Searching for Invisibles

Statistical Lecture

Giovanni Punzi
University of Pisa

Invisibles School '17 - 10/6/17

What is a 'search' ?

- Most physicists immediately understand the notion of “performing a search for a new signal”
- In statistical language, a “search” can be defined as the combination of two base statistical concepts:

  Hypothesis Test + Interval Estimation(limit setting)
- That is, one performs 2 (separate) steps:
  - Test the data for presence of a signal
    1) If significant deviation from “null hypothesis”
       - Claim discovery (and possibly measure properties → Estimation)
    2) Otherwise SET LIMITS on new signal parameters

Optimization of a 'search'

- Obviously, when performing a search, you will want to have the best possible *sensitivity*
- Part of this is getting the best possible *detectors*, and collecting *largest samples of quality data*
- Another part is using the *best possible analysis* procedure → 'optimal search procedure'

This is the main topic of the present lecture

- The two-pronged nature of “searching” easily leads to a dilemma “optimize for discovery, or for limit setting ?”
- We will discuss this problem in some detail, and how it can be solved in a general way.
- I will start by reminding you of the base concepts of Hypothesis Test and Interval Estimation

Hypothesis Testing

What is Hypothesis Testing

- I have several possible <u>alternative</u> hypotheses H_i for the true state of Nature. Each implies a different *pdf* for data xBX: $p_i(x;\mu,\nu\ldots)$
- I observe data *x*. Now I want to infer <u>which of the H_i is true</u>.
  - p_i is 'simple' if NO parameters, otherwise 'composite'
  - the p_i do not need to share the same parameters, but it is also possible that $p_i(x;\mu)=p(x;\mu,\nu_i)$. Even then, **testing≠estimation**
- On widely general grounds, a test T is any function $T(x) : X \rightarrow \{H_i\}$
- Usually one of the H_i is 'default': the 'null hypothesis' H_0
  - Testing is asymmetrical: H_0 is the *accepted conclusion* in the lack of evidence to the contrary. Reflects the base scientific concept of “falsifying” a theory.
  - Natural approach for a search for deviations from SM ($=H_0$)

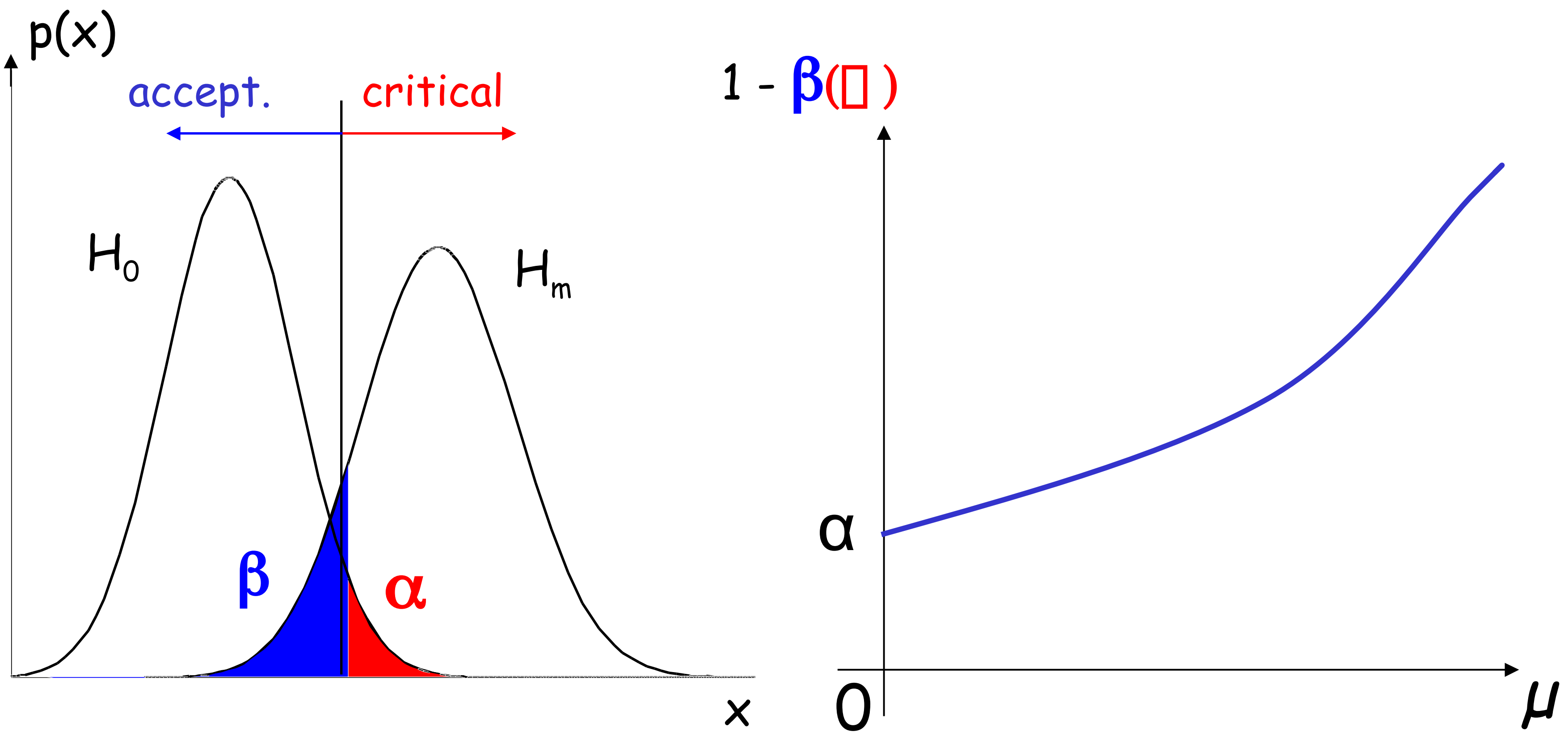
What is Hypothesis Testing

Definitions:

- If $T(x) \neq H_0$ the result is 'significant' otherwise 'not significant'
- Def: **critical region** $\{x:T(x) \neq H_0\}$ **acceptance region**: $\{x:T(x) = H_0\}$
  - α : probability of 'Type I error' : $\sup_\nu p_0(T(x)\neq H_0;\nu)$
  - β(i): probability of 'Type II error' : $p_i(T(x)=H_0)$
- α is also called **size** of the test, and is FIXED **before** the test
  - Desirable *small,* often 0.05 or 0.01. Physicists often use Gaussian tail probabilities for 3σ or 5σ
- β(i) is calculated after fixing α, and it is desirable small.
- Def: **power** of test $pow_T(i) = 1 - \beta(i)$ → ***you want it to be maximum***
- But: power an unambiguous criterion for choosing the best test only in case of just 2 simple hypotheses.

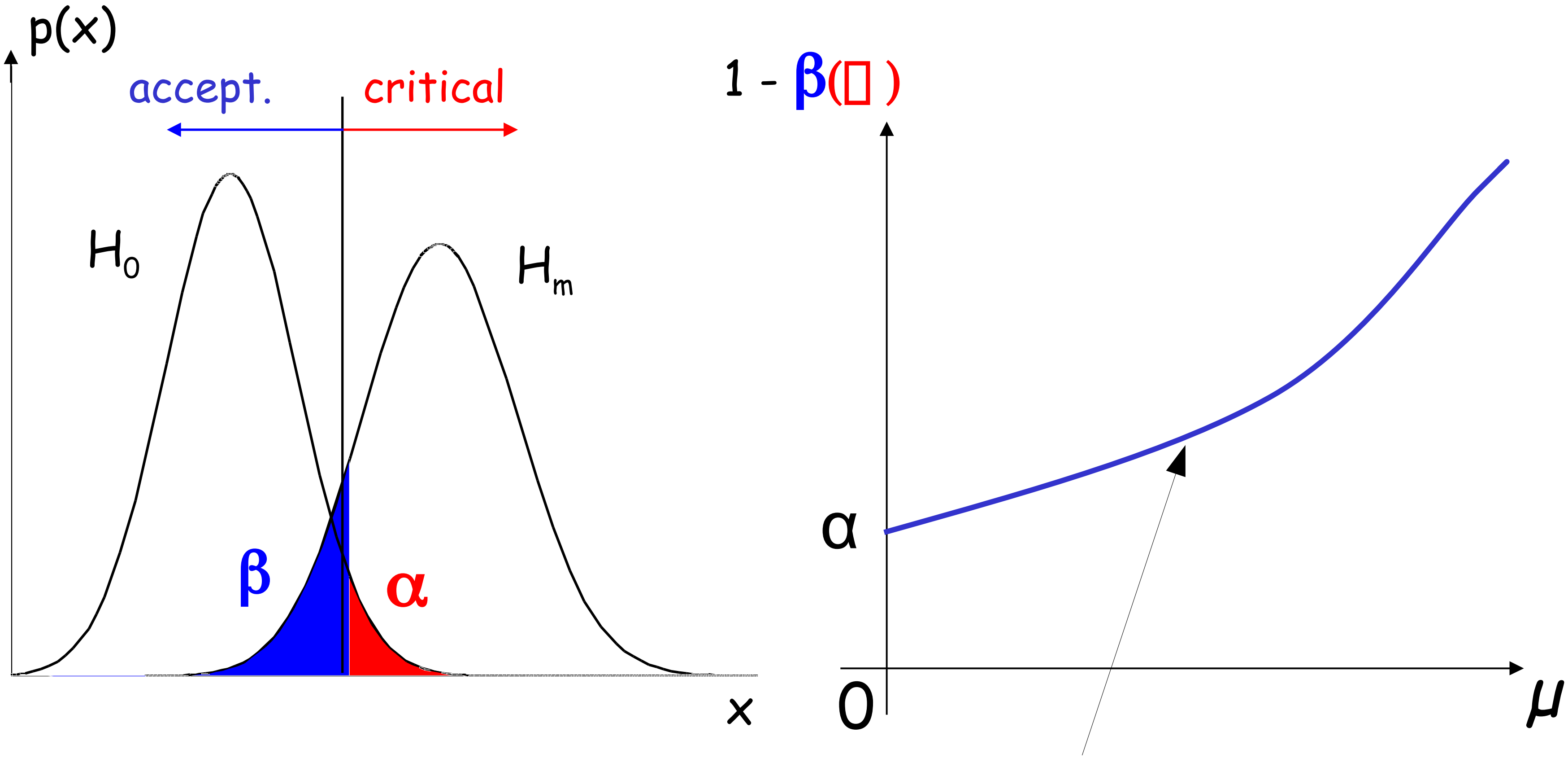
Example sketch in 1-D

- Consider a case where H_i are distinguished by the value of a single real parameter μ



