Statistical Data Analysis 2

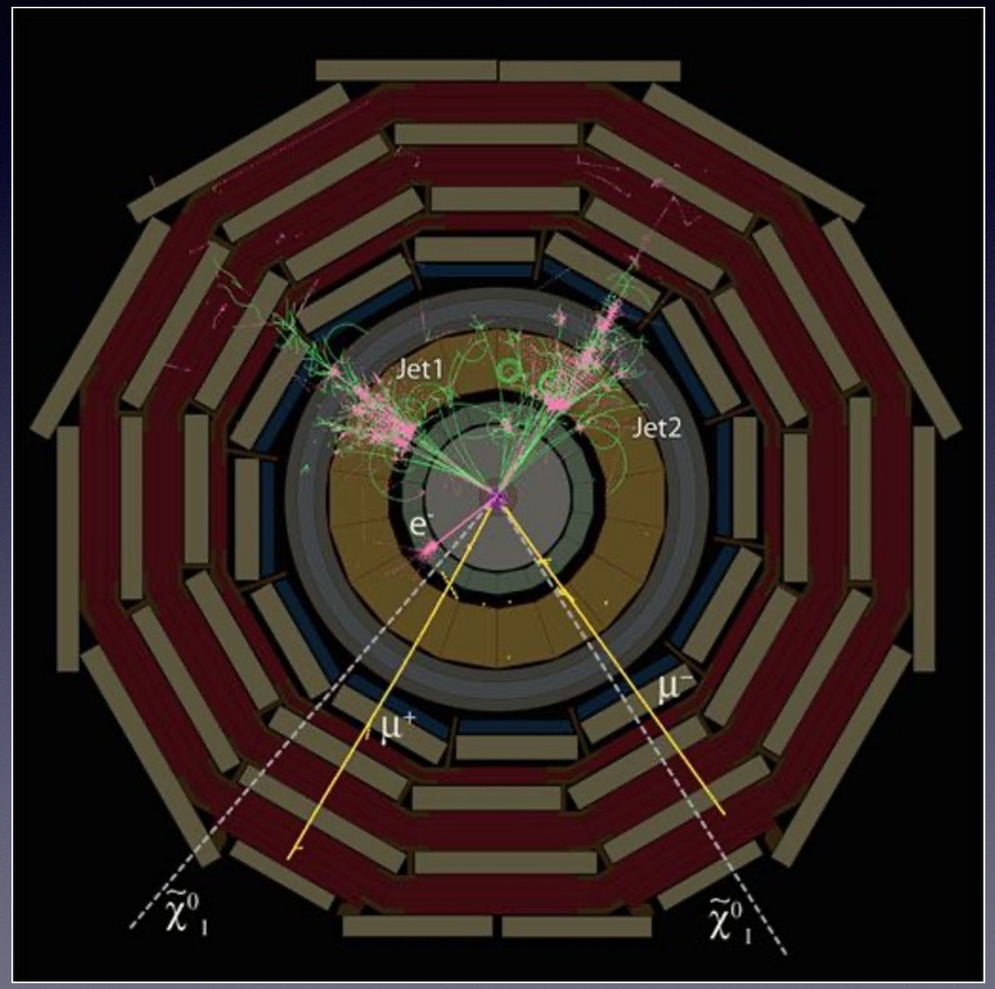
Tae Jeong Kim
Hanyang University

For 32nd Vietnam School of Physics:
Particle Physics and Dark Matter
July 18, 2026

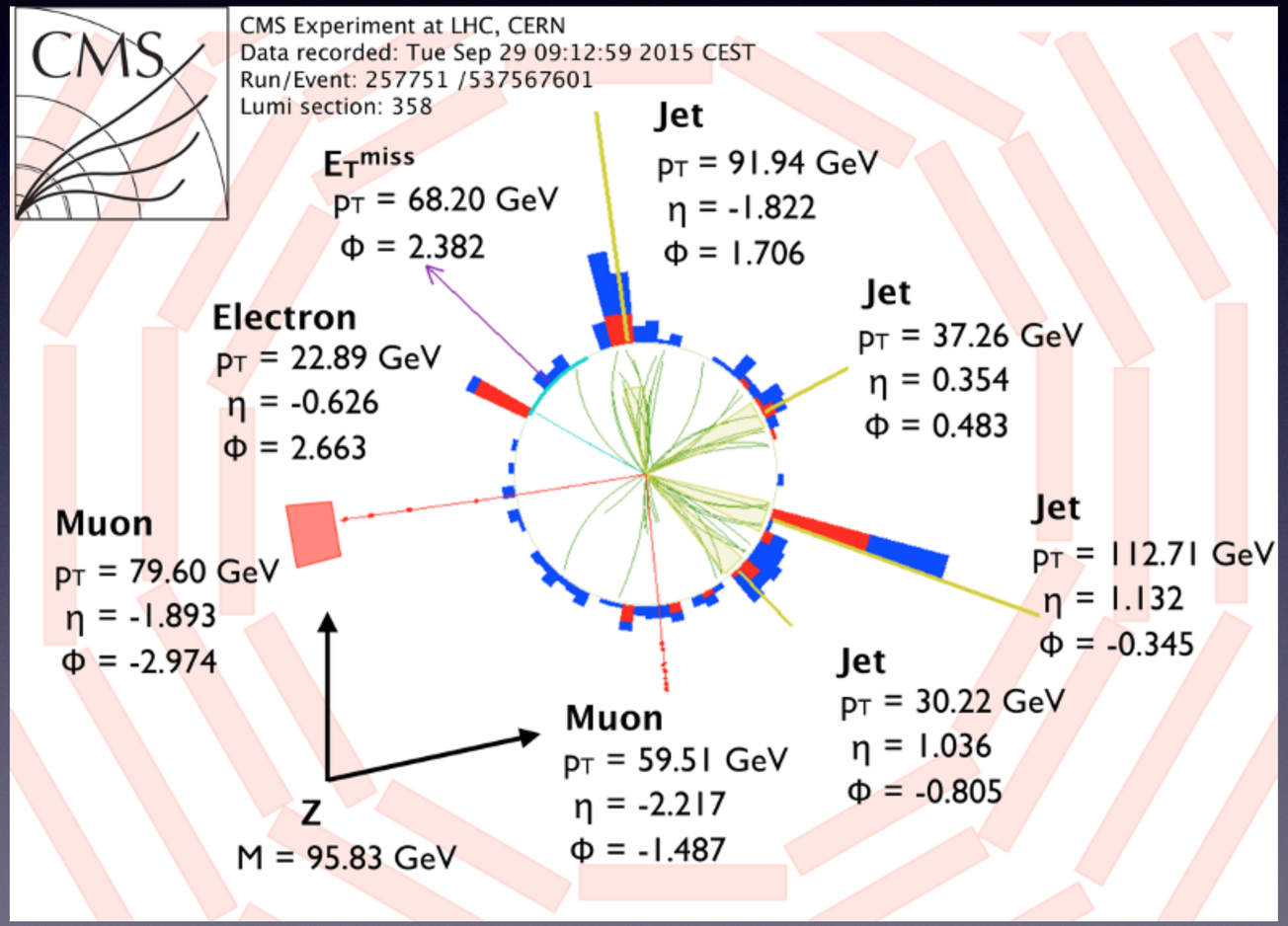
Disclaimer

- Materials for this slide are mainly from the book of “Statistical Data Analysis” by Glen Cowan

A simulated SUSY event



Background event



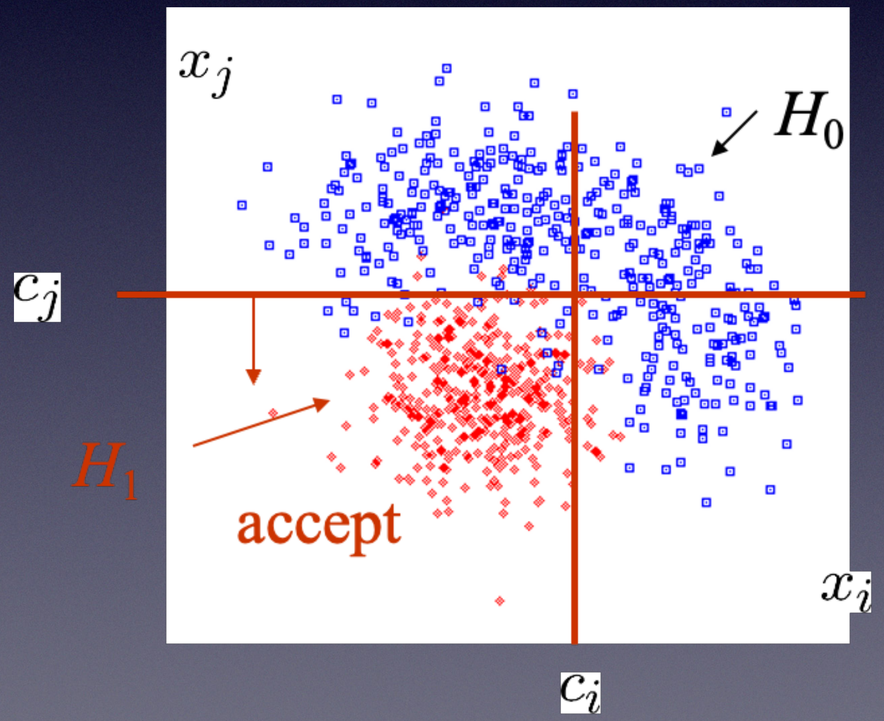
The event from the SM ttbar production associated with Z has also high pT jets and muons, electrons and some missing transverse energy

