Modern Methods of Statistical Data Analysis

From parameter estimation to deep learning — A guided tour of probability

Lecture 6

-

Hypothesis Tests & Neyman Pearson

P.-D. Dr. Roger Wolf

roger.wolf@kit.edu

Dr. Pablo Goldenzweig

pablo.goldenzweig@kit.edu

Dr. Jan Kieseler

jan.kieseler@kit.edu

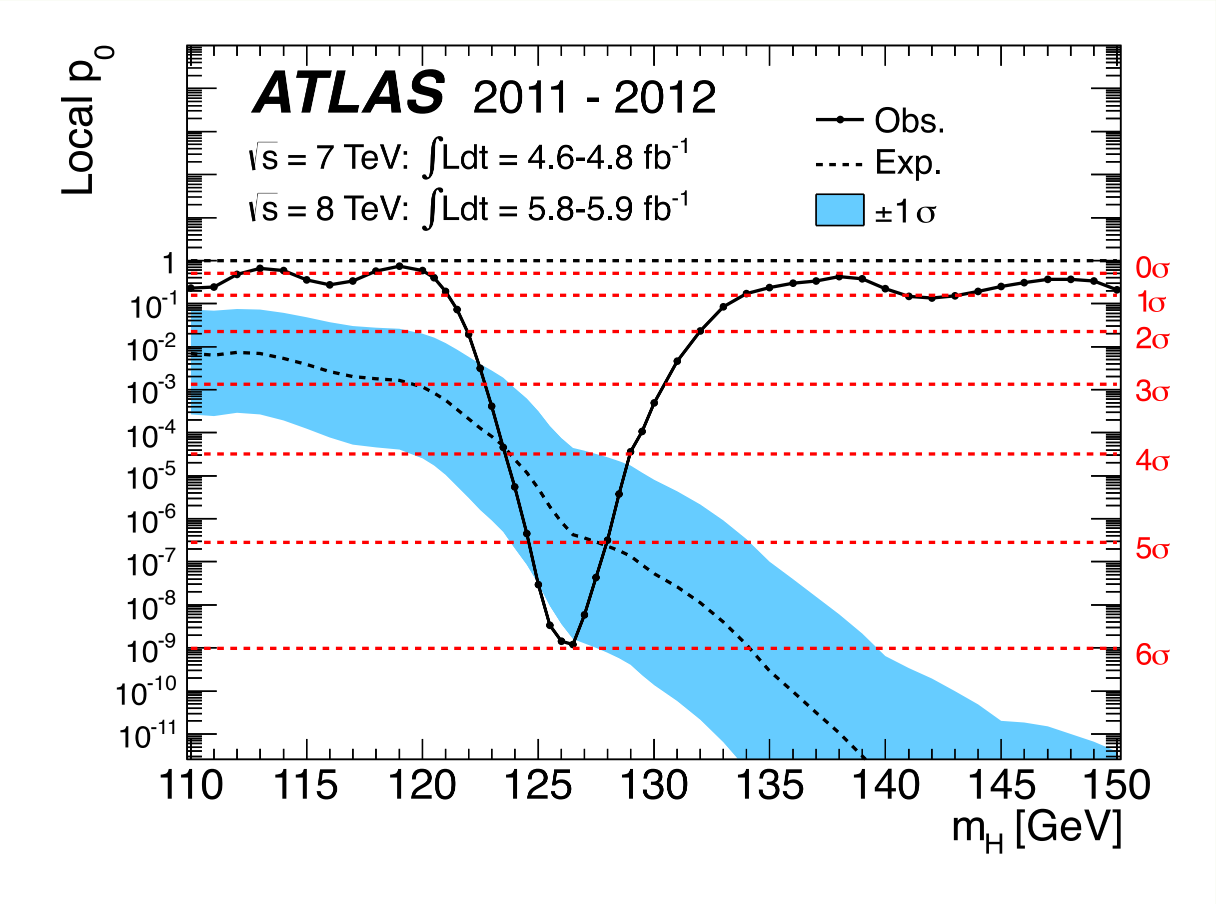
Dr. Slavomira Stefkova

slavomira.stefkova@kit.edu

Karlsruher Institut für Technologie

Today

- Hypothesis testing
  - Particle selection example
- Neyman-Pearson Lemma
- Fisher discriminant function
- Higgs discovery & significance with a P-value



Estimators and Parameter Estimation

Review

Let $x_1, x_2, \ldots, x_n$ be n independent measurements with unknown mean μ (e.g., mass) and variance σ^2. The **estimators** (denoted with $\hat{}$) are:

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

Maximum Likelihood (ML) estimators (MLE) maximize the likelihood function for given data x

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) \qquad \frac{\partial \ln \mathscr{L}}{\partial \theta_i} = 0, \quad i = 1, \ldots, m$$

$\ln \mathscr{L}$ is often more convenient to work with than $\mathscr{L}$, and doesn't change the estimation

Poisson example: MLE for the mean is just the data count [the usual arithmetic mean (avg.) estimator]

Least Squares (LS) χ^2 estimator finds the model parameters that minimize the total squared deviations of the data points (x_i, y_i) from the mean

$$\chi^2(\boldsymbol{\theta}) = -2\ln\mathscr{L}(\boldsymbol{\theta}) + \text{constant} = \sum_{i=1}^{N}\frac{\left(y_i - \lambda(x_i;\boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

Same as $-2\ln\mathscr{L}$ for Gaussian and independent y_i, so you often see $-2\ln\mathscr{L}$ for comparison

When fitting a histogram with Poisson errors ALWAYS perform a ML fit (not a χ^2 fit)

***MLE** better than **LS** at low statistics, but numerical optimization may take longer*

But numerical estimation not really an issues anymore, so ML is the method of choice overall

MLE vs. LS

- For **MLE**, the RCF inequality **turns into an equality**, so the RCF bound is indeed reached. For LS, the RCF bound can only be reached if Gaussian. If not, the estimator variance will be larger than it could be.

- For MLE, you have optimal coverage of a confidence interval (more next week). I.e., 68% really means "68%."

- The inclusion of systematic errors into the likelihood via nuisance parameters (more later today) is straightforward with MLE. Not so easy with LS.

- Don't need to worry about binning with MLE. Binned converges to unbinned when $n_{\mathrm{bins}} \to \infty$.

- When you use LS (for large enough event counts), $-2 \ln \mathscr{L}$ is approximated by the χ^2 distribution. As χ^2 is known, there's no need for large MC.

Answer Time: Quiz 5

Combining measurements with least squares

a) Calculate the average of two measured quantities $y_1 = 5$ and $y_2 = 6$ with a covariance matrix

$$C = \begin{pmatrix} 0.5. & 0.2 \\ 0.2 & 0.7 \end{pmatrix},$$

first using the uncorrelated formula Eq. 7.26 and then using the proper expression Eqs. 7.29, 7.30 Remember that $\sigma_i^2 = V_{ii}$.

Reduces to Eqs. 7.34 - 7.39 for the case of 2 measurements

- Uncorrelated average: 5.41667
- Correlated average: 5.375

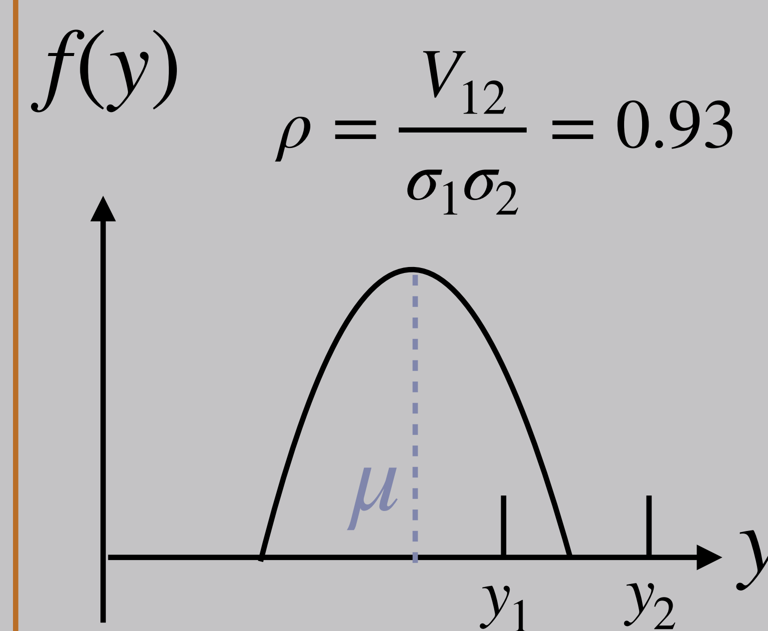
b) Calculate the correlated average of the same measured quantities but assume now that the covariance matrix is given by

$$C = \begin{pmatrix} 0.5. & 0.55 \\ 0.55 & 0.7 \end{pmatrix}.$$

Why is the average (you should get $\hat{\lambda} = 4.5$) not between 5 and 6? What is the variance of the average?

