Statistics

Craig Blocker, Brandeis

NEPPSR, 2006

Huge topic - will pick and choose.

Hope to give you an understanding of how to correctly think about statistical problems.

I assume you know a bit about probability (means, variance, etc.) and some probability distributions (Gaussian, Poisson, binomial).

Probability Distributions

Gaussian: $$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\bar{x} = \mu$$

$$\mathrm{Var}(x) = \overline{(x-\mu)^2} = \sigma^2$$

Poisson: $$P(n \mid \mu) = \frac{e^{-\mu}\mu^n}{n!}$$

$$\bar{n} = \mu$$

$$\mathrm{Var}(x) = \mu$$

Binomial $$P(n \mid N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$\bar{n} = Np$$

$$\mathrm{Var}(x) = Np(1-p)$$

Probability vs Statistics

What is the difference between probability and statistics?

In probability theory, we know the probability distribution (pdf) and predict the results of trials.

For example, if you flip a fair coin 10 times, how often do you get 7 heads?

In statistics, we take a number of trials (usually small) and try to say something about the probability distribution (often an estimate of a parameter).

For example, someone tells you they flipped a coin 10 times and got 7 heads. What is the estimate of the probability of getting heads on a single flip? Do you think the coin is fair? What if it was 70 out of 100, 700 or 1000, or whatever?

What is Random in a Measurement?

You have found the Higgs boson and measure its mass:

$$M_h = 126 \pm 3 \text{ GeV}/c^2.$$

This does NOT mean that the Higgs mass is a random variable whose mean value is 126 and rms is 3.
We often use language that seems to say this ("the Higgs mass is 126 ± 3", "there is a 68% probability that the Higgs mass is from 123 to 129", etc.).

Parameters of nature are not random variables!

This does mean: "the probability that my measured interval of 123 to 129 contains the true value of the Higgs mass is 68%."

Neyman Construction

We want to construct an interval at confidence level CL such that if we could repeat the same experiment many times, our interval would contain the true value of the parameter a fraction CL of the time.

For measurements, CL is almost always 68%, which is the area of a Gaussian within 1 standard deviation of the mean. For limits, people often use a different CL (more on this later).

Neyman construction is a frequentist method.

The probability that the intervals contains the true value is called the coverage.

Ordering Principle

There are 2 end points to an interval and a confidence level provides only 1 criterion.
We need a second criterion, known as the ordering principle.

There are many possibilities:

1. **symmetric about the parameter estimate**
2. **equal coverage on each side**
3. **for a likelihood fit, equal value of likelihood at each point**
4. **etc.**

For a frequentist, the important concept is the coverage, and any reasonable ordering principle is OK.

Bayes Theorem

Let x and y be random variables.

P(x) dx = probability x is between x and x + dx.
P(x|y) = conditional probability of x, given a value of y.
Same for P(y) and P(y|x).

Bayes theorem: $$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

Bayes theorem is a perfectly correct theorem in probability theory.

Bayesian Statistics

Let x be a measurement (or set of measurements), such as the Higgs mass.

Let μ be a parameter (or set of parameters) the measurement probability depends on, such as the true Higgs mass (for example, suppose the measured value is Gaussian distributed about the true mass).

The probability of measuring x is P(x|μ). A Bayesian puts this into Bayes theorem giving

$$P(\mu \mid x) = \frac{P(x \mid \mu)P(\mu)}{P(x)} = \frac{P(x \mid \mu)P(\mu)}{\int_0 P(x \mid \mu)P(\mu)d\mu}$$

P(μ|x) is known as the posterior probability density function (pdf).

This is fine, if we put in the correct probability functions.

P(μ) should be $\delta(\mu-\mu_{true})$, in which case P(μ |x) becomes $\delta(\mu-\mu_{true})$.

Bayesian Statistics II

Of course, we don't know μ_{true} and hence don't know P(μ).

A Bayesian puts in some function he likes, calls it the prior [often symbolized as p(μ)], and proceeds.

