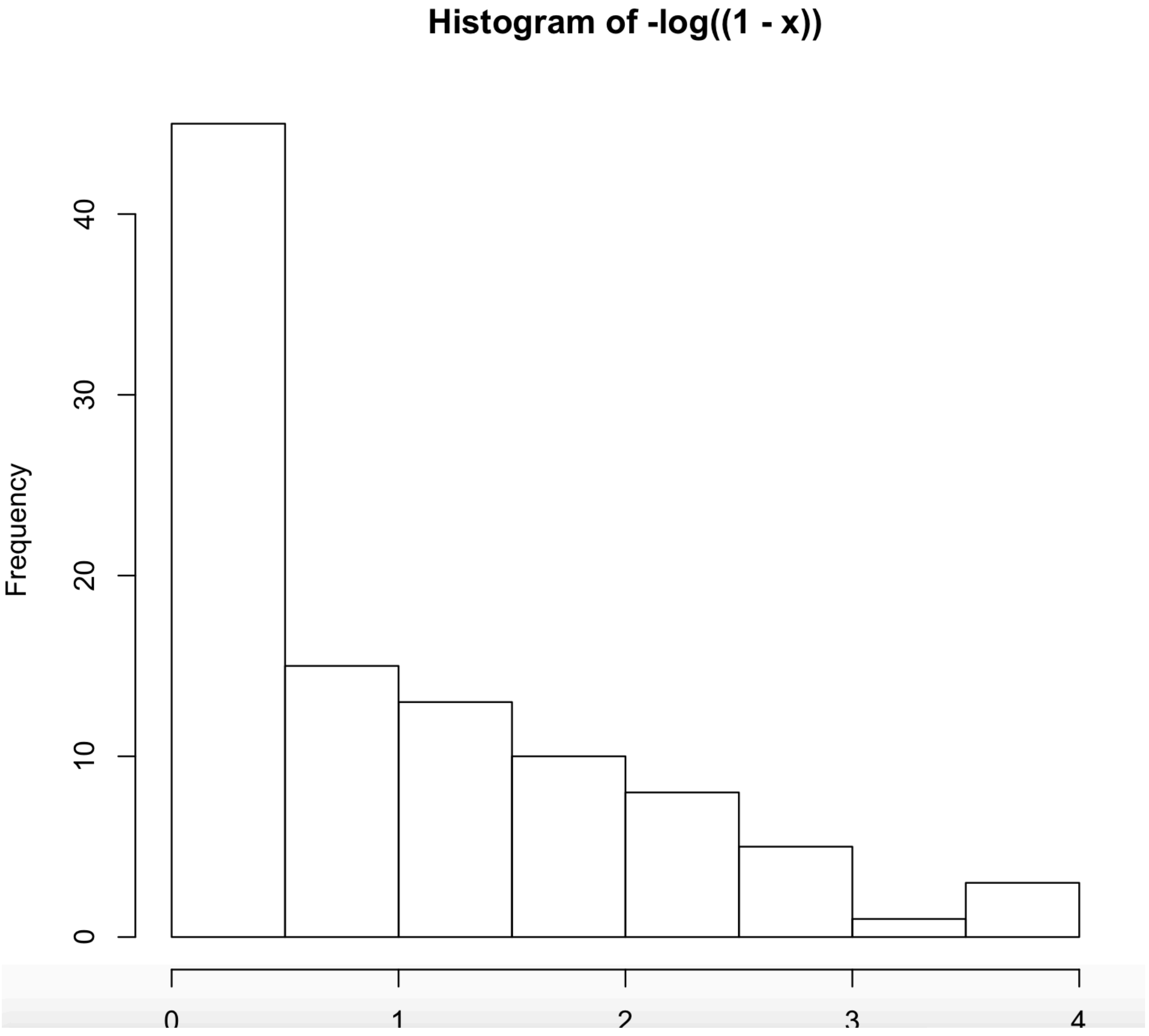
Proposed exercise

Extract N random numbers distributed as an exponential function with lifetime τ

Fill an histogram

Write the likelihood $L(t|\tau)$ in the binned and unbinned cases

```
N=100; x<-runif(N) ; x
  [1] 0.405059710 0.028044254 0.758571449 0.382914253 0.231949128 0.457176317
  [7] 0.736658152 0.038088207 0.104203774 0.513283288 0.742335360 0.368812945
 [13] 0.898926650 0.884993284 0.029905424 0.510855547 0.976764989 0.163296696
 [19] 0.312905139 0.172199152 0.789298260 0.518792378 0.076755612 0.187093519
 [25] 0.613189997 0.007589616 0.476067148 0.091391122 0.254679165 0.642145047
 [31] 0.068187724 0.213190998 0.284391620 0.652574104 0.375936000 0.938753973
 [37] 0.768648992 0.934079373 0.576549295 0.822300084 0.963397188 0.677318145
 [43] 0.804149516 0.278122875 0.918408046 0.161690666 0.816283114 0.219679127
 [49] 0.247514679 0.144359027 0.238819577 0.499138632 0.801599954 0.882881265
 [55] 0.817341159 0.484859340 0.865183191 0.866059658 0.375084123 0.287952191
 [61] 0.832247817 0.392507337 0.292606502 0.018239798 0.980023583 0.892270450
 [67] 0.843237637 0.927634800 0.204098272 0.763523759 0.545941953 0.600462520
 [73] 0.078878091 0.445519178 0.375912647 0.614324038 0.194723071 0.839467755
 [79] 0.265073122 0.870599505 0.696728359 0.085964346 0.004559065 0.710412472
 [85] 0.824518329 0.868817609 0.730170102 0.016328960 0.087571226 0.173662371
 [91] 0.367700928 0.491316323 0.085512807 0.738371863 0.977629644 0.378448315
 [97] 0.194459494 0.754219429 0.376693783 0.939928670
```


Histogram of -log((1 - x))
Frequency
0
10
20
30
40
0
1
2
3
4

Likelihood is NOT a PDF

A Poisson distribution describes
a discrete event count n for
a real valued Meμ.

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

Say, we observe n_o events

What is the likelihood of μ?

The likelihood of μ is given by

$L(\mu) = Pois(n_o | \mu)$

It is a continues function

of μ but it is NOT a PDF

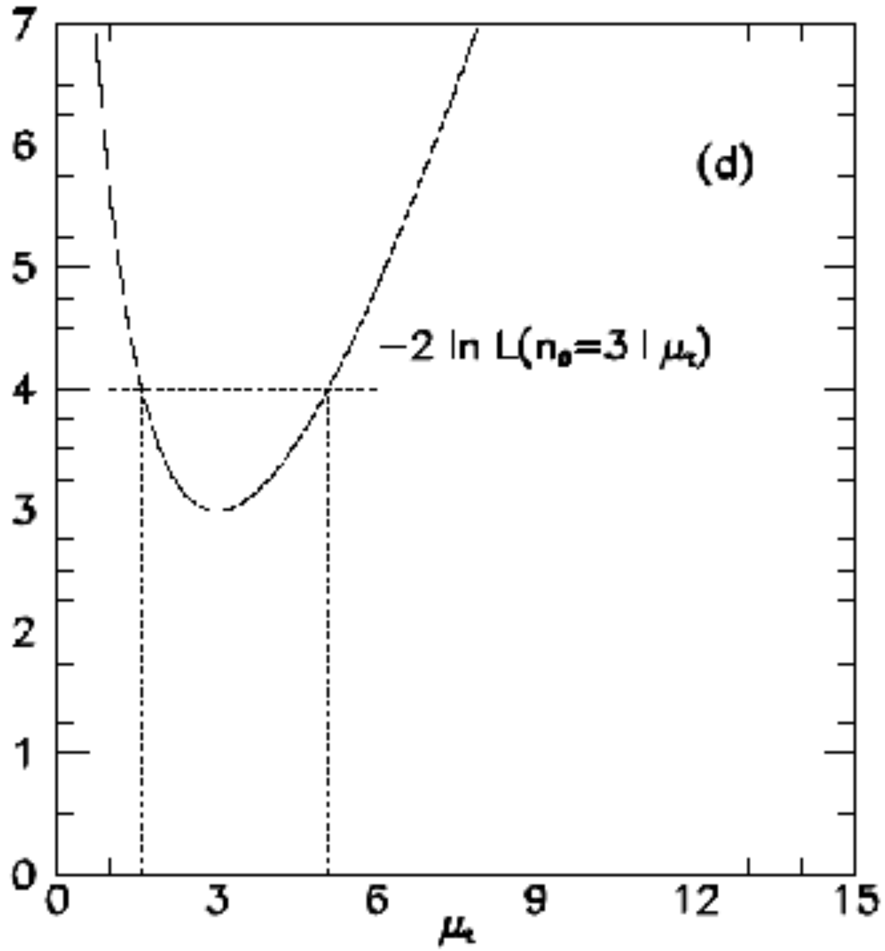


Figure from R. Cousins,
Am. J. Phys. 63 398 (1995)

Testing an Hypothesis (wikipedia...)

- The first step in any hypothesis test is to state the relevant null, H_0 and alternative hypotheses, say, H_1
- The next step is to define a test statistic, q, under the null hypothesis
- Compute from the observations the observed value q_{obs} of the test statistic q.
- Decide (based on q_{obs}) to either
  fail to reject the null hypothesis or
  reject it in favor of an alternative hypothesis
- next: How to construct a test statistic, how to decide?

