Statistics…

> “There are lies, damned lies, and statistics.”
>
> Mark Twain

Hypothesis → Data

- P(data|H) it is the function which describes the experiment. It gives the probability of observing data when the laws of physics are given by the hypothesis **H**.

- H → data occurs in real experiments (where H is unknown but true)

- H → data occurs in simulation (where H is known but generally not true)

- Data are random. Hypotheses are NOT.

Data → Hypothesis

The reverse process is called **statistics**: The Bayesian way and the Frequentist way.

1. *Probabilities (B. vs F.)*
2. *Measuring a parameter: Point Estimation*
3. *Finding the error on the above:* ***Interval Estimation***
4. *Comparing two hypotheses:* ***Hypothesis Testing***
5. *Hands-on exercises.*

Statistics should not be a replacement of common sense…
and common sense should not replace statistics.

Probability

- All statistical methods are based on calculations of probability. We can define three different kinds of probability.

1. **Mathematical probability.** Is the probability defined by some axioms alone.

   i. $P(X_i) \geq 0$ for all i.

   ii. $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$

   iii. $\sum_{\Omega} P(X_i) = 1.$

2. **Frequentist probability.** Is defined as the *limiting frequency* of favorable outcomes in a large number of identical experiments.

3. **Bayesian probability.** Is defined as the *degree of belief* in a favorable outcome of a single experiment.

Frequentist Probability

- First defined by John Venn in 1866.

- The frequentist probability of an event A is defined as the number of times A occurs, divided by the total number of trials N, in the limit of a large number of numerical trials:

$$P(A) = \lim_{N \to \infty} \frac{N(A)}{N}$$

- where A occurs N(A) times in N trials. Frequentist probability is used in most scientific work because it is objective.

Frequentist Probability

Cons:

- It requires an **infinite number of experiments.** But many scientific concepts are defined as limits.
- It requires a repeatable phenomena. Most scientific work involves repeatable phenomena, and frequentist probability is well defined.

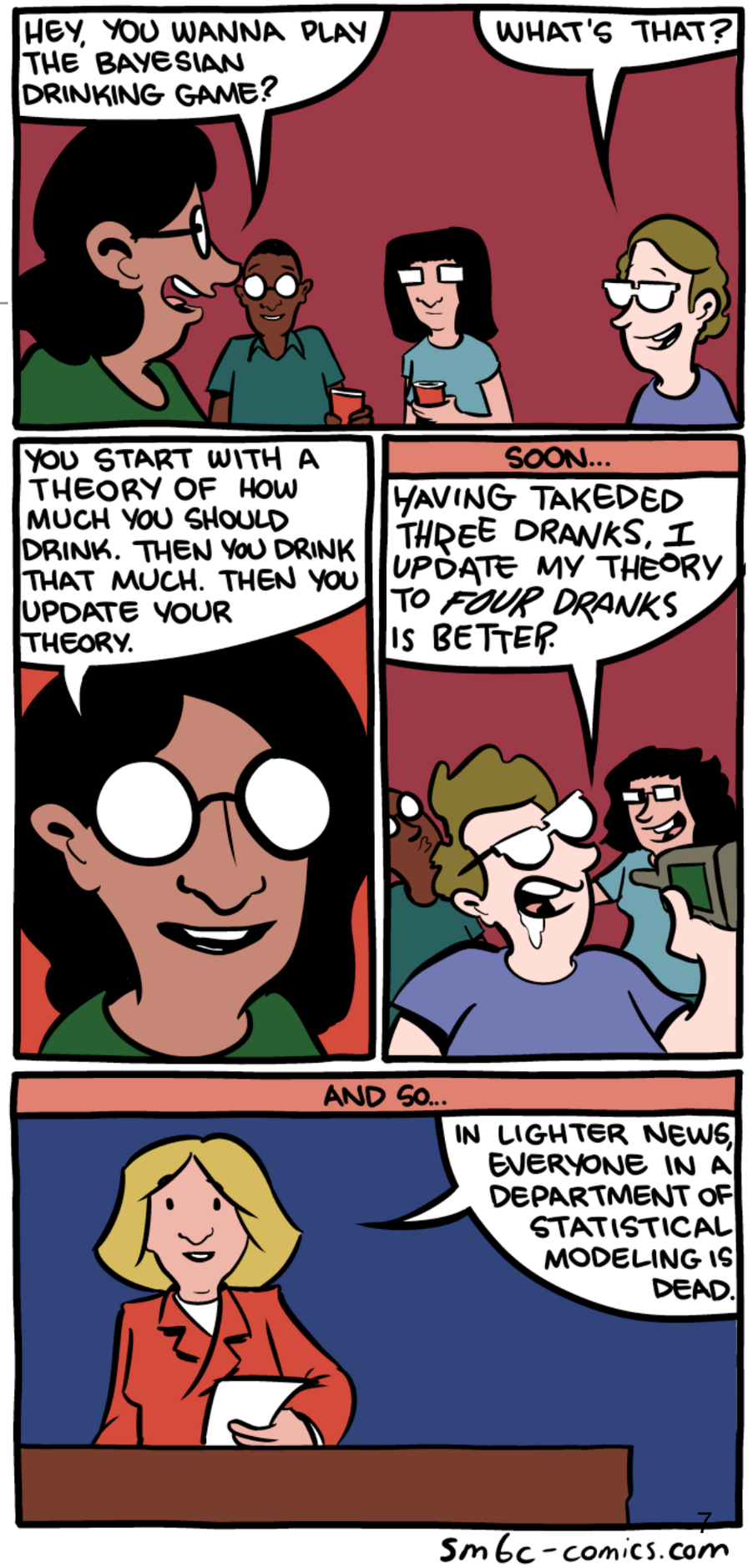
But we need in addition a more general kind or probability if we want to apply it to non-repeatable phenomena.

Bayesian Probability

- For phenomena that are ***not repeatable*** (for example: the probability it will rain tomorrow). It is not easy to define.

- The **Bayesian probability** of A depends on the ***degree of belief*** that A will happen and therefore is not only a property of A but depends also on the state of knowledge and beliefs of the observer. It generally ***changes with time*** as the observer gains more knowledge.

I will focus on the frequentist approach as is generally used in HEP.
More on the debate:
http://ned.ipac.caltech.edu/level5/March01/Dagostini/Dagostini_contents.html
Why isn't every physicist a Bayesian? - Cousins, Robert D. Am.J.Phys. 63 (1995) 398
UCLA-HEP-94-005



Hypothesis

- The Hypothesis is what we want to test, verify, measure.
  - Example of H: Data is described by a point-source signal.
  - P(data|H) is assumed known. For a example in a Poisson process:

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

where N is our data and μ is our Hypothesis.