Then, the best estimate of μ is the one that maximizes P(μ|x) and the CL interval is such that

$$\int_{\mu_L}^{\mu_U} P(\mu \mid x)\, d\mu = CL$$

Bayesian Statistics III

For example, suppose we have one measurement of the Higgs mass M that is Gaussian distributed about the true mass with a width σ (which we assume we know) and assume a constant prior.

$$P(M_h \mid M) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(M - M_h)^2}{2\sigma^2}} p(M_h)}{\int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(M - M_h)^2}{2\sigma^2}} p(M_h) dM_h} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(M_h - M)^2}{2\sigma^2}}$$

This gives exactly the same best estimate and confidence interval as a frequentist analysis (not true for other priors). A constant prior can't be normalized and is known as an improper prior. However, it cancels in the formula, so Bayesians don't let this slow them down.

Bayesian Statistics IV

If pushed hard, a Bayesian will (usually) admit that $P(m|x)$ is not a probability but will call it a "degree of belief" or "betting odds."

However, they treat it like a probability and talk about it that way (if it walks like a duck, quacks like a duck, ...).

I have no idea what the mathematics of a "degree of belief" is.

In fairness to Bayesians (which is not something I feel compelled to be), they correctly point out that how we approach things in life is usually very Bayesian.

Problems with Bayes Statistics

1. **Prior dependence (answer depends on choice of prior).**

2. **Metric dependence (interval depends on whether we analyze M_h or M_h^2, for example).**

3. **Logically (philosophically?) bothersome.**

Quote from Louis Lyons

Bayesians address the question everyone is interested in by using assumptions that no one believes.

Frequentists use impeccable logic to deal with an issue of no interest to anyone.

Louis Lyons
Academic Lecture at Fermilab
August 17, 2004

Nuisance Parameters

Suppose we count the number of events n of some type and want to convert to a cross section, which depends also on things like background b, efficiency e, and integrated luminosity L.

$$s = \frac{(n - b)}{eL}$$

In general, b, e, and L will also have uncertainties, which we must take into account in our statistics. These are known as nuisance parameters (they are necessary, but not what we are primarily interested in).

A Bayesian has no problems - he or she just treats these as having some probability distribution and proceeds.

A frequentist wants a confidence interval that correctly covers for all true values of m, b, e, and L.

Mixed Statistics

Sometimes, people integrate over the nuisance parameters (a very Bayesian way to approach life) and then do frequentist statistics using the resulting distribution.

This is a mixed method.

This is not so common in parameter estimation (here we usually have a separate systematic uncertainty for uncertainties in the nuisance parameters), however, it is quite common for setting limits.

Bias and Coverage

However you define your statistical method, it is very important that you check (unless it is a known method that someone else has already checked):

1. Bias, that is, if you could repeat your experiment many times, the average of the best estimators should be the true value.

2. Coverage, that is, if you could repeat your experiment many times, your confidence interval should contain the true value 68% of the time. Under coverage is considered very bad. Over coverage is mildly bad.

Physicists are very good at inventing their own methods, which must be checked.

