Statistics for Particle Physics
Lecture 1

Terascale Statistics School

https://indico.desy.de/event/43398/

DESY, Hamburg
2-5 April 2024

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

→ Tuesday 11:05	Introduction
	Probability
	Hypothesis tests, parameter estimation
Wednesday 9:15	Confidence limits
	Systematic uncertainties
	General analysis, asymptotics
Thursday 16:00	“Errors on errors”

More resources in the University of London course:

https://www.pp.rhul.ac.uk/~cowan/stat_course.html

Theory ↔ Statistics ↔ Experiment

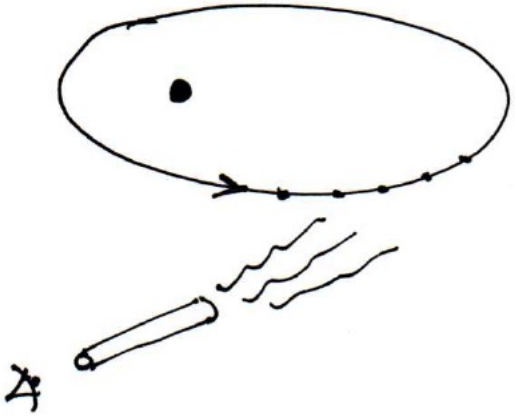
Theory (model, hypothesis):

$$F = -G \frac{m_1 m_2}{r^2} , \ldots$$

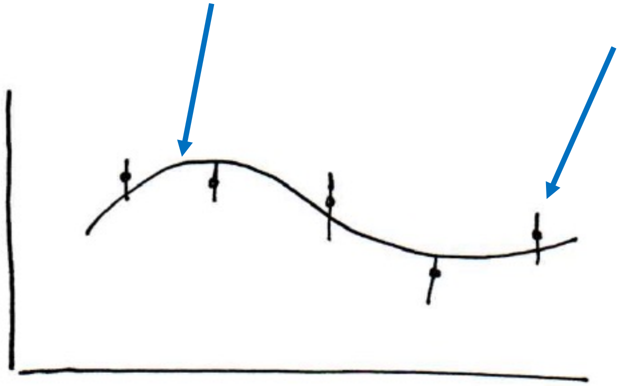
+ response of measurement apparatus

= model prediction

Experiment (observation):



data



Uncertainty enters on many levels

→ quantify with probability

A quick review of probability

Frequentist (A = outcome of repeatable observation)

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

Subjective (A = hypothesis)

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

Conditional probability:

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

E.g. rolling a die,
outcome n = 1,2,...,6:

$$P(n \leq 3 | n \text{ even}) = \frac{P((n\leq 3)\cap n \text{ even})}{P(n \text{ even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

A and B are independent iff:

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

I.e. if A, B independent, then

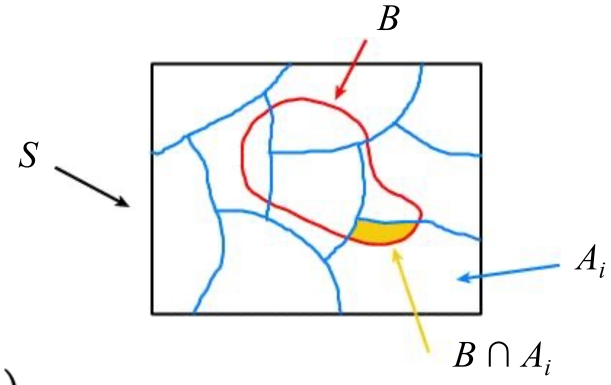
$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes' theorem

Use definition of conditional probability and $P(A \cap B) = P(B \cap A)$

$$\rightarrow \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ (Bayes' theorem)

If set of all outcomes $S = \cup_i A_i$ with A_i disjoint, then law of total probability for $P(B)$ says



$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

so that Bayes' theorem becomes $$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Bayes' theorem holds regardless of how probability is interpreted (frequency, degree of belief...).

Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand: $\boldsymbol{x}$).

Probability = limiting frequency

Probabilities such as

P (string theory is true),
P ($0.117 < \alpha_s < 0.119$),
P (Biden wins in 2024),

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

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

Preferred theories (models, hypotheses, ...) are those that predict a high probability for data "like" the data observed.

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability for hypotheses:

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

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No general prescription for priors (subjective!)

Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers $\boldsymbol{x}$.

A (simple) hypothesis is a rule that assigns a probability to each possible data outcome:

$P(\mathbf{x}|H)$ = the likelihood of H

Often we deal with a family of hypotheses labeled by one or more undetermined parameters (a composite hypothesis):

$P(\mathbf{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta})$ = the “likelihood function”

Note:

1) For the likelihood we treat the data $\boldsymbol{x}$ as fixed.
2) The likelihood function $L(\boldsymbol{\theta})$ is not a pdf for $\boldsymbol{\theta}$.

Frequentist hypothesis tests

Suppose a measurement produces data $\boldsymbol{x}$; consider a hypothesis H_0 we want to test and alternative H_1

H_0, H_1 specify probability for $\boldsymbol{x}$: $P(\boldsymbol{x}|H_0)$, $P(\boldsymbol{x}|H_1)$

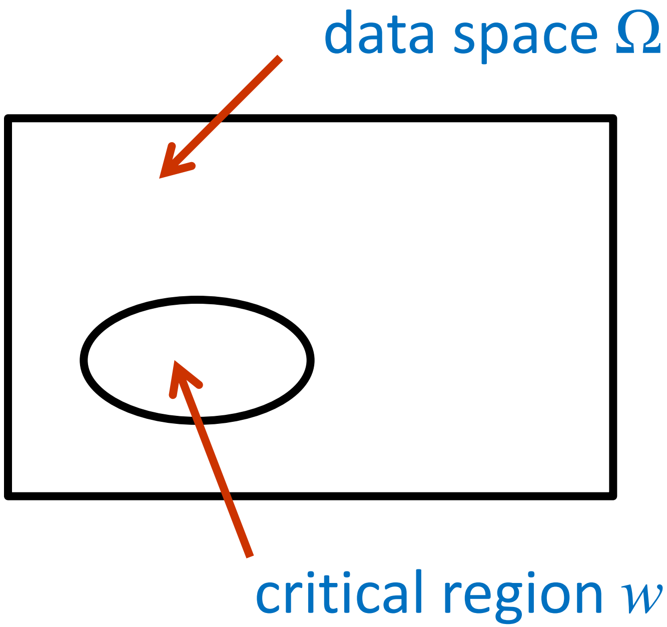
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

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

Need inequality if data are discrete.

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

If $\boldsymbol{x}$ is observed in the critical region, reject H_0.

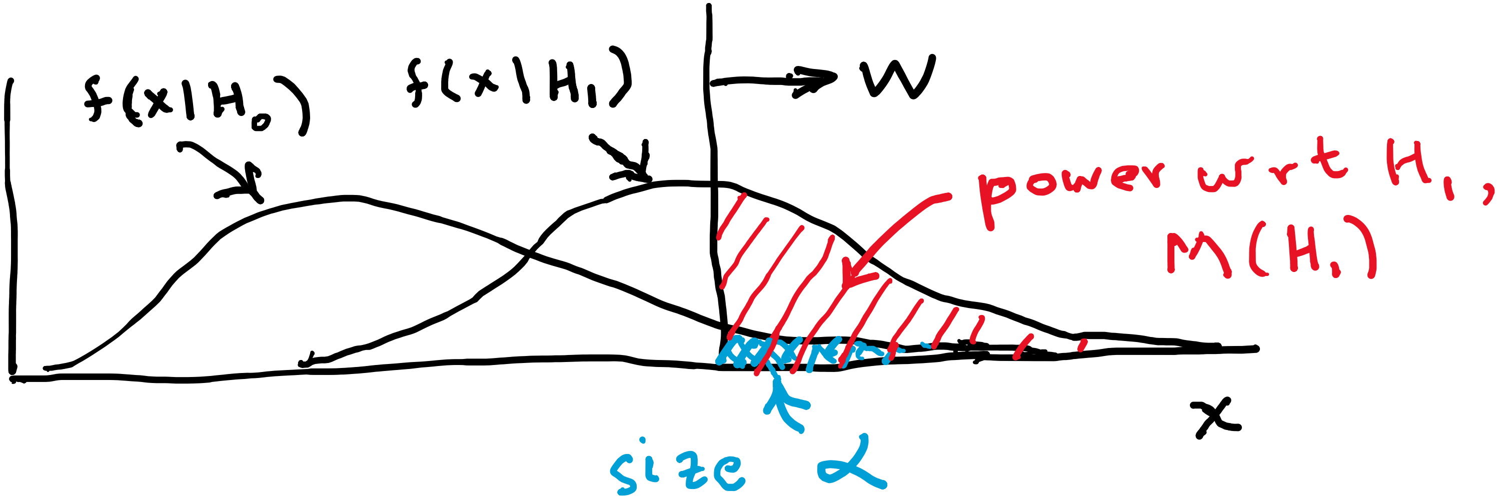


Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same size α.

Use the alternative hypothesis H_1 to motivate where to place the critical region.

Roughly speaking, place the critical region where there is a low probability (α) to be found if H_0 is true, but high if H_1 is true:



Classification viewed as a statistical test

Suppose events come in two possible types:

s (signal) and b (background)

For each event, test hypothesis that it is background, i.e., H_0 = b.

Carry out test on many events, each is either of type s or b, i.e., here the hypothesis is the “true class label”, which varies randomly from event to event, so we can assign to it a frequentist probability.

Select events for which where H_0 is rejected as “candidate events of type s”. Equivalent Particle Physics terminology:

background efficiency $$\varepsilon_{\rm b} = \int_W f(\mathbf{x}|H_0)\, d\mathbf{x} = \alpha$$

signal efficiency $$\varepsilon_{\rm s} = \int_W f(\mathbf{x}|H_1)\, d\mathbf{x} = 1 - \beta = \text{power}$$

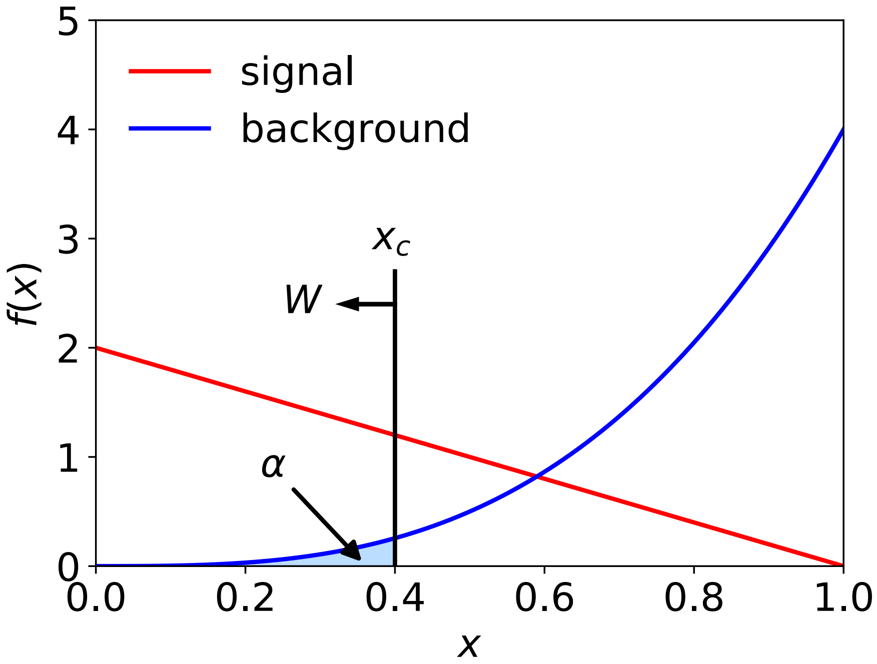
Example of a test for classification

Suppose we can measure for each event a quantity x, where

$$f(x|s) = 2(1-x)$$

$$f(x|b) = 4x^3$$

with $0 \leq x \leq 1$.