Probability Density Function

- When data are continuous, the probability P becomes a Probability Density Function, or pdf, as in:

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

here **μ** is the parameter of interest (Hypothesis) but **σ**, if it is not known, is a **nuisance parameter**: an unknown whose value does not interest us, but is necessary for the calculation of P(data| H)

Likelihood function

- If in P(data | H) we put in the values of the data observed in the experiment Ω_{data}, and consider the resulting function as a function of the unknown parameter (H) it becomes:

$$\prod_{\Omega_{data}} P(data|H) = \mathcal{L}(H)$$

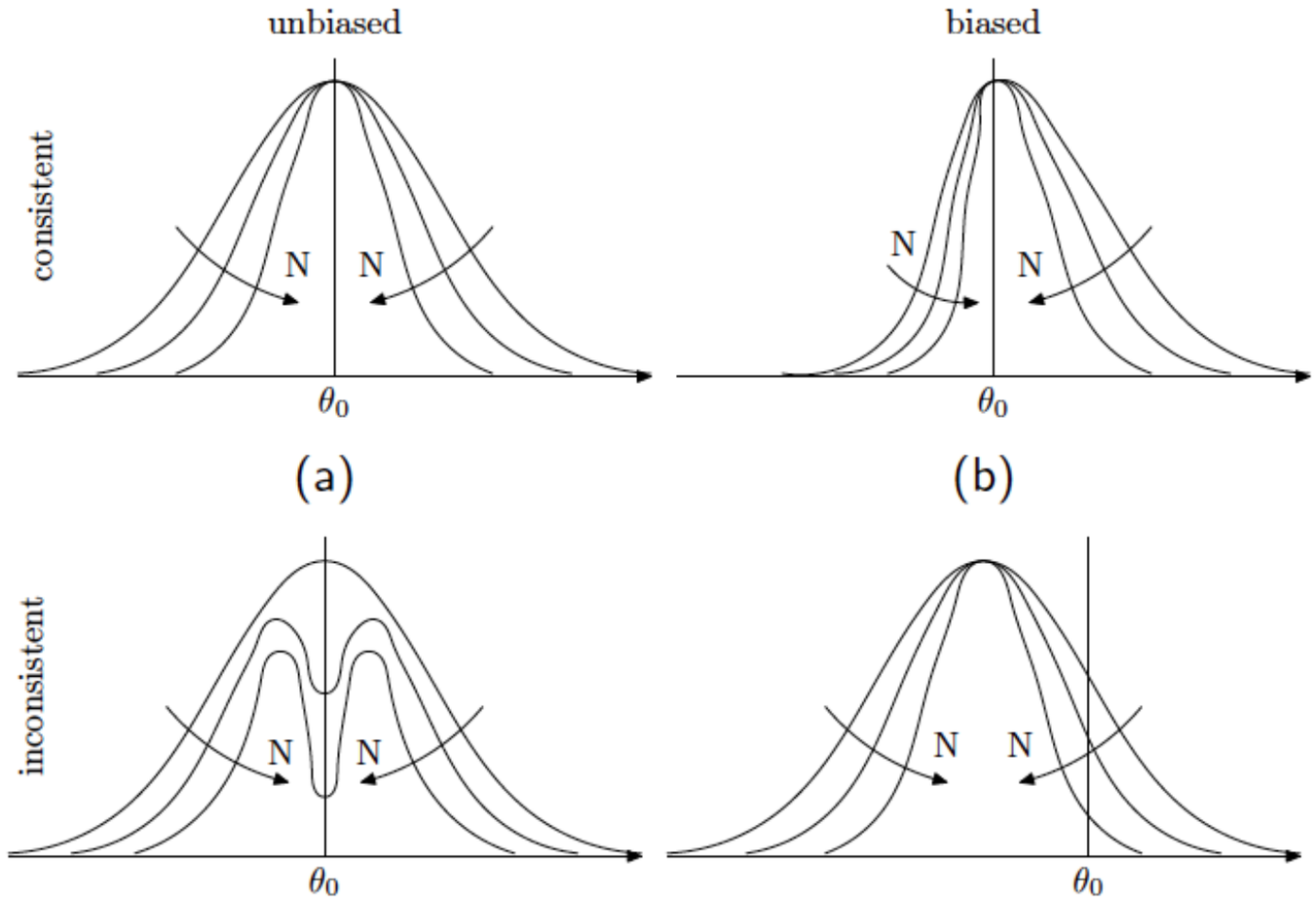
- This is called the **likelihood function**. It is **not a probability** (at least in frequentist approach)

- The likelihood is invariant under transformation H → H'(H) $\mathscr{L}$(H) = $\mathscr{L}$(H'(H)) but a pdf needs to have the ***Jacobian*** of the transformation (X→ Y(X)). Pdf(X) = J(X,Y)Pdf(Y)

Point Estimation

- The goal of Point Estimation is to find the **function of the data X**, $\varepsilon_\mu(X)$, which gives the best estimate μ' (measurement) of the parameter μ, i.e. $\varepsilon_\mu(X) = \mu'$.
- A parameter estimator has to be: consistent, unbiased, and efficient.

its variance as close as the cramer-rao lower limit



Point estimate: Maximum Likelihood

- The maximum likelihood (ML) estimate of a parameter μ is that μ' which $\mathscr{L}$(X|μ) has its maximum given the particular observation X.

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \ln f(x_i, \mu) = \frac{\partial}{\partial \mu} \ln \mathcal{L}(X, \mu) = 0$$

This is the analytic way to find the maximum, but in practice we do it numerically.

Asymptotically (for very large data samples N→∞), the ML estimator is consistent unbiased, and efficient, it follows the properties:

i. The variance is given by the **Cramer-Rao bound**

ii. The estimates are **normally distributed**

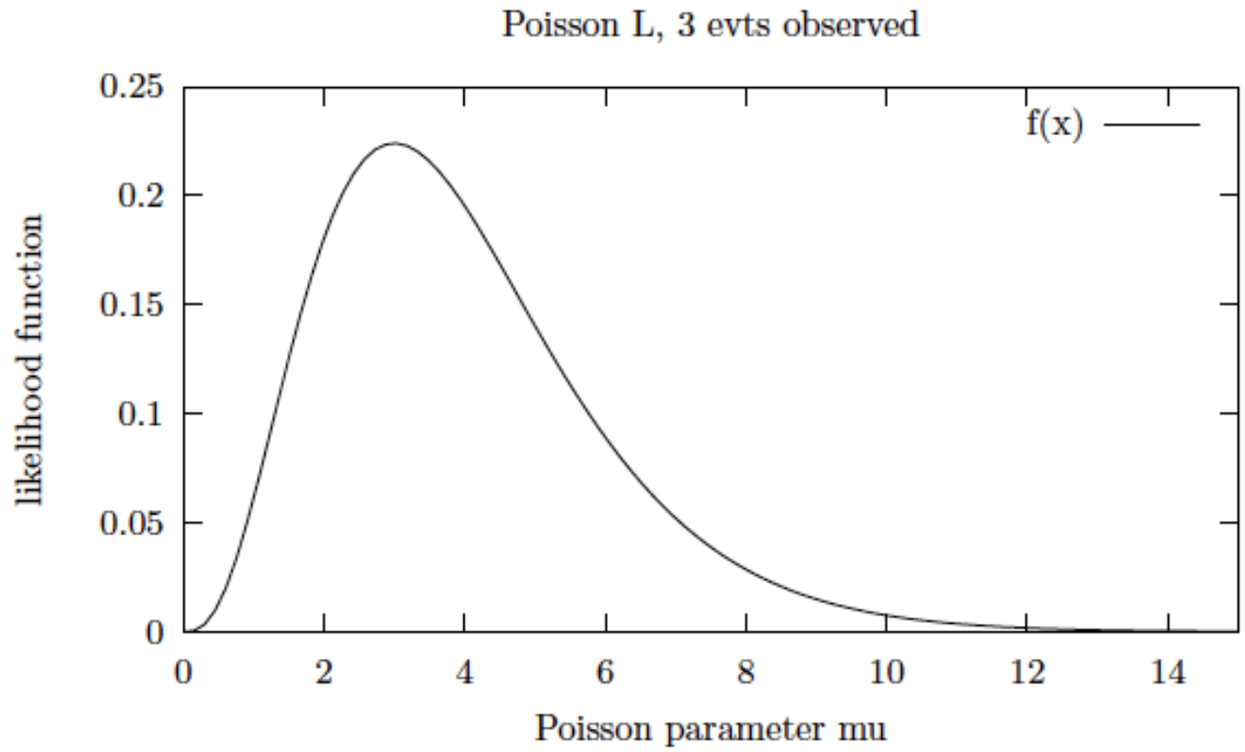
Thus, asymptotically, $\mathscr{L}$ is proportional to a Gaussian function of with mean μ' and variance 1/I(μ')

