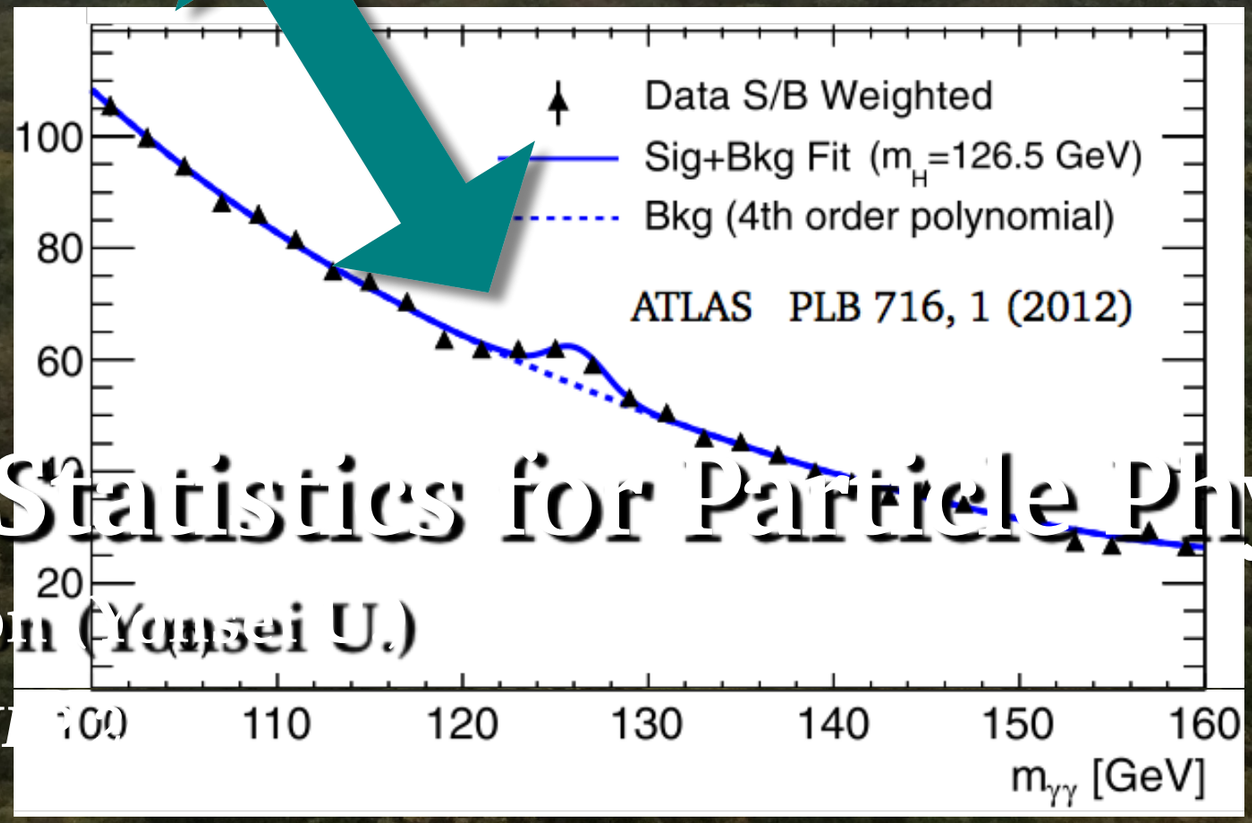


Practical Statistics for Particle Physicists (II)

Youngjoon Kwon (Yonsei U.)

Oct. 13-25, AEPSHI

Outline

- **Basic elements**
  - some vocabulary
  - Probability axioms
  - some probability distributions

Part I

- **Two approaches: Frequentist vs. Bayesian**

- **Hypothesis testing**
- **Parameter estimation**
- **Other subjects — "nuisance", "spurious", "look elsewhere"**

Parts II & III

a quick review of Lecture I

- **Probability vs. Statistics**
- **PDF (and CDF)**
  - expectation values, covariance matrix, correlation coefficients
- **Frequentist vs. Bayesian**
- **Some theorems**
  - LLN, CLT, Neyman-Pearson lemma, Wilks theorem, etc.

Two interpretations of Probability

- While the classic or frequentist approach can lead to a well-defined probability for a given situation, it is not always usable.

  ➔ In such circumstances one is left with only one option: ***Bayesian***.

- When data are scarce ➔ these two approaches can give somewhat different predictions,

  but given sufficiently large data sample, they give pretty much the same conclusion. In that case the choice between the two may be regarded arbitrary.

- Perhaps, we may choose one for the main result, and try the other for a cross-check.

Hypothesis Testing

Probability $P(H|\vec{x})$

- In the frequentist approach, we do not, in general, assign probability of a hypothesis itself.

  Rather, we compute the probability to accept/reject a hypothesis assuming that it (or some alternative) is true.

- In Bayesian, on the other hand, probability of any given hypothesis (*degree of belief*) could be obtained by using the Bayes' theorem:

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

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

Hypothesis Testing

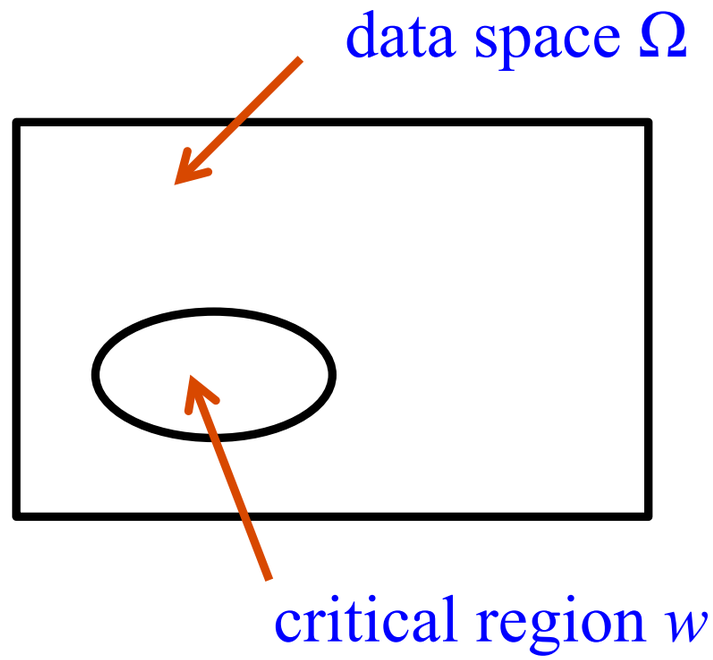
- A hypothesis H specifies the probability for the data
  (*shown symbolically as* $\vec{x}$ *here*),
  often expressed as a function $f(\vec{x}|H)$
- The measured data $\vec{x}$ could be anything:
  * observation of a single particle, a single event, or an entire experiment
  * uni-/multi-variate, continuous or discrete
- the two kinds:
  * simple (or "point") hypothesis – $f(\vec{x}|H)$ is completely specified
  * composite hypothesis – H contains unspecified parameter(s)
- The probability for $\vec{x}$ given H is also called the **likelihood** of the hypothesis, written as $L(\vec{x}|H)$

Critical Region - *what it is*

- Consider e.g. a simple hypothesis H_0 and an alternative H_1
- A (frequentist) test of H_0:

  Specify a critical region w of the data space Ω such that, assuming H_0 is correct, there is no more than some (small) probability α to observe data in w

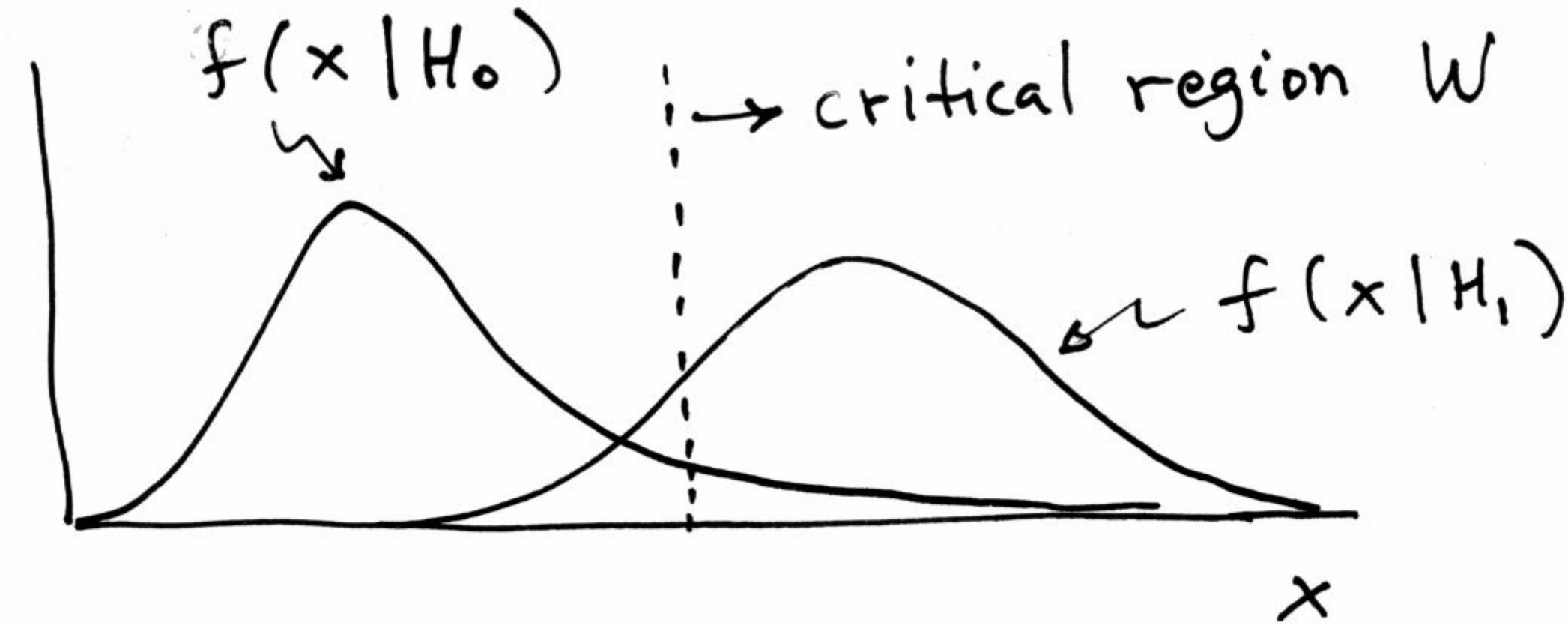
$$P(\vec{x} \in w | H_0) \leq \alpha$$



- α: "size" or "significance level" of the test
- If $\vec{x}$ is observed within w, we reject H_0 with a confidence level $1 - \alpha$

Critical Region - *how to choose*

- In general, $\exists$ an ∞ number of possible critical regions that give the same significance level α
- Usually, we place the critical region where there is a low probability α for $\vec{x} \in w$ if H_0 is true, but high if the alternative (H_1) is true



Test statistic

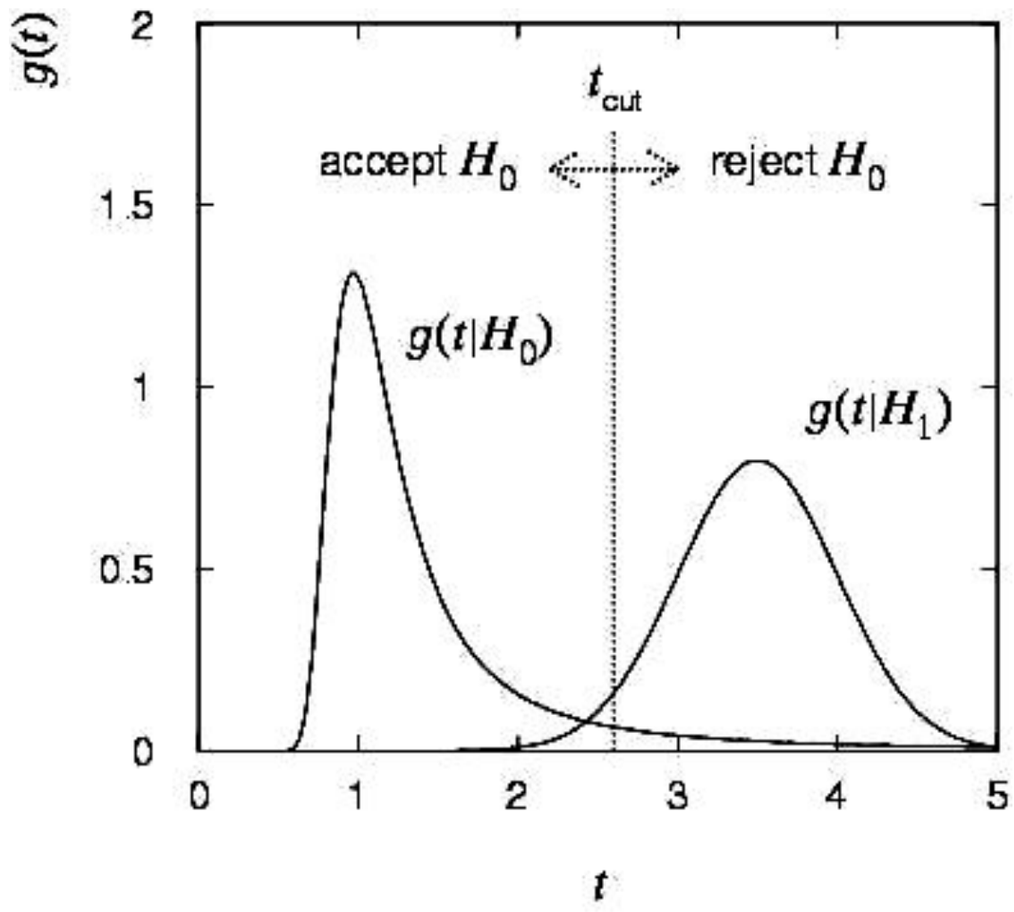
- The boundary surface of the critical region for an n-dim. data space can be defined by an equation of the form:

$$t(x_1, \cdots, x_n) = t_c$$

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

- For the test statistic t, we can work out the PDFs $g(t|H_0)$, $g(t|H_1)$, etc.
- Decision boundary is now given by a signle 'cut' on t, thus defining the critical region

  $\Rightarrow$ for an n-dim. data space, the problem is reduced to a 1-dim. problem



Type-I, Type-II errors

- Rejecting H_0 when it is true is called the **Type-I error**

  (Q) Given the significance α of the test, what is the maximum probability of Type-I error?