Basic Definitions: type I-II errors

- By defining α you determine your tolerance towards mistakes… (accepted mistakes frequency)

- type-I error: the probability to reject the tested (null) hypothesis (H_o) when it is true

- $$\alpha = \Pr ob(reject\ H_0 \mid H_0)$$

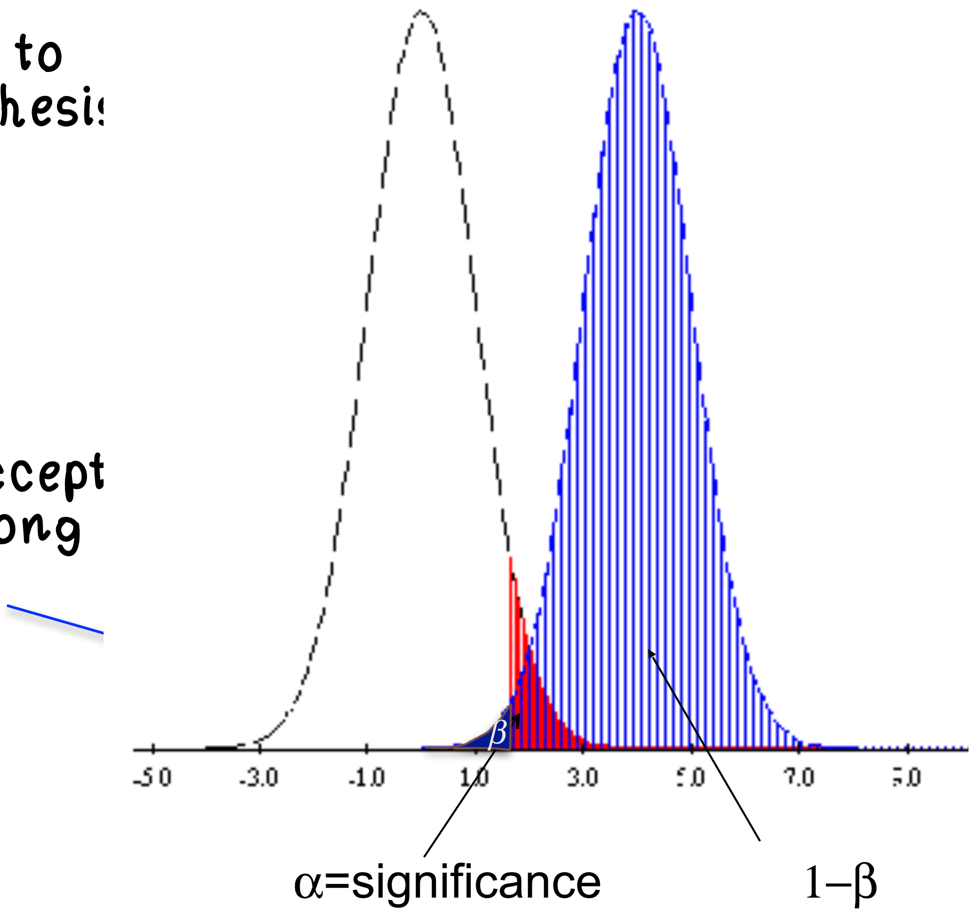
$$\alpha = typeI\ error$$

- Type II: The probability to accept null hypothesis when it is wrong

$$\beta = \Pr ob(accept\ H_0 \mid \bar{H}_0)$$

$$\beta = typeII\ \ error$$

- The pdf of q….



Basic Definitions: POWER

- $\alpha = \Pr ob(reject\ H_0 \mid H_0)$

- The POWER of an hypothesis test is the probability to reject the null hypothesis when it is indeed wrong (the alternate analysis is true)

- $POWER = \Pr ob(reject\ H_0 \mid \bar{H}_0)$

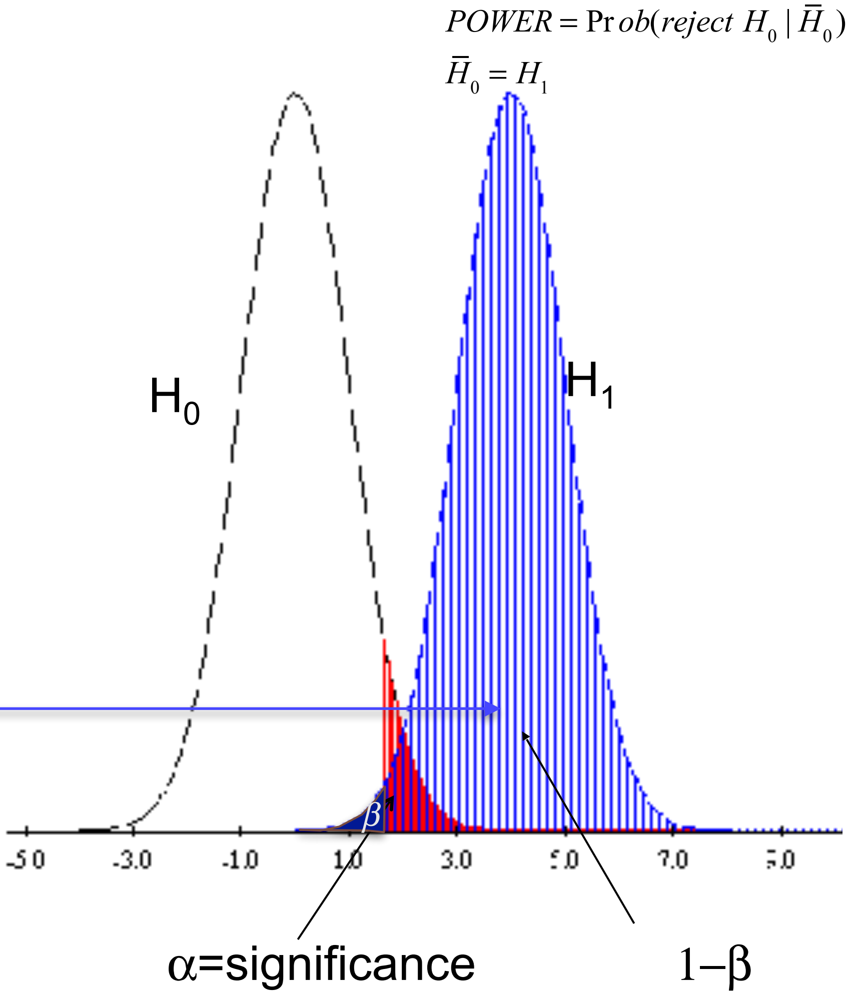
  $\beta = Prob(accept\ H_0 \mid \bar{H}_0)$

  $1-\beta = Prob(reject\ H_0 \mid \bar{H}_0)$

  $\bar{H}_0 = H_1$

  $1-\beta = Prob(reject\ H_0 \mid H_1)$

- The power of a test increases as the rate of type II error decreases



Basic Definitions: POWER

- $\alpha = \Pr ob(rej\,ect\, H_0 \mid H_0)$
- The POWER of an hypothesis test is the probability to reject the null hypothesis when the alternate analysis is true!
- $$POWER = \text{Prob}(reject\, H_0 \mid H_1)$$
  $$\beta = \Pr ob(rej\,ect\, H_1 \mid H_1) \Rightarrow$$
  $$1 - \beta = \Pr ob(accep\,t\, H_1 \mid H_1) \Rightarrow$$
  $$1 - \beta = \Pr ob(rej\,ect\, H_0 \mid H_1) \Rightarrow$$
  $$POWER = 1 - \beta$$
- The power of a test increases as the rate of type II error decreases

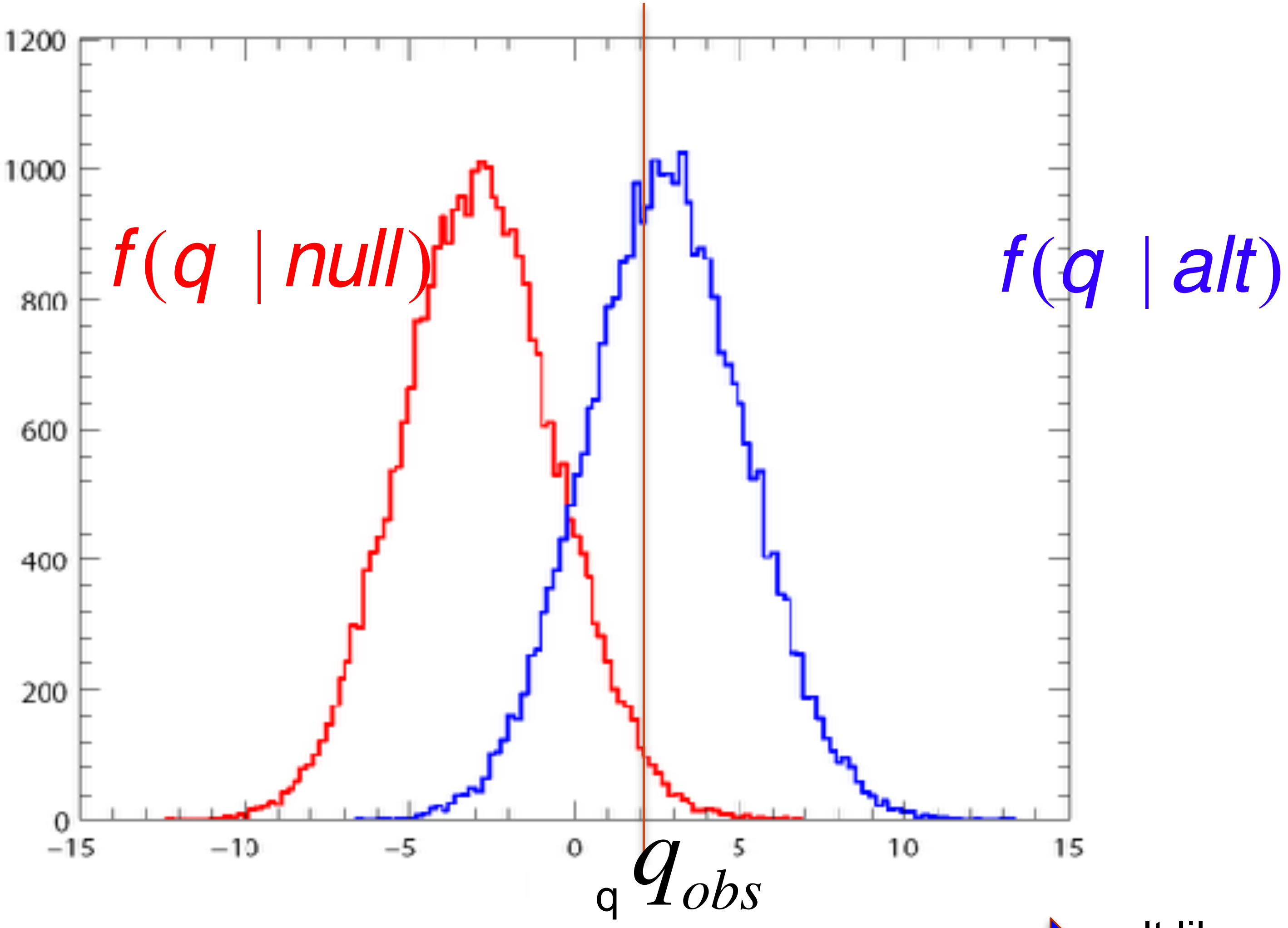
p-Value

- The observed p-value is a measure of the compatibility of the data with the tested hypothesis.

