fit”).

Don’t do this with external inputs that are not determined experimentally. For instance, theoretically based external inputs are not Gaussian in general, and their distribution is unknown.

Bayesian approach

For Bayesians, s is just another parameter. Assume an a priori distribution for s that allows “integrating it out” through marginalization and use the result $p(x|m)$ as model for any subsequent (Bayesian) inference.

$$p(\vec{x}; \vec{m}) = \int p(\vec{x}; \vec{m}, \vec{s}) p(\vec{s}) d\vec{s}$$

- As usual, the dependence of results on the chosen prior $p(s)$ may be an issue
- Results from multiple measurements based on independent data but sharing nuisance parameters may get correlated (by the common prior)
- Do not mix Frequentist and Bayesian approaches. E.g., don't use marginalized $p(x;m)$ to get Neyman confidence intervals. Hybrid results are the hardest to interpret. If you assume the distribution of parameter s known, you enter the Bayesian realm, hence should rather assume known the distributions of *all* parameters.

Frequentist approach — systematic uncertainties in point estimates

You use a ML estimator to get a point estimate $\hat{m}$. The model $p(x|m,s)$ contains a nuisance parameter s . Value of s and $p(s)$ are unknown but can guess, or fit from data, a “realistic / reasonable” default value s_0 and a range around it. Typically two options are available

1. Fit data multiple times, each with a likelihood that assumes a different value for s . The distribution of the resulting values of $\hat{m}$ gives an idea systematic uncertainty. This is **faster**, but the resulting **uncertainties are overestimated**: since the observed variations include statistical fluctuations due to changes of fit model — undesirable if statistical fluctuations are large.
2. Choose a value for m and one for s . Generate many samples of pseudodata from $p(x|m,s)$. Fit each sample with the default model. Repeat for various choices of s . The width of the difference distribution between values of $\hat{m}$ obtained in samples generated with s_0 and samples generated with alternative values of s gives an idea of the statistical uncertainty at a given m . Repeat for various choices of m .
More laborious but more accurate

Frequentist approach — interval estimates

The goal remains to devise a procedure that **guarantees coverage** *whatever* is the value of the unknown nuisance parameters.

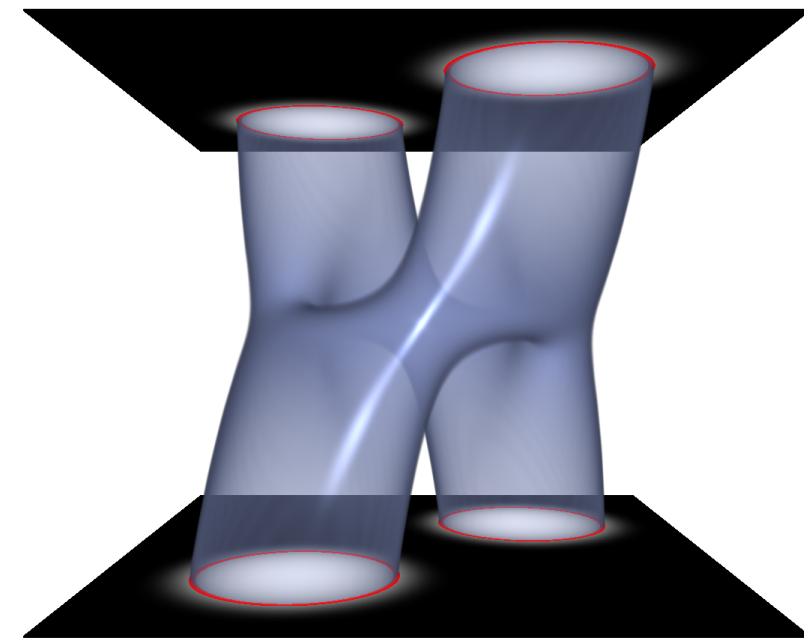
Because the allowed and plausible range of the possible true values of the nuisance parameters might be large, the coverage requirement may result into overcoverage for some values of s .

Finding the optimal procedure to obtain rigorous frequentist confidence intervals in the presence of nuisance parameters is the object of current active research in the statisticians/physicist community (see papers by, e.g., Rolke, Lopez and Conrad or K. Cranmer)

Consensus has not yet been reached on what is the best approach. However, in recent years the generalization of the LR ordering using a profiled likelihood seems to gain momentum. This is the only method we'll discuss here.

Why not just multidimensional LR-ordering?

