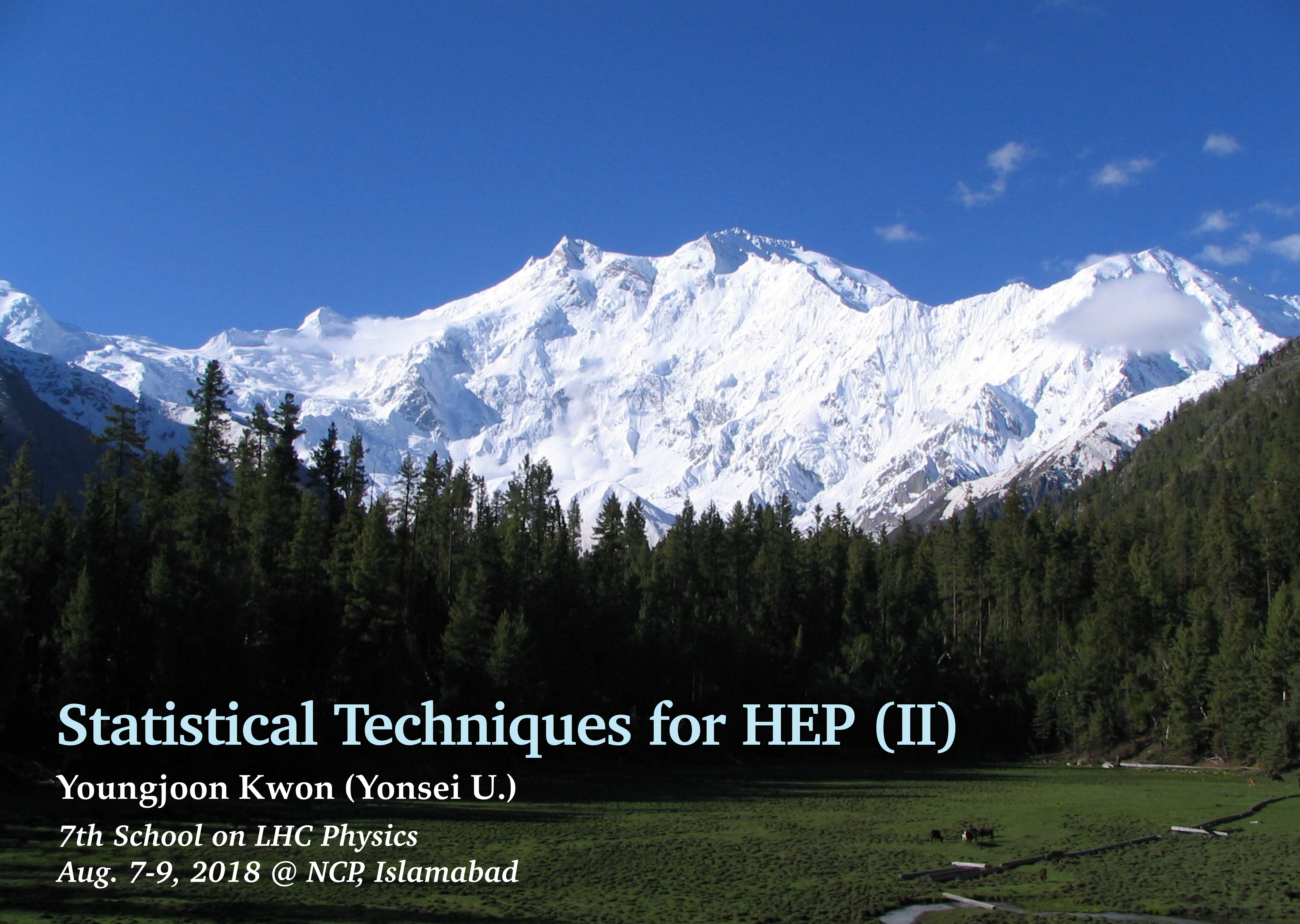


Statistical Techniques for HEP (II)

Youngjoon Kwon (Yonsei U.)

7th School on LHC Physics

Aug. 7-9, 2018 @ NCP, Islamabad

Outline

- **Basic elements**
  - some vocabulary
  - Probability axioms
  - some probability distributions
- **Two approaches: Frequentist vs. Bayesian**

Part I

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

Parts II & III

A TALE OF TWO STATISTICS …

Frequentist vs. Bayesian

"Bayesians address the question everyone is interested in by using assumptions no-one believes,

while Frequentists use impeccable logic to deal with an issue of no interest to anyone."

"*Bayes and Frequentism: a particle physicist's perspective*"
by Louis Lyons, arXiv:1301.1273

Two approaches

Relative frequency

A, B, ... are outcomes of a repeatable experiment **Frequentist**

$$P(A) = \lim_{n\to\infty} \frac{\text{times outcome is } A}{n}$$

Subjective probability

A, B, ... are hypotheses (statements that are true or false) **Bayesian**

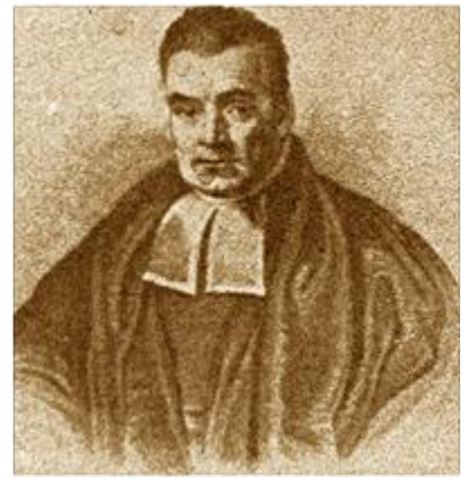
$$P(A) = \text{degree of belief that } A \text{ is true}$$

Frequentist approach is, in general, easy to understand, but some HEP phenomena are best expressed by subjective prob., e.g. systematic uncertainties, prob(Higgs boson exists), ...

Bayes' theorem

From the definition of conditional prob., we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$



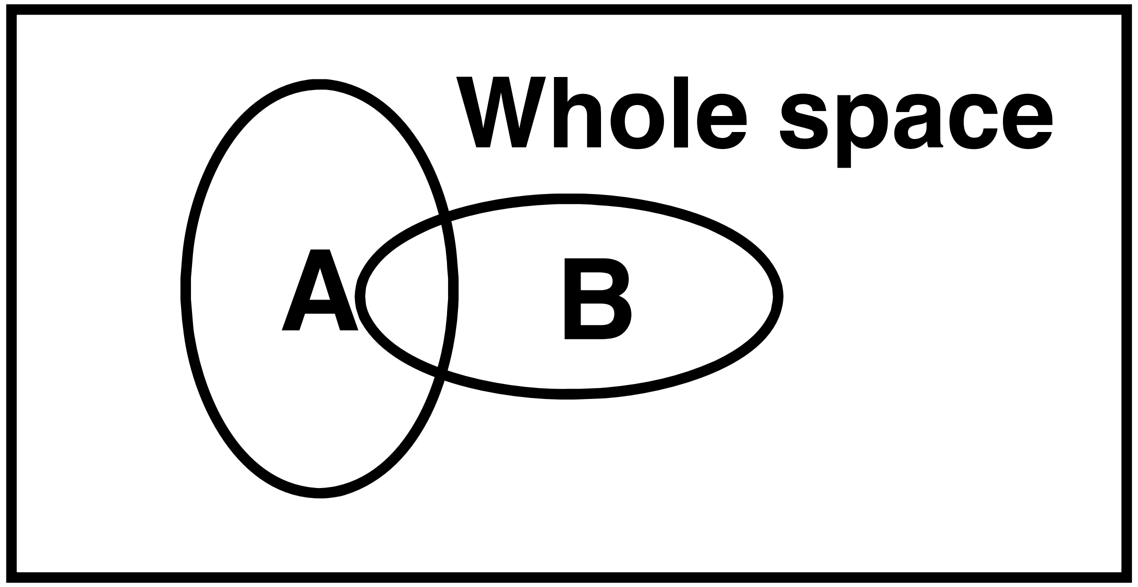
- but $P(A \cap B) = P(B \cap A)$
- therefore,

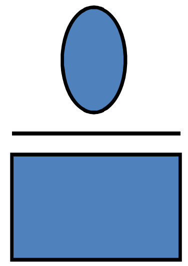
$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

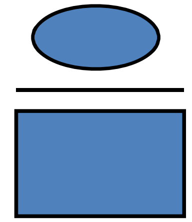
- First published (posthumous) by Rev. Thomas Bayes (1702-1761)

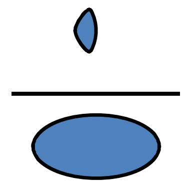
An essay towards solving a problem in the doctrine of chances,
Phil. Trans. R. Soc. 53 (1763) 370.

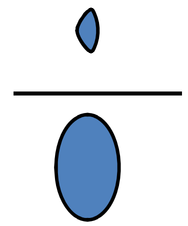
P, Conditional P, and Derivation of Bayes' Theorem in Pictures

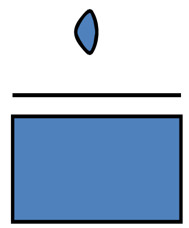


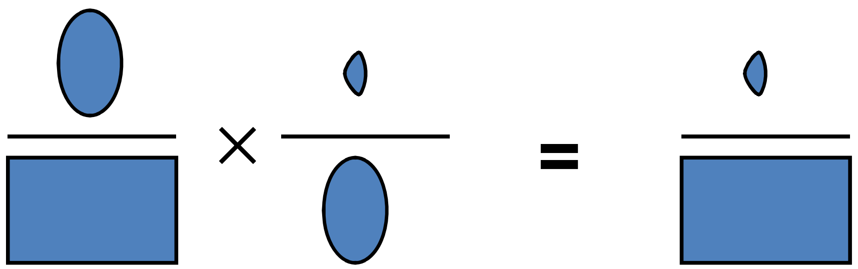
P(A) = 

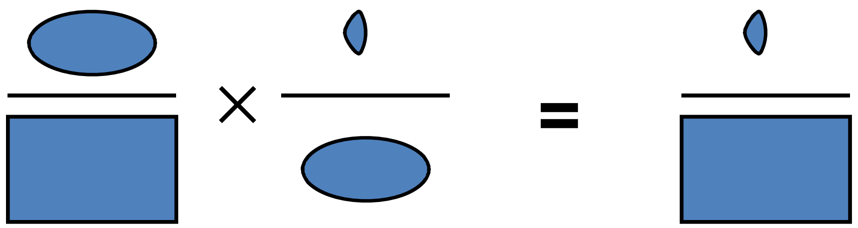
P(B) = 

P(A|B) = 

P(B|A) = 

$P(A \cap B)$ = 

$P(A) \times P(B|A)$ =  = $P(A \cap B)$

$P(B) \times P(A|B)$ =  = $P(A \cap B)$

Bob Cousins, CMS, 2008

$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$

Frequentist statistics – general philosophy

- In frequentist statistics, probabilities such as

  $P(\text{SUSY does exist})$

  $P(0.117 < \alpha_s < 0.121)$

  are either 0 or 1, but we don't have the answer

Bayesian statistics – general philosophy

- In Bayesian statistics, interpretation of probability is extended to the **degree of belief** (*i.e.* subjective).
- suitable for **hypothesis testing** (but no golden rule for priors)

probability of the data assuming hypothesis H (the likelihood)

prior probability, i.e., before seeing the data

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

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

- can also provide more natural handling of non-repeatable things: *e.g.* systematic uncertainties, P(Higgs boson exists)

(Ex) Bayesian answer for coin toss

Suppose I stand to win or lose money in a single coin-toss. My companion gives me a coin to use for the game.

- Do I trust the coin? What is $P(\text{faircoin})$?
- Frequentist answer:
  - toss the coin n times
  - $P(\text{heads}) = \lim_{n\to\infty}(n_{\text{H}}/n)$
  - make a complicated statement about the results, which is *only indirectly* about whether the coin is fair ...
- But I can only test the coin with five throws:
  - What if I get 4H, 1T?
  - Do I trust the coin, or claim that the game is unfair?
- What about Bayesian answer?

(Ex) Bayesian answer for coin toss

Assume: a 'bad' coin has a 75% probability to show 'head'
for a 'fair' coin, it's 50%

Priors:

$$P(\text{fair}|\text{BG}) = 0.50$$
$$P(\text{bad}|\text{BG}) = 0.50$$

Likelihoods:

$$P(4\text{H}, 1\text{T}|\text{fair}) = 0.1563$$
$$P(4\text{H}, 1\text{T}|\text{bad}) = 0.3955$$

Posterior:

$$P(\text{fair}|4\text{H}, 1\text{T}, \text{BG}) = \frac{P(4\text{H}, 1\text{T}|\text{fair}) \cdot P(\text{fair}|\text{BG})}{\sum_i P(4\text{H}, 1\text{T}|i) \cdot P(i|\text{BG})}$$
$$= \frac{0.1563 \cdot 0.50}{0.1563 \cdot 0.50 + 0.3955 \cdot 0.50} = 0.283$$

(Ex) Bayesian answer for coin toss

Assume: a 'bad' coin has a 75% probability to show 'head'
for a 'fair' coin, it's 50%

Priors:
$$P(\text{fair}\,|\,\text{GG}) = 0.95$$
$$P(\text{bad}\,|\,\text{GG}) = 0.05$$

Likelihoods:
$$P(4\text{H}, 1\text{T}|\text{fair}) = 0.1563$$
$$P(4\text{H}, 1\text{T}|\text{bad}) = 0.3955$$

Posterior:

$$P(\text{fair}|4\text{H}, 1\text{T}, \text{GG}) = \frac{P(4\text{H}, 1\text{T}|\text{fair}) \cdot P(\text{fair}|\text{GG})}{\sum_i P(4\text{H}, 1\text{T}|i) \cdot P(i|\text{GG})}$$

$$= 0.88$$

Frequentist or Bayesian, which one to use?