→

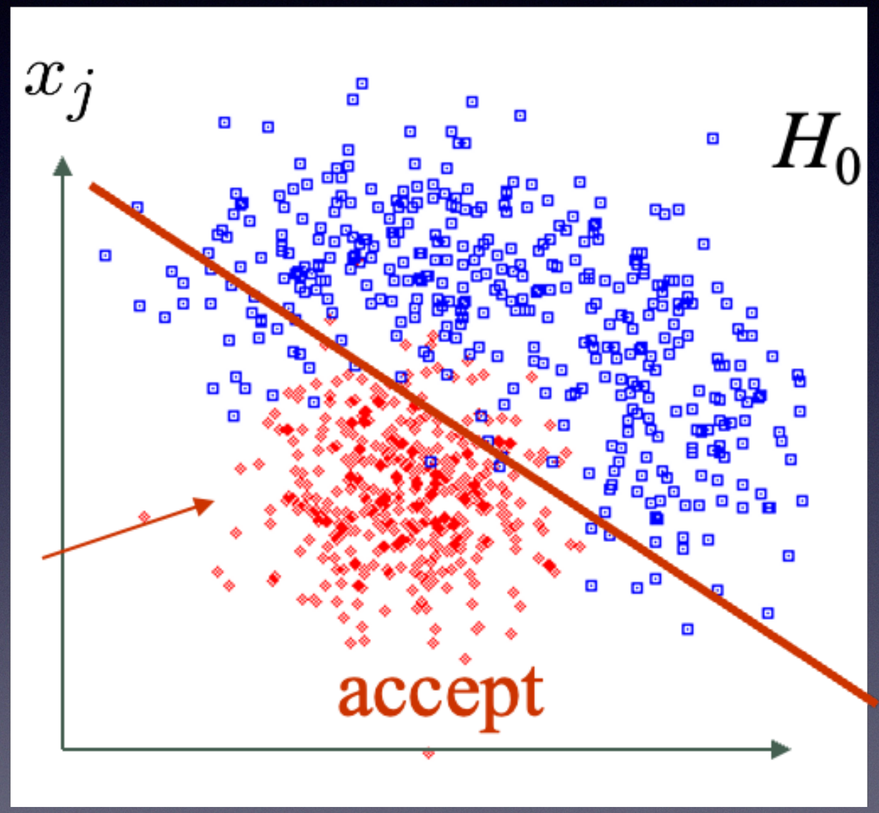
can easily mimic a SUSY event

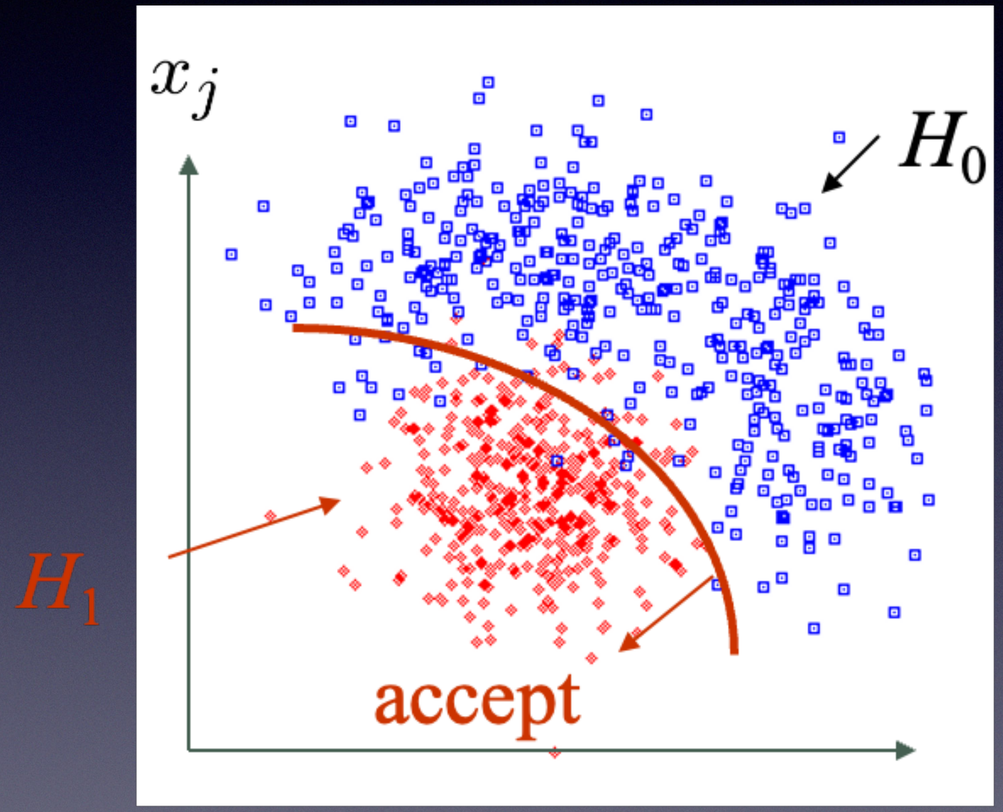
Selecting events

- An individual event is a collection of numbers
  - x_1 = number of muons
  - x_2 = mean p_T of jets
  - x_3 = missing energy
- H_0 is the background hypothesis (we want to reject) and H_1 is the signal hypothesis (event type we want to select)



Other ways to select events





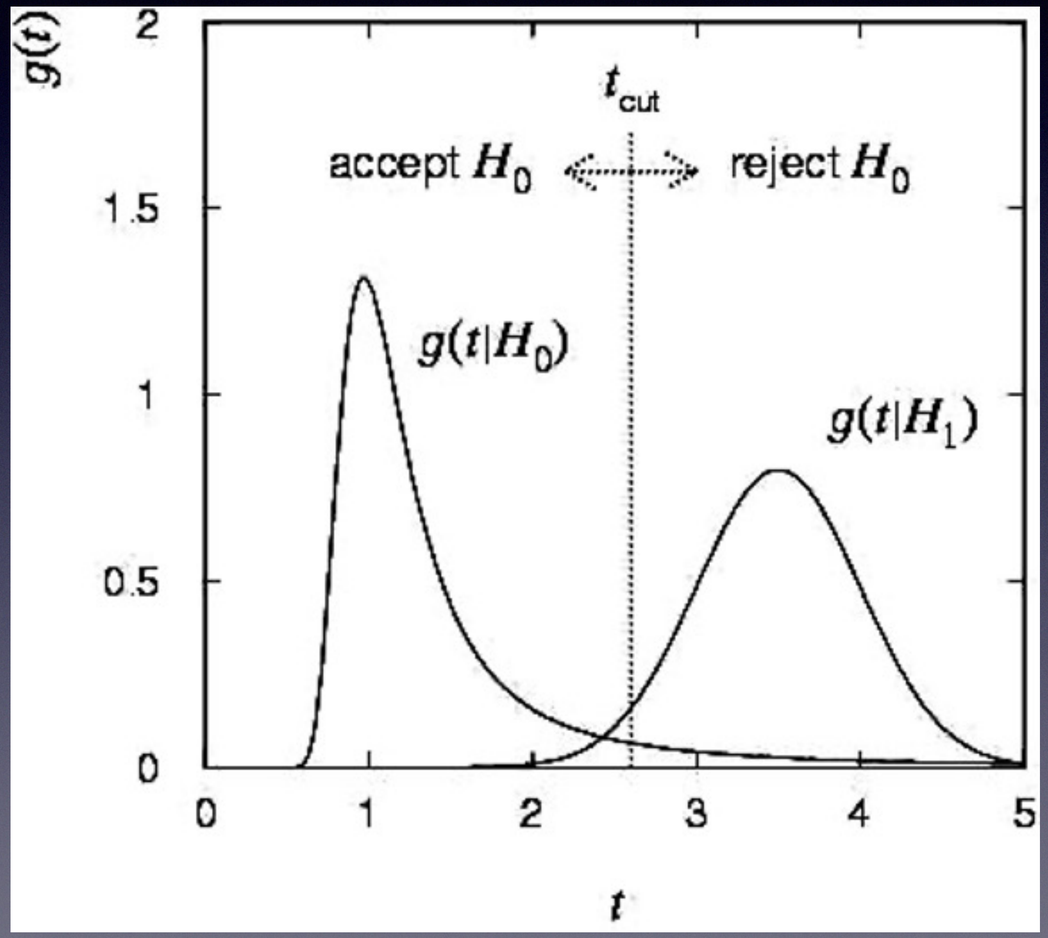
How can we do this in an ‘optimal’ way?

Test statistics

- The decision boundary can be defined by an equation of the form

$t(x_1, \ldots, x_n) = t_{cut}$ ← test statistics

- Now the decision boundary is a single cut on t
- If the data falls in the critical region, we reject H_0



Type I and Type II errors

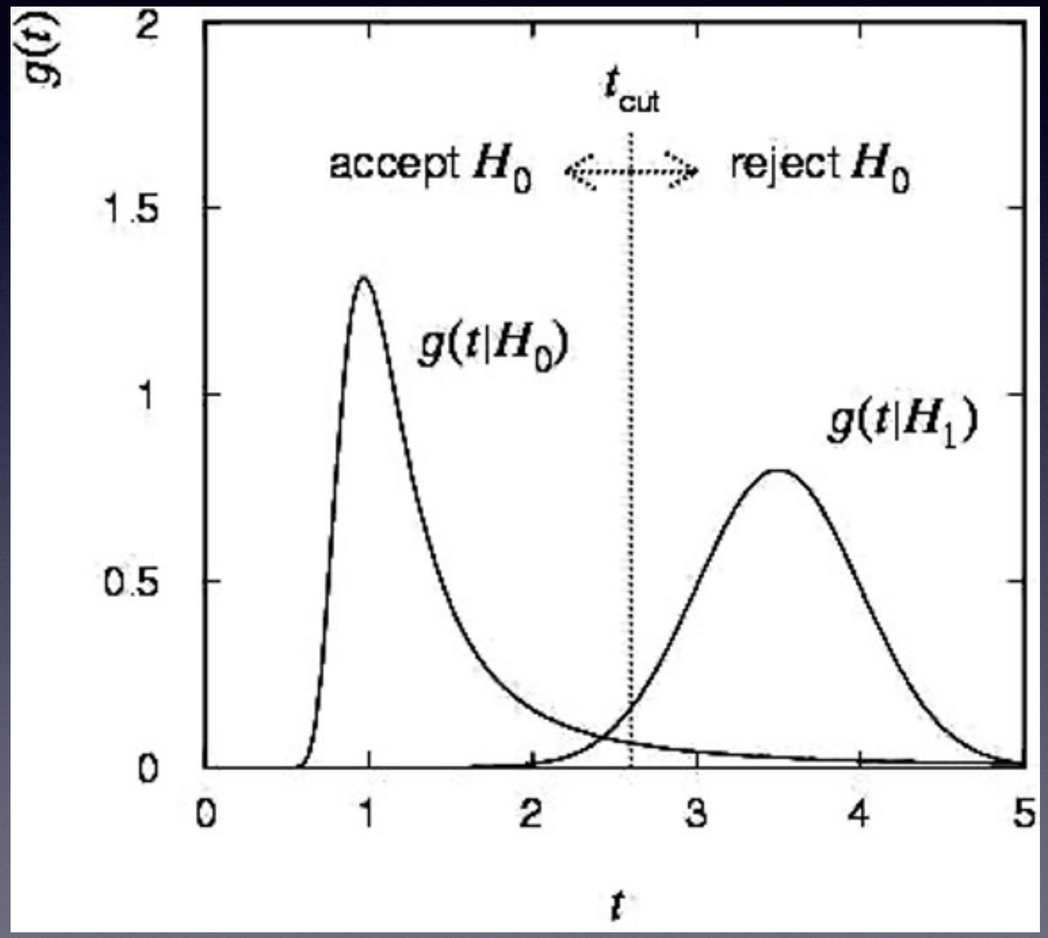
- Type I error: The null hypothesis (H_0) has been rejected while it was actually true
- Type II error: One has failed to reject the null hypothesis (H_0) while it was actually false