Can one generalize the construction based on likelihood-ratio ordering to multidimensional likelihoods, $L(\vec{m}, \vec{s})$? I.e., construct a multidimensional confidence belt whose intersection with observed value $\vec{x}$ is “projected” onto the subspace of interesting parameters $\vec{m}$ to get the interval?



One can. But has to deal with two **serious issues**.

- Due to geometry, projections of higher dimensional structures into lower-dimensional subspaces lead to information loss. Structures in the multidimensional space overlap and get “shadowed”, leading to broader (i.e. lower-precision) intervals that may extend over the whole domain of m , making the inference non informative.
- A LR-based construction implies generation and fit of pseudodata that achieve an adequate sampling of the space of parameters. If such space is multidimensional, the needed computing power quickly diverges,

Profile-likelihood ratio ordering

An attractive and promising approach is to perform ratio-ordering on the likelihood profiled with respect to its nuisance parameters. This is called a profile-likelihood, which *is not a likelihood*, but a lower-dimensional derivation of it: the likelihood is a multidimensional function of the physics and nuisance parameters; The profile likelihood is a lower-dimensional function of the physics parameters only obtained by maximizing the likelihood wrt to the nuisance parameters.

Variable	Meaning
m	Parameters of interest ("physics parameters")
s	Nuisance parameters
$\hat{m}, \hat{s}$	Parameters that maximize $L(x m, s)$
$\hat{s}^*$	Parameter that maximizes $L(x m = m_0, s)$

$$\text{PLR} = \frac{L(x|m=m_0, \hat{s}^*)}{L(x|\hat{m}, \hat{s})}$$

Profile-likelihood ratio ordering

Algorithm is similar to that of the LR ordering.

1. Choose one value m_0 for m and one value s_0 for s , and generate pseudodata accordingly
2. For each observation x (i) maximize $p(x|m=m_0,s)=L(m=m_0,s)$ with respect to s to get $L_x(m=m_0,\hat{s}^*)$ and (ii) maximize the likelihood $L(m,s)$ over the space of m and s to obtain $L_x(\hat{m},\hat{s})$
3. Rank all x in decreasing order of profile likelihood ratio $PLR=L_x(m=m_0,\hat{s}^*)/L_x(\hat{m},\hat{s})$
4. Start from the x with higher PLR and accumulate the others until the desired CL is reached.
5. Repeat for all values of m
6. Repeat for values of s sampled in a plausible range

Need to **fit** each sample **twice**, one with all parameters (physics and nuisance) floating, and another one with physics parameters fixed to their test value m_0 .

It has been shown that the PLR is independent of the true values of the nuisance parameters s and, asymptotically, its distribution too gets independent of them.

When solving a given problem, try to avoid solving a more general problem as an intermediate step.

V.I. Vapnik

Inferring from data

Asymptotic properties of (profile) likelihood ratios

LR- and PLR-ordering a CPU nightmare

Neyman constructions using (P)LR-ordering imply generation and fit of large numbers of simulated experiments. These are necessary to determine the (P)LR distributions needed to build the confidence belt, for a set of true values of physics and nuisance parameters of the likelihood that adequately sample its n -dimensional space.

Given a target sampling density d , needed computing power grows like d^n and becomes soon unmanageable (I dealt with a 27-dimensional case, and that was a nightmare, <http://arxiv.org/pdf/0810.3229v2.pdf>)

Can we use good-enough approximations of the (P)LR distributions for every value of the likelihood parameters, and that are based only on the observed likelihood function?

Wilks' theorem

Asymptotically (i.e. for large N), **the distribution of likelihood ratio**

$$-2 \ln \text{LR}(m_0) = -2 \ln \frac{p(x|m_0)}{p(x|\hat{m})}$$

approaches a χ^2 distribution with a number of degrees of freedom corresponding to the dimensionality of m.



Samuel S. Wilks (1906-1964)

This holds independently of the shape of $p(x|m)$ and on the value of m.

Facilitates enormously usage of likelihood- and profile-likelihood-ratio as ordering quantities in the construction of intervals. If the likelihood is regular enough to be in asymptotic regime, one can avoid massive production of simulated experiments.