- It is the probability, under assumption of the null hypothesis $H_{null,}$ of finding data of equal or greater incompatibility with the predictions of H_{null}

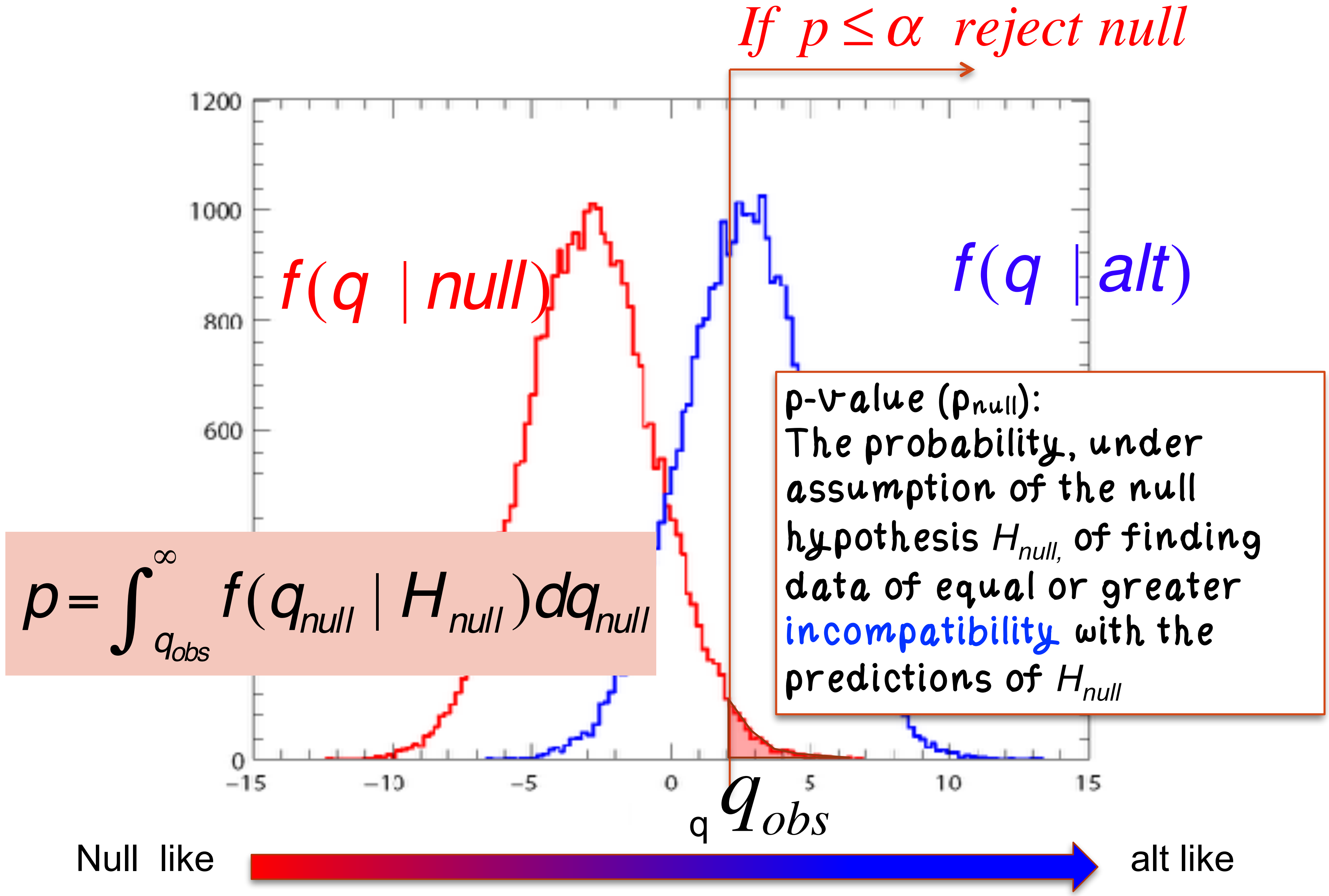
- An important property of a test statistic is that its sampling distribution under the null hypothesis be calculable, either exactly or approximately, which allows p-values to be calculated. (Wiki)

PDF of a test statistic



Null like → alt like

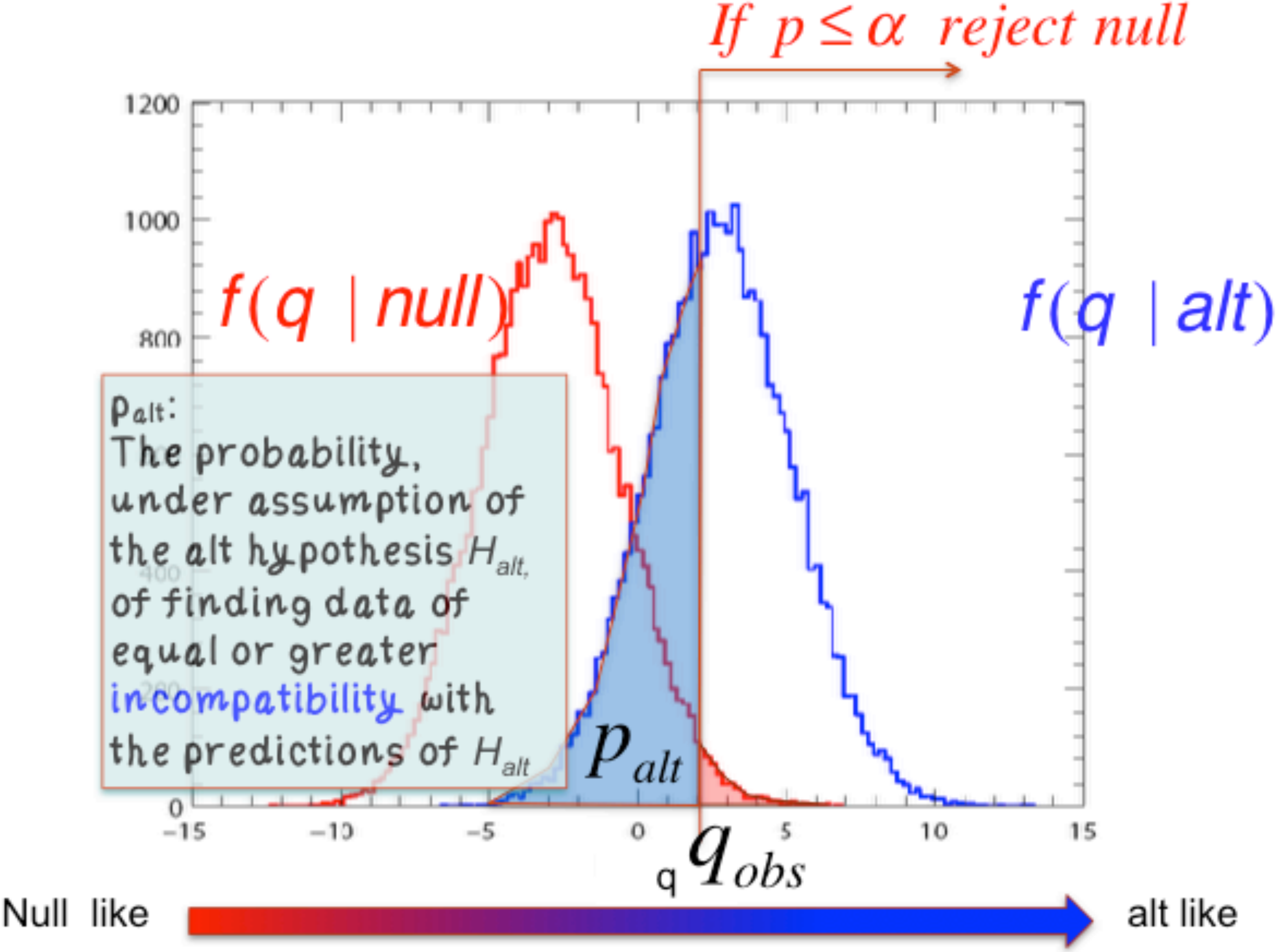
PDF of a test statistic



$$p = \int_{q_{obs}}^{\infty} f(q_{null}\ |\ H_{null})\,dq_{null}$$

p-value (p_{null}):
The probability, under assumption of the null hypothesis $H_{null,}$ of finding data of equal or greater incompatibility with the predictions of H_{null}