- Values are highly correlated (93%), i.e. very likely that true mean lies on the opposite side of the value with the smaller error
- Variance: $\dfrac{1}{\sigma^2} = \dfrac{1}{1-\rho^2}\left[\dfrac{1}{\sigma_1^2} + \dfrac{1}{\sigma_2^2} + \dfrac{2\rho}{\sigma_1\sigma_2}\right] \Rightarrow \sigma = \sqrt{0.475} \approx 0.69$

$$\rho = \frac{V_{12}}{\sigma_1\sigma_2} = 0.93$$



If $\rho > \sigma_1/\sigma_2$, the weight $w < 0$, which means the weighted avg. does not lie between y_1 and y_2

Read more in Cowan section 7.6 & 7.61

P-values

a) What is the interpretation of a p-value and what does a low p-value imply?

P-value: Probability to observe a LS fit result with a χ^2 as large or larger than the observed value, given the underlying (fit) model is true

b) Using the χ^2_{obs} table below decide if the following binned LS or χ^2 fits describe the observed data well:

# of free parameters in the fit

$$(\chi^2_{\text{obs}}, n, m) = (6.2, 5, 3), \qquad (\chi^2_{\text{obs}}, n, m) = (1.2, 2, 1), \qquad (\chi^2_{\text{obs}}, n, m) = (2.2, 10, 3).$$

# of points

Degrees of freedom (df)	χ^2 value[20]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

P-values

$\approx$ 0.30 - 0.20 (dof = 1) **OK**

$\approx$ 0.05 - 0.01 (dof = 2) **Not well**

$\approx$ 0.95 - 0.90 (dof = 7) **Agreement maybe *too good?***

Grounds for checking that the errors have not been overestimated or are not correlated

Hypothesis tests

Hypothesis tests (i)

Goal of a statistical test: Statement about how well the observed data stand in agreement with a given predicted probability = ***Hypothesis!***

Hypothesis under consideration: Null hypothesis or H_0

could be a PDF $f(x)$

If $f(x)$ fully determined → *Simple hypothesis (focus on these for now)*

If $f(x) = f(x; \boldsymbol{\theta})$ → *Composite hypothesis ($\boldsymbol{\theta}$ determined from data)*

Often compare validity of H_0 by comparing to alternate hypotheses ($H_1, H_2, \ldots$)

Notation: $f(x \,|\, H_0)$
$f(x \,|\, H_1)$

Use same notation as for conditional probability

$x = (x_1, x_2, \ldots, x_n)$ **data**

Interpret: *Each hypothesis specifies a joint PDF*

Hypothesis tests (ii)

def statistic ≡ A function of the observed measurements which contains no unknown parameters.

To investigate agreement, construct a test statistic $t(x)$

$x = (x_1, x_2, \ldots, x_n)$ **measured values**

Each of the given hypotheses will imply a PDF for t

Notation: $g(t\,|\,H_0)$
$g(t\,|\,H_1)$

The test statistic can be a scalar $t = t(x)$ or a multidimensional vector

$$t = (t_1(x), t_2(x), \ldots, t_m(x))$$

Question: Why not simply use the original vector of data $x = (x_1, x_2, \ldots, x_n)$?

Answer: *Constructing a statistic of lower dimension $m < n$ reduces the amount of data without losing the ability to discriminate between hypotheses.*

PDF of test statistics

Goal: Formulate a statement about the compatibility between data and various hypotheses in terms of a decision to accept or reject H_0

Define a critical region for t with t_{cut}

$t_{\text{obs}} > t_{\text{cut}}$: reject H_0, accept H_1

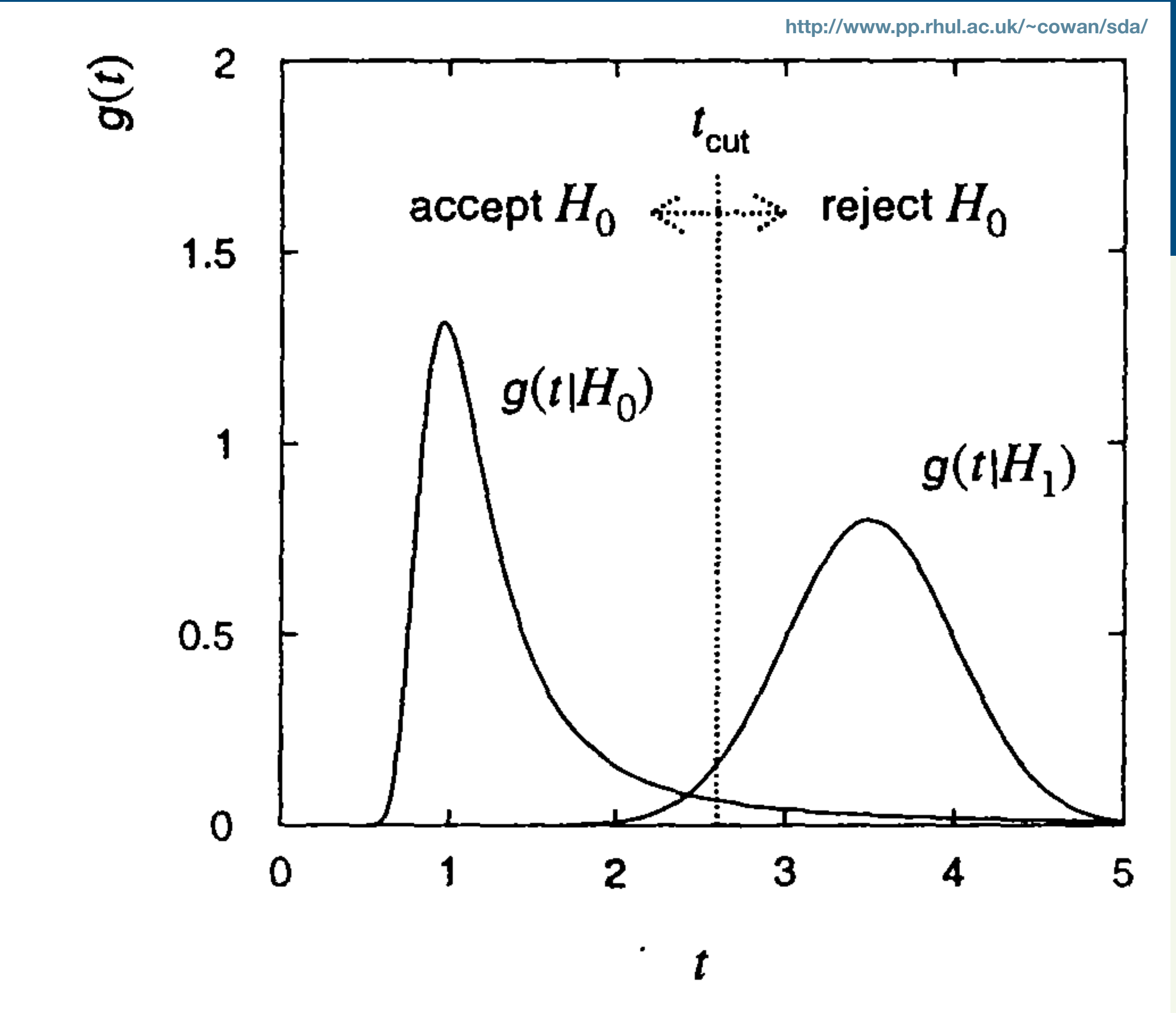
$t_{\text{obs}} < t_{\text{cut}}$: accept H_0

Acceptance region
(complement of critical region)

Choose t_{cut} s.t. the probability to observe $t > t_{\text{cut}}$ is set to some significance level α

$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t\,|\,H_0)dt$$

significance level



There is the probability of α to reject H_0 even if H_0 is true: **Type I error**

