Poisson Statistics

Poisson statistics actually begins with the Binomial Distribution. The binomial distribution applies to situations involving a series of independent and identical trials, where each trial can have only one of two possible outcomes.

Suppose n coins are tossed and x is the number of heads ~~obtat~~ obtained. x can then range from 0 to n.

$$P_x(k) = \begin{pmatrix} \text{\# of ways to} \\ \text{arrange } k \text{ heads} \\ \text{and } n-k \text{ tails} \end{pmatrix} \begin{pmatrix} \text{prob. of any particular} \\ \text{sequence having } k \text{ heads} \\ \text{and } n-k \text{ tails} \end{pmatrix}$$

$$= \begin{pmatrix} \quad '' \quad \end{pmatrix} p^k (1-p)^{n-k}$$

$$= \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Now suppose the limit in which n is large. This is denoted "the Poisson Limit."

<u>Thm.</u> If $n \to \infty$ and $p \to 0$ in such a way that $\lambda = np$ remains constant, then for any non-negative integer k

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-np}(np)^k}{k!}$$

<u>Proof:</u> $$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \lambda^k \left(\frac{1}{n^k}\right) \left(1 - \frac{\lambda}{n}\right)^{-k} \left(1 - \frac{\lambda}{n}\right)^n$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \left(\frac{1}{n^k}\right)\left(\frac{n-\lambda}{n}\right)^{-k} \left(1-\frac{\lambda}{n}\right)^n$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{(n-\lambda)^k} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda} \text{ as } n\to\infty}$$

$$= \frac{\lambda^k}{k!} e^{-\lambda} \lim_{n\to\infty} \frac{n!}{(n-k)!\,(n-\lambda)^k}$$

$$= \frac{\lambda^k}{k!} e^{-\lambda} \lim_{n\to\infty} \underbrace{\frac{n(n-1)\cdots(n-k+1)\cancel{(n-k)!}}{\cancel{(n-k)!}\,(n-\lambda)^k}}_{\to 1 \text{ as } n\to\infty}$$

$$= \frac{(np)^k e^{-np}}{k!}$$

$\blacksquare$

<u>The Poisson Distribution:</u>

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad \forall k \in \{0 \cup \mathbb{Z}\}$$

λ is the expected value.

<u>Assumptions of the Poisson Model:</u>

1. Probability of an occurence for a single event (or interval) is small
2. Occurences are independent.
3. Probability of an occurence remains constant.

CDF Note 8128

Modified Frequentist Confidence Levels

Suppose n independent counting search analyses. We may define a test statistic X which discriminates signal like outcomes from background-like ones. The "likelihood ratio" is a common choice of test statistic.

$$\text{Let} \left\{ \begin{array}{l} s_i \equiv \text{estimated signal} \\ b_i \equiv \text{estimated background} \\ d_i \equiv \text{observed candidates} \end{array} \right\} \quad \text{for } i^{\text{th}} \text{ channel}$$

Then the likelihood ratio is defined by

$$X \equiv \prod_{i=1}^{n} X_i \quad ; \quad X_i \equiv \frac{\frac{1}{d_i!} e^{-(s_i+b_i)} (s_i+b_i)^{d_i}}{\frac{1}{d_i!} e^{-b_i} b_i^{d_i}}$$

Ratio of Poisson Distributions

So if there is no signal ($s_i = 0$), then $X_i = 1$. If there is only signal ($b_i = 0$), then $X_i \propto \frac{e^{-s_i} s_i^{d_i}}{b_i^{d_i} \to 0} \to \infty$. The channels "$i$" are the analyses that may be combined.

The "confidence level" for excluding the possibility of both signal and background production is

$$\underbrace{CL_{s+b} = P_{s+b}(X \le X_{obs})}_{\text{Assuming both background and signal at hypothesized levels, probability the test statistic} \le \text{observed data.}}$$

$$P_{s+b}(X \le X_{obs}) = \sum_{X(\{d_i'\}) \le X(\{d_i\})} \prod_{i=1}^{n} \frac{e^{-(s_i+b_i)} (s_i+b_i)^{d_i'}}{d_i'!}$$

Another similar value is the confidence level for the background alone.

$$\underbrace{CL_b = P_b(X \leq X_{obs})}_{\text{Probability that background processes yield fewer or equal than the number observed.}}$$

"Modified Frequentist Confidence Level"

$$\boxed{CL_s = \frac{CL_{s+b}}{CL_b}}$$

Confidence Level Calculation

Calculating $P_{s+b}(X \leq X_{obs})$ is difficult because there are n channels with m outcomes each. This sum ~~is~~ can be carried out w/ MC, selecting representative outcomes of the experiment and comparing their test statistic with the test statistic computed from data events

Alternatively, we compute a probability distribution function for the test statistic for a set of channels which can be convoluted with the PDF of other channels. The PDF for a test statistic is actually ~~dist~~ discrete. For a single channel, it is the sum of δ functions at the accessible values X_i. The possible outcomes are

$$(X_i^j, p_i^j)$$

where X_i^j is the test statistic for the ith channel if it were to have j events, p_i^j is the Poisson probability of selecting j events in the ~~a~~ i^{th} channel if the average

expected rate is $s_i + b_i$ when computing CL_{s+b} [or b_i when computing CL_b].