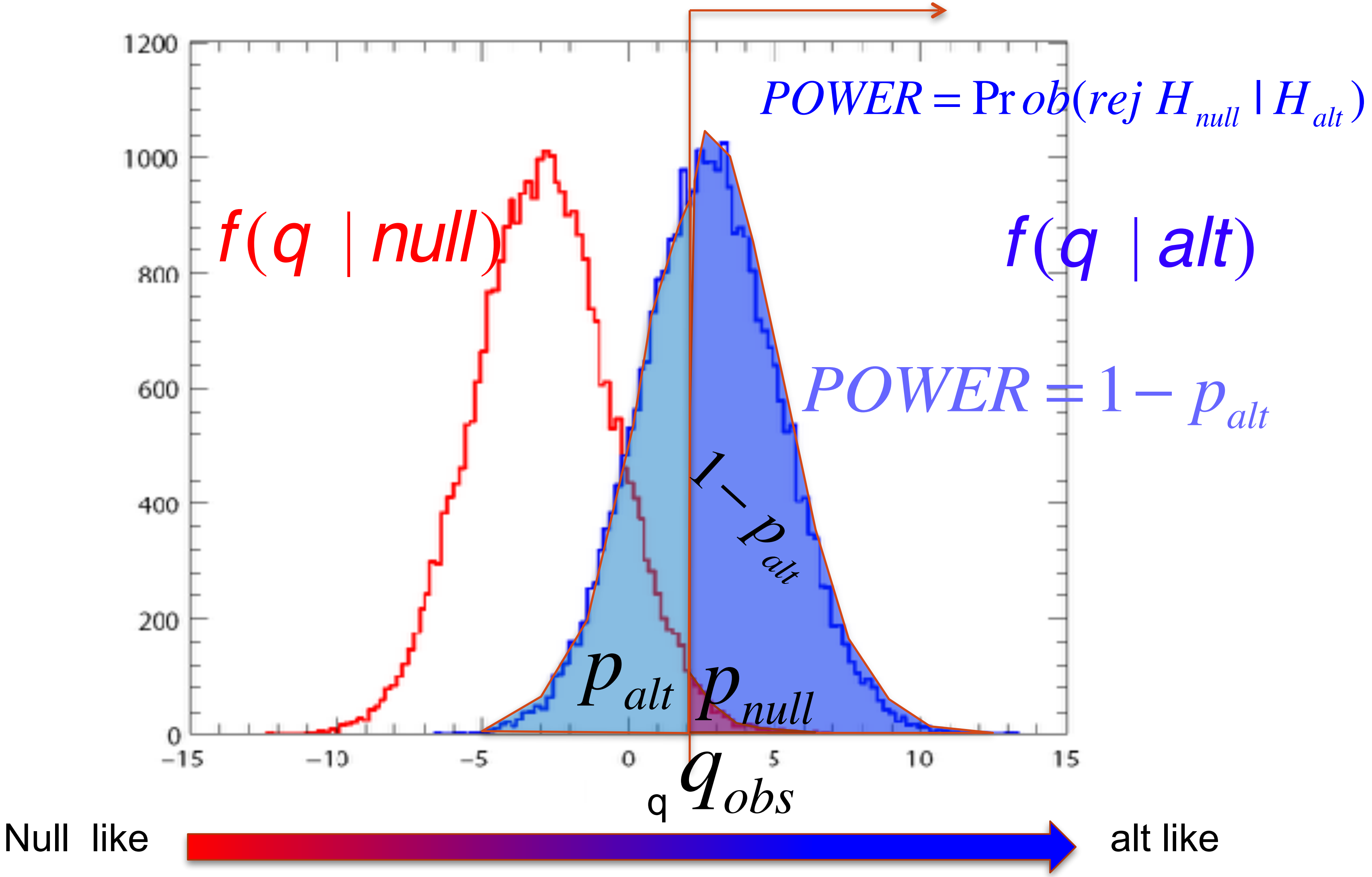
PDF of a test statistic



p_{alt}:
The probability, under assumption of the alt hypothesis $H_{alt,}$ of finding data of equal or greater incompatibility with the predictions of H_{alt}

PDF of a test statistic

$If\ \ p \leq \alpha\ \ reject\ null$



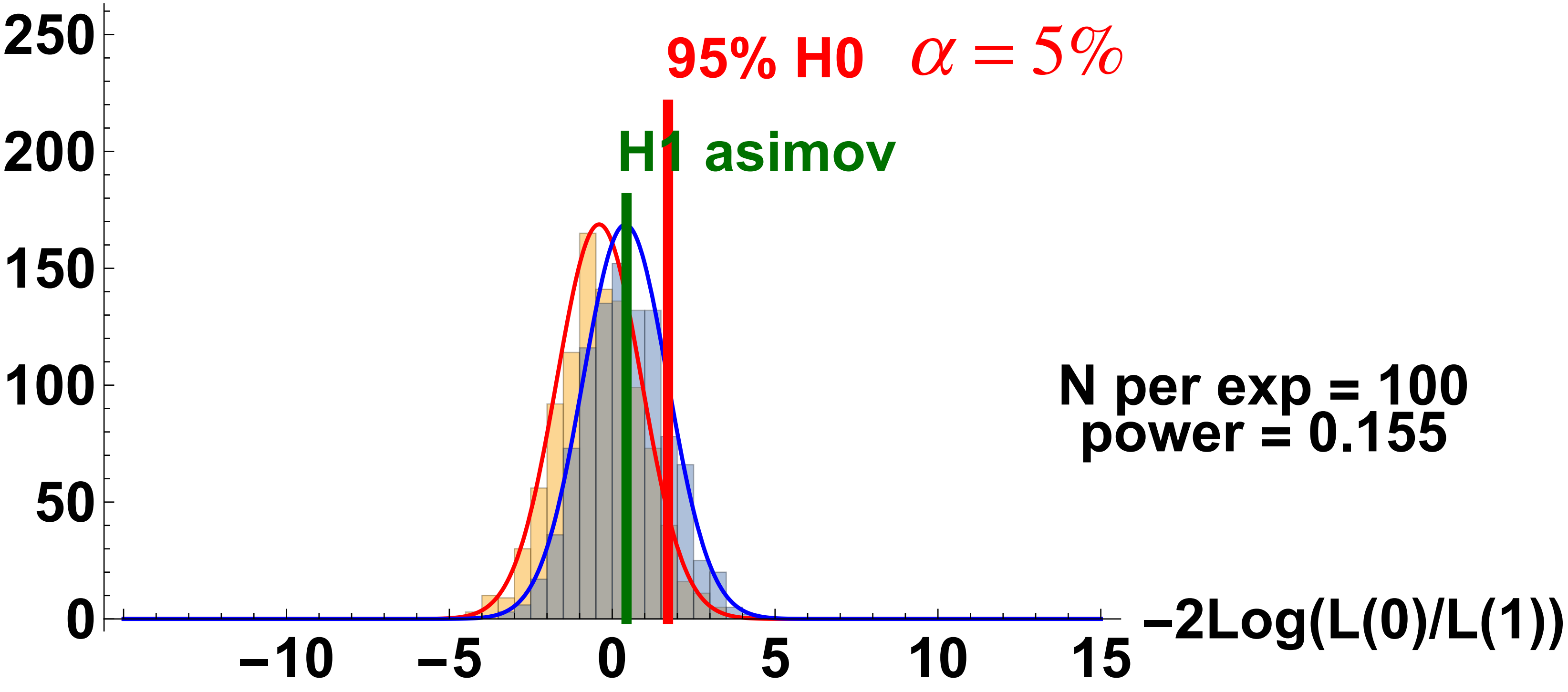
$POWER = \Pr ob(rej\ H_{null} \mid H_{alt})$