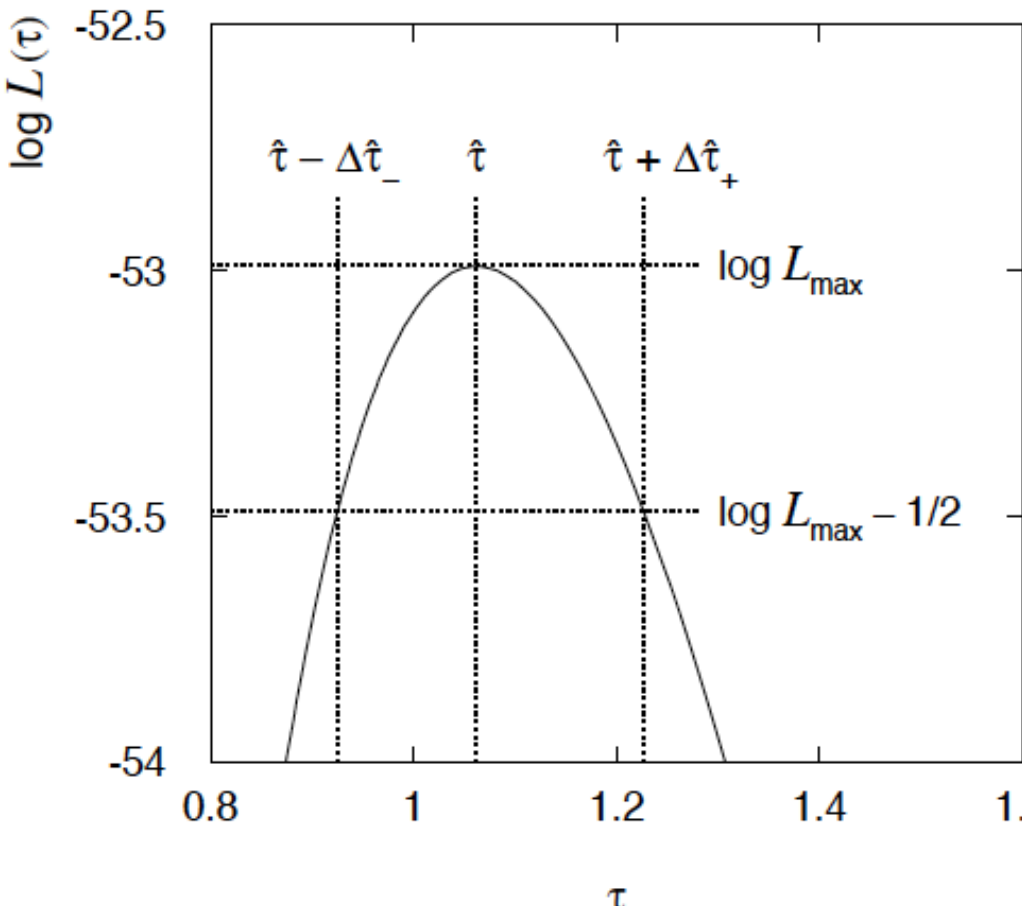
One can use few samples of pseudodata to check if the likelihood is already asymptotic. If it is, then trust Wilks' theorem; if not, only pseudodata allow determining the needed distributions

Wilks' theorem at work

Remember the graphical construction for the variance of a one-dimensional ML estimator?

Wilks' theorem tells us we can use this in any number n of likelihood dimensions to find **approximate central acceptance regions** in Neyman constructions that use (P)LR-ordering