	H_0 is true	H_0 is false
Rejected H_0	Type I error	Correct decision
Did not reject H_0	Correct decision	Type II error

Signal efficiency

- Probability of accepting the signal event

$$\epsilon_s = \int_{t_{cut}}^{\infty} g(t \mid s) dt$$



Purity of event selection

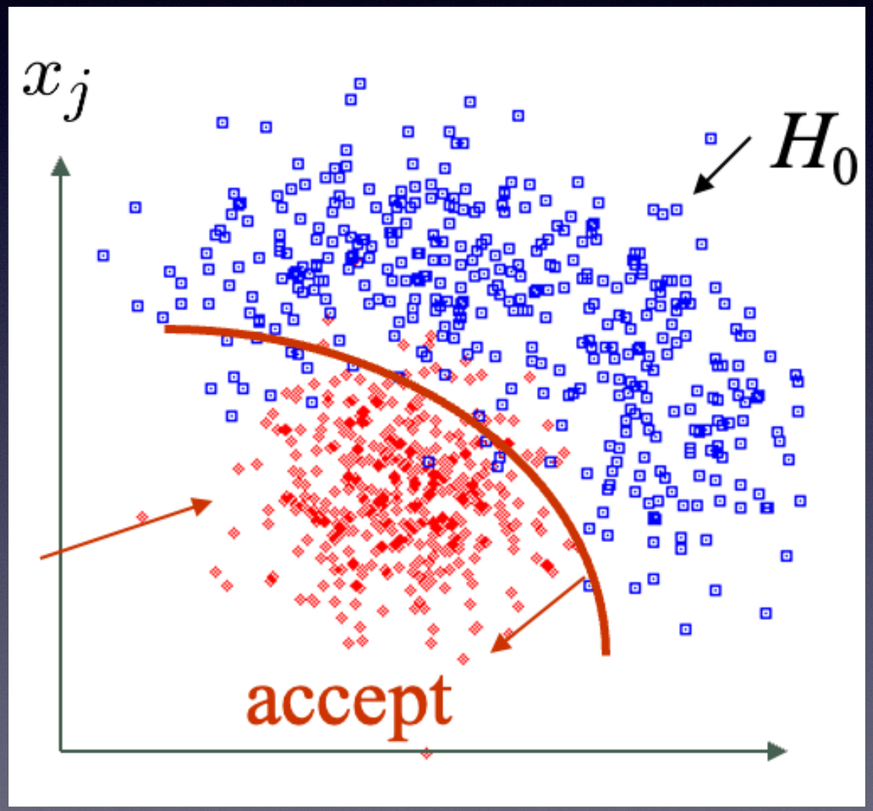
- Suppose only one background type b, the cross sections of the signal and background events are σ_s and σ_b (prior probabilities)
- Purity means the probability to be signal given that the event was accepted. Using Bayes' theorem, we find

$$P(s \mid t_{cut}) = \frac{P(t > t_{cut} \mid s)\sigma_s}{P(t > t_{cut} \mid s)\sigma_s + P(t > t_{cut} \mid b)\sigma_b} = \frac{\epsilon_s \sigma}{\epsilon_s \sigma_s + \epsilon_b \sigma_b}$$

- So the purity depends on the prior probabilities as well as on the signal and background efficiencies

Non-linear test statistics

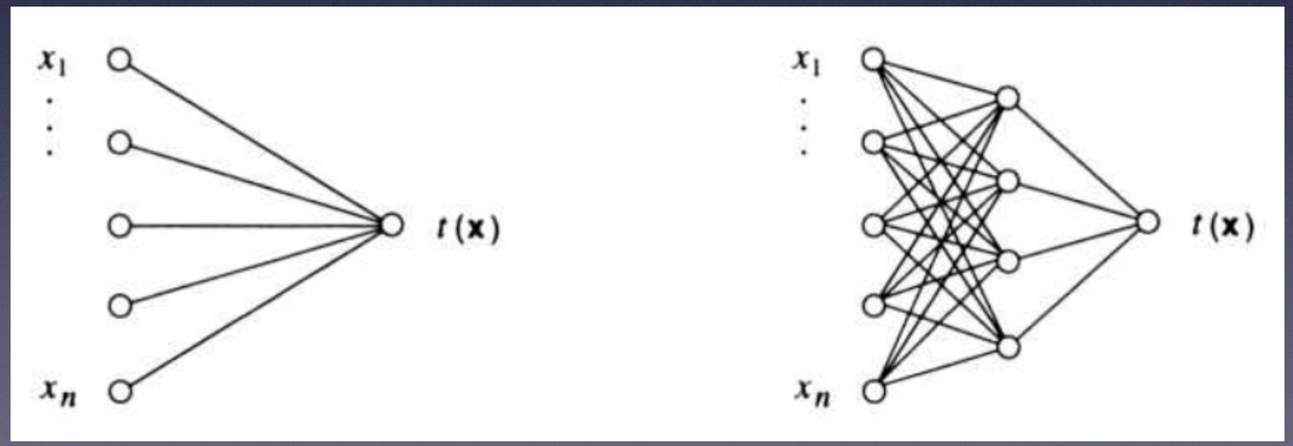
- Multivariate statistical method
  - Neural networks
  - Boosted decision trees
  - Deep neural network
- TMVA, Tensorflow, Keras,….



Artificial Neural Network

$$t(\mathbf{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i x_i\right)$$

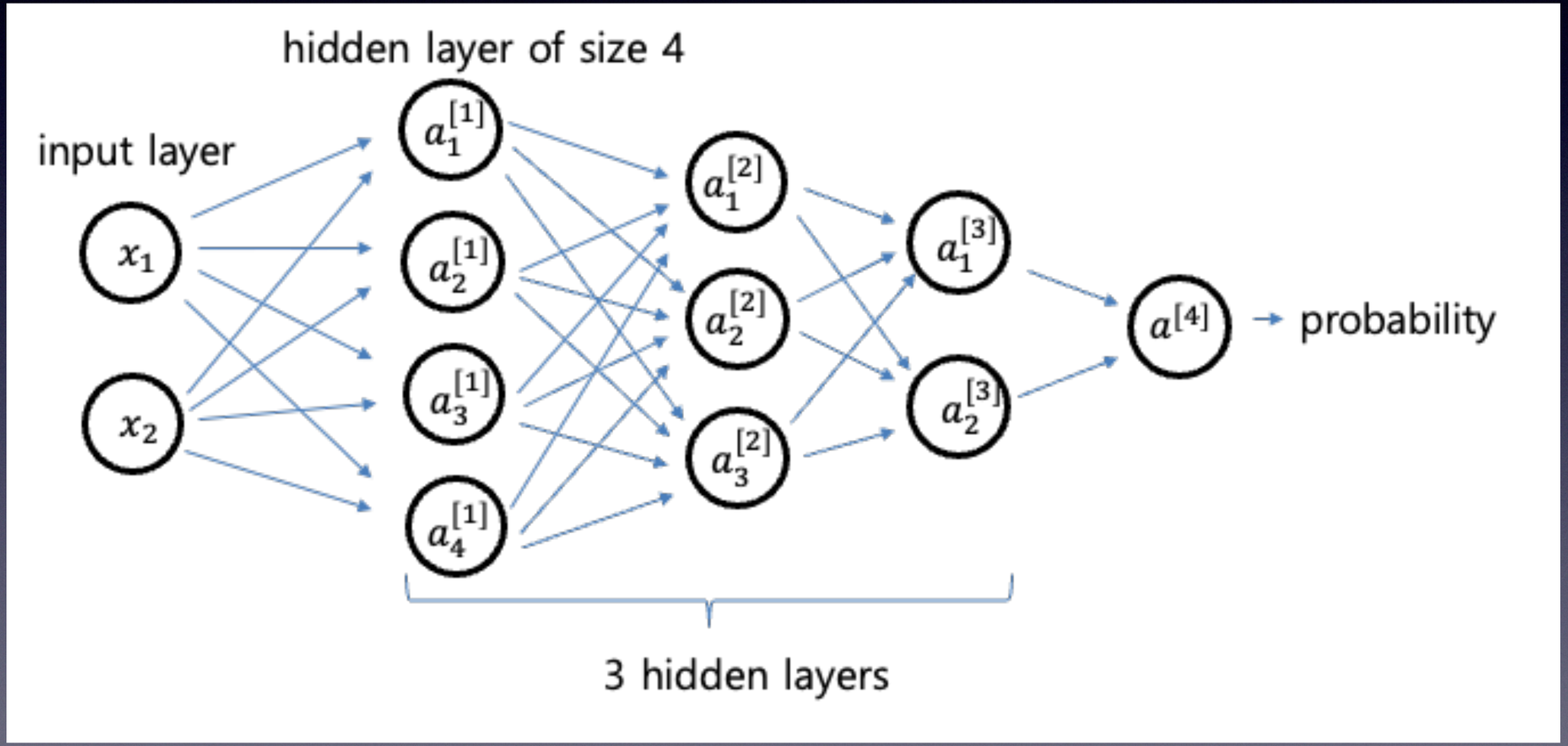
- The function s(·) is called activation function
  - sigmoid or step or Lelu function



single-layer perceptron two-layer perceptron

- Feed-forward neutral network

Deep Neural Network

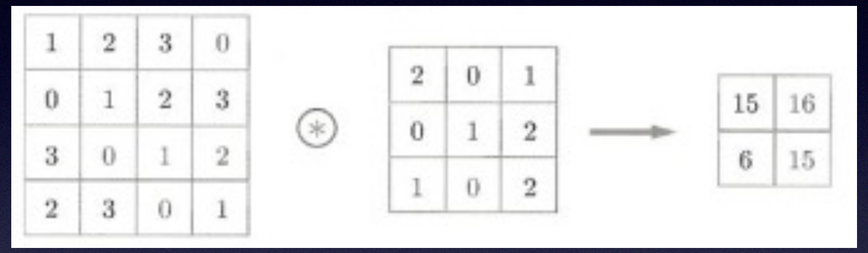


Convolutional Neural Network

- A fully connected network has problems
  - Ignore its spatial information
  - Has too many parameters that should be determined from training
    - e.g. In the MNIST dataset, it (gray) has 28×28=784 weight parameters and for RGB color image, it has 3×28×28=2352 weight parameters
- Convolutional Neural Network
  - Preserves the relationship between nearby pixels
  - Keep the spatial information throughout the layers
  - examples: image recognition, object identification, event classification