Accept H_0 although not true ($t < t_{\text{cut}}$): **Type II error**

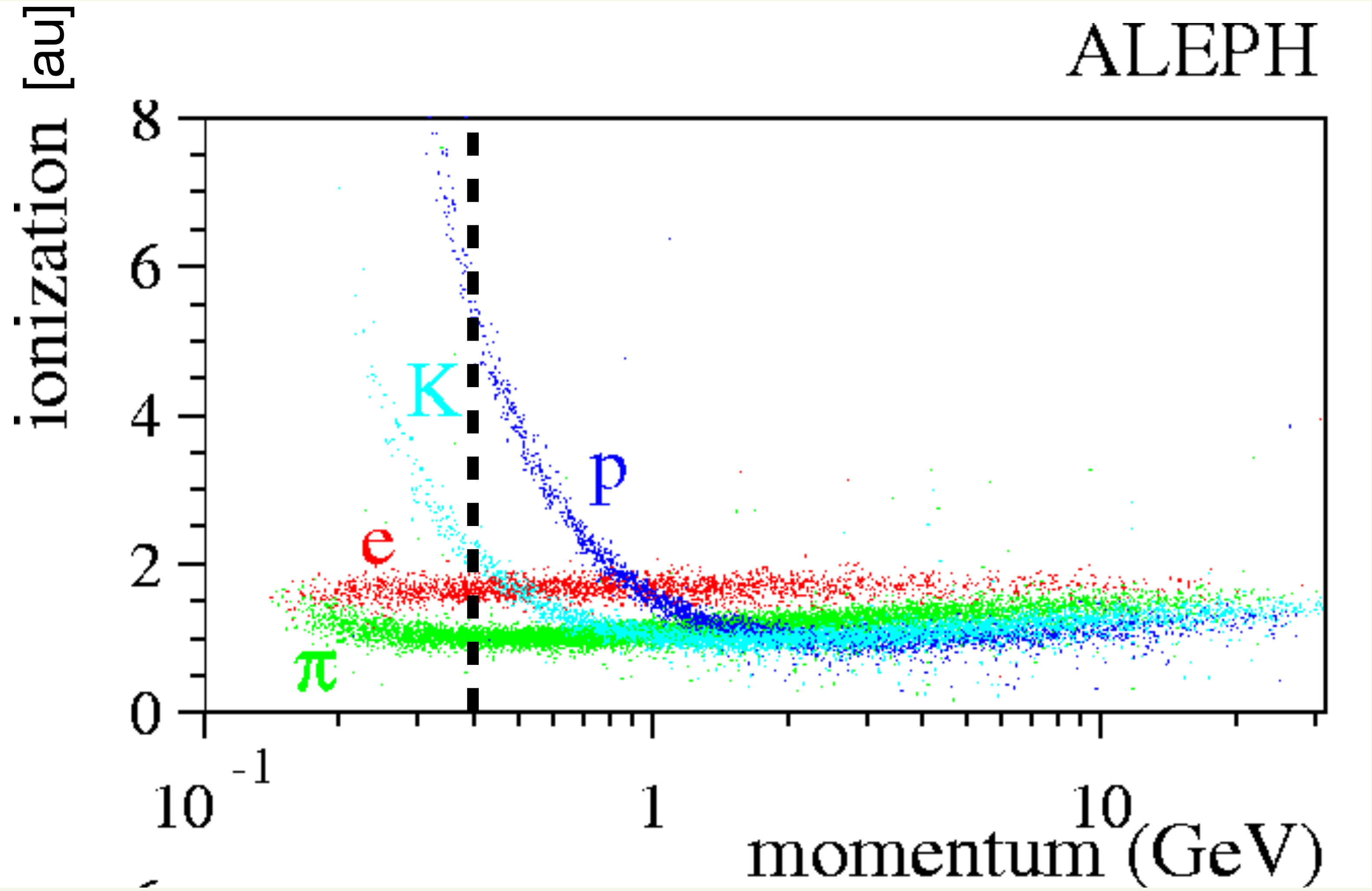
e.g., H_1 is true $\quad \beta = \int_{-\infty}^{t_{\text{cut}}} g(t\,|\,H_1)dt$

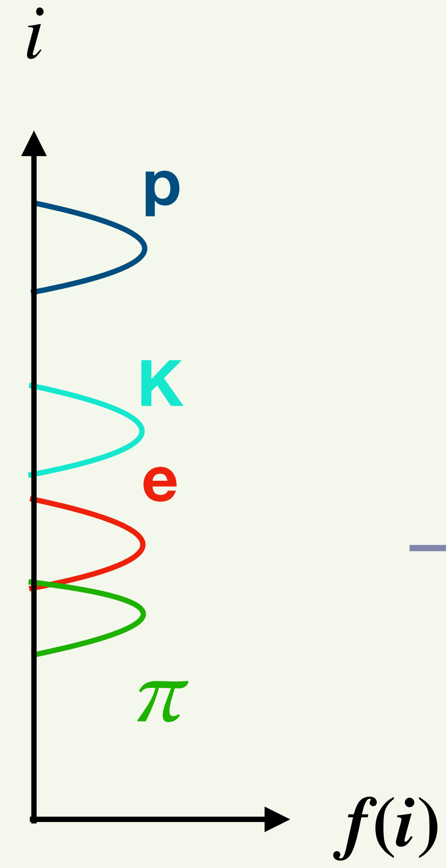
$1 - \beta =$ **Power** of the test to discriminate against H_1

Ex. with particle selection

Example with particle selection

- As an example: our test statistic t represents the measured ionization created by a charged particle of a known momentum traversing a detector
  - The amount of ionization is subject to fluctuations from particle to particle and depends (for a fixed momentum) on the particle's mass





→ *Define some function $t(i)$*

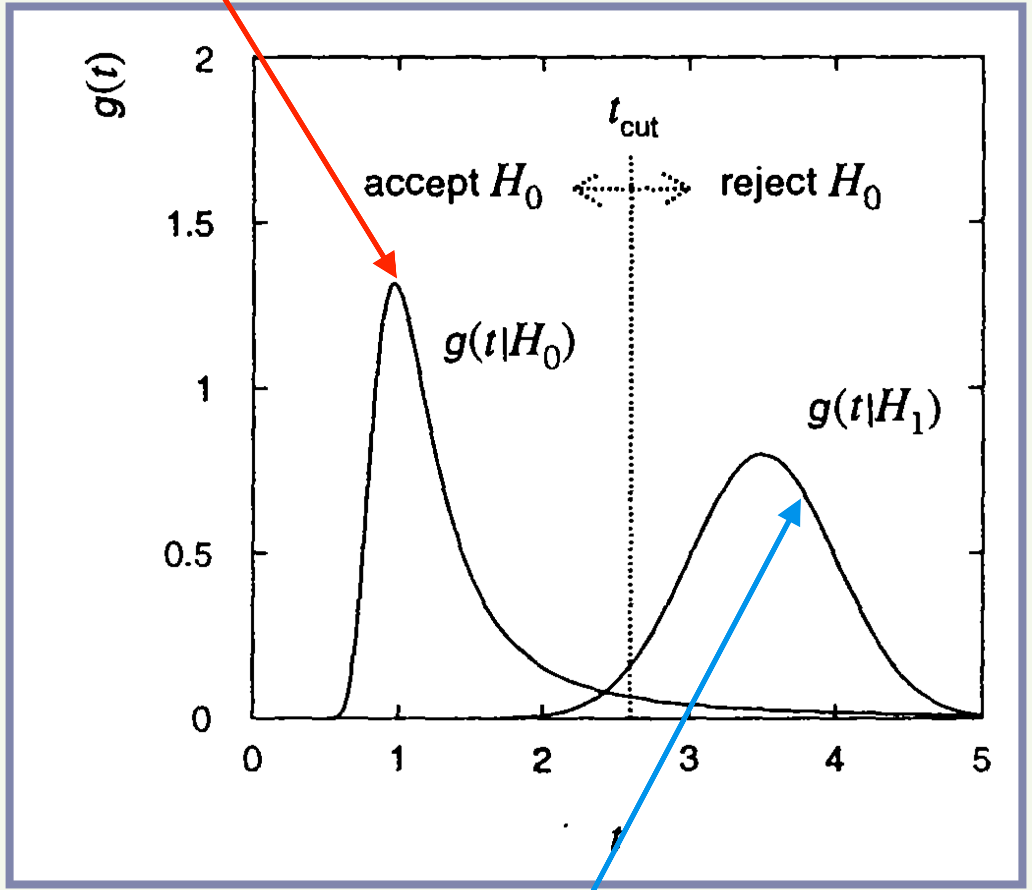
Electron and pion hypothesis

- $H_0 = e$, $H_1 = \pi$
- Selection efficiencies: ϵ_e and ϵ_π

$$\epsilon_e = \int_{-\infty}^{t_\text{cut}} g(t\,|\,e)dt = 1 - \alpha$$

$$\epsilon_\pi = \int_{-\infty}^{t_\text{cut}} g(t\,|\,\pi)dt = \beta$$

PDF of electron hypothesis for test statistic t



PDF of pion hypothesis for test statistic t

- Both can be brought arbitrarily close to zero or unity by appropriate choice of t_cut *(i.e., by making a looser or tighter cut on the ionization)*
    - **However, there is a price:** the higher the signal efficiency, the larger the contamination *(i.e., the purity of the electron sample decreases since some pions are accepted as well)*