- We might also accept H_0 when it is indeed false, and an alternative H_1 is true. This is called the **Type-II error**

  The probability β of Type-II error:

$$P(\vec{x} \in \Omega - w | H_1) = \beta$$

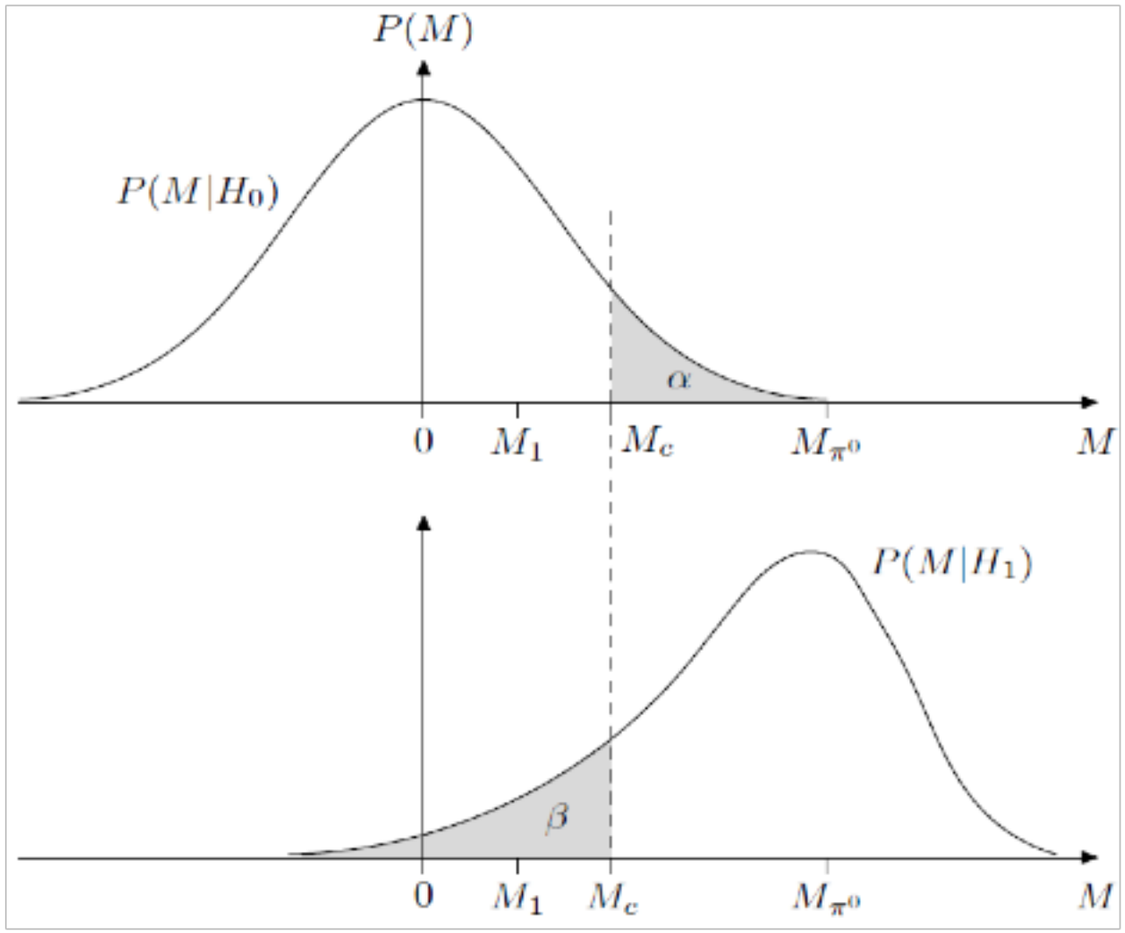
$1 - \beta$ is called the **power** of the test with respect to H_1

Type-I, Type-II errors

	H_0 chosen	H_1 chosen
H_0 true	Correct decision, Prob = 1-α	Type I error, Prob = α
H_1 true	Type II error, Prob = β	Correct decision, Prob = 1-β

Optimal decision: minimize β for given α

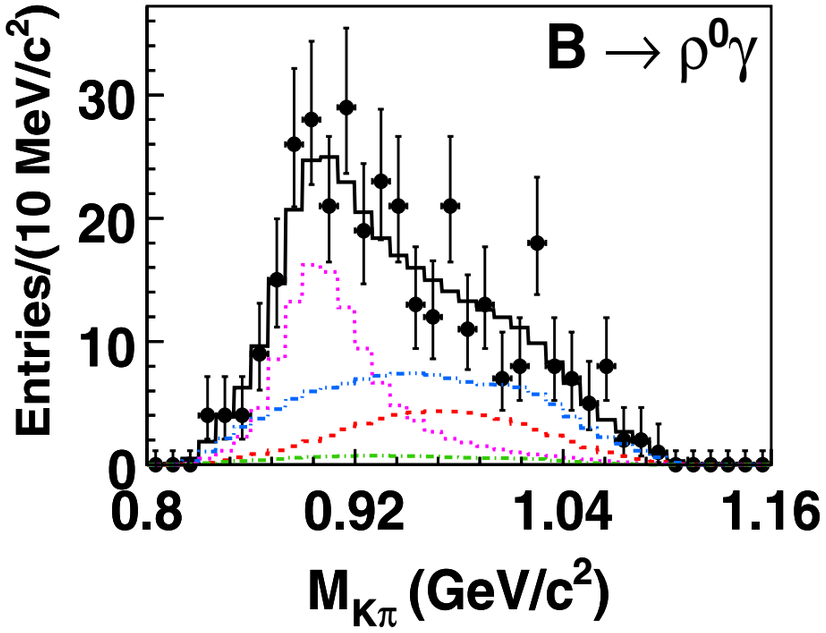
- The size of the test is $Pr_0(Y \in R_\alpha)=\alpha$.
- The power of the test is $Pr_1(Y \in R_\alpha)=1-\beta$.



from an FAPPS09 Lecture by S. T'Jampens

exercise on Type-I, II errors

Since $B \to K^*\gamma$ has much higher branching fraction than $B \to \rho\gamma$, the former can be a serious background to the latter. It is crucial to understand the "efficiency" and "fake rate" of K/π identification system of your experiment in this study. The figure below shows the $M_{K\pi}$ invarianbt mass distribution, where one of the pion mass (in $\rho^0 \to \pi^+\pi^-$ decay) is replaced by the Kaon mass, for the $B^0 \to \rho^0\gamma$ signal candidates (Belle, PRL 2008).



Express the following observables in Type-I & Type-II errors. ***What are H_0 & H_1, for each case?***

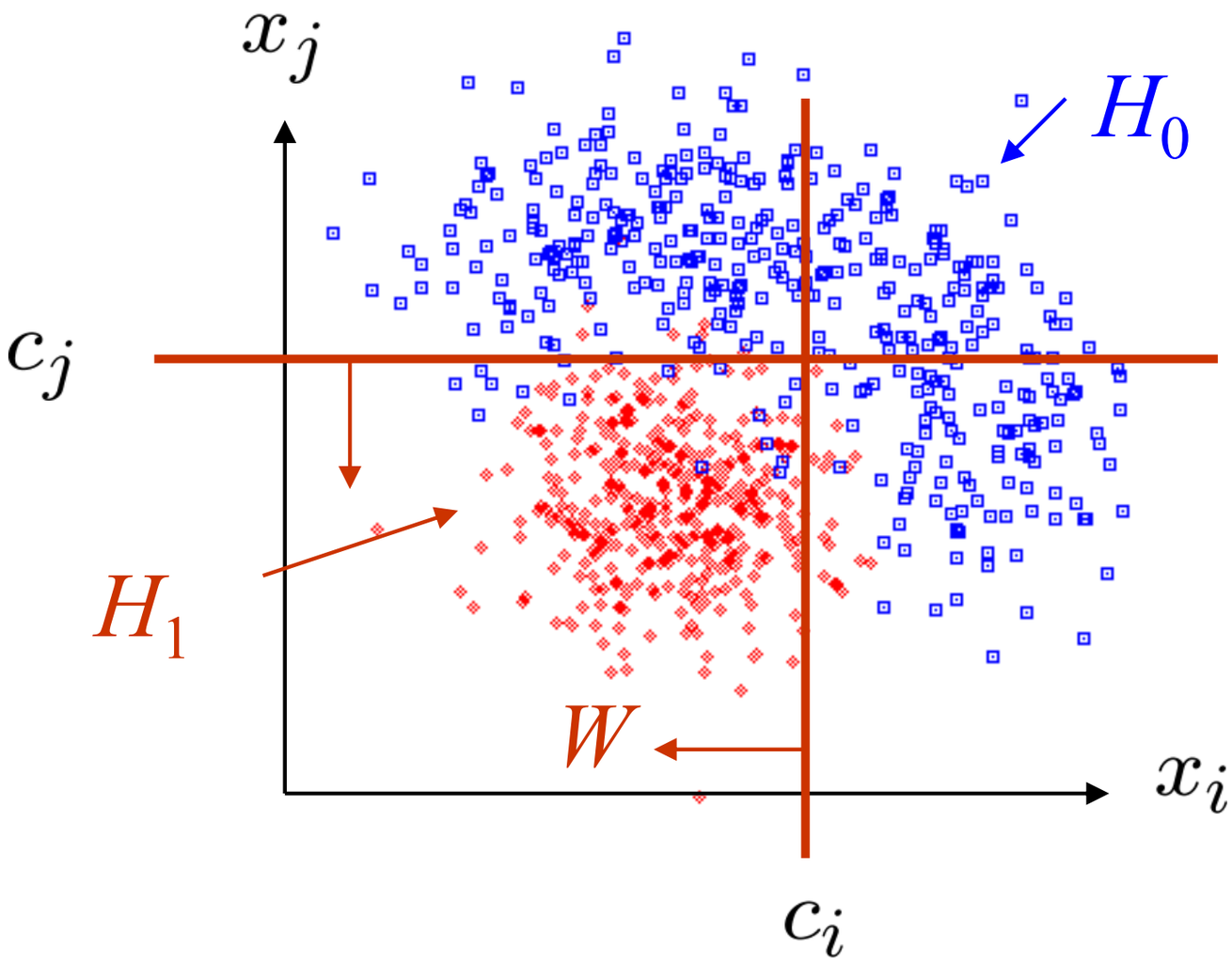
- $f_{\pi^+ \to K^+}$ = probability of misidentifying a π^+ as a K^+
- $f_{K^+ \to \pi^+}$ = probability of misidentifying a K^+ as a π^+
- ϵ_{K^+} = prob. of identifying a K^+ correctly as a K^+
- ϵ_{π^+} = prob. of identifying a π^+ correctly as a π^+

Defining a multivariate critical region

with "square cuts"

$$x_i < c_i$$

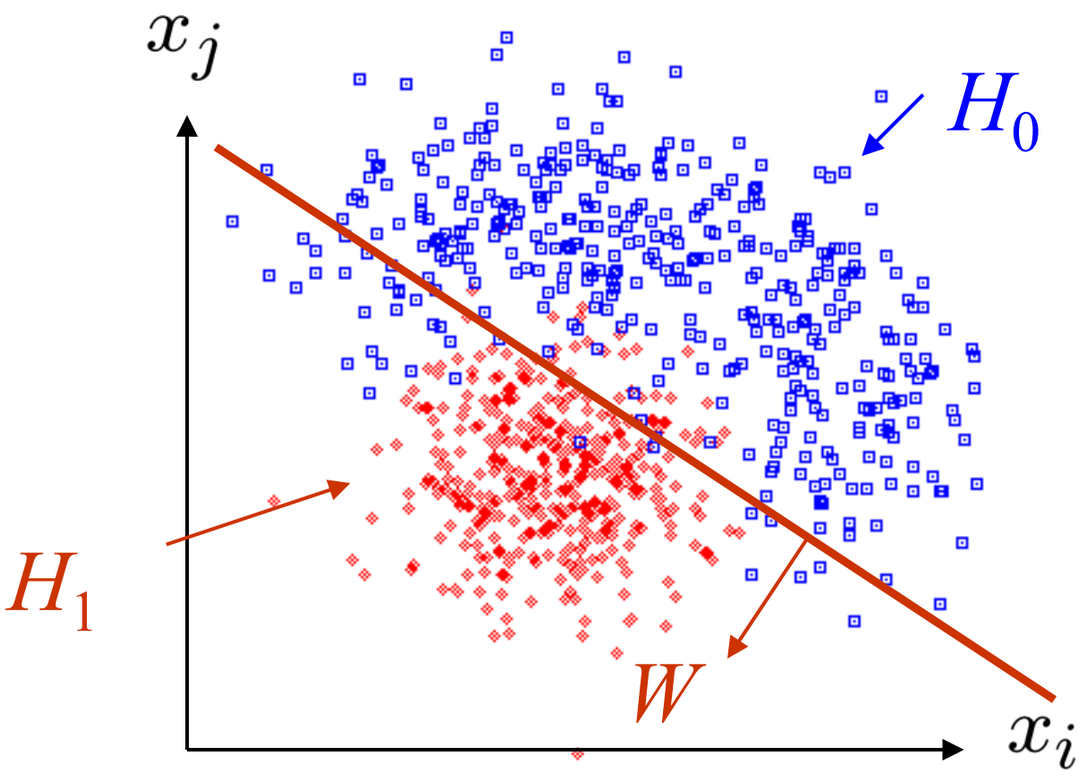
$$x_j < c_j$$



Defining a multivariate critical region

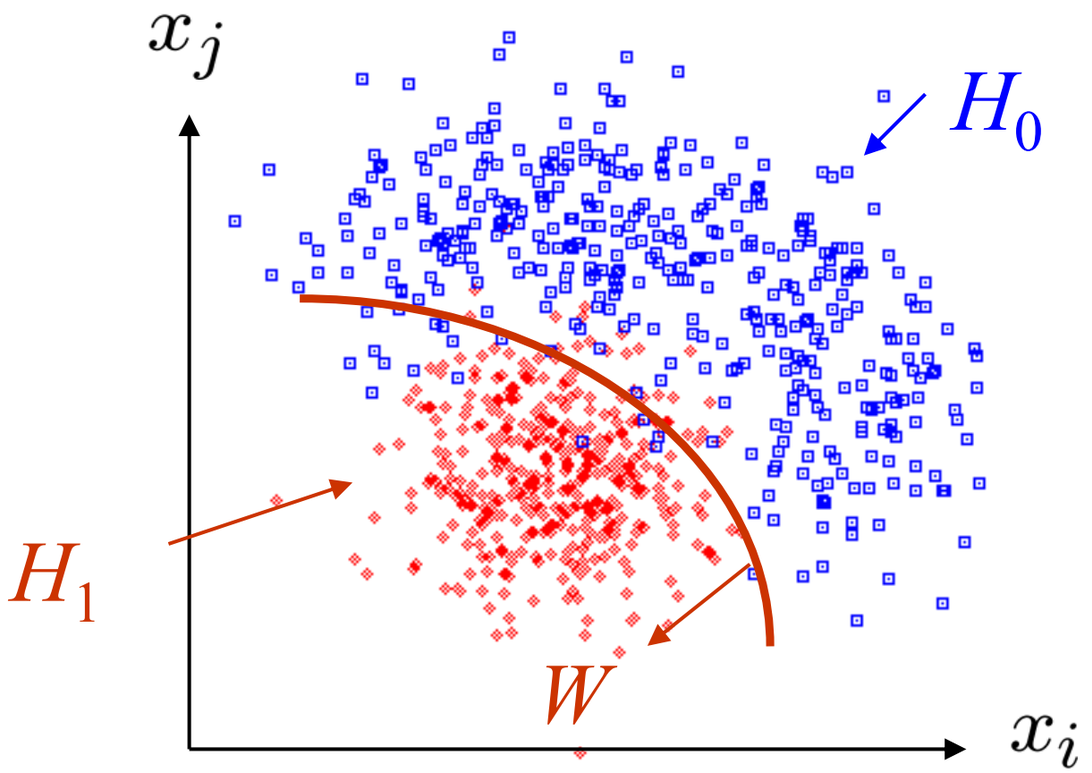
some more sophisticated ways

linear



(ex) Fisher discriminants, etc.