For n channels, the outcomes are $[X_i^j X_{i'}^{j'}, p_i^j p_{i'}^{j'}]$. We still have the problem of enumerating all possible experimental outcomes. An approximation is introduced.

The approximation is to create some number of bins of the PDF instead of dealing with each individual possible experimental outcome.

Expected Confidence Levels:

$$\langle CL_b \rangle = \sum_{i=1}^{N_{blist}} \left[p_i^b \sum_{j=1}^{i} p_j^b \right]$$

N_{blist} = # of entries in the table of the PDF of X.

$$\langle CL_{s+b} \rangle = \sum_{i=1}^{N_{blist}} \left[p_i^b \sum_{X_j^{s+b} \le X_i^b} p_j^{s+b} \right]$$

$$\langle CL_s \rangle = \sum_{i=1}^{N_{blist}} \left[p_i^b \frac{\sum_{X_j^{s+b} \le X_i^b} p_j^{s+b}}{\sum_{j=1}^{i} p_i^b} \right]$$

Systematic Uncertainty on Signal and Background

Frequentist Exclusion and Discovery (6525)

Suppose we have some data that contains a background and may or may not contain a signal. Assume the background has significant systematic uncertainties. Our goal is to determine how to distinguish between data conforming to background and data conforming to background + signal.

In order to discover a signal, the probability of an upward statistical fluctuation of the background needs to be computed. We must calculate both the significance and the expected sensitivity.

Two hypotheses will be tested at a time here:

- Background-only hypothesis
- Background + signal hypothesis

Several cases are possible:

- The hypotheses are sufficiently similar or poorly specified that they cannot be distinguished regardless of data.
- The data may prefer one hypothesis over the other, and/or exclude one.
- The data may prefer neither, excluding both.

Test Statistic, Likelihood Ratio, Confidence Level

Begin with a test statistic: a single real number which is a function of the exponential outcome, and is chosen to maximize the separation between experimental outcomes expected when a signal is present, and those expected when a background contributes.

The optimal choice for a test statistic is the likelihood ratio.

$$Q = \frac{P(\text{data}|s+b)}{P(\text{data}|b)} = \prod_{i=1}^{n_{ch}} \prod_{j=1}^{n_i^{bin}} \frac{\dfrac{e^{-(s_{ij}+b_{ij}^{s})}(s_{ij}+b_{ij}^{s})^{d_{ij}}}{d_{ij}!}}{\dfrac{e^{-b_{ij}^{b}}(b_{ij}^{b})^{d_{ij}}}{d_{ij}!}}$$

The "confidence level" for excluding the production of signal is

$$CL_{s+b} = P_{s+b}(Q < Q_{obs})$$

s_{ij} = Estimated signal rate in bin j of channel i.

b_{ij} = Estimated background rate in bin j of channel i.

d_{ij} = data count...

which is the probability that Q is less than that obtained from data, Q_{obs}.

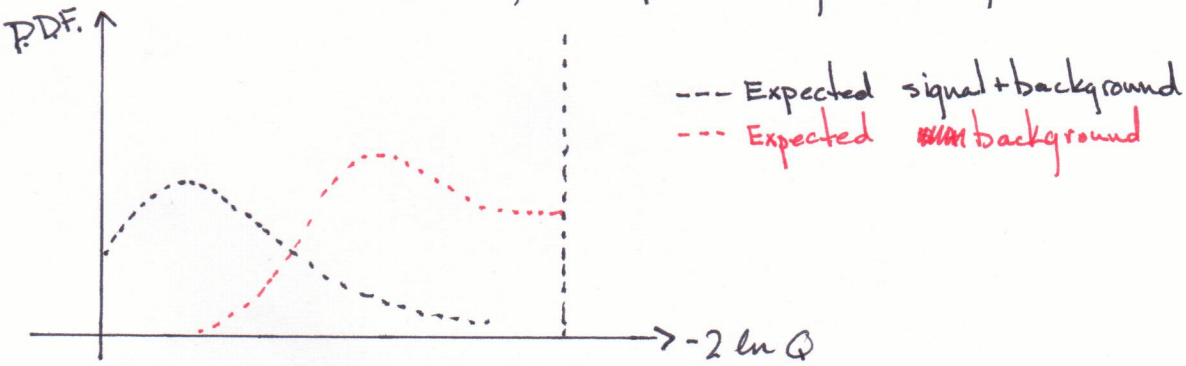
Analogously, the confidence level for excluding the background is another p-value

$$1 - CL_b = P_b(Q \geq Q_{obs}),$$

the probability that the background hypothesis will give an outcome at least as signal-like as the observation.

A problem with the definition of CL_{s+b} is that the signal + background hypothesis is excluded at 95% CL 5% of the time, including when there is no sensitivity in the experiment. If a large number of independent models are tested, then it becomes likely to have exclusions of models for which there is no sensitivity.

Another quantity used is $CL_s \equiv \frac{CL_{s+b}}{CL_b}$. This is not a p-value, confidence level, or probability. With this, it is impossible to exclude a signal just because the data fluctuated below the background. A signal may exist where the prediction is smaller than the observed.

Pseudo-experiments

Confidence levels are computed by generating an ensemble of pseudoexperiments assuming s+b and making a histogram of Q. Then generate another ensemble of pseudoexperiments assuming only b, and make another histogram of Q. Finally, calculate CL_{s+b} by integrating the probability density functions of Q.