Relative fractions

- If the relative fractions of pions and electrons are not known, one can carry out a likelihood fit to the test statistic
  - t is distributed according to

$$f(t; a_e) = a_e g(t \,|\, e) + (1 - a_e) g(t \,|\, \pi)$$

relative fraction of electrons

relative fraction of pions ($a_\pi = 1 - a_e$)

- Knowing a_e allows one to determine the total number of electrons in the sample:

$$N_e = a_e N_{\text{tot}}$$

total number of events

# of electrons

Electron candidates

- Alternatively one may want to select electron candidates by requiring $t < t_{\rm cut}$
  - This leads to $N_{\rm acc}$ accepted out of the $N_{\rm tot}$ particles
  - One then often also wants to determine the total number of electrons before the cut on t. The number of accepted particles is

$$\begin{aligned} N_{\rm acc} &= \epsilon_e N_e + \epsilon_\pi N_\pi \\ &= \epsilon_e N_e + \epsilon_\pi (N_{\rm tot} - N_e) \end{aligned}$$

$$\downarrow$$

$$N_e = \frac{N_{\rm acc} - \epsilon_\pi N_{\rm tot}}{\epsilon_e - \epsilon_\pi}$$

only possible if efficiencies under cut are different

Recall Bayes’ Formalism

Why is Bayes' so important?

Review

It links belief to evidence in probability

$$P(E = \text{Evidence} \,|\, F = \text{Fact})$$

(collected from data)

$$\Big\downarrow \text{Bayes'}$$

$$P(F = \text{Fact} \,|\, E = \text{Evidence})$$

(categorize a new data point)

Given new evidence E, update belief of fact F

Prior belief $\rightarrow$ Posterior belief

$P(F) \rightarrow P(F \,|\, E)$

Bayes' Theorem terminology

Review

- 60% of all email in 2016 is spam. $P(F)$ *prior*
- 20% of spam has the word "Dear." $P(E|F)$ *likelihood*
- 1% of non-spam has the word "Dear." $P(E|F^C)$

You receive an email with the word "Dear in it.

What is the probability that the email is spam? $P(F|E)$ *posterior*

$$\underbrace{P(F|E)}_{\text{posterior}} = \frac{\overbrace{P(E|F)}^{\text{likelihood}}\ \overbrace{P(F)}^{\text{prior}}}{\underbrace{P(E)}_{\text{normalization constant}}}$$

Now let’s use it in our ex.

Bayes again

- The probability that a particle with an observed value of t is an electron or pion, $h(e\,|\,t)$ and $h(\pi\,|\,t)$, can be obtained from the PDFs of $g(t\,|\,e)$and $g(t\,|\,\pi)$ using Bayes' theorem:

prior probability that particle is an *e*

$$h(e\,|\,t) = \frac{\boxed{a_e} g(t\,|\,e)}{a_e g(t\,|\,e) + a_\pi g(t\,|\,\pi)}$$

$$h(\pi\,|\,t) = \frac{a_\pi g(t\,|\,\pi)}{a_e g(t\,|\,e) + a_\pi g(t\,|\,\pi)}$$

Frequentist: fraction of times a particle with a given t will be an electron (pion)

Bayesian: degree of belief that a given particle with a measured value of t is an electron (pion)

Purity

- Often one cares for the purity p_e of a sample of electron candidates selected with $t < t_{\text{cut}}$.
- The purity is given by

$$p_e = \frac{\text{number of electrons with } t < t_{\text{cut}}}{\text{number of all particles with } t < t_{\text{cut}}}$$

$$= \frac{\int_{-\infty}^{t_{\text{cut}}} a_e g(t \,|\, e) dt}{\int_{-\infty}^{t_{\text{cut}}} \left(a_e g(t \,|\, e) + (1 - a_e) g(t \,|\, \pi) \right) dt}$$

$$= \frac{a_e \epsilon_e N_{\text{tot}}}{N_{\text{accepted}}}$$

This is the mean electron probability $h(e \,|\, t)$ averaged over the interval $(-\infty, t_{\text{cut}}]$

Take 5

Neyman-Pearson Lemma

Neyman-Pearson Lemma

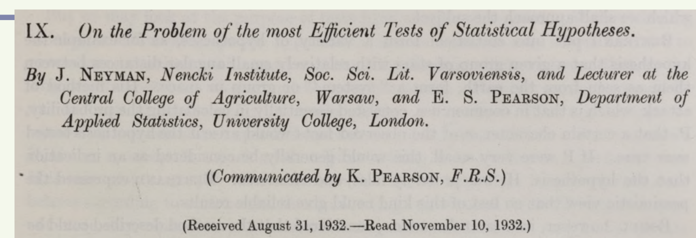
For $t = t(x)$ scalar, choice of t_{cut} is straightforward $\Rightarrow$ *Chosen depending on the efficiency and purity of the selected particles desired for further analysis.*

e.g., can require that they give a max. purity for a given efficiency.

What if $t = (t_1(x), t_2(x), \ldots, t_m(x))$ is a vector?

$\Rightarrow$ *Which $t_{1,2}$ cut offers ideal separation?*

Neyman-Pearson:


IX. *On the Problem of the most Efficient Tests of Statistical Hypotheses.*

By J. Neyman, *Nencki Institute, Soc. Sci. Lit. Varsoviensis, and Lecturer at the Central College of Agriculture, Warsaw, and* E. S. Pearson, *Department of Applied Statistics, University College, London.*

(*Communicated by* K. Pearson, *F.R.S.*)

(Received August 31, 1932.—Read November 10, 1932.)

Acceptance region giving the highest power (and hence highest signal purity) **for a given significance level α is the region in $t-$space s.t.**

$$\frac{g(t\,|\,H_0)}{g(t\,|\,H_1)} > c$$

Constant determined by desired efficiency

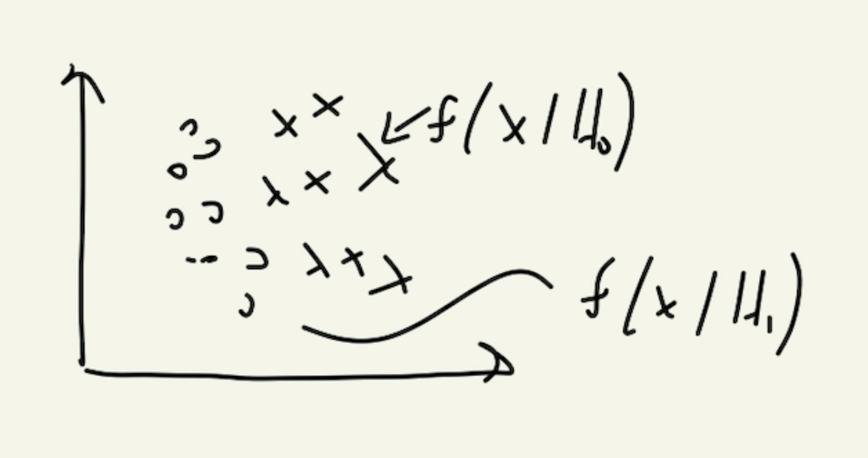
This maps a vector statistic onto a 1D statistic

$$r = \frac{g(t\,|\,H_0)}{g(t\,|\,H_1)}$$

*Called the **likelihood ratio** for simple hypotheses H_0 and H_1 (Corresponding acceptance region given by $r > c$)*

Constructing a test statistic

Given vector of data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, construct a 1D test statistic:



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

The likelihood ratio gives the highest probability to reject H_1 if H_0 is true

To construct t, need to know f $\Rightarrow$ *Very difficult if PDF is multi-dimensional*

In practice, need to use MC to estimate $f(\boldsymbol{x}\,|\,H_i)$ $\Rightarrow$ *Scales terribly:* $\sim M^n$

\# of components

\# of bins

What can we do if we can't determine $f(\boldsymbol{x}\,|\,H_i)$ as nD histograms?