or nonlinear



(ex) artificial neural net, etc.

algorithms for a multivariate critical region

- **Many (*old or new*) methods for finding decision criteria**
  - **Fisher discriminants**
  - **Artificial neural networks**
  - **Boosted decision trees**
  - **Kernel density methods**
  - **...**

Software for multivariate analysis

TMVA, Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

From **tmva.sourceforge.net**, also distributed with ROOT

Variety of classifiers

Good manual

StatPatternRecognition, I. Narsky, physics/0507143

Further info from **www.hep.caltech.edu/~narsky/spr.html**

Also wide variety of methods, many complementary to **TMVA**

Currently appears project no longer to be supported

How to choose an *optimal* test statistic

- Use **Neyman-Pearson lemma**

  For a test of size α of the simple hypothesis H_0,
  to obtain the highest power w.r.t. the simple alternative H_1,
  choose the critical region w such that the likelihoot ratio satisfies

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

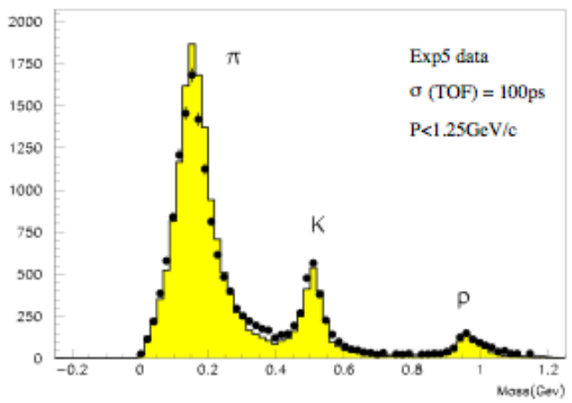
  everywhere in w and is $< k$ elsewhere,
  where k is a constant chosen for each pre-determined size α.

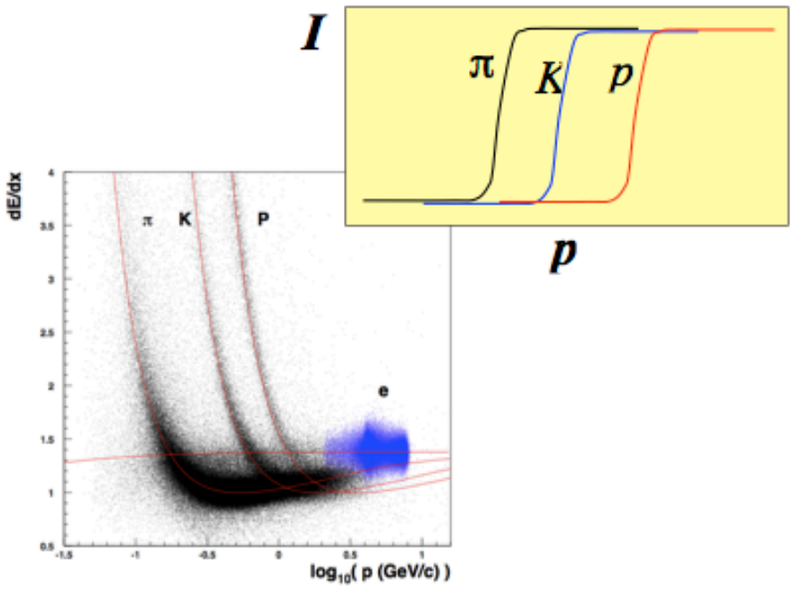
- Equivalently, the optimal scalar test statistic is

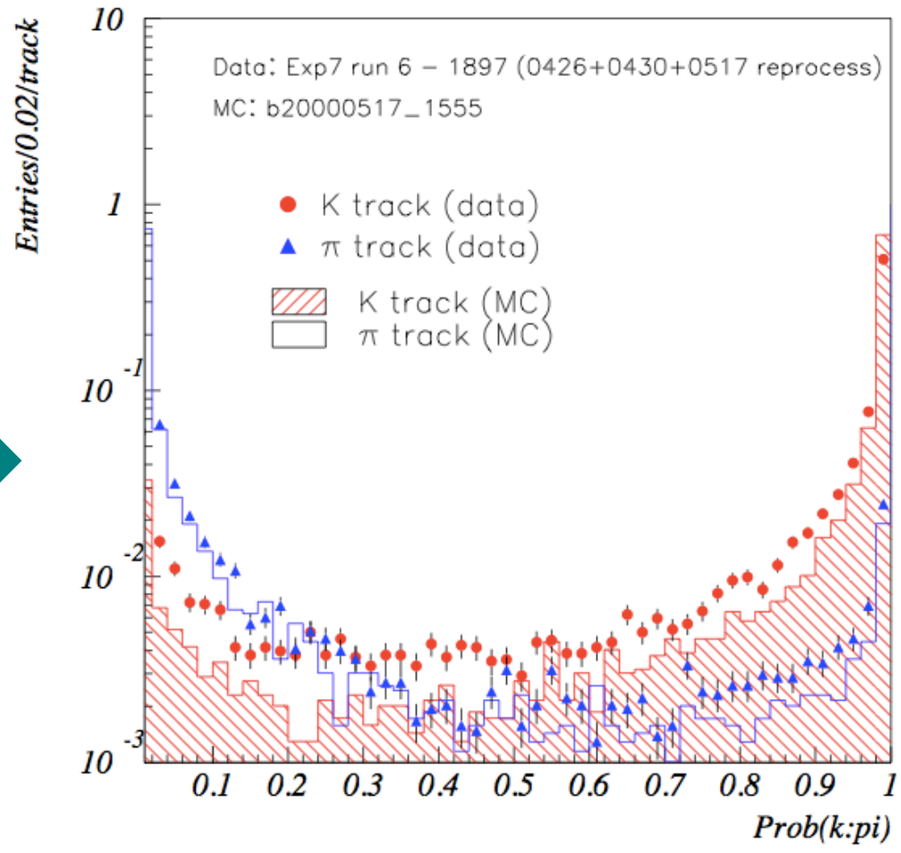
$$t(\vec{x}) = P(\vec{x}|H_1)/P(\vec{x}|H_0)$$

  (Note) Any monotonic function of this leads to the *same test*.

an application of Neyman-Pearson Lemma







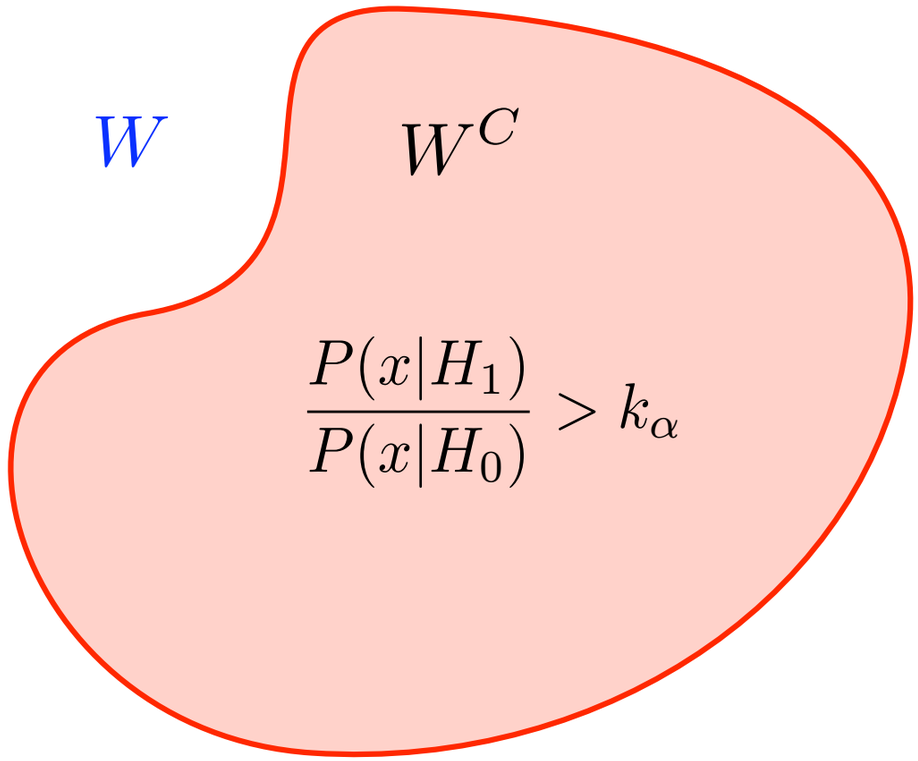
$$P_i \equiv P_i^{dE/dx} \times P_i^{\mathrm{TOF}} \times P_i^{\mathrm{Ch}} \quad \text{e.g. } (i = \pi \text{ or } K)$$

For optimal statistic, construct the likelihood ratio

$R_{K/\pi} = P_K/P_\pi$ (or any ftn. that is monotonic to it)

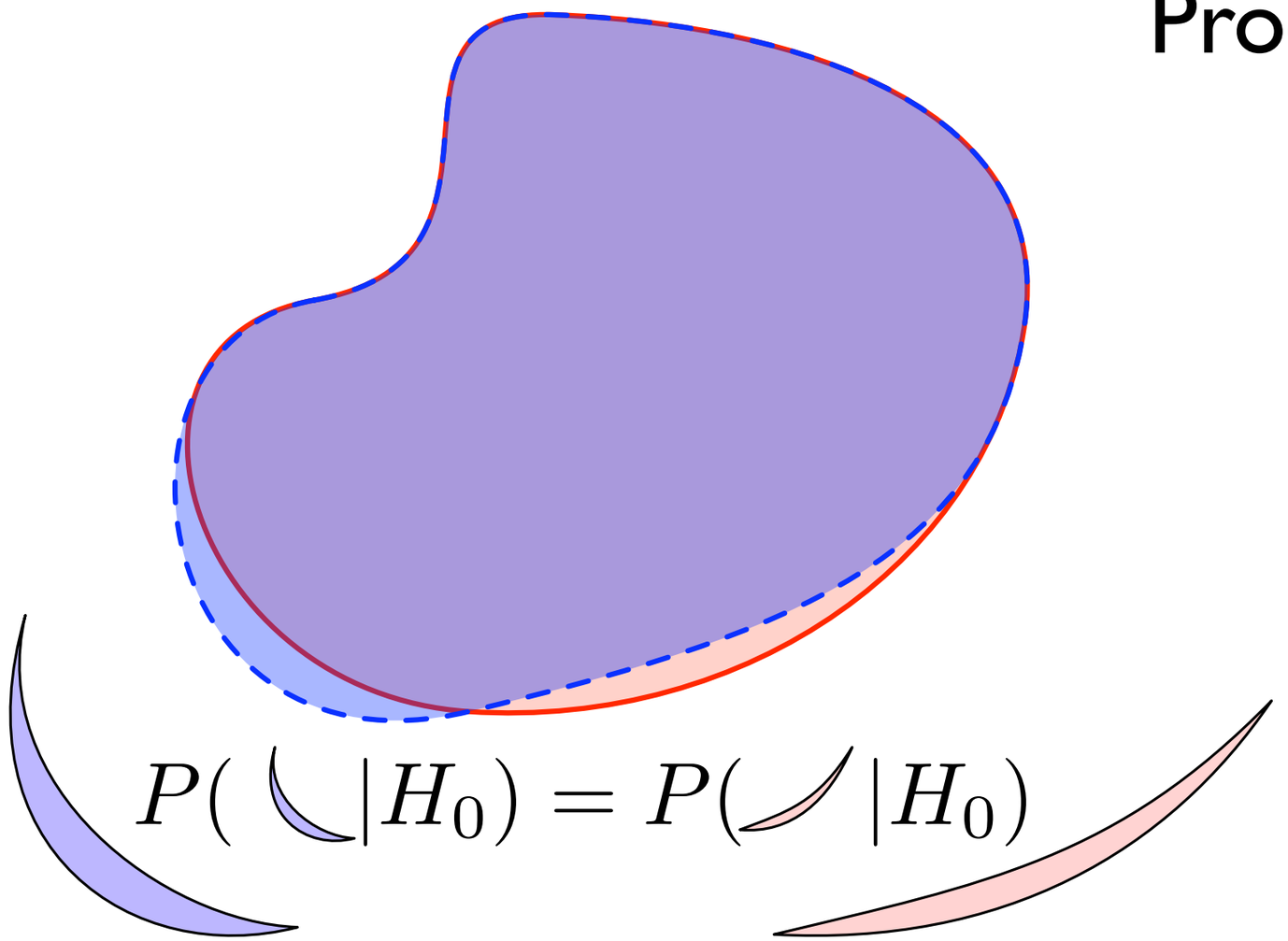
Belle actually used $R_{K/\pi} = P_K/(P_K + P_\pi)$ so that $0 \leq R_{K/\pi} \leq 1$