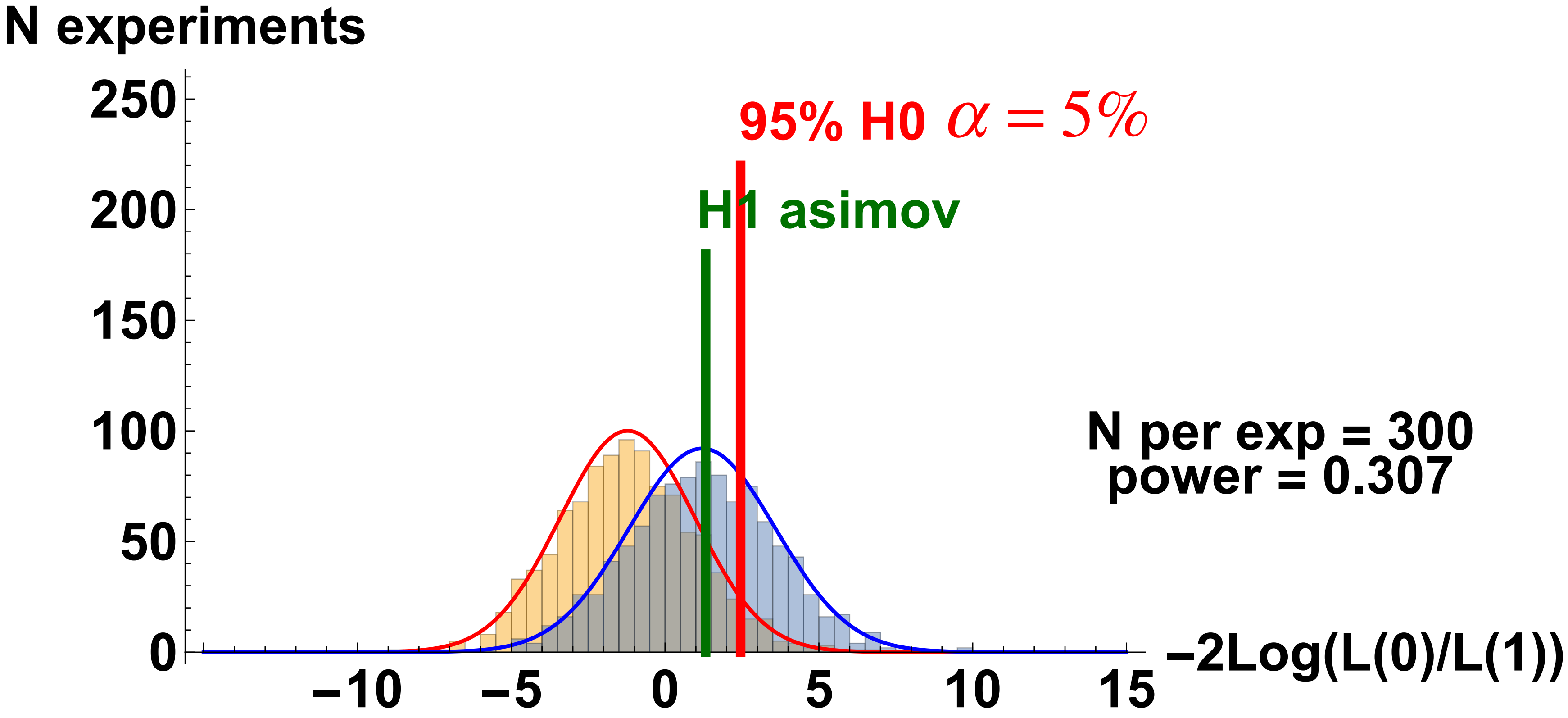
$POWER = 1 - p_{alt}$

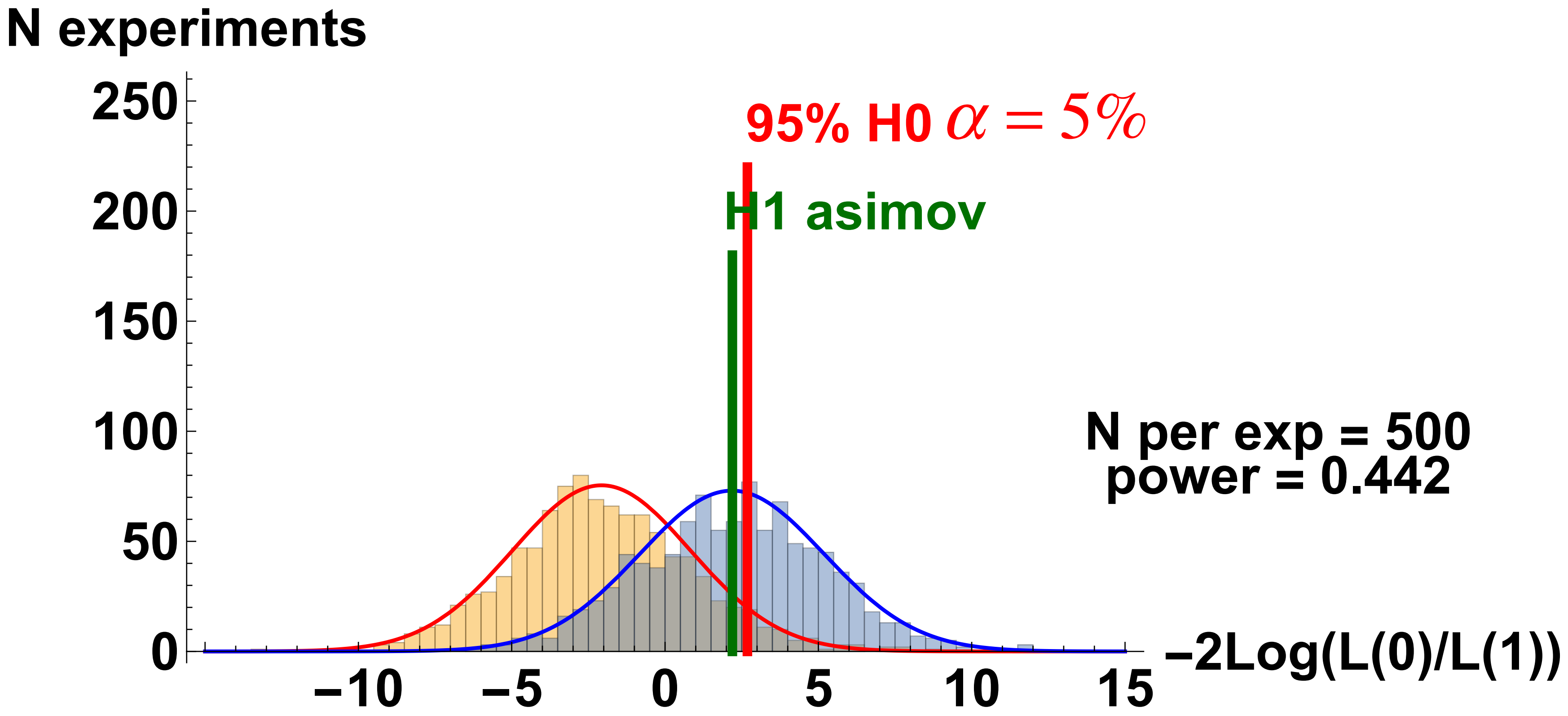
Power and Luminosity

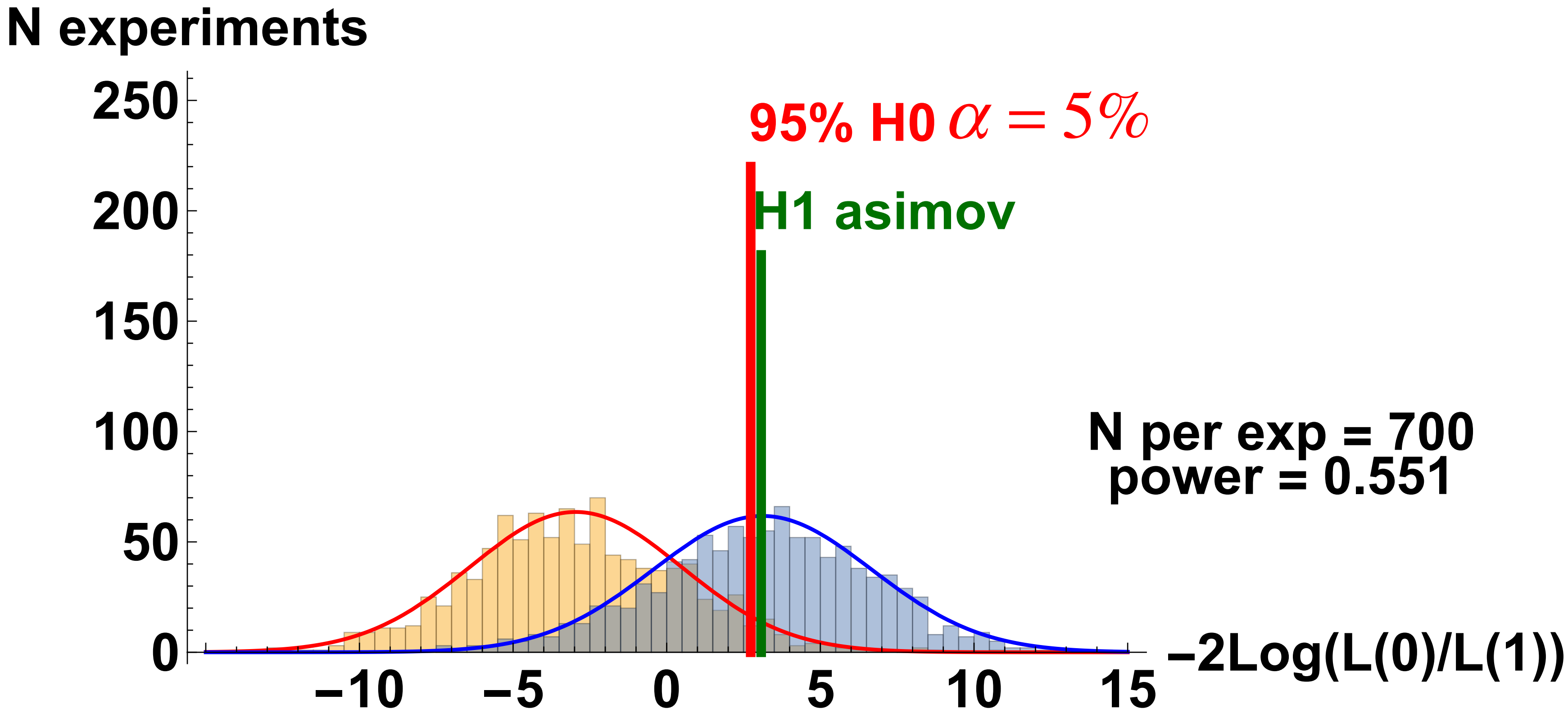
For a given significance the power increases with increased luminosity