$$-2 \ln \text{LR}(m_0) = -2 \ln \frac{p(x|m_0)}{p(x|\hat{m})} = \Delta$$



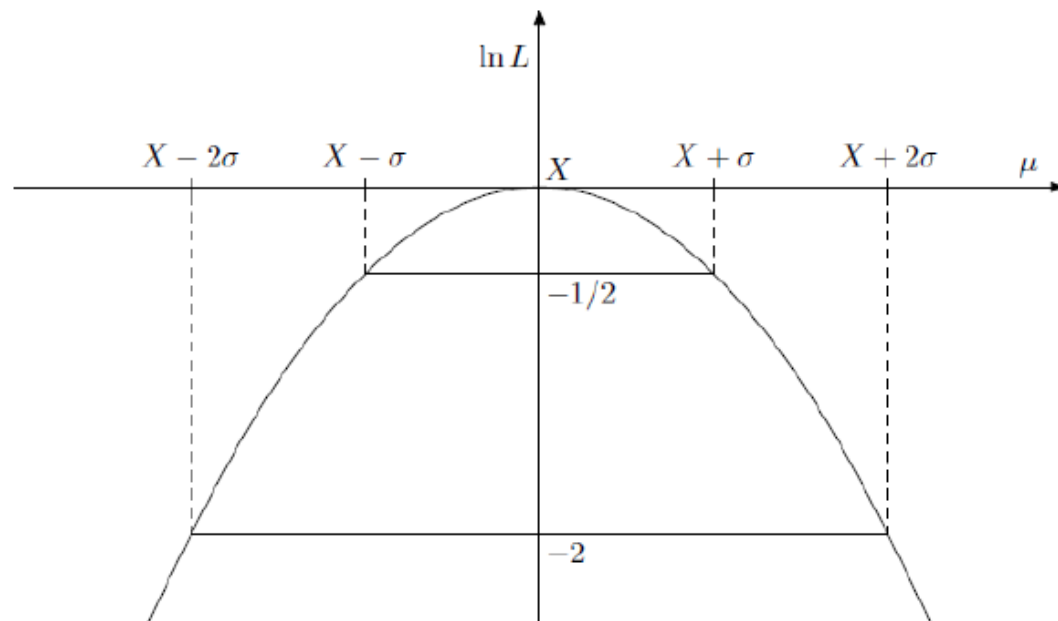
Δ	CL				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

projection onto the space of parameters of a 1(2)-dimensional likelihood at the point where $-2\ln\text{LR}$ varies by 1.0 units identifies a 1(2)-dimensional 68(39)% CL central interval

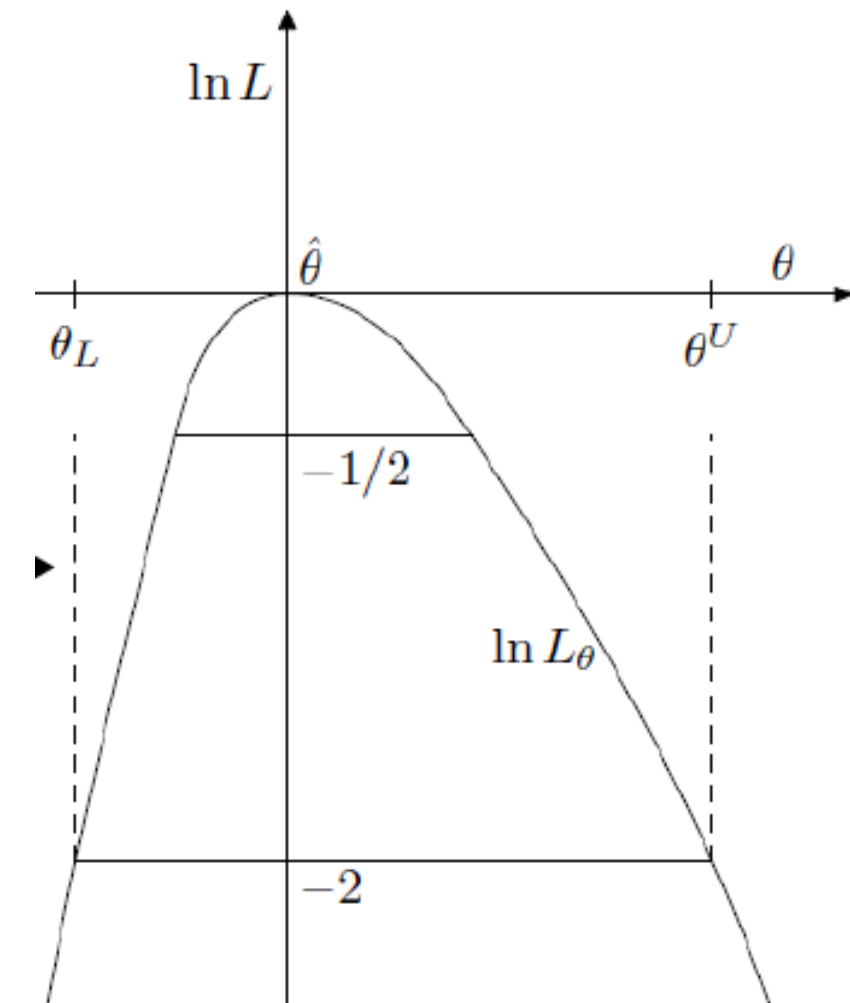
CL	$\hat{\Delta}$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

projection onto the space of parameters of a 3-dimensional likelihood at the point where $-2\ln\text{LR}$ varies by 6.25 units identifies a 3-dimensional 90%CL central interval