For each event in a mixture of signal (s) and background (b) test

H_0 : event is of type b

using a critical region W of the form: $W = \{x : x \leq x_c\}$, where x_c is a constant that we choose to give a test with the desired size α.

Classification example (2)

Suppose we want $\alpha = 10^{-4}$. Require:

$$\alpha = P(x \leq x_{\rm c}|{\rm b}) = \int_0^{x_{\rm c}} f(x|b)\, dx = \left. \frac{4x^4}{4} \right|_0^{x_{\rm c}} = x_{\rm c}^4$$

and therefore $x_{\rm c} = \alpha^{1/4} = 0.1$

For this test (i.e. this critical region W), the power with respect to the signal hypothesis (s) is

$$M = P(x \leq x_{\rm c}|{\rm s}) = \int_0^{x_{\rm c}} f(x|{\rm s})\, dx = 2x_{\rm c} - x_{\rm c}^2 = 0.19$$

Note: the optimal size and power is a separate question that will depend on goals of the subsequent analysis.

Classification example (3)

Suppose that the prior probabilities for an event to be of type s or b are:

$$\pi_{\mathrm{s}} = 0.001$$
$$\pi_{\mathrm{b}} = 0.999$$

The "purity" of the selected signal sample (events where b hypothesis rejected) is found using Bayes' theorem:

$$P(\mathrm{s}|x \le x_{\mathrm{c}}) = \frac{P(x \le x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}}}{P(x \le x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}} + P(x \le x_{\mathrm{c}}|\mathrm{b})\pi_{\mathrm{b}}}$$

$$= 0.655$$

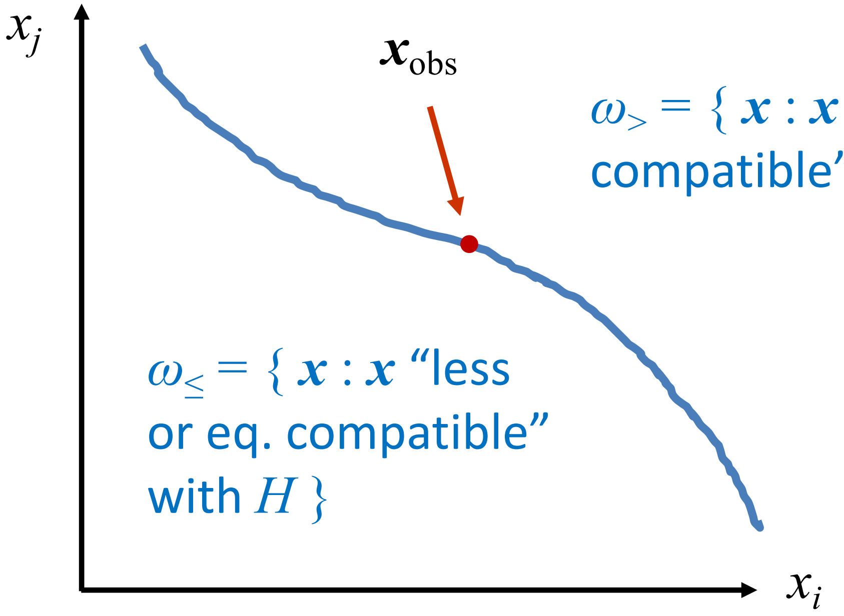
Testing significance / goodness-of-fit

Suppose hypothesis H predicts pdf $f(\boldsymbol{x}|H)$ for a set of observations $\boldsymbol{x} = (x_1,...x_n)$.

We observe a single point in this space: $\boldsymbol{x}_{\rm obs}$.

How can we quantify the level of compatibility between the data and the predictions of H?

Decide what part of the data space represents equal or less compatibility with H than does the point $\boldsymbol{x}_{\rm obs}$. (Not unique!)


x_j
$\boldsymbol{x}_{\rm obs}$
$\omega_> = \{ \boldsymbol{x} : \boldsymbol{x}$ "more compatible" with H $\}$
$\omega_\leq = \{ \boldsymbol{x} : \boldsymbol{x}$ "less or eq. compatible" with H $\}$
x_i

p-values

Express level of compatibility between data and hypothesis (sometimes 'goodness-of-fit') by giving the p-value for H:

$$p = P(\mathbf{x} \in \omega_{\leq}(\mathbf{x}_{\rm obs}) | H)$$

= probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

= probability, under assumption of H, to observe data as discrepant with H as the data we got or more so.

Basic idea: if there is only a very small probability to find data with even worse (or equal) compatibility, then H is "disfavoured by the data".

If the p-value is below a user-defined threshold α (e.g. 0.05) then H is rejected (equivalent to hypothesis test of size α as seen earlier).

p-value of H is not $P(H)$

The p-value of H is not the probability that H is true!

In frequentist statistics we don’t talk about $P(H)$ (unless H represents a repeatable observation).

If we do define $P(H)$, e.g., in Bayesian statistics as a degree of belief, then we need to use Bayes’ theorem to obtain

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

where $\pi(H)$ is the prior probability for H.

For now stick with the frequentist approach;
result is p-value, regrettably easy to misinterpret as $P(H)$.

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

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

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s, e.g.,

test $s = 0$ (rejecting $H_0 \approx$ “discovery of signal process”)

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

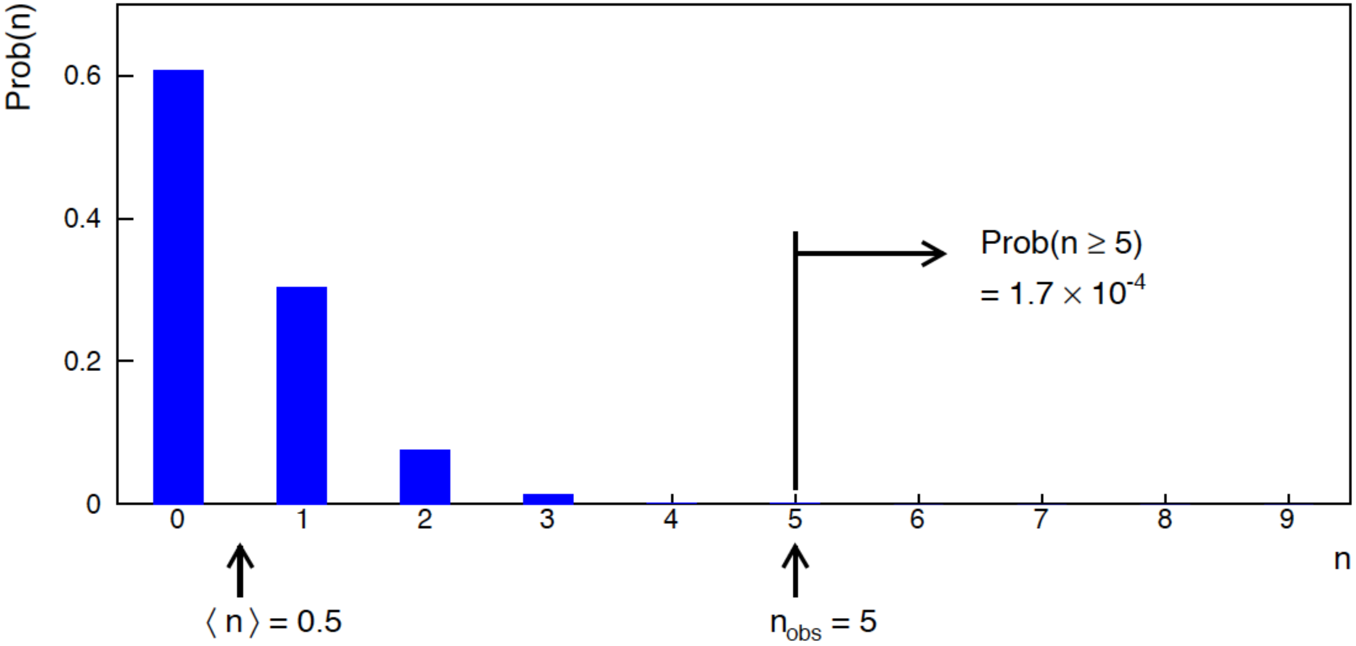
Poisson counting experiment: discovery p-value

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

Should we claim evidence for a new discovery?

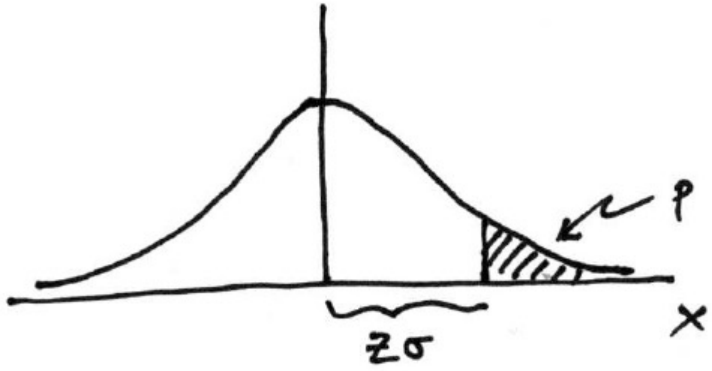
Give p-value for hypothesis $s = 0$, suppose relevant alt. is $s > 0$.