You can find the proof in any statistics book

Point estimate: Poisson example

- In a Poisson process, we observe 3 events.

$$\mathcal{L}(\mu) = P(3|\mu) = \frac{e^{-\mu}\mu^3}{3!}$$



The peak in the likelihood occurs at μ = 3.

Generalizing from 3 to n, we get the expected result:

with n events observed, μ' = n

Interval Estimation

- The goal of interval estimation is to find an interval which will contain the true value of the parameter with a given probability.

- The meaning of this probability, and hence the meaning of the interval, will of course be very different for the Bayesian and frequentist methods.

Interval Estimation: Frequentist (Classical) construction

The Problem: Given β, find the optimal range (θ_a, θ_b) in θ-space such that:

$$P(\theta_a \leq \theta_{\text{true}} \leq \theta_b) = \beta\,.$$

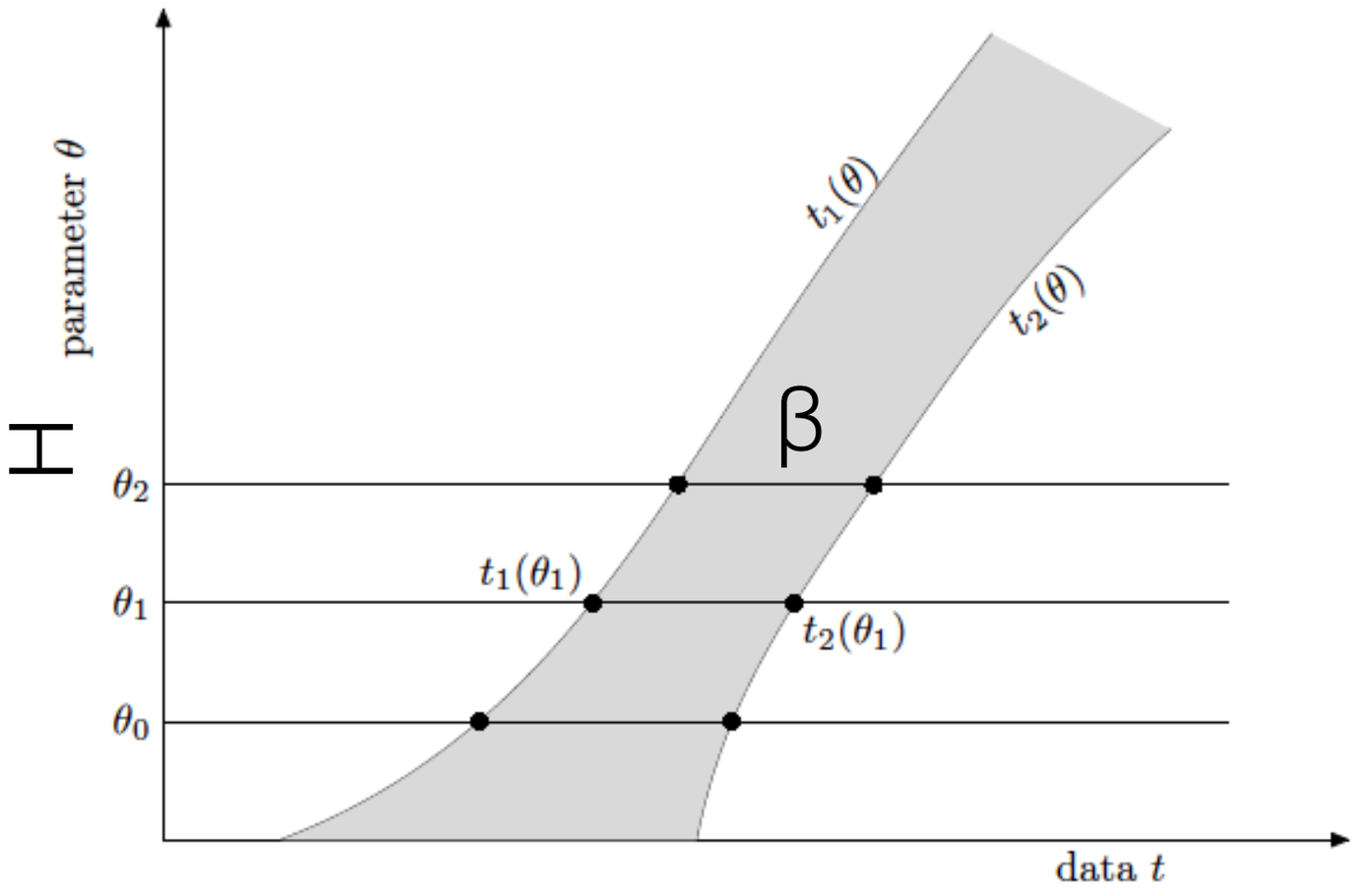
The interval (θ_a, θ_b) is then called a confidence interval. A method which yields intervals (θ_a, θ_b) satisfying the above equation is said to possess the property of coverage.

- Overcoverage occurs when **P > β**
- Undercoverage occurs when **P < β**

Interval Estimation: Exact Frequentist Theory

- Neyman and Pearson created the Neyman construction of confidence intervals to solve the general case of confidence interval (and not only the normal theory case).