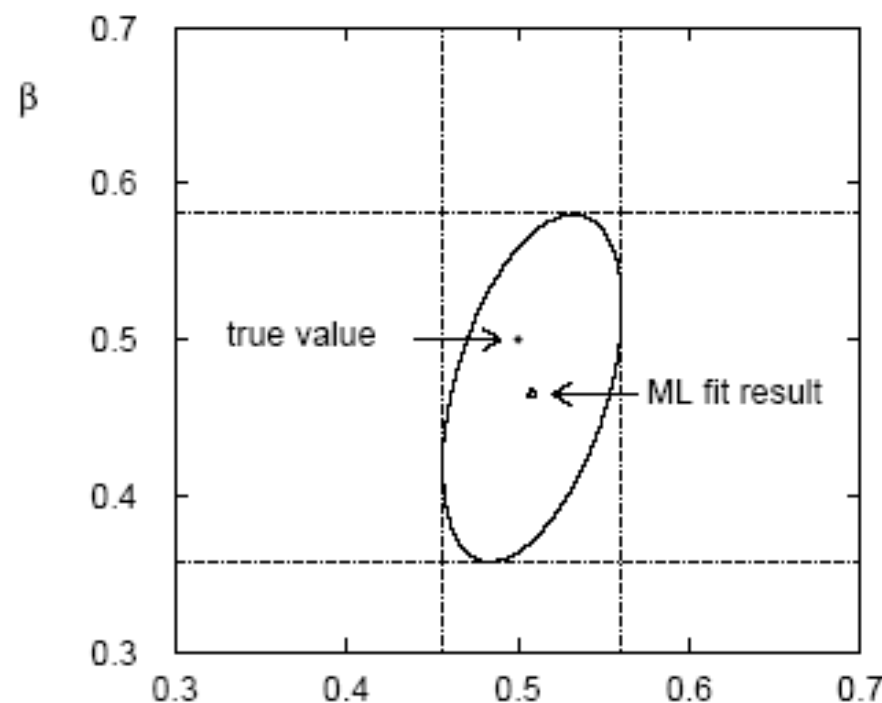
Wilks' theorem at work (F. James)



One-dimensional Gaussian likelihood



One-dimensional non-Gaussian likelihood



Two-dimensional Gaussian likelihood

In god we trust, all others bring data

W. E. Deming

Hypothesis testing

Goodness-of-fit

Significant deviation?

Often experimentalists are confronted with the need to evaluate if an observed effect constitutes a significant deviation from the expectations or if it is likely to arise from a statistical fluctuation of known phenomena.