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



Significance from p-value

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



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

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

```
in ROOT:
p = 1 - TMath::Freq(Z)
Z = TMath::NormQuantile(1-p)
```

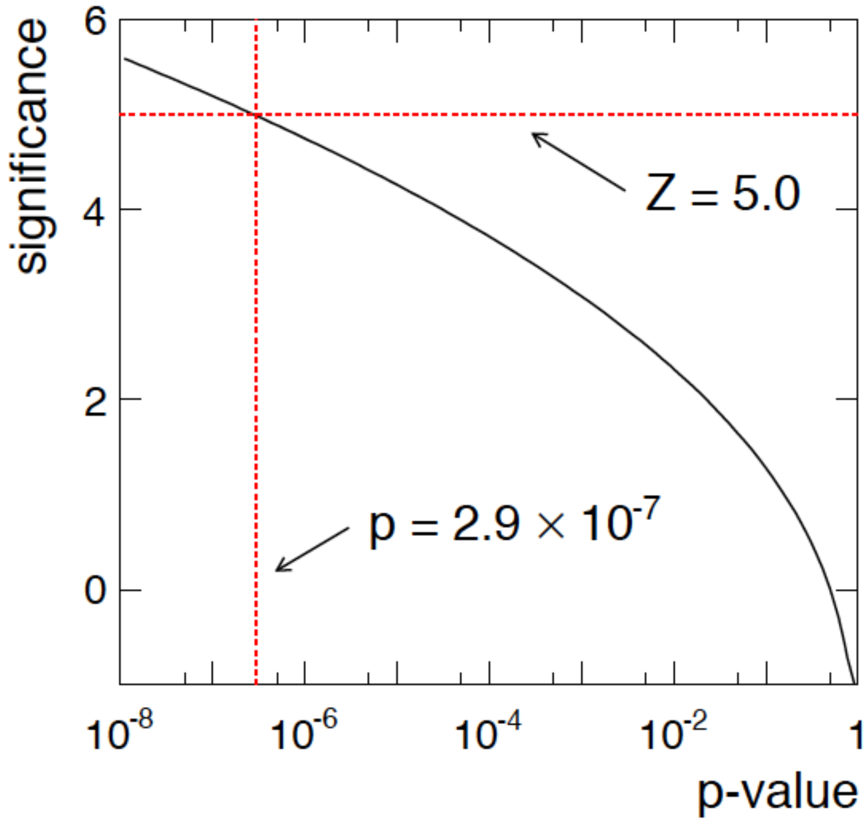
```
in python (scipy.stats):
p = 1 - norm.cdf(Z) = norm.sf(Z)
Z = norm.ppf(1-p)
```

Result Z is a “number of sigmas”. Note this does not mean that the original data was Gaussian distributed.

Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1-p) = 3.6$

Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



In fact this tradition should be revisited: p-value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

Parameter estimation

The parameters of a pdf are any constants that characterize it,

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

r.v. parameter

i.e., θ indexes a set of hypotheses.

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

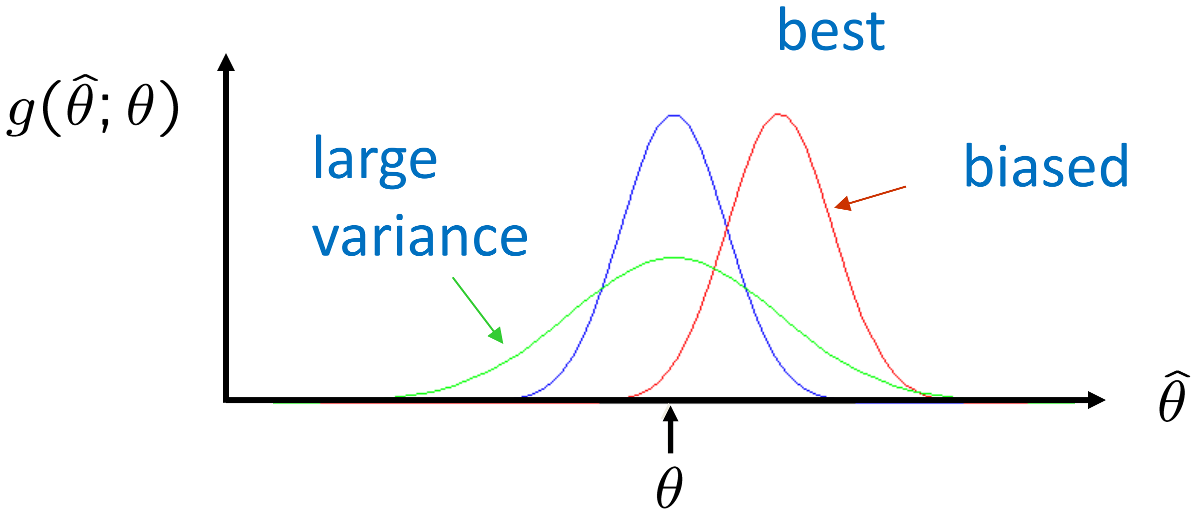
$$\boxed{\widehat{\theta}(\vec{x})}$$

$\leftarrow$ estimator written with a hat

Sometimes we say ‘estimator’ for the function of $x_1, \ldots, x_n$;
‘estimate’ for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



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

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

The likelihood function for i.i.d.* data

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

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

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

In this case the likelihood function is

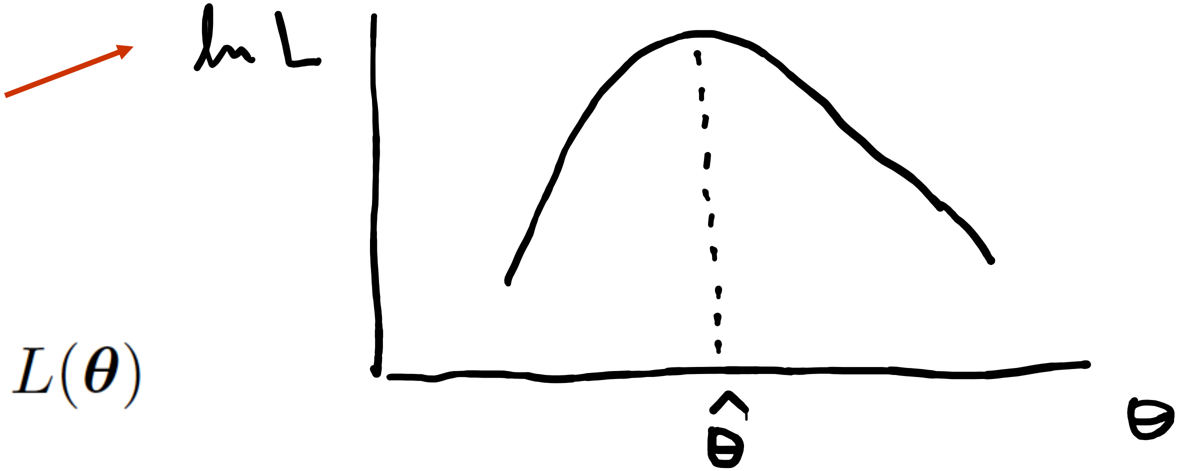
$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i;\vec{\theta})$$

(x_i constant)

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$



$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$$

Could have multiple maxima (take highest).

MLEs not guaranteed to have any ‘optimal’ properties, (but in practice they’re very good).

MLE example: parameter of exponential pdf

Consider exponential pdf, $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

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

MLE example: parameter of exponential pdf (2)

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

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

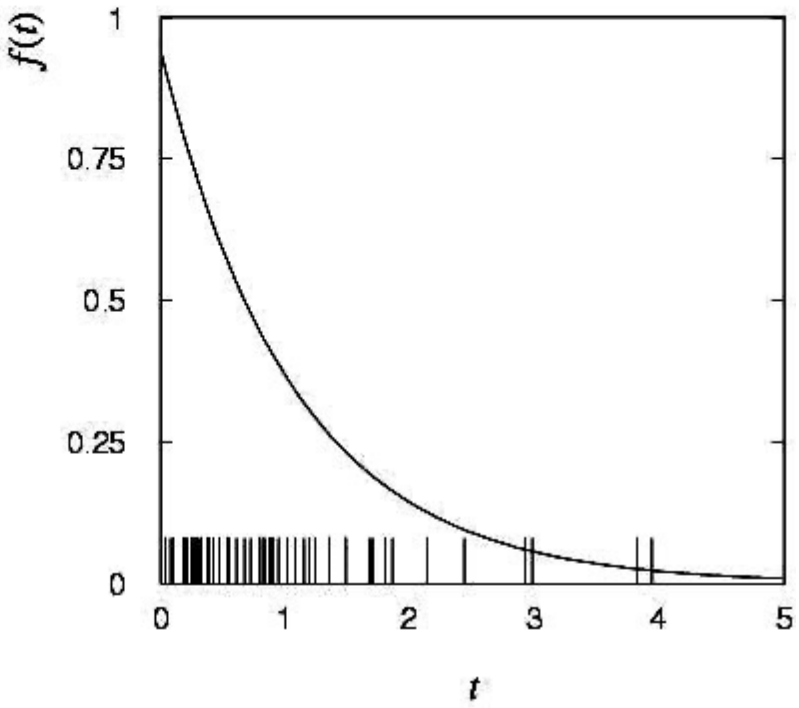
Monte Carlo test:

generate 50 values

using $\tau = 1$:

We find the ML estimate:

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



MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t - \tau)^2 \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n} \sum_{i=1}^{n} t_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n} \sum_{i=1}^{n} t_i \right] = \frac{1}{n^2} \sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

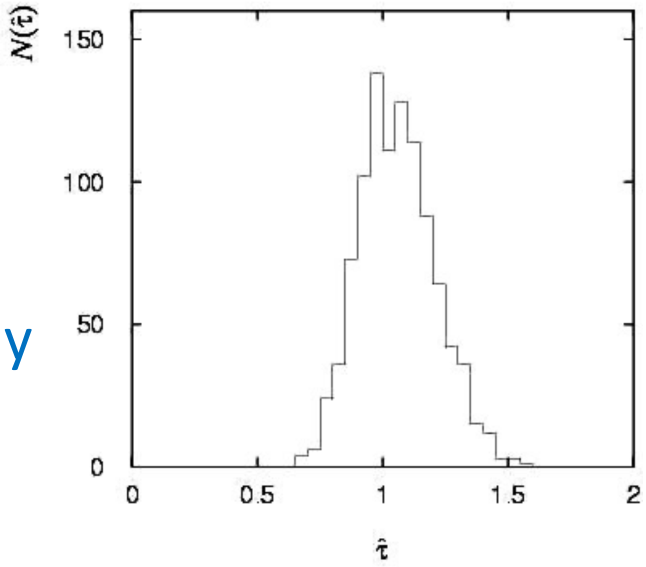
Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its ‘statistical error’, i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\widehat{\sigma}_{\widehat{\tau}} = 0.151$$



Note distribution of estimates is roughly Gaussian − (almost) always true for ML in large sample limit.

The information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only MLE). For a single parameter:

$$(b = E[\hat{\theta}] - \theta)$$

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] = \text{MVB}$$

(Minimum Variance Bound)

where $E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = \int \frac{\partial^2 \ln P(\mathbf{x}|\theta)}{\partial \theta^2} P(\mathbf{x}|\theta)\, d\mathbf{x}$

Proof in Exercise 6.6 of SDA, https://www.pp.rhul.ac.uk/~cowan/sda/prob/prob_6.pdf

“Efficiency” of an estimator = MVB / actual variance.

An estimator whose variance equals the MVB is said to be efficient.

MVB for MLE of exponential parameter

Find $$\mathrm{MVB} = -\left(1+\frac{\partial b}{\partial \tau}\right)^2 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right]$$

We found for the exponential parameter the MLE $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$

and we showed $b = 0$, hence $\partial b / \partial \tau = 0$.

