Statistical Methods in the context of the two problems

Ali Al Kadhim

June 13, 2019

1 Probability: Axioms and Interpretations

Statistics in particle physics in the landscape of particle physics is the liaison between theory and experiment. Theories make prediction for some observable, which will have a certain number of approximations. Also the predictions will not be very accurate and will include some adjustable or free parameters. The measurement will then be taken with some error associated (random fluctuations). So we would like to estimate the values of the parameters. Furthermore, we want to quantify the uncertainty in the parameter estimates. Also, we want to test the theory or model.
There will be uncertainties associated with it, some by randomness and some by the fact that the world is not deterministic (quantum mechanics).To quantify these uncertainties, we have to use probability.
Let me give a quick summary of the basic axioms of probability (derived by Kolmogorov). We have a set S called the sample space, and we have subsets $A, , B, ...$ then

1. for all $A \in S, P(A) \geq 0$
2. P(S = 1)=1
3. If $A \cap B = \emptyset$ then $P(A\ cupB) = P(A) + P(B)$

The first axiom says any subset of the sample space has probability greater or equal to 0. The second says the probability of the sample space is 1. The third says if we have two sets with no intersection (disjoint), then the probability of their union is the sum of the probability of each set.
There are two interpretations of probability: one is the frequentist, where elements of the sample space S represents possible outcomes of the repeatable experiments (or observation). We say that $P(A_) = \lim_{n\to\infty} \frac{timesoutcomeisA}{n}$. Where n is the number of experiments or observations.
The second interpretation is the Bayesian interpretation where $A, B, , ...$ are hypotheses (true or false), and $P(A)$ = degree of belief that hypothesis A is true. The majority of tools in particle physics use the frequentist interpretation, but

if we were concerned whether a particular theory is *true* we need to use the Bayesian interpretation. In certain cases, the two interpretations will lead to different different numbers and in others they will agree on the statistical numbers.
Next. let's define conditional Probability. If we have two subsets A and B, the conditional probability, or the probability of A given (that we know) B is

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

Or B is the restriction on S where the intersection is the logical "and".
Independence is defined as: A and B are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

.
Baye's Theorem:

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

Suppose we stary with the sample space, and restrict it with B, and suppose we restrict it with some disjoint collection of A_i such that the union of them gives the sample space $\cup_i = S$ then $P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$ which is called The Law of total probability. Then Baye's theorem can be rewritten as

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

In the frequentist interpretation the probability is only given for data, so we can't say if a model is preferred. A preferred model predicts a high probability for data "like" the data that you got in this interpretation.
In Bayesian statistics, we can extend interpretation of probability to include subjective probability for hypotheses. We would like to know $P(H|\vec{x})$ where H is some hypothesis, and x is a vector of data, so using Baye's theorem

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\sum_i P(\vec{x}|H_i)\pi(H_i)}$$

Where $\pi(H) = P(H)$ is the probability of a a particular hypothesis before we got any data (prior probability), and $P(\vec{x}|H)$ is the "likelihood", $P(H|\vec{x})$ is the probability of the hypothesis after we got the data, or posterior probability. The denominator basically says if we sum the hypothesis over all hypotheses then it equals 1 or one of them has to be true (like a normalization constant).
So where do we get this prior $\pi(H)$? In Objective Bayesian Statistics we use a formal rule or an invariance principle. In subjective Bayesian Statistics, we use an individual's subjective opinion or "illicitation of expert opinion". This is a long topic and has been a source of debate between th two schools. One

of the applications of Bayesian statistics in the LHC is setting upper or lower limits for parameters. One thing to mention about Bayesian statistics is that the probability of the data given θ $P(\vec{x}|\theta)$ gets narrrower as the accuracy of the data improves (more data taken), so if we have a wide or uninformative prior, the posterior follows the likelihood. In the limit that we have more (or more accurate) data, the posterior follows the likelihood, and the exact nature of the prior is not important. Therefore we have to start with a sufficiently wide prior.

