

Introduction to Statistics and Machine Learning

Hadron Collider School HASCO 2021

Federico Meloni
(Deutsches Elektronen-Synchrotron DESY)

20/07/2021

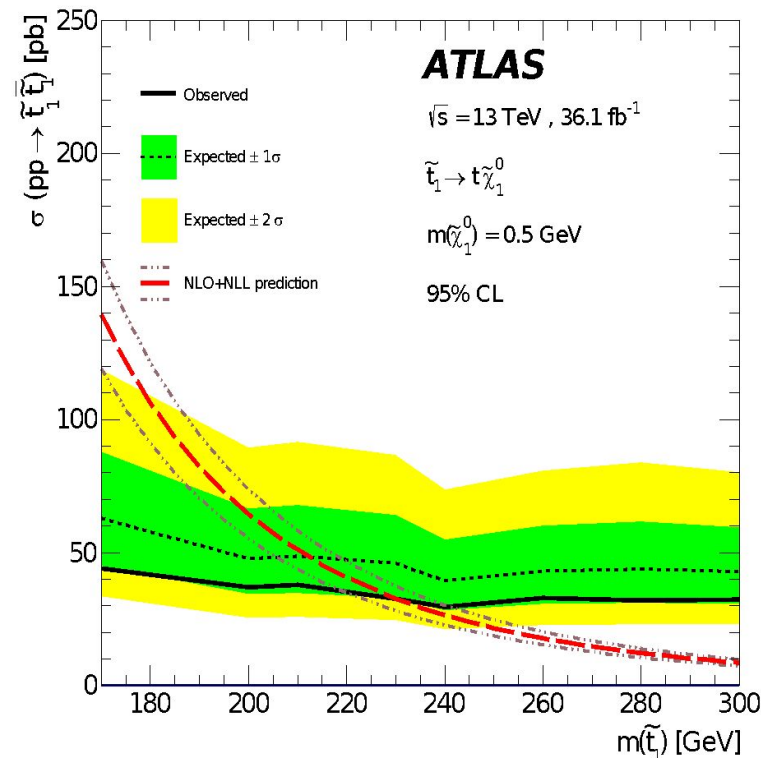


"LISTEN, SCIENCE IS HARD.
BUT I'M A SERIOUS
PERSON DOING MY BEST."

Introduction

We will discuss the most important concepts and methods of statistical data analysis used in HEP.

- Basic formalism
- Discovery and Limits
- Estimators and parameter estimation
- Machine Learning Basics



The aim is allow you to discuss HEP limit setting and understand the basic concepts of machine learning techniques

Useful references

[1] G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998

- These slides are mostly stolen from/inspired by the set of Prof. Cowan's academic training lectures (<https://indico.cern.ch/event/77830/>)

[2] F. James, *Statistical methods in experimental physics*, 2nd ed., World Scientific, 2006

[3] L. Lyons, *Statistics for nuclear and particle physicists*, Cambridge University Press, 1986

[4] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (2nd ed.), Springer Series in Statistics, 2001

- Statistics Tools (ROOT based packages): RooStats, RooFit, HistFitter
- Machine learning (python based): TensorFlow, PyTorch, ...
(fun ML playground <http://playground.tensorflow.org>)

Definitions (lookup table)

- x observable
- H hypothesis
- $P(x|H)$ probability of x given the hypothesis H
- $f(x|H)$ probability density function / probability model
- $L(x|H)$ likelihood of the hypothesis
- W data space
- w critical region
- α test size / test significance level

BASIC FORMALISM

- Frequentist and Bayesian Probability
- Hypothesis testing
- Type-I, Type-II errors and statistical power
- Test statistics
- *p*-values

“Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone.”
- Louis Lyons

Frequentist statistics

In frequentist statistics, probabilities are associated only with the data, i.e. outcomes of repeatable observations.

$$P(x) = \lim_{n \rightarrow \infty} \frac{\text{number of outcomes of } x}{n}$$

Probabilities such as

- P (Higgs boson exists)
- $P(0.117 < \alpha_s < 0.121)$

are either 0 or 1, but we don't know which.

- The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian statistics

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability).

Probability of the data
assuming the hypothesis H

prior
probability

posterior
probability

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H) dH}$$

Normalize over all possible hypotheses

- Bayesian methods can provide more natural treatment of non- repeatable phenomena: e.g. systematic uncertainties
- No golden rule for priors

Nomenclature

A hypothesis H specifies the probability for the data x , i.e. the outcome of the observation.

- $f(x|H)$ is the p.d.f. or **probability model** (or just model)
- $f(x|H)$ could be **uni-/multivariate, continuous or discrete**
- $f(x|H)$ could be the observation of a single particle, a single event, or an entire “experiment”
- Possible values of x form the **data space W**
- Simple hypothesis: $f(x|H)$ completely specified
- Composite hypothesis: H contains unspecified parameters

- The probability for x given H is also called the likelihood of the hypothesis, written $L(x|H)$.

Hypothesis test

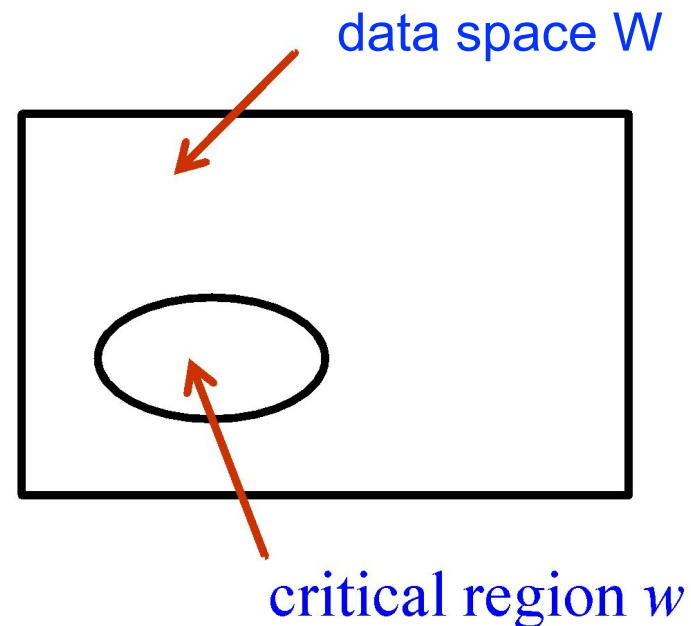
- Consider a simple hypothesis H_0 and alternative H_1 .
- A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α , assuming H_0 is correct, to observe the data there.

$$P(x \in w | H_0) \leq \alpha$$

(just $<$ if data are discrete)

α is called the **size or significance level of the test**.

- If x is observed in the critical region, reject H_0 .



Type-I, type-II errors and power of the test

Type-I error: reject the hypothesis H_0 when it is true.

- The maximum probability for this is the size of the test

$$P(x \in w \mid H_0) \leq \alpha$$

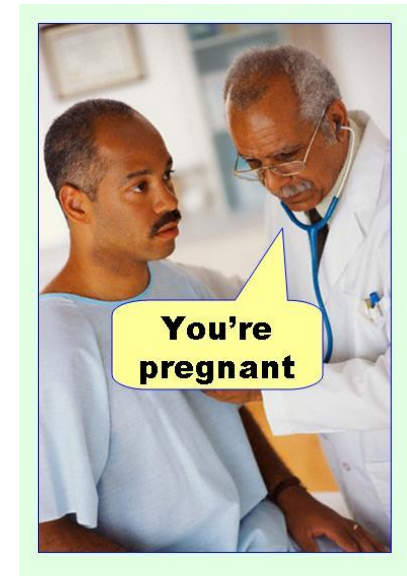
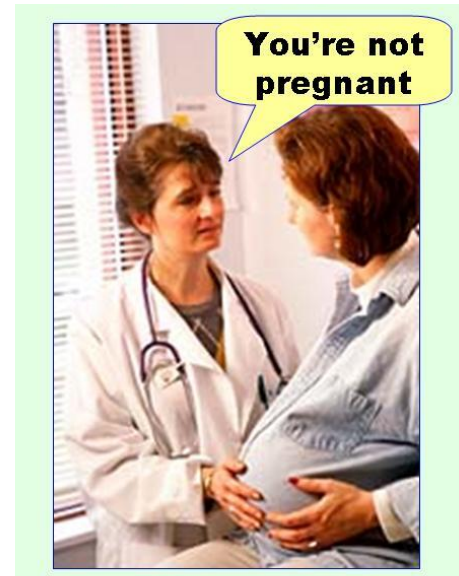
Type-II error: accept H_0 when it is false, and an alternative H_1 is true.

- This occurs with probability

$$P(x \in W - w \mid H_1) = \beta$$

One minus this is called the **power of the test** with respect to the alternative H_1 :

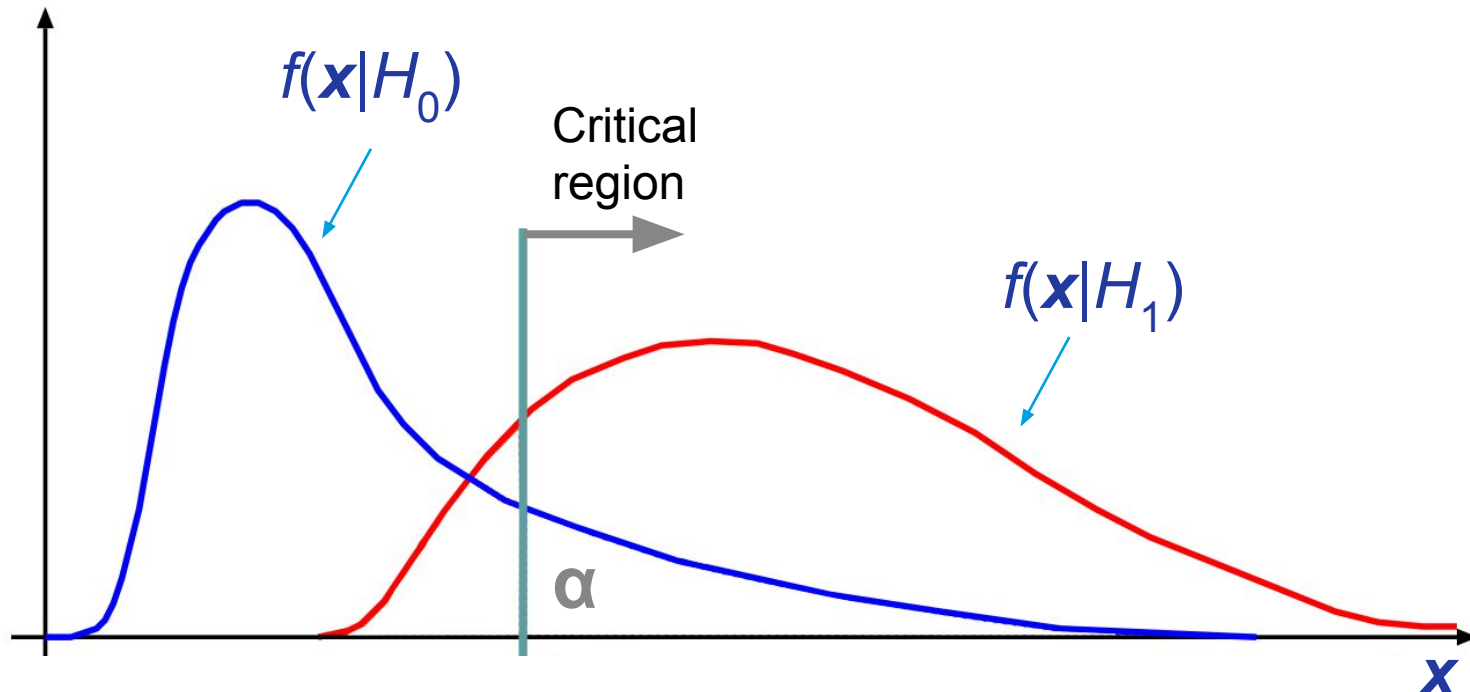
$$\text{Power} = 1 - \beta$$



Test definition

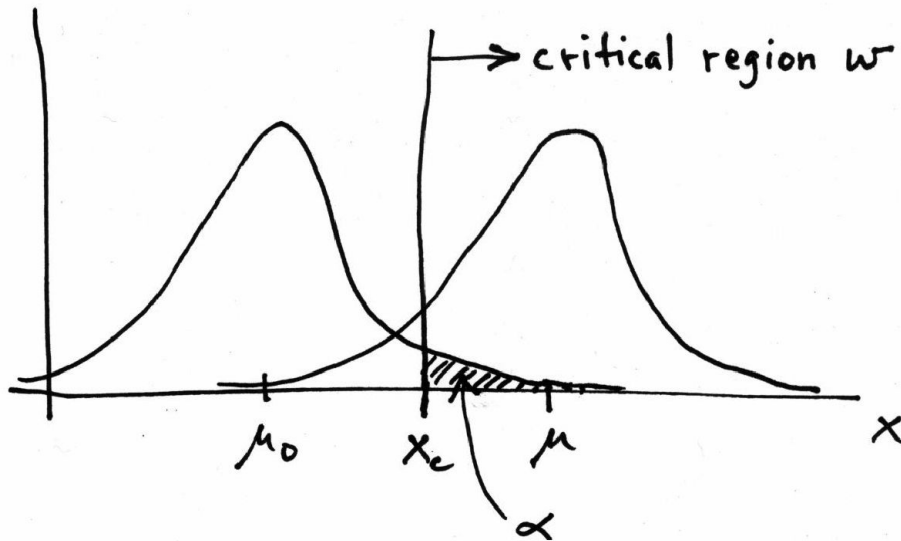
In general there are an infinite number of possible critical regions that give the same significance level α .

- Maximize power with respect to H_1 - maximize probability to reject H_0 if H_1 is true.