- 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

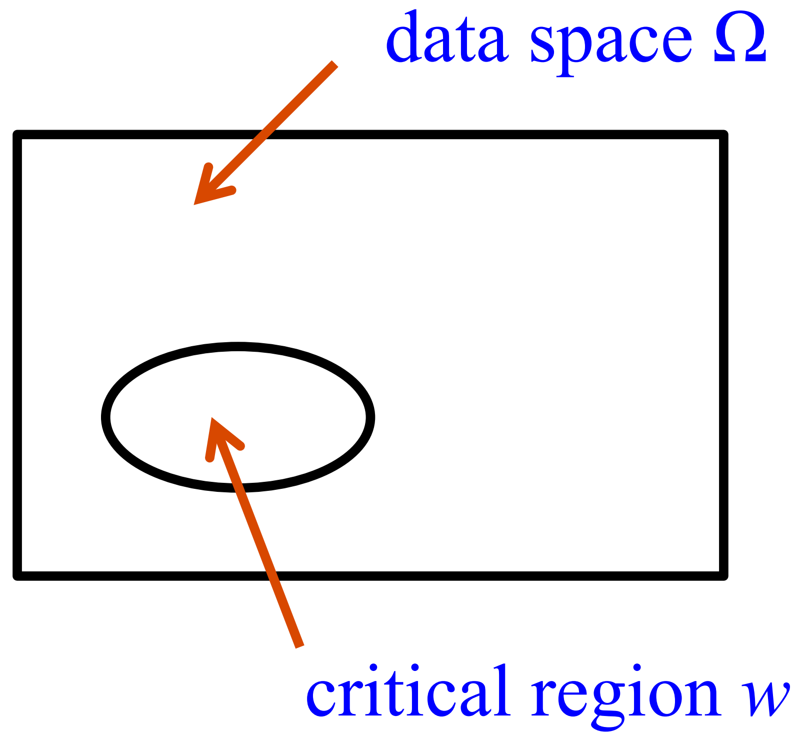
- A hypothesis H specifies the probability for the data
  (*shown symbolically as* $\vec{x}$ *here*),
  often expressed as a function $f(\vec{x}|H)$
- The measured data $\vec{x}$ could be anything:
  - observation of a single particle, a single event, or an entire experiment
  - uni-/multi-variate, continuous 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 is it?*

- 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

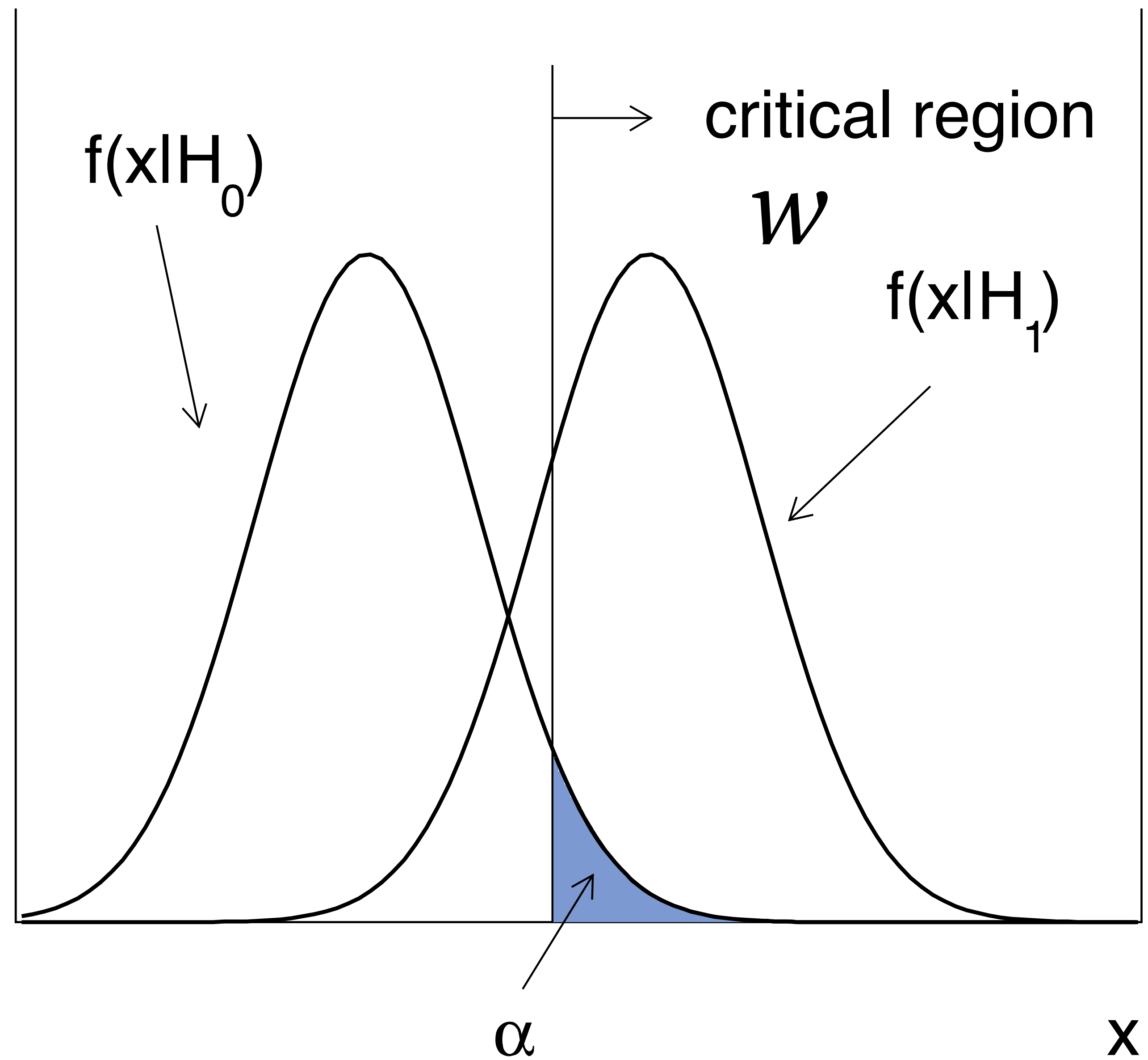
$$P(\vec{x} \in w | H_0) \leq \alpha$$



- 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, ∃ an ∞ number of possible critical regions that give the same significance level α.
- Usually, we place the critical region against an alternative hypothesis H_1 such that the probability to find an event in *w* is low (α) if H_0 is true, but high if the alternative (H_1) is true.



Test statistic

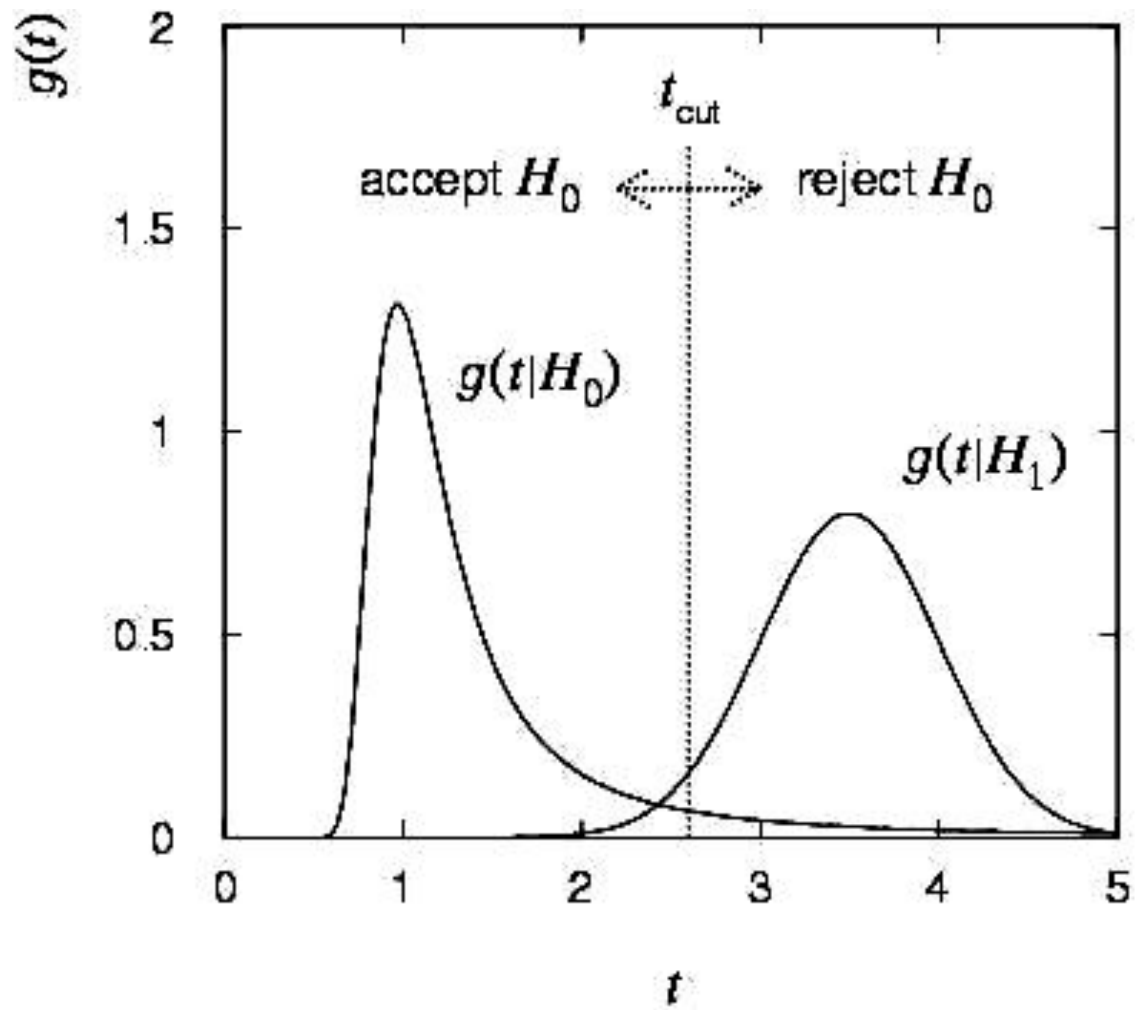
- The boundary surface of the critical region for an n-dim. data space can be defined by an equation of the form:

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

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**

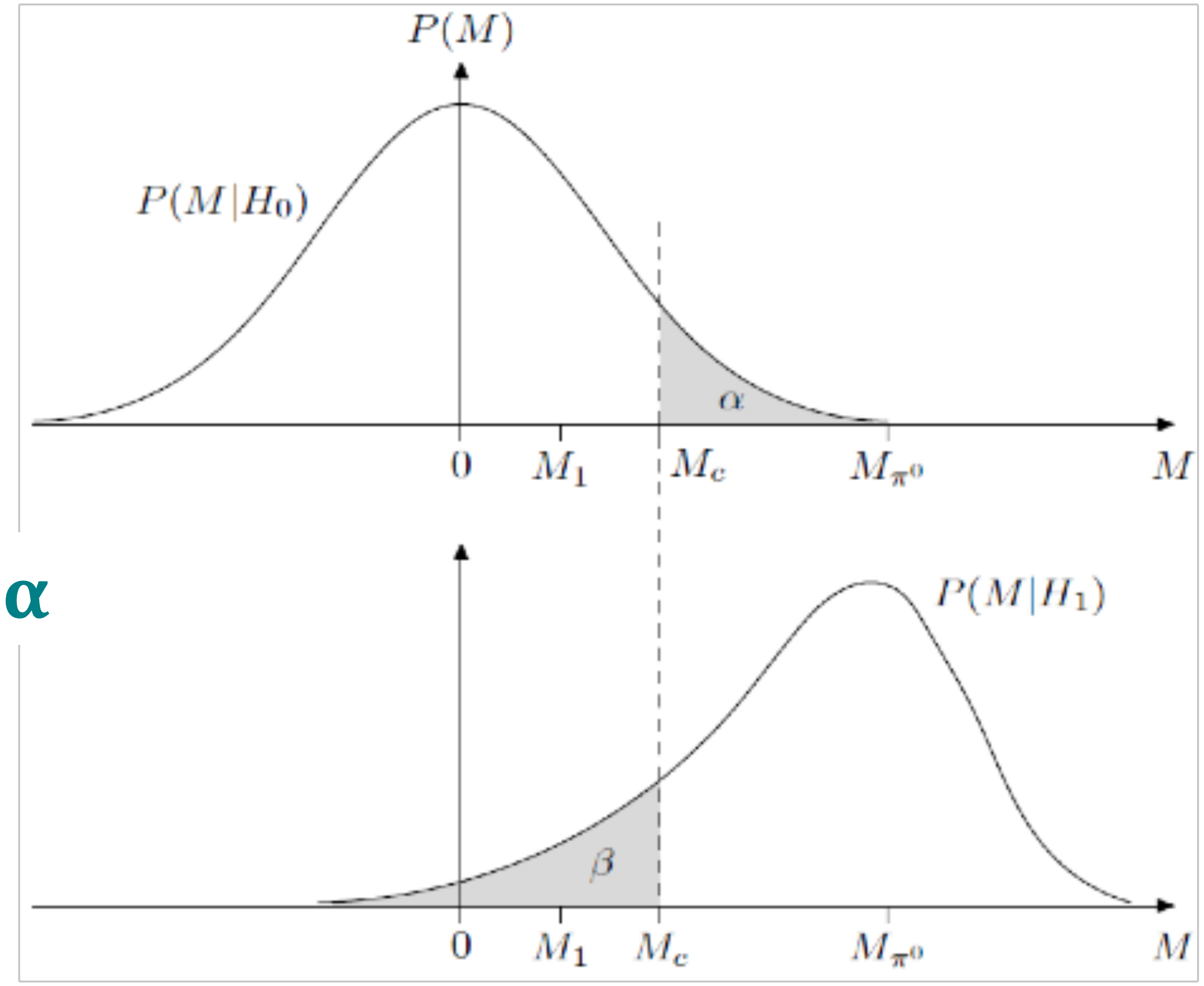
  The probability β of Type-II error:

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

  $1 - \beta$ 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



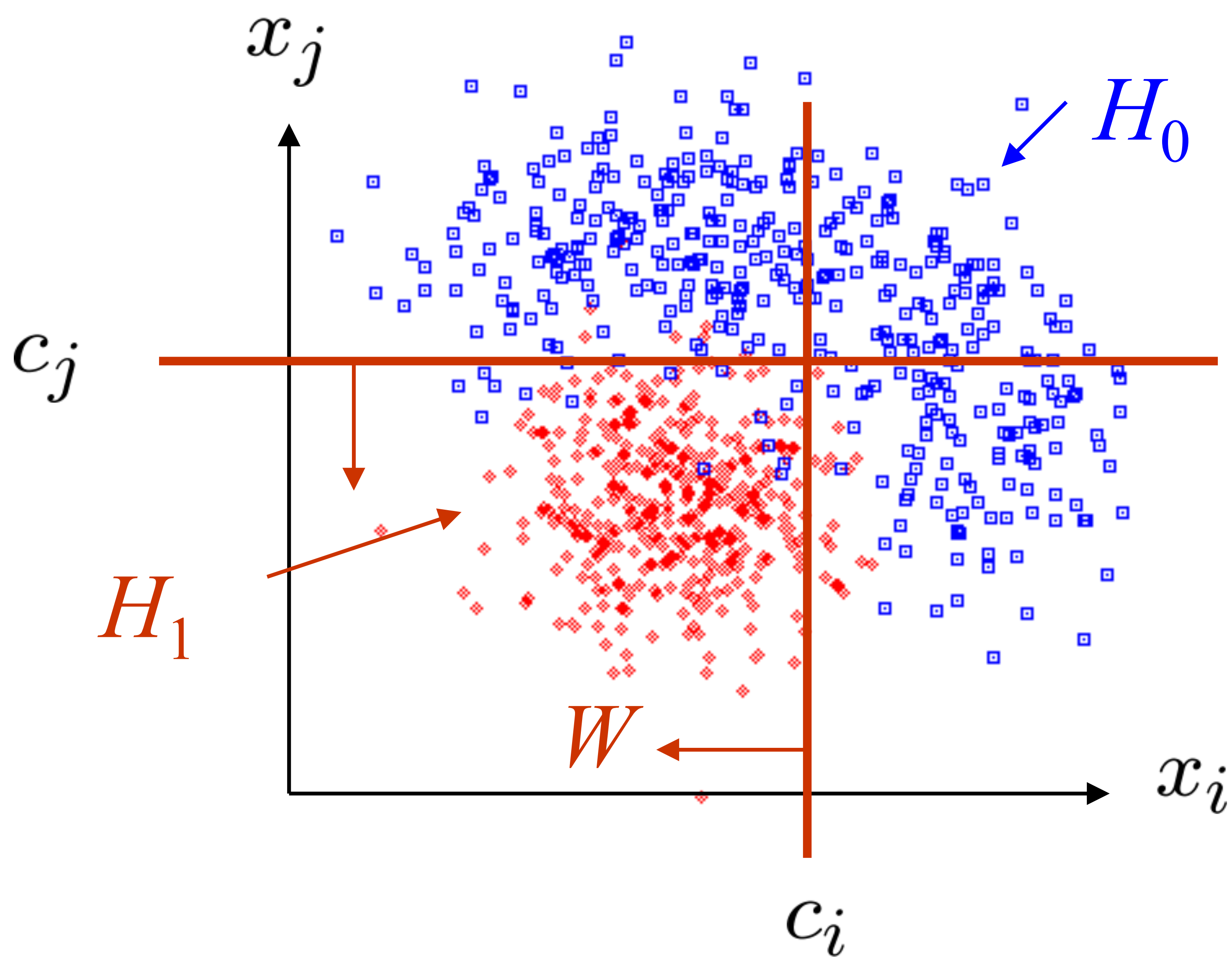
Optimal decision: minimize β for given α

from an FAPPS09 Lecture by S. T'Jampens

Defining a multivariate critical region

with "square cuts"

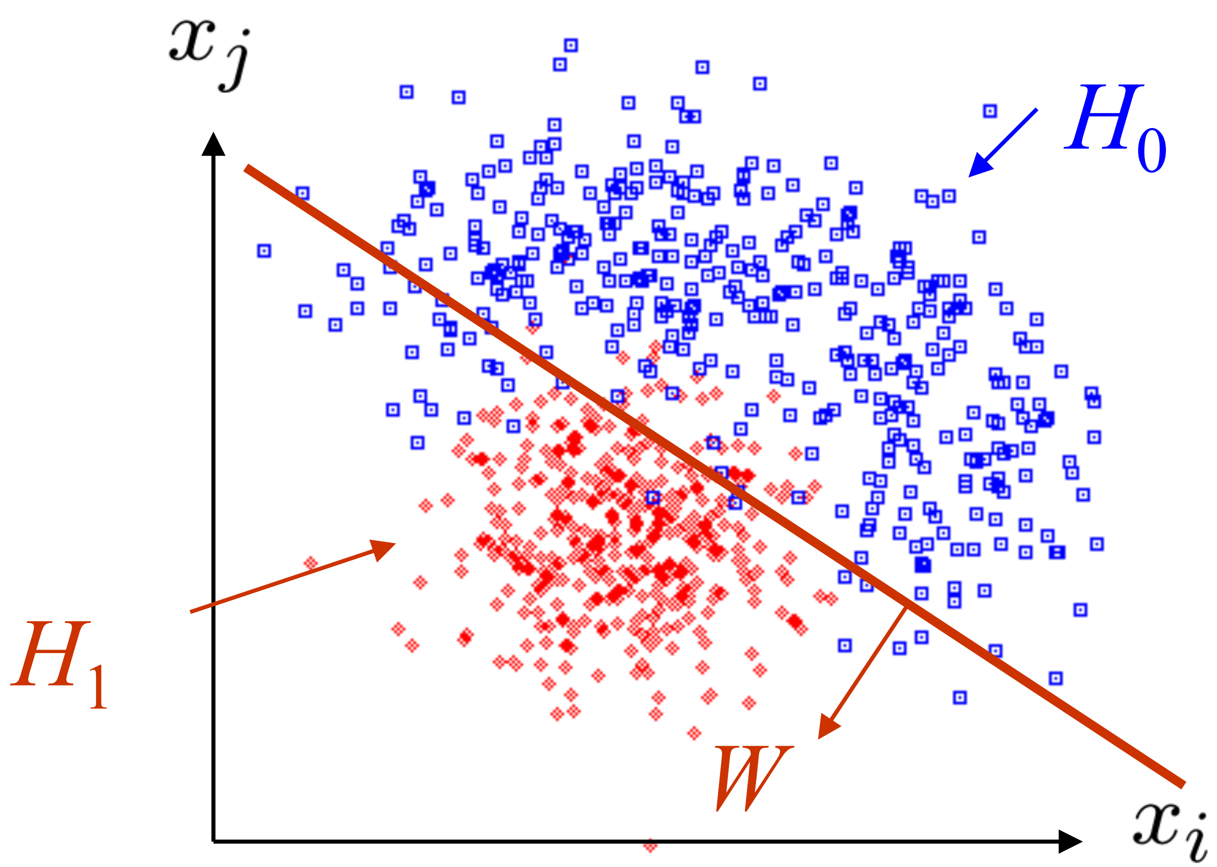
$$x_i < c_i$$
$$x_j < c_j$$



Defining a multivariate critical region

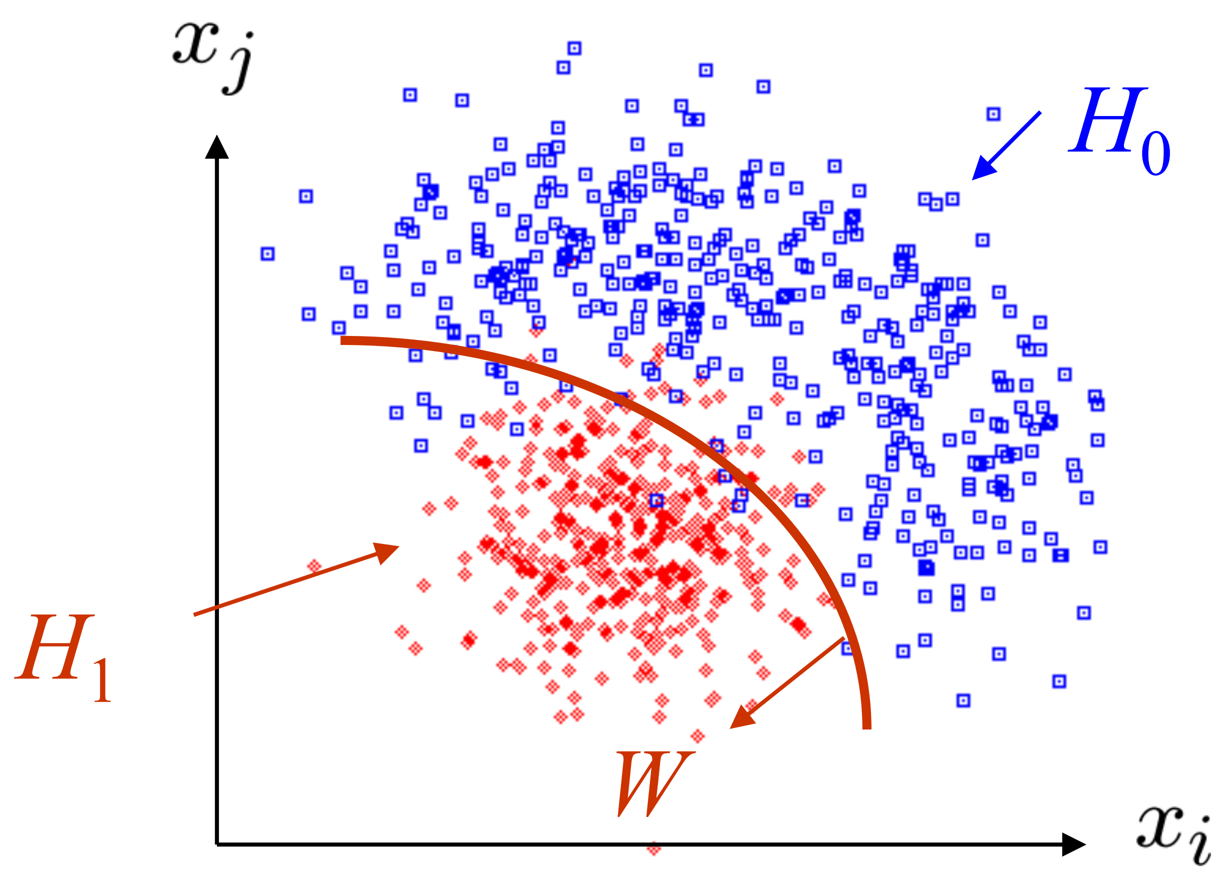
some more sophisticated ways

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
- ...

∃ many excellent software tools to do multivariate analysis.
Please explore yourself!

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.

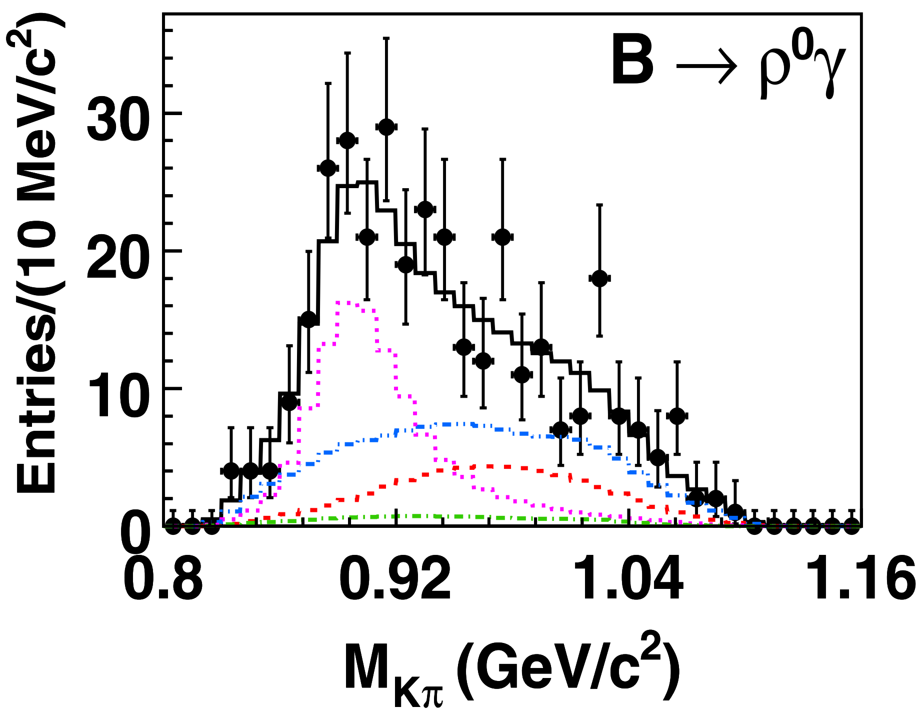
- Equivalently, the optimal scalar test statistic is

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

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

exercise on Type-I, II errors

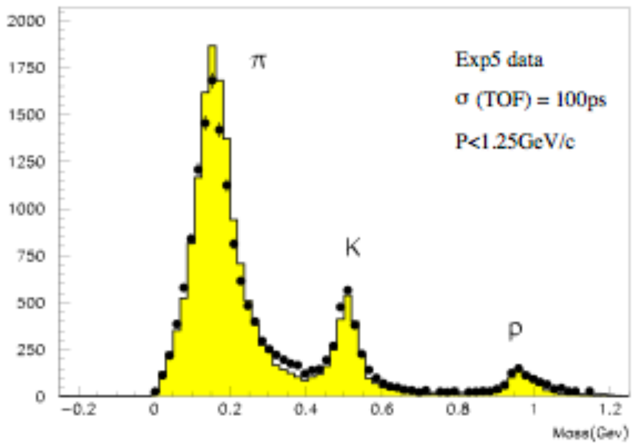
Since $B \to K^*\gamma$ has much higher branching fraction than $B \to \rho\gamma$, the former can be a serious background to the latter. It is crucial to understand the "efficiency" and "fake rate" of K/π identification system of your experiment in this study. The figure below shows the $M_{K\pi}$ invarianbt 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).