Simple (aka “Toy”) Monte Carlo

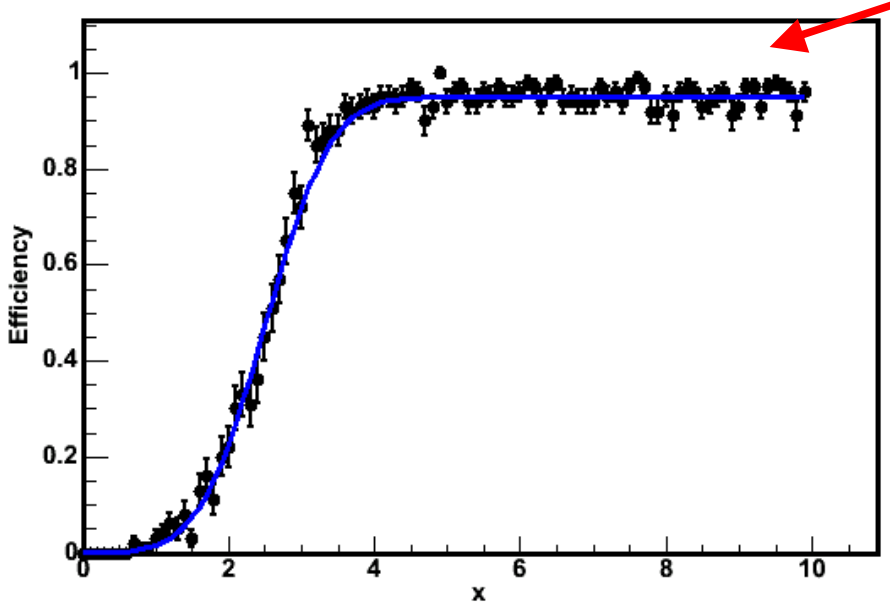
A good way to check bias and coverage is a “toy” (I prefer “simple”) Monte Carlo, in which a simple model of the statistics is tested using computer random number generators.

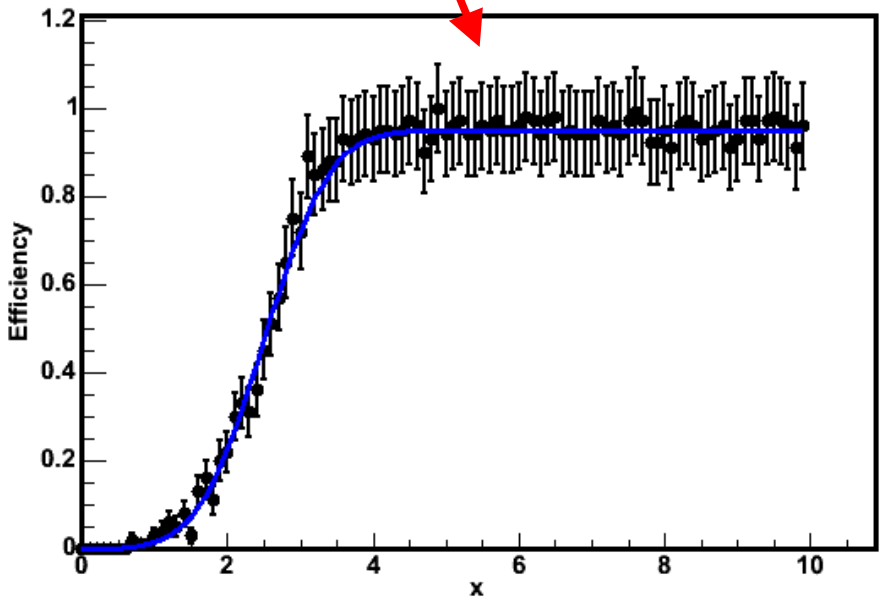
Be sure to use a good random generator, a good model of your problem, and proper ensembles.

Simple Monte Carlos are extremely power tools for checking and understanding bias, coverage, sensitivity, etc.

Binomial vs Poisson Uncertainties

You calculate efficiency in a bin by dividing number of good events n by the number of candidate events N, e = n/N.
Sometimes people will say that since they are counting the number of good events, the uncertainty is Ön, giving plots like this.
Since there is a maximum number of good events, the correct uncertainty is binomial, giving this plot.





Limits - Frequentist

Suppose we are counting some type of event (e.g., how many Higgs events we see). If μ is the expected number, then

$$P(n) = \frac{e^{-\mu}\mu^{n}}{n!}$$

If n_{obs} events are observed, we may want to set an upper limit μ_L on μ. We define the upper limit at a given confidence level (CL) as the μ such that the probability of getting the observed number of events or fewer is 1 - CL.

$$\sum_{n=0}^{n_{obs}} \frac{e^{-\mu_L}\mu_L^{n}}{n!} = 1 - CL$$

Zero Intervals

Suppose we are doing a limit, and we know that the background level is 3 events. Furthermore, suppose we observe 0 events and use CL = 90%.

$$\sum_{n=0}^{0} \frac{e^{-(m_L+b)}(m_L+b)^n}{n!} = e^{-(m_L+3)} = 1 - CL = 0.1$$

There is no non-negative m_L that satisfies this equation!