Example

- Measuring $x \sim \text{Gauss}(\mu, \sigma)$ we may test $H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$
- The highest power with respect to any $\mu > \mu_0$ is obtained by taking the critical region $x \geq x_{\text{cut-off}}$ where $x_{\text{cut-off}}$ is determined by the significance level such that $\alpha = P(x \geq x_{\text{cut-off}} \mid \mu_0)$.



$$\alpha = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma}\right)$$

$$x_c = \mu_0 + \sigma\Phi^{-1}(1 - \alpha)$$

where:

Φ Standard Gaussian cumulative distribution

Φ^{-1} Standard Gaussian quantile

Model independence

- In HEP we often construct a test of H_0 : Standard Model (or “background only”) such that we have a well specified “false discovery rate” (α) and high power with respect to some interesting alternative H_1 : SUSY, Z' , etc.
- **But there is no such thing as a “model independent” test.**
- No “Uniformly Most Powerful” test: any statistical test will have high power with respect to some alternatives and less power with respect to others.

Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true.

- In frequentist statistics only associate probability with outcomes of repeatable observations.
- In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

$$P(H|x) = \frac{P(x|H)\pi(H)}{\int P(x|H)\pi(H) dH}$$

which depends on the prior probability $\pi(H)$.

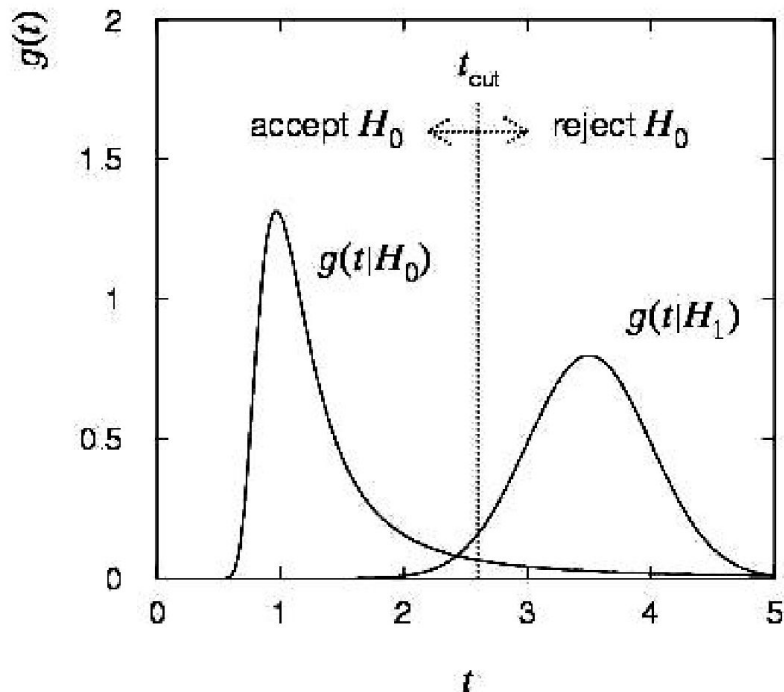
- What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.

Test statistics

- The boundary of the critical region for an n -dimensional data space $\mathbf{x} = (x_1, \dots, x_n)$ can be defined by an equation of the form

$$t(x_1, \dots, x_n) = t_{\text{cut}}$$

where $t(x_1, \dots, x_n)$ is a scalar test statistic.



The decision boundary is now a single 'cut' on t , defining the critical region.

- For an n -dimensional problem we have a corresponding 1-dimensional problem.**

The likelihood function

- Suppose the entire result of an experiment (set of measurements) is a collection of numbers \mathbf{x} , and suppose the pdf for is a function that depends on a set of parameters θ : $f(\vec{x}; \vec{\theta})$
- Now evaluate this function with the data obtained and regard it as a function of the parameters. This is the likelihood function:

$$L(\vec{\theta}) = f(\vec{x}; \vec{\theta}) \text{ for fixed data } \mathbf{x}$$

- Consider n independent observations of x : x_1, \dots, x_n , where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

- In this case the likelihood function is $L(\vec{\theta}) = \prod_{i=1}^n f(x_i; \vec{\theta})$

The optimal critical region

The Neyman-Pearson lemma states:

- For a test of size α of the simple hypothesis H_0 , to obtain the highest power with respect to the simple alternative H_1 , choose the critical region w such that the likelihood ratio satisfies

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq k$$

everywhere in w and is less than k elsewhere, where k is a constant chosen such that the test has size α .

- The optimal scalar test statistic is then

$$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

N.B. any monotonic function of this leads to the same test.

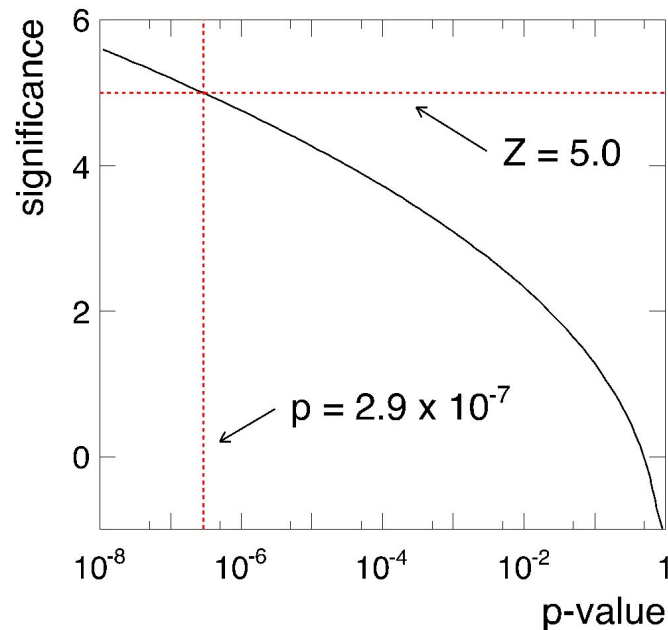
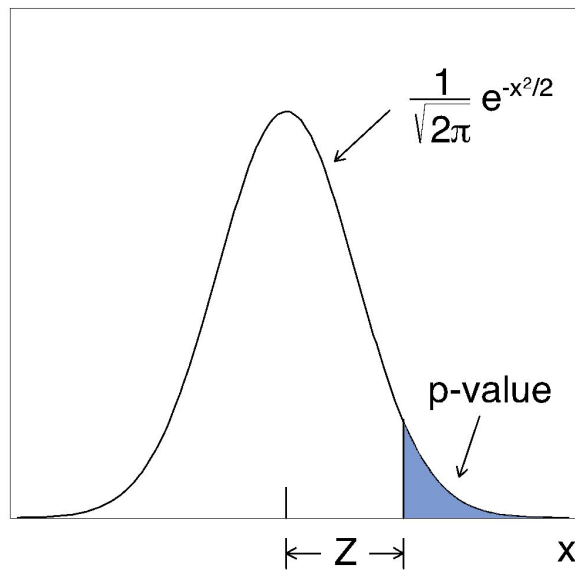
p-values

We can express level of agreement between data and a hypothesis H with a *p*-value:

- p = probability, under assumption of H , to observe data with equal or lesser compatibility with H
- This is not the probability that H is true!
- In frequentist statistics we don't talk about $P(H)$ (unless H represents a repeatable observation).

Significance from p-value

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



z (one tail)	p-value
1.00	0.16
2.00	0.023
3.00	0.0013
4.00	3.2e-05
5.00	2.9e-07
6.00	9.9e-10

Significance of an observation

Suppose we observe n events; these can consist of:

- n_b events from known processes (background)
- n_s events from a new process (signal)
- If n_s, n_b are distributed like a Poisson with means s, b , then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

$$P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

- Suppose $b = 0.5$, and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

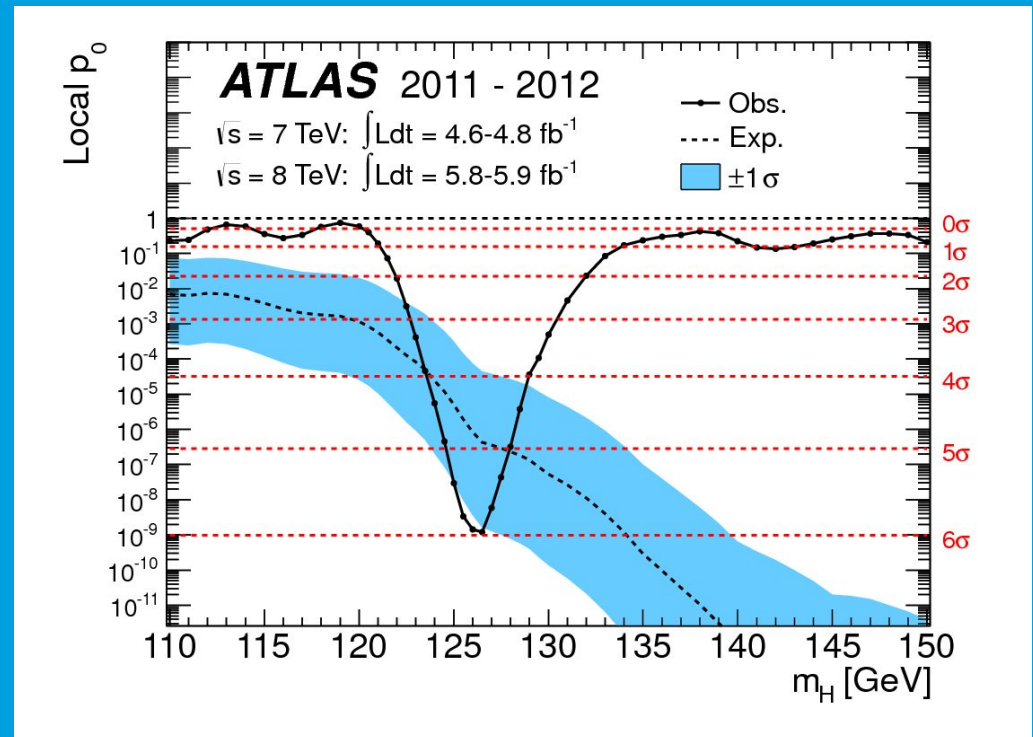
- consider hypothesis $s = 0$:
 $p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$
 $= 1.7 \times 10^{-4} \neq P(s = 0)!$

Test of H_0 using a p-value

- We started by defining critical region in the original data space W , then reformulated this in terms of a scalar test statistic $t(x)$.
- We can now define the critical region of a test of H_0 with size α as the set of data space where p_0 (p-value of H_0) $\leq \alpha$.
- Formally the p -value relates only to H_0 , but the resulting test will have a given power with respect to a given alternative H_1 .

Discovery and limits in HEP

- The likelihood ratio
- Discovery
- Exclusion



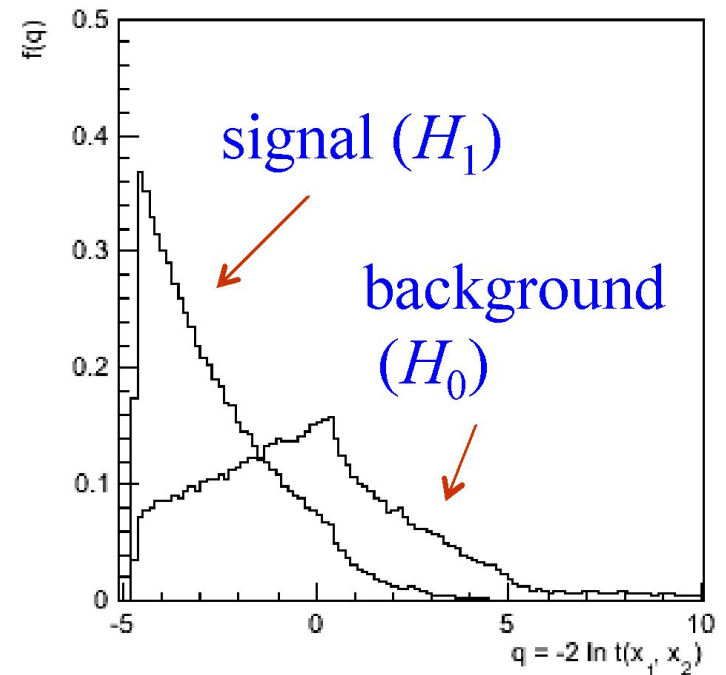
The likelihood ratio

In a real-world problem we usually wouldn't have the pdfs $f(\mathbf{x} | H_0)$ and $f(\mathbf{x} | H_1)$, so we wouldn't be able to evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}$$

For a given observed \mathbf{x} , we need methods to approximate this with some other function.

- Using Monte Carlo, we can find the distribution of the likelihood ratio



Claiming a discovery

If the signal process is not known to exist and we want to search for it, the relevant hypotheses are therefore:

- H_0 : all events are of the background type
- H_1 : the events are a mixture of signal and background
- **Rejecting H_0 with $Z > 5$ constitutes “discovering” new physics.**
- Suppose that for a given integrated luminosity, the expected number of signal events is s , and for background b .

The observed number of events n will follow a Poisson distribution:

$$P(n|b) = \frac{b^n}{n!} e^{-b}$$

$$P(n|s + b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

Claiming a discovery

We observe n events, and thus measure n instances of $\mathbf{x} = (x_1, x_2)$.

The likelihood function for the entire experiment assuming the background-only hypothesis (H_0) is

$$L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^n f(\mathbf{x}_i | b)$$

and for the “signal plus background” hypothesis (H_1) it is

$$L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)} \prod_{i=1}^n (\pi_s f(\mathbf{x}_i | s) + \pi_b f(\mathbf{x}_i | b))$$

where π_s and π_b are the (prior) probabilities for an event to be signal or background, respectively.

Claiming a discovery

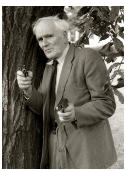
We can define a test statistic Q monotonic in the likelihood ratio as

$$Q = -2 \ln \frac{L_{s+b}}{L_b} = -s + \sum_{i=1}^n \ln \left(1 + \frac{s}{b} \frac{f(\mathbf{x}_i|s)}{f(\mathbf{x}_i|b)} \right)$$

To compute p-values for the b and $s+b$ hypotheses given an observed value of Q we need the distributions $f(Q|b)$ and $f(Q|s+b)$.