Proof (*graphical*)



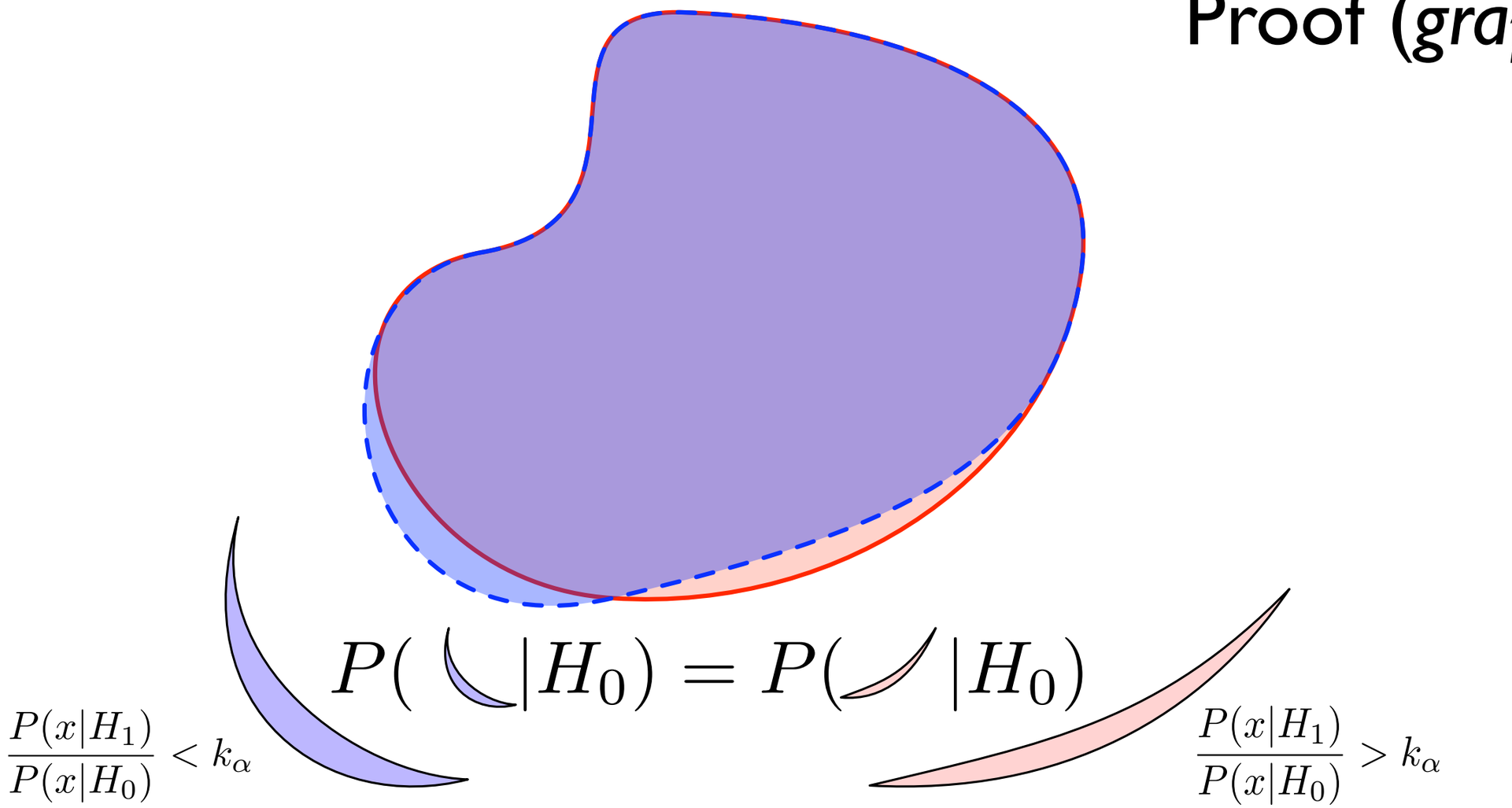
Consider the contour of the likelihood ratio that has size a given size (eg. probability under H_0 is 1-α)

Proof (*graphical*)



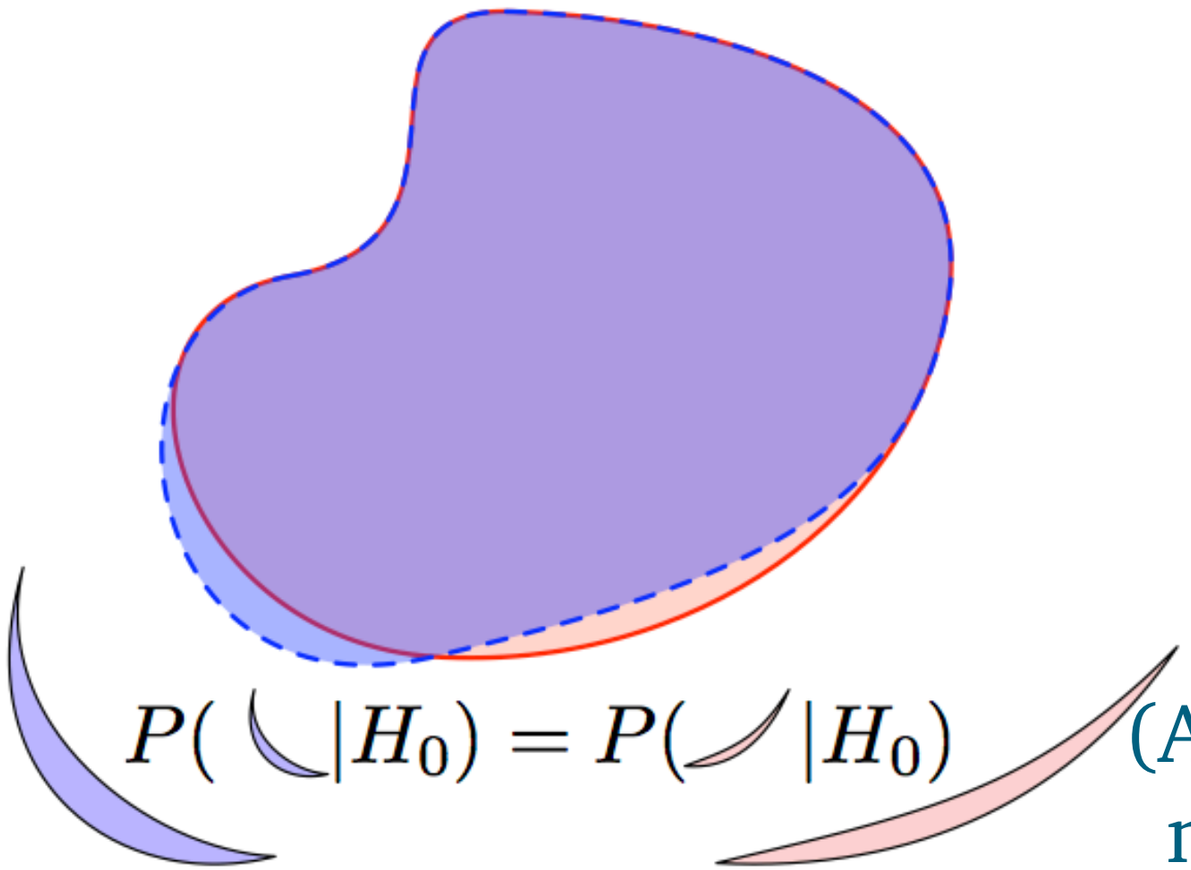
Now consider a variation on the contour that has the same size (eg. same probability under H_0)

Proof (*graphical*)



$$P(\ |H_1) < P(\ |H_1)$$

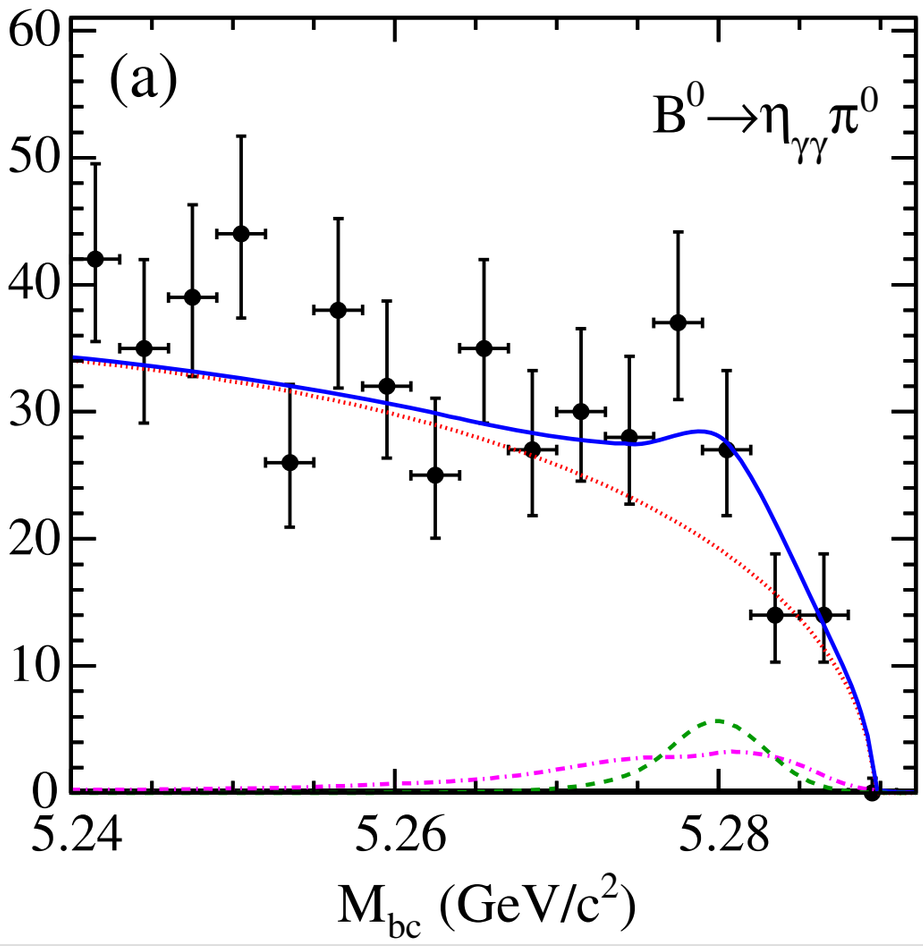
(Quiz) With Neyman-Pearson lemma, we may have THE way to optimize the critical region ("cut"). Then why should we bother with multivariate analyses such as artificial neural network, etc.?

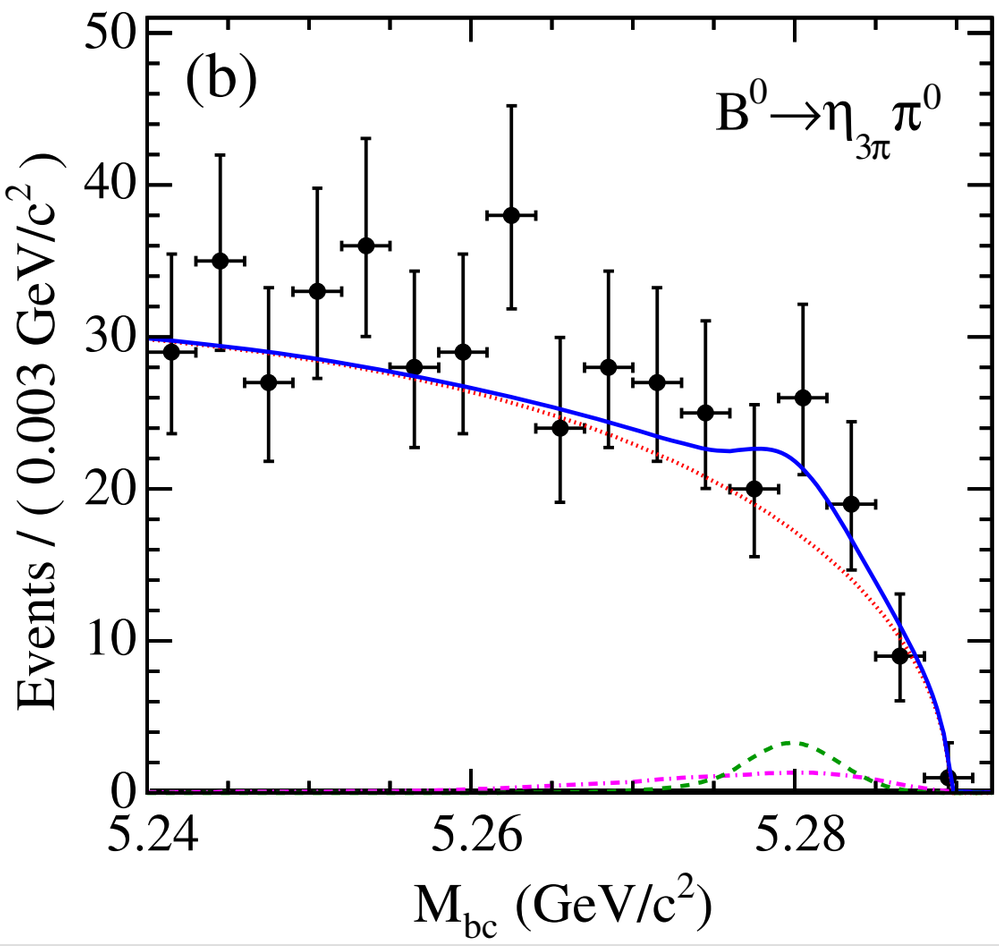


(Ans.) The modeling of P(**x**|H) may not be perfect, if the correlations are not taken properly into account. This will become more serious for higher dimensions of **x**.