Feldman-Cousins

To solve this problem and others for parameters near physical boundaries, Feldman and Cousins proposed a new ordering principle in the Neyman construction.

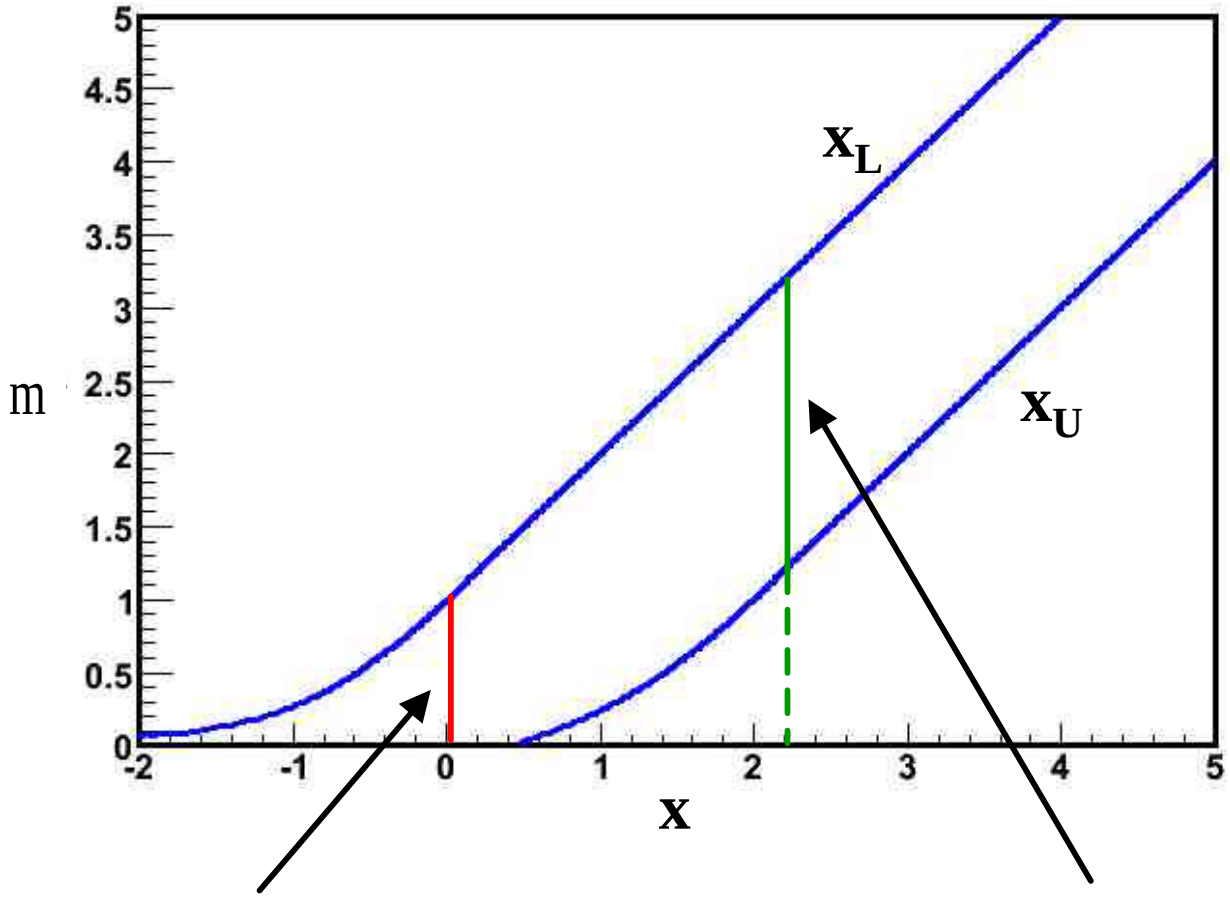
$$R(x) = \frac{P(x|\mu)}{P(x|\mu^*)}$$

Here, μ^* is the most likely value of μ for a given x. Consider a case where P is a Gaussian with unit width, μ is a parameter that cannot be negative (like m_n), and CL = 68%.