- the term $-s$ in front is a constant and can be dropped.
- the rest is a sum of contributions for each event, and each term in the sum has the same distribution.

Q

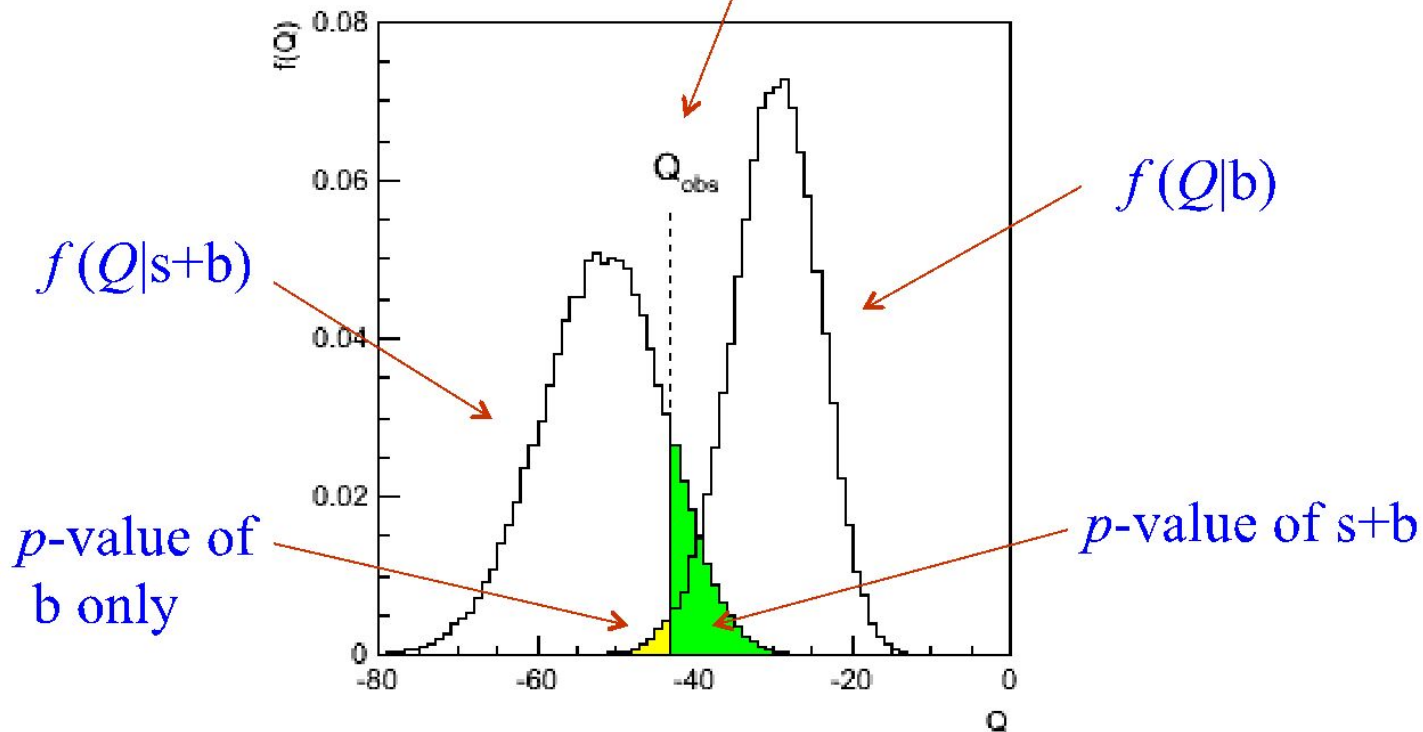


If $p_{s+b} < \alpha$, reject signal model s at confidence level $1 - \alpha$.

If $p_b < 2.9 \times 10^{-7}$, reject background-only model (significance $Z = 5$).

e.g. $b = 100$, $s = 20$.

Suppose in real experiment Q is observed here.



The profile likelihood ratio

Assume a counting experiment to search for signal in a region of phase space

- result is n_i Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$



strength parameter

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters m_i

- Also Poisson distributed

$$E[m_i] = u_i(\boldsymbol{\theta})$$



nuisance parameters ($\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\text{tot}}$)

- Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^N \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^M \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

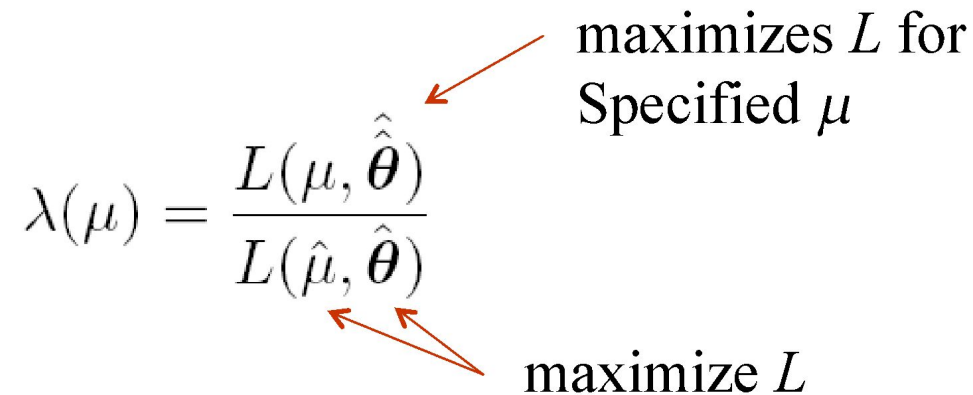
The profile likelihood ratio

Base significance test on the profile likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\theta})}{L(\hat{\mu}, \hat{\theta})}$$

maximizes L for Specified μ

maximize L

The diagram shows the profile likelihood ratio formula. A red arrow points from the text 'maximizes L for Specified mu' to the theta-hat symbol in the numerator. Two red arrows point from the text 'maximize L' to the mu-hat and theta-hat symbols in the denominator.

The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

- The profile LR in the present analysis with variable μ and nuisance parameters θ is expected to be near optimal.

Choice of a test for discovery

If μ represents the signal rate, then discovering the signal process requires rejecting $H_0 : \mu = 0$.

- Often our evidence for the signal process comes in the form of an excess of events above the level predicted from background alone, i.e. $\mu > 0$ for physical signal models.
- So the relevant alternative hypothesis is $H_1 : \mu > 0$.
- In other cases the relevant alternative may also include $\mu < 0$ (e.g., neutrino oscillations).
- The critical region giving the highest power for the test of $\mu = 0$ relative to the alternative of $\mu > 0$ thus contains high values of the estimated signal rate.

Test statistic for discovery

Try to reject background-only ($\mu = 0$) hypothesis using

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

- only regard upward fluctuation of data as evidence against the background-only hypothesis.

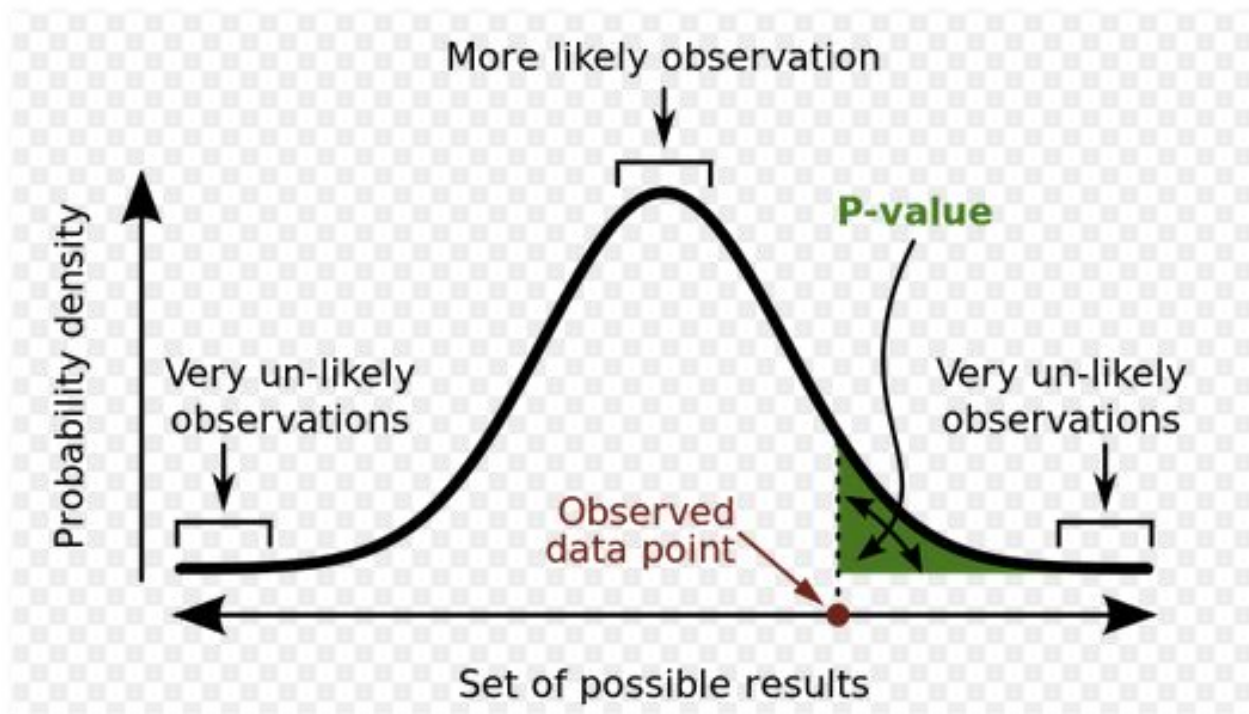
Note that even though here physically $\mu \geq 0$, we allow μ to be negative.

- In large sample limit its distribution becomes Gaussian, and this allows us to write down simple expressions for distributions of our test statistics.

p-value for a discovery

- Large q_0 means increasing incompatibility between the data and hypothesis, therefore p -value for an observed $q_{0,obs}$ is

$$p_0 = \int_{q_{0,obs}}^{\infty} f(q_0|0) dq_0$$



The look elsewhere effect

In the frequentist approach, the correct p -value of the no-peak hypothesis is the probability, assuming background only, to find a peak as significant as the one found anywhere in the search region.

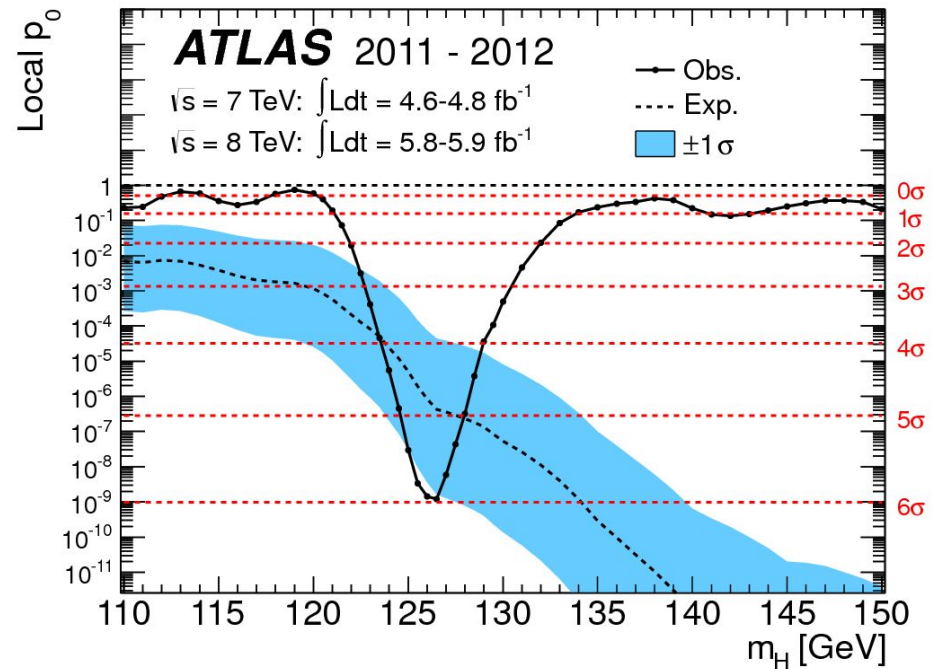
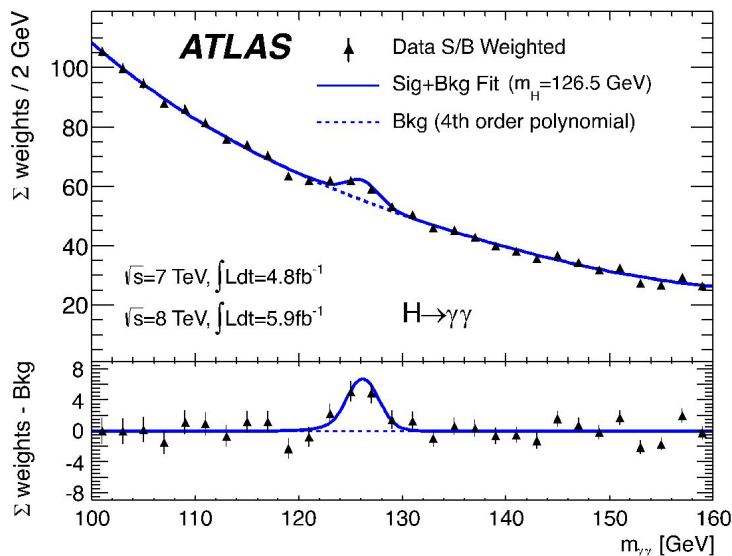
- This can be substantially higher than the probability to find a peak of equal or greater significance in the particular place where it appeared.

There is no look-elsewhere effect when considering exclusion limits.