2 Random Variables

Random variables are numerical lables for any element in the sample (or hypothesis) space. The probability density function is

$$f(x) = \int_a^b f(x)\ dx = P(a < x < b)$$

The joint PDF $f(x, y, ...)$, is if it has more than one argument. If we only want one of the variables, then we integrate over the unwanted variables. This is called the Marginal PDF. If we want the PDF of x, and we had the PDF of x and y then

$$f_x(x, y, ...) = \int f(x, y)\ dy$$

The conditional PDF is the PDF of x given y is held to a fixed value

$$f(x|y) = \frac{f(x, y)}{f_y(y)}$$

The expectation value

$$E[x] = \int x f(x)\ dx = \mu$$

The variance

$$V[x] = E[(x - \mu)^2] = E[x^2] - \mu^2$$

The Covariance

$$Cov[x, y] = E[xy] - \mu_x \mu_y = E[(x - \mu_x)(y - \mu_y)]$$

3 Parameter Estimation

Very often the data that we have might include some unknown parameters, and we'd like to estimate those. Say we have a random variable x which is sampled from a PDF which contains some unknown parameter θ or which follows

an exponential distribution, or $x \sim f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}$ and we have data $\vec{x} = (x_1, ...x_n)$ our goal is to construct a function $\hat{\theta}$ by an "estimator". Since $\hat{\theta}$ is a function of a random variable, it itself is a random variable which can be described by a random variable. There is not unique best estomator and we have to choose a best estimator according to parameters. One of the properties is the Bias of an esitmator, which is defined as

$$b = E[\hat{\theta}] - \theta$$

So we want to choose an estimator that minimizes the bias (is close to the real value of θ).
The width of the distribution of θ represents how reproducible the measurement of θ is of that type, so the width of that distribution represents the statistical error $\sigma_{\hat{\theta}}$ (standard deviation) is statistical error of $\hat{theta}$ and we want it to be small. We also want as small as possible variance. However, in general we cannot optimize with respect to more than one (conflicting) criterion.

3.1 The Method of Maximum Likelihood

If I have a measurement that produces data, and it contains a parameter θ, that function is called "the model" $P(\vec{x}|\theta)$ which assigns the probability for data for every point in the parameter space. If we only look at the dependence on the hypothesis or the parameter θ it is called the likelihood Ł(θ). THe method says the maximum likelihood estimator is the value of θ that maximizes the likelihood.

$$\hat{\theta}_{ML} = argmax_\theta \; L(\theta)$$

The value of θ that maximizes the likelihood is equivalent to the value that maximizes $\ln L(\theta)$.
Say we have the PDF $f(t;\tau) = \frac{1}{\tau}e^{t/\tau}$ which describes the decay time of particles. Then say we have data for n independent identically distributed distributions. Then the joint probability of the data is just the product. Of we were to only consider the parameter τ

$$f_{joint} = \prod_{i=1}^{n} f(t;\tau) = L(\tau)$$

To find the value of τ which maximizes the likelihood, we take the natural log of the distribution and maximize it for τ

$$\ln L(\tau) = \sum_{i=1}^{n} (\ln \frac{1}{\tau} - \frac{t_i}{\tau}) = 0$$

so $\hat{\tau}(t_1, .., t_n) = \frac{1}{n}\sum_{i=1}^{n} t_i$ We need to check the properties of this estimator.

$$E[\hat{\tau}] = E[\frac{1}{n}\sum_{i=1} t_i] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau$$

$$var[\hat{\tau}] = var[\frac{1}{n}\sum_{i=1} t_i] = \frac{1}{n^2}\sum_{i=1} var[t_i] = \frac{\tau^2}{n}$$

and standard deviation $\sigma = \sqrt{var}$

$$\sigma = \frac{\tau}{\sqrt{n}}$$

In general, the PDF's could be more complicated so that we need some approximation methods. We can estimate the variance from the "Information Inequality", which says

$$V[\hat{\theta}] \geq \frac{(1 + \frac{\partial L}{\partial \theta})^2}{-E[\frac{\partial^2 \ln L}{\partial \theta^2}]}$$

Where

$$-E[\frac{\partial^2 \ln L}{\partial \theta^2}] = -\int \frac{\partial^2 \ln L}{\partial \theta^2} P(\vec{x}|\tau)\ d\vec{x}$$

