A Statistician's Approach to Setting Limits,

...computing intervals, and detection

David A. van Dyk

Department of Statistics, University of California, Irvine

PhyStat, January 2011

Outline

Outline

Detection Problem

Consider a simple Poisson model

$$\begin{aligned} n_B \mid (\lambda_B, r, \tau_B) &\sim \text{Poisson}(r\tau_B\lambda_B) \\ n \mid (\lambda_S, \lambda_B, \tau_S) &\sim \text{Poisson}\Big(\tau_S(\lambda_S + \lambda_B)\Big) \end{aligned}$$

where λ_B is known.

We use a standard hypothesis testing framework:

H_0 There is no source: $\lambda_S = 0$

H_A There is a source: $\lambda_S > 0$.

Detection Threshold

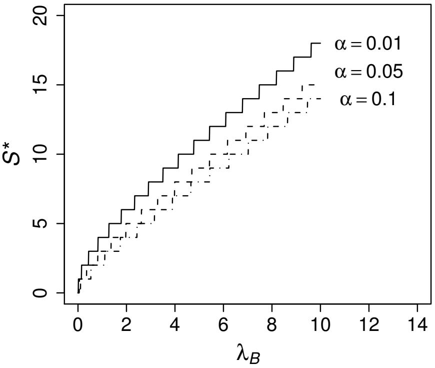
The detection threshold $n^\star$ is the smallest value such that

$$\Pr(n > n^\star | \lambda_S = 0, \lambda_B, \tau_S, \tau_B, r) \leq \alpha,$$

If $n \leq n^\star$ we conclude there is insufficient evidence to declare a source detection.

If $n > n^\star$ we conclude there is sufficient evidence to declare a source detection.

Detection Threshold



α-level detection threshold $n^\star$ as a function of the background intensity λ_B.

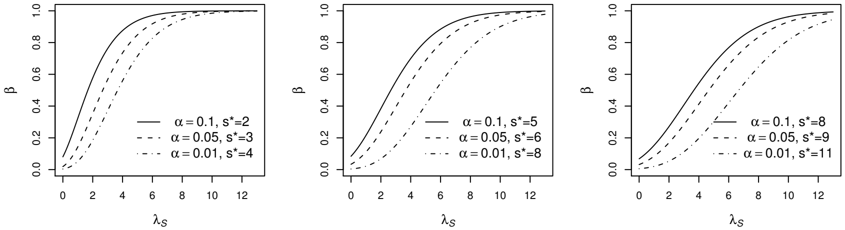
Power

The *power* of the test to detect a source as a function of its intensity is

$$\beta(\lambda_S) = \Pr(n > n^\star | \lambda_S, \lambda_B, \tau_S, \tau_B, r)\,.$$

Note $\beta(\lambda_S = 0) \leq \alpha$.

Power



Power for $\lambda_B = 1, 3, 5$ and given α

Typical Detection Prodedure

When there is a detection astronomers often

1. Report a detection
2. Report a confidence interval for λ_S

When there is not a detection astronomers often

1. Report no detection
2. Report an "Upper Limit" for λ_S

What is the difference?

Upper Limits

What is an "upper limit"?

In astronomy upper limits are inextricably bound to source detection: by an upper limit, an astronomer means

> *The maximum intensity that a source can have without having at least a probability of $\beta_{\min}$ of being detected under an α-level detection threshold.*

or conversely,

> *The smallest intensity that a source can have with at least a probability of $\beta_{\min}$ of being detected under an α-level detection threshold.*

Requires two probability calculations.

Upper Limits

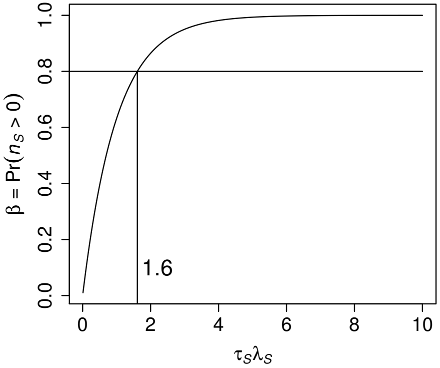
Upper Limits are analogous to sample sizes as follows:

If you don't have a detection, the sample size indicates how much you should worry.