We find $$\frac{\partial^2 \ln L}{\partial \tau^2} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - \frac{2t_i}{\tau^3}\right)$$

and since $E[t_i] = \tau$ for all i, $E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right] = -\frac{n}{\tau^2}$,

and therefore $\mathrm{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]$. So here the MLE is efficient.

Large-sample (asymptotic) properties of MLEs

Suppose we have an i.i.d. data sample of size n: $x_1,\ldots,x_n$

In the large-sample (or "asymptotic") limit ($n \to \infty$) and assuming regularity conditions one can show that the likelihood and MLE have several important properties.

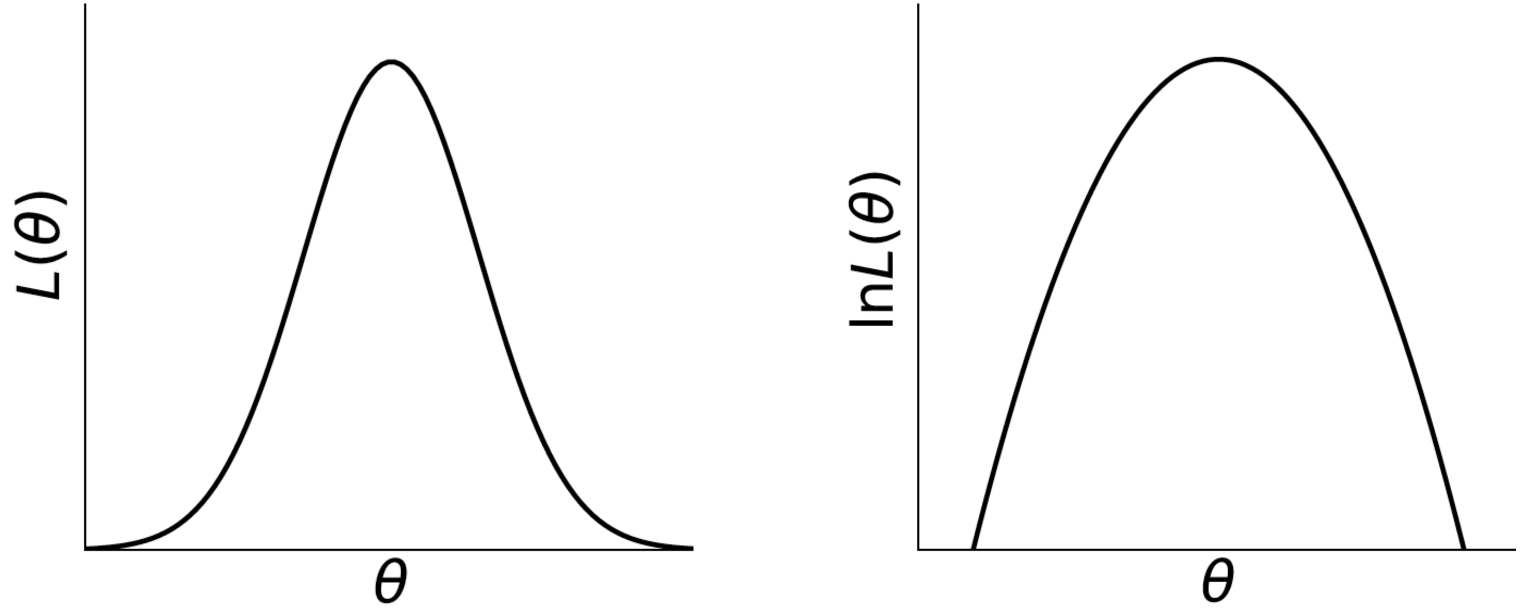
The regularity conditions include:

- the boundaries of the data space cannot depend on the parameter;
- the parameter cannot be on the edge of the parameter space;
- $\ln L(\theta)$ must be differentiable;
- the only solution to $\partial \ln L / \partial \theta = 0$ is $\hat{\theta}$.

In the slides immediately following, the properties are shown without proof for a single parameter; the corresponding properties hold also for the multiparameter case, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)$.

log-likelihood becomes quadratic

The likelihood function becomes Gaussian in shape, i.e. the log-likelihood becomes quadratic (parabolic).



The MLE becomes increasingly precise as the (log)-likelihood becomes more tightly concentrated about its peak,
but $L(\theta) = P(\boldsymbol{x}|\theta)$ is the probability for $\boldsymbol{x}$, not a pdf for θ.

The MLE converges to the true parameter value

In the large-sample limit, the MLE converges in probability to the true parameter value.

That is, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

The MLE is said to be *consistent*.

MLE is asymptotically unbiased

In general the MLE can be biased, but in the large-sample limit, this bias goes to zero:

$$\lim_{n \to \infty} E[\hat{\theta}] - \theta = 0$$

(Recall for the exponential parameter we found the bias was identically zero regardless of the sample size n.)

The MLE's variance approaches the MVB

In the large-sample limit, the variance of the MLE approaches the minimum-variance bound, i.e., the information inequality becomes an equality (and bias goes to zero):

$$\lim_{n\to\infty} V[\hat{\theta}] = -\frac{1}{E\left[\frac{\partial^2 \ln L}{\partial\theta^2}\right]}$$

The MLE is said to be *asymptotically efficient*.

The MLE's distribution becomes Gaussian

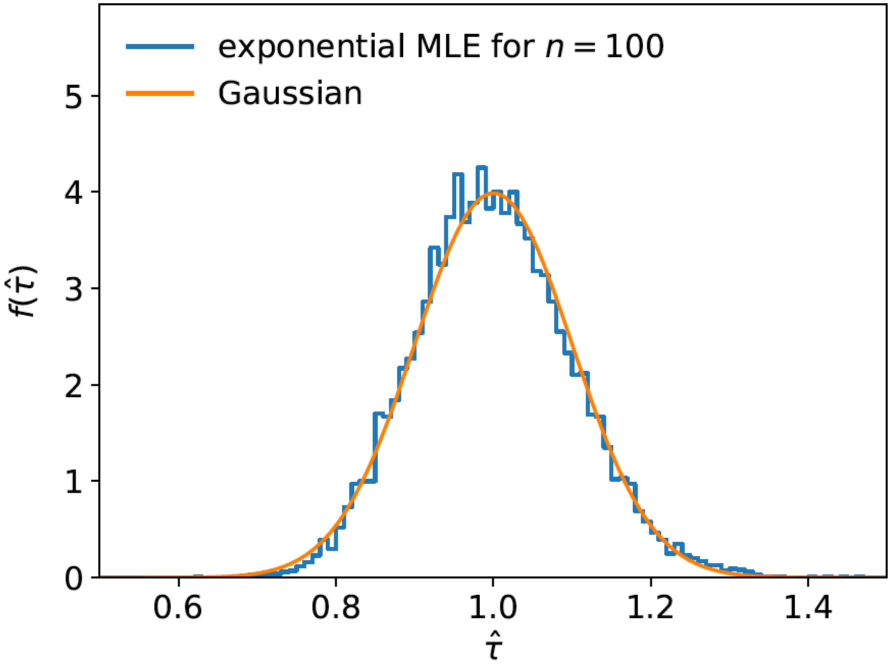
In the large-sample limit, the pdf of the MLE becomes Gaussian,

$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\theta}}} e^{-(\hat{\theta}-\theta)^2/2\sigma_{\hat{\theta}}^2}$$

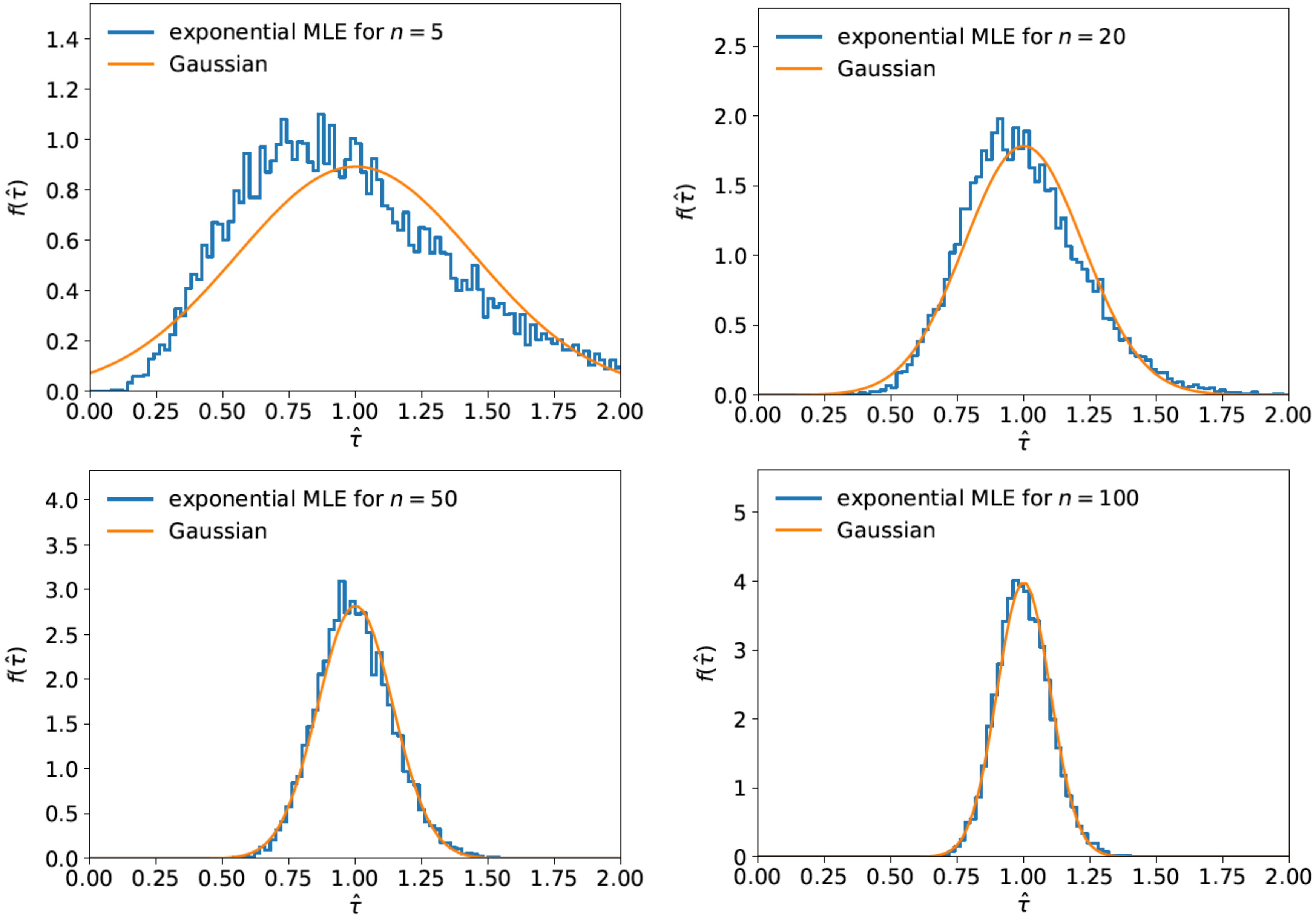
where $\sigma_{\hat{\theta}}^2$ is the minimum variance bound (note bias is zero).