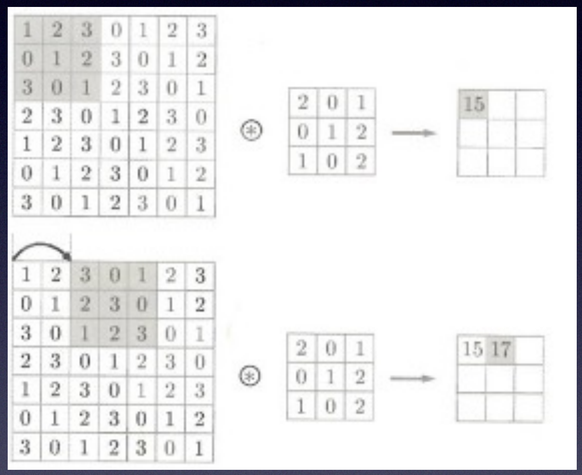
Convolutional Neural Network

Convolution

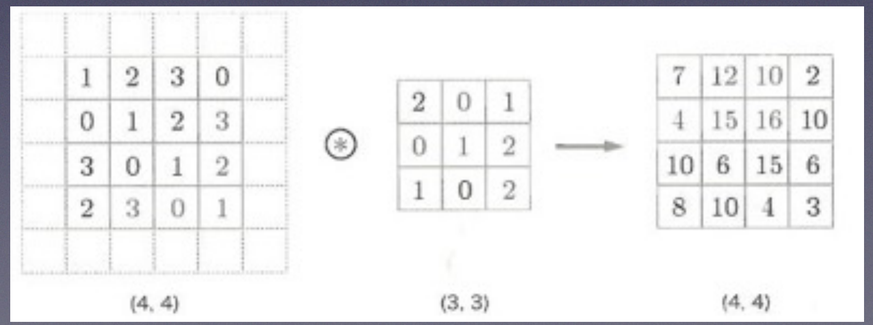


Filter (F)

Stride (S)



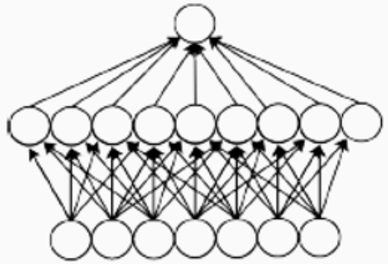
Padding (P)



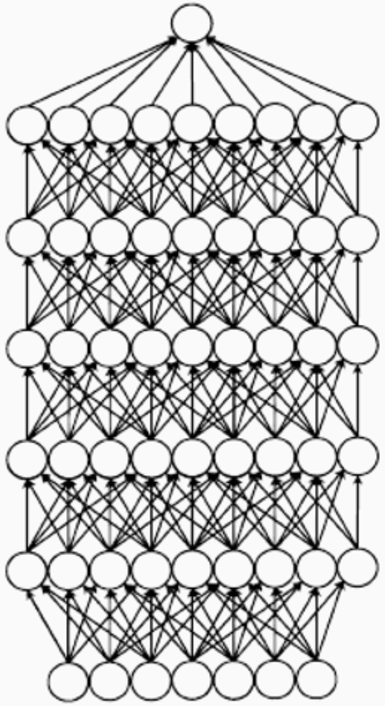
Output Height (or Width)

$$OH = \frac{H + 2P - FH}{S} + 1$$

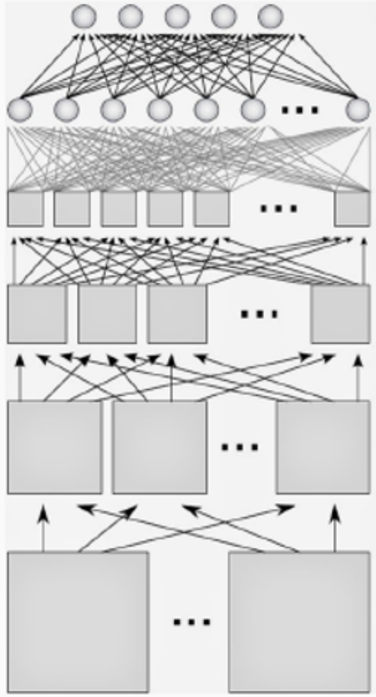
Deep neural network



Neural Network (NN)



Deep NN



Convolutional NN

p-value

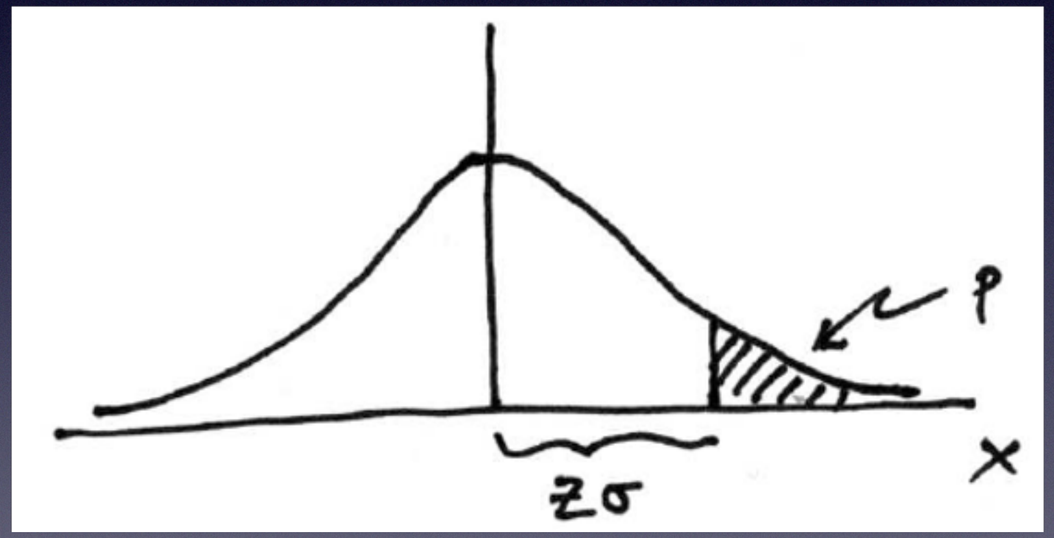
- Probability to observe n heads in N coin tosses is binomial

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

- Hypothesis H: the coin is fair(p=0.5)
- Suppose we toss the coin N=20 times and get n=17 heads
  - P(n=0,1,2,3,17,18,19, or 20)=0.0026
  - p-value=0.0026 is the probability of obtaining such a bizarre result by chance under the assumption of H