$\Rightarrow$ *Make a simpler assumption for the functional form of $t(\boldsymbol{x})$, and choose the best function having this form*

Today: Consider linear functions of the x_i

Later in the semester: Non-linear functions (e.g., Neural Networks)

Linear test statistic

Linear test statistic (i)

Fisher discriminant function

Simplest form is a linear function: $$t(\boldsymbol{x}) = \sum_{i=1}^{n} a_i x_i = \boldsymbol{a}^T \boldsymbol{x}$$

Goal: determine the a_i to maximize the separation between the PDFs $g(t\,|\,H_0)$ & $g(t\,|\,H_1)$

Characterizes the data

Mean values and covariance matrix of the data $\boldsymbol{x}$, for each hypothesis k

$$(\mu_k)_i = \int x_i f(\boldsymbol{x}\,|\,H_k) dx_1 \ldots dx_n$$

$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(\boldsymbol{x}\,|\,H_k) dx_1 \ldots dx_n$$

Characterizes the hypotheses

Each hypothesis k is characterized by an expectation value and variance

$$\tau_k = \int t g(t\,|\,H_k) dt = \boldsymbol{a}^T \boldsymbol{\mu}_k$$

Subtle connection between μ and τ

$$\Sigma_k^2 = \int (t-\tau_k)^2 g(t\,|\,H_k) dt = \boldsymbol{a}^T \boldsymbol{V}_k \boldsymbol{a}$$

What now?

Maximize separation $|\tau_1 - \tau_2|$

Minimize spread

Linear test statistic (ii)

Fisher discriminant function

Quantified by:

Separation between the 2 classes corresponding to H_0 and H_1

$$J(\boldsymbol{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2} = \frac{\sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j}{\sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij}} = \frac{\boldsymbol{a}^T B \boldsymbol{a}}{\boldsymbol{a}^T W \boldsymbol{a}}$$

Sum of the covariance matrices within the classes

To find maximum separation:

$$\frac{\partial J(\boldsymbol{a})}{\partial a_i} = 0 \quad \Rightarrow \quad \boldsymbol{a} \propto W^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

To determine the coefficients, need the matrix W and the expectation values $\boldsymbol{\mu}_{(0,1)}$

Often estimated from a set of training data (e.g., MC simulation)

Key point: *one does not need to determine the full joint PDFs $f(\boldsymbol{x} \,|\, H_0)$ & $f(\boldsymbol{x} \,|\, H_1)$ as nD histograms; only the means μ_k and variances V_k must be found.*

Fisher discriminant for multi-D Gaussians

- If $f(x \mid H_0)$ and $f(x \mid H_1)$ are both multi-D Gaussians with common covariances $V = V_0 = V_1$, the Fisher discriminant has some interesting properties:

$$f(\boldsymbol{x} \mid H_k) = \frac{1}{(2\pi)^{n/2} \mid V \mid^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right]$$

Recall the definition of the multi-D Gaussian in L03, slide 33

- Linear Fisher becomes : $t(\boldsymbol{x}) = a_0 + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T V^{-1} \boldsymbol{x}$

- The (exact) Likelihood ratio is then given by

$$r = \frac{f(\boldsymbol{x} \mid H_0)}{f(\boldsymbol{x} \mid H_1)} = \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_0)^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)\right]$$

$$= \exp\left[(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T V^{-1} \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_0^T V^{-1} \boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T V^{-1} \boldsymbol{\mu}_1\right]$$

$$\propto e^t$$

$$\boxed{\Rightarrow \quad t \propto \underbrace{\log r + \text{const}}_{\text{Monotonic function of } r}\,.}$$

The Fisher discriminant is as good a test statistic as the full likelihood

Bayes' again

- We can again make statements regarding the probability of H_0 given the data $\boldsymbol{x}$:

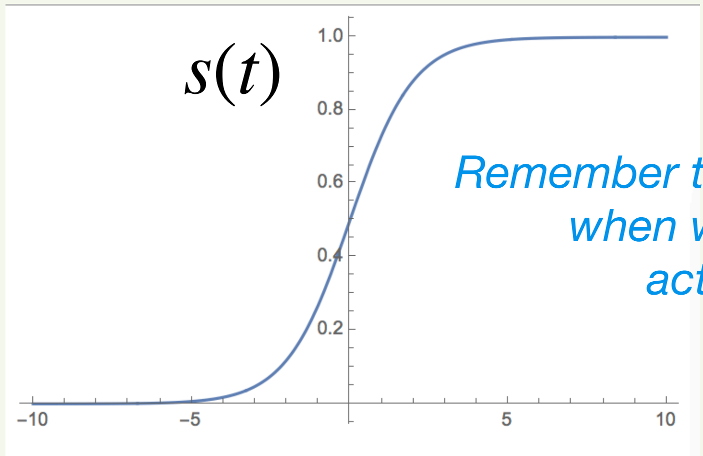
$$P(H_0\,|\,\boldsymbol{x}) = \frac{f(\boldsymbol{x}\,|\,H_0)\pi_0}{f(\boldsymbol{x}\,|\,H_0)\pi_0 + f(\boldsymbol{x}\,|\,H_1)\pi_1} = \frac{f(\boldsymbol{x}\,|\,H_0)\pi_0}{f(\boldsymbol{x}\,|\,H_0)\pi_0\left(1 + \frac{f(\boldsymbol{x}\,|\,H_1)}{f(\boldsymbol{x}\,|\,H_0)}\frac{\pi_1}{\pi_0}\right)} = \frac{1}{1 + \frac{\pi_1}{r\;\pi_0}}$$

prior probabilities

- Now substitute $r \propto e^t$ from the last slide:

$$P(H_0\,|\,\boldsymbol{x}) = \frac{1}{1+e^{-t}} \equiv s(t)$$

This function $s(t)$ is called a **sigmoid** function



t

Remember this function later when we talk about NN activation functions

The prior probabilities have been absorbed into the offset

$$a_0 = \log\frac{\pi_0}{\pi_1} - \frac{1}{2}\boldsymbol{\mu}_0^T V^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T V^{-1}\boldsymbol{\mu}_1$$

Remaining terms from final expression for r in the previous slide, also absorbed into a_0

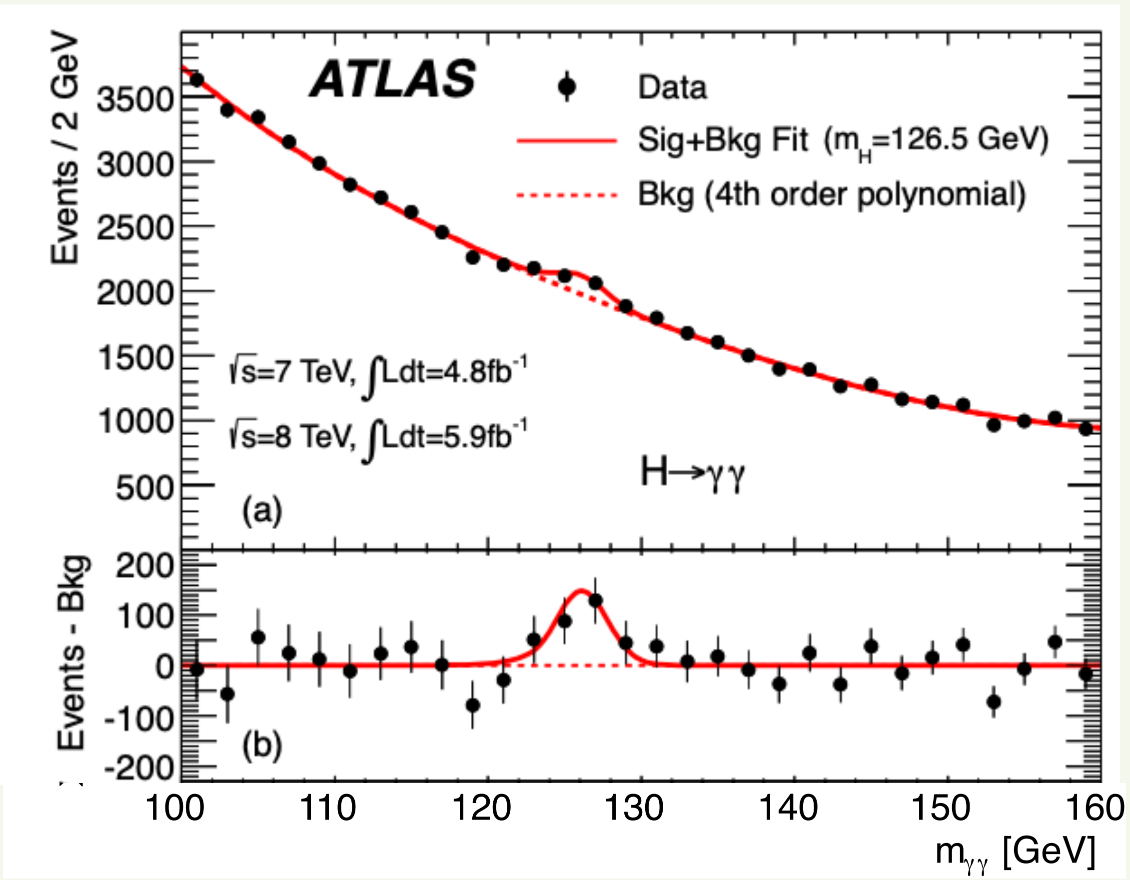
Next lecture: *What if f are not Gaussian or don't share a common covariance?*