For example, exponential MLE with sample size $n = 100$.

Note that for exponential, MLE is arithmetic average, so Gaussian MLE seen to stem from Central Limit Theorem.



Distribution of MLE of exponential parameter



Variance of estimators: graphical method

Expand $\ln L(\theta)$ about its maximum:

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

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

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

i.e.,

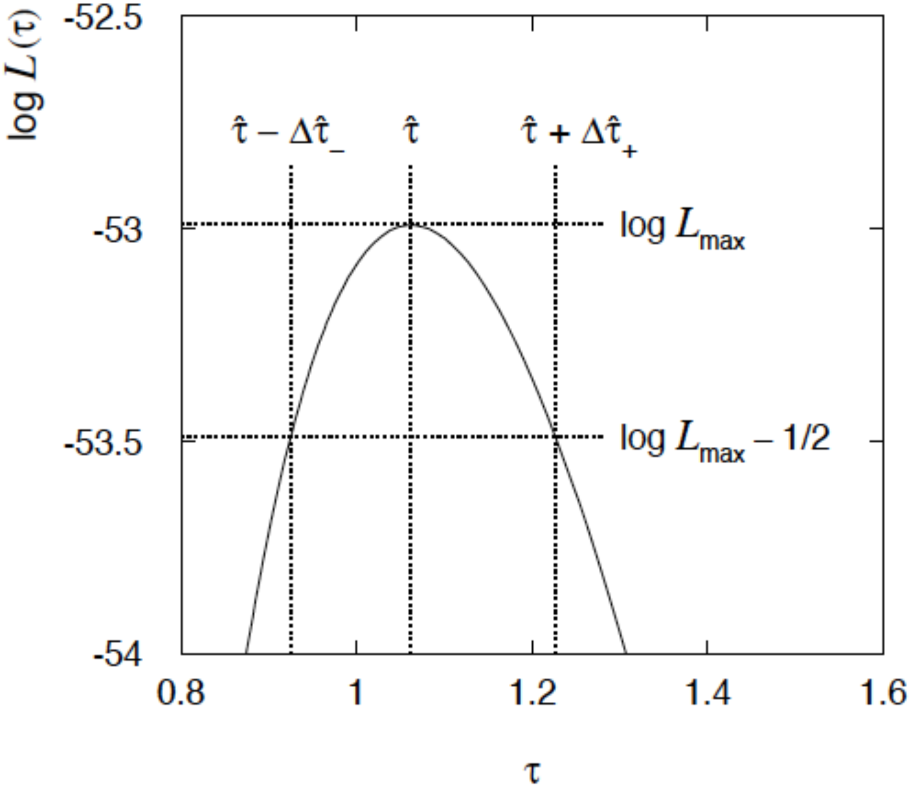
$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

→ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by 1/2.

Example of variance by graphical method

ML example with exponential:

$$\begin{aligned}\hat{\tau} &= 1.062 \\ \Delta\hat{\tau}_{-} &= 0.137 \\ \Delta\hat{\tau}_{+} &= 0.165 \\ \hat{\sigma}_{\hat{\tau}} &\approx \Delta\hat{\tau}_{-} \approx \Delta\hat{\tau}_{+} \approx 0.15\end{aligned}$$



Not quite parabolic $\ln L$ since finite sample size (n = 50).

Extra slides

Statistical Data Analysis
Lecture 9-2

- Goodness-of-fit from the likelihood ratio
- Wilks’ theorem
- MLE and goodness-of-fit all in one

Goodness of fit from the likelihood ratio

Suppose we model data using a likelihood $L(\boldsymbol{\mu})$ that depends on N parameters $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$. Define the statistic

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})}$$

where $\hat{\boldsymbol{\mu}}$ is the ML estimator for $\boldsymbol{\mu}$. Value of $t_{\boldsymbol{\mu}}$ reflects agreement between hypothesized $\boldsymbol{\mu}$ and the data.

Good agreement means $\boldsymbol{\mu} \approx \hat{\boldsymbol{\mu}}$, so $t_{\boldsymbol{\mu}}$ is small;

Larger $t_{\boldsymbol{\mu}}$ means less compatibility between data and $\boldsymbol{\mu}$.

Quantify "goodness of fit" with p-value: $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu})\, dt_{\boldsymbol{\mu}}$

need this pdf

Likelihood ratio (2)

Now suppose the parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ can be determined by another set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$, with $M < N$.

E.g., curve fit with $\mu_i = E[y_i] = \mu(x_i; \boldsymbol{\theta})$, $i = 1, \ldots, N$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$.

Want to test hypothesis that the true model is somewhere in the subspace $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ versus the alternative of the full parameter space $\boldsymbol{\mu}$. Generalize the LR test statistic to be

fit M parameters

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}$$

fit N parameters

To get p-value, need pdf $f(t_{\boldsymbol{\mu}} | \boldsymbol{\mu}(\boldsymbol{\theta}))$.

Wilks' Theorem

Wilks' Theorem: if the hypothesized $\mu_i(\boldsymbol{\theta})$, $i = 1,\ldots,N$, are true for some choice of the parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_M)$, then in the large sample limit (and provided regularity conditions are satisfied)

MLE of $(\theta_1,\ldots,\theta_M)$

$$t_{\boldsymbol{\mu}} = -2\ln\frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}$$

follows a chi-square distribution for $N - M$ degrees of freedom.

MLE of $(\mu_1,\ldots,\mu_N)$

The regularity conditions include: the model in the numerator of the likelihood ratio is "nested" within the one in the denominator, i.e., $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a special case of $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)$.

Proof boils down to having all estimators ~ Gaussian.

S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

Wilks' Theorem (2)

The chi-square pdf for $-2\ln\lambda$ breaks down:

if the sample size is too small;

if the true value of a parameter is on the boundary of the allowed parameter space;

if the model in the numerator is not a special case of the denominator (models must be "nested");

if variance of estimators of any components of $\boldsymbol{\mu}$ too large (e.g., parameter refers to location of a feature not present in the null hypothesis, such as the position of a peak).

Goodness of fit with Gaussian data

Suppose the data are N independent Gaussian distributed values:

$$y_i \sim \mathrm{Gauss}(\mu_i, \sigma_i) \; , \qquad i = 1, \ldots, N$$

want to estimate

known

N measurements and N parameters (= "saturated model")

Likelihood:

$$L(\boldsymbol{\mu}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2}$$

Log-likelihood:

$$\ln L(\boldsymbol{\mu}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C$$

ML estimators:

$$\hat{\mu}_i = y_i \qquad i = 1, \ldots, N$$

Likelihood ratio for Gaussian data

Now suppose $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$, e.g., in an LS fit with $\mu_i(\boldsymbol{\theta}) = \mu(x_i; \boldsymbol{\theta})$.

The goodness-of-fit statistic for the test of the hypothesis $\boldsymbol{\mu}(\boldsymbol{\theta})$ becomes

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})} = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\hat{\boldsymbol{\theta}}))^2}{\sigma_i^2} \sim \chi^2_{N-M}$$

chi-square pdf for N-M degrees of freedom

Here $t_{\boldsymbol{\mu}}$ is the same as $\chi^2_{\min}$ from an LS fit.

So Wilks' theorem formally states the property that we claimed for the minimized chi-squared from an LS fit with N measurements and M fitted parameters.

Likelihood ratio for Poisson data

Suppose the data are a set of values $\boldsymbol{n} = (n_1,\ldots, n_N)$, e.g., the numbers of events in a histogram with N bins.

Assume $n_i \sim$ Poisson(ν_i), $i = 1,\ldots, N$, all independent.

First (for LR denominator) treat $\boldsymbol{\nu} = (\nu_1,\ldots, \nu_N)$ as all adjustable:

Likelihood:
$$L(\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Log-likelihood:
$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} [n_i \ln \nu_i - \nu_i] + C$$

ML estimators:
$$\hat{\nu}_i = n_i \ , \qquad i = 1, \ldots, N$$

Goodness of fit with Poisson data (2)

For LR numerator find $\boldsymbol{\nu}(\boldsymbol{\theta})$ with M fitted parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$:

$$t_{\boldsymbol{\nu}} = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} - \nu_i(\hat{\boldsymbol{\theta}}) + n_i \right]$$

if $n_i = 0$, skip log term

Wilks' theorem: in large-sample limit $t_{\boldsymbol{\nu}} \sim \chi^2_{N-M}$

Exact in large sample limit; in practice good approximation for surprisingly small n_i (~several).

As before use t_ν to get p-value of $\boldsymbol{\nu}(\boldsymbol{\theta})$,

independent of $\boldsymbol{\theta}$

$$p_{\boldsymbol{\nu}} = \int_{t_{\boldsymbol{\nu},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\nu}} | \boldsymbol{\nu}(\boldsymbol{\theta}))\, dt_{\boldsymbol{\nu}} = 1 - F_{\chi^2}(t_{\boldsymbol{\nu},\mathrm{obs}}; N - M)$$

Goodness of fit with multinomial data

Similar if data $\boldsymbol{n} = (n_1,\ldots, n_N)$ follow multinomial distribution:

$$P(\mathbf{n}|\mathbf{p}, n_{\rm tot}) = \frac{n_{\rm tot}!}{n_1! n_2! \ldots n_N!} p_1^{n_1} p_2^{n_2} \ldots p_N^{n_N}$$

E.g. histogram with N bins but fix: $n_{\rm tot} = \sum_{i=1}^{N} n_i$

Log-likelihood: $\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_{\rm tot}} + C$ $\quad (\nu_i = p_i n_{\rm tot})$

ML estimators: $\hat{\nu}_i = n_i$ (Only N–1 independent; one is $n_{\rm tot}$ minus sum of rest.)

Goodness of fit with multinomial data (2)

The likelihood ratio statistics become:

$$t_{\boldsymbol{\nu}} = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i}$$

if $n_i = 0$, skip term

Wilks: in large sample limit $t_{\boldsymbol{\nu}} \sim \chi^2_{N-M-1}$

One less degree of freedom than in Poisson case because effectively only $N–1$ parameters fitted in denominator of LR.

Estimators and g.o.f. all at once

Evaluate numerators with $\boldsymbol{\theta}$ (not its estimator); if any $n_i = 0$, omit the corresponding log terms:

$$\chi^2_{\mathrm{P}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N}\left[n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} - \nu_i(\boldsymbol{\theta}) + n_i\right] \qquad \text{(Poisson)}$$

$$\chi^2_{\mathrm{M}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N} n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} \qquad \text{(Multinomial)}$$

These are equal to the corresponding $-2 \ln L(\boldsymbol{\theta})$ plus terms not depending on $\boldsymbol{\theta}$, so minimizing them gives the usual ML estimators for $\boldsymbol{\theta}$.