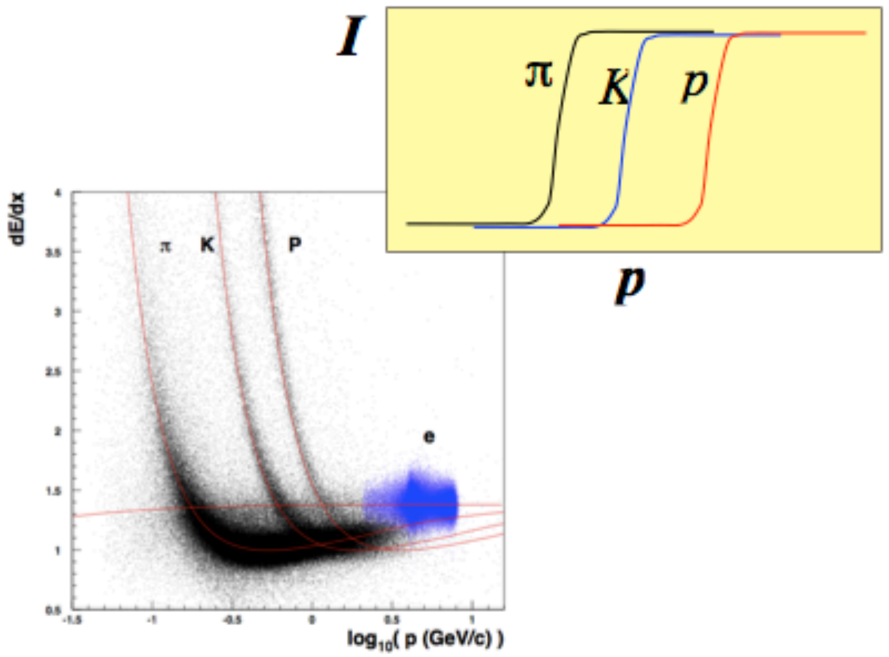


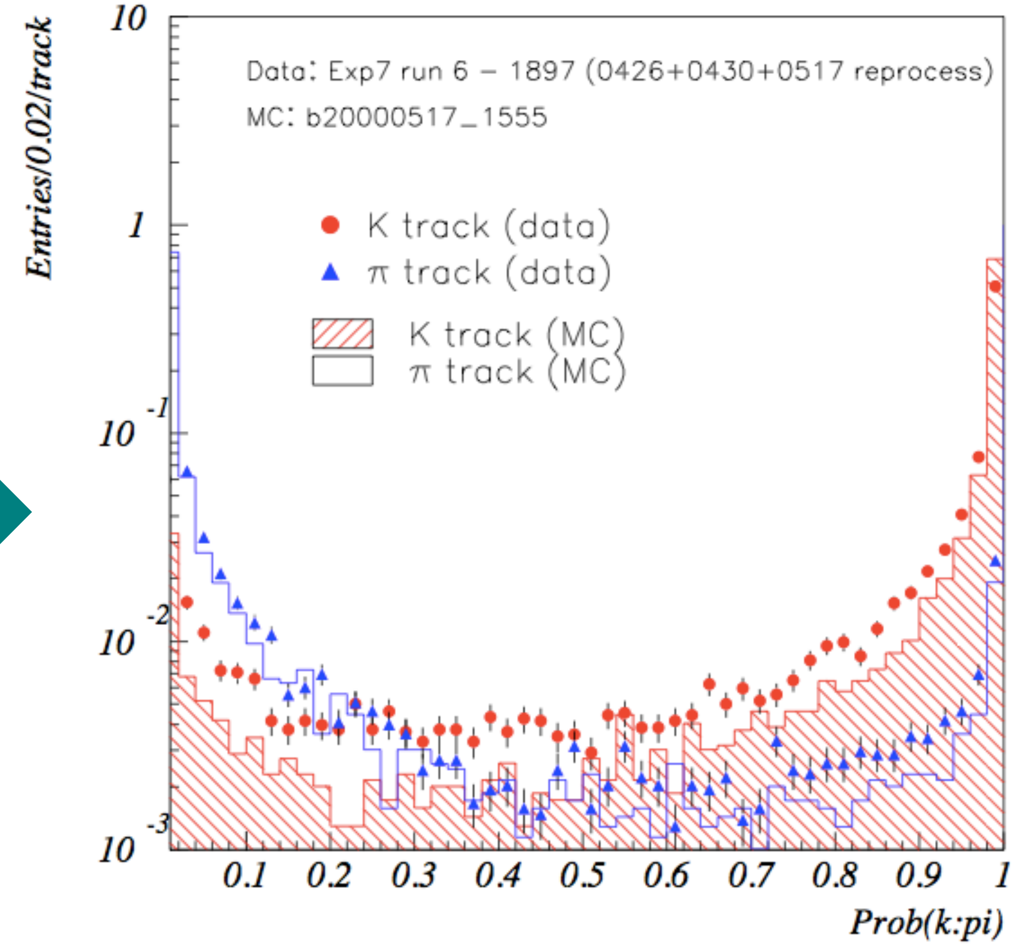
Express the following observables in Type-I & Type-II errors. ***What are H_0 & H_1, for each case?***

- $f_{\pi^+\to K^+}$ = 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^+

an application of Neyman-Pearson Lemma







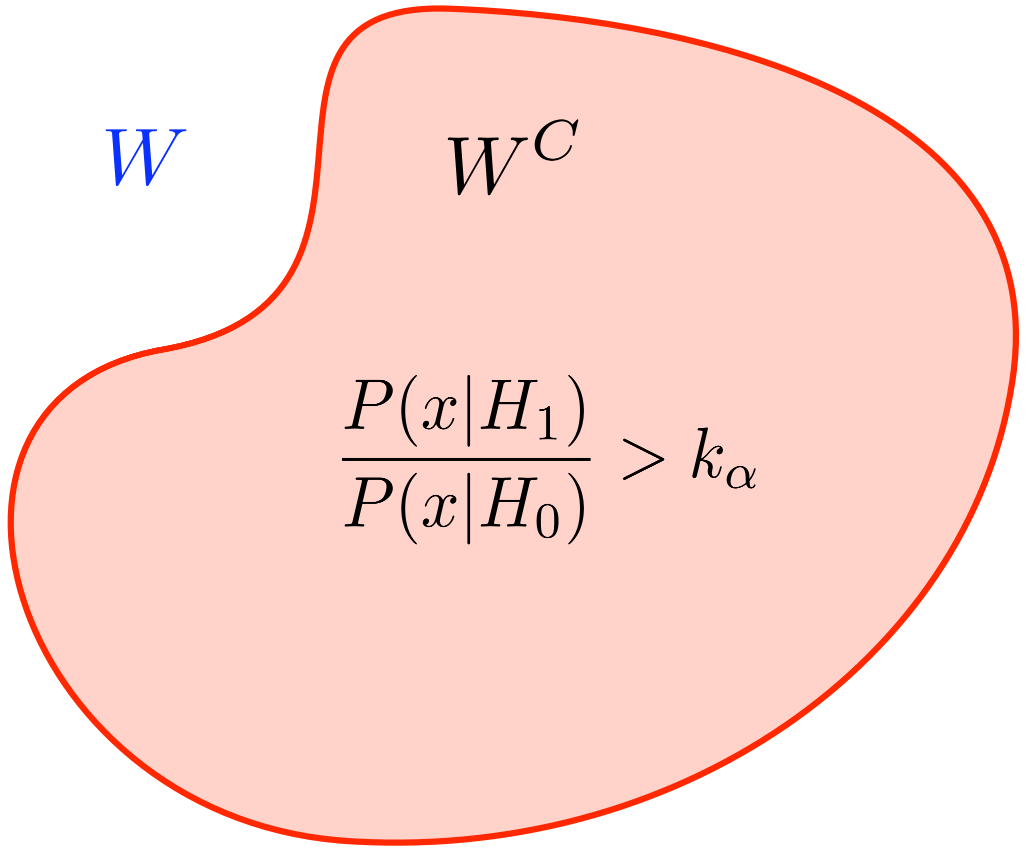
combined likelihood

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

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 $0 \leq R_{K/\pi} \leq 1$

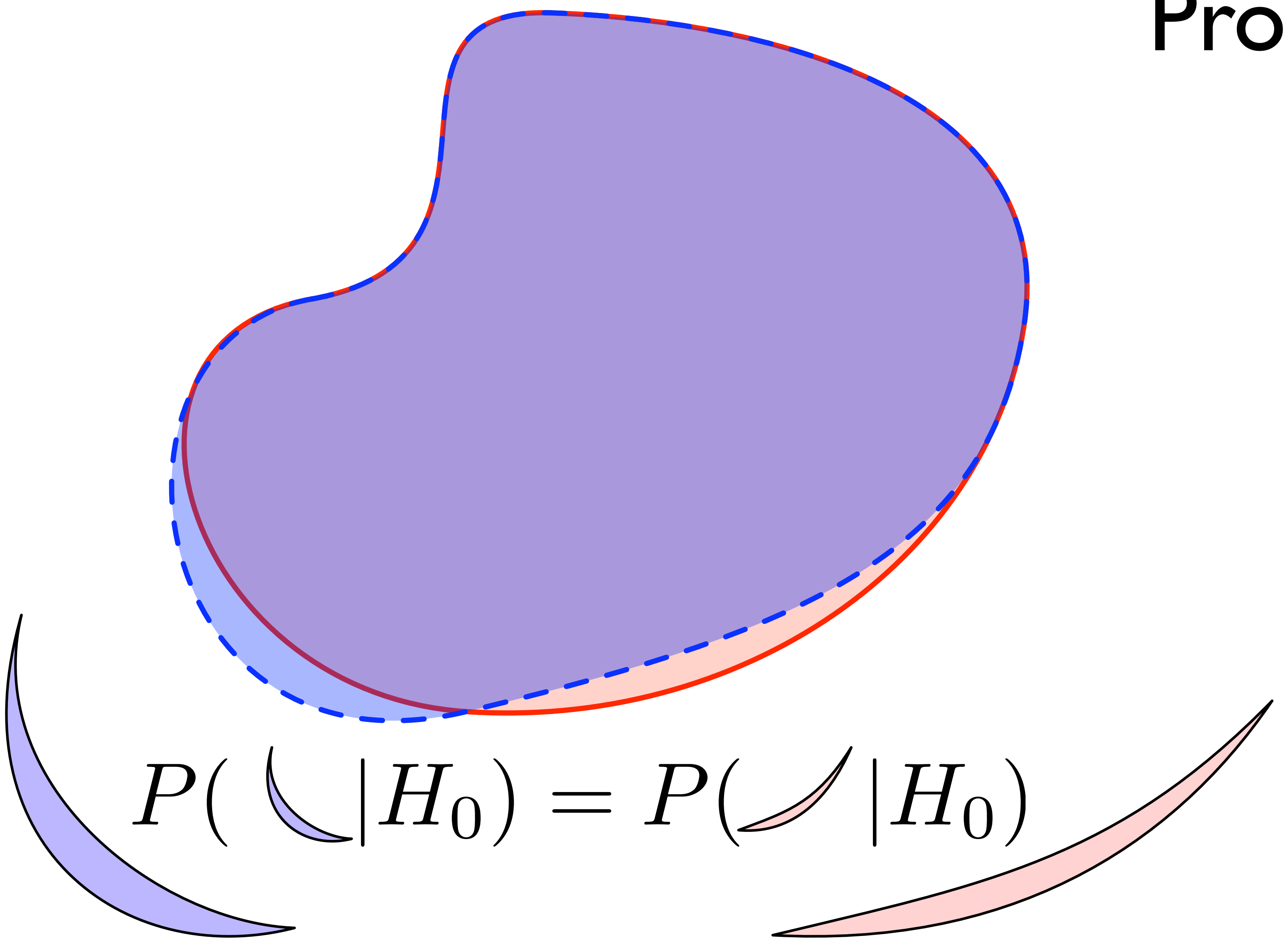
Proof (*graphical*)