- But keep in mind most real cases are N-D
  - There are differences that may confuse you

Example sketch in 1-D



- Let's now discuss how to OPTIMIZE this

What to ask from a H-Test

- **UNBIASEDNESS:** $\mathrm{pow}_T(\mu) > \alpha$ ´Eµ (very desirable)
- **CONSISTENCY:** $\lim_{N\to\infty} \mathrm{pow}_T(\mu) = 1$ ´Eµ (desirable)
  - NB: Consistency → asymptotic Unbiasedness.

- **Maximum power (MP):** $\mathrm{pow}_T(\mu) > \mathrm{pow}_{T'}(\mu)$ ´ET' (at fixed µ)
- **Local Maximum Power (LMP):** $\mathrm{pow}_T(\mu) > \mathrm{pow}_{T'}(\mu)$ ´ET' if $\mu \to \mu_0$
- **Uniform Maximum Power (UMP):** $\mathrm{pow}_T(\mu) > \mathrm{pow}_{T'}(\mu)$ ´ET'´Eµ
  - **If** a UMP test **exists**, it is obviously the method of choice.
- “Simplicity”: can often have T(x) = T(t(x)); t(x) is called **test statistic.** It is convenient to use tests based on a statistic with simple properties]

The Neyman-Pearson Lemma

- If H_0 and H_1 are simple hypotheses, the **MP test exists:** it is based on the Likelihood-Ratio statistic:

  $T(x) = H_1$ ´Ex: $p_1(x)/p_0(x) > c_\alpha$

  where c_α is determined by the chosen size of the test:

$$\int_{T(x)=H_1} p_0(x)dx = \alpha$$

- This simple result is of great importance
- You can't beat N-P: any possible sophisticated MVA analysis you can do, will only be successful in as much as it approximates the N-P statistics

Testing of nested families of *pdf*

- Suppose there is a common pdf underlying our hypotheses p(x;μ,ν), and that our test can be put in the form H_0: $\nu=\nu_0$ against H_ν: $\nu\neq\nu_0$

  then a helpful test statistic is provided by the following, slightly generalized form of Wilk's theorem, stating that under the usual regularity conditions, the statistics:

  $$\lambda(x,\nu) = 2 \log [(\sup_{\mu,\nu} p(x;\mu,\nu))/p(x;\mu,\nu_0)]$$

  has asymptotically the chisquare distribution with dim(ν) degrees of freedom. In addition, E[λ] is larger for $\nu\neq\nu_0$ than for $\nu=\nu_0$ (the distribution is actually known)

- This form of Likelihood-ratio test is widely used for its simplicity – it is important to remember that the result is only asymptotically valid, and it is NOT necessarily **UMP !**

Special case: exponential family

- Theorem: consider a pdf of a 1D parameter μ, having the form:

$$p(\mathbf{x};\mu) = \Pi_i F(x_i)G(\mu)\exp[A(x_i)B(\mu)]$$

where $B(\mu)$ is strictly monotonic.

Then a **UMP** test exists for H_0: $\mu=\mu_0$ against H_μ: $\mu>\mu_0$, based on:

$$\sum_i A(x_i) > c_\alpha$$

- NB: no 2-sided ($\mu\neq\mu_0$) UMP test exists

Pitfalls of some common recipes

Consider a classical “counting experiment”:

- H_0: Pois(B,n) ; H_μ: Pois(B+S,n).
- Suppose expected 'Background' and 'Signal' depend on some parameters **t** under the experimenter's control: B=B(**t**) , S=S(**t**,**μ**)

- How do you choose the optimal ***t*** for conducting the experiment ? Let's examine some popular methods.

1) Maximize the “significance” **S(t)/√B(t)**

1. It is NOT a significance !
   - “significance” is a post-experiment quantity: (Obs – B)/√B(t)
   - Might be argued to be ~average expected significance
2. You want high probability of discovery, NOT high average significance
3. Assumes Gaussian approximation → not good for low statistics
4. Diverges for B→0 (B=0.0001 with S=0.01 considered 'good sensitivity')

Pitfalls of some common recipes

Consider a classical “counting experiment”:

- H_0: Pois(B,n) ; H_μ: Pois(B+S,n).
- Suppose expected 'Background' and 'Signal' depend on some parameters **t** under the experimenter's control: B=B(**t**) , S=S(**t**,**μ**)

- How do you choose the optimal ***t*** for conducting the experiment ? Let's examine some popular methods.

2) Maximize the “significance” **S(t)/√(S(t)+B(t))**

1. It is NOT a significance ! Not even 'average expected significance'
2. It is actually the inverse of expected relative uncertainty on S
3. This is good if you want to measure a BR with precision – not what you are looking for in a search !
4. Does not diverge, but still assumes Gaussian approximation

Pitfalls of some common recipes

Consider a classical “counting experiment”:

- H_0: Pois(B,n) ; H_μ: Pois(B+S,n).
- Suppose expected 'Background' and 'Signal' depend on some parameters **t** under the experimenter's control: B=B(**t**) , S=S(**t**,**μ**)

- How do you choose the optimal ***t*** for conducting the experiment ? Let's examine some popular methods.

3) Maximize “signal-to-noise” **S(t)/B(t)**

Considered intuitive by some, but NO REASON for being a good idea

Summary of H-testing

- Beware of those “intuitive” formulas: they are almost never what you need.
  Make use of the concept of *power* instead.
- Use UMP when it exists, LMP when meaningful
- Optimal choice is *undefined* in most cases; that is, it requires more knowledge about the searched signals that you typically have
- **Anyhow:** optimizing H-test does not tell you anything about the tightness of the limits you will be able to set on the phenomena being searched → this will be our next topic

Plan B:
Setting Limits

“Interval Estimation”: what is it ?

- A more correct expression would actually be “region estimation” or “set estimation” (to allow for multi-D)
- Sometimes also referred to as “limit setting” (another expression biased towards 1-D problems only)
- A “region estimator” **f(x)** for a parameter **μ** ∈ **A** is any function of the data **f: X** → عر **(A)** that may be considered a “useful estimate” of the region where the true value of **μ** might belong.
- Defining what are “good properties” for region estimators is more complicated with respect to point estimators
- I will only deal with *frequentist* interval estimation in this lecture
  - H-testing is frequentist (its Bayesian counterpart is more involved)
  - I am not aware of the existence of a unified optimization method that makes use of Bayesian intervals

Interval Estimation ≠ Point Estimation

- Suppose I have a point estimate $e(x) = a \pm \sigma_a$
  - isn't this an 'interval' already ? $[a - \sigma_a , a+\sigma_a]$
  - The answer is **NO !**
- The quantity σ_a is a statement about the variability of the *statistic* e(x) (= central value 'a') in repeated experiments.
- This gives *no special meaning* to the values $(a - \sigma_a, a+\sigma_a)$
- Remember also that $\sigma_a=\sigma_a(\mu)$ - depends on *true value* μ
  - Usual (silent) convention is to quote $\sigma_a=\sigma_a(a)$
  - Can be a trouble if $\sigma_a(a)$ is very different from $\sigma_a(a\pm\sigma_a)$

Interval Estimation ≠ Point Estimation

- Example: Poisson *pdf:* $p(n;\mu)= e^{-\mu}\,\mu^{n}/n!$
  - MLE estimate of $\mu = n$ (unbiased)
  - $\mathrm{Var}(n) = \mu$
- If I measure n=1 → estimate $\mu = 1 \pm 1$
  - Does not mean that $\mu \le 0$ has a 33% probability...
- If I measure n=0 → estimate $\mu = 0 \pm 0$
  - Hum ? ….
- This is a common problem with histogram error bars
- AND it is not due to the pdf being “non-gaussian” !