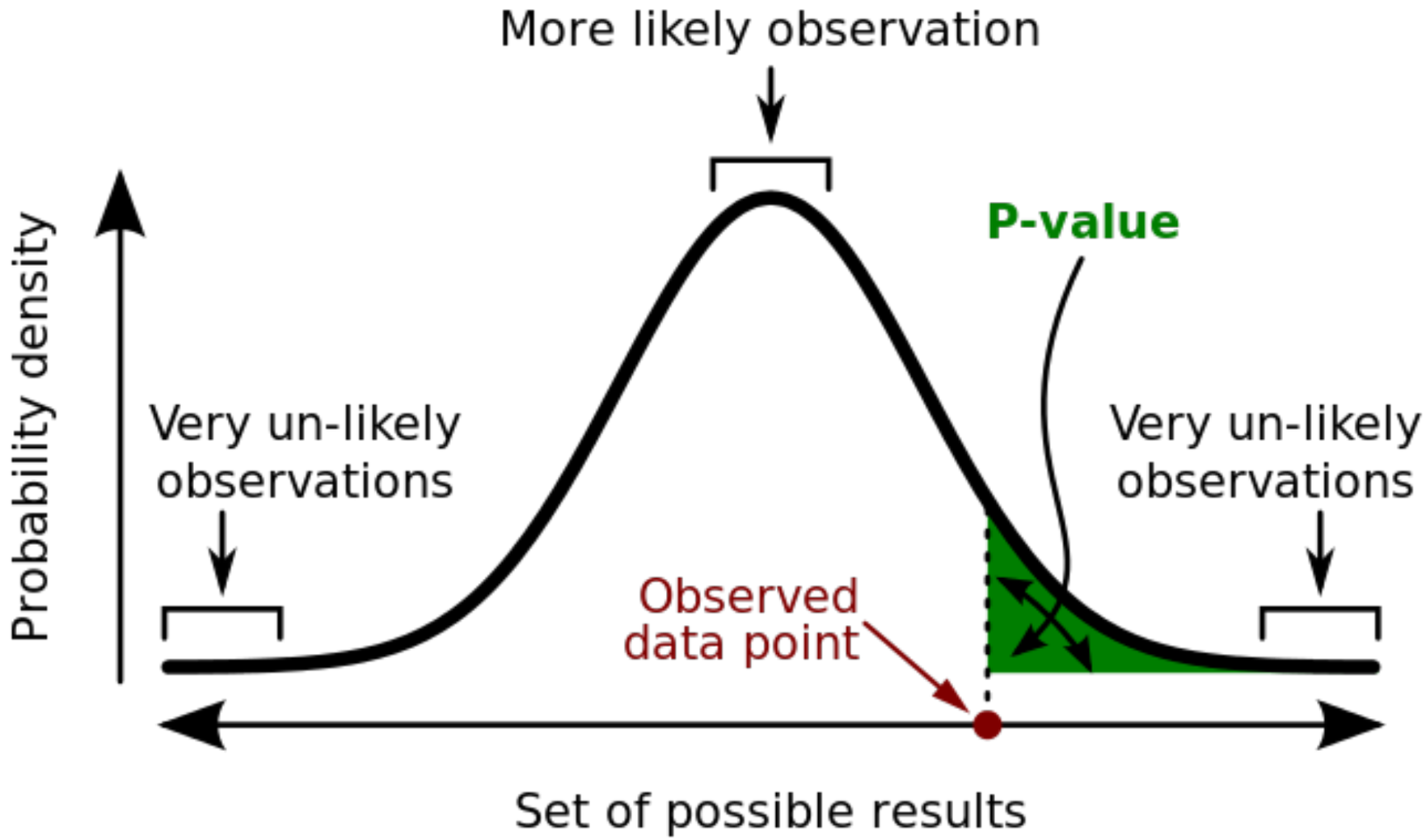
Significance of signal

	Mode	$Y_{\rm sig}$	ϵ (%)	$\mathcal{B}_\eta$ (%)	Significance	$\mathcal{B}(10^{-7})$
(a)	$B^0 \to \eta_{\gamma\gamma}\pi^0$	$30.6^{+12.2}_{-10.8}$	18.4	39.41	3.1	$5.6^{+2.2}_{-2.0}$
(b)	$B^0 \to \eta_{3\pi}\pi^0$	$0.5^{+6.6}_{-5.4}$	14.2	22.92	0.1	$0.2^{+2.8}_{-2.3}$
	Combined				3.0	$4.1^{+1.7}_{-1.5}$





the *p*-value



By User:Repapetilto @ Wikipedia & User:Chen-Pan Liao @ Wikipedia - File:P value.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=36661887

In short, *p*-value is the 'size' of a test against a given hypothesis.

the p-value

- With p-value, we express the level of agreement b/w data and H

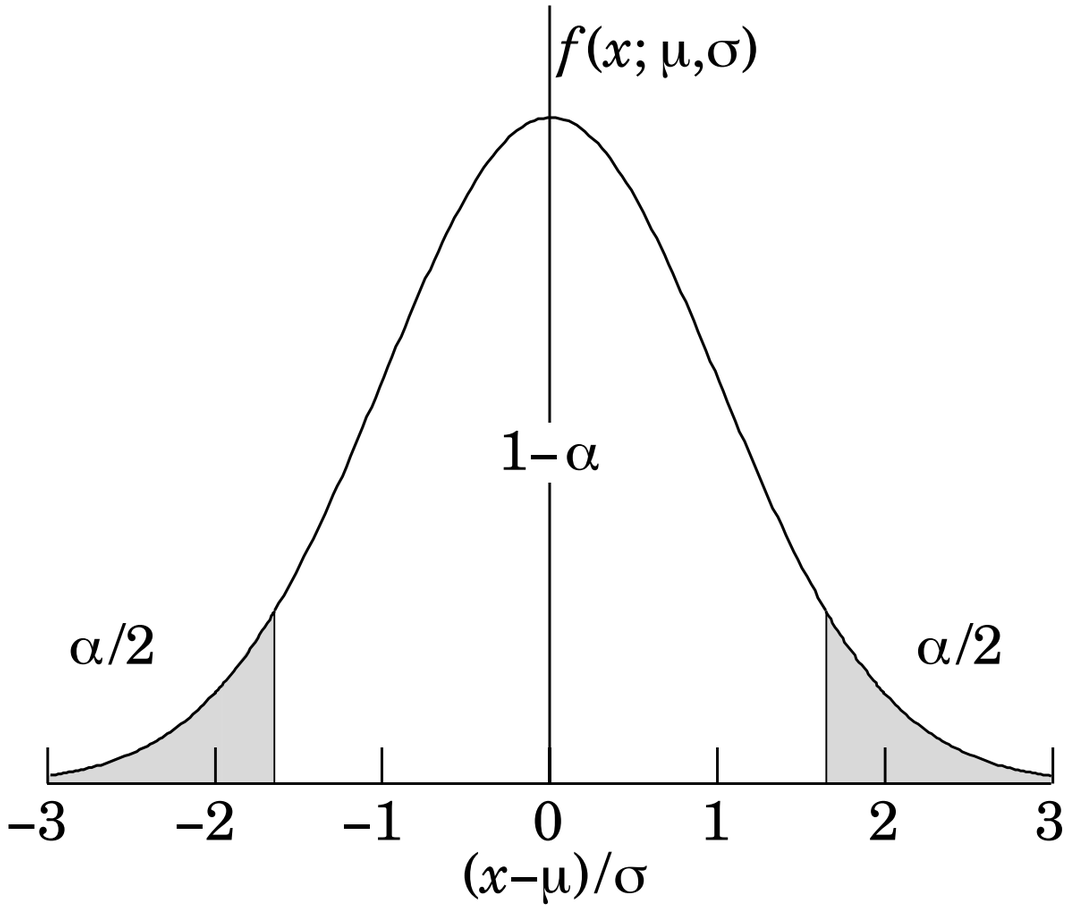
  $p =$ probabilty, under assumption of H, to observe data with equal or lesser compatibility with H, in comparison to the data we obtained

  $\neq$ *the probability that H is true* 

P(observation | hypothesis) ≠ P(hypothesis | observation)

Remember?

Gaussian (Normal) distribution



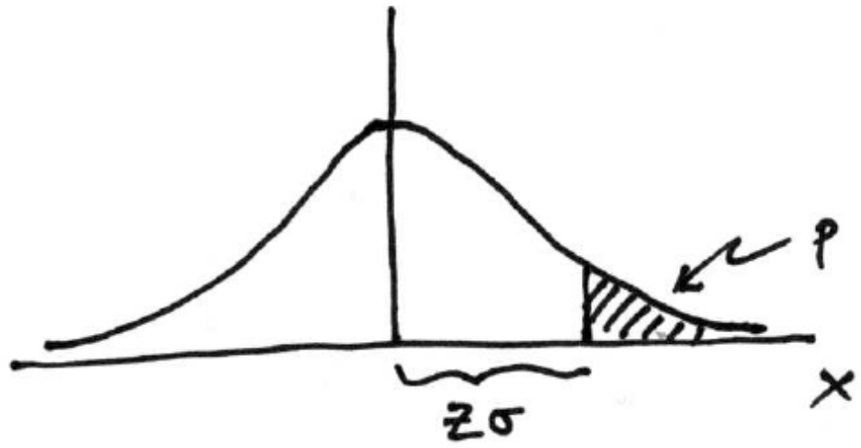
TMath::Prob(δ^2,1)

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

Table 36.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

Significance and the *p*-value

- **Often we quote the significance Z, for a given *p*-value**
  - ***Z*** **= the number of standard dev. that a Gaussian random variable would fluctuate in one direction to give the same *p*-value**



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

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

(Ex) *Z* = 5 (a "5-sigma effect") ⇔ *p* = 2.9 x 10^{-7}

p-value example: a fair coin?

We toss a coin $N = 20$ times and get $n = 17$ heads.
Test whether this coin is 'fair' or not.

Hypothesis H_0: the coin is fair ($\mu = 50\%$ chance for head)

Example: significance of a signal

We observe n events: $n = n_b + n_s$

n_b events from known background

n_s signal events (to be inferred from data)

Assume both n_s, n_b are Poisson.

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

Suppose $b = 0.5$ (assume precise), and we observe $n_{\text{obs}} = 5$.

Can we **claim evidence for a signal excess?**

Give ***p*-value** for the null hypothesis $s = 0$.

Quiz

1983 Korean Baseball Champion HaiTai Tigers' Lineup

			BA	OBA	SLG	HR	RBI	SB
1	김일권	CF	.275	.345	.364	6	26	48
2	서정환	SS	.257	.320	.339	3	34	13
3	김성한	1B	.327	.401	.448	7	40	13
4	김봉연	DH	.280	.371	.552	22	59	2
5	김종모	LF	.350	.404	.524	11	44	7
6	김준환	RF	.248	.308	.362	10	43	11
7	김무종	C	.262	.313	.453	12	60	2
8	양승호	3B	.236	.292	.309	2	11	3
9	차영화	2B	.266	.308	.323	1	23	16

- (observation) Six out of 9 starting hitters have family name 'Kim'.
- (fact) According to census, ~20% of all Koreans have family name 'Kim'.
- (Hypothesis to test) The manager of 1983 Tigers (himself a 'Kim') has a bias toward players with family name 'Kim'.

Example: comparison of hypotheses

Given a set of data resulting from the measurement of some observable

$$\Omega = \{-1.0, -0.9, -0.7, -0.1, 0.0, 0.1, 0.2, 0.5, 0.6, 1.0\}$$

where the total number of data $N = 10$, determine which of the following models is a better description of the data:

- H_0 : the data are distributed according to a Standard Gaussian ($\mu = 0,\ \sigma = 1$);
- H_1 : the data are uniformly distributed over $[-1, +1]$

For a test statistic, we may use

$$R = \frac{P(H_0|\text{data})}{P(H_1|\text{data})} = \frac{P(\text{data}|H_0)P(H_0)}{P(\text{data}|H_1)P(H_1)} \qquad R \gtrless 1 \ ??$$

Example: comparison of hypotheses

Given a set of data resulting from the measurement of some observable

$$\Omega = \{-1.0, -0.9, -0.7, -0.1, 0.0, 0.1, 0.2, 0.5, 0.6, 1.0\}$$

$$R = \prod_{i=1}^{N} R_i$$

$$= \prod_{i=1}^{N} \frac{P(data|H_0)\,P(H_0)}{P(data|H_1)\,P(H_1)}$$

$$= \prod_{i=1}^{N} NG(x_i; 0, 1)$$

ω_i	$P(data\|H_0)$	$P(data\|H_1)$	R_i
−1.0	0.242	0.1	2.42
−0.9	0.266	0.1	2.66
−0.7	0.312	0.1	3.12
−0.1	0.397	0.1	3.97
0.0	0.399	0.1	3.99
0.1	0.397	0.1	3.97
0.2	0.391	0.1	3.91
0.5	0.352	0.1	3.52
0.6	0.333	0.1	3.33
1.0	0.242	0.1	2.42

taken from A. Bevan's book

model-independent test?