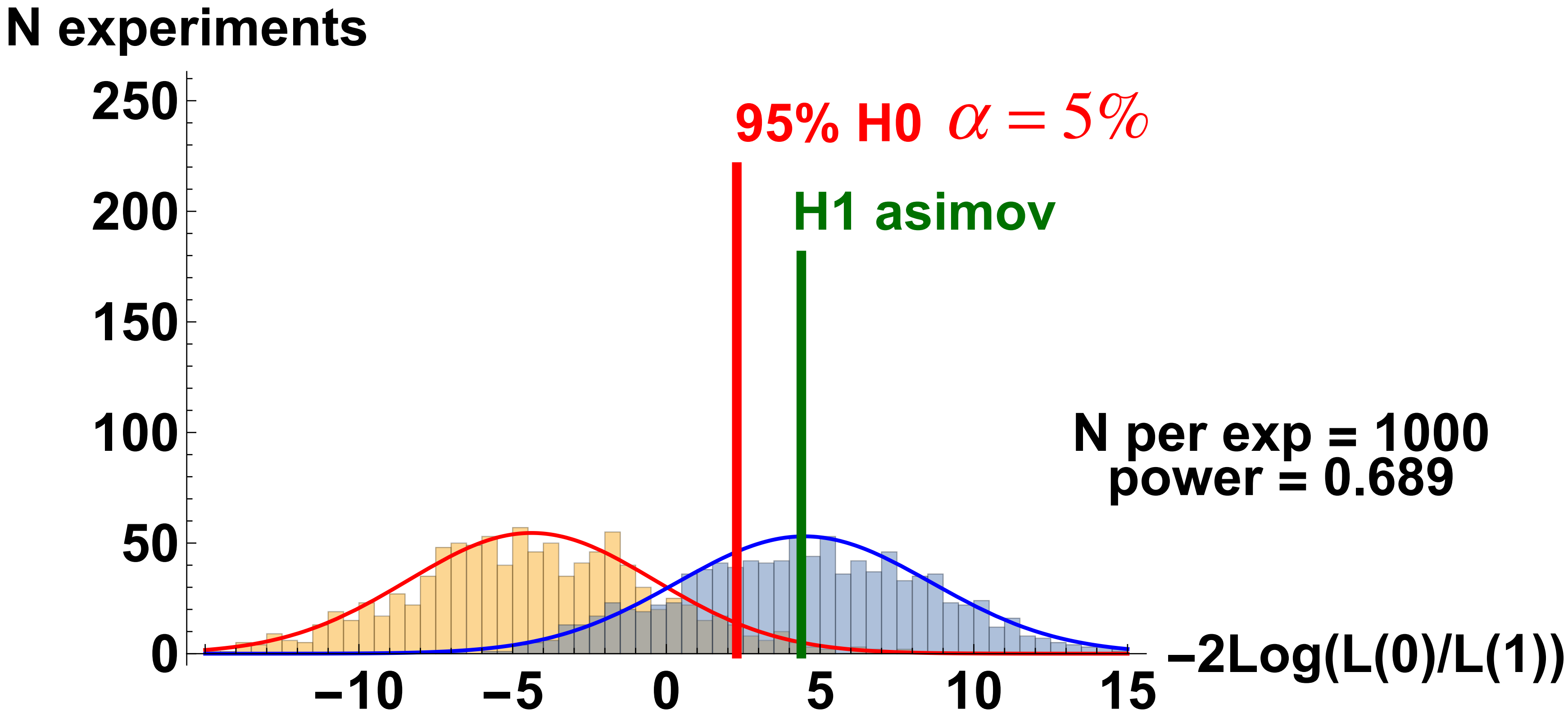
Luminosity ~ Total number of events in an experiment


95% H0
$\alpha = 5\%$
H1 asimov
N per exp = 100
power = 0.155
−2Log(L(0)/L(1))
250
200
150
100
50
0
−10
−5
0
5
10
15


N experiments
250
200
150
100
50
0
−10
−5
0
5
10
15
−2Log(L(0)/L(1))
95% H0 $\alpha = 5\%$
H1 asimov
N per exp = 300
power = 0.307


N experiments
250
200
150
100
50
0
−10
−5
0
5
10
15
−2Log(L(0)/L(1))
95% H0 $\alpha = 5\%$
H1 asimov
N per exp = 500
power = 0.442


N experiments
250
200
150
100
50
0
95% H0 $\alpha = 5\%$
H1 asimov
N per exp = 700
power = 0.551
−10
−5
0
5
10
15
−2Log(L(0)/L(1))


N experiments
250
200
150
100
50
0
95% H0 $\alpha = 5\%$
H1 asimov
N per exp = 1000
power = 0.689
−2Log(L(0)/L(1))
−10
−5
0
5
10
15

Parameter estimation: Maximum likelihood method

Best estimate of parameters to fit theory to data $\hat{\boldsymbol{\theta}}_i$

It is obtained maximizing the likelihood $L(x/\underline{\theta})$ = $L(\underline{\theta})$

We get $t^*(\hat{\underline{\theta}})$

Problem of finding the maxima of a K-dimensional function

- Analytically, by doing the derivatives of the function (of the logarithm of the function to simplify the calculations) with respect to the parameters and putting them equal to 0.

$$\frac{\partial \ln L}{\partial \theta_k} = 0$$

system of M equations with M unknowns

- Numerically, in all cases. The "hystorical" program MINUIT developed at CERN in the '70s is still now the most used package for this kind of problems.

Parameter estimation: Maximum likelihood method

ML estimators properties:

(1) **Unbiasness**: the mean of the estimator should be equal to the ”true” value of the parameter $E[\hat{\theta}] = \theta_{true}$.

(2) **Consistency**: the estimator should converge to the ”true” value once the number of measurements increases $Var[\hat{\theta}] \rightarrow 0$ for $N \rightarrow \infty$.

(3) **Efficiency**: the estimator variance should be the minimum, any other estimator of the same parameter should have a larger variance.

Parameter estimation: Maximum likelihood method

$\hat{\theta}$ Is a random variable with its own pdf's: $\mathrm{E}[\hat{\theta}]\ \mathrm{Var}[\hat{\theta}]$

Central values of the parameter estimation and interval estimation $\hat{\theta} \pm \sigma_{\hat{\theta}}$
(for the moment with probability content in the frequentist approach)

In general maximizing : $\mathrm{L}(x|\underline{\theta})$

We get central values : $\underline{\hat{\theta}}$

with covariance matrix : $V_{jk} = cov[\hat{\theta}_j, \hat{\theta}_k]$

Parameter estimation: Maximum likelihood method

$\hat{\theta}$ Is a random variable with its own pdf's: $\mathrm{E}[\hat{\theta}]\ \mathrm{Var}[\hat{\theta}]$

Central values of the parameter estimation $\hat{\theta} \pm \sigma_{\hat{\theta}}$
and interval estimation
(for the moment with probability content in the frequentist approach)

In general maximizing : $\mathrm{L}(x|\underline{\theta})$

We get central values : $\underline{\hat{\theta}}$

with covariance matrix : $V_{jk} = cov[\hat{\theta}_j, \hat{\theta}_k]$

Parameter estimation: Cramer-Rao inequality

(K=1). The variance of an unbiassed estimator $\hat{\theta}$ obeys the following inequality:

$$Var[\hat{\theta}] \geq \frac{1}{E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

the denominator is also called **Fisher information** factor, and is usually indicated as $I(\theta)$.

(K >1). Given the "Fisher information" matrix

$$I(\underline{\theta})_{jk} = E\left[-\frac{\partial^2 \ln L}{\partial \theta_j \theta_k}\right]$$

each term of the covariance matrix V_{jk} obeys the following inequality

$$V_{jk} \geq I^{-1}(\underline{\theta})_{jk}$$

The Fisher information matrix is also called Hessian matrix of the function L .

The Cramer-Rao inequality states that the inverse of the Fisher information is the minimum variance attainable for an estimator. When the inequality becomes an equality, the estimator is said to be **fully efficient**.

$I^{-1}(\underline{\theta})_{jk}$ is the inverse of the Hessian matrix.

ML Parameter estimators:

Theorems:

- If, for a given parameter, at least a fully efficient estimators exists, such an estimator is the ML estimator.
- For estimators based on a large number of observation $N \to \infty$, ML estimators are fully efficient.
- In case of fully efficient estimators, it is possible to replace the mean of the second derivative with the second derivative evaluated at the estimator central value:

$$E\left[-\frac{\partial^2 \ln L}{\partial\theta^2}\right] = -\left.\frac{\partial^2 \ln L}{\partial\theta^2}\right|_{\theta=\hat{\theta}}$$

The last two theorems are particularly important in practice. Second derivatives evaluated at the central values allow to get the covariance matrix for all ML estimators with a reasonably large number of observations. This method is extensively used to get the covariance matrix of the parameters.

ML Parameter estimators:

1-dimensional example:

f(θ)=-ln L(x|θ)

Taylor expansion around the minimum $\theta = \hat{\theta}$

$$f(\theta) = f(\hat{\theta}) + \left.\frac{df}{d\theta}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2}\left.\frac{d^2 f}{d\theta^2}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

The first order term vanishes, the second order coefficient (~ 1/width of the parabola), according to ML estimators and Cramer-Rao inequality:

$$Var[\hat{\theta}] \geq \frac{1}{E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

$$\left.\frac{d^2 f}{d\theta^2}\right|_{\theta=\hat{\theta}} = \frac{1}{\sigma_{\theta}^2}$$

$$E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] = -\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\hat{\theta}}$$

ML Parameter estimators: profile likelihood method

1-dimensional example:

f(θ)=-ln L(x|θ)

$$f(\theta) = f(\hat{\theta}) + \left.\frac{df}{d\theta}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2}\left.\frac{d^2 f}{d\theta^2}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

$$\left.\frac{d^2 f}{d\theta^2}\right|_{\theta=\hat{\theta}} = \frac{1}{\sigma_\theta^2}$$

f(θ)= $-\ln L_{max} + \frac{1}{2} n^2$

$$\frac{1}{2} n^2 = \frac{1}{2} \frac{(\theta - \hat{\theta})^2}{\sigma_\theta^2}$$