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

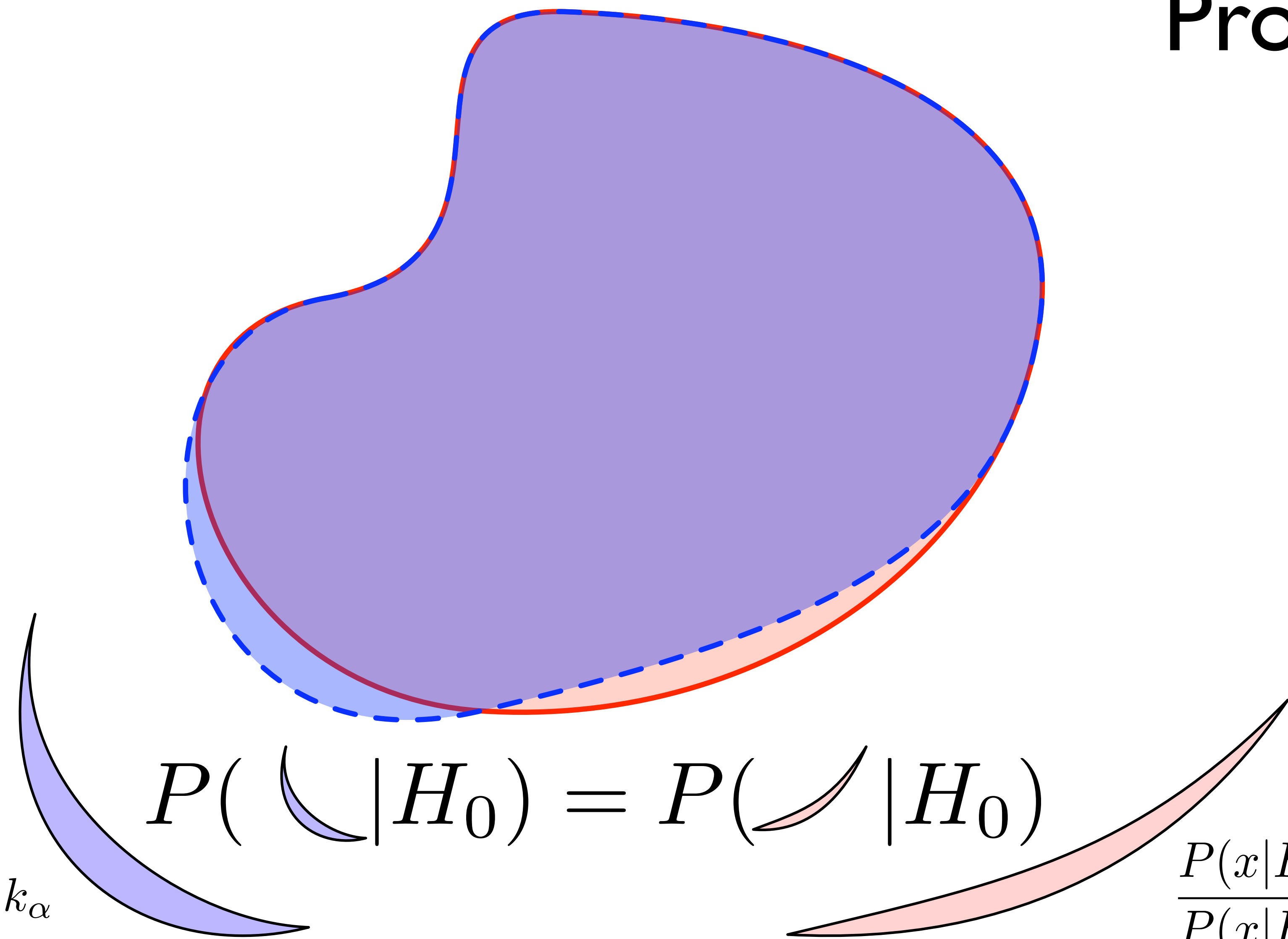
Kyle Cranmer (NYU) CERN School HEP, Romania, Sept. 2011

Proof (*graphical*)



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

Proof (*graphical*)



$$P(\ \ |H_0) = P(\ \ |H_0)$$

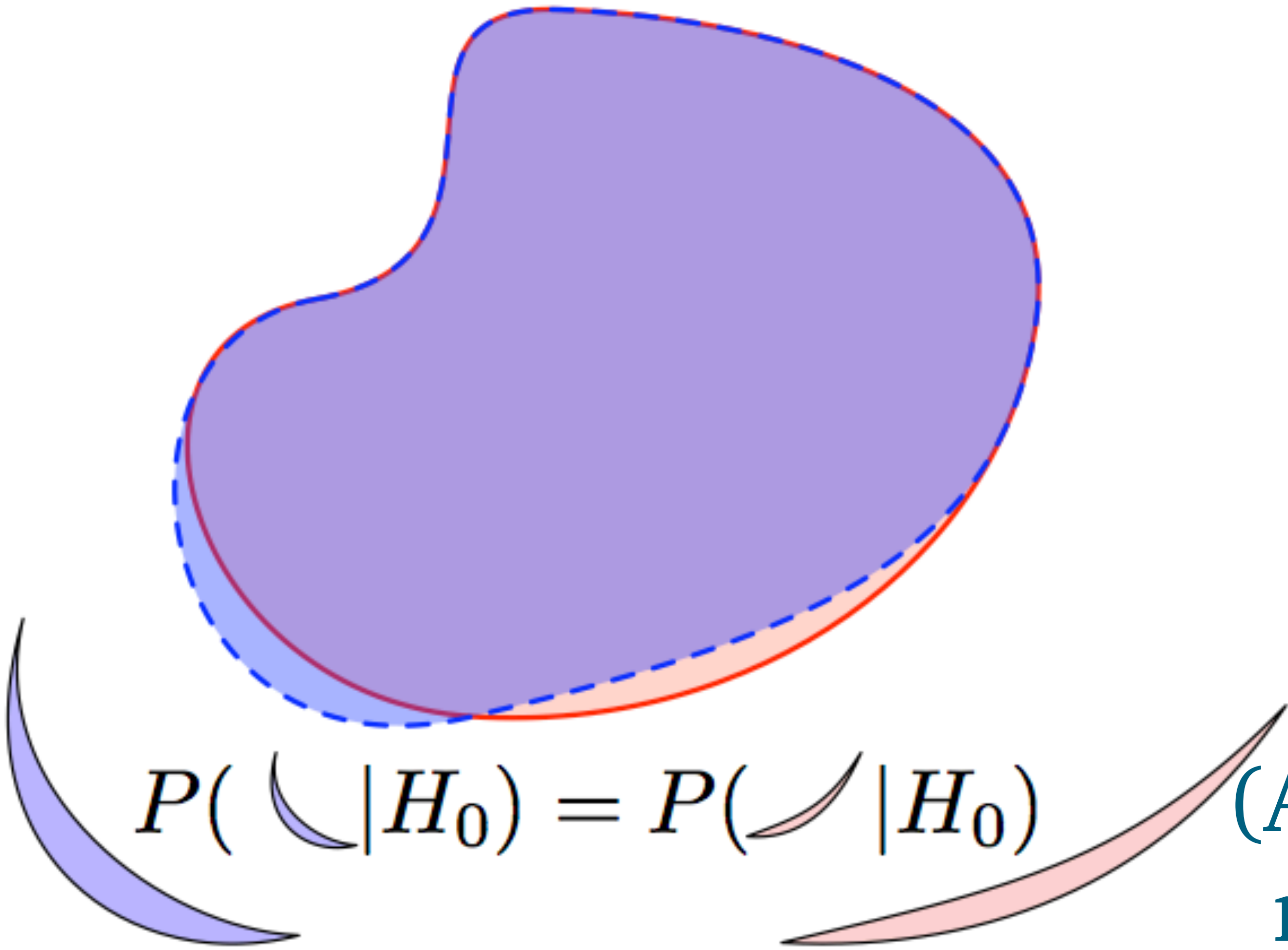
$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \qquad \frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

$$P(\ \ |H_1) < P(\ \ |H_0)k_\alpha \qquad P(\ \ |H_1) > P(\ \ |H_0)k_\alpha$$

$$\boxed{P(\ \ |H_1) < P(\ \ |H_1)}$$

Kyle Cranmer (NYU) CERN School HEP, Romania, Sept. 2011

(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.?

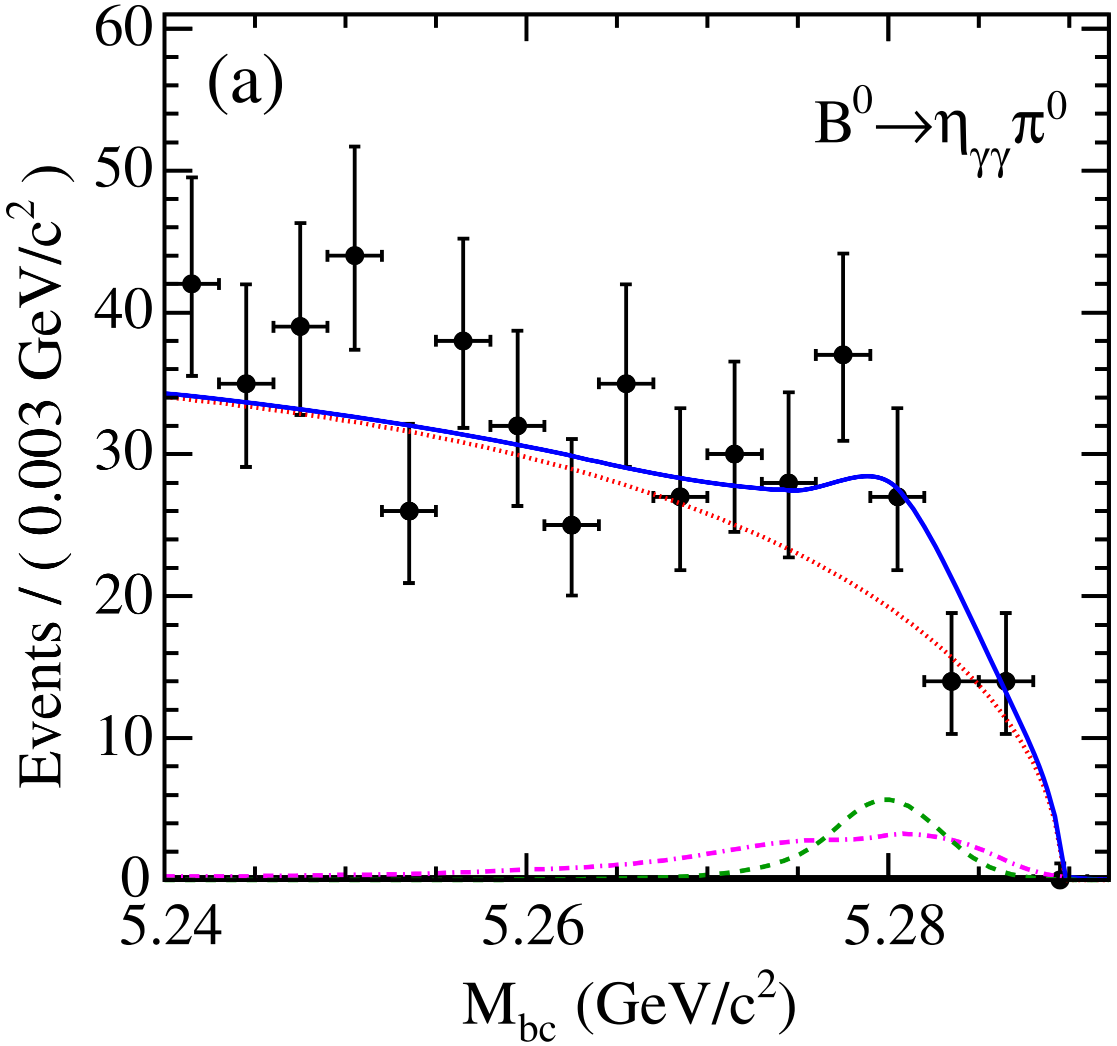


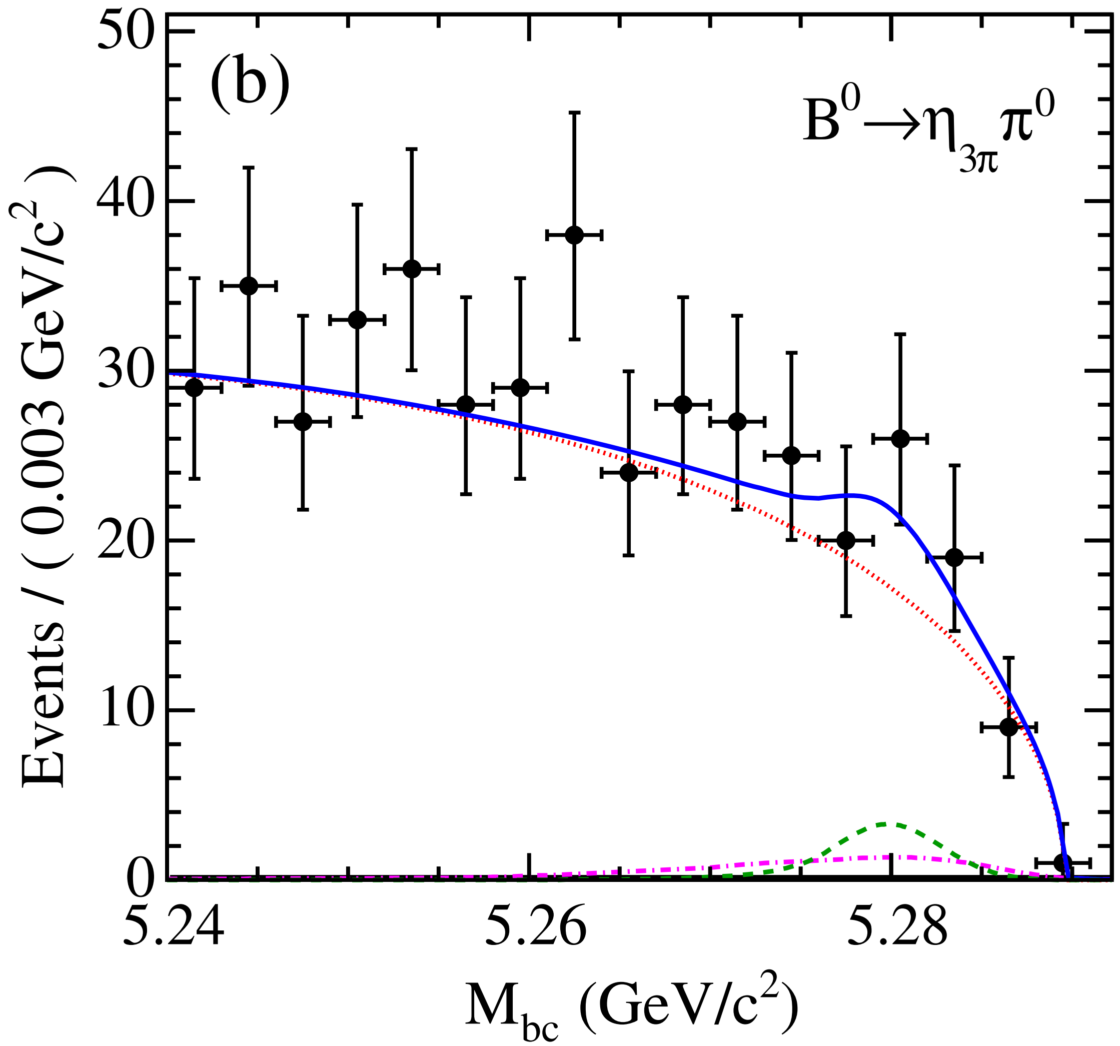
(Ans.) The modeling of P(**x**|H) may not be perfect, if the correlations are not taken properly into account. This will become more serious for higher dimensions of **x**.