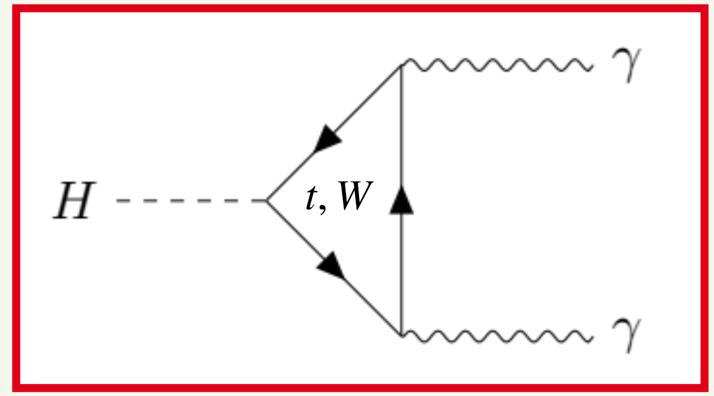
Recall the Higgs discovery

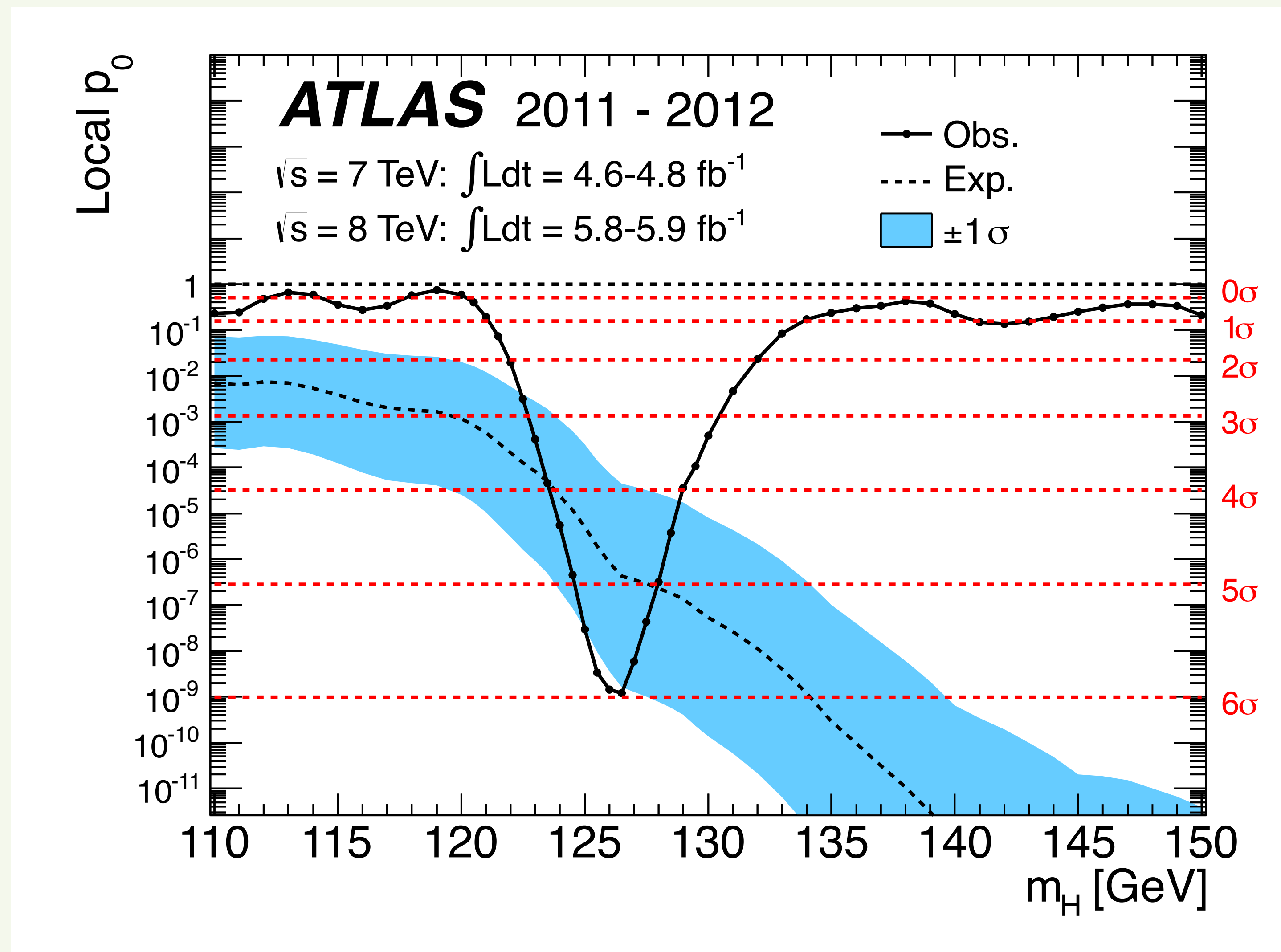
Adapted from
https://indico.cern.ch/event/508168/contributions/2028747/attachments/1307803/1962991/Statistical-Reasoning-HASCO16.pdf