For a given μ, find the x_L and x_U such that

$$\int_{x_L}^{x_U} P(x|\mu)dx = CL \text{ and } R(x_L) = R(x_U)$$

Feldman-Cousins II



For a measured value of x, the x_U and x_L curves give the confidence interval in m.

Feldman-Cousins avoids zero intervals, covers correctly, smoothly transitions from 1-sided to 2-sided intervals, and approaches "standard" intervals for x >> 0.

For x = 0, interval is 0 to 0.99

For x = 2.2, interval is 1.2 to 3.2

Flip-flopping

From Feldman, NEPPSR IV, "many physicists like to

1. **pretend they measure 0 if x < 0**
2. **give upper limit if x < 3s**
3. **give 2-sided interval if x > 3s".**

This is known as flip-flopping. In this case, it is impossible to construct proper frequentist intervals.

Feldman-Cousins gives a smooth transition from 1-sided intervals (limits) to 2-sided intervals (measurements).

However, you have to understand that if your interval does not contain 0, this doesn't mean you have good evidence for a discovery.

Bayesian Limits

A Bayesian determines an upper limit μ_L by

$$\int_0^{\mu_L} P(\mu \mid x)\, d\mu = CL$$

For a Poisson process with no background and using a flat prior, this gives the same limit as a frequentist calculation.

Bayesian limits

1. **depend on the prior**
2. **are metric dependent**
3. **in general, do not cover correctly.**

Nuisance Parameters

Nuisance parameters usually are quite important in limits.

For example, we may look for a number of events of some type and wish to put a limit on the cross section by correcting for the background and efficiency, both of which usually have uncertainties.

Frequentists require correct coverage for any true value of the parameters, including the nuisance parameters. This usually leads to very long and intensive Monte Carlo calculations (particularly if there many nuisance parameters).

Bayesian just include the nuisance parameter distributions in their "probability" functions and then happily integrate over them.

Truncated Gaussian Efficiency

Bayesians can get undesired effects.

For example, if you

1. **are doing a counting experiment (i.e., Poisson process),**
2. **assume the efficiency e is Gaussian distributed,**
3. **truncate the Gaussian so that e ^{3} 0, and**
4. **you use a flat prior, then**

you will find that the posterior density is not normalizable.

People often put in a cutoff (either knowingly or unknowingly), but then the limit is cutoff dependent.
You can also change the flat prior or change what you assume for an efficiency distribution.

Sensitivity

In Feldman-Cousins and other methods, it is possible that fluctuations will cause an experiment with larger background to have a lower limit that a better experiment with less background.

In order that people can evaluate your experimental results properly, it is important to also give the sensitivity, which is defined as the average limit that would be set if the true parameter value is 0.

A simple Monte Carlo is a good way to calculate the sensitivity.

Recommendations

Whether you are a Frequentist, Bayesian, or mixed (Frayesian??), you should

1. **Check your statistics process for bias and coverage, particularly if the method is not standard (even if you are a Bayesian).**
2. **Remember what your measurement is actually saying.**
3. **Tell people explicitly what you did.**
4. **Quote your sensitivity.**
5. **Not undercover.**

Statistics References

Louis Lyons' Lectures:
http://www-ppd.fnal.gov/EPPOffice-w/Academic_Lectures/
Colin Gay, Likelihood methods, NEPPSR V, IV, (...?)
C. Blocker, Likelihood methods, NEPPSR IV, III, I
Gary Feldman, Statistics of Small Numbers, NEPPSR IV

PDG: http://pdg.lbl.gov
CDF Statistics:
http://www-cdf.fnal.gov/physics/statistics/ statistics_home.html
Babar: http://www.slac.stanford.edu/BFROOT/www/Statistics/

and references therein.