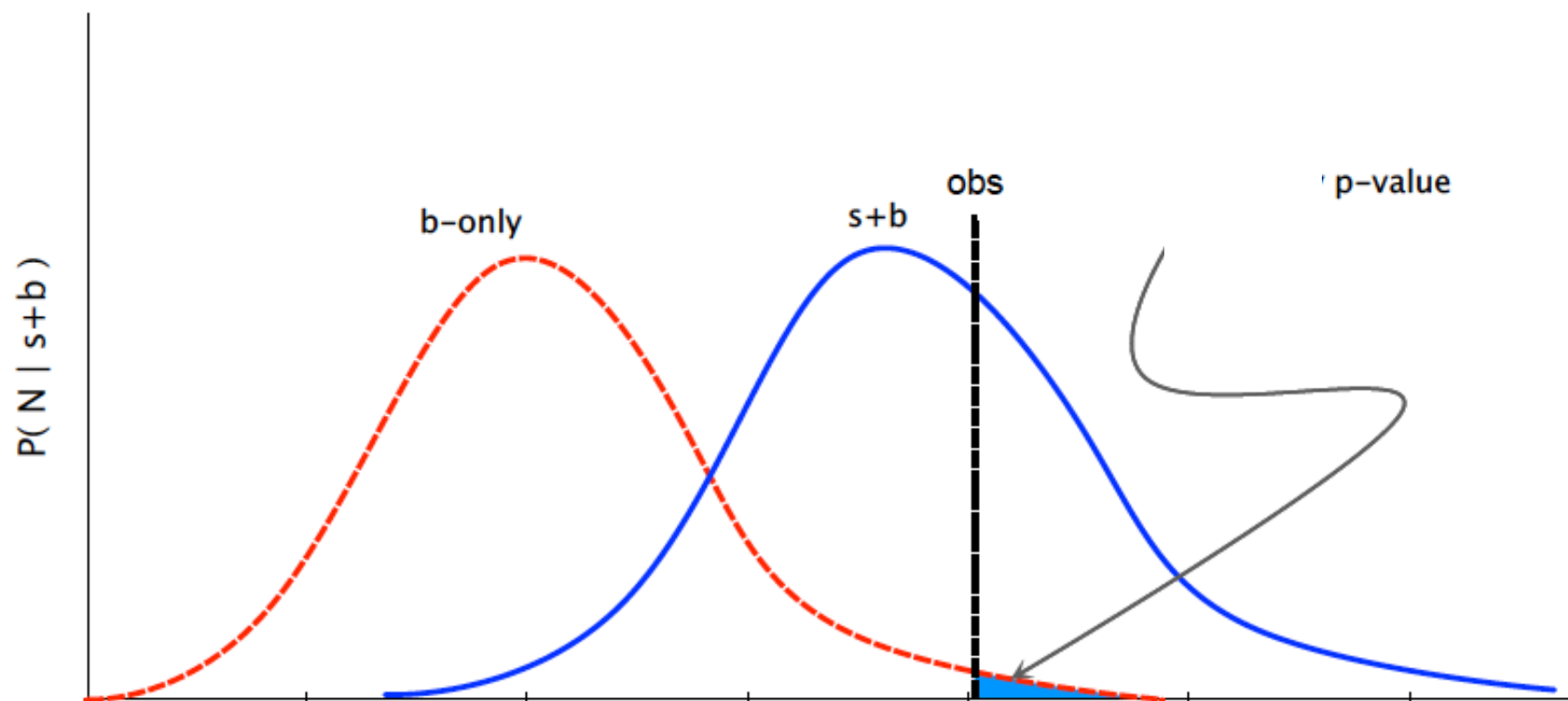
First thing you will do if you suspect you may have a discovery (but will also be useful in many other less exciting situations).

Frequentists use p-values, random variables that provide a quantitative evaluation of the probabilities to be observing a genuine anomaly or a fluke.

Need a $p(x|m_0)$ for the null hypothesis H_0 , which corresponds the expectation (“background only”). Need an observation in data, x_0 . Need a way to compare $p(x|m_0)$ with the observed x_0 to expressing the probabilities for observing the actual data x_0 (or more extreme) if H_0 is true thus quantifying the probability that the observation is due to a fluctuation under H_0 .

p-values

p-value is the relative fraction of the integral of the null model over values of data as extreme as observed and more extreme. The smaller the p-value, the higher the evidence against H_0 .



Arbitrary function of the data that allows separation between H_0 and H_1

p-value is not a probability! It is a random variable (function of the data). Important: **does not express the probability that H_0 is true or false!** It is a quantity related to the probability that, if H_0 were true, one would observe the value of x observed in data or more extreme.

p-values in coin tossing (G. Cowan)

Check if a coin is fair. The probability to observe j heads in n trials is binomial

$$f(j; n, p) = \binom{n}{j} p^j (1 - p)^{n-j} = \frac{n!}{(n-j)!j!} p^j (1 - p)^{n-j}$$

Hypothesis H_0 , the coin is fair ($p=0.5$). We do 20 trials and get 17 heads.

Regions of data space with equal or lesser compatibility with H_0 , relative to $j=17$ include $n=17, 18, 19, 20, 0, 1, 2, 3$.

$P(n=0,1,2,3,17,18,19,\text{or } 20) = 0.26\%$ Hence, if H_0 were true (coin is fair) and we would repeat the experiment many times, only 0.26% of the times we would obtain a result as extreme or more than that observed.

Often p-values are transformed into equivalent number of standard deviations a Gaussian variable should fluctuate to give the same p-value (a “ $n\sigma$ deviation”) by using the known area of the tails of the Gaussian. 0.26% corresponds to $> 3\sigma$, suggesting evidence against H_0 (coin seems not fair)

p-values in signal significance (G. Cowan)

Suppose you observe n events in a counting experiment, n_b of which are expected from background and n_s may be due to an unknown new process

If n_s and n_b are Poisson-distributed with means s and b , then n_s+n_b is also Poisson-distributed with mean $s+b$.

$$P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

Suppose $b=0.5$ is well known, and you observe 5 events. Have you made a discovery? Evaluate p-value for hypothesis $s=0$.