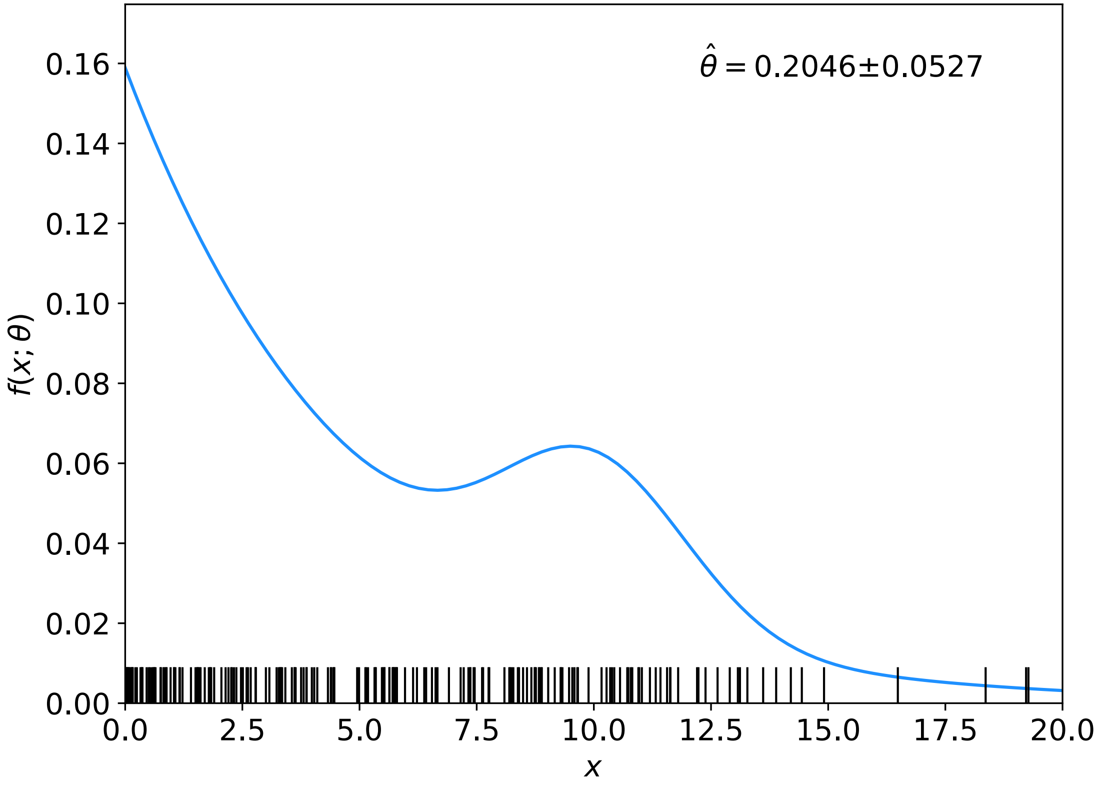
The minimized value gives the statistic t_ν, so we get goodness-of-fit for free.

Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.

Examples of ML/LS fits

Unbinned maximum likelihood (mlFit.py, minimize negLogL)

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

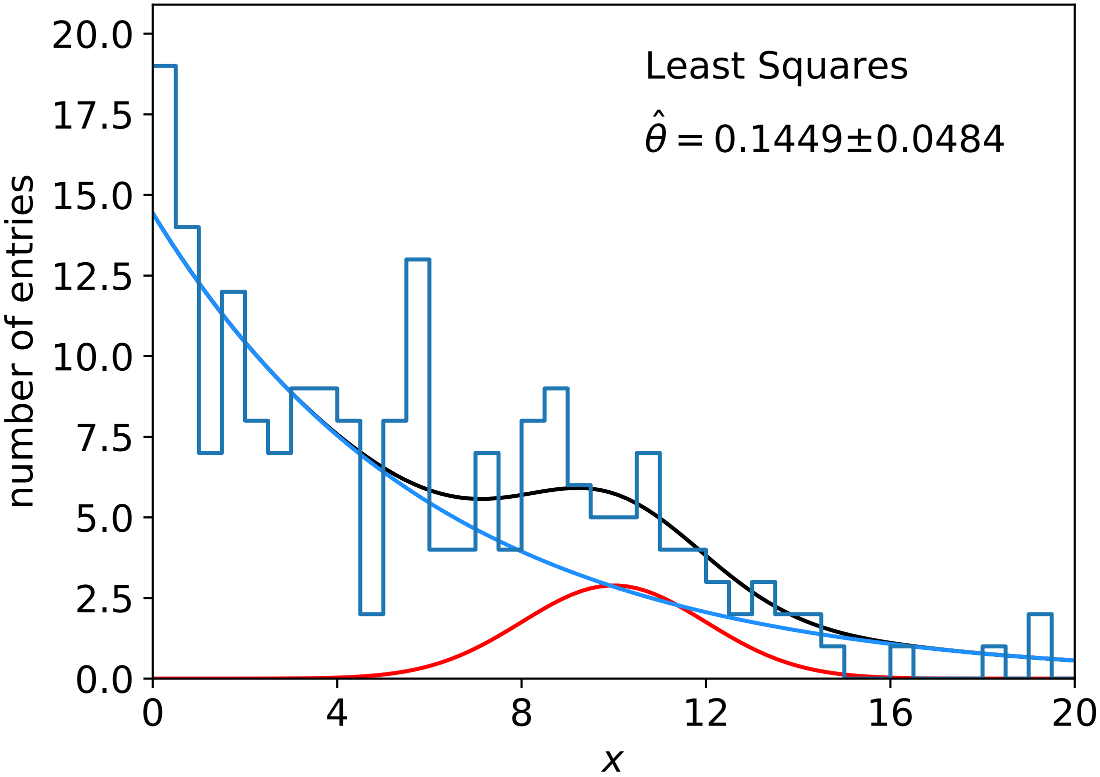


No useful measure of goodness-of-fit from unbinned ML.

Examples of ML/LS fits

Least Squares fit (histFit.py, minimize chi2LS)

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\mu_i(\boldsymbol{\theta})}$$



$\chi^2_{\min} = 32.7$

$n_{\mathrm{dof}} = 38$

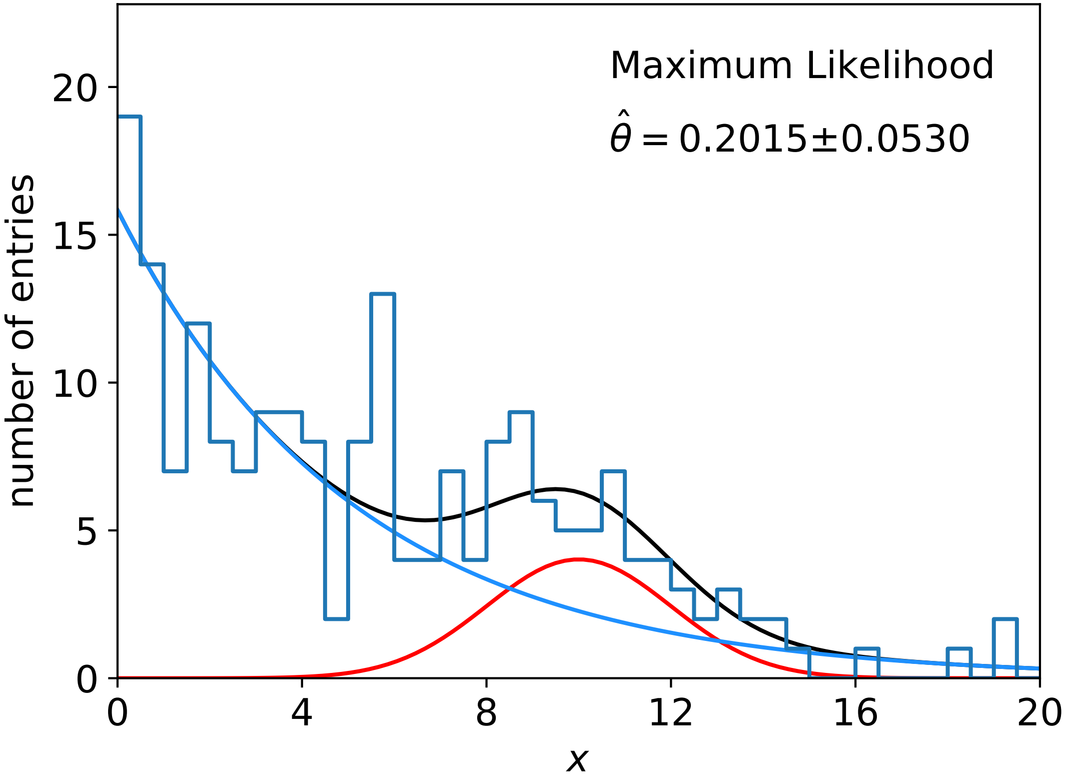
$p = 0.71$

Many bins with few entries, LS not expected to be reliable.

Examples of ML/LS fits

Multinomial maximum likelihood fit (histFit.py, minimize chi2M)

$$\chi^2_{\mathrm{M}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N} n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i}$$



$\chi^2_{\mathrm{min}} = 35.3$

$n_{\mathrm{dof}} = 37$

$p = 0.55$

Essentially same result as unbinned ML.

Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

Luca Lista, *Statistical Methods for Data Analysis in Particle Physics,* Springer, 2017.

L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

S. Brandt, *Statistical and Computational Methods in Data Analysis,* Springer, New York, 1998.

R.L. Workman et al. (Particle Data Group), Prog. Theor. Exp. Phys. 083C01 (2022); **`pdg.lbl.gov`** sections on probability, statistics, MC.

Some distributions

Distribution/pdf	Example use in Particle Physics
Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy	Mass of resonance
Landau	Ionization energy loss
Beta	Prior pdf for efficiency
Gamma	Sum of exponential variables
Student’s t	Resolution function with adjustable tails

Binomial distribution

Consider N independent experiments (Bernoulli trials):

outcome of each is ‘success’ or ‘failure’,

probability of success on any given trial is p.

Define discrete r.v. n = number of successes ($0 \leq n \leq N$).

Probability of a specific outcome (in order), e.g. ‘ssfsf’ is

$$pp(1-p)p(1-p) = p^n (1-p)^{N-n}$$

But order not important; there are $\displaystyle \frac{N!}{n!(N-n)!}$

ways (permutations) to get n successes in N trials, total probability for n is sum of probabilities for each permutation.