Reminder: **Confidence Level**

$$CL(f) \equiv \inf_{\mu\in^{\circ}} \int_{x:\mu\in f(x)} p(x;\mu)$$

- In words: *any algorithm providing correct bounds (at least) a fraction CL of the times, <u>independently</u> of the actual value of μ - is said to have confidence level CL*
- CL is a property of the *algorithm,* not of individual region
- Does NOT comply with the Likelihood Principle
- It is a statement about probability of f(x)I μ – not μBf(x)
- Still requires some criteria for choosing the region - commonly from some *ordering principle*

Coverage

- One can also write:

$$CL(f) \equiv \inf_{\mu\in^{\circ}} C(\mu)$$

where:

$$C(\mu) = \int_{x:\mu\in f(x)} p(x;\mu)$$

- C(μ) is the **coverage** of the algorithm **f** at point μ. It can be informative to plot coverage as function of μ.
- Where C(μ) > CL the algorithm is said to ***overcover*** (***undercovering*** is a failure to attain the required CL)
- Ideally, it would be *optimal* to have constant C(μ) = CL
- This may not be possible in practice – due to discretization, for instance)

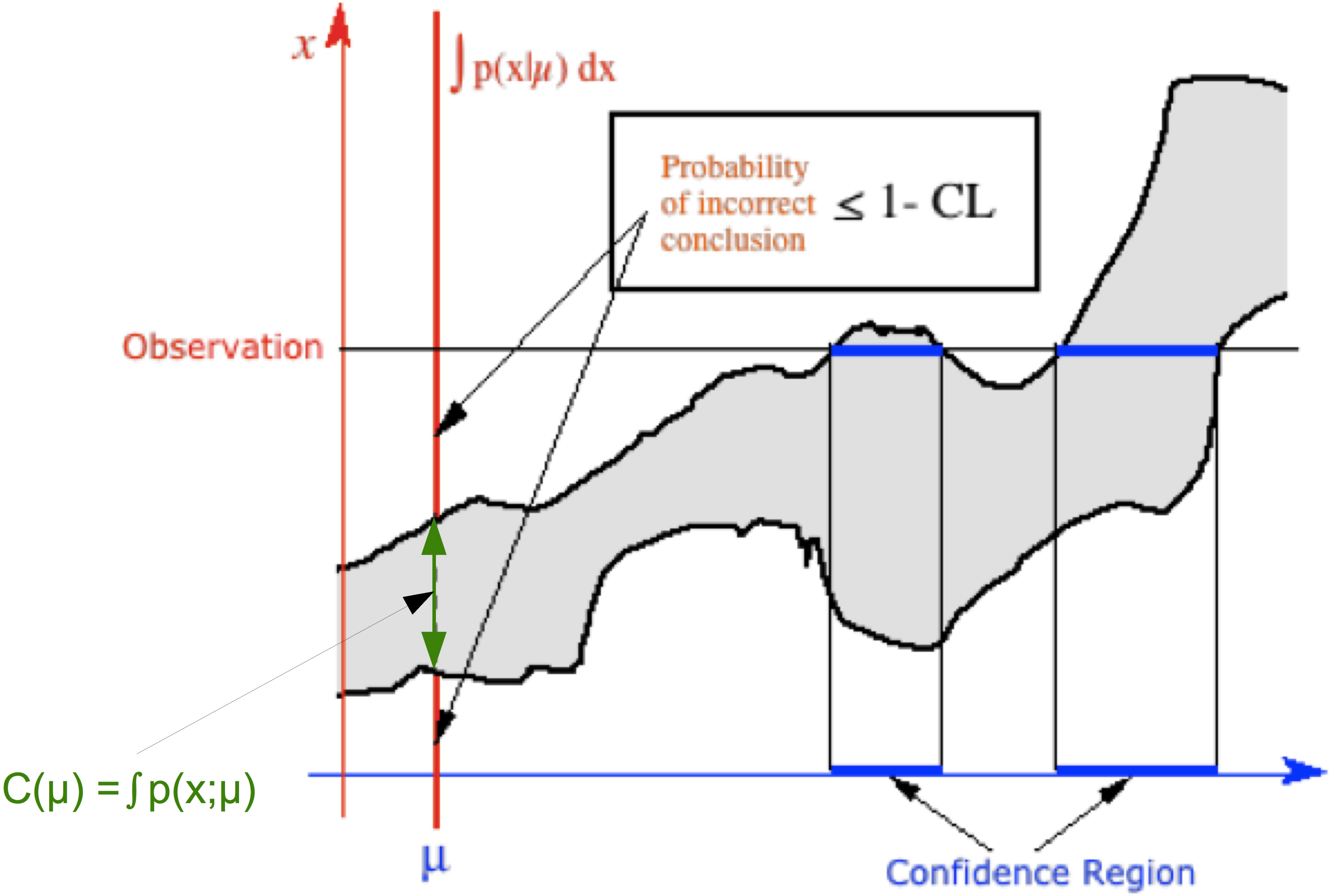
Elementary example of what 'confidence' means

- Two types of bags: A, B
- A contains 90% white balls, 10% black
- B contains 10% white balls, 90% black
  - Extract a ball, and classify A or B (make two piles:OA, OB)
  - Probability of wrong assignment of B-classified bag

    p(A|OB) = p(A,OB)/p(OB) = pA*f/(pA*f+(1-f)*(1-pA))

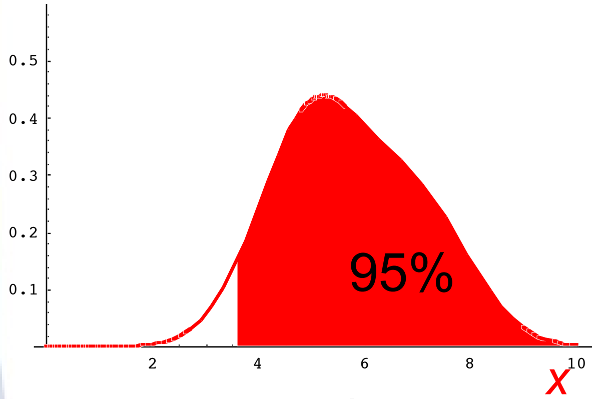
    = f/(f+(1-f)/pA-(1-f)) = f/(2f-1+(1-f)/pA)
  - p(A|OB) spans the full [0,1] range
- But, p(any error) = 10% independent of pA !
  - This is our *confidence* in the classification procedure

Construction of a **Confidence Band**

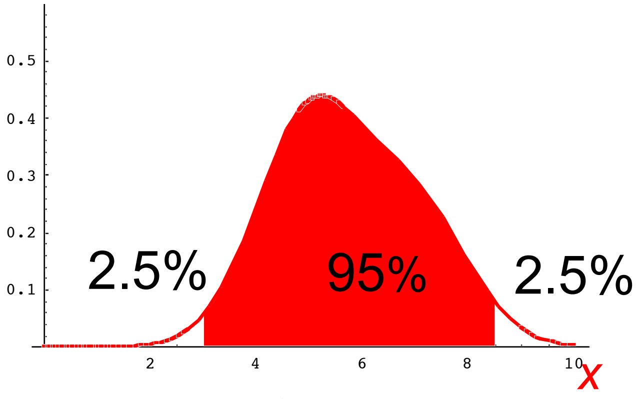


Ordering Algorithms

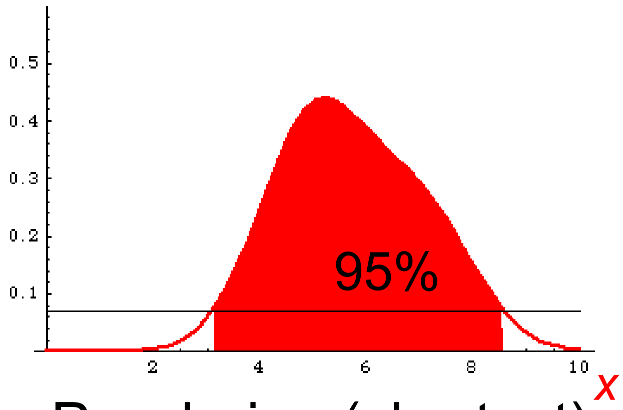
- Ample freedom in choosing the CB asks for some criteria
- Typically, an “ordering function” o(x) is used, so that the band is built by requiring: $\int_{o(x)>c} p(x;\mu) \geq CL$
- A different method is used for “central” limits
- Low/high ($o(x)=\pm x$), central, or P-ordering ($o(x)=p(x;\mu)$), have been the only methods in use for quite some time