- The first important step in finding an exact theory was to work in the right space: P(data| H), with one axis (or set of axes) for data, and another for hypotheses

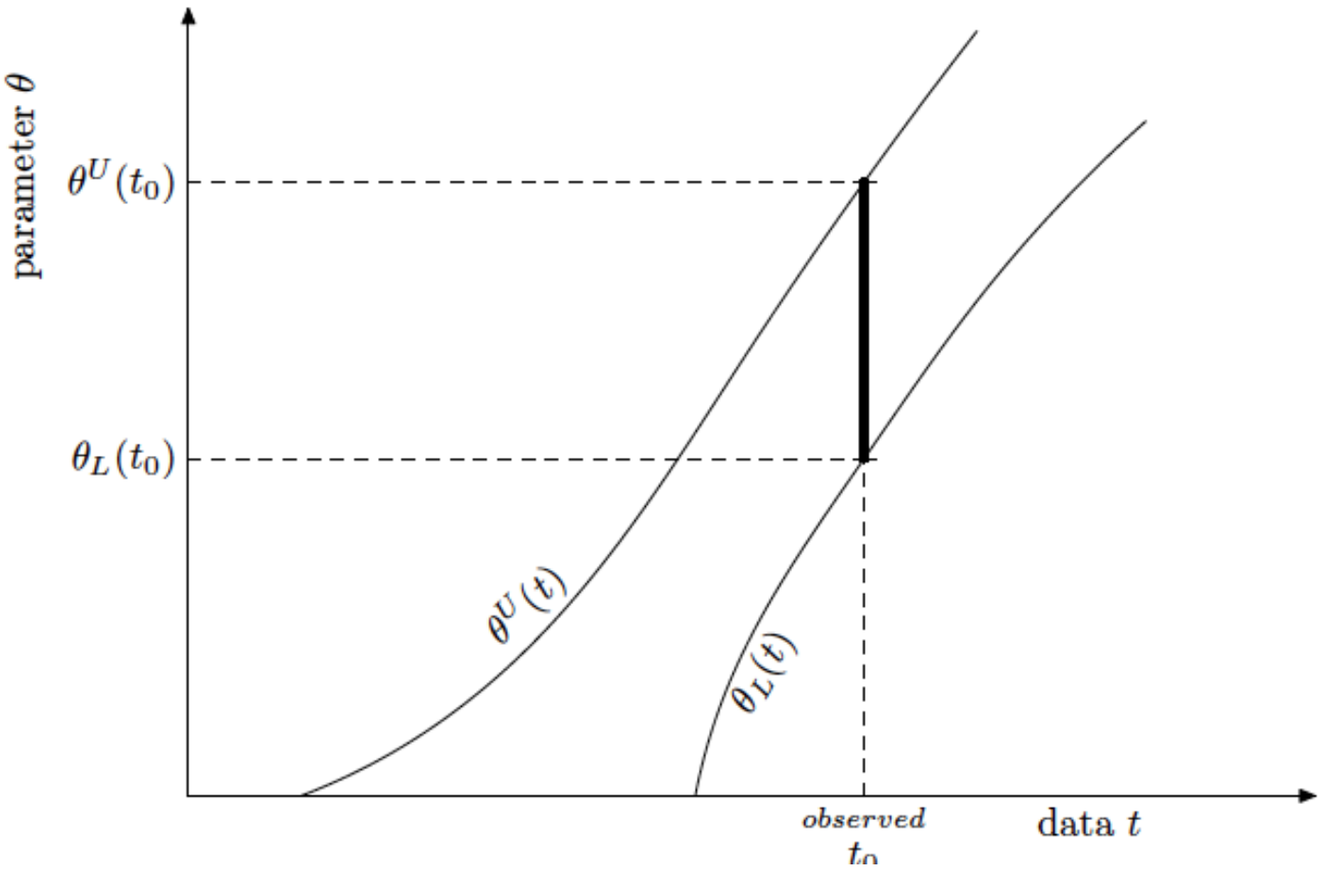


H

DATA

Interval Estimation: Neyman construction

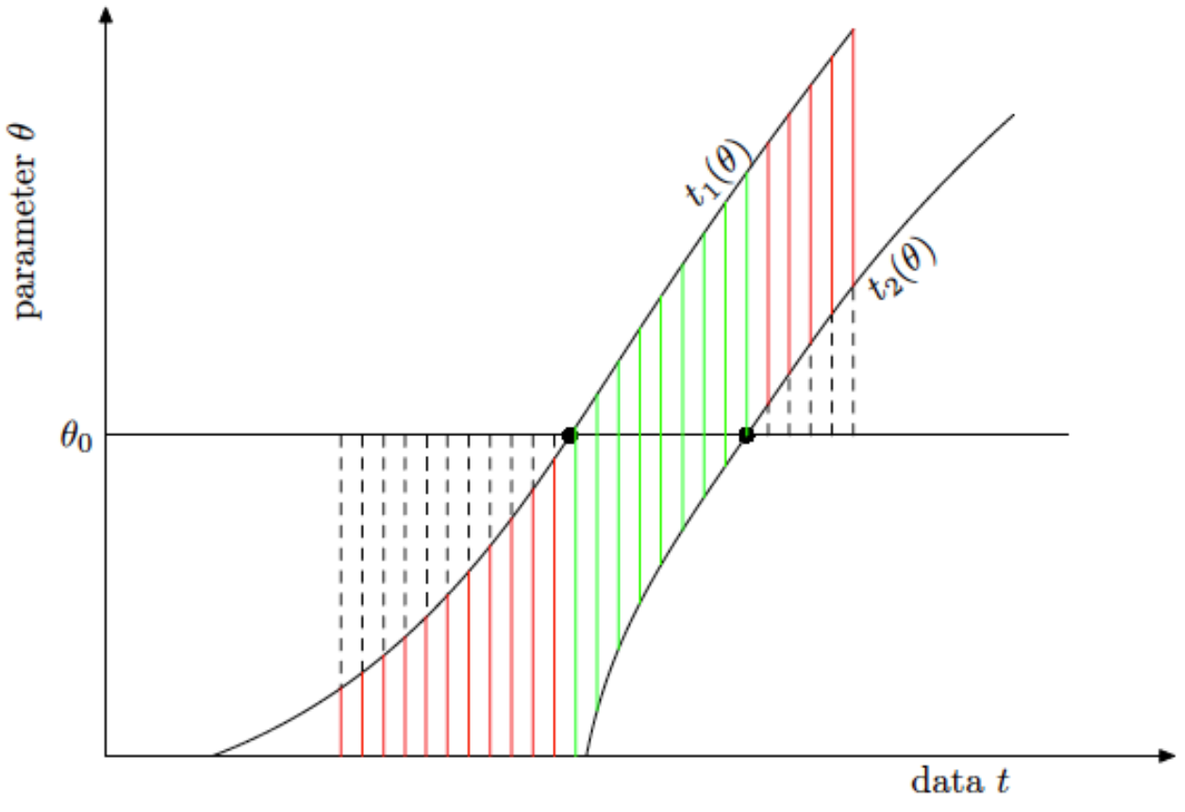
- The two curves of t(θ) are re-labelled as θ(t), and the confidence limit is read vertically.