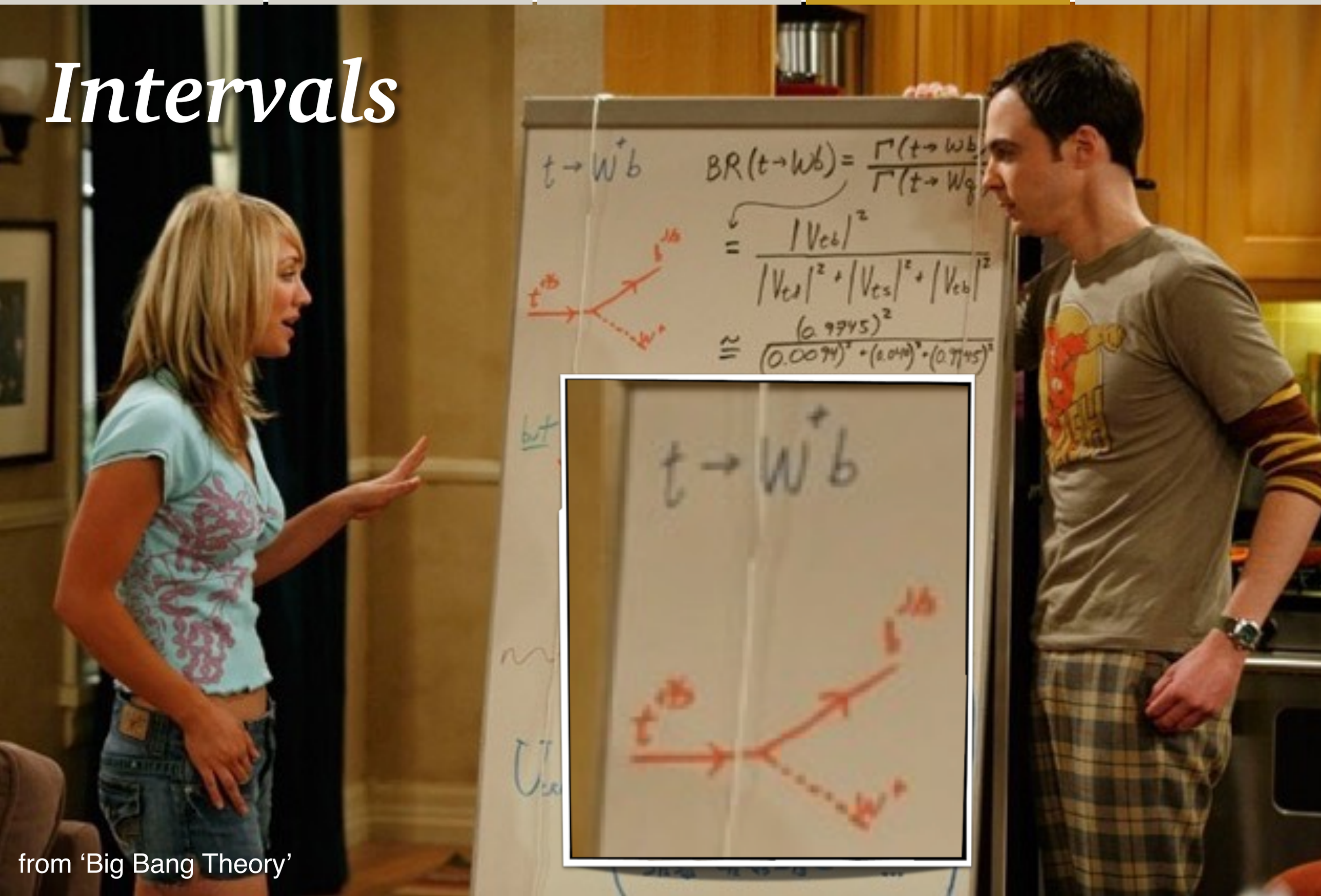
- In general, we cannot find a single critical region that gives the maximum power for all possible alternatives (no "uniformly most powerful" test)
- In HEP, we often try to construct a test of the Standard Model as H_0 (or sometimes called "background only")

  such that we have a well specified *false discovery rate* α (=prob. to reject H_0 when it is true),

  and high power w.r.t. some interesting alternative H_1, e.g. SUSY, Z', etc.
- But, there is no such thing as a *model-independent* test.

  Any statistical test will inevitably have high power w.r.t. some alternatives and less for others

Intervals



from ‘Big Bang Theory’

2016 Review of Particle Physics.
Please use this CITATION: C. Patrignani *et al.*(Particle Data Group), Chin. Phys. C, **40**, 1000

$t \to Wb$

$\Gamma(t \to Wb)/\Gamma(t \to Wq\,(q = b, s, d))$

OUR AVERAGE assumes that the systematic uncertainties are uncorrelated.

VALUE	DOCUMENT ID		TECN	COMMENT
0.957 ± 0.034	**OUR AVERAGE** Error includes scale factor of 1.5.			
0.87 ±0.07	1 AALTONEN	2014G	CDF	$\ell\ell + \not{E}_T +$ ≥ 2j (0,1,
1.014 ±0.003 ±0.032	2 KHACHATRYAN	2014E	CMS	$\ell\,\ell + \not{E}_T$ + 2,3,4j (0 −
0.94 ±0.09	3 AALTONEN	2013G	CDF	$\ell + \not{E}_T +$ ≥ 3jets (≥
0.90 ±0.04	4 ABAZOV	2011X	D0	

Measurement with errors

- Let's say we are reporting a single measurement

$$x = a \pm b$$

- **Frequentist interpretation**
  - Repeating the measurement many times under identical conditions ("ensemble"), the estimated interval will vary each time. **In 68.3% of those results, the true value of *x* will lie within the interval.**
- **Result of each measurement is a sampling from a Gaussian distribution G(μ,σ)**
  - We may not know μ
  - We have some idea about σ -- experimental sensitivity

when $\mu \pm \sigma$ is not enough…

If the PDF of the estimator is not Gaussian, or
if there are physical boundaries on the possible values of the parameter,

one usually quotes an interval given a confidence level.

a Bayesian procedure for intervals

$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\boldsymbol{x})\, d\theta$$

If the physical value is non-negative, one may choose a prior:

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases}$$

Likelihood for s, given b, is $\quad P(n|s) = \dfrac{(s+b)^n}{n!} e^{-(s+b)}$

If what we seek is of a very low (or no) signal, interval ➔ UL

Then,

$$1-\alpha = \int_{-\infty}^{s_{\rm up}} p(s|n) ds = \frac{\int_{-\infty}^{s_{\rm up}} P(n|s)\,\pi(s)\,ds}{\int_{-\infty}^{\infty} P(n|s)\,\pi(s)\,ds}$$

$F^{-1}_{\chi^2}$: inverse of the CDF ➔ $s_{\rm up} = \dfrac{1}{2} F^{-1}_{\chi^2}[1-\alpha; 2(n+1)] - b$

(Ex) UL on Poisson parameter

- Consider again the case of observing $n \sim \text{Poisson}(s+b)$.
  Suppose $b = 4.5$ and $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.
- Relevant alternative is $s = 0$, resulting in critical region at low n.
- The p-value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$.
  Therefore, the upper limit s_{up} at CL $= 1 - \alpha$ is obtained from

Frequentist "confidence intervals"

- **on repeated measurements**

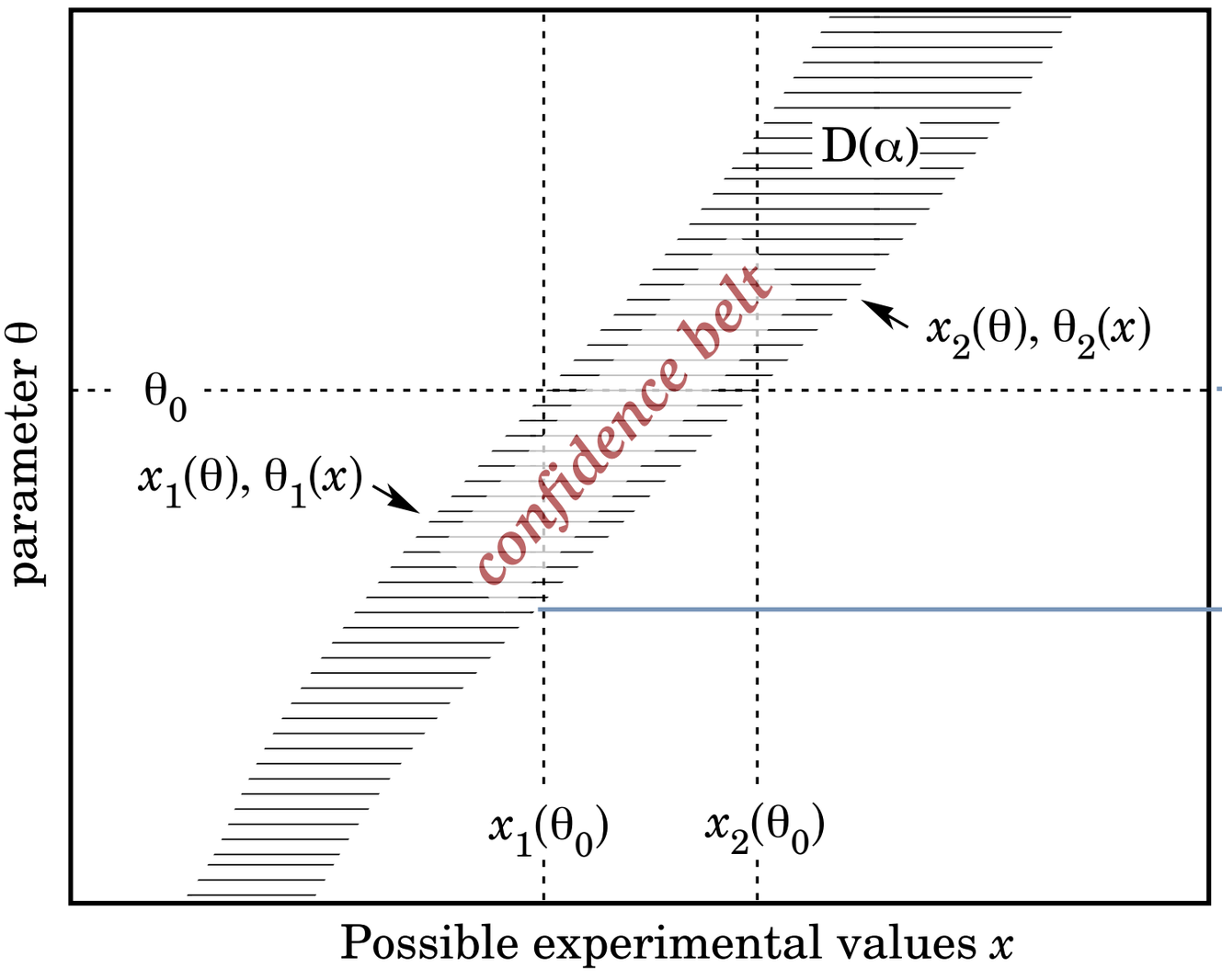
  Remember frequentist approach is always about repeated measurements!

- **"confidence interval"**

  = intervals constructed to include the true value of the parameter with a probability ≥ (*a specified value*)

Frequentist "confidence intervals"

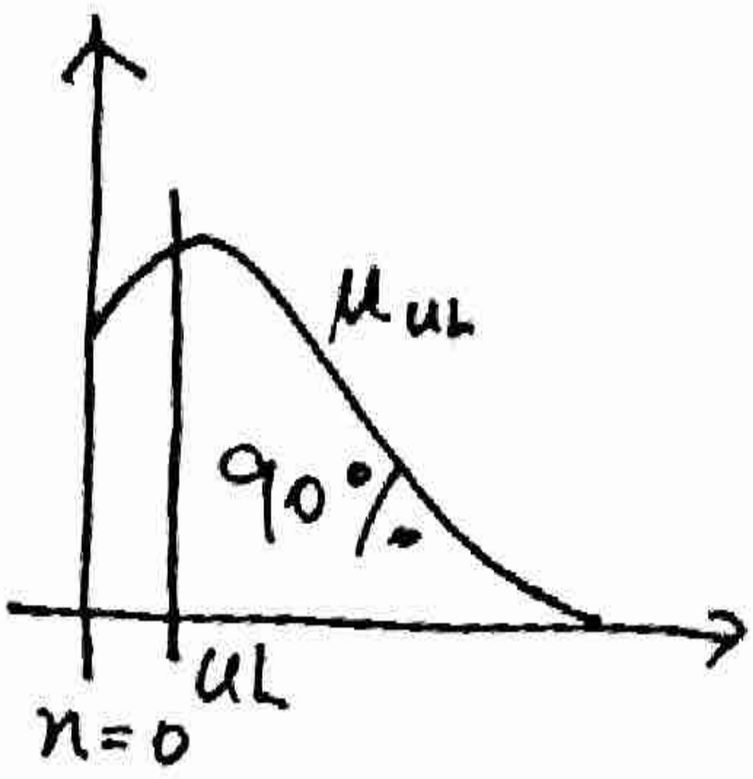
- **Consider a pdf $f(x;\theta)$** $\quad P(x_1 < x < x_2; \theta) = 1 - \alpha = \int_{x_1}^{x_2} f(x;\theta)\, dx$
  - x : outcome of an experiment
  - θ : unknown parameter for which we set the interval

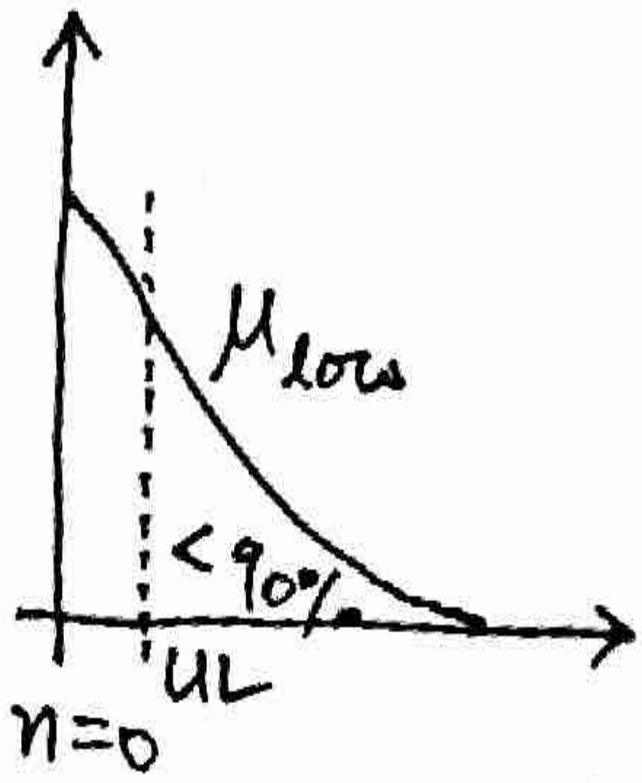


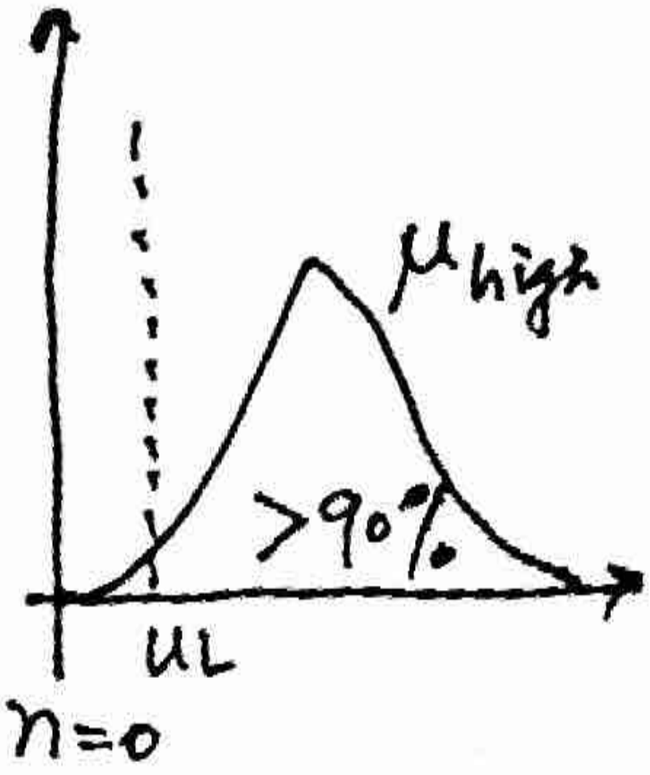
the confidence interval, given a measured outcome of $x_0 = x_1(\theta_0)$