The Upper Limit aims to directly calibrate this.

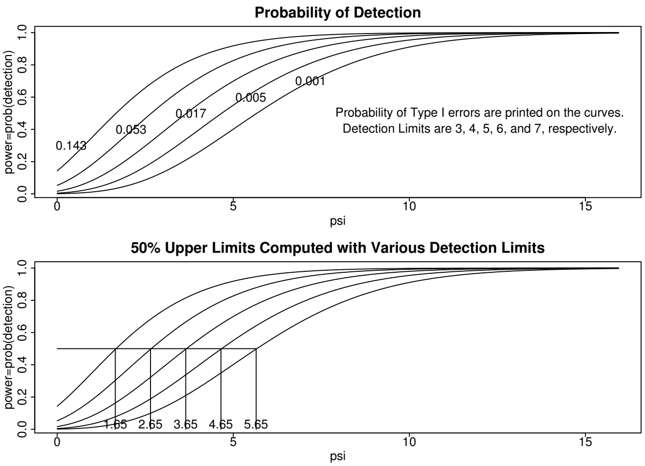
Physicists generally refer to this as the "sensitivity" of the detection.

Illustrating Upper Limits/Sensativity

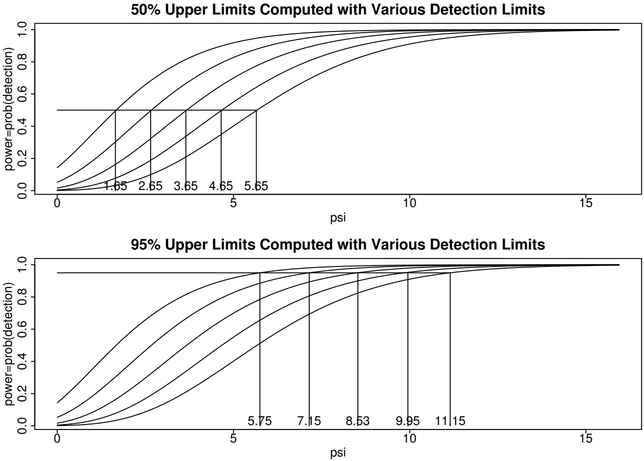


Upper limit with no background contamination.

Effect of Detection Threshold on UL/Sensativity



Effect of UL probability on UL/Sensativity



Upper Limits/Sensativity and Power

- In a typical power calculation, we would find the minimum τ_S so

$$\beta(\lambda_S) = \Pr(n > n^{\star} | \lambda_S, \lambda_B, \tau_S, \tau_B, r)$$

  achieves a given value for a given λ_S. Say 90% for $\lambda_S = 2$.
- For an upper limit we solve the same equation, but fixing τ_S and solving for λ_S.

Like power, an upper limit does not depend on the data and can be computed in advance.

Upper Bounds

The upper end point of the (one-sided) interval:

The largest plausible value of the source intensity consistent with the observed data.

This quantity is referred to as the

Upper Limit by physicists and the
Upper Bound by astronomers.

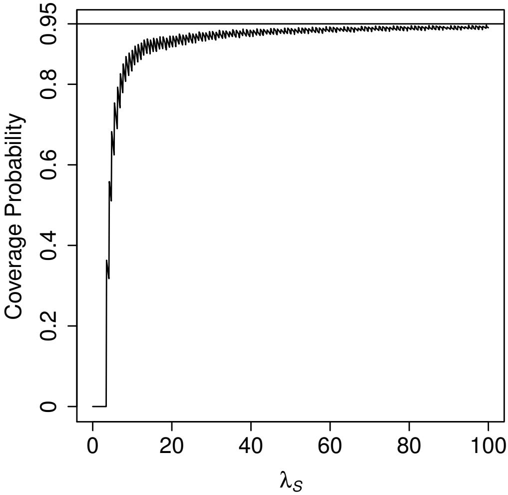
Outline

The Typical Procedure

- In the typical astronomy procedure, the confidence interval is only reported if a source is detected.
- With Power-Constrained Limits, UL is only reported if data is above a threshold. Otherwise the sensitivity is reported.
- But deciding whether to report the interval or UL *based on the data* alters the frequency properties.