Recall this example from the intro. lecture

- Discovery of the **Higgs boson**





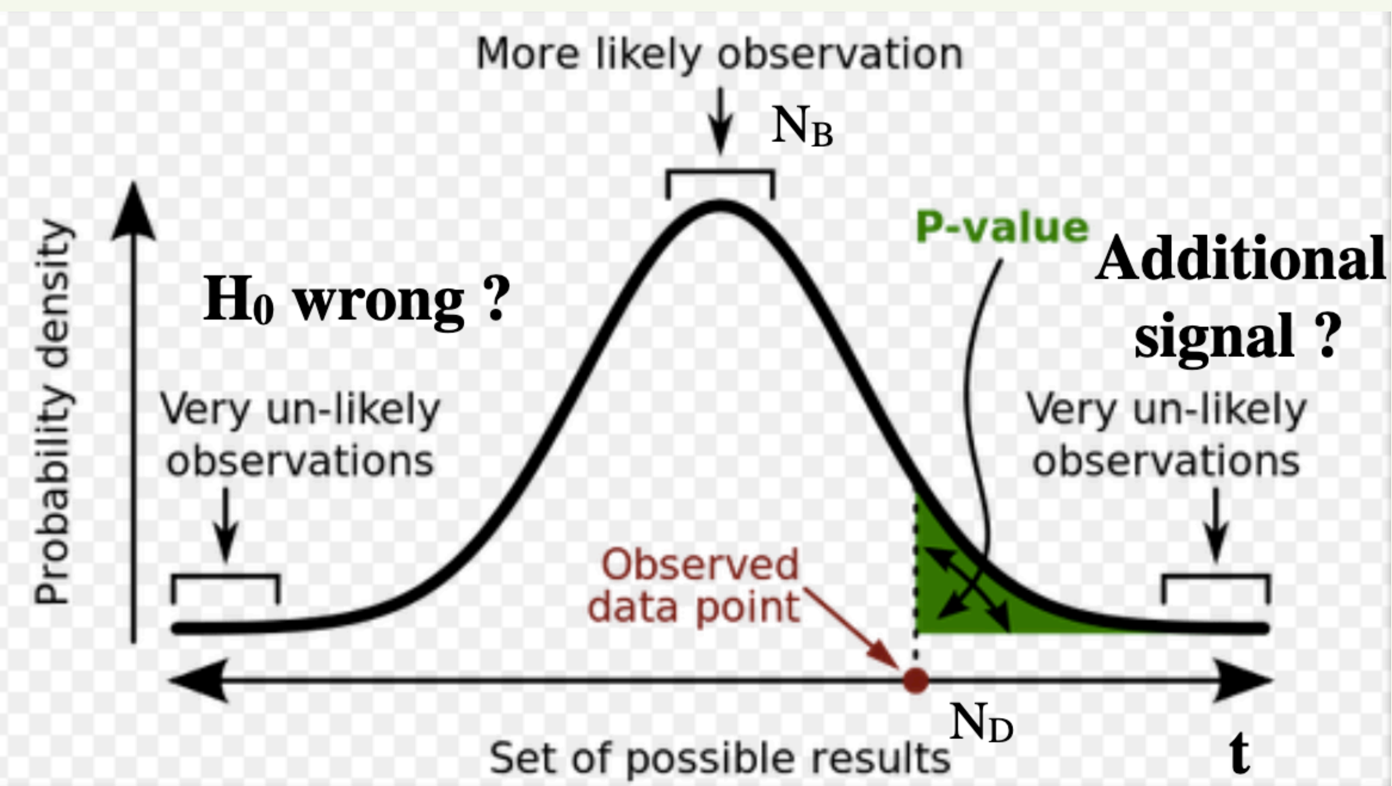


- When did this peak become a discovery?
  I.e. when did we consider it as incompatible with the background hypothesis (SM without Higgs)?
- Estimate N_{Bkgd} and N_{Data} under the peak, then calculate the significance (goodness of fit)

Significance statement with a P-value

- Quantify the compatibility of data with hypothesis, e.g., the Standard Model (SM) of particle physics
  - Define a test statistic t (e.g., # of events)
  - Calculate the P-value on the PDF, the likelihood $f(t\,|\,H_0)$ for t given a hypothesis H_0 (e.g. background/SM without Higgs)

$$P = \int_{t_{\mathrm{obs}}}^{\infty} f(t\,|\,H_0)\;dt$$



Conventional thresholds:

$P \lesssim 0.03,\quad 2\sigma,\qquad \Rightarrow$ *Happens often*

$P \lesssim 0.002,\quad 3\sigma,\qquad \Rightarrow$ *Evidence*

$P \lesssim 10^{-7},\quad 5\sigma,\qquad \Rightarrow$ ***Discovery!***

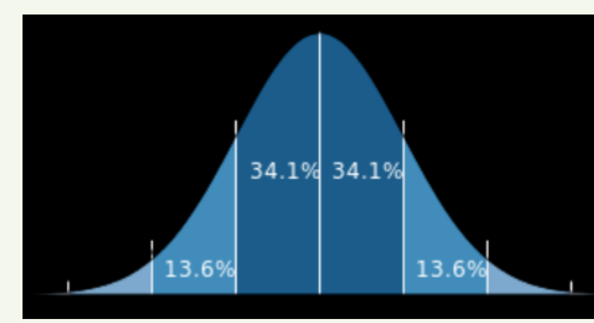
Significance statement with σ

- The P-value can be transformed into the number of sigma:

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

  - Φ = the cumulative (integral) of the Normal distribution
  - Φ^{-1} = the inverse (quantile)
    - With root: sigma = ROOT:Math::normal_quantile_c(p-value,1)

	α	δ	α	δ
	0.3173	1σ	0.2	1.28σ
	4.55×10^{-2}	2σ	0.1	1.64σ
Evidence	2.7×10^{-3}	3σ	0.05	1.96σ
	6.3×10^{-5}	4σ	0.01	2.58σ
Discovery	5.7×10^{-7}	5σ	0.001	3.29σ
	2.0×10^{-9}	6σ	10^{-4}	3.89σ



Area α of the tails outside $\pm\delta$ from the mean of a Normal distribution

The P-values of the Higgs Discovery

According to the NP lemma the likelihood ratio of two alternative models/hypotheses H_1 and H_0 is the best test statistic