Low/High

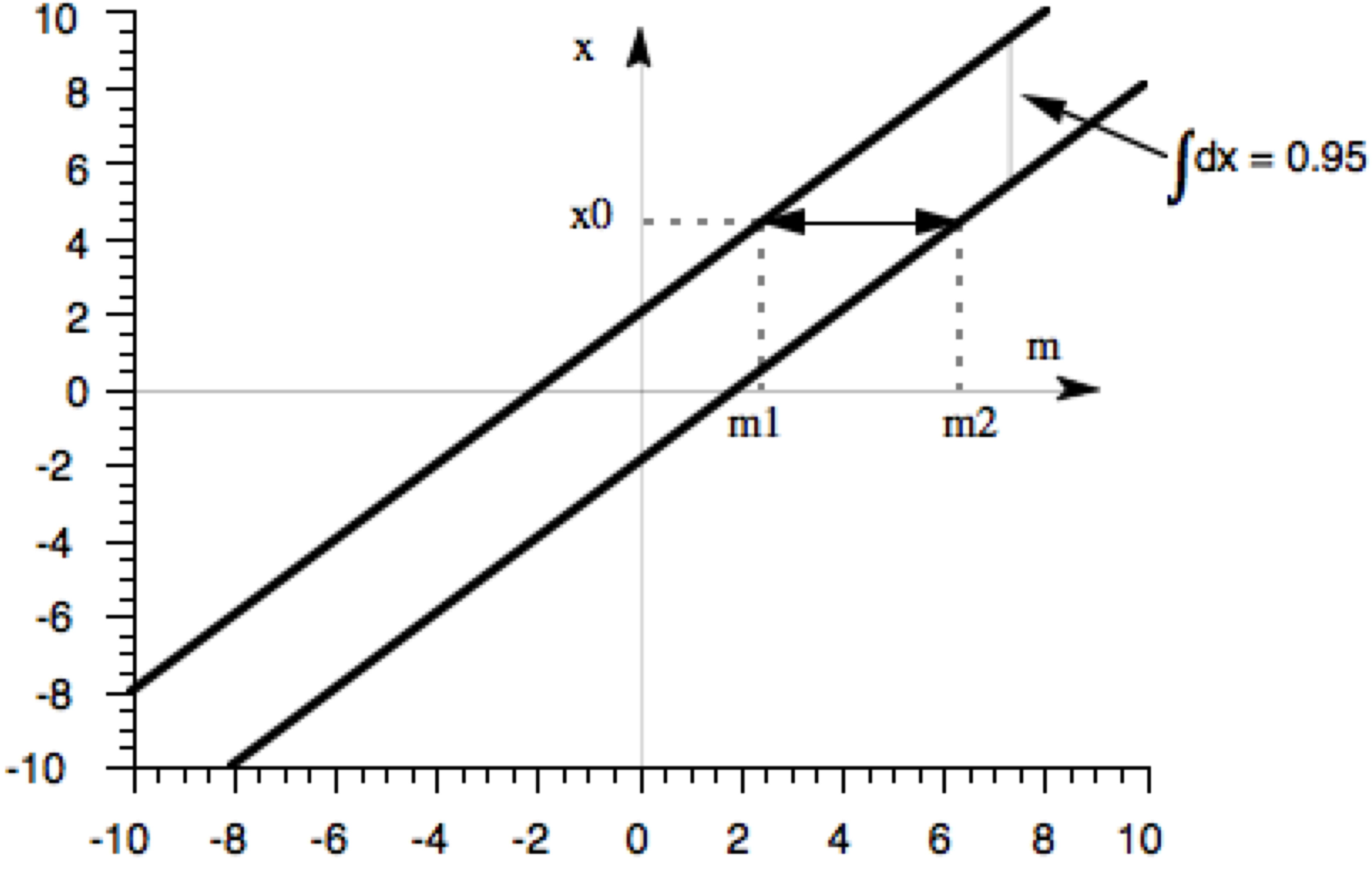


“Central”

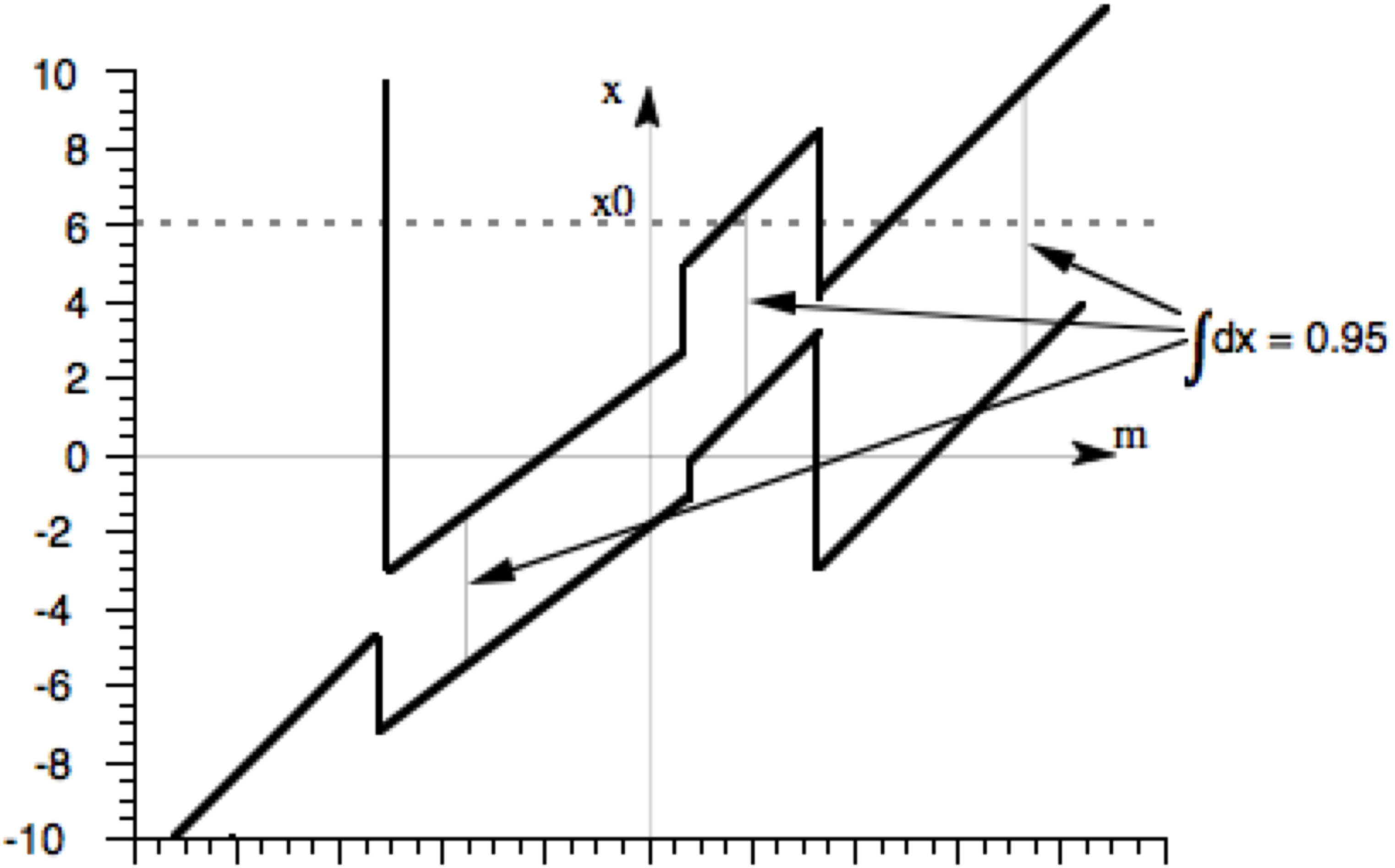


P-ordering (shortest)

Confidence Band example:
Gaussian central



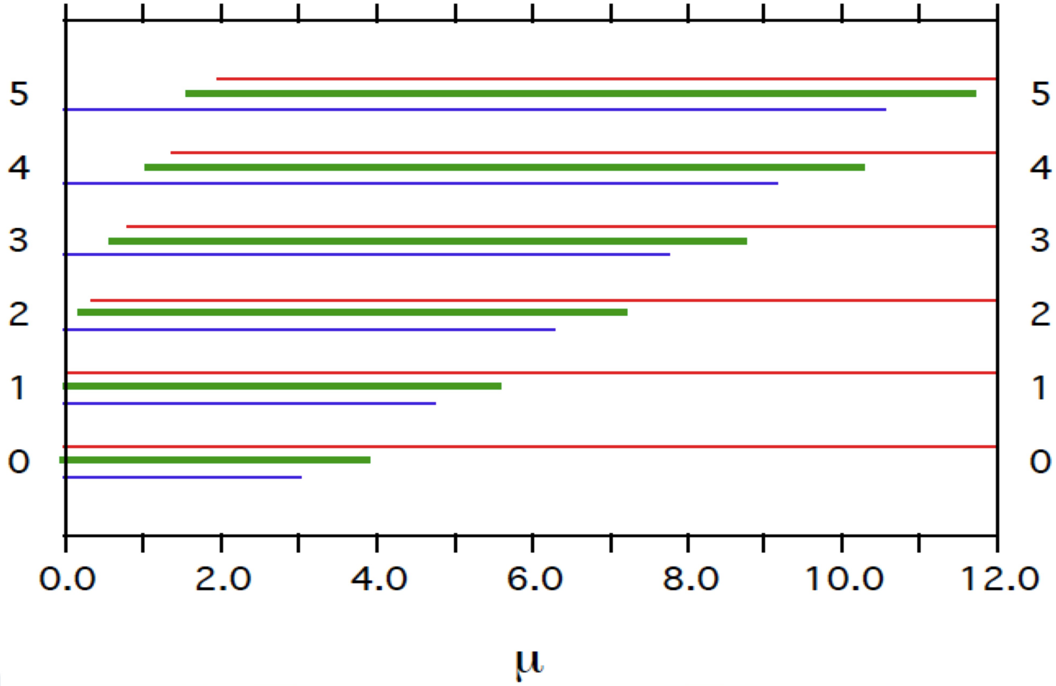
Confidence Band example:
Gaussian “mixed choices”



The “flip-flopping issue”