Binomial distribution (2)

The binomial distribution is therefore

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

random variable

parameters

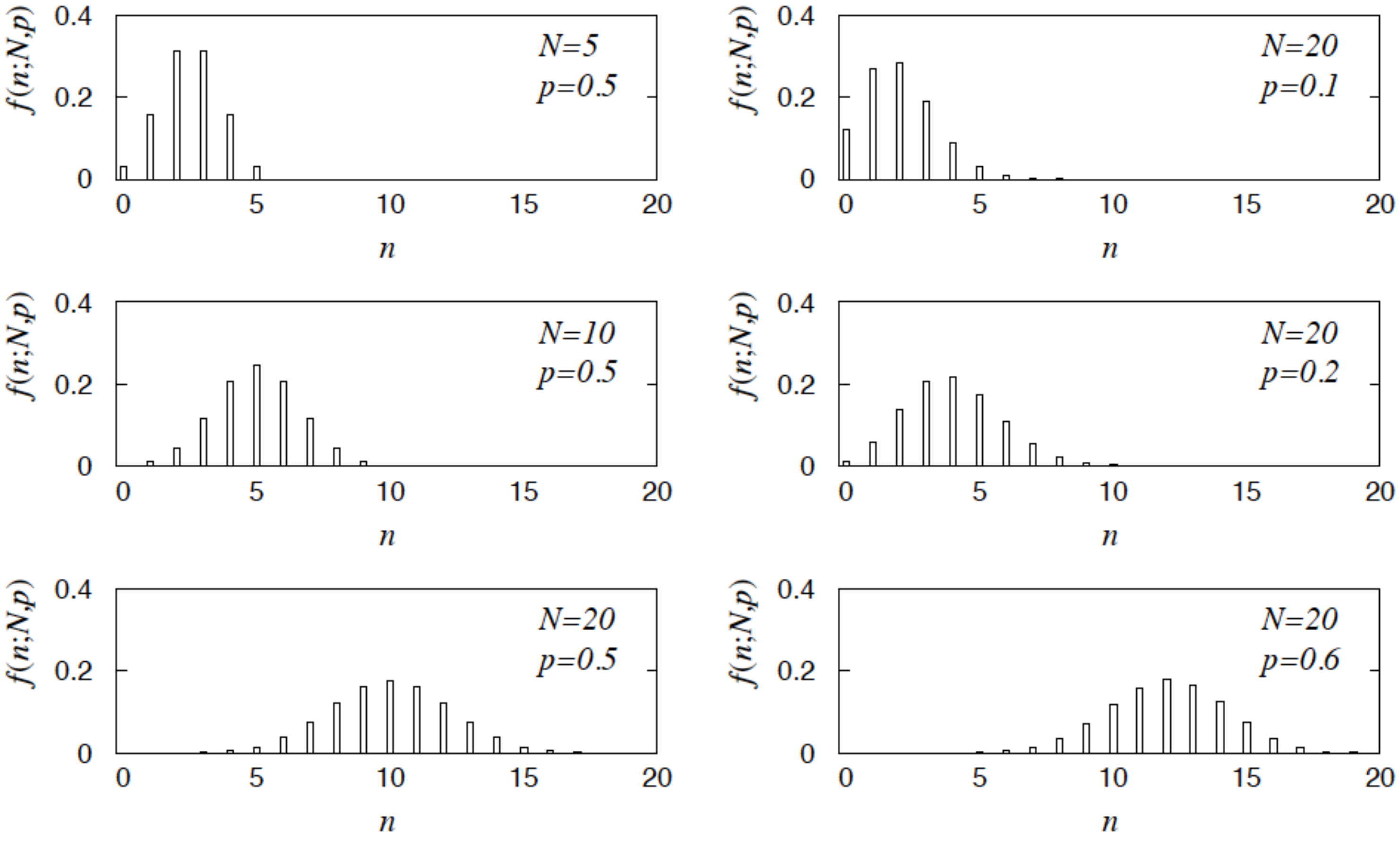
For the expectation value and variance we find:

$$E[n] = \sum_{n=0}^{N} n f(n;N,p) = Np$$

$$V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$

Binomial distribution (3)

Binomial distribution for several values of the parameters:



Example: observe N decays of $W^{\pm}$, the number n of which are $W \to \mu\nu$ is a binomial r.v., p = branching ratio.

Multinomial distribution

Like binomial but now m outcomes instead of two, probabilities are

$$\vec{p} = (p_1, \ldots, p_m)\,, \quad \text{with} \; \sum_{i=1}^{m} p_i = 1\,.$$

For N trials we want the probability to obtain:

n_1 of outcome 1,
n_2 of outcome 2,
$\vdots$
n_m of outcome m.

This is the multinomial distribution for $\vec{n} = (n_1, \ldots, n_m)$

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

Multinomial distribution (2)

Now consider outcome i as 'success', all others as 'failure'.

→ all n_i individually binomial with parameters N, p_i

$$E[n_i] = Np_i, \qquad V[n_i] = Np_i(1-p_i) \qquad \text{for all } i$$

One can also find the covariance to be

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

Example: $\vec{n} = (n_1, \ldots, n_m)$ represents a histogram

with m bins, N total entries, all entries independent.

Poisson distribution

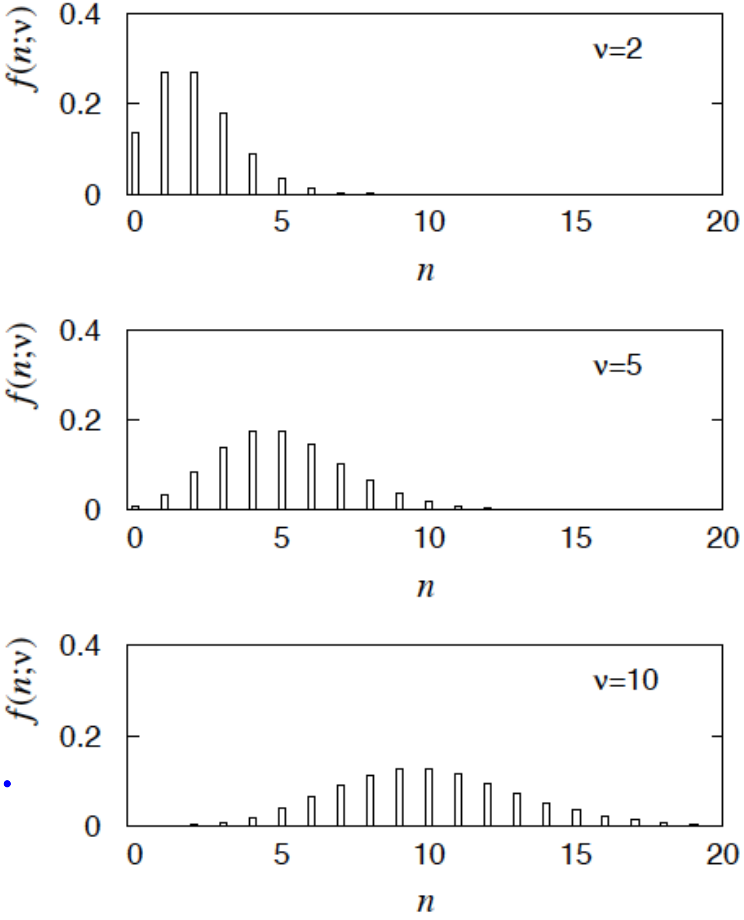
Consider binomial n in the limit

$$N \to \infty, \qquad p \to 0, \qquad E[n] = Np \to \nu \, .$$

→ n follows the Poisson distribution:

$$f(n;\nu) = \frac{\nu^n}{n!} e^{-\nu} \qquad (n \geq 0)$$

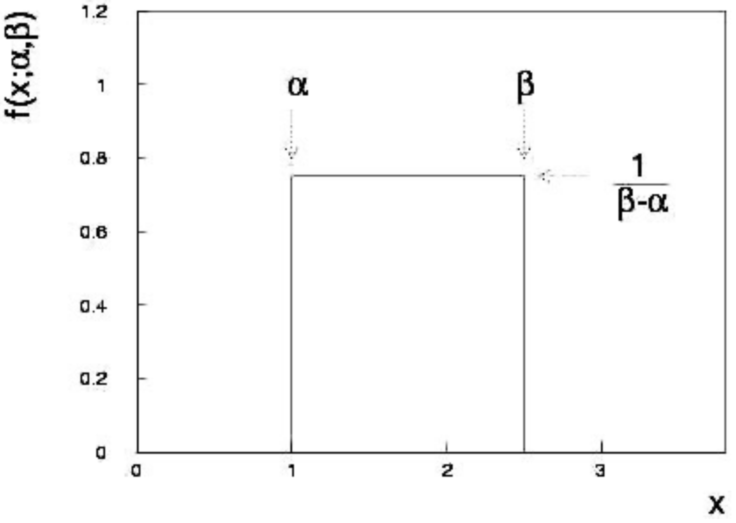
$$E[n] = \nu \, , \qquad V[n] = \nu \, .$$



Example: number of scattering events n with cross section σ found for a fixed integrated luminosity, with $\nu = \sigma \int L \, dt$.

Uniform distribution

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$



$$E[x] = \frac{1}{2}(\alpha + \beta)$$

$$V[x] = \frac{1}{12}(\beta - \alpha)^2$$

Notation: x follows a uniform distribution between α and β

write as: $x \sim \mathrm{U}[\alpha,\beta]$

Uniform distribution (2)

Very often used with $\alpha = 0, \beta = 1$ (e.g., Monte Carlo method).

For any r.v. x with pdf $f(x)$, cumulative distribution $F(x)$, the function $y = F(x)$ is uniform in $[0,1]$:

$$g(y) = f(x) \left| \frac{dx}{dy} \right| = \frac{f(x)}{|dy/dx|}$$

$$= \frac{f(x)}{|dF/dx|} = \frac{f(x)}{f(x)} = 1 \,, \quad 0 \leq y \leq 1$$

because $f(x) = dF/dx = dy/dx$

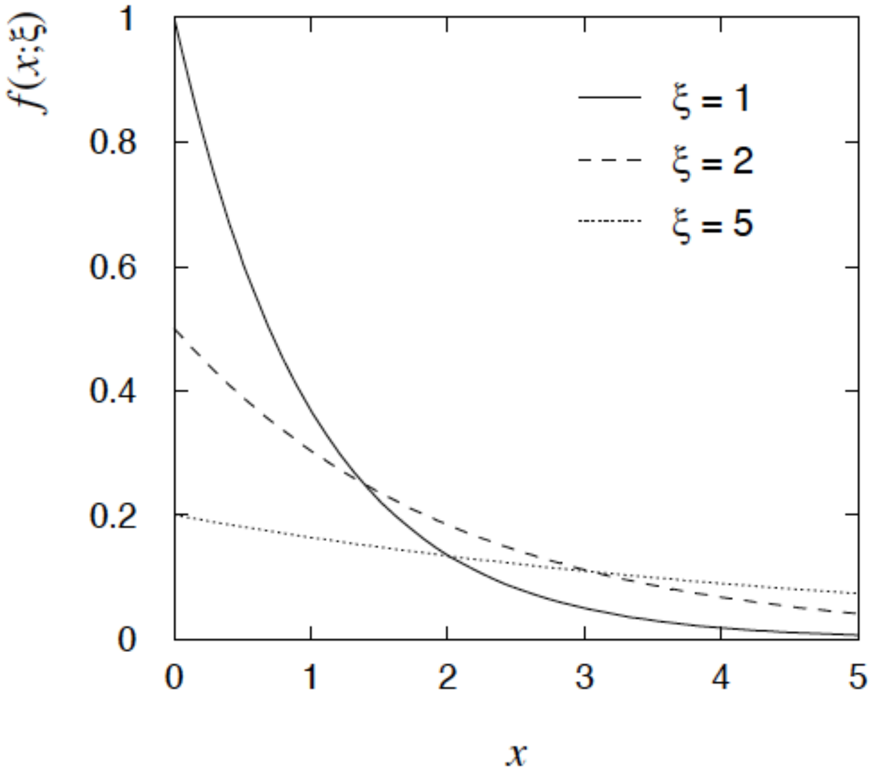
Exponential distribution

The exponential pdf for the continuous r.v. x is defined by:

$$f(x;\xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \xi$$

$$V[x] = \xi^2$$



Exponential distribution (2)

Example: proper decay time t of an unstable particle

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \qquad (\tau = \text{mean lifetime})$$

Lack of memory (unique to exponential): $f(t - t_0 | t \geq t_0) = f(t)$

Question for discussion:

A cosmic ray muon is created 30 km high in the atmosphere, travels to sea level and is stopped in a block of scintillator, giving a start signal at t_0. At a time t it decays to an electron giving a stop signal. What is distribution of the difference between stop and start times, i.e., the pdf of $t - t_0$ given $t > t_0$?