Significance of signal

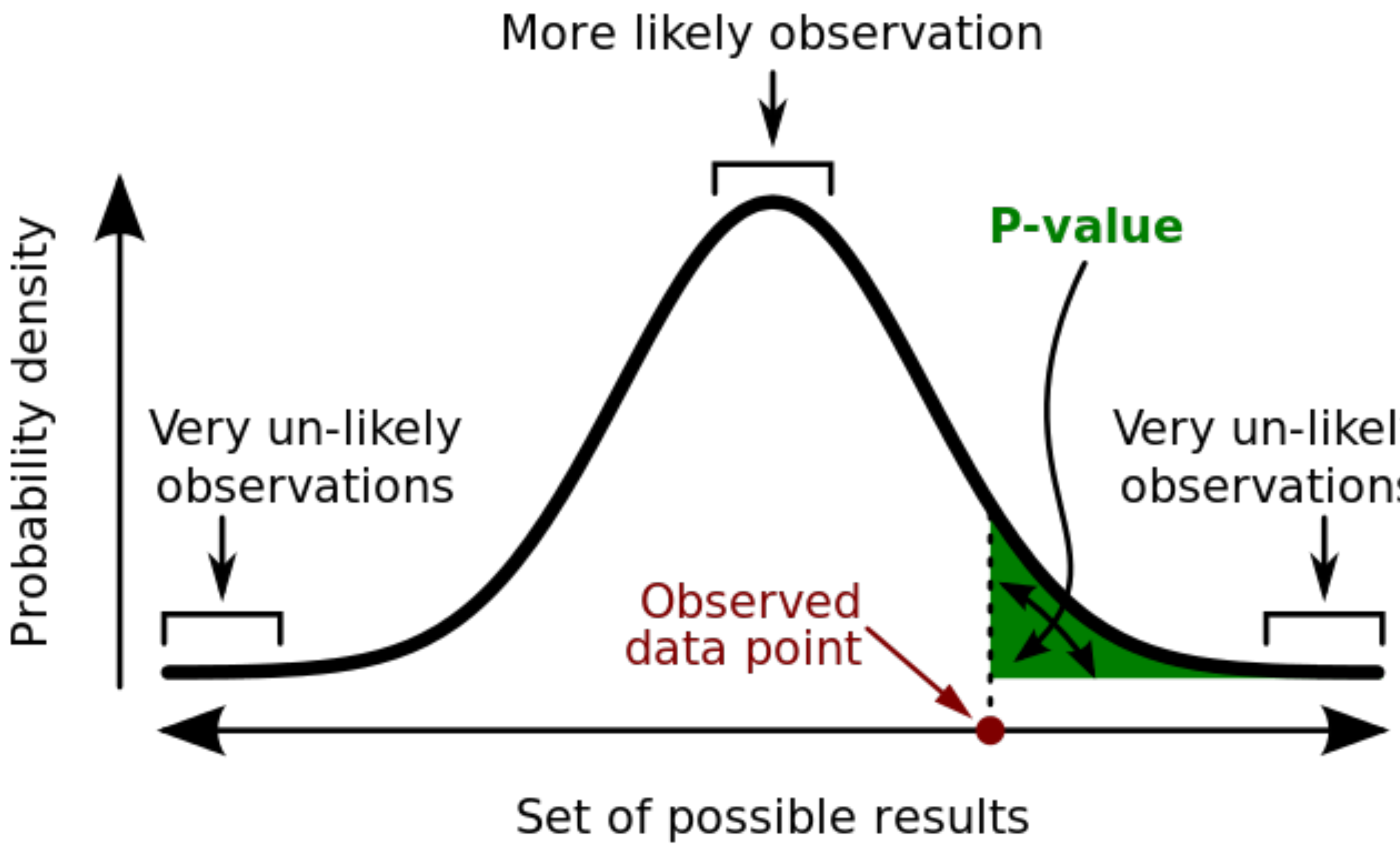
PRD 92, 011101 (2015)

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

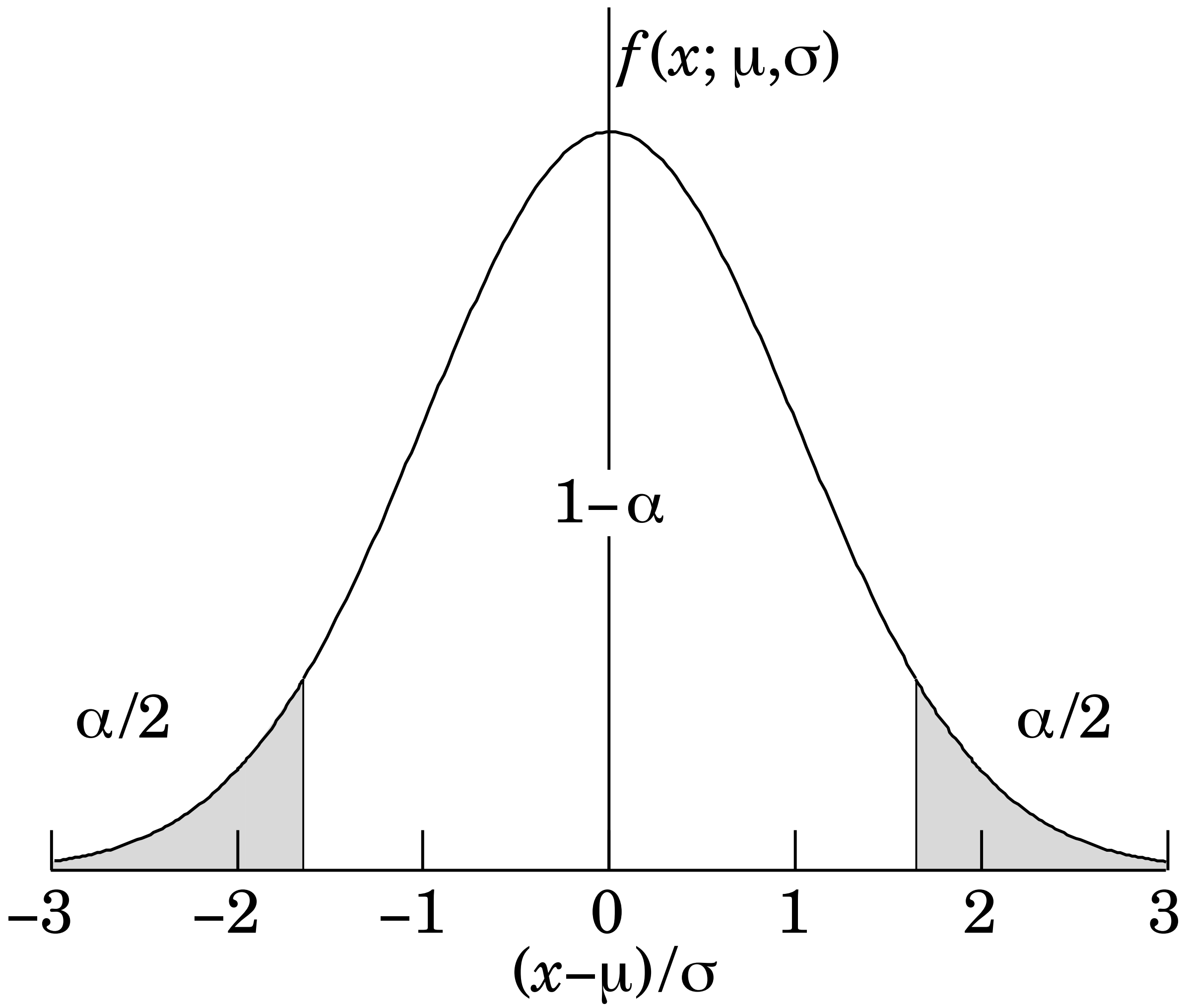
≠ the probability that *H* is true

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

- (Note) The *p*-value, under null hypothesis, is uniformly distributed.

Remember?

Gaussian (Normal) distribution



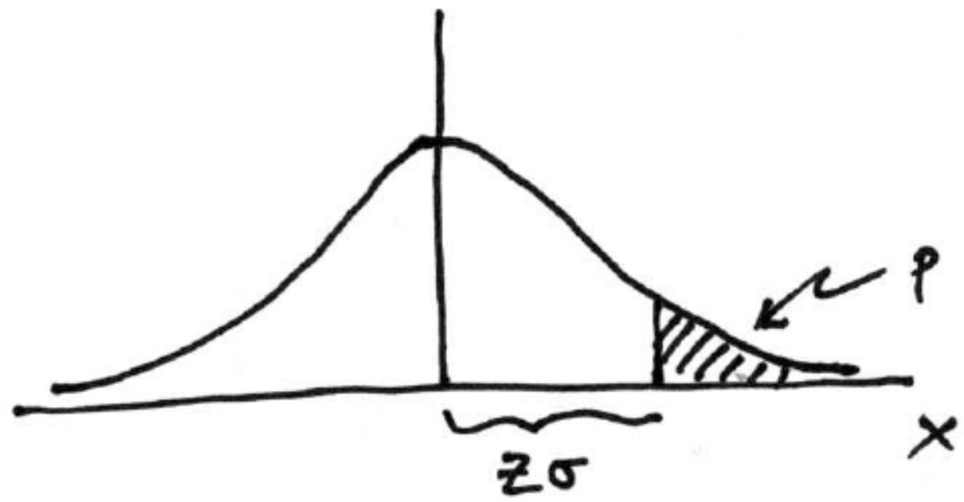
TMath::Prob(δ², 1)

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 (μ = 50% chance for head)

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

binomial probability for n heads in N toss

Critical region w = data space with values equal or lesser compatibility with H in comparison to n = 17

w = {n = 17, 18, 19, 20, 0, 1, 2, 3}

$P(n \in w) = 0.0026 \Leftarrow$ This is the p-value.

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.

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

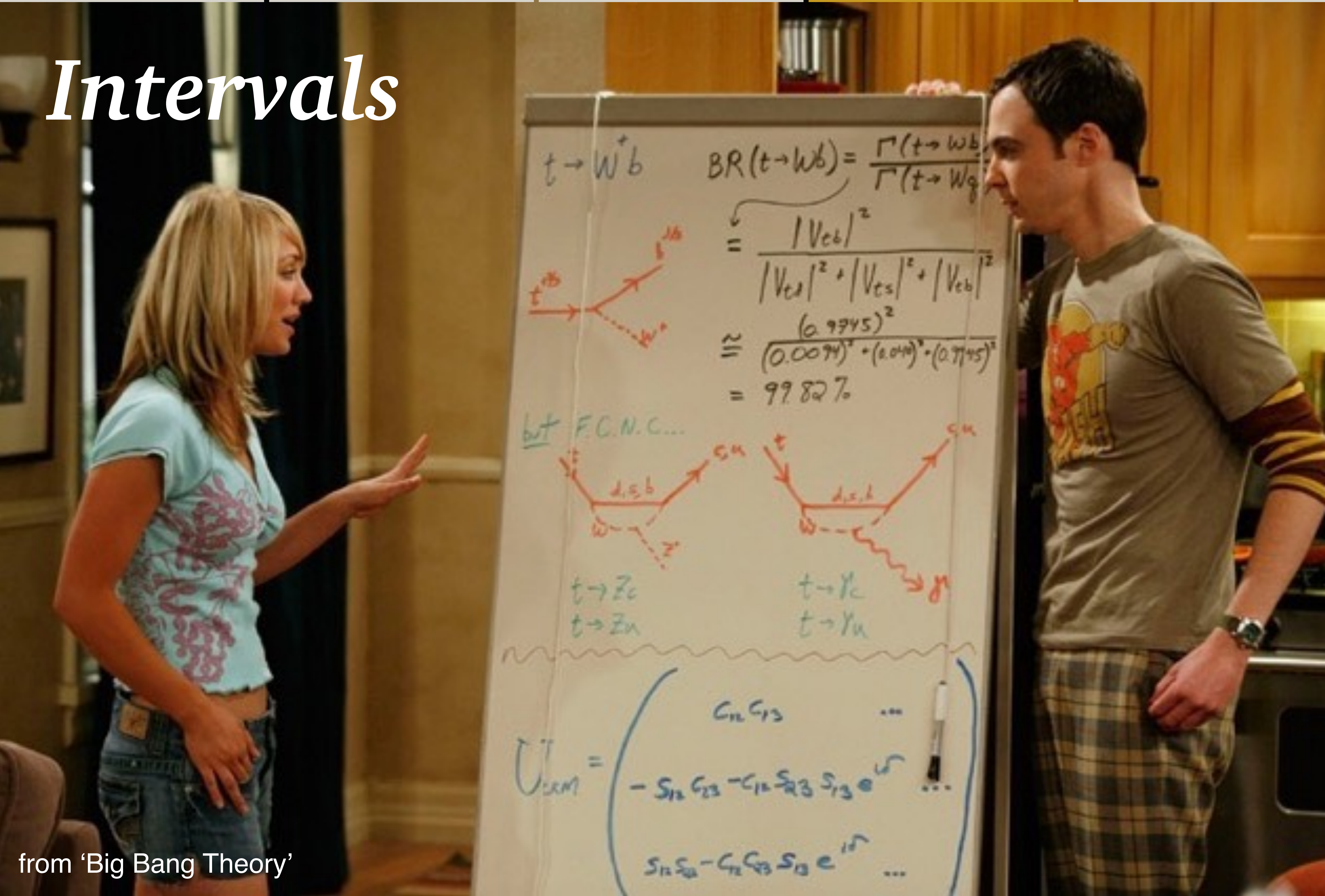
Suppose $b = 0.5$ (assume precise), and we observe $n_{\text{obs}} = 5$.
Can we **claim evidence for a signal excess?**
Give ***p*-value** for the null hypothesis $s = 0$.