Significance from p-value

- Oftern define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value

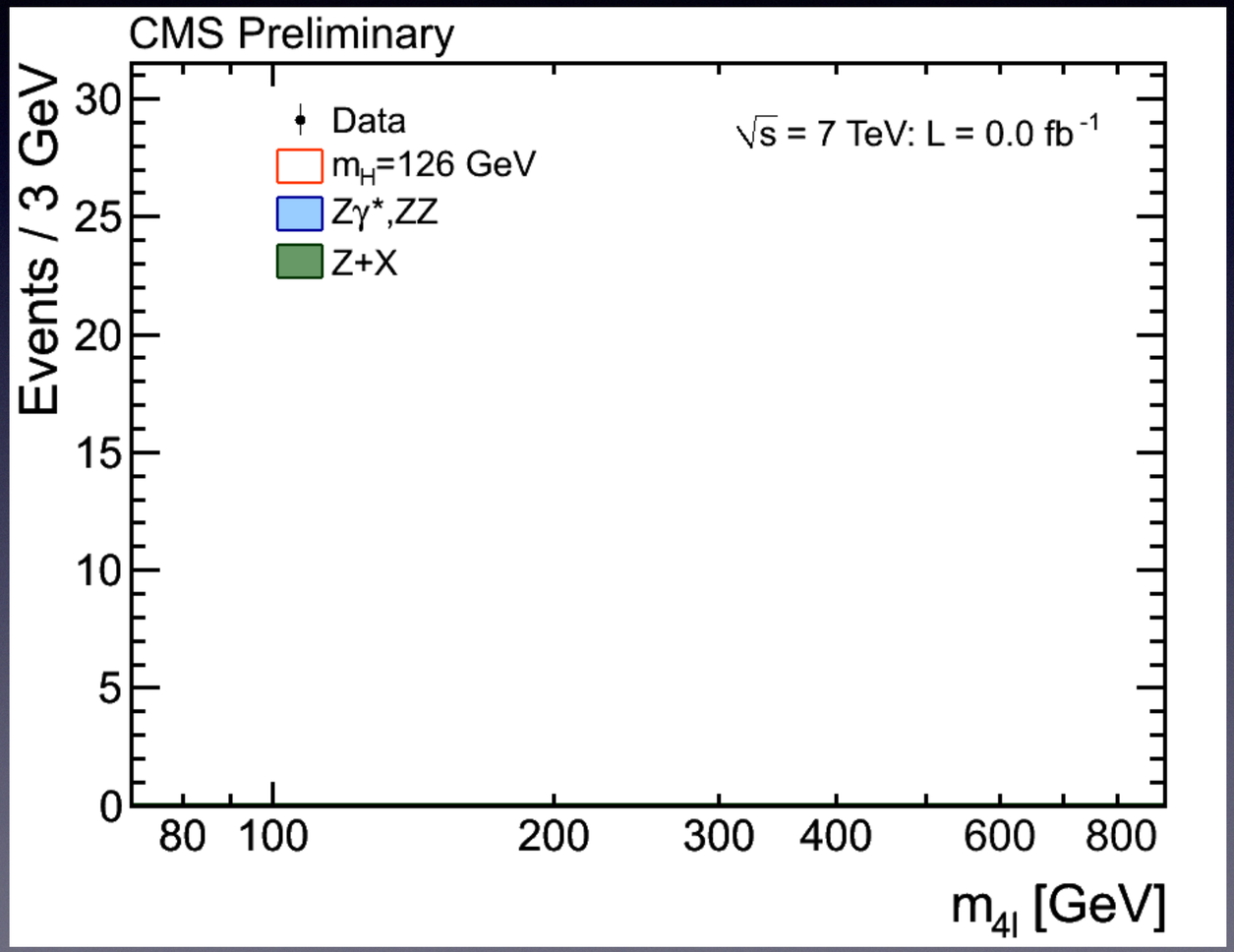


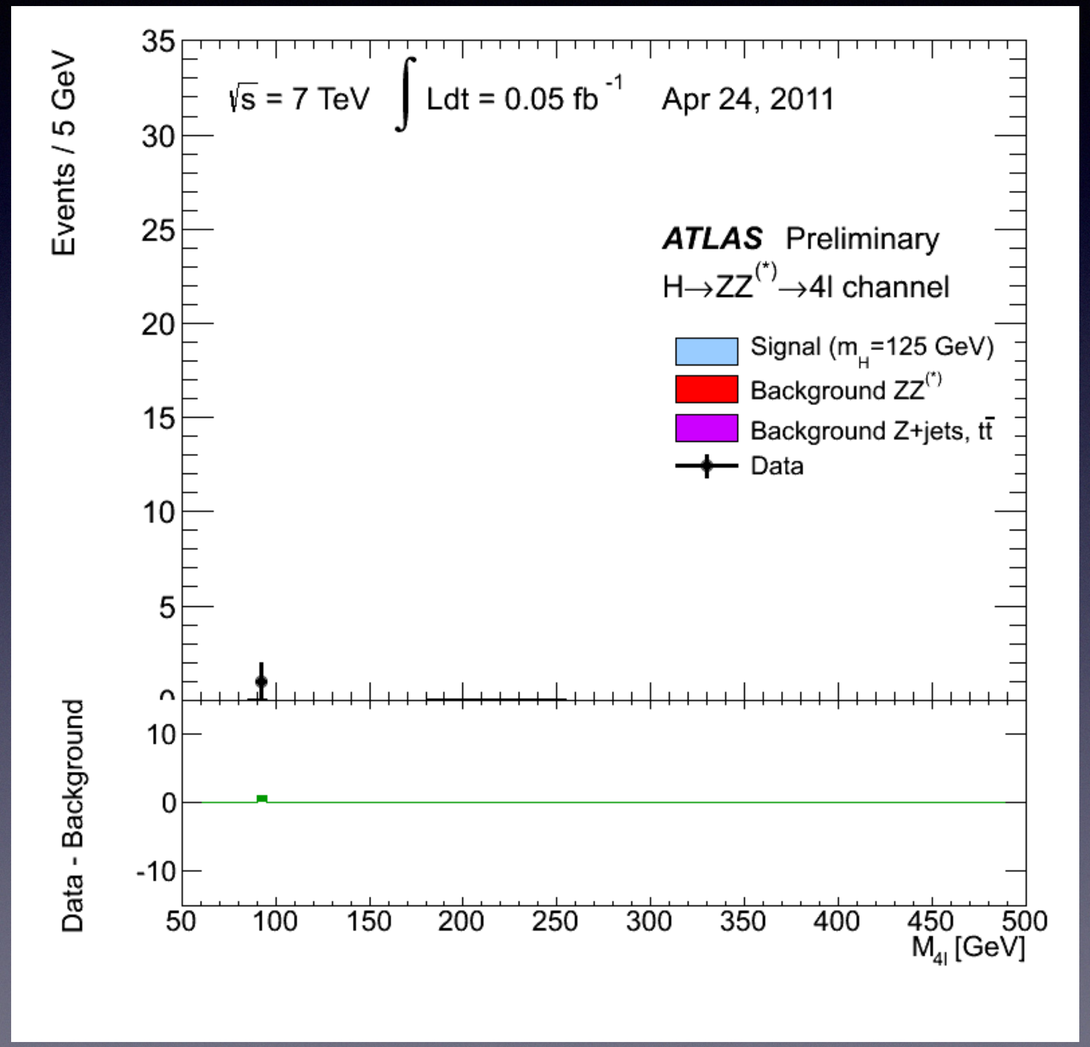
$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1-p)$$

- $Z = 5$ (a 5 sigma deffect) means $p = 2.9 \times 10^{-7}$

Higgs discovery (5σ)





Parameter estimation

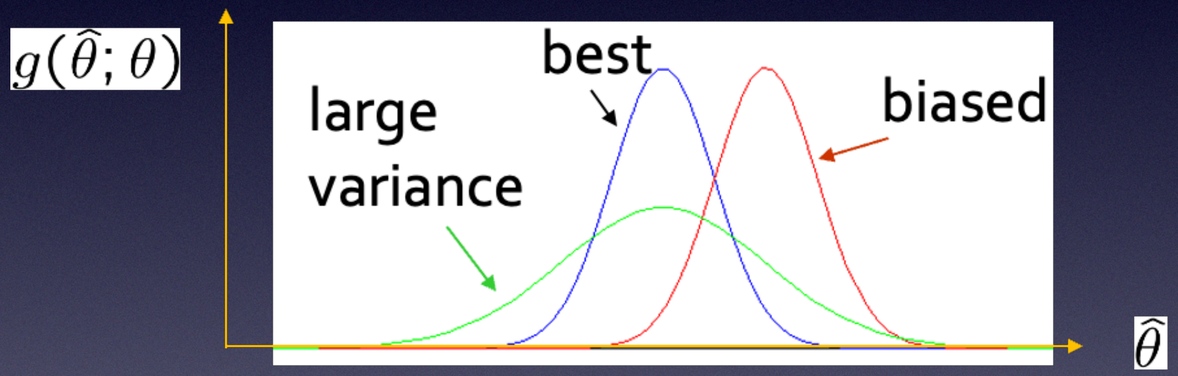
- The parameters of a PDF are constants that characterize its shape, e.g.

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

x:random variable θ:random variable

Properties of estimators

- If we repeat the entire measurements, the estimates follow a PDF



- We want a small bias and a small variance
- Generally there is a tradeoff between bias and variance

An estimator for the mean

- Parameter: $\mu = E[x]$
- Estimator: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \overline{x}$

we find expectation value so that bias $b = E[\hat{\mu}] - \mu = 0$

repeat n times: uncertainty $\sigma_{\hat{\mu}} = \frac{\mu}{\sqrt{n}}$

The likelihood function

- Let's assume that the outcome of an experiment is $x_1, \ldots, x_n$ which is modeled as a sample from a joint PDF with parameters: θ
  $f(x_1, \ldots, x_n; \theta)$

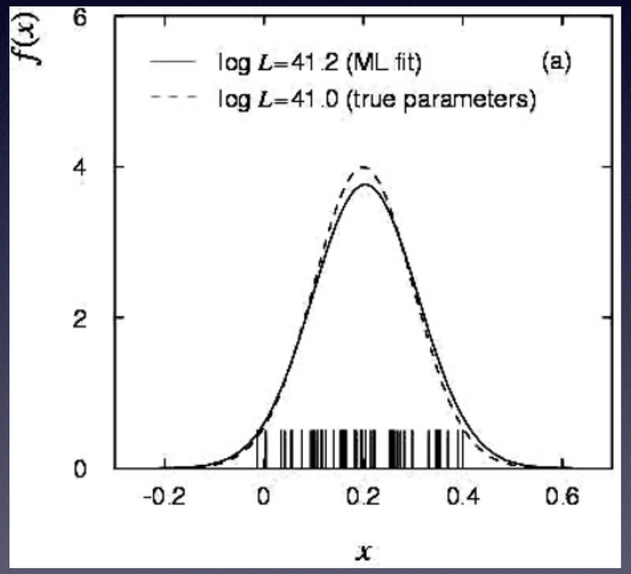
- If the x_i are independent observations, then

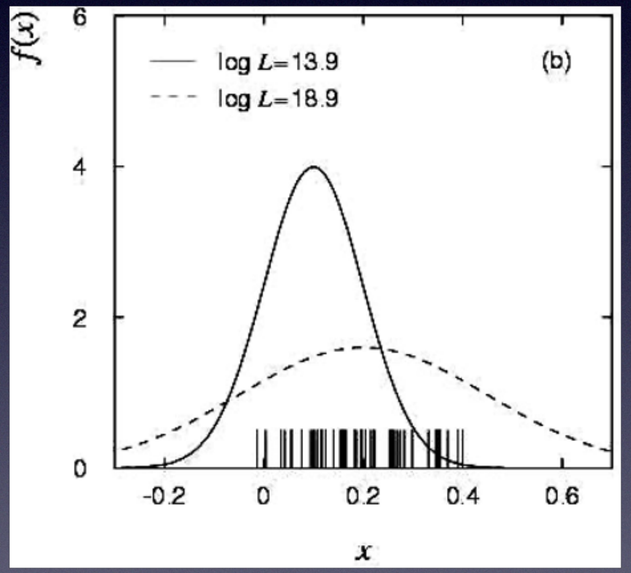
$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$ Likelihood function

- We evaluate this with data and regard it as a function of the parameters

Maximum likelihood estimators

- If the parameters are close to the true value, we expect a high probability





- So we define the maximum likelihood (ML) estimators to be the parameter values for which the likehood is maximum