"Neyman construction"

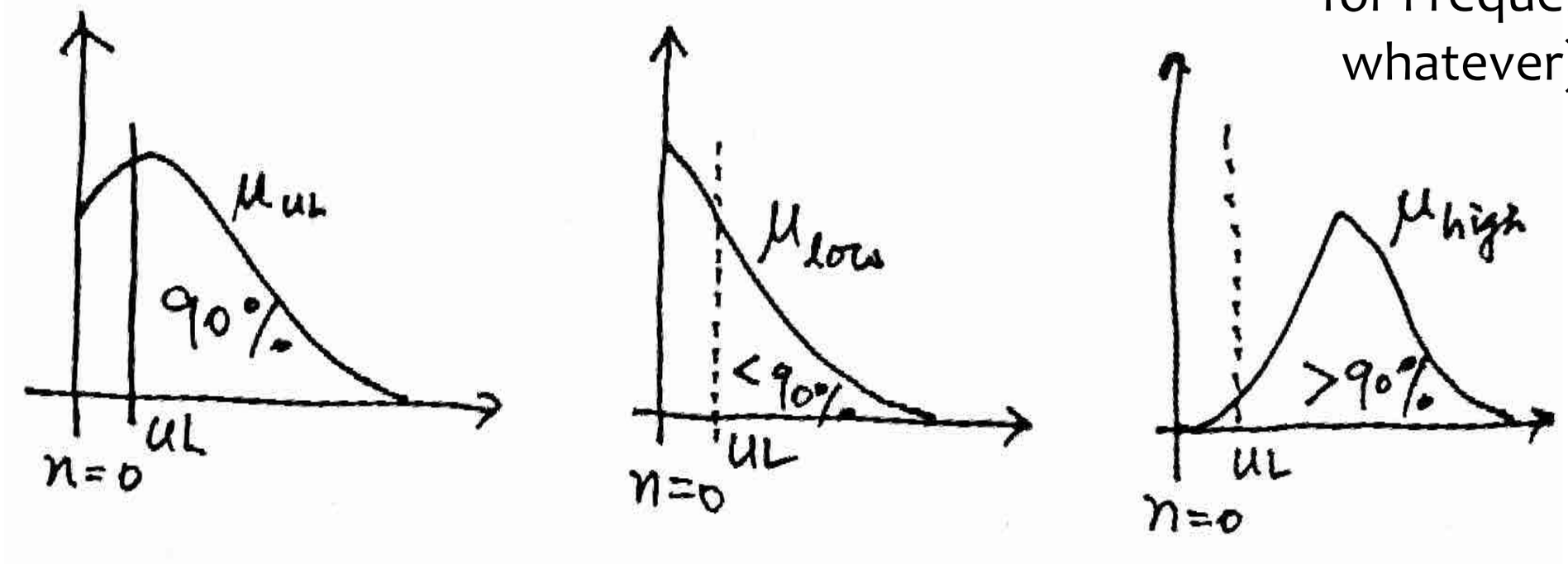
for Frequentist UL, the 90% (or whatever) integration is done above the UL







for Frequentist UL, the 90% (or whatever) integration is done above the UL



	$1-\alpha = 90\%$		$1-\alpha = 95\%$	
n	$\mu_{\rm lo}$	$\mu_{\rm up}$	$\mu_{\rm lo}$	$\mu_{\rm up}$
0	–	2.30	–	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15

Table 38.3:
naive frequentist interval

	$1-\alpha = 90\%$		$1-\alpha = 95\%$	
n	μ_1	μ_2	μ_1	μ_2
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76

Table 38.4: Feldman-Cousins interval
Phys. Rev. D57, 3873 (1998)

"unified approach"

Confidence interval from inversion of a test

- For confidence intervals for a parameter θ, define a test of size α for the hypothesized value θ (*repeat this for all* θ)
  - *If the observed data falls in the critical region, reject the value* θ.
  - The values that are *not rejected* constitutes a **confidence interval** for μ at confidence level $\mathrm{CL} = 1 - \alpha$.

- By construction the confidence interval will contain the true value of θ with probability $\geq 1 - \alpha$.
  * The interval depends on the choice of the test (critical region).
  * If the test is formulated in terms of a p-value, p_θ, then the confidence interval represents those values of θ for which $p_\theta > \alpha$.
  * To find the end points of the interval, set $p_\theta = \alpha$ and solve for θ.

Coincidence of frequentist and Bayesian intervals

- If the expected background is zero, the Bayesian upper limit (for a Poisson RV) becomes equal to the limit determined by frequentist approach.

$$s_{\mathrm{up}} = \frac{1}{2} F^{-1}_{\chi^2}\left[p, 2(n+1)\right] - b$$

$$= \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha; 2(n+1))$$

For more details, you may read e.g.
a statistics review in PDG. http://pdg.lbl.gov/2015/reviews/rpp2015-rev-statistics.pdf

Parameter Estimation

Basics of parameter estimation

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

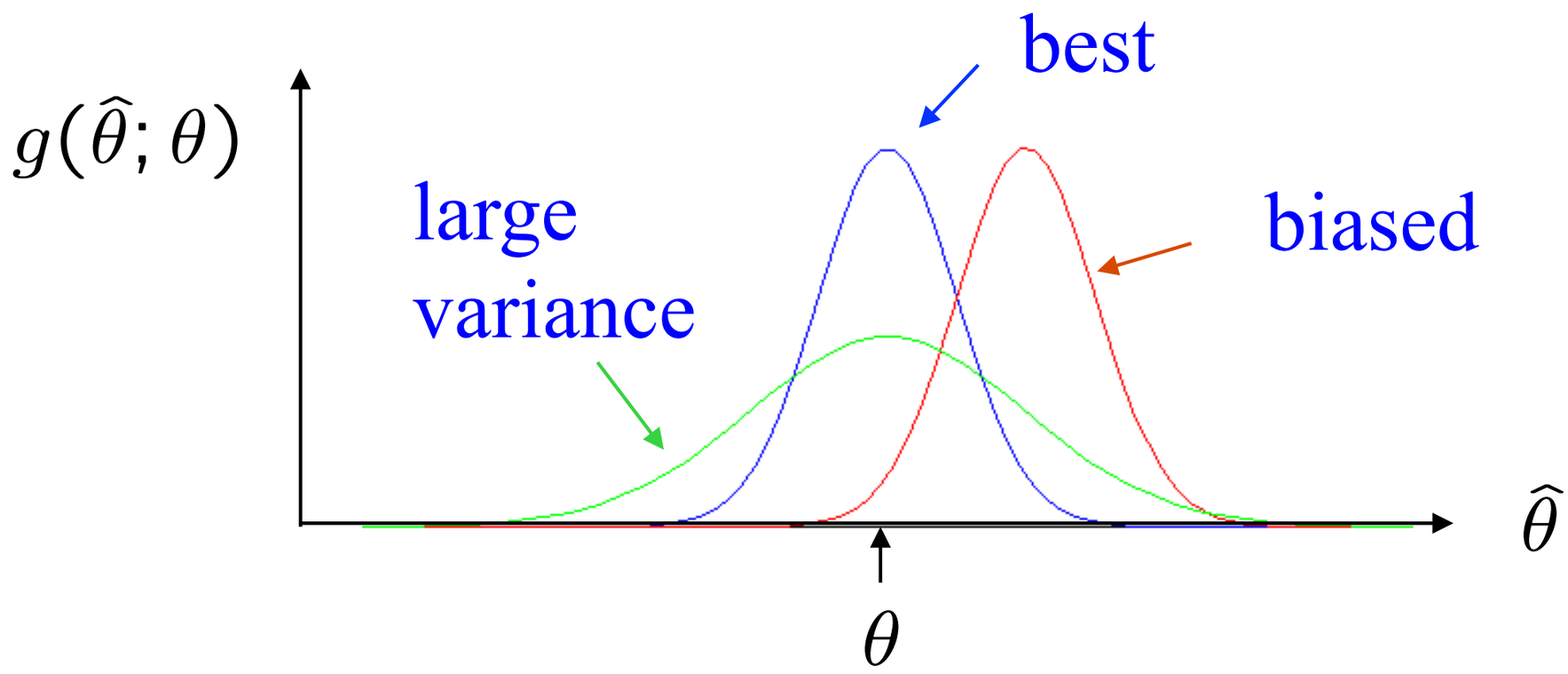
$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

  where θ is the parameter, while x is the random variable.

- Suppose we have a sample of observed values, $\vec{x}$.
  We want to find some function of the data to *estimate* the parameter(s): $\hat{\theta}(\vec{x})$.
  Often $\hat{\theta}$ is called an **estimator**.

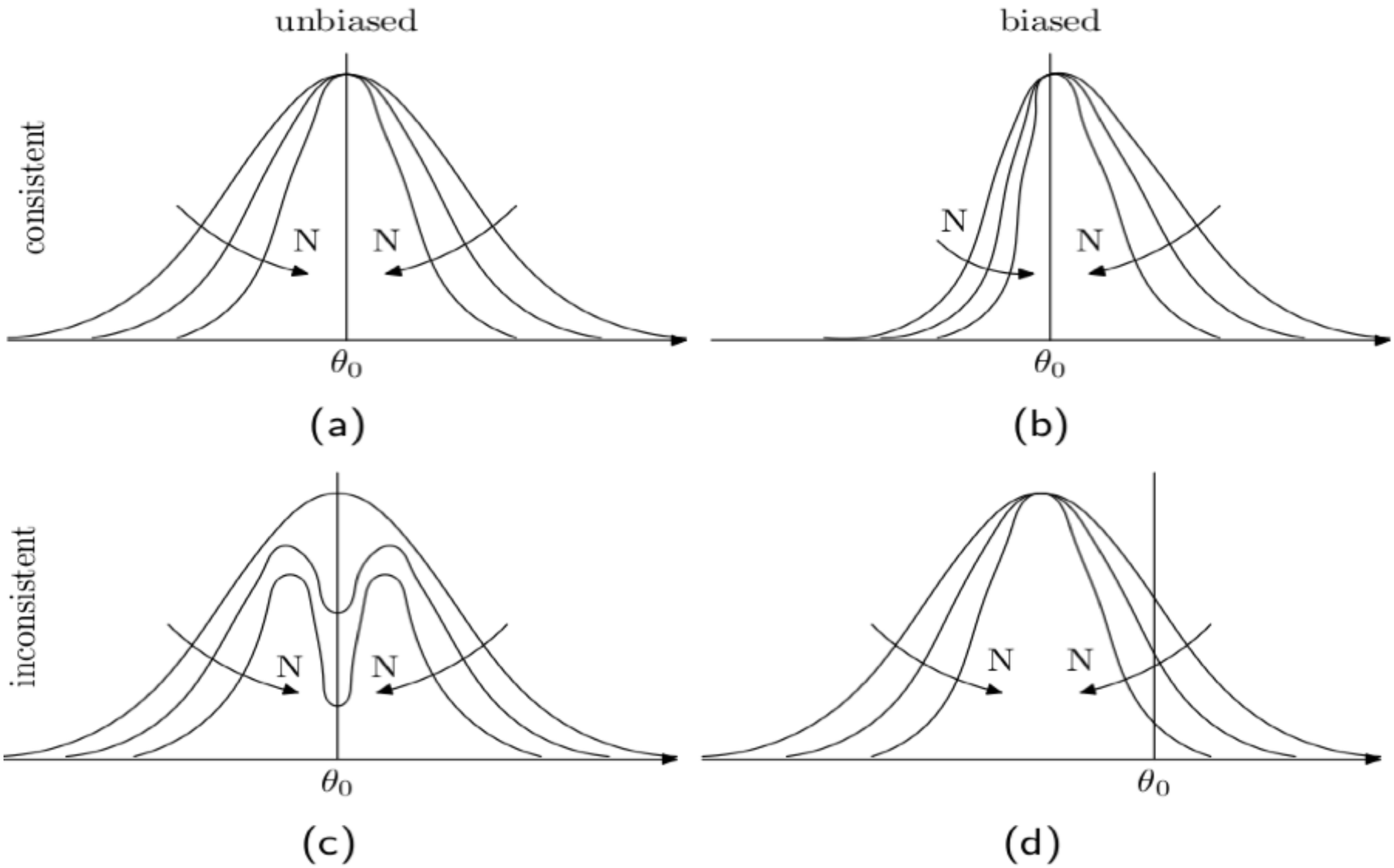
Properties of estimators

- If we were to repeat the entire measurement, the set of estimates would follow a PDF:



  - We want small (or zero) bias ($\Rightarrow$ syst. error): $b = E[\hat{\theta}] - \theta$
  - and we want a small variance ($\Rightarrow$ stat. error): $V[\hat{\theta}]$

Bias vs. Consistency



Likelihood function

- Suppose the entire result of an experiment (*set of measurements*) is a collection of numbers $\vec{x}$, and suppose the joint PDF for the data $\vec{x}$ is a function depending on a set of parameters $\vec{\theta}$: $f(\vec{x};\vec{\theta})$
- Evaluate this function with the measured data $\vec{x}$, regarding this as a function of $\vec{\theta}$ only. This is the **likelihood function**.