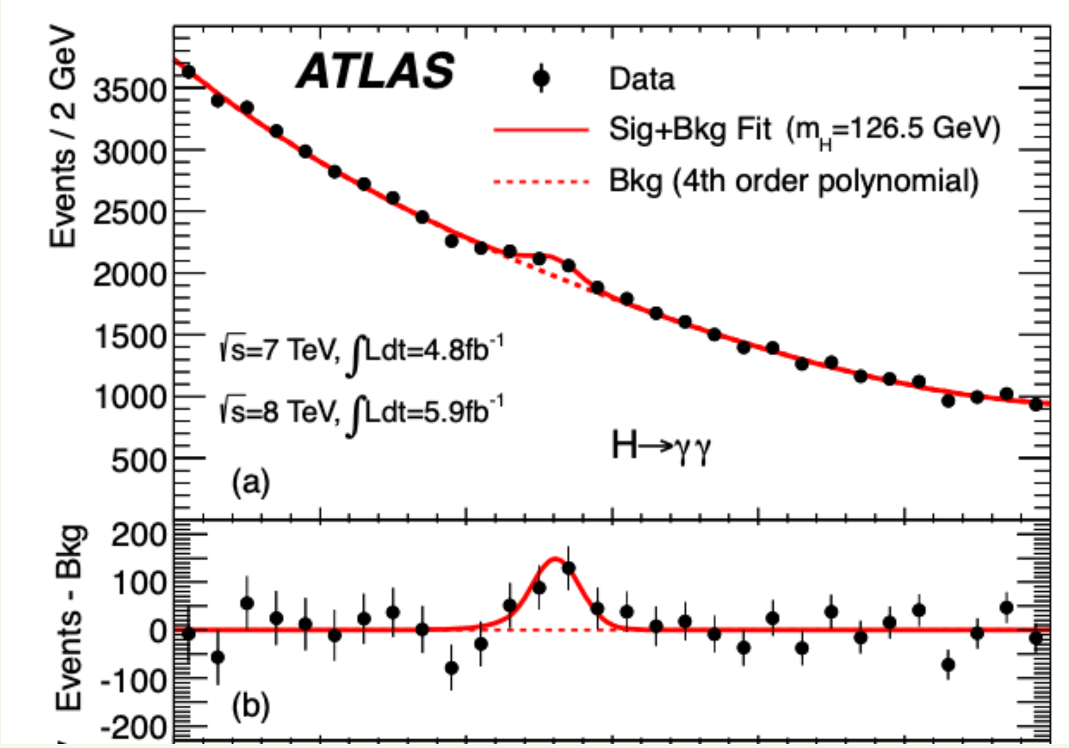
likelihood ratio

signal strength

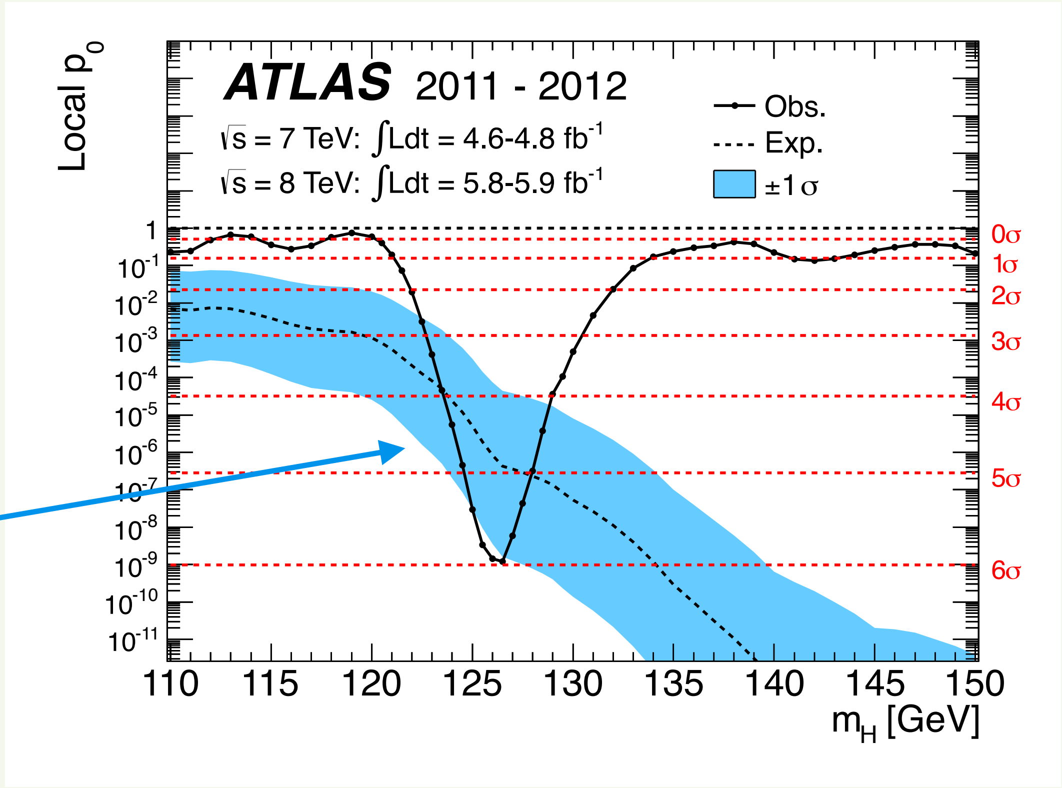
$$t(\boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid H_1)}{f(\boldsymbol{x} \mid H_0)} \quad \Rightarrow \quad r = \frac{\mathscr{L}(\mu_1)}{\mathscr{L}(\mu_0)}$$

$$\mathscr{L}(\mu) = \prod_i P(N_{\text{Data}}\ ;\ \mu N_{\text{Sig}} + N_{\text{Bkg}})$$

The Poisson distribution models the statistical fluctuations of the data



Test for deviations from the background only model (e.g., SM w/out Higgs). Put $\mu_1 = 0$ (bkg only), fit μ_0 to data, and integrate from N_{Data} to ∞ to obtain the P-value.
A 5σ deviation from $\mu = 0$ was achieved around 125 GeV.



The same calculation with $\mu = 1$(SM *with* Higgs) fits the data well in that region, within 1σ

I.e., First find a deviation ($\mu = 0$),
then check alternative models ($\mu = 1$)

For next time

- Required reading
  - Cowan textbook: Chapter 4 (through 4.4.1)
  - Reading material / L06 / Statistical-Reasoning-HASCO16
  - Reading material / L05 / L03_Statistics_Fitting_II

- Extra reading for fun: /Reading material / L06 /
  - NeymanPearson (original paper)

Next time

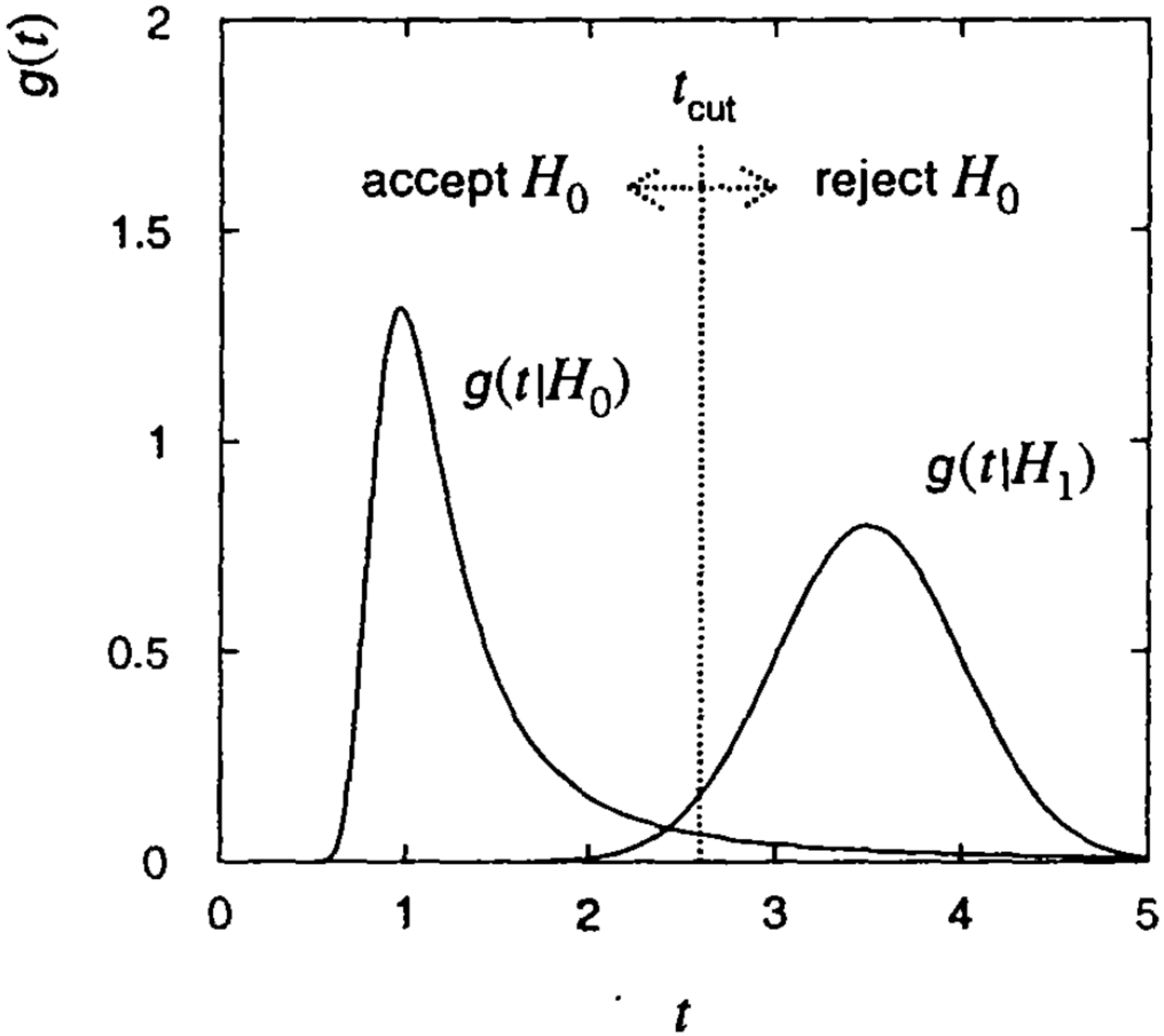
- Classical Confidence intervals
  - Exact method
  - Examples:
    - Gaussian distributed estimator
    - Poisson distributed estimator
    - Correlation coefficient, transformation of parameters
  - Likelihood and LS Confidence intervals

- Limits near a physical boundary
  - Shifted and Bayesian approaches
  - Example: Upper limit on the mean of a Poisson variable with background

Quiz Time: 6th Round

Type I vs. type II errors

1. You have two hypotheses H_0 and H_1 and a test statistics t distributed according to $g(t|H_0)$ and $g(t|H_1)$, as shown in the figure below. You now choose a certain value t_{cut} to accept H_0 / reject H_0. Using the figure, explain the meaning of Type I and Type II errors.



Fischer / LS hypothesis tests

2. Properties of the Fisher discriminant / LS / Bonus

   a) Write down the definition of a Fisher discriminant of n data points $x = (x_1, x_2, \ldots, x_n)$.

   b) When is it beneficial to construct a Fisher discriminant instead using the full likelihood ratio?

   c) You carried out least square fits (LS) using two hypotheses and obtained χ_0^2 and χ_1^2. What is the equivalent of the likelihood ratio for binned data?

   d)* Explain step-by-step how you would obtain $g(t|H_0)$ and $g(t|X_1)$ using MC techniques

Bibliography

- Part of the material presented in this lecture is taken from the following sources. See the active links (when available) for a complete reference
  - Recall the Higgs Discovery section adapted from https://indico.cern.ch/event/508168/contributions/2028747/attachments/1307803/1962991/Statistical-Reasoning-HASCO16.pdf
  - **Statistical Data Analysis** textbook by G. Cowan (U. London): all figures & equations with white background