Maximum likelihood estimators

- Outcome of experiment can be modeled as a set of random variable x
- Theory and detector effects can be described according to some parameters θ which are unknown
- The PDF evaluated for the observation x is called likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

n independent measurements

- The best fit value for θ can be found by maximizing the likelihood function

$$\frac{\partial L}{\partial \theta} = 0$$

Example

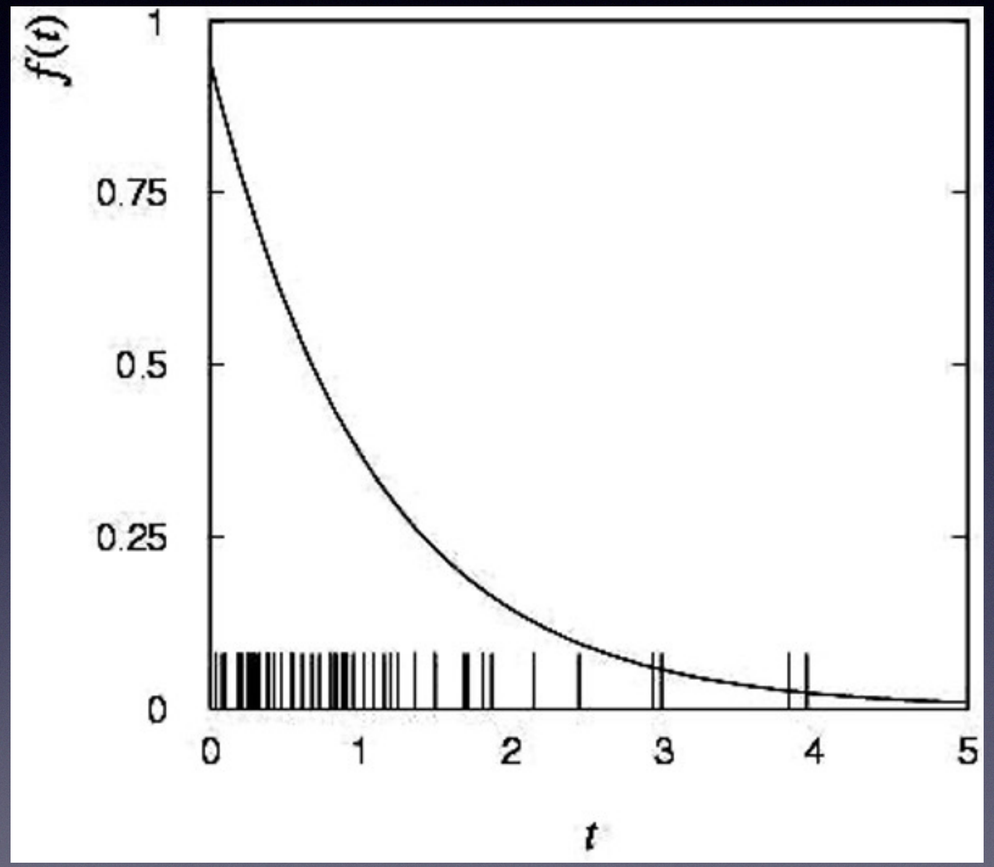
- Consider exponential pdf $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$
- The likelihood function is $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$
- Log-likelihood function

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

Find its maximum by setting $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

ML estimate: $\hat{\tau} = 1.062$



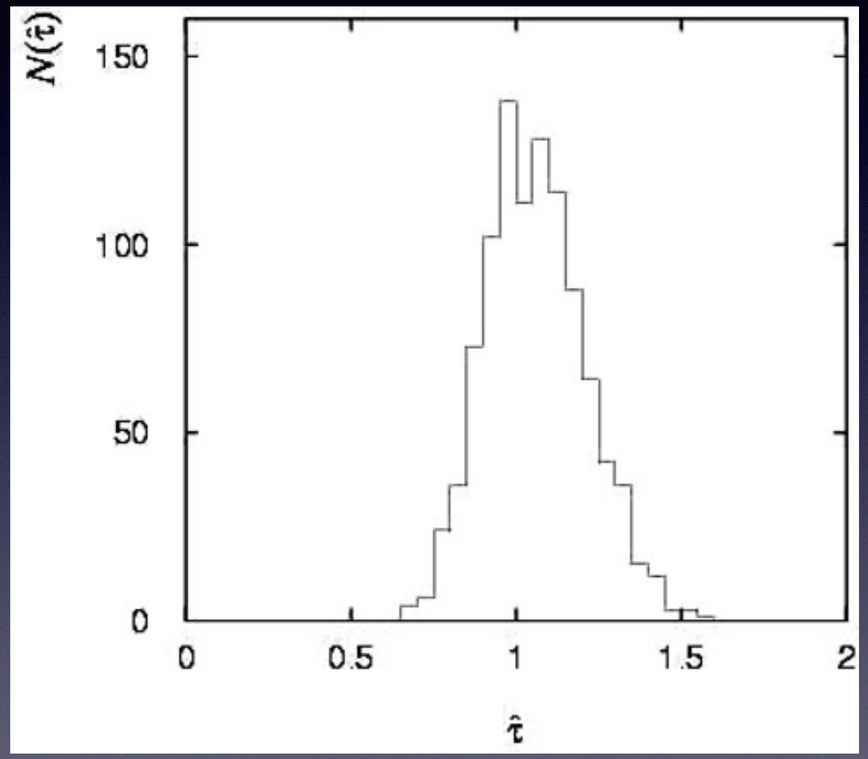
Monte Carlo 50 events using $\tau = 1$

Binned maximum likelihood

- For very large data samples, the log-likelihood function becomes difficult to compute
- Binned maximum likelihood is much faster than unbinned ML
- We use histogram as PDF
- Then, the likelihood function is the product of Poisson probabilities for all bins

Uncertainty on the parameter

- It would be more important to report its uncertainty (statistical error) on our estimation
- One way to do this would be to simulate the entire experiment many times with a Monte Carlo program
- The distribution of estimates is roughly Gaussian



For exponential example, from sample variance of estimates, we find $\hat{\sigma}_{\hat{\tau}}$=0.151

Variance by graphical method

- Using the information unequality (RCF), $\widehat{V}[\widehat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\widehat{\theta}}$
- Expand $L(\theta)$ about its maximum:

$$\ln L(\theta) = \ln L(\widehat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta})^2 + \ldots$$

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \widehat{\theta})^2}{2\widehat{\sigma^2}_{\widehat{\theta}}}$$

i.e., $$\ln L(\widehat{\theta} \pm \widehat{\sigma}_{\widehat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

- To get the uncertainty of $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by 1/2

Variance by graphical method

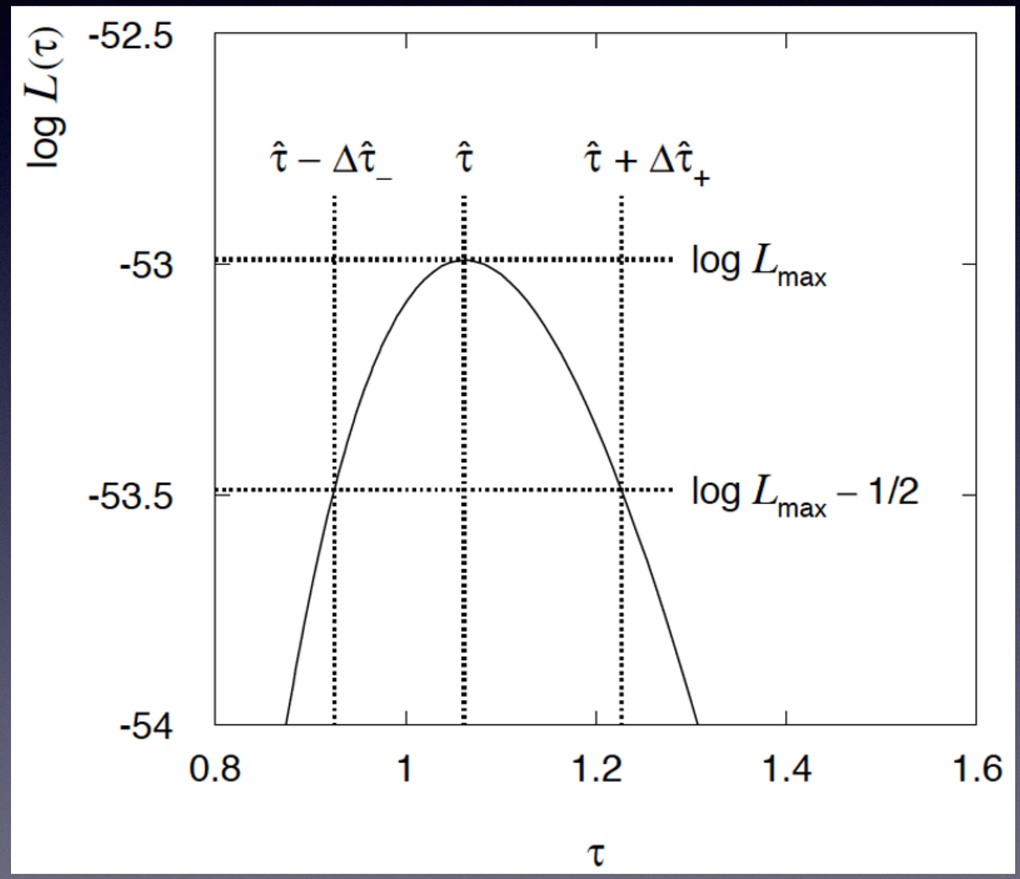
- ML example with exponential:
- not quite parabolic $\ln L$ when the sample size is not big

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

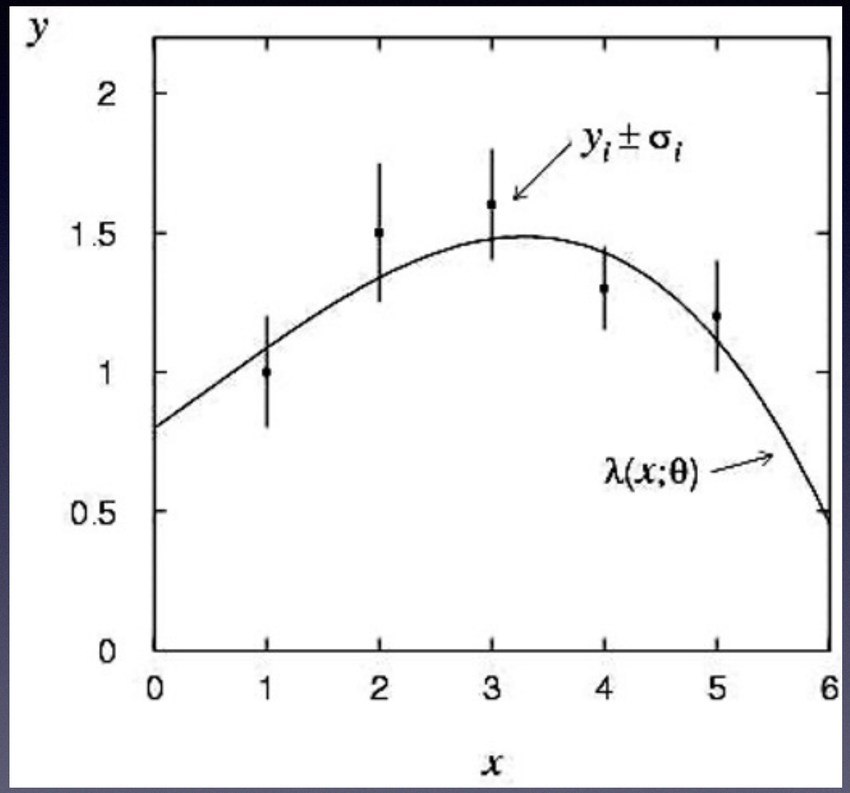
$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$