- For observed data t_0, the confidence interval is $(\theta_L(t_0); \theta^U(t_0))$: $P(\theta_L < \theta_{true} < \theta^U) = \beta$

Interval Estimation: Neyman construction

- Suppose the true value is θ_0. Then, depending on the observed data, we could get the intervals indicated as red and green vertical lines below:



Only the green confidence intervals cover the true value.

The probability of getting a green confidence interval is β.

Question?

- What of the two statements is correct?

  a) Given a confidence interval the probability that the true value is in the interval is β

  b) Given a confidence interval, the probability that the interval covers the true value is β

Question?

- What of the two statements is correct?

  a) Given a confidence interval the probability that the true value is in the interval is β

  b) Given a confidence interval, the probability that the interval covers the true value is β

Exercises

Download the scripts from icecube SVN

```
> export SVN="http://code.icecube.wisc.edu/svn"

> svn co $SVN/sandbox/aguilar/BootCamp14

> cd BootCamp14/Stats

> offline-software

> nb
```

Click on statsNormal

Problems with the Neyman construction.

- We will deal with them after Hypothesis Testing.

Hypothesis Testing

- Compare two hypotheses to see which one better explains the data.

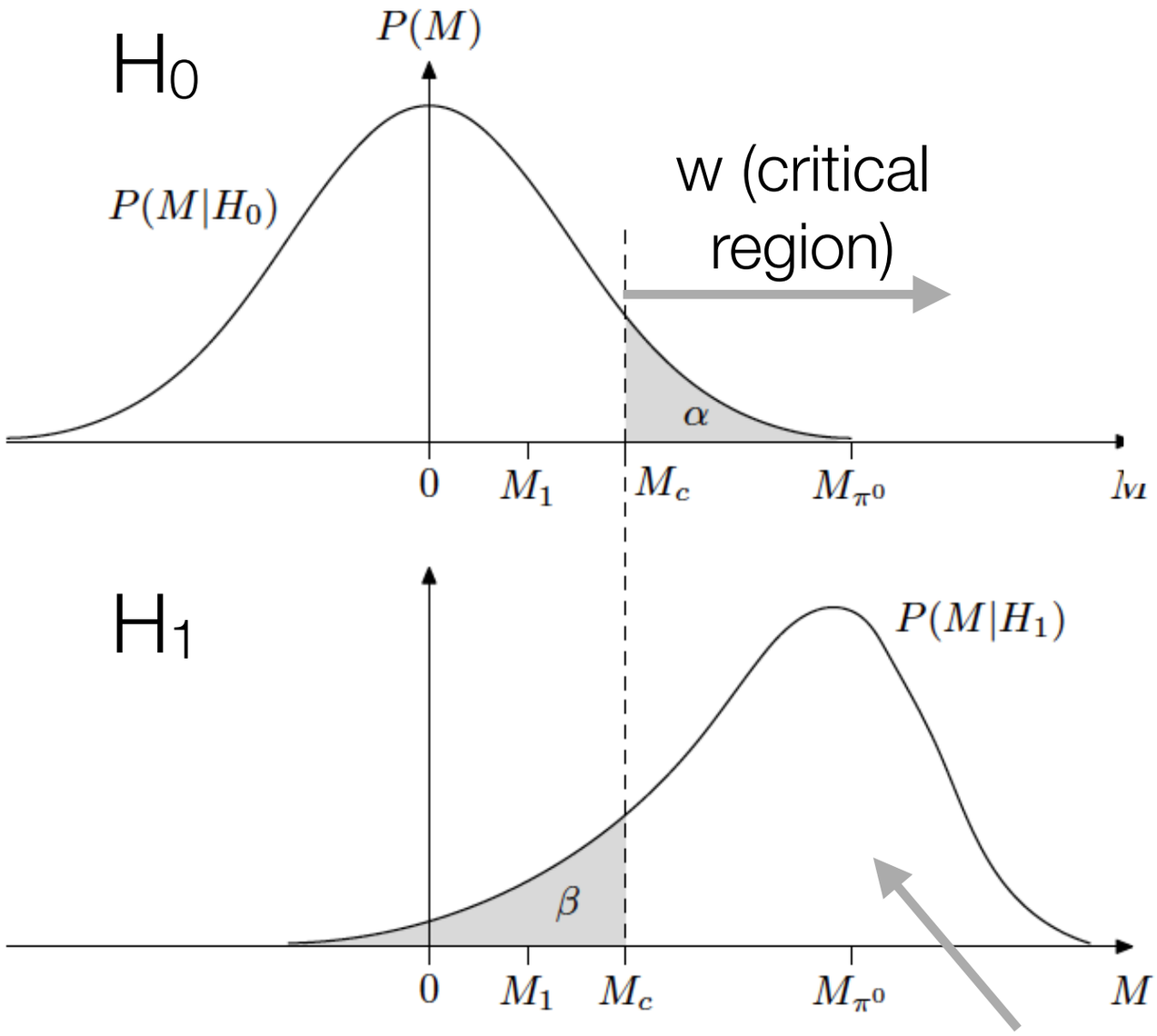
- Or, alternatively, what is the best way to separate events into two classes, those originating from each of two hypotheses.

- The two hypotheses are traditionally called:

  - H_0: null hypothesis (or background hypothesis)

  - H_1: alternative hypothesis (or signal hypothesis)

Hypothesis testing



		H_0 TRUE	H_1 TRUE
$X \notin w$	ACCEPT H_0	Acceptance good Prob $= 1 - \alpha$	Contamination Error of the second kind Prob $= \beta$
$X \in w$ (critical region)	REJECT H_0	Loss Error of the first kind Prob $= \alpha$	Rejection good Prob $= 1 - \beta$

Loss: Error of first type; α

Contamination: Error of second type; β

power of the test: 1 - β

Which test to use?

- We will deal with one case: Hypothesis of the same parametric family.

- This means that we can go from H_0 to H_1 by changing the value of 1 or more parameters. Ex:

$$H_0 : \boldsymbol{\theta} \in \nu$$

$$H_1 : \boldsymbol{\theta} \in \theta - \nu$$

- We saw that the maximum likelihood is a good point-estimate. Closely related the likelihood ratio method proposed by Neyman is a good test for hypothesis testing:

$$\lambda = \frac{\max\limits_{\boldsymbol{\theta} \in \nu} L(\mathbf{X}|\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta} \in \theta} L(\mathbf{X}|\boldsymbol{\theta})} .$$