Unfortunately, frequency properties depend on what you would have done, had you had a different data set.

Under Coverage



Proposed Procedure

Always report

1. Whether the source was detected.
2. A Confidence Interval for the source intensity.
    - This may be a one-sided interval taking the form of an upper limit.
3. The sensitivity, in order to quantify the strength of the experiment.

Advantage of Proposed Procedure in HEP

Corrections to standard UL

- PCL mixes a standard UL with the sensitivity.
- CL_S alters the UL for a smoothed version of PCL.

Both

- *sacrifice frequency properties and*
- *are rather difficult to interpret.*

By reporting both the UL and the Sensitivity.

- We report the largest value consistent with the data (UL)
- and the smallest value we have sensitivity to detect.

UL < Sensitivity

Question:

What does it mean when UL is less than sensitivity?

Answer:

Something other than data is constraining the intensity.

- Assumption that $\mu \geq 0$.
- Assumption about λ_B.

In any case, knowing UL and Sensitivity is more informative than knowing max(UL, sensitivity).

Concerns with Existing Intervals / Limits

- Frequentist methods can give empty intervals for λ_S.
- Frequentist methods can give very short intervals that seem to imply a very sensitive experiment.
- The upper limit may increase as *n* decreases.
- “Goldilocks effect": Frequency coverage should be above a minimum, but no more than the minimum.
- Apprehension about Bayesian methods and their priors.

Frequency Intervals

Confidence Interval:

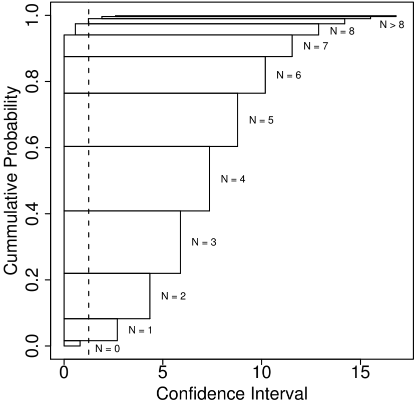
$$\{\lambda_S : n \in \mathcal{I}(\lambda_S)\},$$

where $\lambda_B = 2.9$ and

$$\Pr(n \in \mathcal{I}(\lambda_S) | \lambda_S) \geq 95\%.$$

Values of λ_S with given propensity to generate data.

Sampling Dist'n of 95% CI



The CI gives plausible values of λ_S given the data.

Short or Empty Intervals

What do they mean?

There are few plausible values of λ_S given the data.

What they do not mean?

Experimental uncertainty is small. (SE or Risk of $\hat{\lambda}_S$??)

What if intervals are repeatedly short or empty?

Short intervals should be uncommon. If they are not, blame the model, not the interval— regardless of the strength of the subjective prior belief in the model.

Pre-Data and Post-Data Probabilities

- Frequency Interval has 95% chance of covering true λ_S.
- What is the chance that $\emptyset$ contains λ_S?

There is a 95% chance that Confidence Intervals computed in this way contain the true value of λ_S.

- Frequency-based probability says nothing about a particular interval.
- Bayesian methods can quantify this type of probability.
- Precise probability statements may not be relevant statements.

Our intuition leads us to condition on the observed data.

Some thoughts on 5σ

Are we really worried about making one Type-1 error in 1.7 million results??

No. We are worried about:

- The look elsewhere effect.
- Calibration and systematic errors.
- Statistical error rates that are not well calibrated due to general model misspecification. (E.g., David Cox)

Model misspecification

is the same problem that leads to ubiquitous short or empty intervals and UL < sensitivity.

Problems with 5σ

Using 5σ is really not the answer:

- We don't know the actual effect of Systematics and LEE.
- "No distribution is valid to the 5σ tail!"
- Sampling distributions are only asymptotic approximations.
- Must calculate extreme-tail probabilities.

*We have **no** idea what the actual level is.*

5σ simply sweeps the problem under the rug.

What Should We Do?

Real solutions require real work:

- Deal with systematics, LEE, and general model misspecification directly.
- Model diagnostics and model improvement will improve statistical properties of detection, intervals, and limits.
- Hiding assumptions and ad hoc fixes do not eliminate assumptions—but makes evaluating their effect difficult.
- Bayesian methods lay their assumptions out for all to see.
- Model specification is more fundamental than the choice of Bayesian/Frequency/Other procedure.