Often, the bias can be neglected. Also, the inequality is an approximate equality in the limit $\lim_{n\to\infty}$, so

$$V[\hat{\theta}] \approx -\frac{1}{E[\frac{\partial^2 \ln L}{\partial \theta^2}]}$$

So often we can estimate the variance as

$$V[\hat{\theta}] = -\frac{1}{E[\frac{\partial^2 \ln L}{\partial \theta^2}]}\Big|_{\hat{\theta}=\theta}$$

If we have vector of data $\hat{\theta} = (\theta_1, .., \theta_m)$ and we have the covariance of each pair of esitmator, then we have the covariance matrix $Cov[\hat{\theta_i}, \hat{\theta_j}] = V_{ij}$.
To obtain an estimate of the covariance matrix, we use the Fischer matrix.

$$I_{ij} = -E[\frac{\partial^2 \ln L}{\partial \theta_i\ \partial \theta_j}]$$

So that the information inequality becomes the statement $\vec{V} - \vec{I}^{-1}$ is positive definite. So we approximate

$$V_{ij} \approx I_{ij}$$

So we find $\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}$ numerically at $\vec{\hat{\theta}} = \vec{\theta}$ and then invert to get the covariance matrix.
Say we have two events, i.e. signal and background as our data, where each event is characterized by a continuous variable x. If the signal events $f_s(x)$ is Gaussian distributed with mean μ_s, and the background events $f_b(x)$ is exponential distributed, with mean μ_b, then the PDF of x is

$$f(x) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_b + \mu_s} f_b(x)$$

3.2 Hypothesis Tests

A hypothesis H is a statement that is either true or false. A hypothesis also gives you a rule to find the probability for the data $P(\vec{x}|H)$. Very often often the hypothesis is parameterized by the parameters of a particular model θ, so $P(\vec{x}|H) = P(\vec{x}|H, \theta)$.
Consider a single hypothesis H_0, a hypothsis which is going to be either accepted or regected, often called the "null hypothesis". Consider an alternative hypothesis H_1. The critical region is in $\vec{x}$ space which has the property $P(\vec{x}|H_0) \leq \alpha$ where α is a pre-specified small number, like 5%, called the size (significance level). Then, if x is in the critical region w, reject H_0. The regions of the two hypothesis overlap, and α is defined as some cut that discriminates between the two hypotheses, the area of $f(\vec{x}|H_1)$ is the power of test with respect to H_1. SO we put $w$so as to maximize the power $M = P(x \in w|H_1)$ for some alternative H_1.
When we say that we reject H_0 we mean that we believe it to be false, meaning the probability $P(H_0)$ is low. But a statement about the degree of belief in a hypothesis necessarily has to involve Baye's theorem, which includes the prior.

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\ dx}$$

Optimal Choice for critical region is given by the Neyman-Pearson lemma. Say $P(\vec{x} \in w|H_0) = \alpha$. Consider a specific H_1. Suppose that for all $x \in w$. then $\frac{P(x|H_1)}{P(x|H_0)} \geq c_\alpha$ and $\frac{P(x|H_1)}{P(x|H_0)} < c_\alpha$ for x not in w. This region gives the best test, i.i. highest power with respect to H_1 for a a test of H_0 with size α. In practice, we have multidimensional data, say $\vec{x}$. The events according to one hypothesis are $f(\vec{x}, H_0)$ and we want to carry out a statistical test for each event, and determine if it of type H_0 or of type H_1 which has some overlap with H_1. What is the optimal critical region for test of H_0. The optimal boundary of the critical region is defined by forming a function $t(x_1, ...x_n) = t_c$ where t_c is some constant. The optimal $t(\vec{x})$ is given by the likelihood ratio $t(\vec{x}) = \frac{P(\vec{x}|H_1)}{P(\vec{x}|H_0)}$. This is not useful in practice, however, because of the inabaility to numerically calculate these probabilities. We have instead Monte Carlo programs that generate points in $\vec{x}$ space that follow, say the standard model events and other simulated points by some other model. But suppose now we measure a real event from a real experiment; we don't know whether it was created by a standard model process or some other process. We then want to know that for that point in $\vec{x}$ space, whats the value of the probability $P(\vec{x}|H_1)$ which we are unable to do from the simulations. What to do instead is writing some ansatz for $t(\vec{x})$ which has some adjustable parameters, and then we can use the simulated data to find the optimal parameters from the simulated data.
If we want an assumption of the form of this test statistic that contains a smaller number of parameters. Suppose we want an n-dimensional histogram with M cells per dimension (M bins), then we have M^n cells, which is the number of parameters to determine. Then we could try a linear ansatz for the test statistic

so that

$$t(\vec{x}) = \sum_i^n a_i x_i$$

which is called a Fisher discriminant. Another is called a Neural Network, which describes a non-linear surface in this multidimensional space. Another is called Boosted Decision Trees, and Support Vector Machines.