The method of least squares

- We measure N values $y_1, \ldots, y_N$, which are assumed to be independent Gaussian random variables with $E[y_i] = \lambda(x_i; \theta)$
- Each point has variances $V[y_i] = \sigma_i^2$
- Now we want to estimate θ in the fit function $f(y_i; \theta)$
- The likelihood function is



$$L(\theta) = \prod_{i=1}^{N} f(y_i; \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - \lambda(x_i; \theta))^2}{2\sigma_i^2}\right]$$

The method of least squares

- The log-likelihood function is

$$\ln L(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2} + \text{terms not depending on } \theta$$

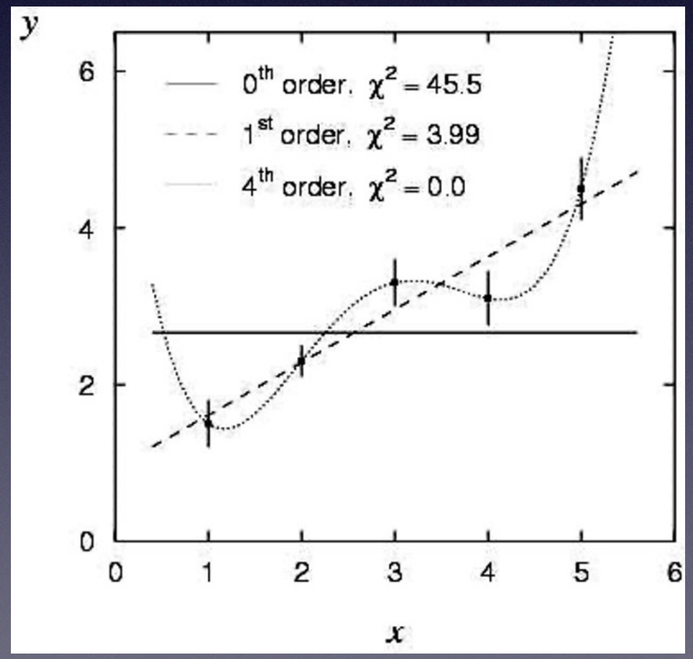
- Maximizing the likelihood is equivalent to minimizing $\chi^2(\theta)$

$$\chi^2(\theta) = \sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

- Minimum of this quantity defines the least squares estimater $\hat{\theta}$
- We rely on the program (MINUIT2) to minimize χ^2 numerically

Example of least squares fit

- Fit a polynomialof order p: $\lambda(x;\theta_0,\ldots,\theta_p) = \sum_{n=0}^{p} \theta_n x^n$



Variance of Least Square estimators

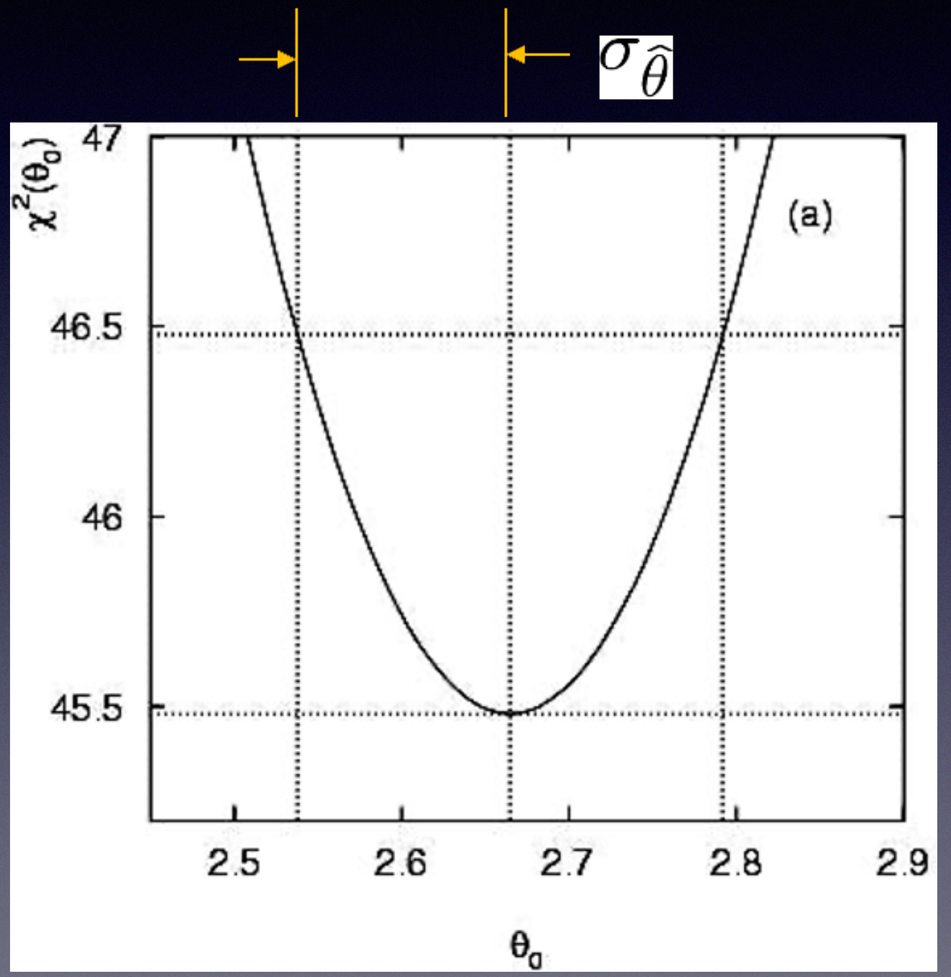
- We can obtain the variance in the same way for Maximum Likelihood method

$$\chi^2(\theta) = -2 \ln L(\theta)$$

and so $$\widehat{\sigma^2}_{\hat{\theta}} \approx 2 \left[\frac{\partial^2 \chi^2}{\partial \theta^2} \right]^{-1}_{\theta=\hat{\theta}}$$

- We can use the graphical method

$$\chi^2(\theta) = \chi^2_{\text{min}} + 1$$



Goodness-of-fit with least squares

- We can show that if the hypothesis is correct, then the statistics $t = \chi^2_{min}$ follows the chi-square pdf

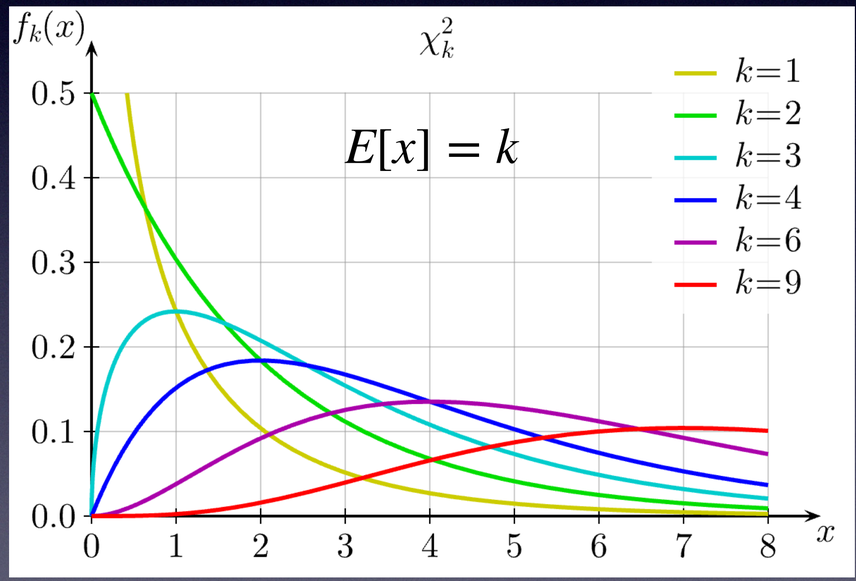
$$f(t; n_{\mathrm{d}}) = \frac{1}{2^{n_{\mathrm{d}}/2}\Gamma(n_{\mathrm{d}}/2)} t^{n_{\mathrm{d}}/2-1} e^{-t/2}$$

where n_d is the number of degrees of freedom

n_d = number of data points - number of fitted parameters

The χ^2pdf has an expectation value equal to n_d

If $\chi^2_{min}/n_d \approx 1$, the fit is good!

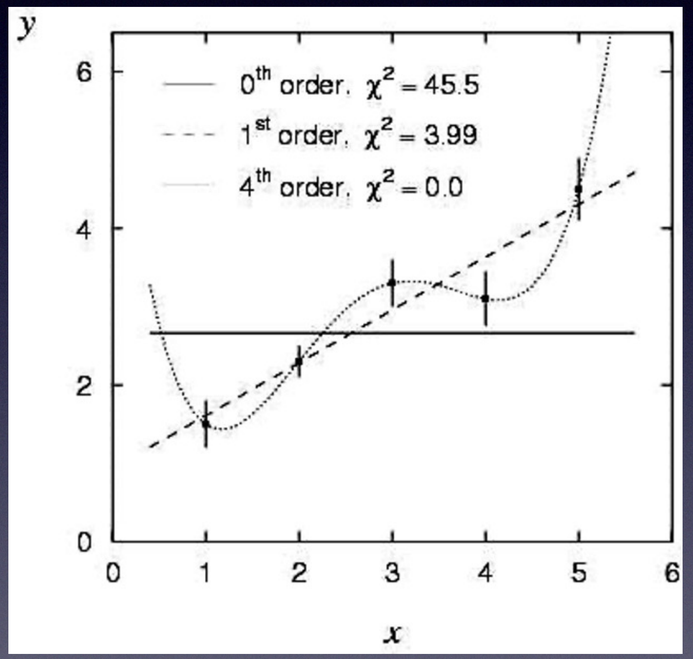


k in this plot is the number of degrees of freedom

Goodness of fit with least squares

- More generally, find the p-value: $p = \int_{\chi^2_{\text{min}}}^{\infty} f(t; n_{\text{d}})\, dt$

  This is the probability of obtaining a χ^2_{min} as high as the one we got, or higher if the hypothesis is correct