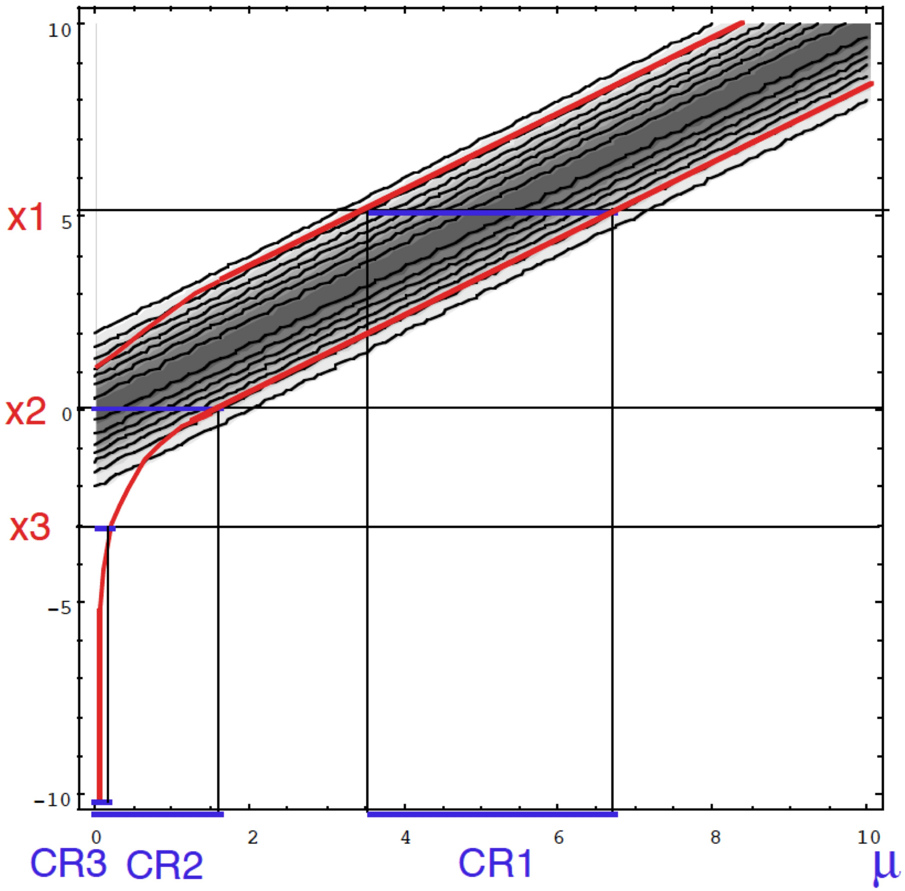
- I might well want upper limits only for small N
- I might be tempted to switch method at some N (“flip-flopping”)
- Flipflopping **undercovers** This breaks the CL ! **Don't** do it !
- Need to redesign band entirely and coherently

N	max	min	central
0	2.99572	0	0 - 3.68888
1	4.74385	0.051293	0.02532 - 5.57164
2	6.29579	0.355362	0.24221 - 7.22469
3	7.75365	0.817691	0.61867 - 8.76727
4	9.15352	1.36632	1.08987 - 10.2416
5	10.5130	1.97016	1.62349 - 11.6683
6	11.8424	2.61302	2.20189 - 13.0595
7	13.1481	3.28532	2.81436 - 14.4227
8	14.4346	3.98082	3.45385 - 15.7632
9	15.7052	4.69523	4.11537 - 17.0848



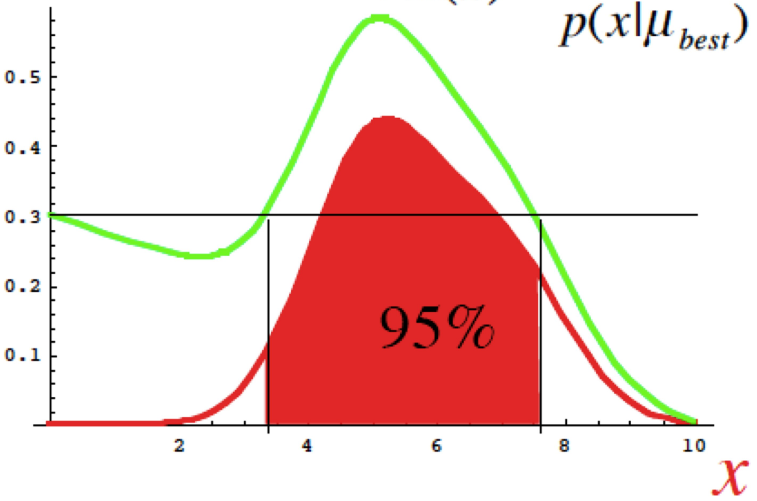
Feldman-Cousins "Unified approach"

[Phys Rev D 57,3873 (1998)]



- Use a different ordering algorithm: *Likelihood-Ratio ordering.* NB: depends on the pdf for other values of μ

$$LR(x) = \frac{p(x|\mu)}{p(x|\mu_{best})}$$



- Removes unpleasant empty intervals and avoids flip-flopping
- Invariant for change of *observable* (not mentioned in paper!)
- Current most popular solution (although still not perfect!)

Optimization of Limit-setting ?

Consider again the classical “counting experiment”:

- H_0: Pois(B,n) ; H_μ: Pois(B+S,n).
- Suppose expected 'Background' and 'Signal' depend on some parameters **t** under the experimenter's control: B=B(**t**) , S=S(**t**,**μ**)

- How do you choose the optimal ***t*** for conducting the experiment , *when the desired objective is to obtain the strongest limits ?*

1. Minimize *average expected* upper limit on S
   - problem: not invariant for transformation of observable

2. Minimize *median expected* upper limit on S
   - invariant; however, only esists for 1-D problems

Unfortunately, unlike H-testing no general concept of “power” exists

ASIDE

Feldman-Cousins on “sensitivity”

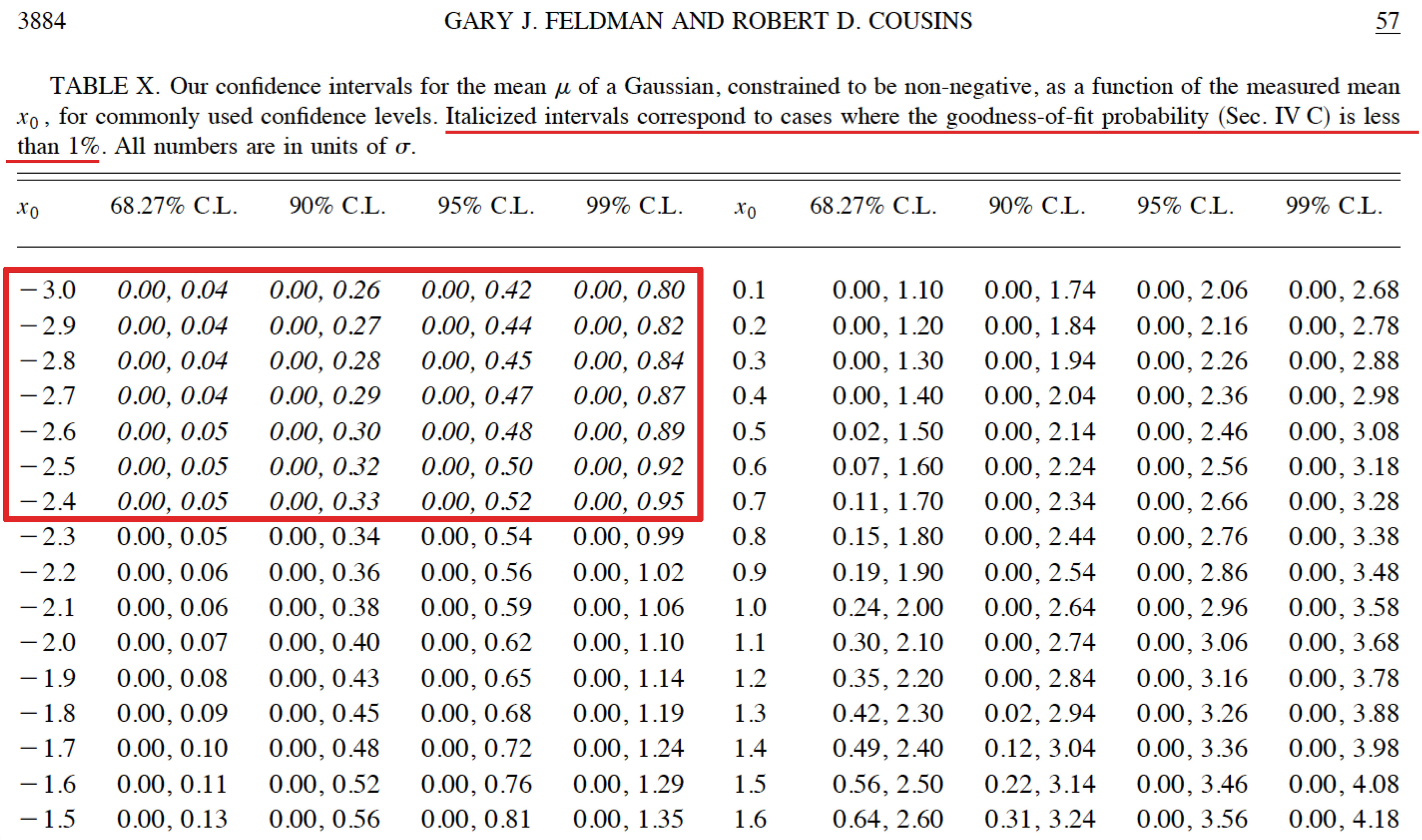
Quoting from the paper:

“Our suggestion [...] is that in cases in which the measurement is less than the estimated background, the experiment reports both the upper limit and the ‘‘sensitivity’’ of the experiment, where the ‘‘sensitivity’’ is defined as the average upper limit that would be obtained by an ensemble of experiments with the expected background and no true signal. […] we suggest that the sensitivity curve be displayed as well as the upper limit”

- This important recommendation is rarely followed ! 💣
  - NB: unfortunately formulated for 1-D problems only
- One should always check whether the results of the experiment are particularly “unlikely” - in which case the limits will be suspicious