3.3 P-values

A hypothesis H gives $P(\vec{x}|H)$. The outcome of the experiment will be a single point in the data space $\vec{x}$, and we want to quantify or give a numerical value to the level of agreement between that observation, and the prediction of the hypothesis. The analysis will form a boundary where on one side of the boundary $\omega_<$ where $\vec{x}$ has a level of compatibility that is $\leq H$, and the other side of the boundary $\omega_>$ gives greater compatibility to the prediction of H. Less compatible with H means more campatible with some alternative, which we don't have to specify. the p-value of H is $P(\vec{x} \in \omega_<(\vec{x}_{obs}|H)$. If the p-value is small, that means there is a small probability to get even worse compatibility, therefore the model would be bad.
If we want the probability of the hypothesis we have to use the prior. $P(H|\vec{x}) \propto P(\vec{x}|H)\pi(H)$. We can show that the PDF of the p-value under the assumption that the hypothesis is correct $f(p_H|H)$ is uniformly distributed between 0 and 1.
We can use the criteria to define the critical region w

$$P(p_H < \alpha|H_0) = \alpha$$

so that $w = \{\vec{x} : p_0 \leq \alpha\}$ We can define the significance Z as $p = 1-\Phi(z)$ where Φ is the cumulative distribution of a Gaussian. So $Z = \Phi^{-1}(1-p)$. We can reject the no-signal hypothesis If $Z \geq 5$ which corresponds to $p = 2.9 \times 10^{-7}$.
In a search region in particle physics we count the number of event n with some signal and background events. n $Poisson(s+b)$ where n_b $Poisson(b)$, n_s $Poisson(s)$. But we only measure the sum of the signal and background events. Therefore $P(n|s+b) = \frac{(s+b)}{n!}e^{-(s+b)}$. Suppose the expected number of background events is known from Monte Carlo b. Then we do the experiment and find five events $n = 5$. P-value of the hypothesis that there is only background "the background-only hypothesis" $s = 0$ p_0. $p_b = P(n \geq 5|s = 0, b = 0.5)$. The alternative to $s = 0$ is $s > 0$ then the p-value is $\sum_{n<5} \frac{b^n}{n!}e^{-b} = 1.7 \times 10^{-4}$ which can be converted to the corresponding significance.

3.4 Interval Estimation

Say we have a vector of parameters $\vec{\theta} = (\theta_1, ...\theta_n)$. We want to find a region for the true value of θ with some specified probability. $Prob(\theta_{True} \in region) \geq$

$1-\alpha$. In the frequentist approach, we find p-value for all θ. If $p_\theta < \alpha$, regect θ so that the set of θ where $p_\theta > \alpha$ is the confidence interval.

3.5 Generic LHC variable

Say we are searching for a kinematic variable x. We plot the histogram with some expected background b and expected signal s for each bin $n_i = 1, ..., N$. Where $n_i \tilde{} Poisson(\mu s_i + b_i)$. So the likelihood function $L(mu) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)}$. Then to define the ctitical region we use the test statistic λ

$$\lambda(\mu) = \frac{L(\mu)}{L(\hat{\mu})}$$

We can define $t_\mu = -2 \ln \lambda(\mu)$ so if λ is 1, t is 0, and if it is low, t is high, so that high values of t correspond to bad compatibility. Then the p-value is $p_\mu = \int_{t_{\mu,obs}}^{\infty} f(t_\mu|\mu)\ dt_\mu$.
Wilk's Theorem
$f(t_\mu|\mu) \tilde{} chi - square\ dist$ for $n = 1$ degree of freedom. Where the chi-square distribution for n degrees of freedom is $f_{\chi_n^2}(t_\mu) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} t_\mu^{n/2-1} e^{-t_\mu/2}$ in large sample limit (large amount of data), so that $p_\mu = 1 - F_{\chi_1^2}(t_\mu) =^{set} \alpha$