- This defines the **test statistic for H_0**

Maximum Likelihood Ratio Test

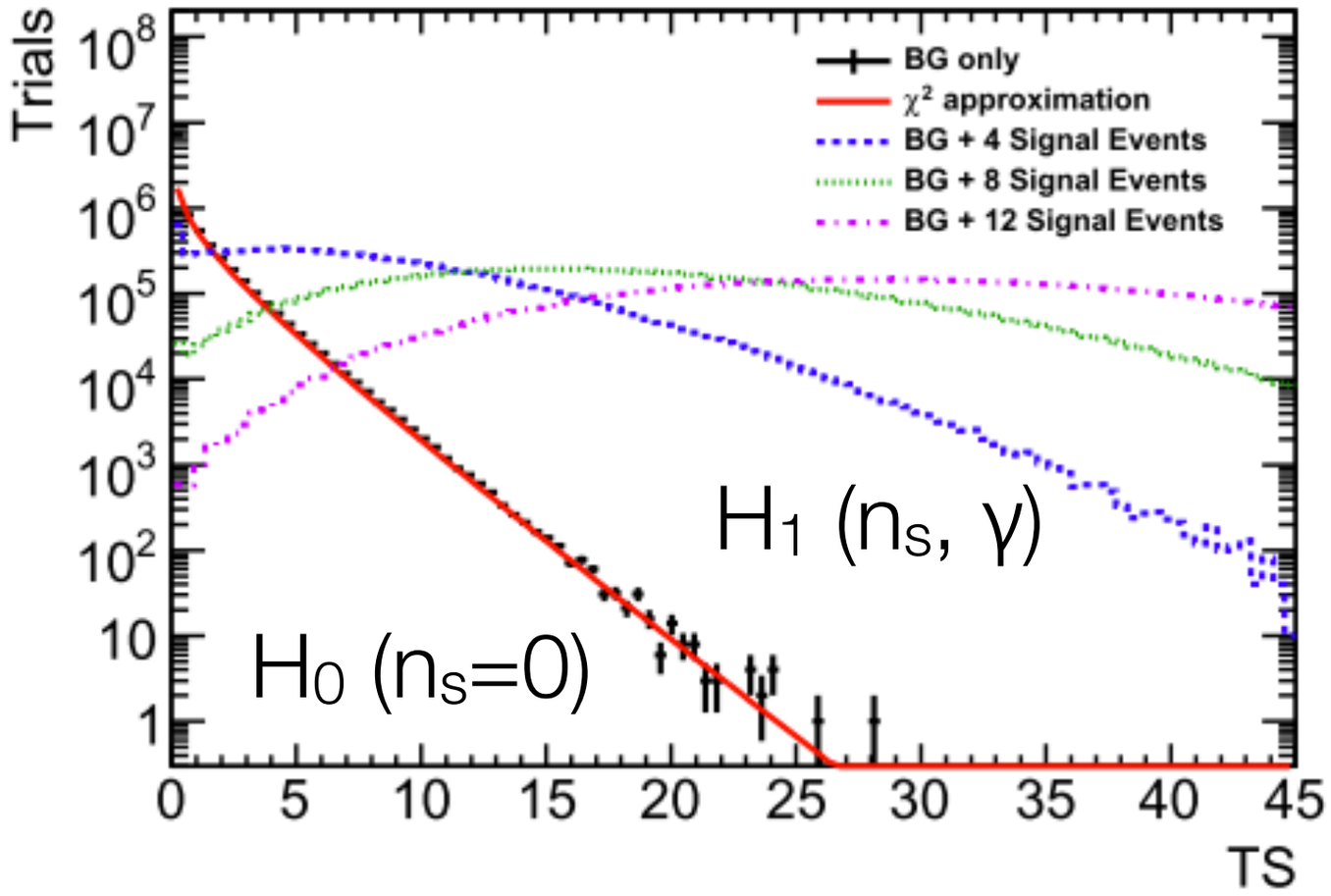
- The importance of the maximum likelihood ratio comes from the fact that **asymptotically** it behaves as a χ^2

- If H_0 imposes r constraints on the s + r parameters in H_0 and H_1, then

-2 ln λ is distributed as $\chi^2(r)$ under H_0

- This means we can read the confidence level from a table of χ^2.

Example from point-sources



- This is the test statistics used in point-sources analysis. The H_0 is BG only events, the alternate hypothesis, H_1, is BG + signal.

$$\log \lambda = \log\left(\frac{L(\hat{\gamma},\hat{n}_s)}{L(n_s = 0)}\right)$$ Our TS is distributed with a χ^2 of 2 dof.

More on ps tomorrow..

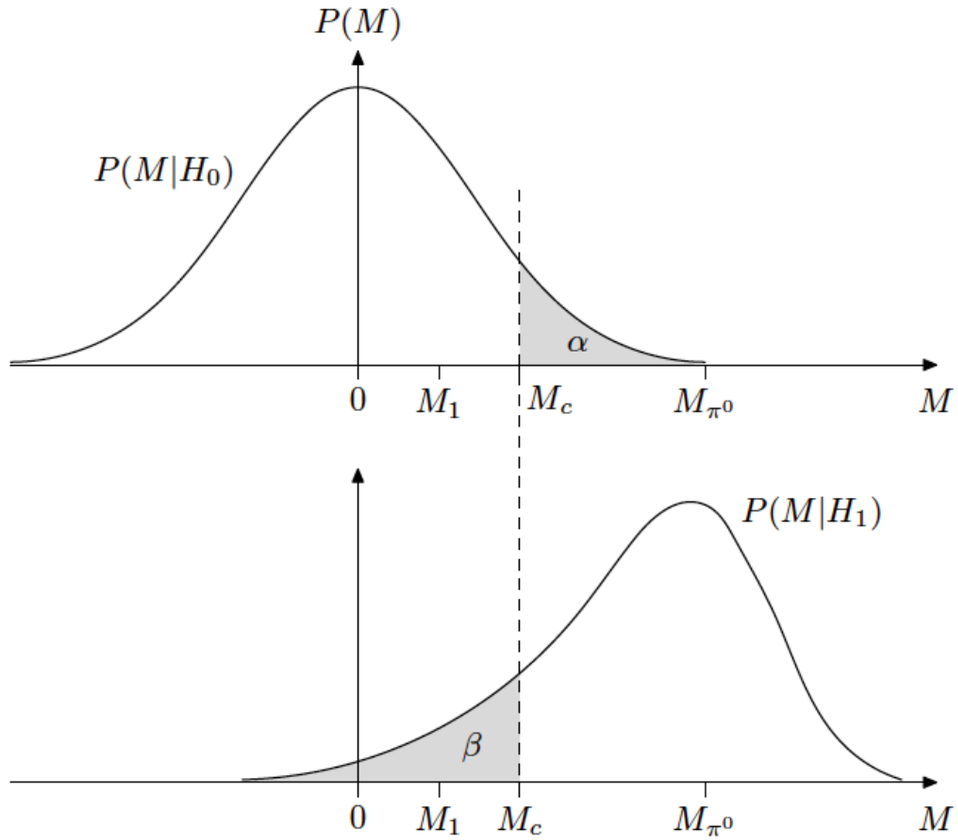
Hypothesis testing

How to define α and β:

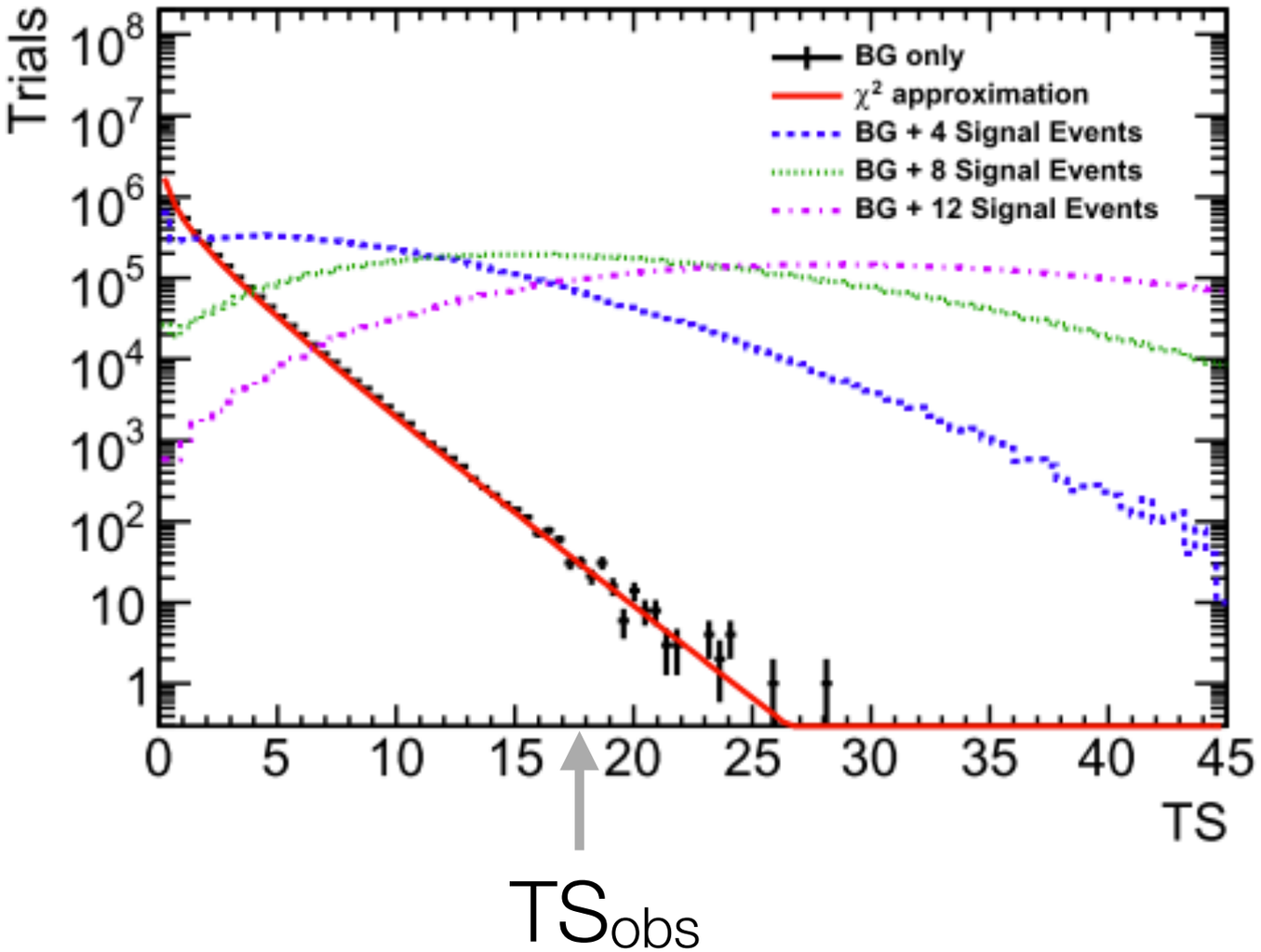
- **Discovery.** Low background loss H_0: α = (1 - 5σ) ~ 2.87e-7 (one-sided). Power 1 - β is selected to 50% chance of accepting/ rejecting H_1

- **Sensitivity.** A representative value of H_0 observation. A common choice is the value at α = 50%. To reject H_1 the power is selected to be 90% (C.L.)

- **Upper limit.** The observation is been made, the value of TS is taken so α equals the p-value.



P-values



- We usually quote a p-value, which is after observation the probability that the observed value comes from H_0.
- It is usually quoted in number of sigmas (even though distributions are not gaussians): $5\sigma = 2.87e\text{-}7$

A p-value is the probability (integral to the right) from the observed value TS_{obs} under the H_0 hypothesis.

Some wording…

- In HEP in general some wording has been adopted as when to claim discovery.
  - 5σ: Discovery
  - 4σ: Evidence
  - 3σ: Observation

- However, will you believe the same a 4σ astrophysical neutrino as a 4σ faster than speed of light neutrino?

side note: recently BICEP2 claimed a 5.2 σ B-mode signal detection. What is wrong about this statement?

Some wording…

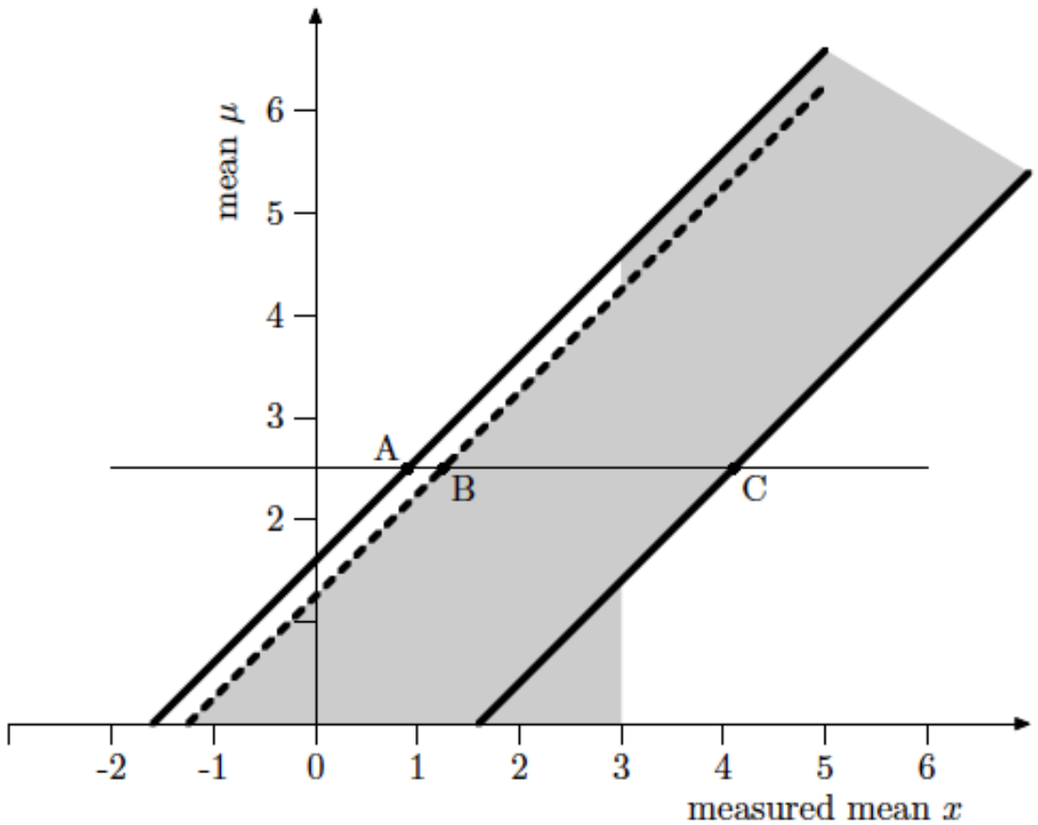
- In HEP in general some wording has been adopted as when to claim discovery.
  - 5σ: Discovery
  - 4σ: Evidence
  - 3σ: Observation

- However, will you believe the same a 4σ astrophysical neutrino as a 4σ faster than speed of light neutrino?

side note: recently BICEP2 claimed a 5.2 σ B-mode signal detection. What is wrong about this statement?