=> Detemination of

$$\hat{\theta} \pm n\sigma_\theta$$

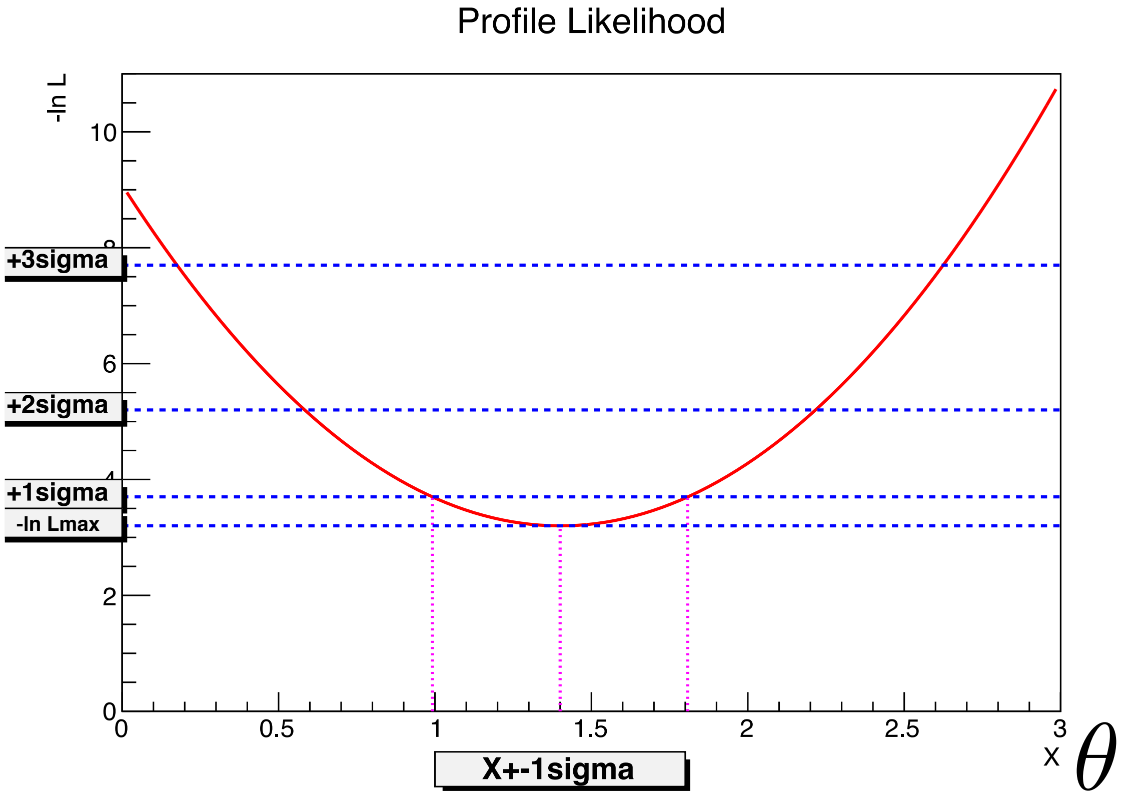


FIGURE 11. Scheme of principle of a profile likelihood method. A $-\ln L$ with parabolic shape is shown for a given variable X. Horizontal lines are shown for $-\ln L_{max} + \frac{1}{2}n^2$ for $n = 0, 1, 2, 3$ and a ± 1 σ is shown for the X variable.

ML Parameter estimators: profile likelihood method

1-dimensional example:

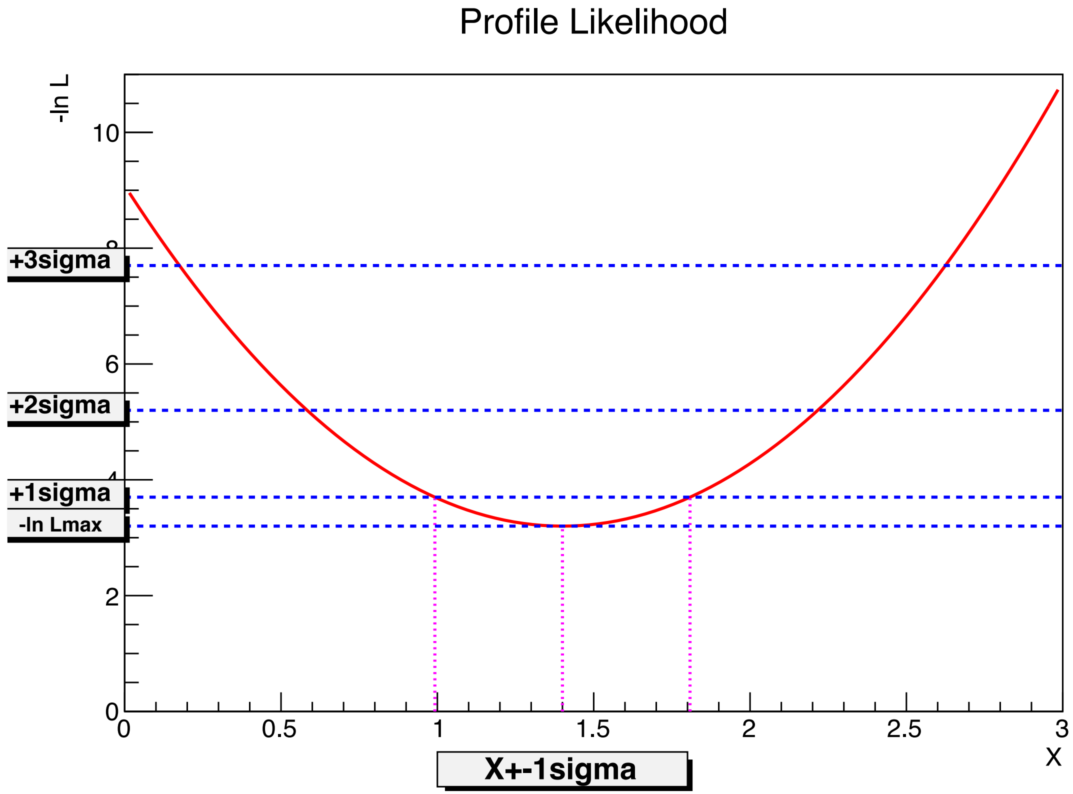


FIGURE 11. Scheme of principle of a profile likelihood method. A $-\ln L$ with parabolic shape is shown for a given variable X. Horizontal lines are shown for $-\ln L_{max} + \frac{1}{2}n^2$ for $n = 0, 1, 2, 3$ and a $\pm\,1\ \sigma$ is shown for the X variable.

$$\hat{\theta} \pm n\sigma_{\theta}$$

$$\theta$$

If 2nd order terms can be neglected => gaussian limit => confidence intervals with gaussian probability content (n=1,2,3 => 68%, 95%, 99.7%)

ML Parameter estimators: profile likelihood method

1-dimensional example:

If we are not in the gaussian limit, the profile likelihood method can be used as well, and the probability content remains to a good approximation the same of the gaussian case. In this case, as shown in the example of fig.12, the intervals can be asymmetric and the result will be written as

$$\hat{\theta}^{+\sigma_\theta^+}_{-\sigma_\theta^-}$$

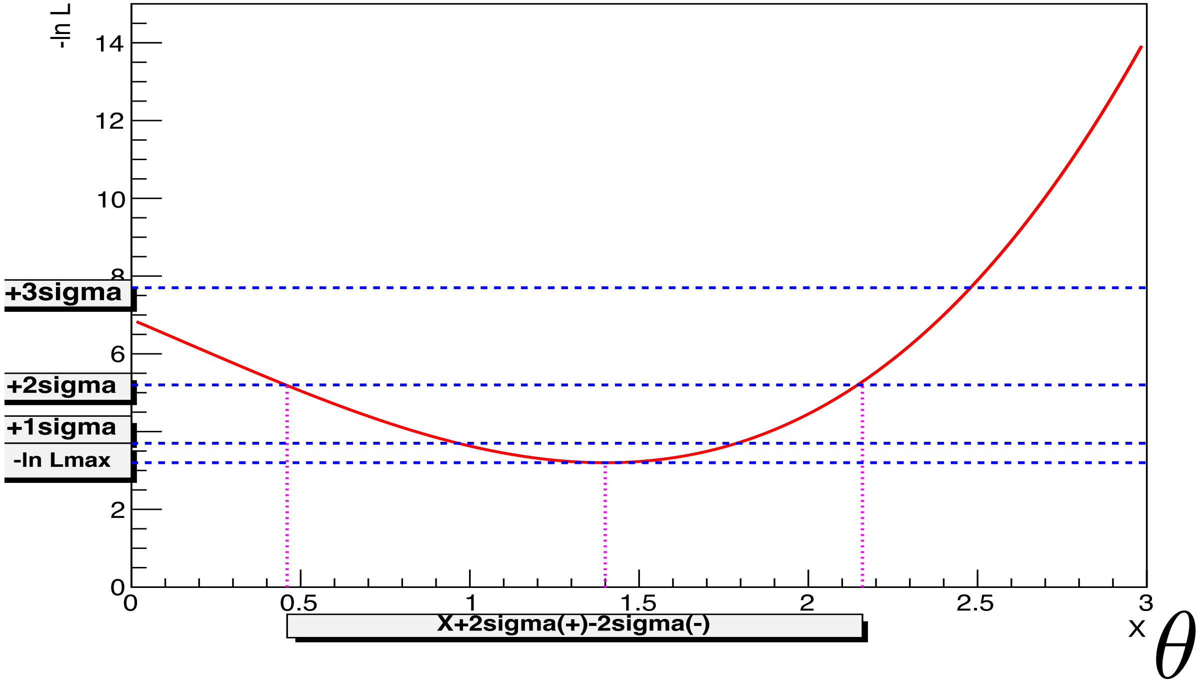
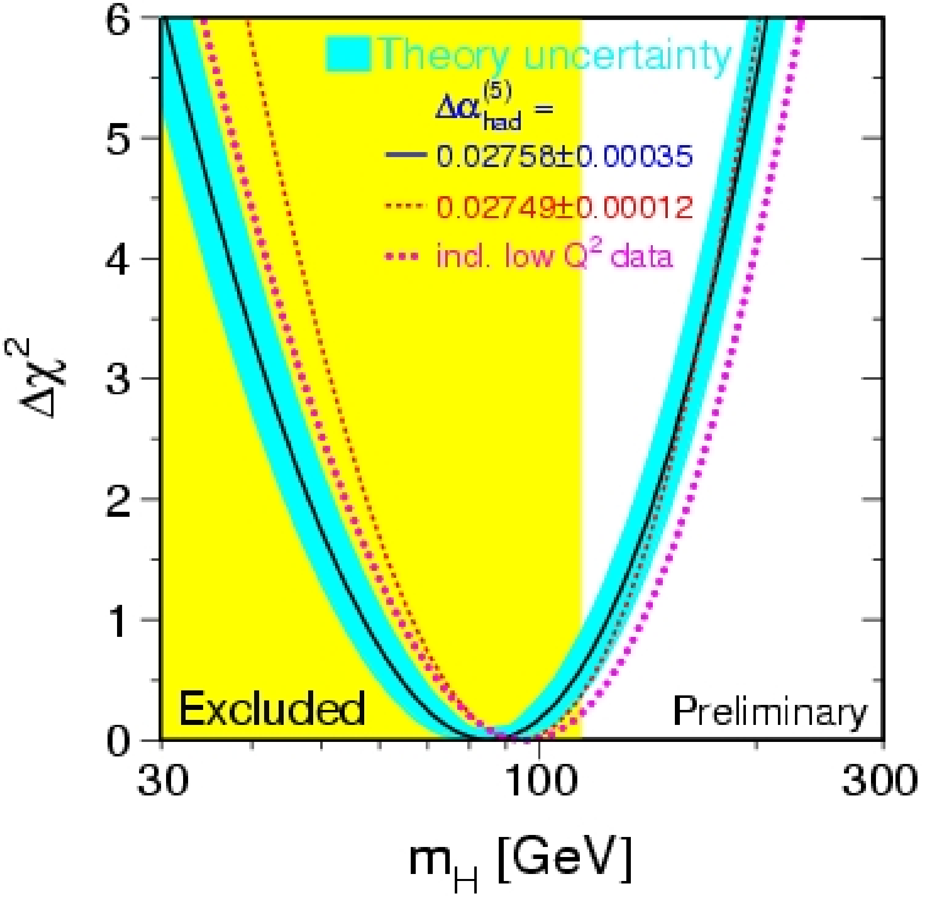