Goal: Honest frequency error rates or a calibrated Bayesian procedure.

Outline

Standard Hypothesis Testing

Consider again the standard hypothesis testing framework:

H_0 There is no source: $\lambda_S = 0$

H_A There is a source: $\lambda_S = k > 0$.

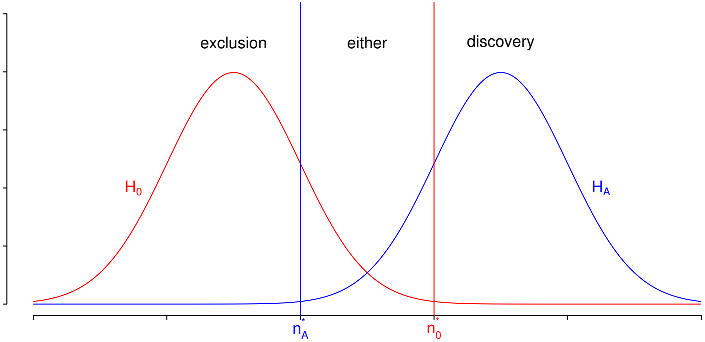
The typical (Neyman-Person) strategy:

1. Set the detection threshold to limit the probability of a false positive (Type-I error).
2. Compute power via prob of false negative (Type-II error).
3. Compute interval, e.g., by inverting the test.
4. Compute Limits from interval or via a power calculation.

Hypothesis Testing in HEP

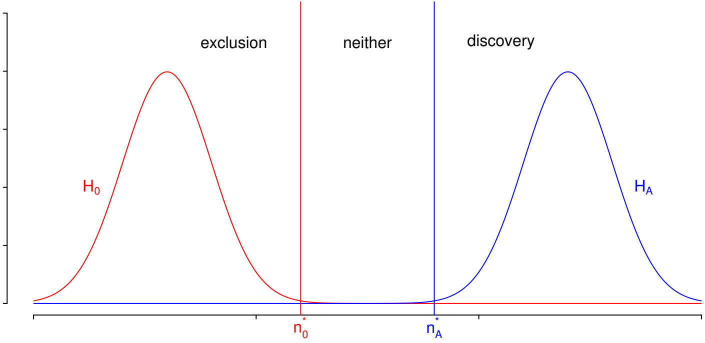
Along with standard test, conduct test interchanging H_0 and H_A:

- H_0 There is a source: $\lambda_S = k > 0$.
- H_A There is no source: $\lambda_S = 0$



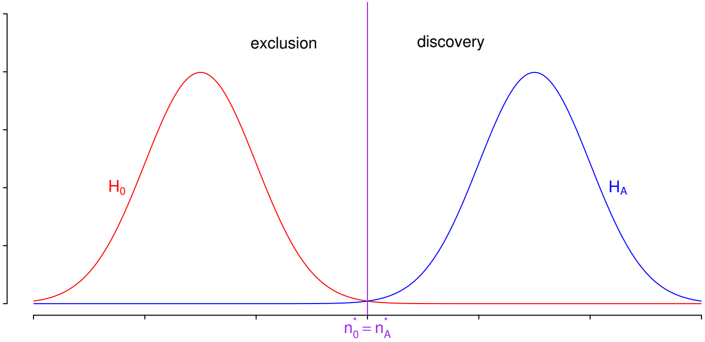
Hypothesis testing in HEP

This results in 4 possible outcomes: exclusion, discovery, no decision (either is possible), or excluding both hypotheses.



Hypothesis testing in HEP

2 or 3 are possible depending on the order of n_0^* and n_A^*.



Louis wondered if there is anything like this in the statistics literature.

HEP in Statistics Literature

- It is unusual to treat the hypotheses in a symmetric fashion.
- Typically the alternative parameter space contains the null parameter space.
- Using various values of λ_S in the null to compute the UL corresponds to inverting a test.
- This is standard, except that in HEP a different test is inverted than the test used for discovery.
- Other tail is used with 2σ rather than 5σ.
- But the formal symmetric testing seems unusual if not unique to HEP.