ASIDE

Feldman-Cousins: Gaussian case



3884 GARY J. FELDMAN AND ROBERT D. COUSINS 57

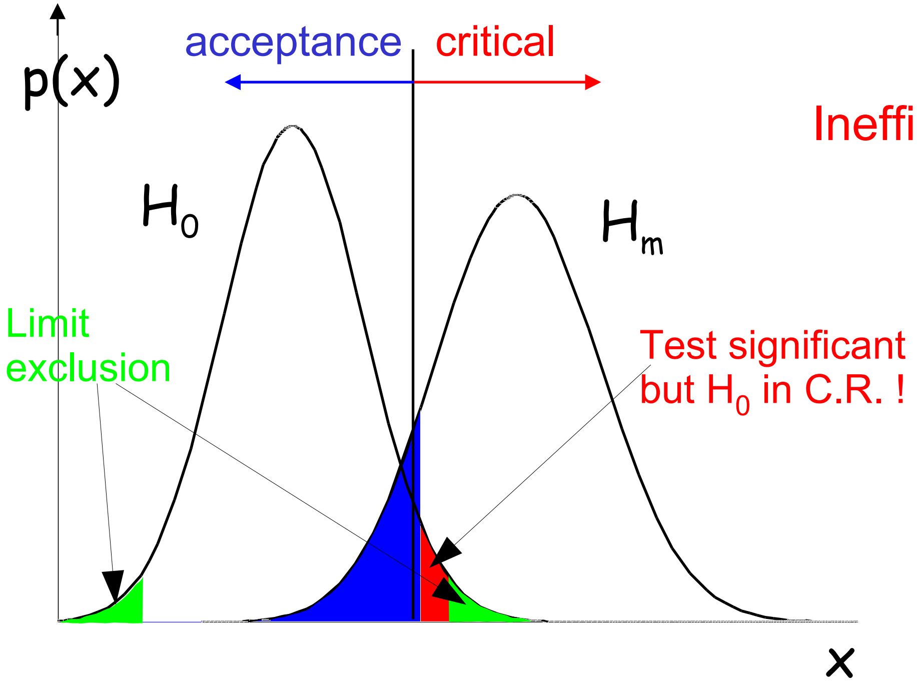
TABLE X. Our confidence intervals for the mean μ of a Gaussian, constrained to be non-negative, as a function of the measured mean x_0, for commonly used confidence levels. Italicized intervals correspond to cases where the goodness-of-fit probability (Sec. IV C) is less than 1%. All numbers are in units of σ.

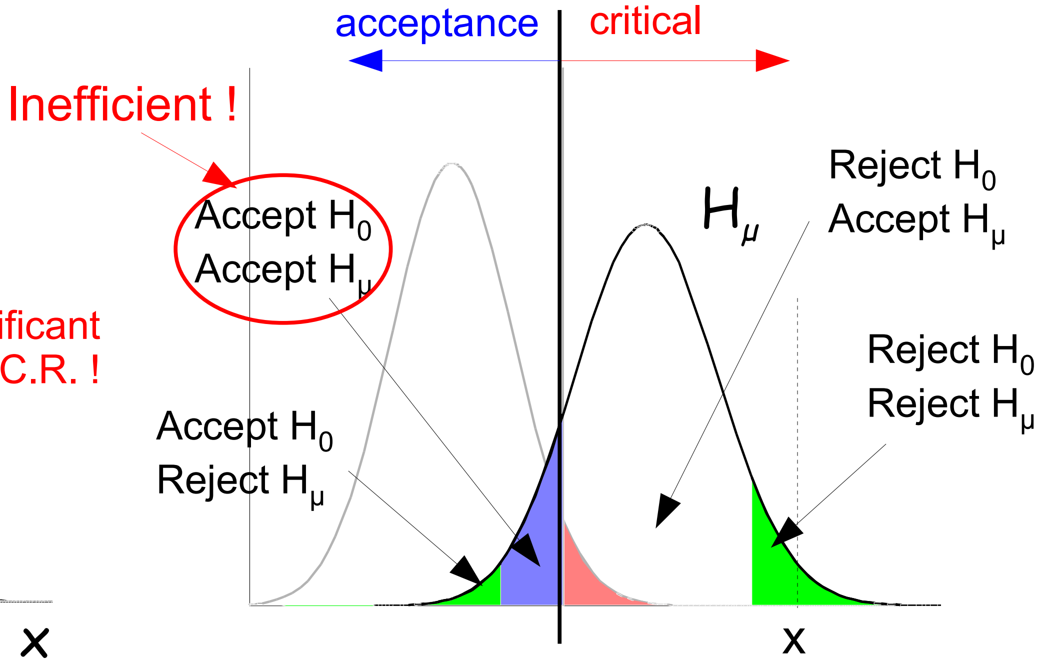
x_0	68.27% C.L.	90% C.L.	95% C.L.	99% C.L.	x_0	68.27% C.L.	90% C.L.	95% C.L.	99% C.L.
−3.0	*0.00, 0.04*	*0.00, 0.26*	*0.00, 0.42*	*0.00, 0.80*	0.1	0.00, 1.10	0.00, 1.74	0.00, 2.06	0.00, 2.68
−2.9	*0.00, 0.04*	*0.00, 0.27*	*0.00, 0.44*	*0.00, 0.82*	0.2	0.00, 1.20	0.00, 1.84	0.00, 2.16	0.00, 2.78
−2.8	*0.00, 0.04*	*0.00, 0.28*	*0.00, 0.45*	*0.00, 0.84*	0.3	0.00, 1.30	0.00, 1.94	0.00, 2.26	0.00, 2.88
−2.7	*0.00, 0.04*	*0.00, 0.29*	*0.00, 0.47*	*0.00, 0.87*	0.4	0.00, 1.40	0.00, 2.04	0.00, 2.36	0.00, 2.98
−2.6	*0.00, 0.05*	*0.00, 0.30*	*0.00, 0.48*	*0.00, 0.89*	0.5	0.02, 1.50	0.00, 2.14	0.00, 2.46	0.00, 3.08
−2.5	*0.00, 0.05*	*0.00, 0.32*	*0.00, 0.50*	*0.00, 0.92*	0.6	0.07, 1.60	0.00, 2.24	0.00, 2.56	0.00, 3.18
−2.4	*0.00, 0.05*	*0.00, 0.33*	*0.00, 0.52*	*0.00, 0.95*	0.7	0.11, 1.70	0.00, 2.34	0.00, 2.66	0.00, 3.28
−2.3	0.00, 0.05	0.00, 0.34	0.00, 0.54	0.00, 0.99	0.8	0.15, 1.80	0.00, 2.44	0.00, 2.76	0.00, 3.38
−2.2	0.00, 0.06	0.00, 0.36	0.00, 0.56	0.00, 1.02	0.9	0.19, 1.90	0.00, 2.54	0.00, 2.86	0.00, 3.48
−2.1	0.00, 0.06	0.00, 0.38	0.00, 0.59	0.00, 1.06	1.0	0.24, 2.00	0.00, 2.64	0.00, 2.96	0.00, 3.58
−2.0	0.00, 0.07	0.00, 0.40	0.00, 0.62	0.00, 1.10	1.1	0.30, 2.10	0.00, 2.74	0.00, 3.06	0.00, 3.68
−1.9	0.00, 0.08	0.00, 0.43	0.00, 0.65	0.00, 1.14	1.2	0.35, 2.20	0.00, 2.84	0.00, 3.16	0.00, 3.78
−1.8	0.00, 0.09	0.00, 0.45	0.00, 0.68	0.00, 1.19	1.3	0.42, 2.30	0.02, 2.94	0.00, 3.26	0.00, 3.88
−1.7	0.00, 0.10	0.00, 0.48	0.00, 0.72	0.00, 1.24	1.4	0.49, 2.40	0.12, 3.04	0.00, 3.36	0.00, 3.98
−1.6	0.00, 0.11	0.00, 0.52	0.00, 0.76	0.00, 1.29	1.5	0.56, 2.50	0.22, 3.14	0.00, 3.46	0.00, 4.08
−1.5	0.00, 0.13	0.00, 0.56	0.00, 0.81	0.00, 1.35	1.6	0.64, 2.60	0.31, 3.24	0.00, 3.56	0.00, 4.18

Putting everything together: *Combined* Optimization

Coordinating H-Test with Limits

- In principle, limit-setting may be done in a way totally unconnected with the test of hypothesis. But there can be undesirable consequences.





Exclude H_0 despite non-significant test

Lack of power in setting limits on μ

→ **Optimality requirement: the ordering algorithm, for each μ, must first exclude the acceptance region for H_0.**

Definition of “sensitivity region” for a search

ArXiv:physics/0308063

- Differs from usual notion of sensitivity as a *number:* here we consider ***the range of hypotheses for which the experiment will provide a “definite conclusion”***
- *Def:* The **sensitivity region** for a *search* is defined as:

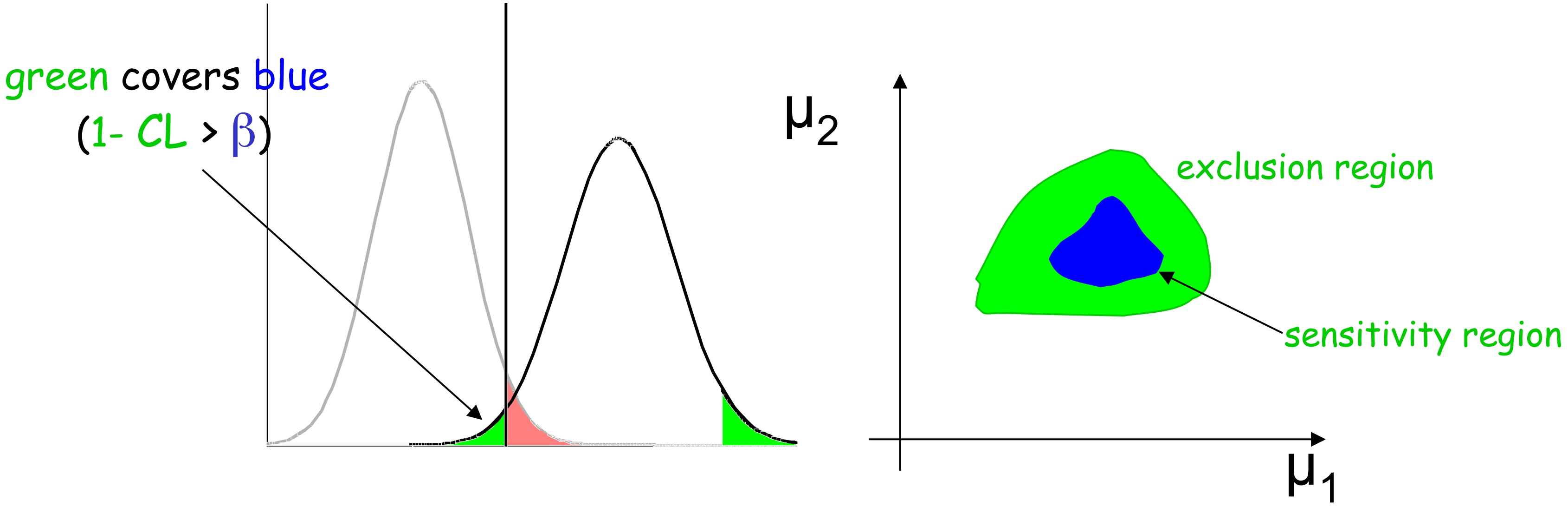
$$\mathcal{S} = \{\mu: 1 - \beta_\alpha(\mu) > CL\}$$