4 Classification

Classification is about seperating observations of different catergories from each other. If observations are grouped in just two categories, this is binary classification. In binary classification, one group is often more interesting than the other. The first category (interesting) is described as "signal" and the second is described as "background". Formally, we have a scalar random variable Y and a vector random variable X. At any point X in the multivariate space, class label Y is distributed according to a mass function P(y jx), the probability of observing y at x. The goal of statistical classification is to learn the distribution P(y jx). This learning is accomplished by building (training) a predictive model on data with known class labels (labeled data). The constructed model can predict y for data without known class labels (unlabeled data).

4.1 Loss Functions

Suppose we train a predictive model on labeled data. How do we know if the predicted distribution $\hat{P}(y|x)$ is a good approximation to the true distribution $P(y|x)$? This is a tough problem. Some classifiers do not offer a straightforward way of computing $\hat{P}(y|x)$ from the learned model. In practice, the quality of learned model at point $\vec{x}$ is measured using a loss function, $l(y, f(\vec{x}))$. It can be thought of as a distance between the true class label y and predicted response

$f(\vec{x})$. Classification loss for the leaned model $f(\vec{x})$ is the expected distance,

$$L(X,Y) \equiv E_{X,Y}\ell(Y,f(X)) = \sum_{y\in\mathcal{Y}}\int_{\mathcal{X}}\ell(y,f(x))P(x,y)dx$$

over the entire domain of Y and X for the joint probability density function. The expected loss is usually estimated by averaging $l(y,f(\vec{x}))$ over the labeled data $\{(x_n,y_n)\}_{n=1}^{N}$ drawn from the joint pdf $P(\vec{x},y)$

$$\hat{L} = \frac{1}{N}\sum_{n=1}^{N}\ell\left(y_n, f\left(x_n\right)\right)$$

Assuming a distribution$P(y)$, we can model the joint pdf using $P(\vec{x},y) = P(\vec{x}|y)P(y)$. Often, the analyst would simulate the signal and background observations by Monte Carlo, based on the known forms of $P(\vec{x}|y)$ and then mix the two sets in a certain proportion. This proportion defined the prior probability $P(y)$, and mixing the two sets in this proportion amoints to simulating $P(\vec{x},y)$.
A soft classification score always expresses the level of classification confidence, but its nature varies from one model to another. For linear Discriminant analysis, the score for class y is the posterior probability $\hat{P}(y|\vec{x})$ estimated under the assumption of multivariate normality. For binary classification by support vector machines, the score is the signed distance to the hyperplane seperating the two classes: the score is: +1 if $\vec{x}$ is a support vector for the positive class, -1 if $\vec{x}$ is a support vector for the negative class, and 0 if $\vec{x}$ lies on hyperplane of optimal seperation. The simplest measire of the predictive power is 0-1 loss, or classification error:

$$\ell(y,\hat{y}) = \begin{cases} 0 & if \quad y = \hat{Y} \\ 1 & if \quad y \neq \hat{p} \end{cases}$$

The expected loss $L(\vec{X},Y)$ is then minimized by classifying every x into the most probable class, $y^*(x) = \arg\max_{y\in\mathcal{Y}} P(y|x)$. This loss equals the probability of observing one of the less probable classes,

$$\epsilon^* = 1 - \int_{\mathcal{X}} P\left(y^*|x\right)P(x)dx$$

$P(y|x)$ is often called the posterior probability and the minimal classification error is often called the Bayes error. In physics analysis, a common goal is to optimize a figure of merit expressed as a function of the expected signal and background, s and b. An example of such a figure is $s/\sqrt{s+b}$. We need an optimization algorithm! In practice, one can construct a classifier by minimizing a loss function and then find the best threshold on the soft score by optimizing the chosen figure of merit.
The predictive power of a regression error is often measured bu the mean squared

error. Its expectation, $E_{X,Y}\left[(Y-f(X))^2\right]$ is what we aim to minimize. Let $y^*(x)$ and $f^*(x)$ be the expectations of Y and F at x, respectively. Then

$$\begin{aligned} E_{Y,F|X}\left[(Y-F)^2\right] =& E_{Y|X}\left[(Y-y^*)^2\right] \\ &+ (y^*-f^*)^2 \\ &+ E_{F|X}\left[(f^*-F)^2\right] \end{aligned}$$