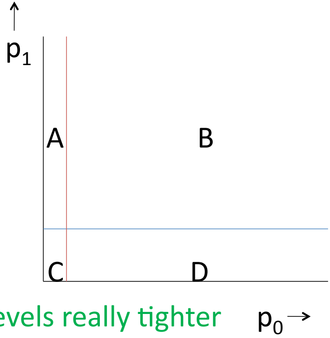
Possible outcomes

A = Reject H_0

B = Make no choice

C = ?

D = Exclude H_1



N.B. Reject/exclude levels really tighter

If H_1 true: D = false exclusion, B+D = Error of 2^{nd} kind (for H_0)

From L Lyons 2010 BIRS talk.

Ultimate Goals

Recall concerns about standard methods:

1. Intervals may be short or empty.
2. Detailed observations about the character of procedures under certain circumstances.
3. Desire for precise frequency coverage.
4. Apprehension about Bayesian priors.

When compared with the ultimate goals:

1. Detection if and only if source exists.
2. Intervals that contain actual intensities.
3. Upper Limits that bound the actual intensities.

concerns appear superficial.

Costs of Errors in HEP Detection

The “Loss” Function:

	Decision			
Truth	H_0	H_A	either	neither
H_0	0	C_{01}	C_{0e}	C_{0n}
H_A	C_{10}	0	C_{1e}	C_{1n}

- C_{01} is the cost of a false positive.
- The other costs are likely significantly smaller than C_{01}.
- We might add another row for "Truth = Neither".

A Simplified Cost Structure

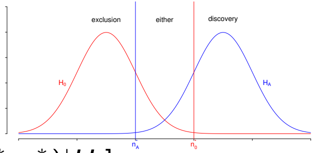
The “Loss” Function:

	Decision			
Truth	H_0	H_A	either	neither
H_0	0	C	c	c
H_A	c	0	c	c

Here we assume

1. The costs of all errors except a false positive are more-or-less equal.
2. $C \gg c > 0$
3. $C + c = 1$ (This is just a choice of scale.)

Minimize Expected Loss



$$\begin{aligned}
\text{Risk}(n_0^*, n_A^* | H_0) &= \text{E}(\text{Loss} | H_0) \\
&= C \Pr[n > \max(n_0^*, n_A^*) | H_0] \\
&+ c \Big\{ \Pr[n_0^* < n < n_A^* | H_0] + \Pr[n_A^* < n < n_0^* | H_0] \Big\}
\end{aligned}$$

$$\begin{aligned}
\text{Risk}(n_0^*, n_A^* | H_1) &= \text{E}(\text{Loss} | H_1) \\
&= c \Pr[n > \min(n_0^*, n_A^*) | H_1] \\
&+ c \Big\{ \Pr[n_0^* < n < n_A^* | H_1] + \Pr[n_A^* < n < n_0^* | H_1] \Big\}
\end{aligned}$$

We want to find thresholds that minimize Risk.

Bayes Risk: Averaging Over the Truth

$$\text{Bayes Risk}(n_0^*, n_A^* | \pi) = (1-\pi)\,\text{Risk}(n_0^*, n_A^* | H_0) + \pi\,\text{Risk}(n_0^*, n_A^* | H_A)$$

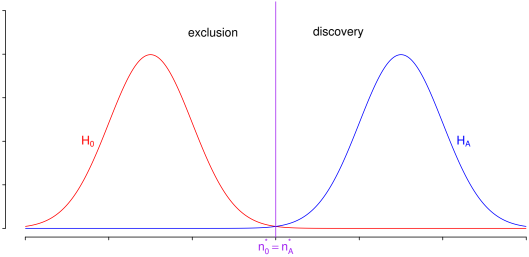
- Here π is the prior probability of H_A.
- Bayes Risk in minimized either when

$$C = \frac{(1-\pi)f_0(n_0^*) + \pi f_A(n_0^*)}{2(1-\pi)f_0(n_0^*) + \pi f_A(n_0^*)} = \frac{(1-\pi)f_0(n_A^*) + \pi f_A(n_A^*)}{2(1-\pi)f_0(n_A^*) + \pi f_A(n_A^*)}$$

or at a point Bayes Risk is not differentiable: $n_0^* = n_A^*$

Back to Basics

We minimize the Bayes Risk, by setting $n_0^* = n_A^*$.