- Theorem: the following two facts hold simultaneously:

1) If the true value of μ is inside $\mathcal{S}$, the probability of discovery (excluding H_0 @signif. α) is ***at least*** CL

2) In case of non-significant result, the excluded region @CL will ***contain*** $\mathcal{S}$ (independently of the true value of μ !)

Definition of “sensitivity region”

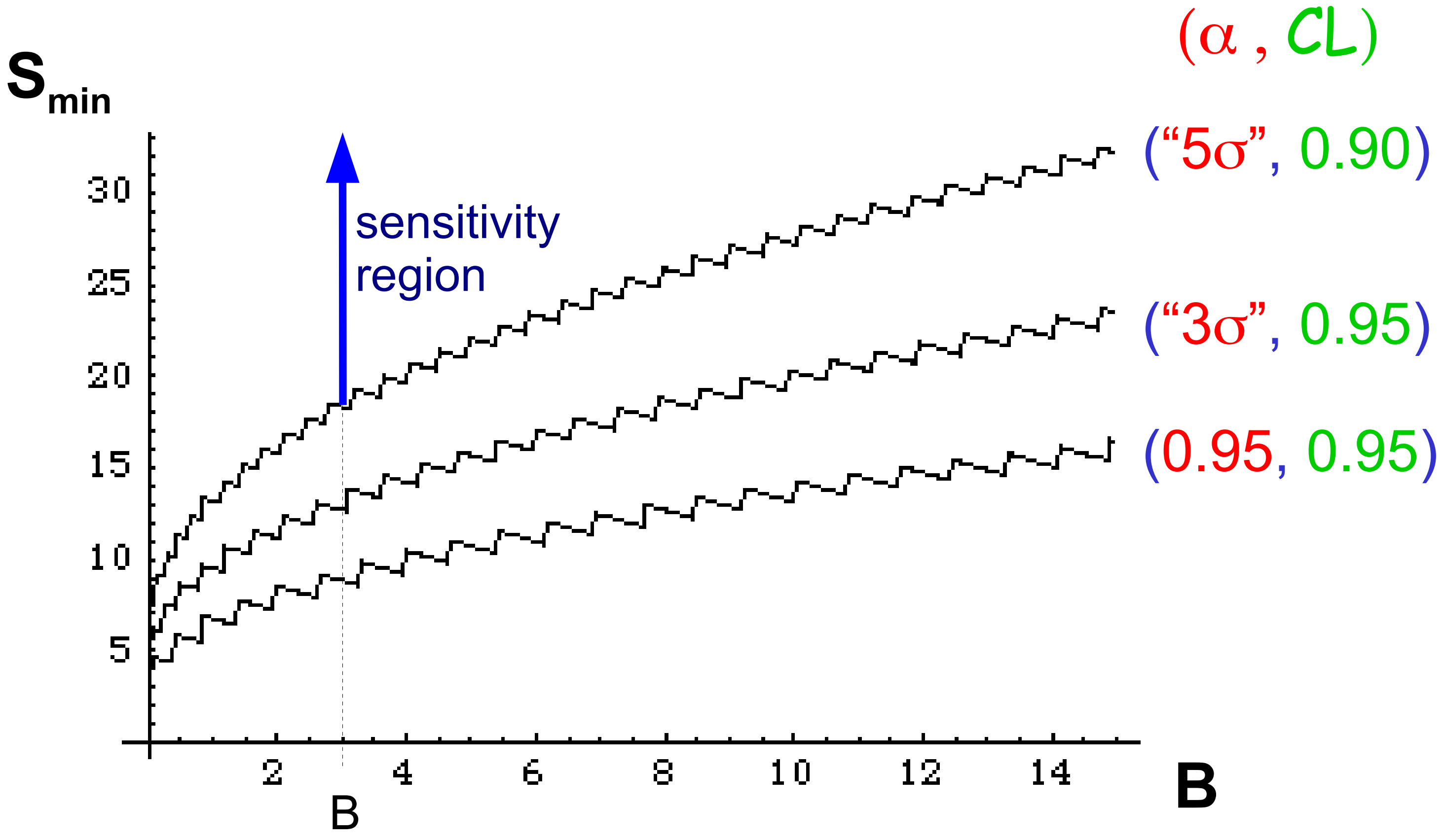


- “Unified” view of sensitivity: independently of expectations on H_μ , you want the sensitivity region to be as large as possible,
- Does not contain elements of randomness (“absolute sensitivity”). If the sensitivity region covers the whole parameter space of a theory, the experiment is conclusive.
- No dependence on metric or priors (purely frequentist) – or expected signal.
- No dependence on the limits ordering function. *Apart from the “optimality requirement”, can use any method with frequentist coverage (including CLs).*
- Valid in any number of dimensions

Application to 'counting experiment'

Sensitivity region takes the simple form: $S(\mu,t) > S_{min}(B(t))$

It depends on both the chosen **α** and **CL**



Note small decreasing tracts – due to intrinsic discrete problem

Can take advantage of a smooth interpolation

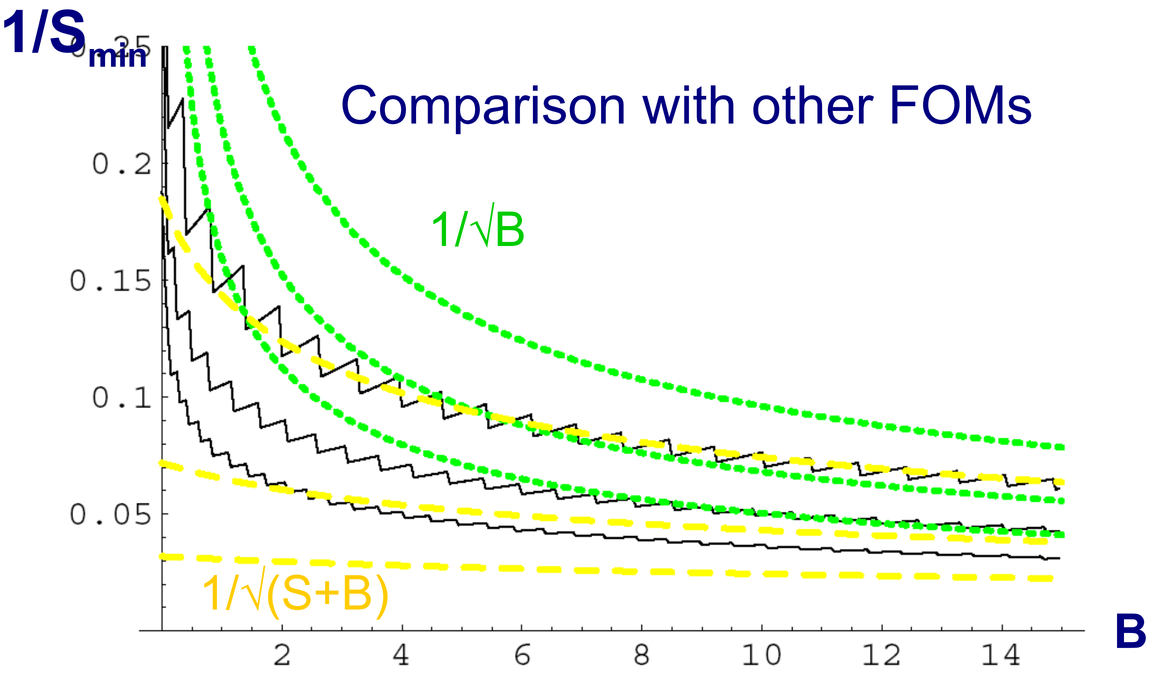
Use in Optimization of a Search

- S depends on selection cuts **t**: $S(\mu,t)=\varepsilon(t)*L*\sigma(\mu) > S_{min}$