- With exclusion, when testing many signal models (or parameter values) we might exclude some even in the absence of signal (“spurious exclusion”)
- <https://xkcd.com/882/>

How to read the p_0 plot

- The “local” p_0 means the p -value of the background-only hypothesis obtained from the test of $\mu = 0$ at each individual m_H , without any correct for the Look-Elsewhere Effect.
- The “Sig. Expected” (dashed) curve gives the median p_0 under assumption of the SM Higgs ($\mu = 1$) at each m_H .



arXiv:1207.7214

Choice of a test for exclusion

Suppose the existence of the signal process ($\mu > 0$) is not yet established.

- The interesting alternative in this context is $\mu = 0$.
- We want to know what values of μ can be excluded on the grounds that the implied rate is too high wrt to what is observed in the data.
- The critical region giving the highest power for the test of μ relative to the alternative of $\mu = 0$ contains low values of the estimated rate.
- Test based on one-sided alternative \rightarrow upper limit.

- In other cases (not treated here) we want to exclude μ on the grounds that some other measure of incompatibility between it and the data exceeds some threshold. E.g. the process may be known to exist, and thus $\mu = 0$ is no longer an interesting alternative.

Test statistic for exclusion

For purposes of setting an upper limit on μ one may use

$$q_{\mu} = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

- When setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ .
- From observed q_{μ} find p-value:

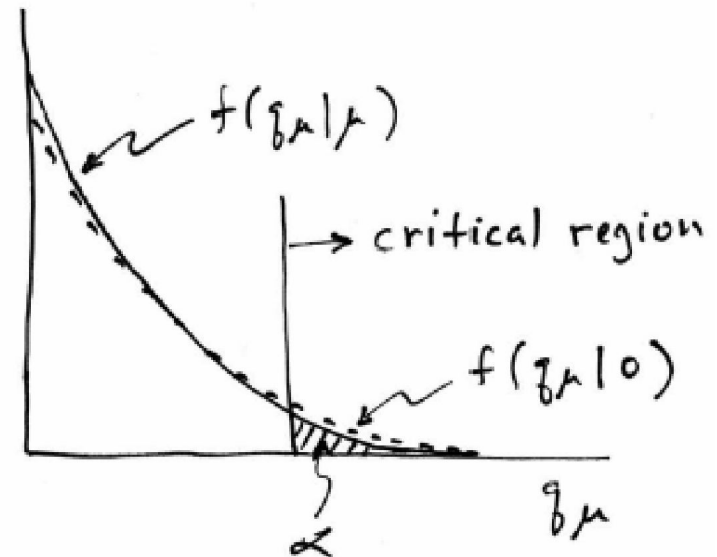
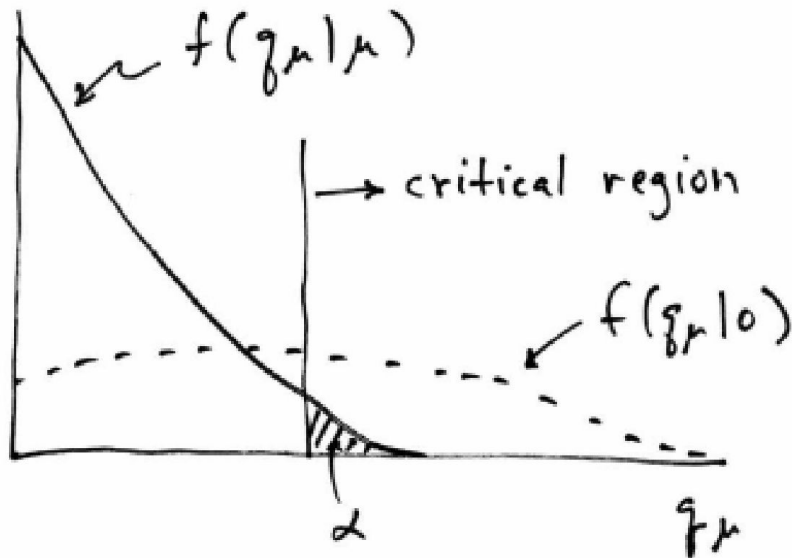
$$p_{\mu} = \int_{q_{\mu, \text{obs}}}^{\infty} f(q_{\mu} | \mu) dq_{\mu}$$

- The 95% CL upper limit on μ is highest value for which p-value is not less than 0.05.

Spurious Exclusion

Use the power (P to reject μ if $\mu = 0$) as measure of sensitivity

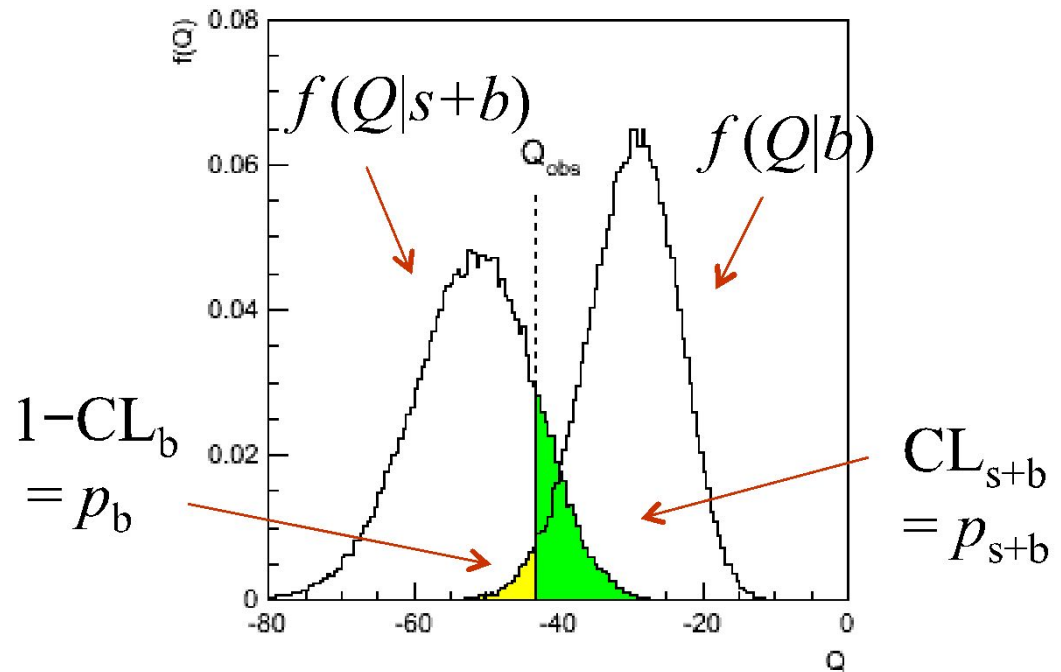
- Having sensitivity to μ means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ are well separated. The power is substantially higher than α .
- In the case of low sensitivity, the power is only slightly greater than α .
- “Spurious exclusion”: with probability of around $\alpha = 5\%$ (slightly higher), one excludes hypotheses to which one has essentially no sensitivity



The CLs procedure

The CLs solution (A. Read et al.) is to base the test not on the usual p-value (CL_{s+b}), but rather to divide this by CL_b (one minus the p-value of the b-only hypothesis)

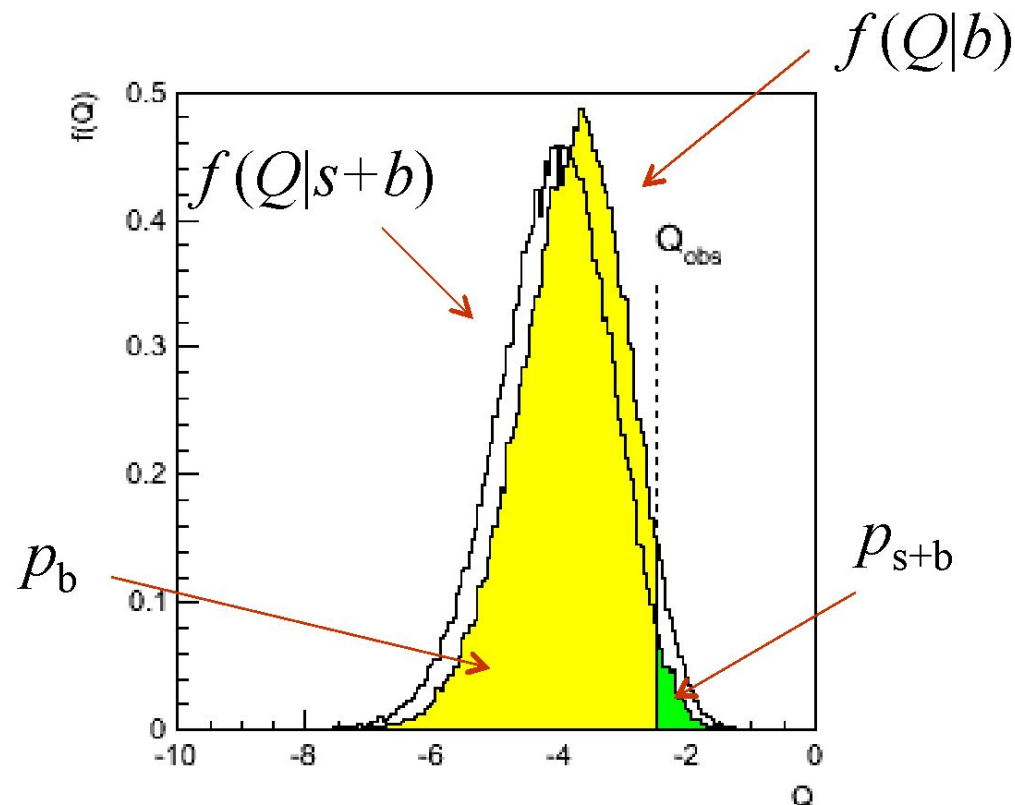
- Reject the s+b hypothesis if $CL_s = (CL_{s+b}/CL_b) < \alpha$



- Increases the “effective” p-value when the two distributions become close (prevents exclusion if sensitivity is low).

The CLs procedure

- Like we saw before, having low sensitivity means that the distributions of Q under b and $s+b$ are very close.

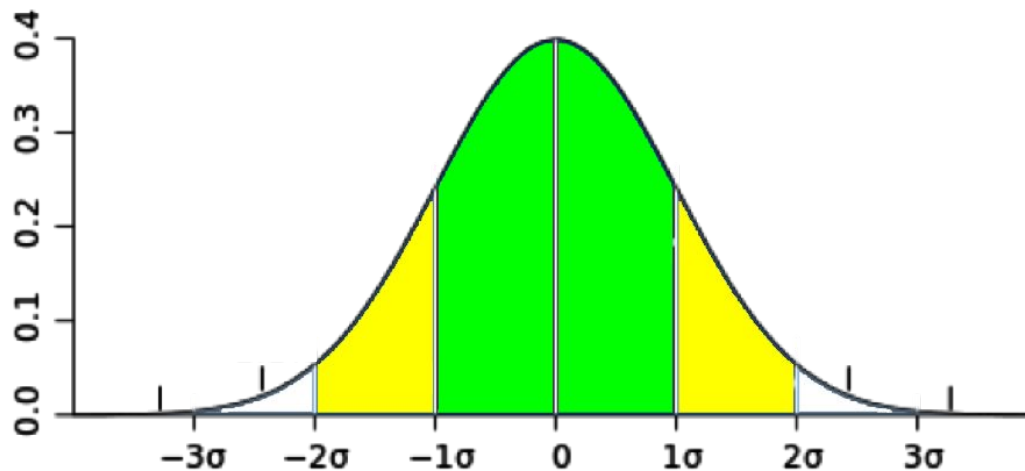


Setting upper limits

Carrying out the CLs procedure for the parameter $\mu = \sigma/\sigma_{\text{theory}}$ results in an upper limit μ_{up} .

E.g. in the Higgs search, this was done for each value of m_H .

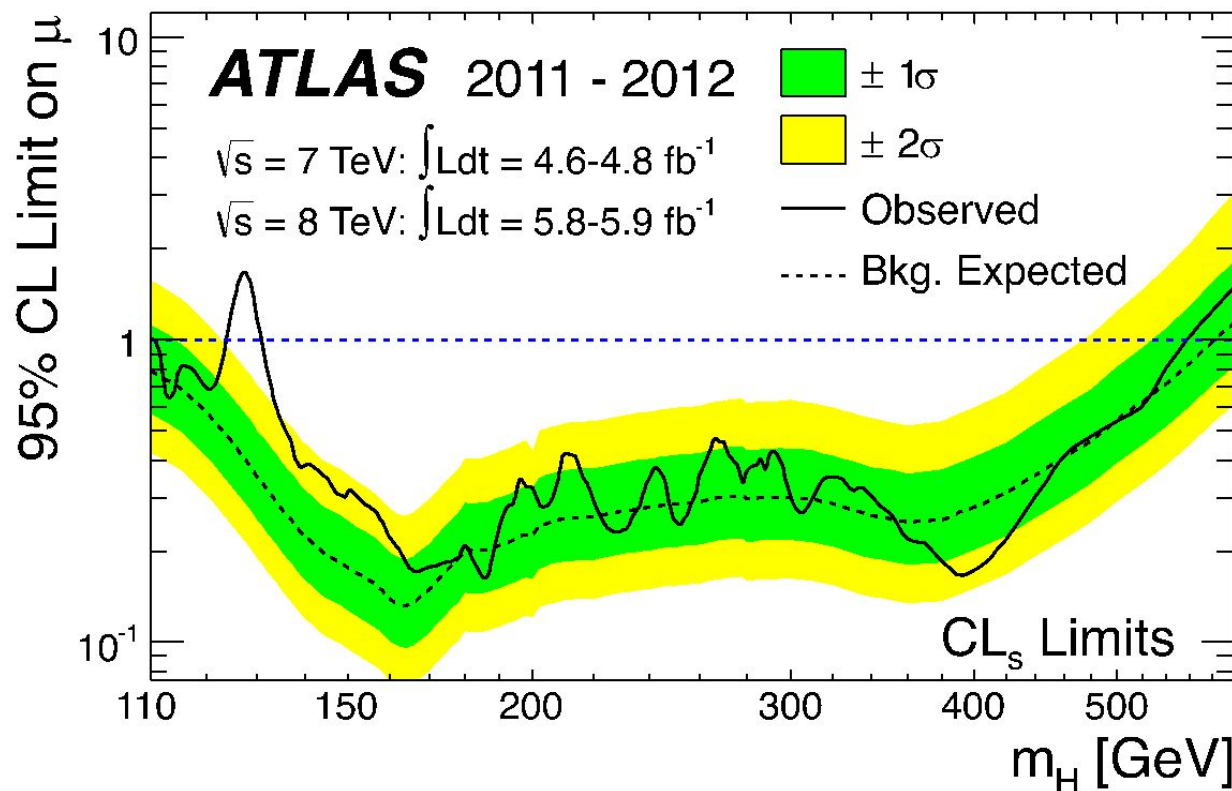
- At a given value of m_H , we have an observed value of μ_{up} , and we can also find the distribution $f(\mu_{\text{up}}|0)$.
- $\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands



How to read exclusion plots

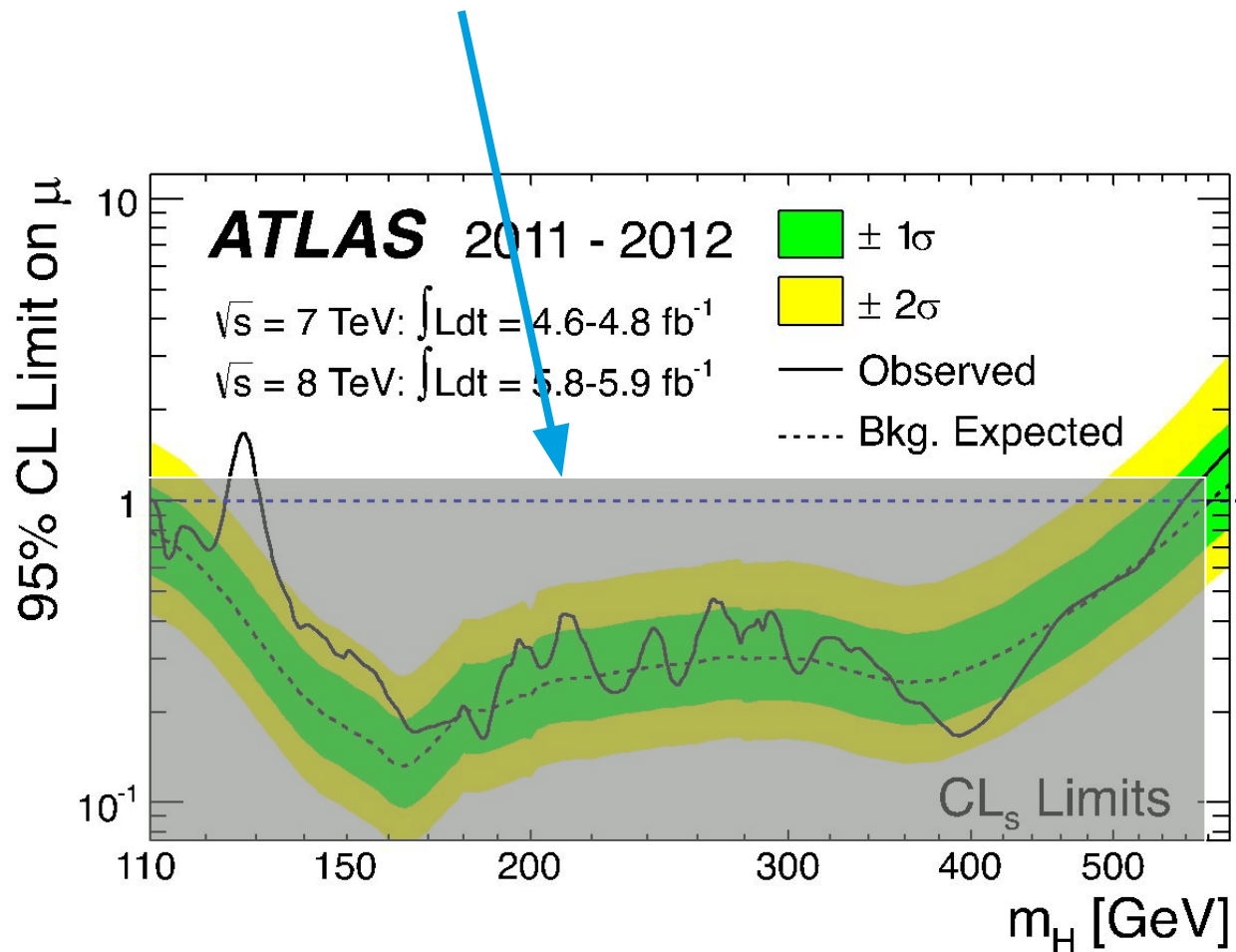
For every value of m_H , find:

- The CLs upper limit on μ and the distribution of upper limits μ_{up} for $\mu = 0$.
- The dashed curve is the median μ_{up} , and the green (yellow) bands give the $\pm 1\sigma$ (2σ) regions of this distribution.



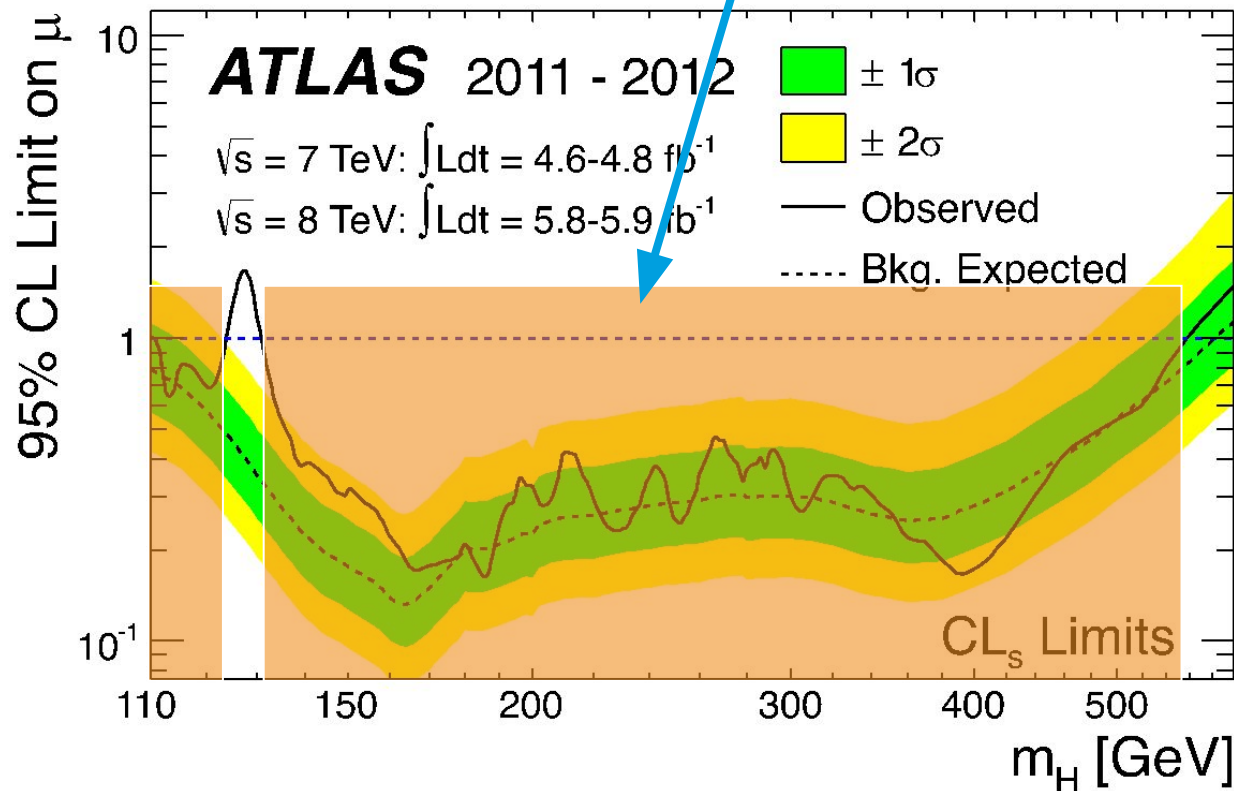
How to read exclusion plots

Expected excluded mass range



How to read exclusion plots

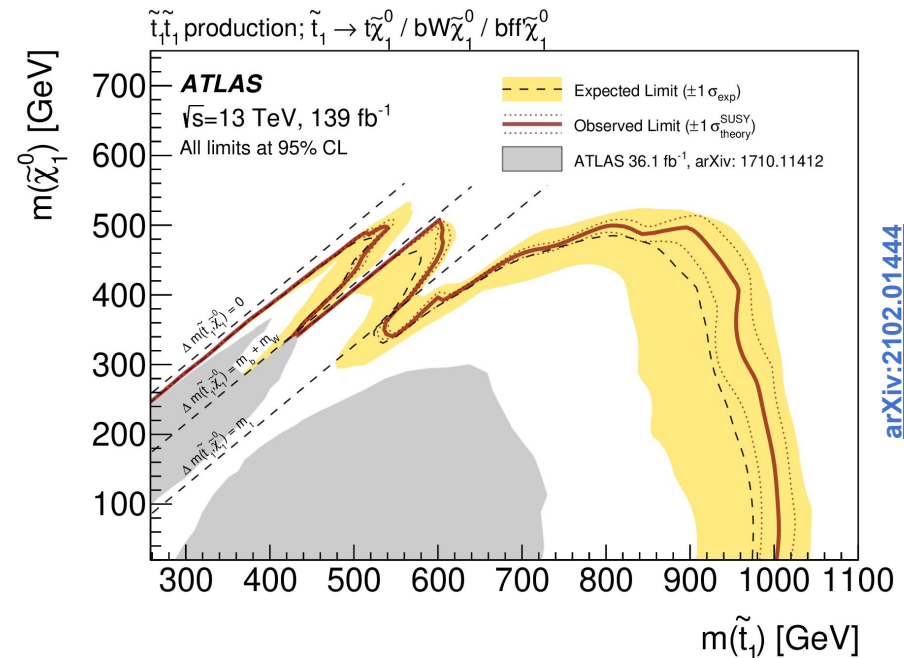
Observed excluded mass range



How to read 2D exclusion plots

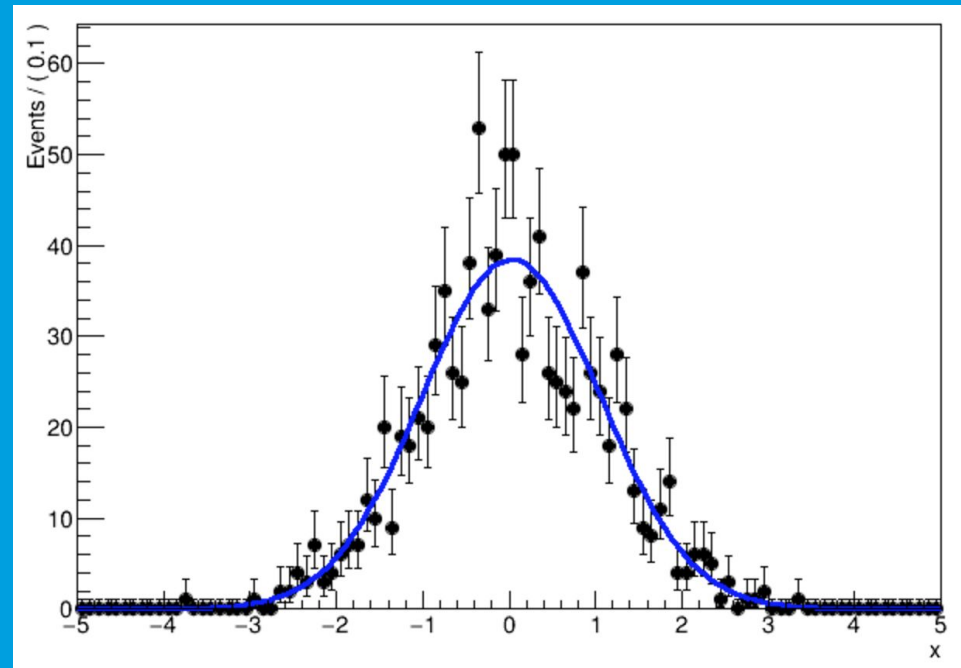
A typical SUSY exclusion plot.

- More complex signals with dependencies from multiple unknown parameters.
- Often the two axes represent two sparticle masses.
- Here, for every choice of $m(\chi_2)$, $m(\chi_1)$ find the CLs upper limit on μ .
- The lines and bands show the contours of $\mu = 1$ (or CLs = 0.05)
- The dashed curve is the median $\mu_{\text{up}} = 1$, with the yellow bands giving the $\pm 1\sigma$ regions (for SM uncertainties)
- Dashed red lines are the $\pm 1\sigma$ regions (for signal theory uncertainties)



Parameter estimation

- Parameter fitting
- Properties of estimators
- Likelihood and maximum likelihood estimators



Estimators and parameter estimation

The central problem of statistics is to infer the properties of $f(x)$ based on a sample of observations $\mathbf{x} = (x_1, \dots, x_n)$.

- Specifically, one would like to construct functions of the x_i to estimate the various properties of the p.d.f. $f(x)$.
- The parameters of a pdf are constants that characterize its shape, e.g.