- In the previous example with 1st order polynomial

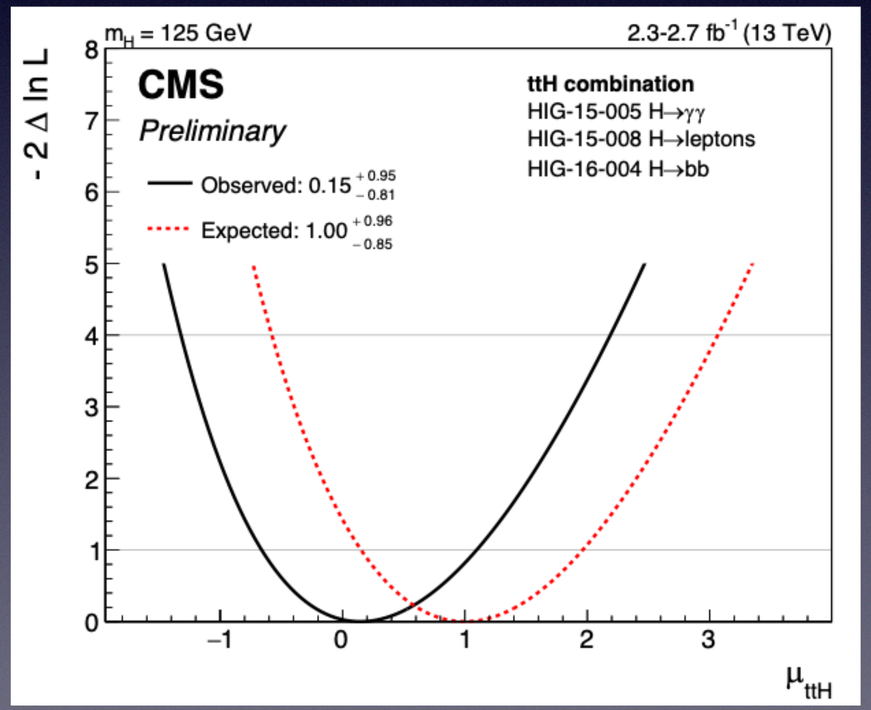
$$\chi^2_{\text{min}} = 3.99\,, \qquad n_{\text{d}} = 5-2 = 3\,, \qquad p = 0.263$$

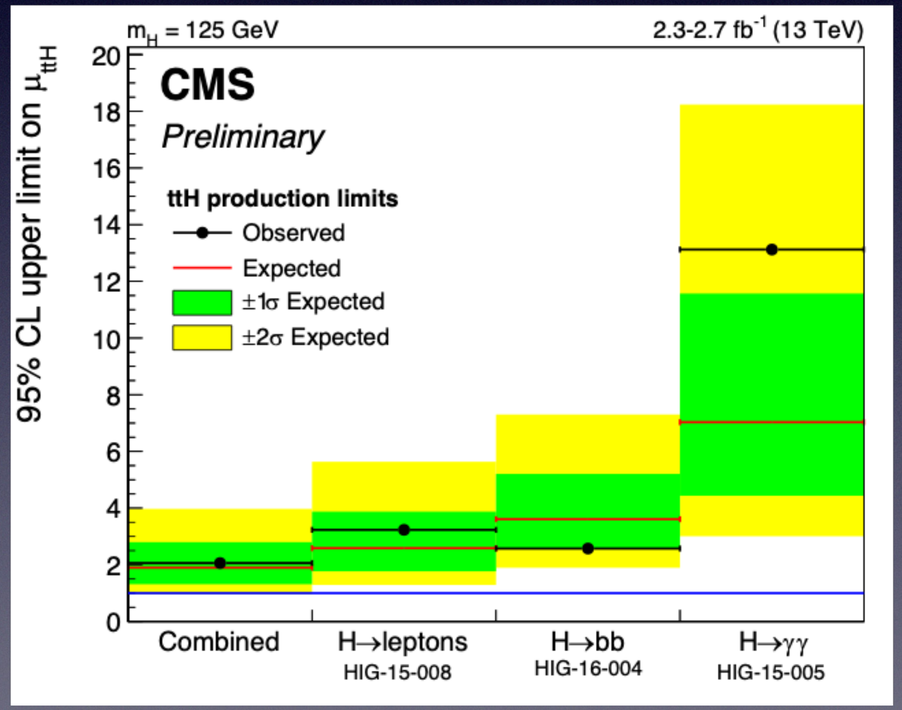
with 0th order polynomial

$$\chi^2_{\text{min}} = 45.5\,, \qquad n_{\text{d}} = 5-1 = 4\,, \qquad p = 3.1\times10^{-9}$$

Setting limits

- What is the signal strength and upper limit?





Upper limit

- Consider the case of fining $n = n_s + n_b$ events (n_s from signal, n_b from background) which are following the Poission distribution
- And we are searching for the evidence of the signal process but the number of events found is roughly equal to the expected background events (n_b=4.6 and n_{obs} = 5)
- There is no access of the signal over background
- What values of n_s can we exclude on the grounds that they are incompatible with the data ⟶ We set the upper limit on n_s

Confidence interval

- Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ)
  - Values of data disfavoured by θ, P(data in critical region) $\leq \alpha$, e.g. α=0.05 or 0.1
  - If data observed in the critical region, we reject the value θ
- Invert the test to define a confidence level as:
  - set of θ values that would not be rejected in a test of size α (confidence level is 1-α)
- The interval will cover the true value of θ with probability $\geq 1 - \alpha$

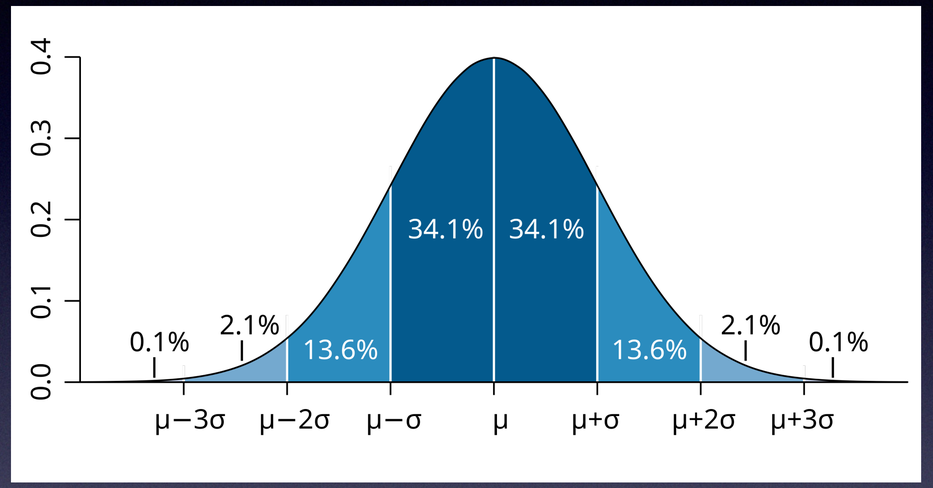
Relation between confidence interval and p-value

- We can considder a significance test for each hypothesized value (q) resulting in a p-value (p_q)

p-valye $p_q < \alpha$ then reject the hypotheis (q)

- The confidence interval at $CL = 1 - \alpha$ consists of those values of q that are not rejected

Confidence interval for a Gaussian distribution



- For the Gaussian distribution, the 68.3% central confidence itnerval is given by the estimated value ± one standard deviation

- The final result is then simply reported as $\theta_{obs} \pm \sigma_{\theta}$

$$[a, b] = [\hat{\theta}_{obs} - \sigma_{\hat{\theta}}, \hat{\theta}_{obs} + \sigma_{\hat{\theta}}]$$

Example of an upper limit

- Find the hypothetical value of s such that there is a given small probability, say, α=0.05, to find as few events as we did or less

$$\gamma = P(n \leq n_{\text{obs}}; s, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

- Solve numerically for $s = s_{up}$, this gives an upper limit on s at a confidence level of $1 - \alpha$
- Example: suppose $b = 0$ and we find $n_{obs} = 0$. For $1 - \alpha = 0.95$

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \rightarrow s_{\text{up}} = -\ln \alpha \approx 3.00$$

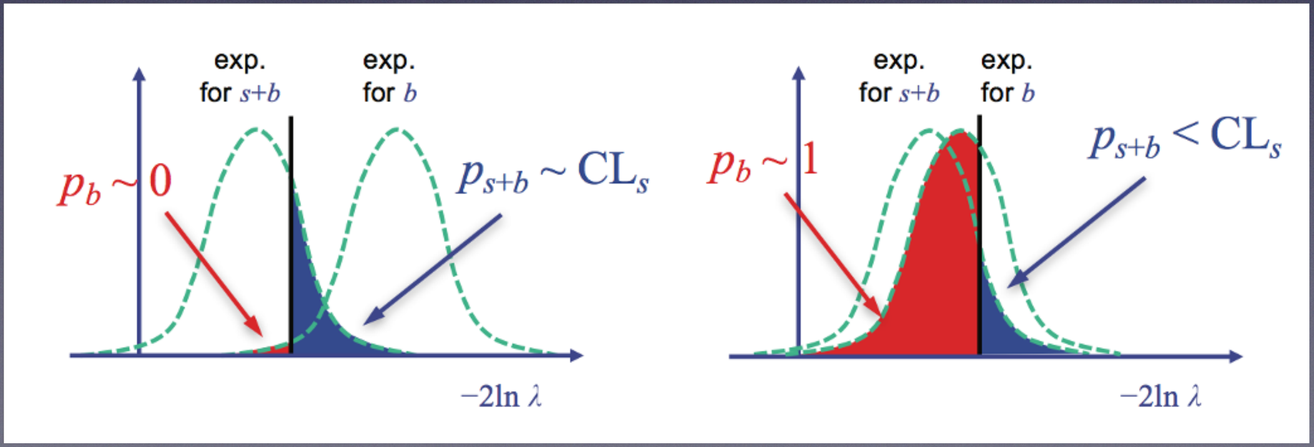
- The interval $[0, s_{up}]$ is an example of a confidence interval, designed to cover the true value of s with a probability $1 - \alpha$

Modified frequentist approach

- p_{s+b} is the probability to obtain a result which is less compatible with the signal than the observed result, assuming the signal hypothesis

- p_b is the probability to obtain a result less compatible with the background-only hypothesis than the observed one, assuming background only

$$CL_S = \frac{p_{s+b}}{1 - p_b}$$



$$\lambda(x) = \frac{L(x \mid H_1)}{L(x \mid H_0)}$$