The first term on the right side is irreducuble noise, the second term is the square of the regression bias and the third term is the regression variance. The irreducible noise is the property of the data, and we cannot eliminate it. We can hope to build a predictive model with minimal bias and variance. Let's take a look at the simplest case, 0-1 binary classification. Let $y^*(x)$ be the most probable class at x and therefore the optimal classification at x. The minimal possible (Bayes) error is $1-P(y^*|x)$. If $\hat{y}(x)$ is the predicted label, the error at x is $\epsilon(x) = 10P(\hat{y}=y|x)$. The decomposition proposed by Breiman(1998) is

$$\begin{aligned} \epsilon(x) =& \left[1-P\left(y^*|x\right)\right] \\ &+ \left[P\left(y^*|x\right)-P\left(\hat{y}^*|x\right)\right] P_F\left(\hat{Y}^*|x\right) \\ &+ \left[P\left(y^*|x\right)-P\left(\hat{y}^{**}|x\right)\right] P_F\left(\hat{y}^{**}|x\right) \end{aligned}$$

. James (2003) obtains, for binary classification with 0-1 loss:

$$\begin{aligned} \epsilon(x) =& \left[1-P\left(y^*|x\right)\right] \\ &+ \left[P\left(y^*|x\right)-P\left(\hat{p}^*|x\right)\right] \\ &+ \left[2P\left(\hat{Y}^*|x\right)-1\right]\left[1-P_F\left(\hat{Y}^*|x\right)\right] \end{aligned}$$

One of the most important parameters for decision tree training is the minimal leaf size, that is, the minimal number of observations in a terminal node. If the training set is composed of unique observations, a binary decision tree with leaf size one usually finds a perfect seperation between the two classes in a training set.
To summarize, to optimize the predictive power of a classifier, we need to carefully select its training parameters and use as many observations as possible for training.

4.2 Optimal Splitting?

The training error increases with leaf size; it is optimistic for small leaves and approaches the test error as the leaf size approaches the value maximizing the predictive power.
The test error decreases with leaf size until the optimal value is reached. In practice, to obtain a relationship between the training and test errors, we need

to do this empirically, for a specific classifier applied to a specific problem. The training error is usually not a good estimate of the generalization error. Obviously, we would like to learn the model from abailable data as accurately as possible. At the same time, we would like to estimate its predictive power as accurately as possible. These two requirements are in conflict. The predictive power of a model would improve if we provided more training data. On the other hand, the accuracy of the estimate of the predictive power would improve if we provided more test data.

5 Asides

We have proton-proton collision. If one of the gluons happens to be having a lot of energy, the gluons could annihilate producing t top quark-antiquark together (virtual particles in a loop), which have a chance of producing the Higgs. So indirectly we've produced the Higgs with Top quarks!
Next: look into classification (quark-induced or gluon-induced) gluon fusion to produce the Higgs.
Grid Search (signal-background discrimination).
Probability Density estimation.
Support Vector Machines.
Bayes Discriminant Function.
The Fisher discriminant (FLD), random grid search (RGS), probability density estimation (PDE), neural network (ANN) and support vector machine (SVM) are simply different algorithms to approximate the Bayes discriminant function D(X), or a function thereof.
LEP Higgs Working group developed formalism to combine channels and take advantage of discriminating variables in the likelihood ratio

$$Q = \frac{L\left(x|H_1\right)}{L\left(x|H_0\right)} = \frac{\prod_i^{N_{chan}} \text{Pois}\left(n_i|s_i+b_i\right) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} \text{Pois}\left(n_i|b_i\right) \prod_j^{n_i} f_b\left(x_{ij}\right)}$$

$$q = \ln Q = -s_{\text{tot}} \sum_i^{N_{\text{chan}}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s\left(x_{ij}\right)}{b_i f_b\left(x_{ij}\right)}\right)$$