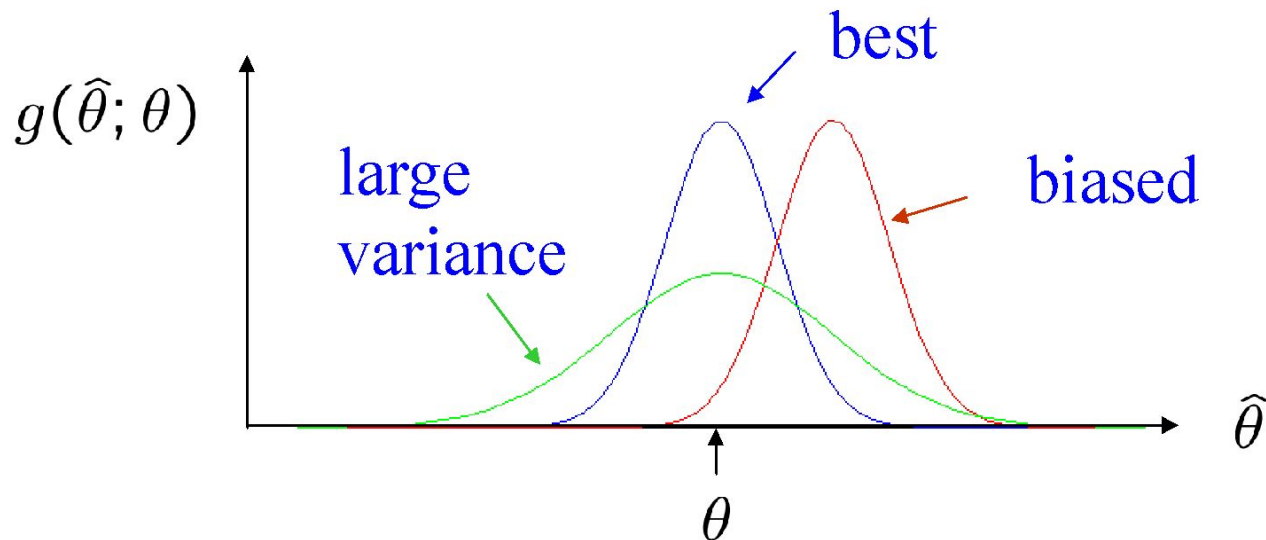
$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$$

random variable parameter

- $\hat{\theta}(\vec{x})$ estimators are written with a hat
- Estimator is typically the function of x_1, \dots, x_n
- Estimate the value of the estimator with a particular data set.

Estimator properties

The estimates from each experiment repetition follow a pdf:



We want **small (or zero) bias** (systematic error; $b = E[\hat{\theta}] - \theta$)

- average of repeated measurements should tend to true value

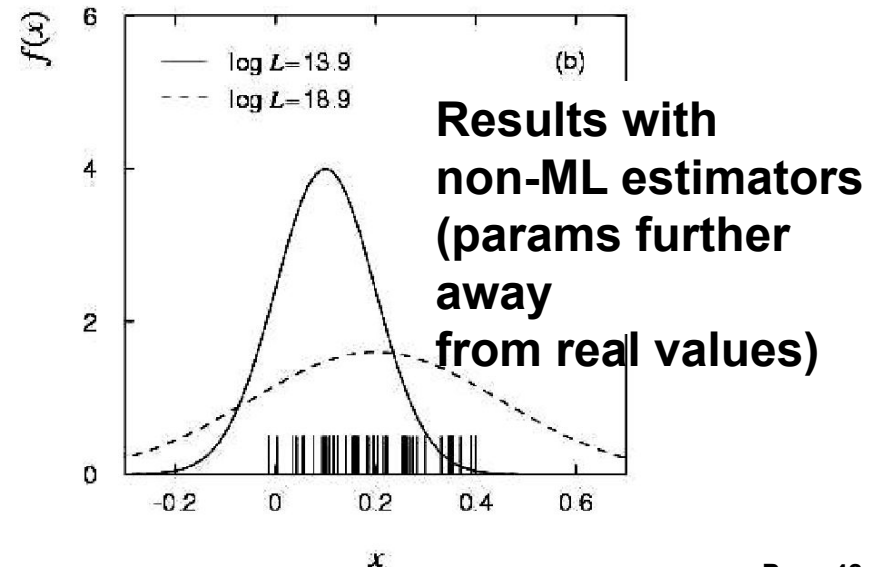
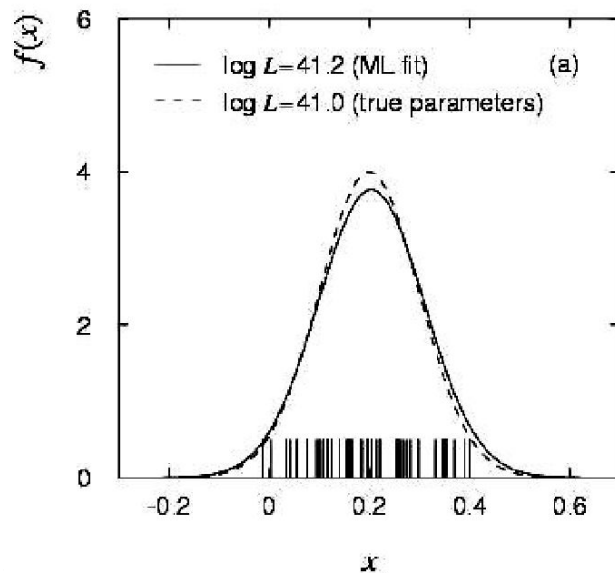
And we want a **small variance** (statistical error):

- small bias & variance are in general conflicting criteria

Maximum likelihood estimators

The maximum likelihood (ML) estimators for the parameters are those which maximize the likelihood function.

- If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.
- ML estimators are not guaranteed to have any 'optimal' properties, (but in practice they're very good).



Example

Assume our data is distributed according to a Gaussian(μ, σ).

- Let's compute the MLE for μ
- Find the minimum of the $-\ln L(x | \mu, \sigma)$
- The MLE estimator for μ is the sample mean

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

$$\mathcal{L}(x) = \prod_{i=1}^n f(x_i|\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$-\ln \mathcal{L}(x) = -\sum_{i=1}^n \ln f(x_i|\mu, \sigma) = -\sum_{i=1}^n \left\{ \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

$$\frac{\partial}{\partial \mu} (-\ln \mathcal{L}(x)) = 0$$

$$\frac{\partial}{\partial \mu} (-\ln \mathcal{L}(x)) = -\sum_{i=1}^n \frac{\partial}{\partial \mu} \left(\left\{ \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right) =$$

$$= -\sum_{i=1}^n \frac{2(x_i - \mu)}{2\sigma^2} = -\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

$$-\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\sum_{i=1}^n x_i - n\mu = 0$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

MACHINE LEARNING

- Basic terminology
- Classical approaches to prediction
- Bias-variance trade-off
- Introduction to Neural Networks

“In God we trust, all others bring data.”

- William Edwards Deming

Basic Terminology

The goal of machine learning is to predict results based on incoming data.

Features (also parameters, or variables): these are the factors for a machine to look at. E.g.: cartesian coordinates, pixel colors, a car mileage, user's gender, stock price, word frequency in the text.

- Quantitative ($x = \{1.02, 0.21, 0.12, 2\}$)
- Qualitative *discrete* ($x = \{\text{medium, small, large}\}$) or *categorical* ($x = \{\text{red, blue, green}\}$)

Algorithms (also models): Any problem can be solved in different ways. The method you choose affects the precision, performance, and size of the final model.

- If the data is insufficient/inappropriate (e.g. statistically limited or missing important features), even the best algorithm won't help. Pay attention to the accuracy of your results only when you have a good enough dataset.

CLASSICAL MACHINE LEARNING

Data is pre-categorized
or numerical

SUPERVISED

Predict
a category

CLASSIFICATION

«Divide the socks by color»



Predict
a number

REGRESSION

«Divide the ties by length»



OUR FOCUS

Data is not labeled
in any way

UNSUPERVISED

Divide
by similarity

CLUSTERING

«Split up similar clothing
into stacks»



Identify sequences

Find hidden
dependencies

ASSOCIATION

«Find what clothes I often
wear together»



DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»



Image credit: https://vas3k.com/blog/machine_learning/

Prediction: Least squares

The linear model is one of our most important tools in statistics.

- Given a vector of inputs $X^T = (X_1, X_2, \dots, X_p)$, we predict the output Y via

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j$$

- The term β_0 is the intercept, also known as the *bias* in machine learning

How do we fit the linear model to a set of data?

- The most popular method is the method of least squares: pick the coefficients β to minimize the residual sum of squares (RSS)

$$\text{RSS}(\beta) = \sum_{i=1}^N (y_i - x_i^T \beta)^2$$

- $\text{RSS}(\beta)$ is a quadratic function of the parameters, and hence its minimum always exists, but may not be unique.

Prediction: Least squares

Linear Regression of 0/1 Response

Data were simulated with outputs being either **BLUE** or **ORANGE**.

A linear regression model was fit to the data, used here as *training dataset*.

The fitted values \hat{Y} are converted in a classification according to

$$\hat{G} = \begin{cases} \text{ORANGE} & \text{if } \hat{Y} > 0.5 \\ \text{BLUE} & \text{if } \hat{Y} \leq 0.5 \end{cases}$$

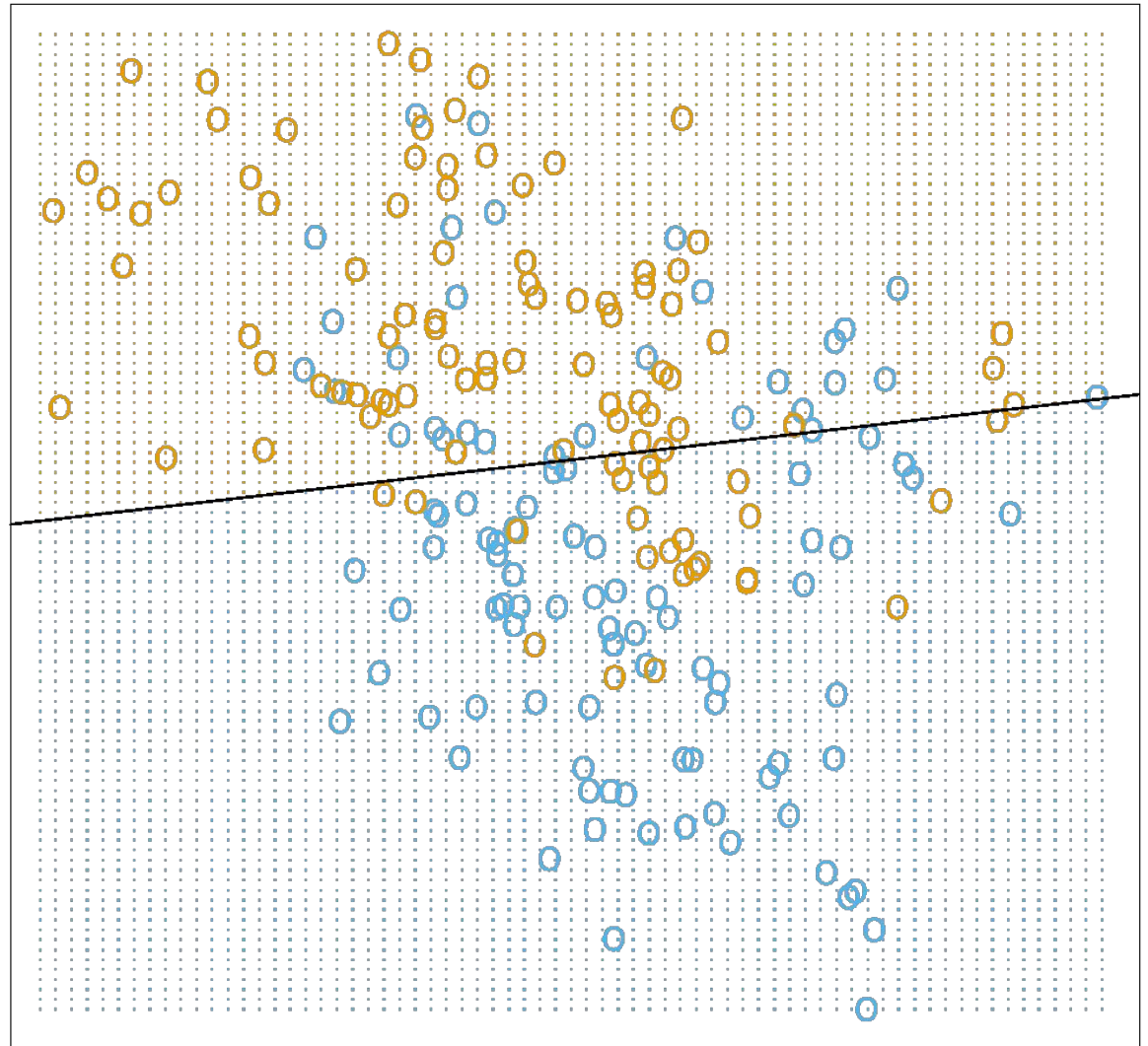


Image from [4]

Prediction: nearest neighbor classifier

An alternative algorithm for classification is the method of nearest neighbors.

Nearest-neighbor methods use those observations in the training set closest in input space to x to form Y .

The k -nearest neighbor fit for Y is defined as:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

where $N_k(x)$ is the neighborhood of x defined by the k closest points x_i in the training sample.

Closeness implies a metric, which in our case we assume is Euclidean distance.

In words, we find the k observations with x_i closest to x in input space, and average their responses.