1. This corresponds to the standard hypothesis setting.
2. The optimal value of $n_0^* = n_A^*$ is determined by C and c.

But.... with more complicated Losses...

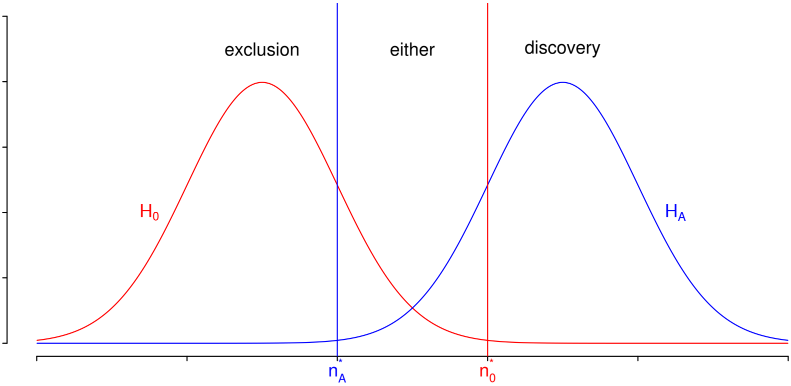
The situation may be different with more complicated loss.

E.g., the losses associated with false exclusion, either, and neither may be different for H_0 and/or H_A.

The "Loss" Function:

	Decision			
Truth	H_0	H_A	either	neither
H_0	0	C	c	c
H_A	c	0	c	c

Understanding the Result



Holding n_0^* fixed, increasing n_A^* decreases the expected loss.

Empirical Results

Consider a Normal Model:

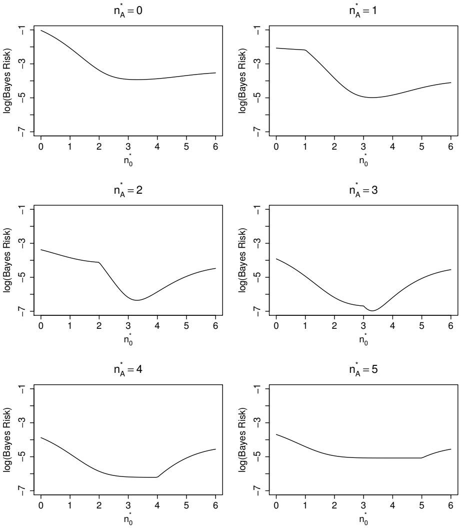
Under H_0: $n \sim \mathrm{N}(0, 1)$

Under H_A: $n \sim \mathrm{N}(5, 1)$

1. We reject H_0 if $n > n_0^*$.
2. We reject H_A if $n < n_A^*$.

Bayes Risk is computed with

- $\pi = 0.25$ and
- $C = 0.95$.



Frequency Properties

Hypothesis Testing

- Decision theoretic tests do not aim at control the probability of Type I error.
- Instead they aim to control the overall expected loss.
- $C \gg c \rightarrow$ Type I errors far less frequent that Type II errors.

Intervals can be constructed by inverting a test

- Set of values of λ_0 such that we cannot reject $H_0 : \lambda_S = \lambda_0$.
- Pr(Type I error) $\leq \alpha \rightarrow$ coverage $\geq 1 - \alpha$.
- If we don't control Type I error, coverage may fluxuate.

A Better Strategy

Derive Loss functions that quantify desired properties of interval and limits.

Intervals: $\text{Loss} = b \times \text{length}(\text{interval}) - I_{\text{interval}}(\theta)$.

Limits: $\text{Loss} = b \times \text{limit} - I\{\theta < \text{limit}\}$.

Compute Risk & Bayes Risk and minimize over interval or limit.

These are just examples. Some high-energy physicists find intervals that are too short undesirable.

Summary

Parting Words

- Focus on model diagnostics and model improvement.
- View prior distributions as a way to illuminate assumptions, not as a source of assumptions.
- Focus on ultimate scientific goals, not superficial properties of procedures.