The probability is in fact that the B-mode cannot be a fluctuation of the E-mode (H_0) but there were other null hypotheses not considered (synchrotron, dust) with probabilities of 2.3 σ and 2.2 σ respectively.

Flip-flopping & Empty Intervals.



Flip-flopping for a Gaussian measurement. The shaded area represents the effective confidence belt resulting from choosing to report an upper limit only when the measurement is less than 3σ above zero. This effective belt undercovers for $1.2 < \mu < 4.3$, for example at $\mu = 2.5$ where the intervals AC and $B\infty$ each contain 90% probability but BC contains only 85%.

The unified approach (Feldman-Cousins)

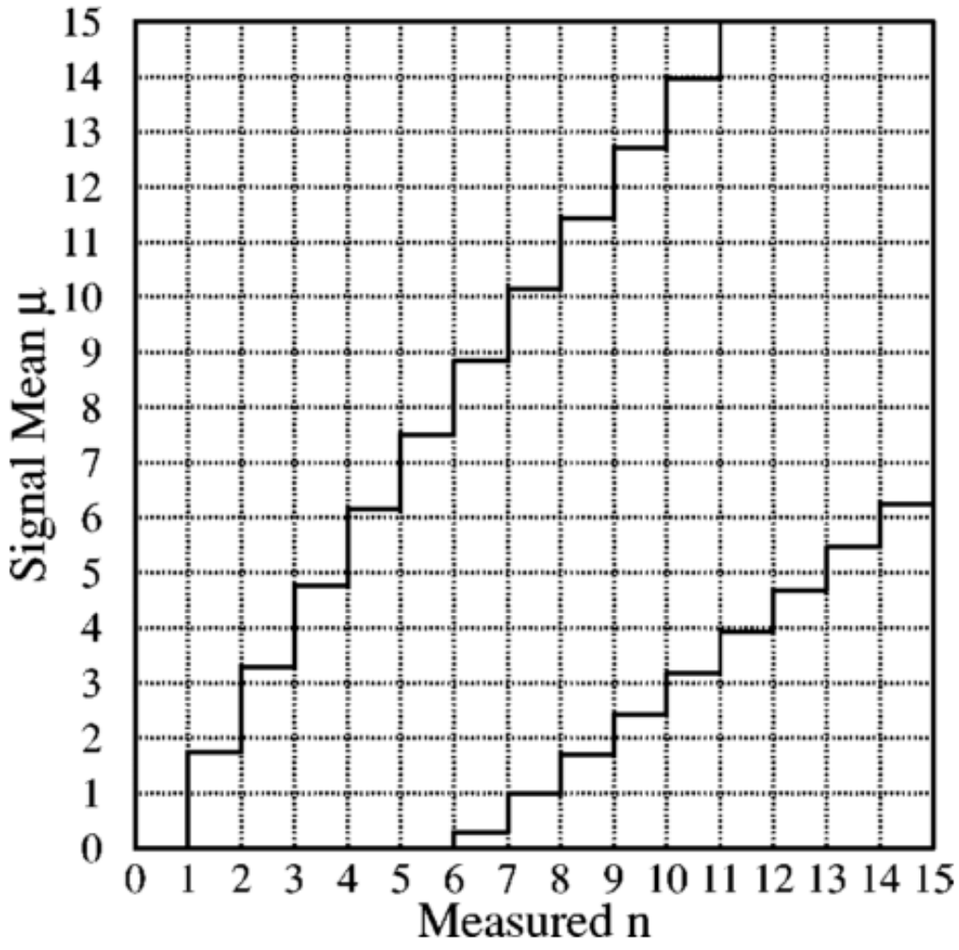
- The elegant way to solve all the problems (flip-flopping and empty intervals) would be to find an ordering principle.

- Feldman and Cousins proposed the likelihood ratio ordering: Feldman and Cousins, Unified Approach to the classical statistical analysis of small signals, Phys. Rev. D 57 (1998) 3873

The unified approach (Feldman-Cousins)

Poisson process with background of 0.3

Neyman classical approach



FC construction for μ = 0.5

n	$P(n\|\mu)$	μ_{best}	$P(n\|\mu_{\text{best}})$	R	rank	U.L.	central
0	0.030	0.0	0.050	0.607	6		
1	0.106	0.0	0.149	0.708	5	√	√
2	0.185	0.0	0.224	0.826	3	√	√
3	0.216	0.0	0.224	0.963	2	√	√
4	0.189	1.0	0.195	0.966	1	√	√
5	0.132	2.0	0.175	0.753	4	√	√
6	0.077	3.0	0.161	0.480	7	√	√
7	0.039	4.0	0.149	0.259		√	√
8	0.017	5.0	0.140	0.121		√	
9	0.007	6.0	0.132	0.050		√	
10	0.002	7.0	0.125	0.018		√	
11	0.001	8.0	0.119	0.006		√	

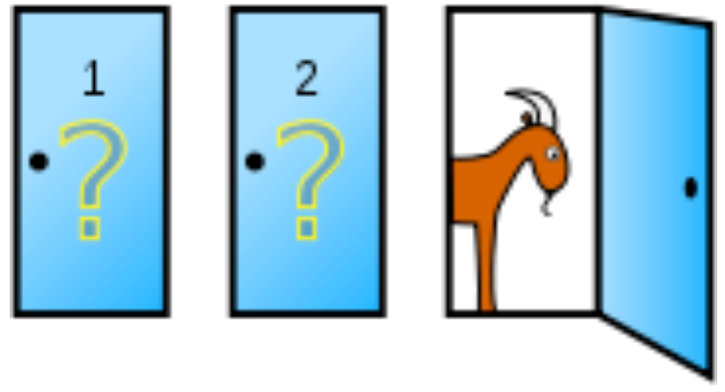
- Exercise: Back to the notebook.

Final words.

- FC construction is an elegant way to solve the problems of flip-flopping and empty intervals based on the likelihood ratio estimate. However, it is very computational intense.

- In point-sources we don’t use the FC construction. In my option, guaranteeing an exact coverage is important when doing precise measurements (ie, mass of particle), but it is less important for astrophysical fluxes (where theories live in an order of magnitude domain).

Monty Hall problem



Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?