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

2018 Korean Ladies Curling team (Olympic Silver)

- (observation) All 5 members of the team are of family name ‘Kim’.
- (fact) According to census, ~20% of all Koreans have family name ‘Kim’.
- (Hypothesis to test) The coach of the Team Kim (herself a ‘Kim’) has a bias toward players with family name ‘Kim’.

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	$\ell\ell + \not{E}_T + \geq$ 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	$\ell + \not{E}_T + \geq$ 3jets ($\geq$
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.

Frequentist “confidence intervals”

- **on repeated measurements**

  Remember frequentist approach is always about repeated measurements under identical conditions!

- **“confidence interval”**

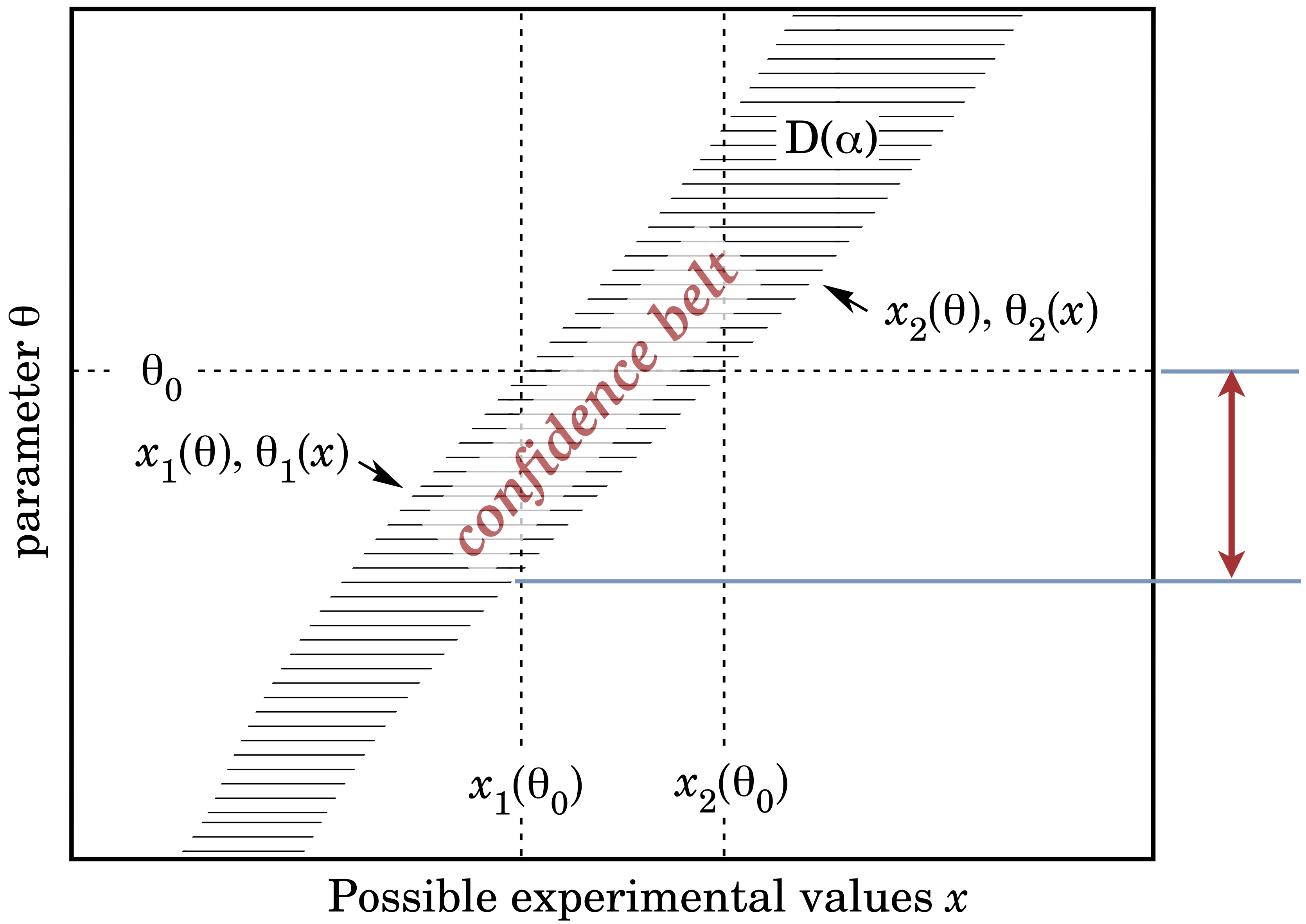
  = intervals constructed to include the true value of the parameter with a probability ≥ (*a specified value*)

Frequentist "confidence intervals"

- **Consider a pdf $f(x;\theta)$**

$$P(x_1 < x < x_2; \theta) = 1 - \alpha = \int_{x_1}^{x_2} f(x;\theta)\, dx$$

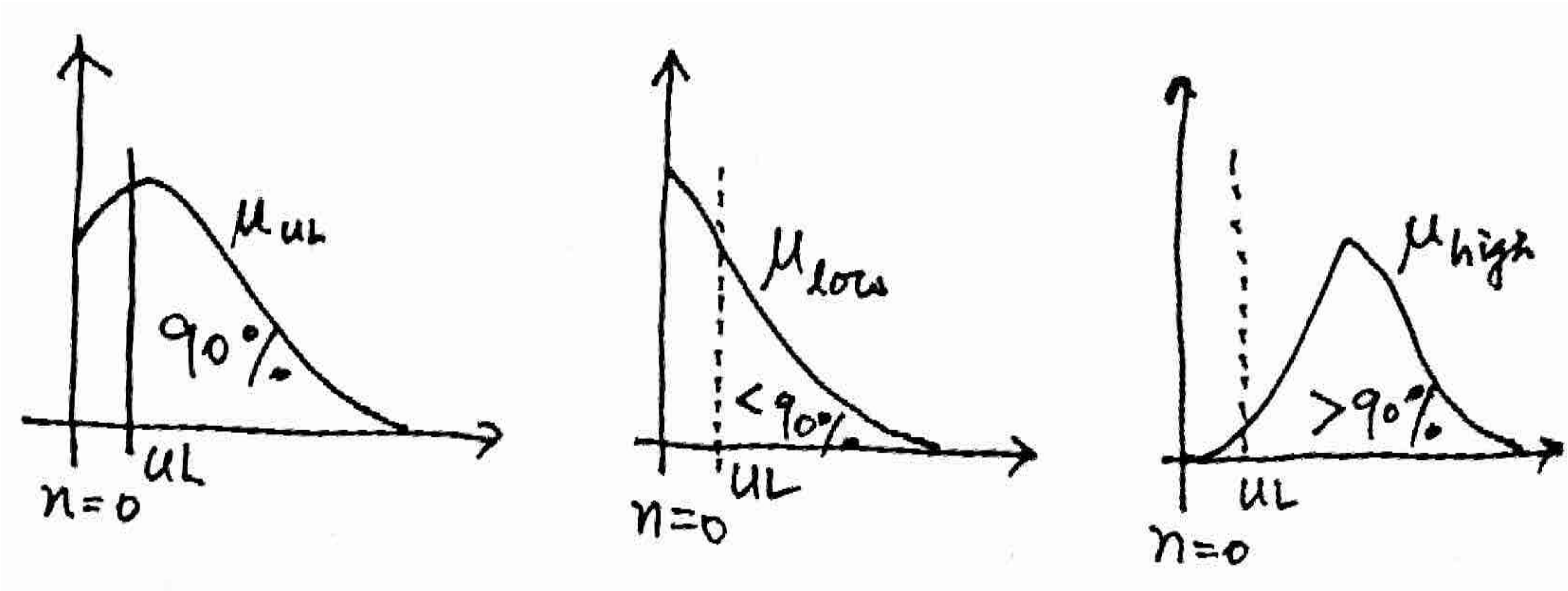
  - x : outcome of an experiment
  - θ : unknown parameter for which we set the interval

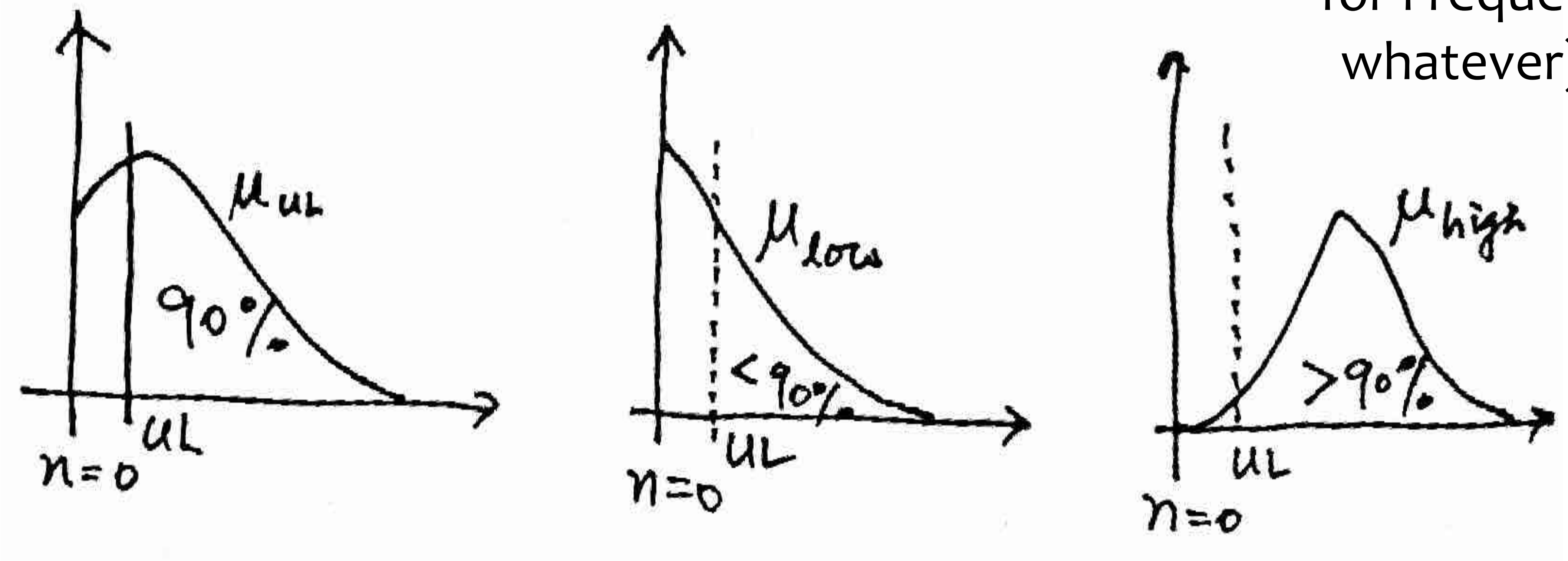


the confidence interval, given a measured outcome of $x_0 = x_1(\theta_0)$

"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	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

Feldman-Cousins interval **Phys. Rev. D57, 3873 (1998)**

"unified approach"

PHYSICAL REVIEW D VOLUME 57, NUMBER 7 1 APRIL 1998

Unified approach to the classical statistical analysis of small signals

Gary J. Feldman*
Department of Physics, Harvard University, Cambridge, Massachusetts 02138

Robert D. Cousins†
Department of Physics and Astronomy, University of California, Los Angeles, California 90095
(Received 21 November 1997; published 6 March 1998)

We give a classical confidence belt construction which unifies the treatment of upper confidence limits for null results and two-sided confidence intervals for non-null results. The unified treatment solves a problem (apparently not previously recognized) that the choice of upper limit or two-sided intervals leads to intervals which are not confidence intervals if the choice is based on the data. We apply the construction to two related problems which have recently been a battleground between classical and Bayesian statistics: Poisson processes with background and Gaussian errors with a bounded physical region. In contrast with the usual classical construction for upper limits, our construction avoids unphysical confidence intervals. In contrast with some popular Bayesian intervals, our intervals eliminate conservatism (frequentist coverage greater than the stated confidence) in the Gaussian case and reduce it to a level dictated by discreteness in the Poisson case. We generalize the method in order to apply it to analysis of experiments searching for neutrino oscillations. We show that this technique both gives correct coverage and is powerful, while other classical techniques that have been used by neutrino oscillation search experiments fail one or both of these criteria.
[S0556-2821(98)00109-X]

PACS number(s): 06.20.Dk, 14.60.Pq

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

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

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

Then,

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

$F_{\chi^2}^{-1}$: inverse of the CDF

➔ $s_{\rm up} = \frac{1}{2}F_{\chi^2}^{-1}[1-\alpha;2(n+1)]-b$

(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 $P(n \le n_{\text{obs}}; s, b)$. Therefore, the upper limit s_{up} at $\text{CL} = 1 - \alpha$ is obtained from

$$\alpha = P(n \le n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F_{\chi^2}^{-1}(0.95; 2(5+1)) - 4.5 = 6.0$$

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. pdg.lbl.gov/2018/reviews/rpp2018-rev-statistics.pdf