Prediction: nearest neighbor classifier

15-Nearest Neighbor Classifier

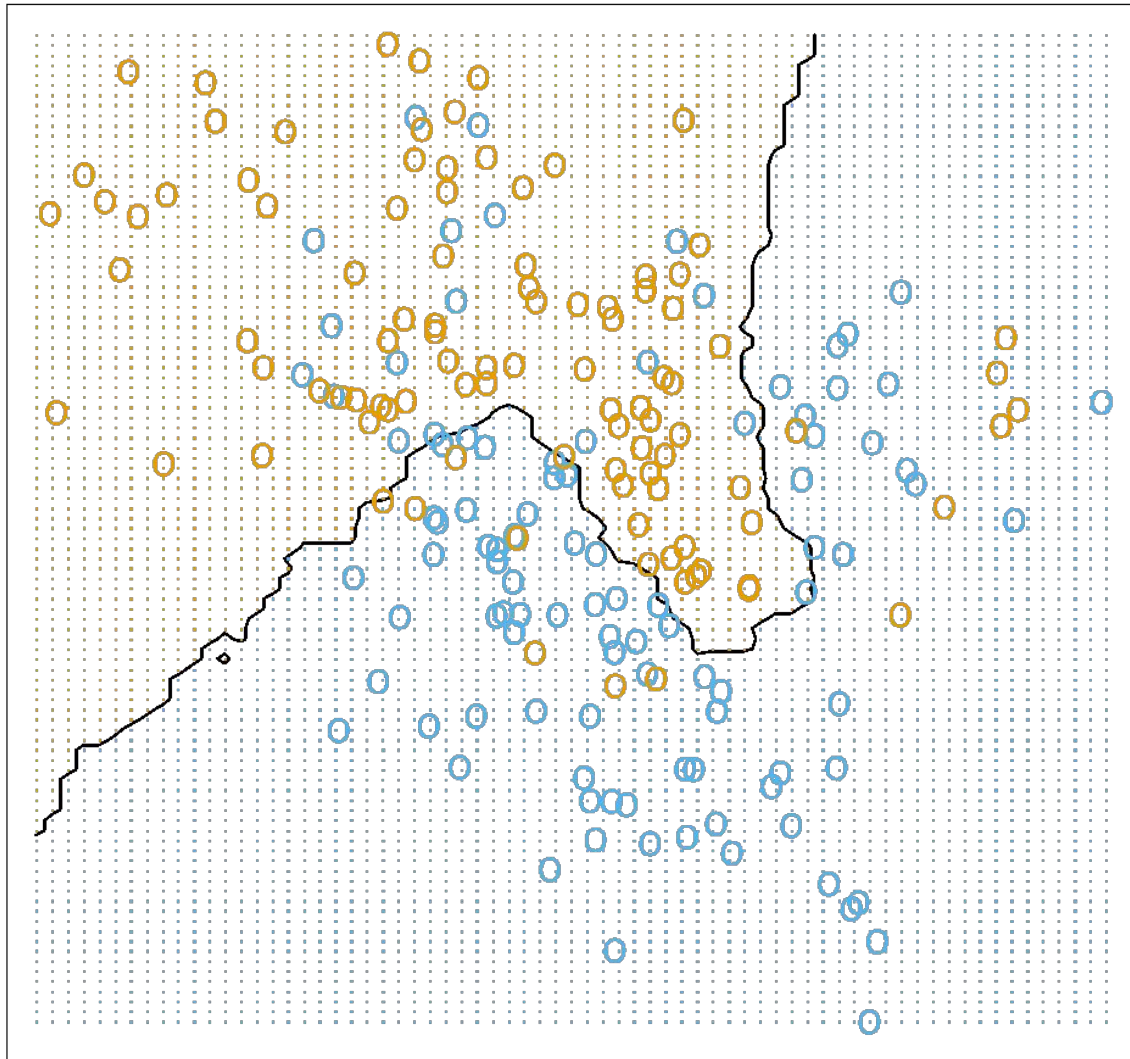


Image from [4]

Perfect classification?

1-Nearest Neighbor Classifier

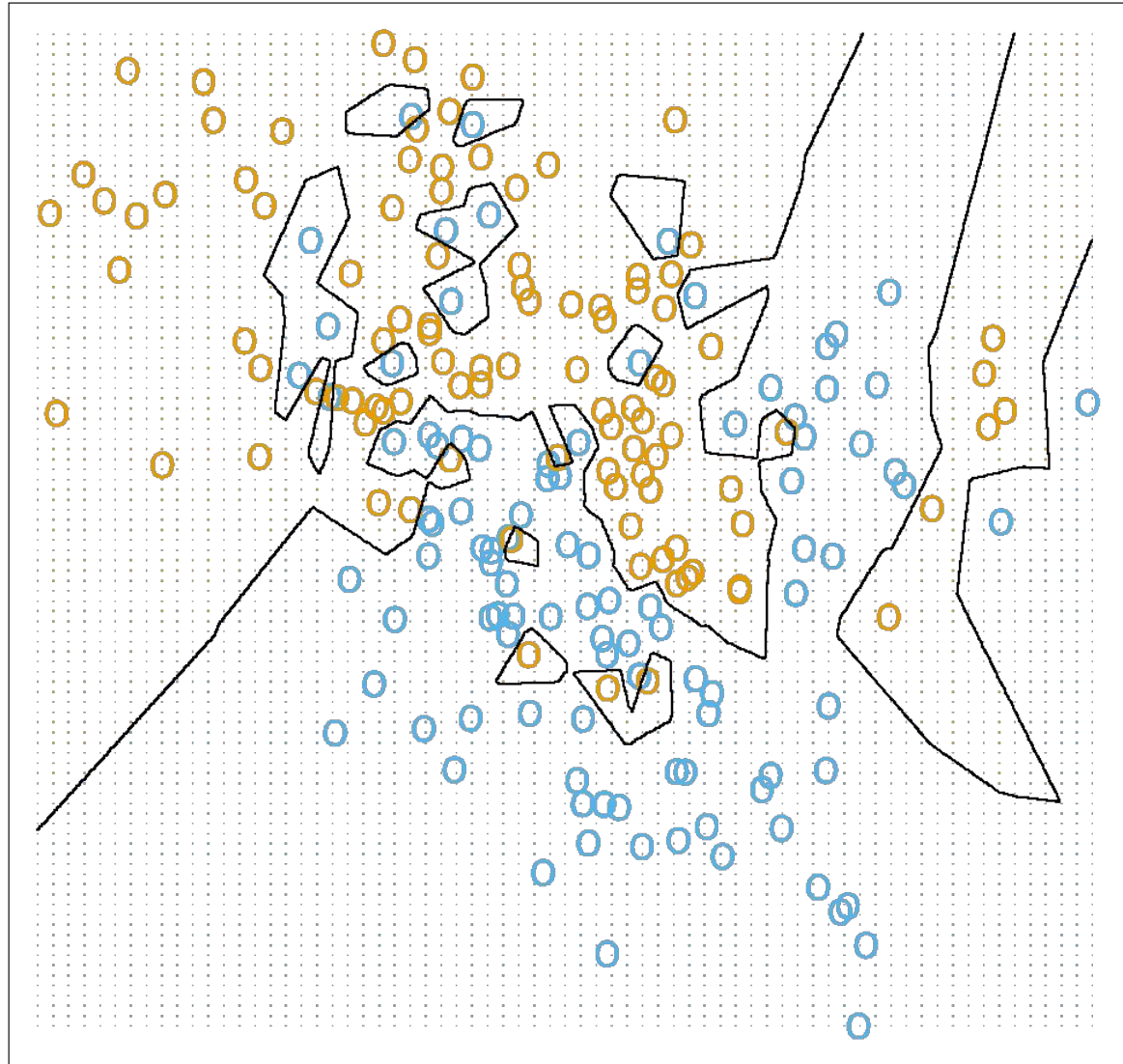


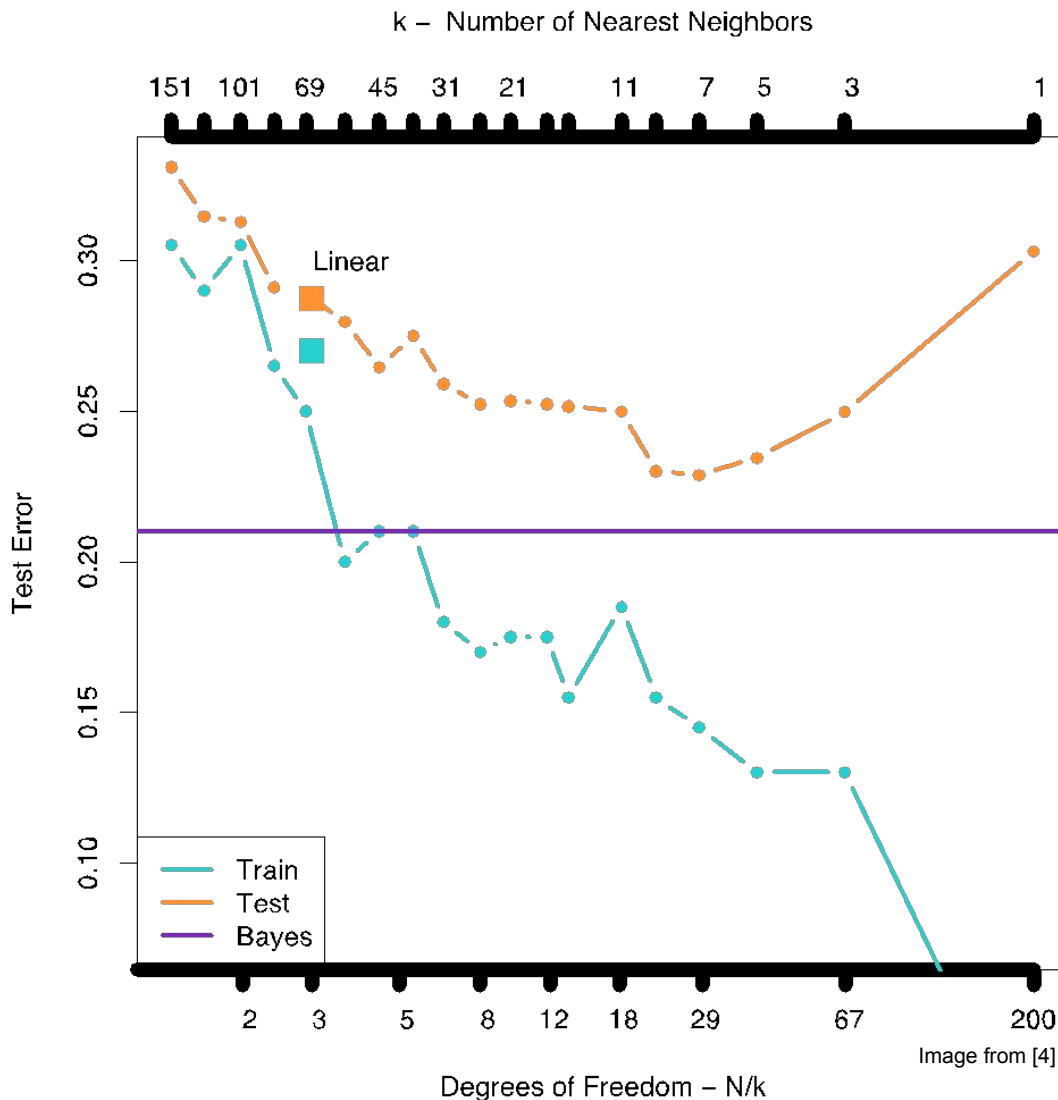
Image from [4]

Comparison

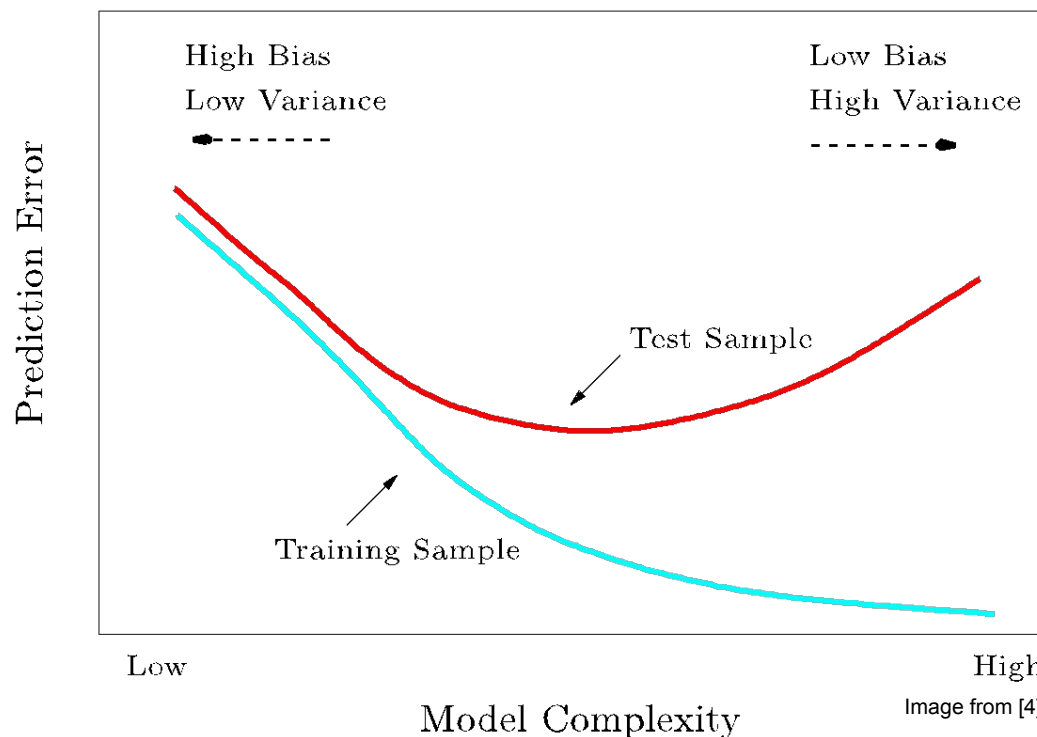
To compare the different algorithms, let's define a *loss* (or cost) criterion.

- Here, we can take the rate of misclassifications

In order to compare the performances, let's introduce a second, independent, dataset to evaluate the performance: the *test dataset*.



Bias-variance tradeoff



The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder.

- With too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error).
- In contrast, if the model is not complex enough, it will underfit and may have large bias, again resulting in poor generalization.

Where are the fancy neural networks?