  → $\sigma(\mu) > S_{min}/(\varepsilon(t)*L)$

- Want to be sensitive to the smallest possible x-section $\sigma(\mu)$

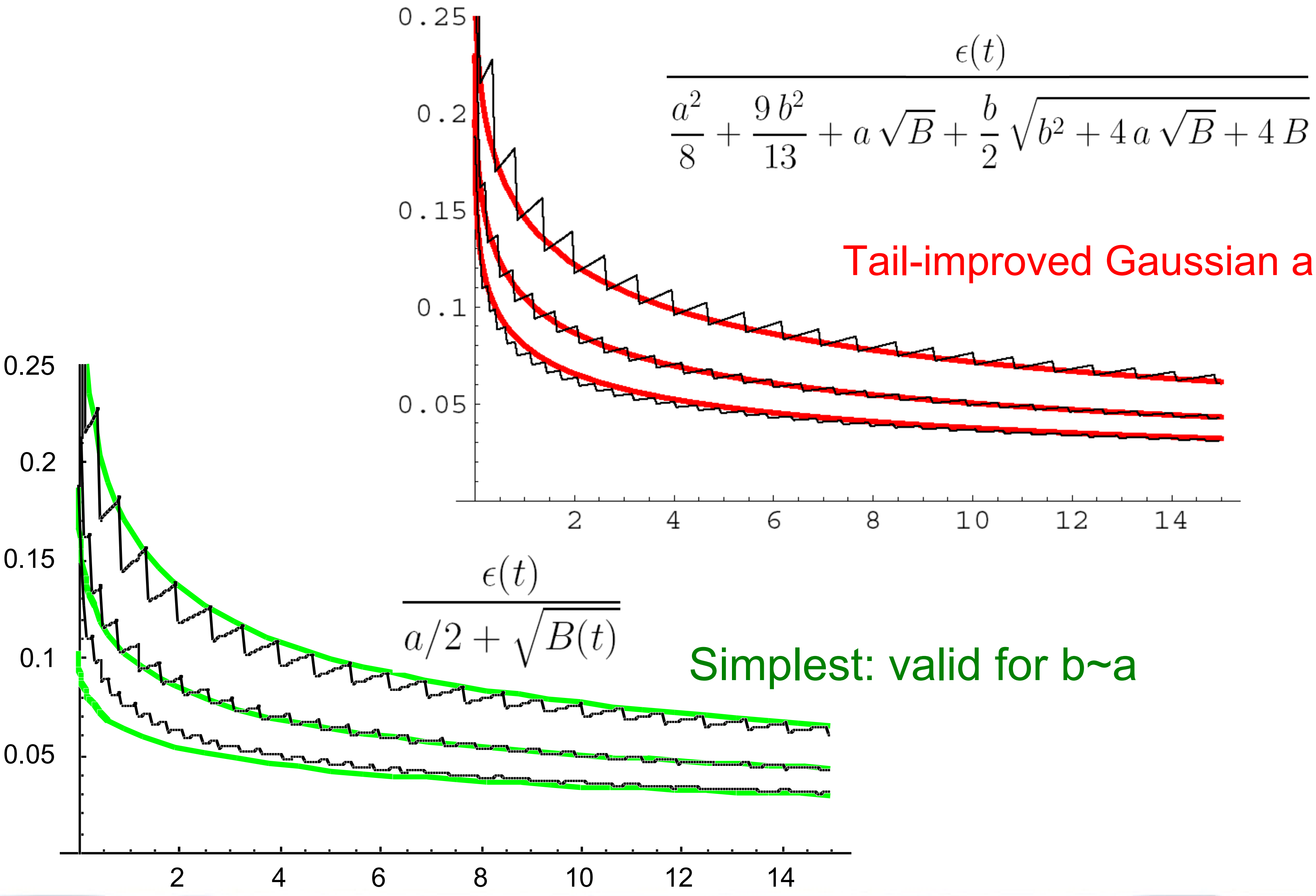
  → For best sensitivity, **maximize** the FOM: $\varepsilon(t)/S_{min}$



- ***Independent*** from expected cross section for signal
- Does not diverge for small B

Approximate formulas

(a,b) = # of sigmas for (α, β)



$$\frac{\epsilon(t)}{\frac{a^2}{8}+\frac{9\,b^2}{13}+a\sqrt{B}+\frac{b}{2}\sqrt{b^2+4\,a\sqrt{B}+4\,B}}$$

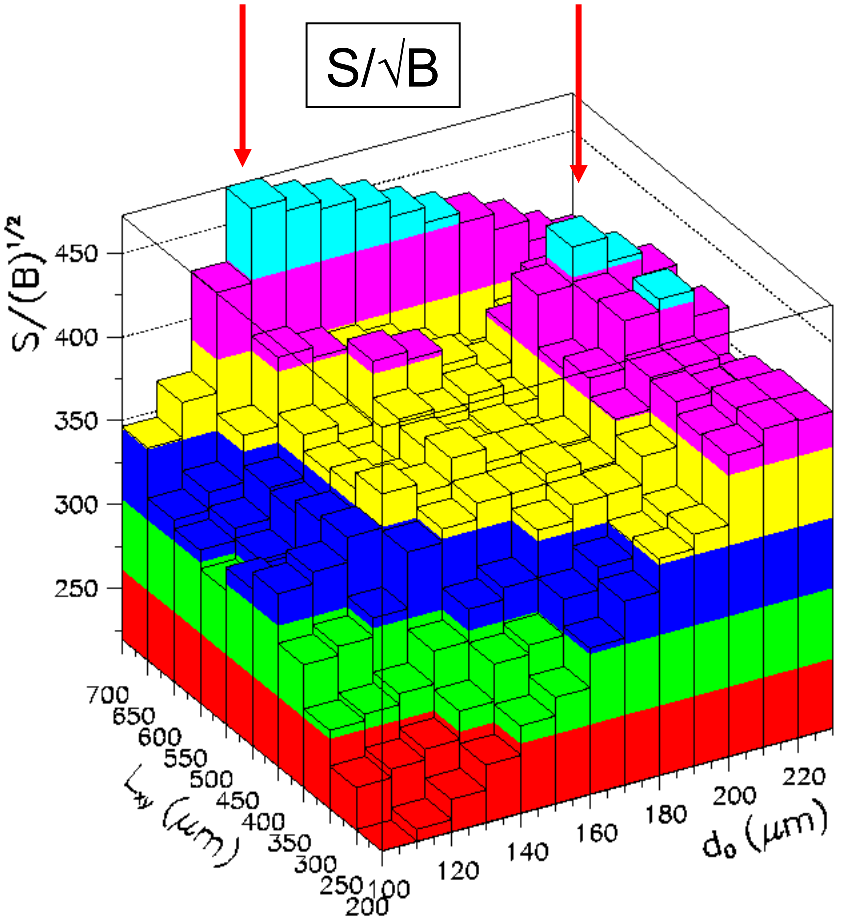
Tail-improved Gaussian approx.

$$\frac{\epsilon(t)}{a/2+\sqrt{B(t)}}$$

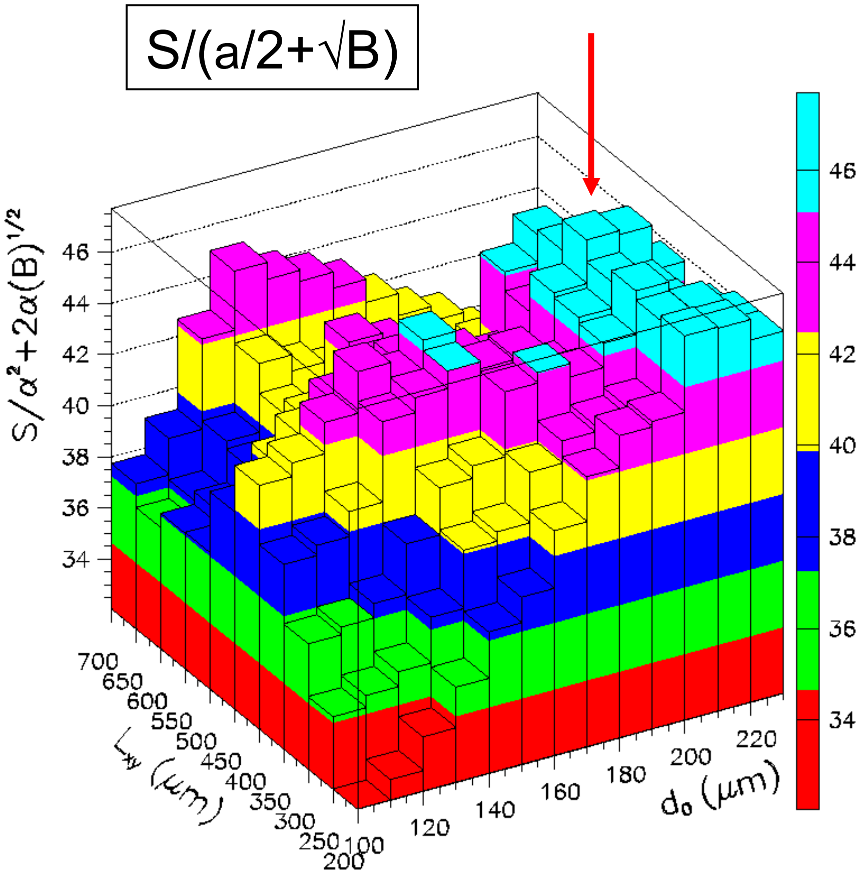
Simplest: valid for b~a

A real-life example

Optimizing a search for rare decays



S/√B



S/(a/2+√B)

Whe $\alpha = 3$

The new formula eliminates fake solution with tight cuts

Summary

$$\frac{\epsilon(t)}{a/2 + \sqrt{B(t)}}$$

- Maximization of this FOM turned out very useful in searches – it has been used in >100 papers (and counting) by various experiments
- Independent of absolute signal size, and of absolute efficiency
- Simple, and easy to evaluate
- It is the result of applying the general criteria of largest sensitivity region (1-β > CL) to a "counting experiment"
- For full details, see ArXiv:physics/0308063 - you can optimize the method for your specific application if you wish to.