Parameter Estimation

Basics of parameter estimation

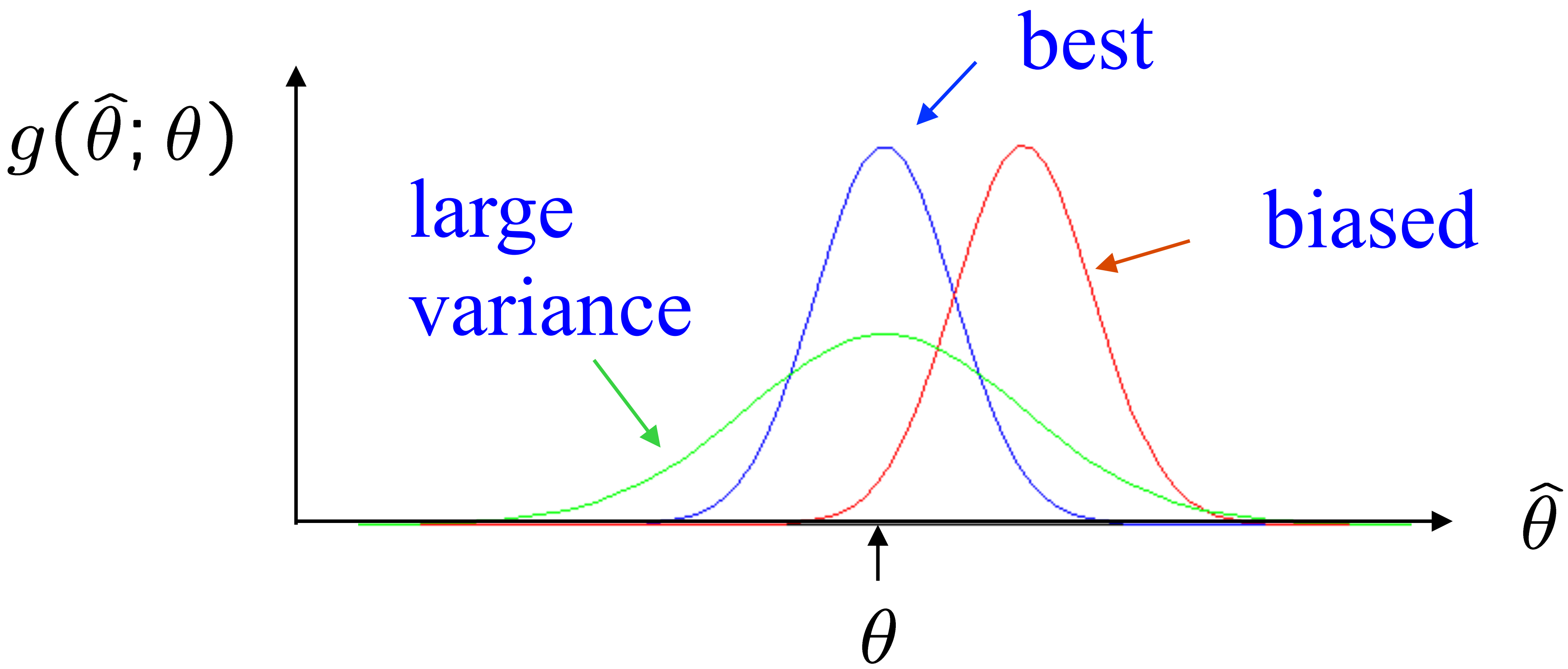
- The parameters of a PDF are constants characterizing its shape, e.g.

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

  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**.

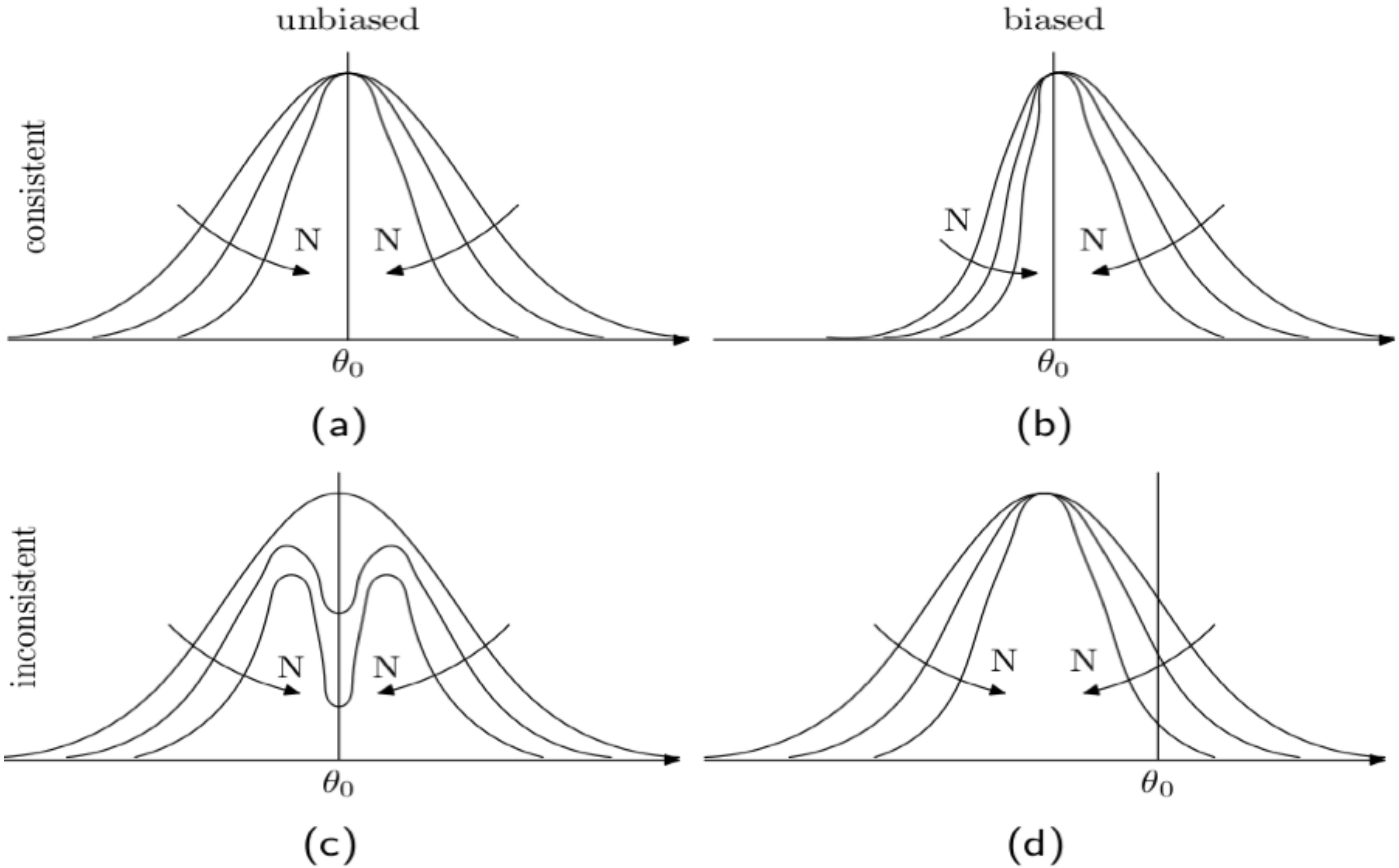
Properties of estimators

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



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

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

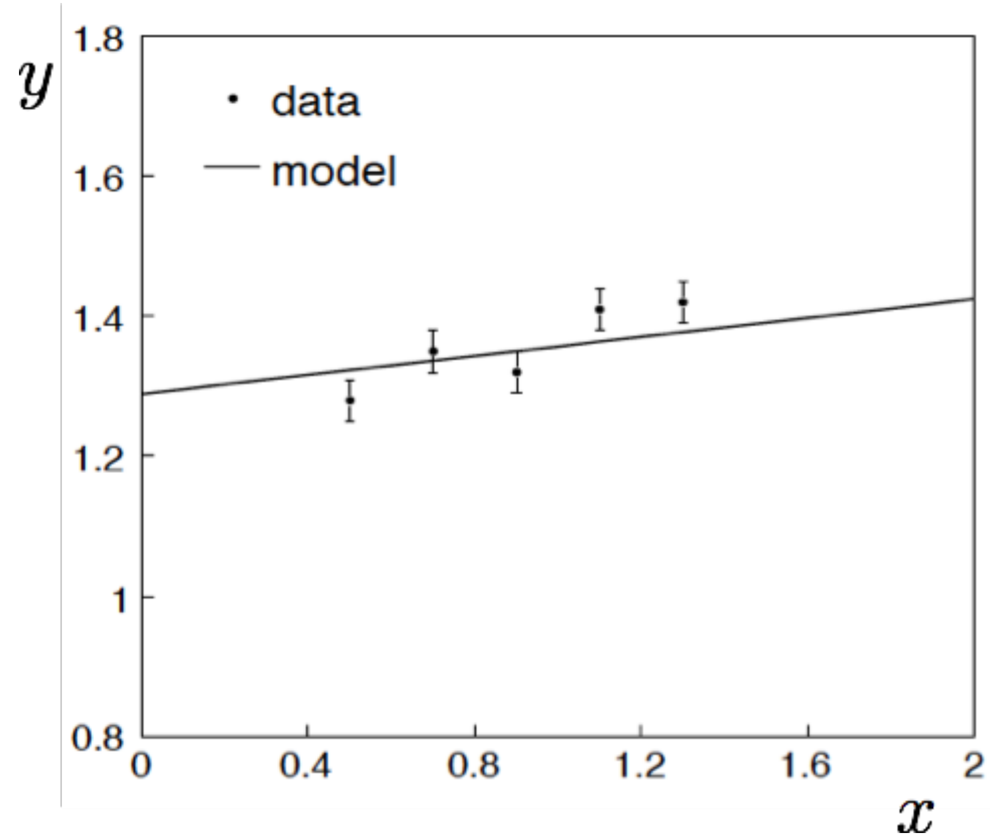
- In this case, the likelihood function is

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

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

the Wilk’s theorem

- *We will encounter it later when we discuss the “likelihood ratio” …*

THE LARGE-SAMPLE DISTRIBUTION OF THE LIKELIHOOD RATIO FOR TESTING COMPOSITE HYPOTHESES[1]

By S. S. Wilks

By applying the principle of maximum likelihood, J. Neyman and E. S. Pearson[2] have suggested a method for obtaining functions of observations for testing what are called *composite statistical hypotheses*, or simply *composite*

…

[1] Presented to the American Mathmatical Society, March 26, 1937.

…

We can summarize in the

Theorem: If a population with a variate x is distributed according to the probability function $f(x, \theta_1, \theta_2 \cdots \theta_h)$, such that optimum estimates $\bar{\theta}_i$ of the θ_i exist which are distributed in large samples according to (3), *then when the hypothesis H is true that $\theta_i = \theta_{0i}$, $i = m + 1, m + 2, \cdots h$, the distribution of $-2 \log \lambda$, where λ is given by* (2) *is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h - m$ degrees of freedom.*

the Wilk's theorem

http://wwwusers.ts.infn.it/~milotti/Didattica/StatisticaAvanzata/Cowan_2013.pdf

Suppose we model the data $\vec{X}$ with a likelihood $L(\vec{\mu})$ that depends on a set of N parameters $\vec{\mu} = (\mu_1, \cdots, \mu_N)$. (For simplicity, let's just consider a single parameter μ.)

- Define the statistic $t_\mu = -2\ln[L(\mu)/L(\hat{\mu})]$, where $\hat{\mu}$ is the ML estimator.
- The value of t_μ is a measure of how well the hypothesized parameter μ stand in agreement with the observed data.
- Larger values of t_μ indicate increasing incompatibility between the data and the hypothesized μ.
- According to Wilk's theorem, if the parameter value μ is true, then in the asymptotic limit of a large data sample, the PDF of t_μ is a χ^2 distribution for N d.o.f.

$$f(t_\mu|\mu) \sim \chi^2_N$$

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:

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

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

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

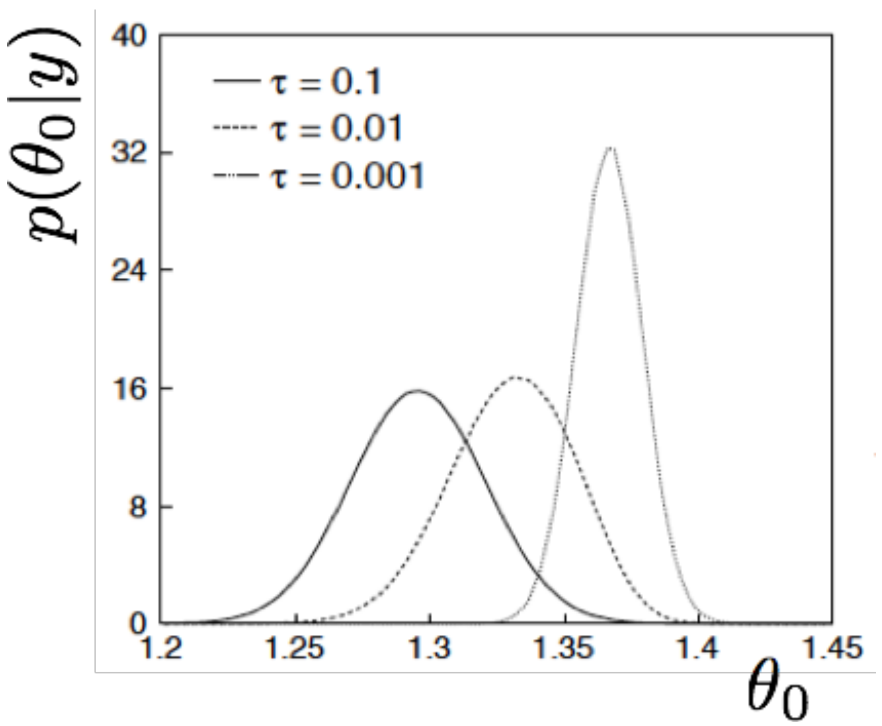
- Then, $p(\theta_0|\vec{x}) = \int p(\theta_0, \theta_1|\vec{x}) \, d\theta_1$

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.