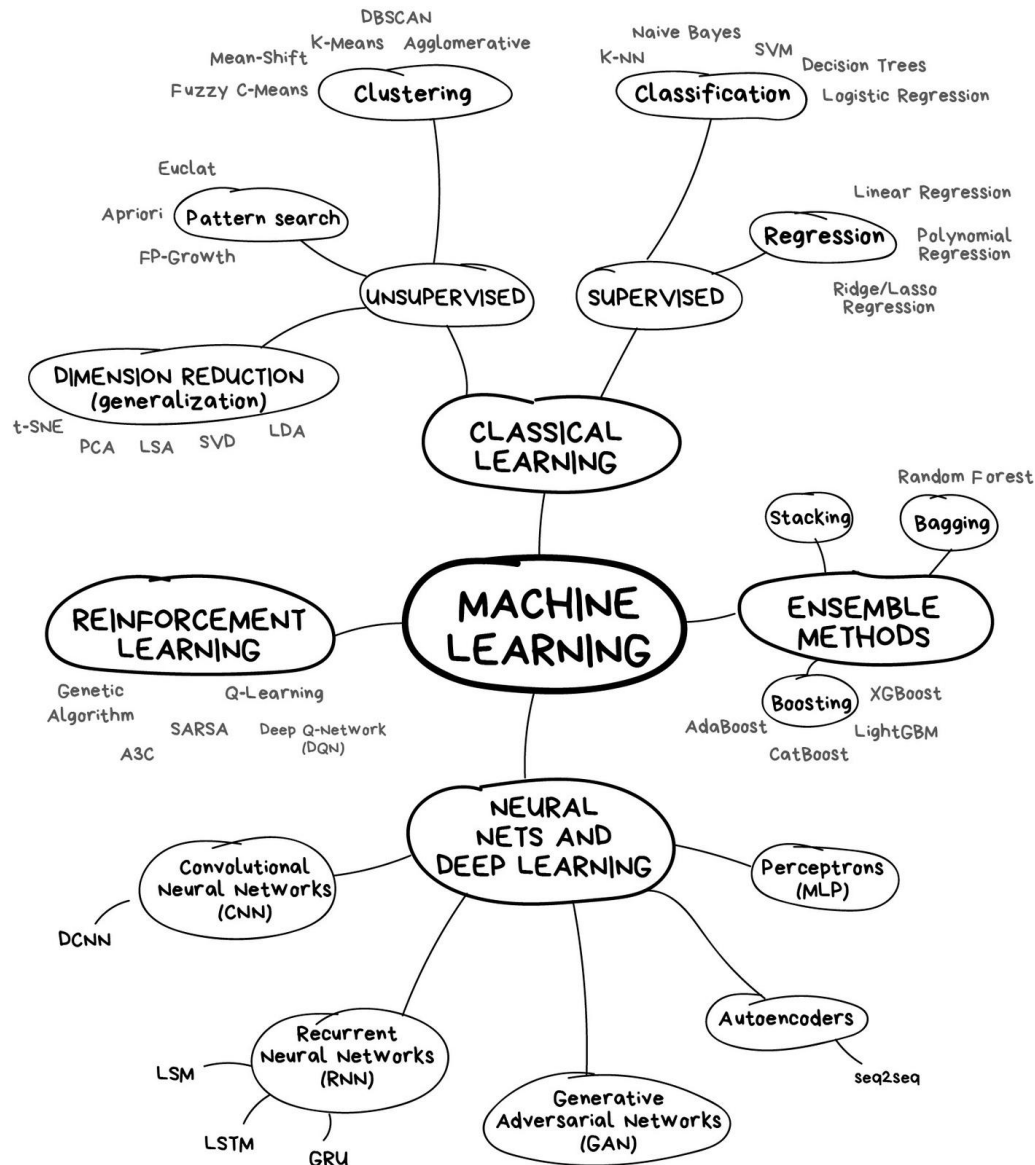


Image credit: https://vas3k.com/blog/machine_learning/

Neural Networks

Any neural network is a collection of **neurons** and **connections** between them.

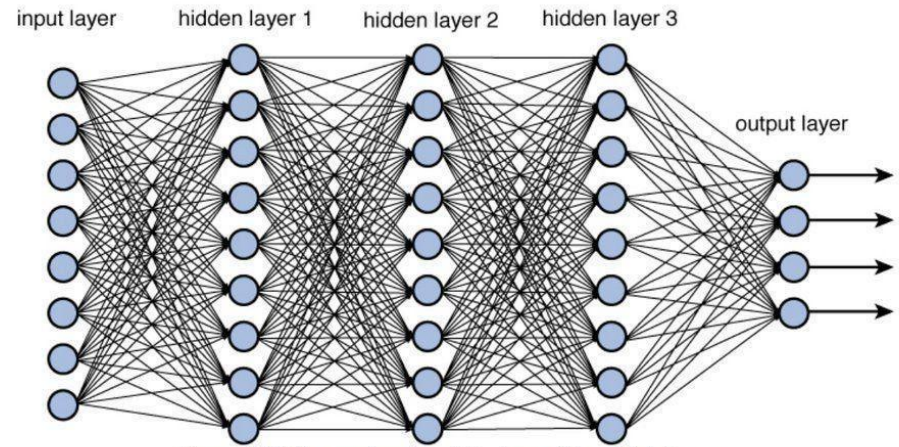
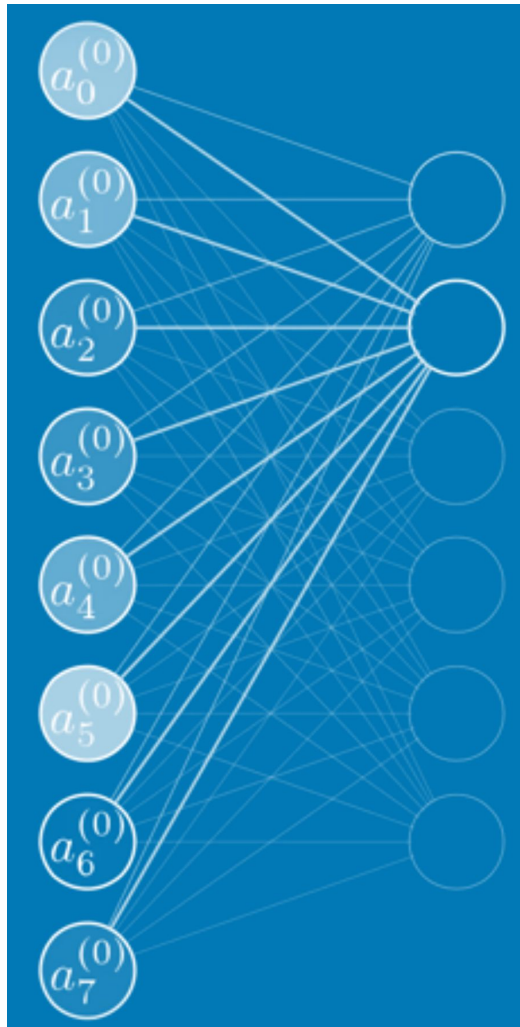
Neuron is a function with a set of inputs and one output. Its task is to take all numbers from its input, apply a function on them and send the result to the output.

- Example: sum up all numbers from the inputs and if that sum is bigger than N give 1 as a result. Otherwise return zero.

Connections are like channels between neurons. They connect outputs of one neuron with the inputs of another so they can send digits to each other. Each connection has only one parameter the *weight*.

- These weights tell the neuron to respond more to one input and less to another. Weights are adjusted when training — that's how the network learns.

How do NNs work?



layer

$$a_0^{(1)} = f(w_{0,0} a_0^{(0)} + w_{0,1} a_1^{(0)} + \dots + w_{0,n} a_n^{(0)} + b_0)$$

activation function

weights

bias

$$\begin{bmatrix} a_0^{(1)} \\ \dots \\ a_n^{(1)} \end{bmatrix} = f \left(\begin{bmatrix} w_{0,0} & \dots & w_{0,n} \\ \vdots & \ddots & \vdots \\ w_{k,0} & \dots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ \dots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ \dots \\ b_n \end{bmatrix} \right)$$

How do NNs learn?

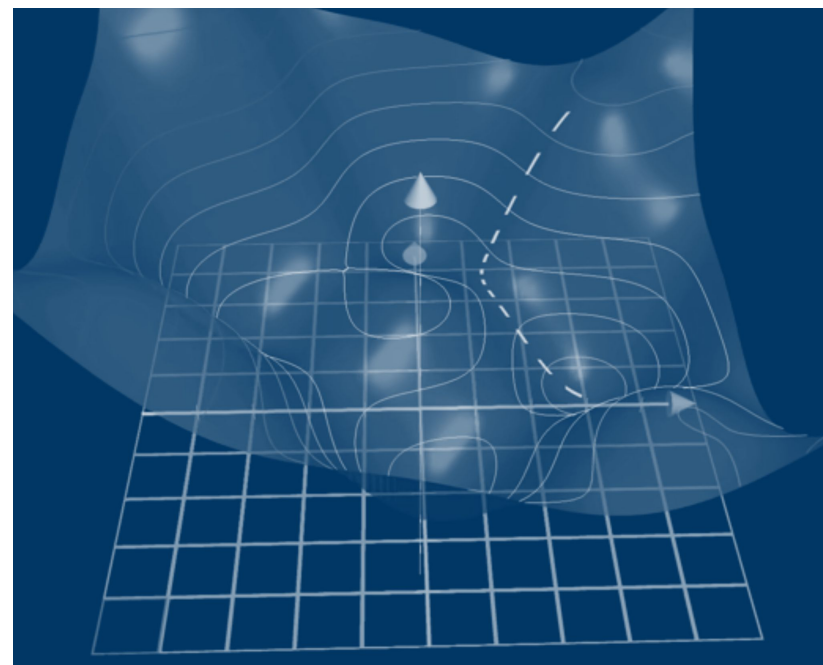
After we constructed a network, our task is to **assign proper weights** so neurons will react correctly to incoming signals.

- define a loss function to measure how far the response is from the truth

This function is a function of all the weights and biases in the NN (a priori a very large number), and the goal of training is to find its minimum.

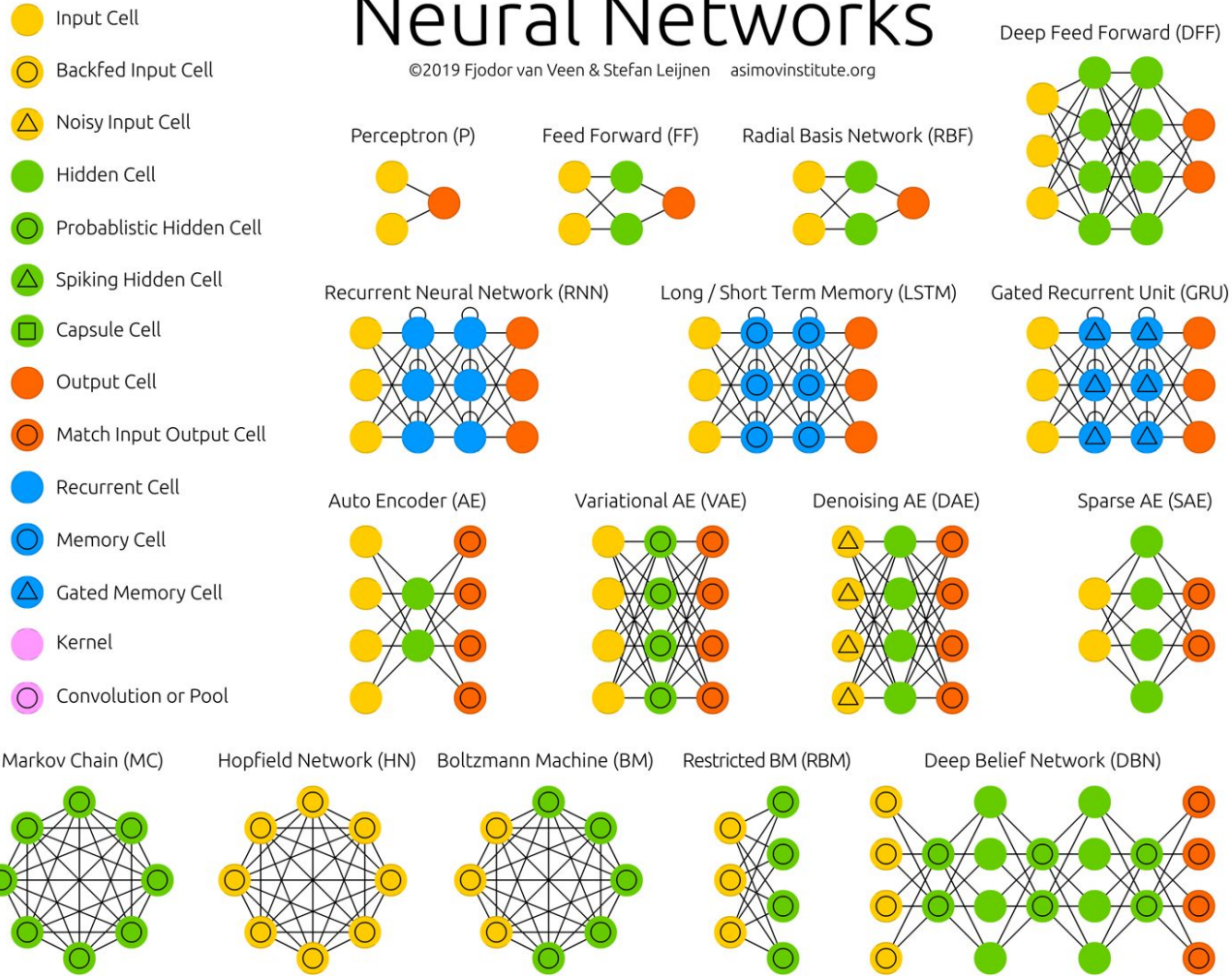
- To start with, all weights are assigned randomly.
- After evaluating the NN on the training dataset, we can compute all the per-neuron differences with respect to the correct result.
- Computing the gradient of the loss, gives us a direction in which to tune the weights towards a local minimum

The process of correcting the weights is called **backpropagation of an error**.



A mostly complete chart of Neural Networks

©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org



There are many more...

We'll see some together in the lab tomorrow!

Done!

Thanks a lot for
your attention!

Questions?

I'm here until Friday! Or:
federico.meloni@desy.de