FIGURE 12. Example of a profile likelihood method when $-\ln L$ has not a parabolic shape. As in fig.11, horizontal lines are shown for $-\ln L_{max} + \frac{1}{2}n^2$ for $n = 0, 1, 2, 3$. A "2sigma" interval is shown for X clearly asymmetric.

ML Parameter estimators: profile likelihood method

1-dimensional example:



If the likelihood profile
is far from the parabolic shape
=> far from the gaussian limit

FIGURE 13. 1-dimensional χ^2 of the Standard Model fit to get an interval for the unknown Higgs boson mass. Notice that the horizontal axis is in logarithmic scale, so that the minimum is strongly asymmetric. (very "popular" plot, taken e.g. from www.zfitter.com).

ML Parameter estimators: contour likelihood method

2-dimensional example:

5.5.4. *Contour Likelihood.* The Profile Likelihood method described above can be applied to the single parameter case only. However when $K = 2$ a graphical method is also available providing an interesting insight into the fit result: the so called **contour likelihood method**. The function $-\ln L$ is, in this case, a 2D function $f(\theta_1, \theta_2)$ that, around the minimum $\hat{\theta}_1, \hat{\theta}_2$ has a 2-D paraboloid shape. For a given probability content β, regions S_β can be defined in the $\theta_1 - \theta_2$ plane with the property:

$$p([\theta_1, \theta_2] \subset S_\beta) = \beta \tag{134}$$

that is regions containing the point θ_1, θ_2 with probability β. Such regions can be obtained by intersecting the surface $f(\theta_1, \theta_2)$, with planes of constant $-\ln L$ at values (compare to eq.130)

$$-\ln L_{max} + \Delta \ln L_\beta \tag{135}$$

The equivalent of eq.128 for the two parameters case, is, in the gaussian limit

$$-\ln L = -\ln L_{max} + \frac{1}{2}(\theta - \hat{\theta})^T V^{-1}(\theta - \hat{\theta}) \tag{136}$$

where we have used directly the matrix formalism (T means transposed). By comparing eq.136 with eq.117 we see that $-\ln L + \ln L_{max}$ has a χ^2 distribution with 2 degrees of freedom. This allows to evaluate the values of $\Delta \ln L_\beta$ of eq.135. Table 2 gives the values of $\Delta \ln L_\beta$ for $K = 1$, 2 and 3 for three different values of β. For $K = 3$ or more, the graphical contour representation is not available, but regions S_β can be built with the same method.

ML Parameter estimators: contour likelihood method

2-dimensional example:

TABLE 2. For 3 different values of probability levels (corresponding to the usual 1,2 and 3 gaussian std.deviations) the values of $\Delta \ln L_\beta$ are given for one-parameter (K=1) and two or three-parameters fits.

β (%)	$2\Delta \ln L_\beta$ (K=1)	$2\Delta \ln L_\beta$ (K=2)	$2\Delta \ln L_\beta$ (K=3)
68.3	1	2.30	3.53
95.4	4	6.18	8.03
99.7	9	11.83	14.16

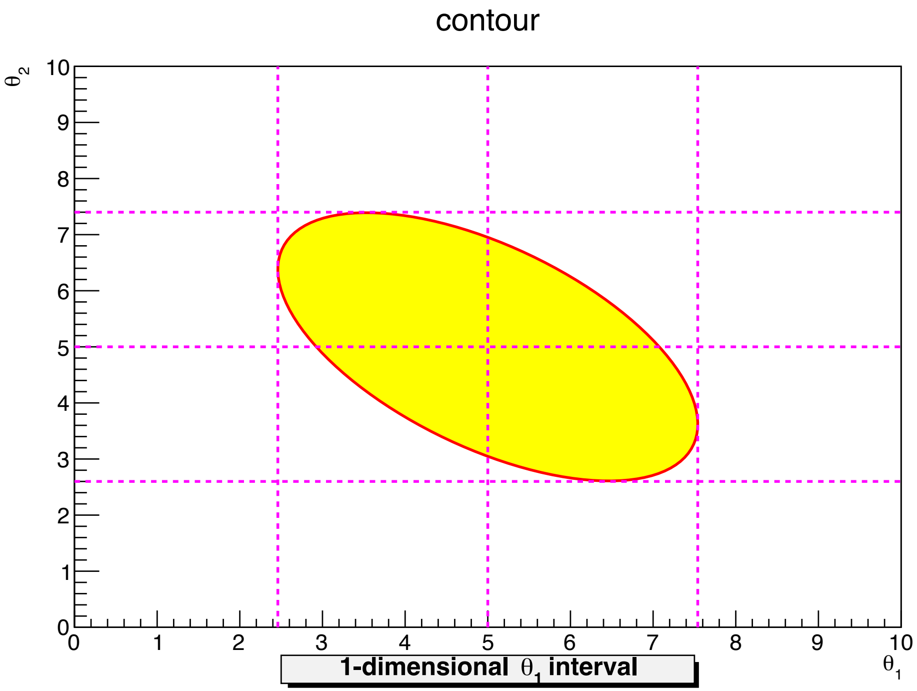


FIGURE 14. Contour plot of two correlated parameters in the gaussian limit. The ellipse shown in yellow, is the S_β region described in the text. The horizontal and vertical bands allow to get 1-dimensional intervals for the two variables. The probability contents of these intervals is different from β.

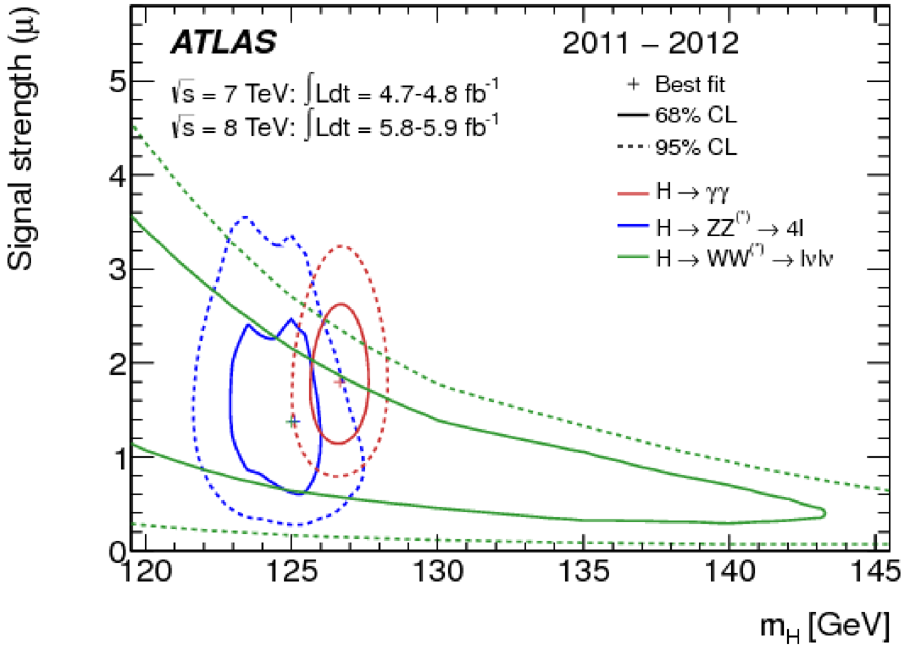


FIGURE 15. From the ATLAS experiment. Results of the fits of 3 different Higgs decay channels (namely $\gamma\gamma$, ZZ and WW) in a 2-dimensional plane, mass vs. signal strength. For each fit, both 68% and 95% probability regions are shown. Notice that in all the cases apart from the $\gamma\gamma$, we are very far from the gaussian limit. (taken from ATLAS collaboration, Phys.Lett. B716 (2012) 1-29).