$$L(\vec{\theta}) = f(\vec{x};\vec{\theta}) \ (\vec{x}, \text{fixed})$$

The likelihood function for i.i.d. data

i.i.d. = *independent and identically distributed*

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

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

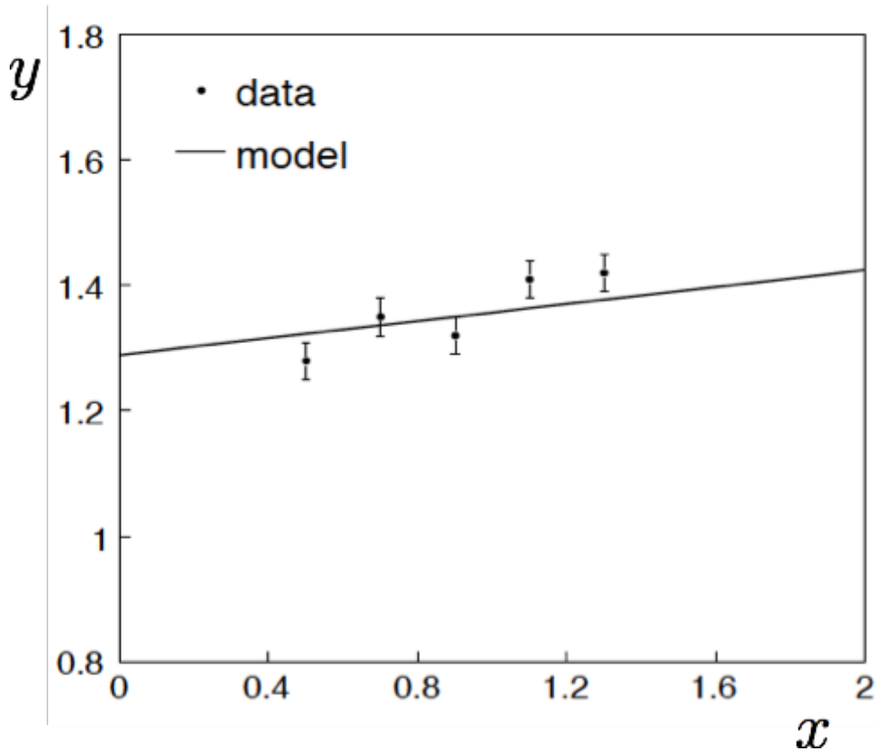
- In this case, the likelihood function is

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

So we define the max. likelihood (ML) estimator(s) to be the parameter value(s) for which the L becomes maximum.

ML estimator example: fitting to a straight line

- Suppose we have a set of data:
  $(x_i, y_i, \sigma_i),\ i = 1, \cdots, n.$
- Modeling: y_i are independent and follow $y_i \sim G(\mu(x_i), \sigma_i)$ (G: Gaussian) where $\mu(x_i)$ are modelled as
  $\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x$
  Assume x_i and σ_i are known.
- Goal: to estimate θ_0
  Here, let's suppose we don't care about θ_1 (an example of a *nuisance parameter*)



ML fit with Gaussian data

- In this example, the y_i are assumed independent, so that likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right]$$

- Then maximizing L is equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2\ln L(\theta_0, \theta_1) + C = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}$$

i.e., for Gaussian data, ML fitting is the same as the method of least squares

Wilk's theorem

ML fit or Least-square fit?

- **Consider we have a random variable $x \in [0, 3]$, and a distribution $f(x)$.**
- **In a series of measurements, we obtained**
  - 9 events in [0,1), 10 events in [1,2), and 8 events in [2,3]
  - We have a model of uniform $f(x)$, and would like to estimate the mean value of $\int f(x)\, dx$ for each histogram bin.
- **Run a thought-experiment, comparing**
  - maximum likelihood method, and least-square method
  - *Do they give the same result?*

Bayesian likelihood function

- Suppose our L-function contains two parameters θ_0 and θ_1, where we have some knoweldege about the prior probability on θ_1 from previous measurements:

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\pi_1(\theta_1)$$
$$\pi_0(\theta_0) = \text{const.}$$
$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_p} e^{-(\theta_1-\theta_p)^2/2\sigma_p^2}$$

- Putting this into the Bayes' theorem gives the posterior probability:

$$p(\theta_0, \theta_1|\vec{x}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i-\mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2} \; \pi_0 \; \frac{1}{\sqrt{2\pi}\sigma_p} e^{-(\theta_1-\theta_p)^2/2\sigma_p^2}$$

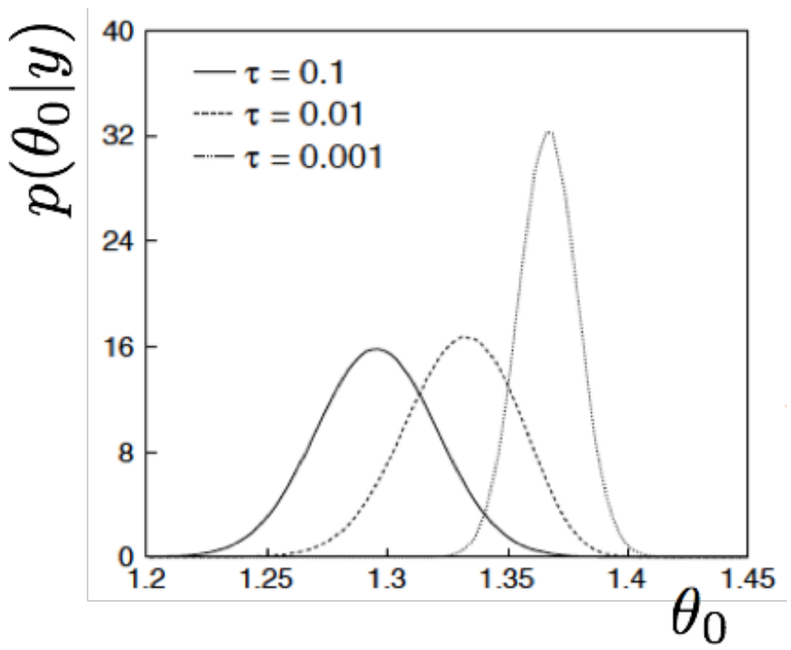
- Then, $p(\theta_0|\vec{x}) = \int p(\theta_0, \theta_1|\vec{x})\, d\theta_1$

with alternative priors

- Suppose we don't have a previous measurement of θ_1 but rather a theorist saying that θ_1 should be > 0 and not too much greater than, say, 0.1 or so. In that case, we may try modeling the prior for θ_1 as something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau},\ \theta_1 \geq 0,\ \tau = 0.1$$

- From this we obtain (numerically) the posterior PDF for θ_0

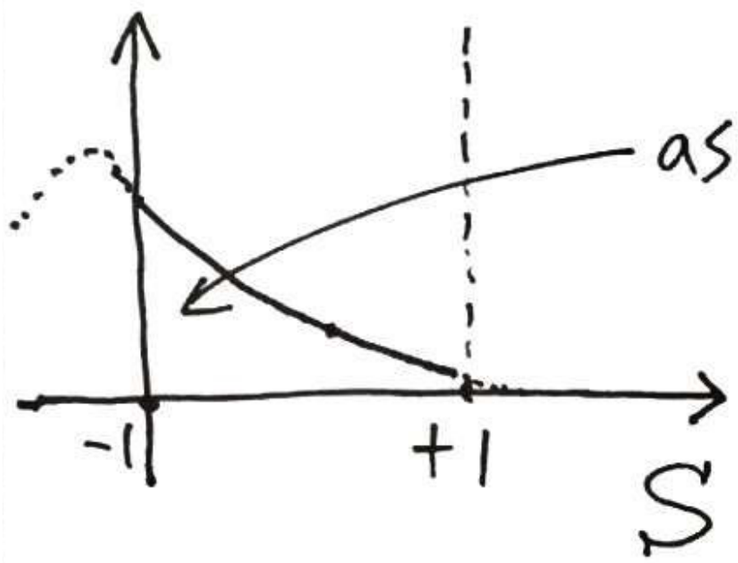


- This plot summarizes all knowledge about θ_0.

Exercises

1. Setting limits:

   (a) The parameter S is measured to be -1.1 ± 0.4. What is the Bayesian 90% CL upper limit on S given that S is physically bound within the interval $[-1, +1]$?



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

$\pi(H)$ = prior probability; flat over [-1, +1]

$P(\vec{x}|H) =$

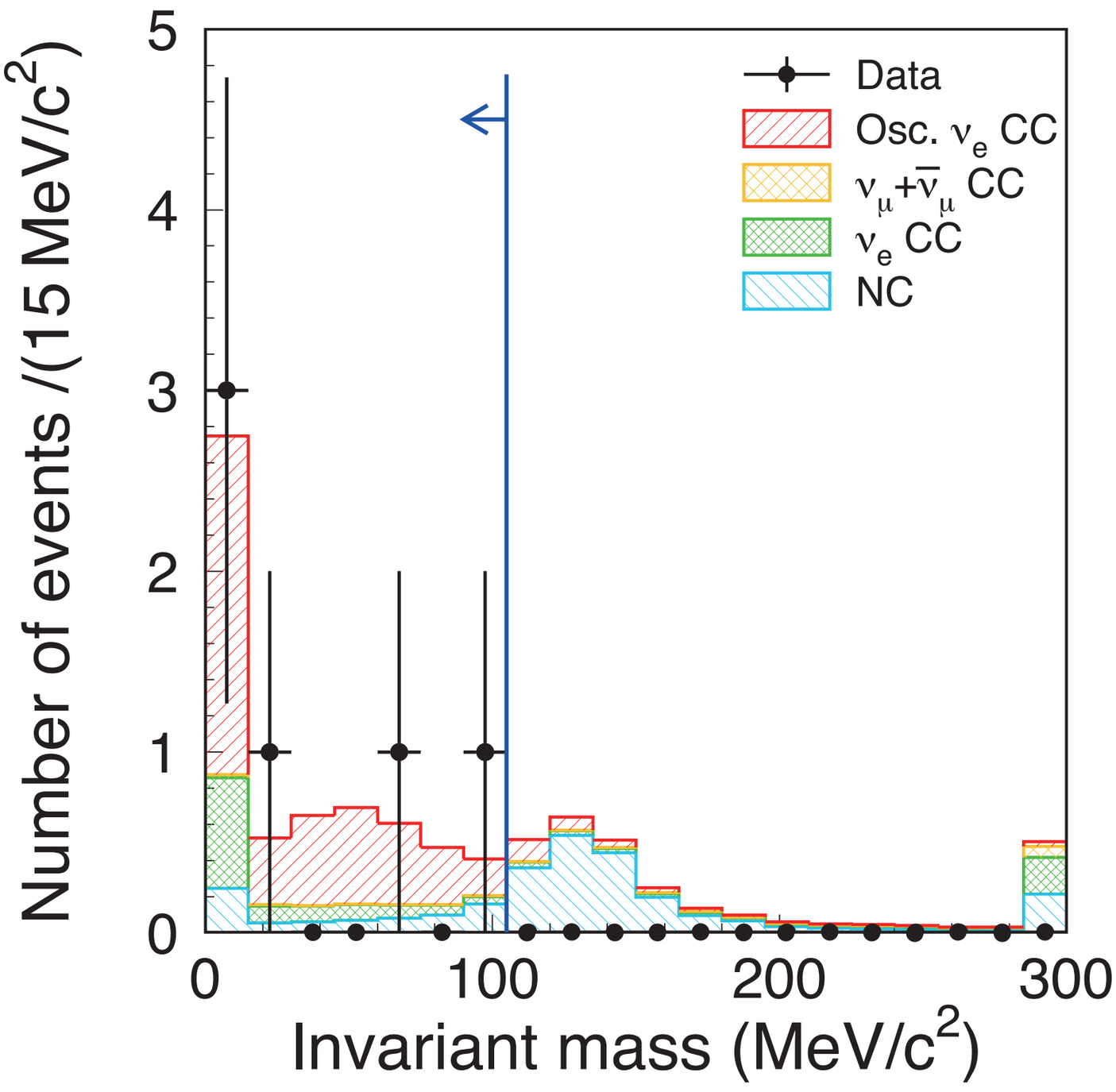
$\int P(\vec{x}|H)\,\pi(H)\,dH =$

But, what should we use for μ, σ ?

(b) Determine the 90% CL upper limit on the signal yield μ given a background expectation of one event (ignore systematic uncertainties) and an observed yield of one event for a rare decay search.

(Example) a T2K result

PRL 107, 041801 (2011)



T2K observed 6 candidate events of $\nu_\mu \rightarrow \nu_e$ while a background of 1.5±03 events is expected.

- How significant is this signal?
- How to include the systematic uncertainty in the analysis?
- What is the relevant 'limit' from this result?

(Ex) "Bayesian classifier"

3. The ratio of energy and momentum (E/p) for a charged particle is used to identify electrons against other particle types. Assume that it takes a value within the interval $[0.0, 1.2]$.

   (a) Assuming that the production of all three types of charged particles are equal, define a Bayesian classifier based on a Gaussian distributions with $\mu = 1.0$ for an electron, 0.38 for $\pi^{\pm}$ and 0.08 for $\mu^{\pm}$. For simplicity, let's assume all three PDFs have a width of $\sigma = 0.2$. Classify the tracks with the following measured values of E/p: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0.

   (b) Repeat the above problem, but with the assumption that the expected fraction of charged particles in the experiment is 80% for $\pi^{\pm}$, 10% for $\mu^{\pm}$ and 10% for $e^{\pm}$.

(Ex) "Bayesian classifier"

3. The ratio of energy and momentum (E/p) for a charged particle is used to identify electrons against other particle types. Assume that it takes a value within the interval $[0.0, 1.2]$.

   (a) Assuming that the production of all three types of charged particles are equal, define a Bayesian classifier based on a Gaussian distributions with $\mu = 1.0$ for an electron, 0.38 for $\pi^{\pm}$ and 0.08 for $\mu^{\pm}$. For simplicity, let's assume all three PDFs have a width of $\sigma = 0.2$. Classify the tracks with the following measured values of E/p: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0.

   (b) Repeat the above problem, but with the assumption that the expected fraction of charged particles in the experiment is 80% for $\pi^{\pm}$, 10% for $\mu^{\pm}$ and 10% for $e^{\pm}$.

- Given some data that can be tested against a set of classifications given by hypotheses H_i.
- For each event ω_j in the data set we compute $P(\omega_j|H_i)$ for all i.
- Find i where H_i gives the largest value of P over all i's, and we classify the event ω_j as belonging to the category j.