Gaussian (normal) distribution

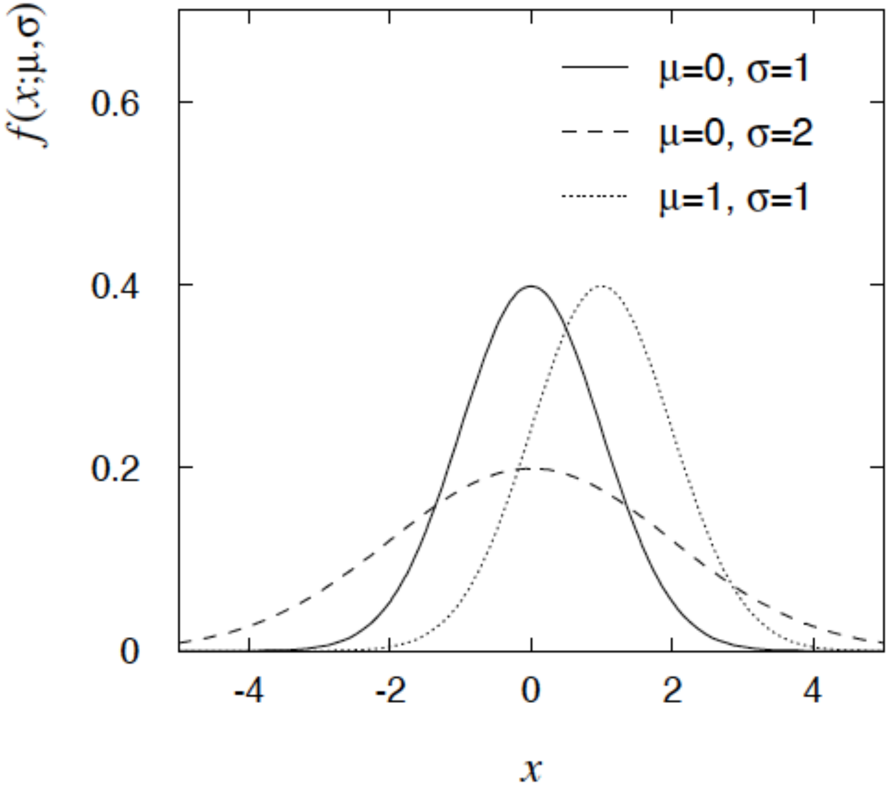
The Gaussian (normal) pdf for a continuous r.v. x is defined by:

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

$$E[x] = \mu$$

$$V[x] = \sigma^2$$

N.B. often μ, σ^2 denote mean, variance of any r.v., not only Gaussian.



Standardized random variables

If a random variable y has pdf $f(y)$ with mean μ and std. dev. σ, then the *standardized* variable

$$x = \frac{y - \mu}{\sigma} \quad \text{has the pdf} \quad g(x) = f(y(x)) \left| \frac{dy}{dx} \right| = \sigma f(\mu + \sigma x)$$

has mean of zero and standard deviation of 1.

Often work with the *standard* Gaussian distribution (μ = 0. σ = 1) using notation:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \,, \qquad \Phi(x) = \int_{-\infty}^{x} \varphi(x')\, dx'$$

Then e.g. $y = \mu + \sigma x$ follows

$$f(y) = \frac{1}{\sigma} \varphi \left(\frac{y - \mu}{\sigma} \right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

Multivariate Gaussian distribution

Multivariate Gaussian pdf for the vector $\vec{x} = (x_1, \ldots, x_n)$:

$$f(\vec{x}; \vec{\mu}, V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1} (\vec{x}-\vec{\mu}) \right]$$

$\vec{x}, \ \vec{\mu}$ are column vectors, $\vec{x}^T, \ \vec{\mu}^T$ are transpose (row) vectors,

$$E[x_i] = \mu_i, \ , \qquad \mathrm{cov}[x_i, x_j] = V_{ij} \ .$$

Marginal pdf of each x_i is Gaussian with mean μ_i, standard deviation $\sigma_i = \sqrt{V_{ii}}$.

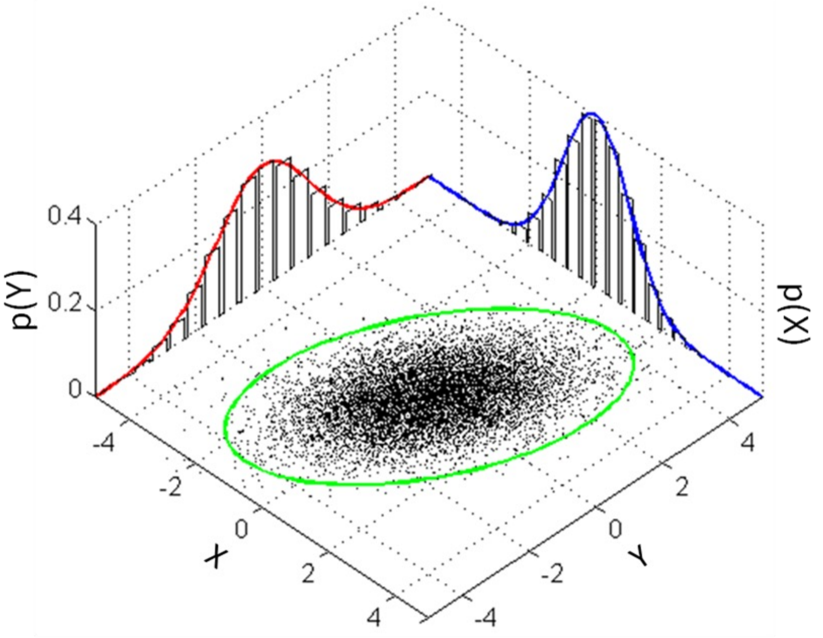
https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Two-dimensional Gaussian distribution

$$f(x_1, x_2, ; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right] \right\}$$

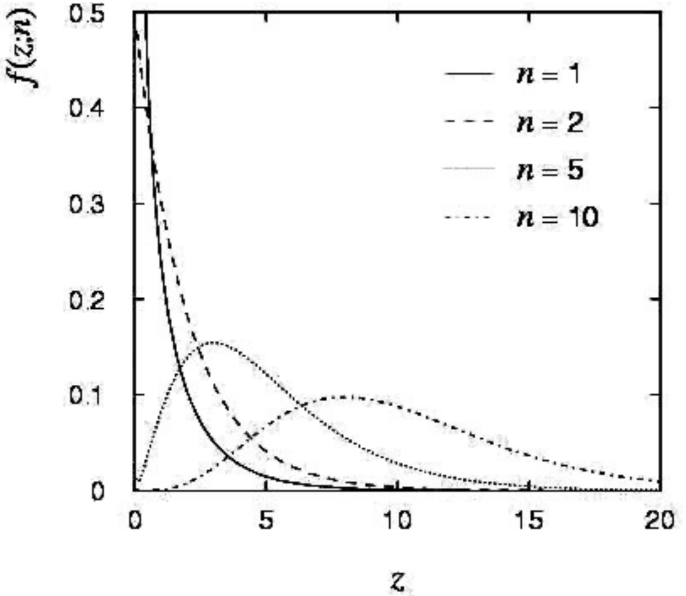
where $\rho = \mathrm{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ is the correlation coefficient.



Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z ($z \geq 0$) is defined by

$$f(z;n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$



n = 1, 2, ... = number of 'degrees of freedom' (dof)

$$E[z] = n \, , \qquad V[z] = 2n \; .$$

For independent Gaussian x_i, i = 1, ..., n, means μ_i, variances σ_i^2,

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

follows χ^2 pdf with n dof.

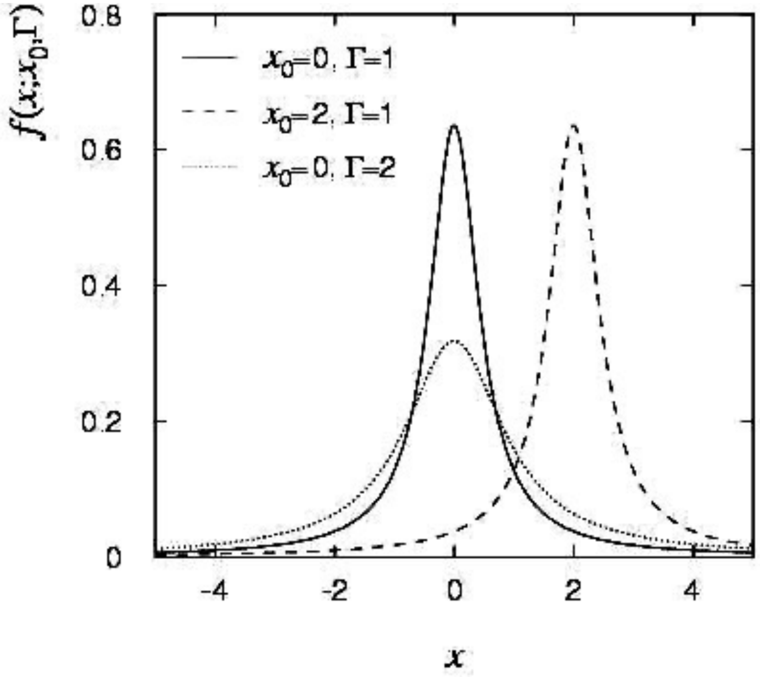
Example: goodness-of-fit test variable especially in conjunction with method of least squares.

Cauchy (Breit-Wigner) distribution

The Breit-Wigner pdf for the continuous r.v. x is defined by

$$f(x;\Gamma,x_0)=\frac{1}{\pi}\frac{\Gamma/2}{\Gamma^2/4+(x-x_0)^2}$$

($\Gamma = 2$, $x_0 = 0$ is the Cauchy pdf.)



$E[x]$ not well defined, $V[x] \to \infty$.

x_0 = mode (most probable value)

Γ = full width at half maximum

Example: mass of resonance particle, e.g. ρ, K^*, φ^0, ...

Γ = decay rate (inverse of mean lifetime)

Landau distribution

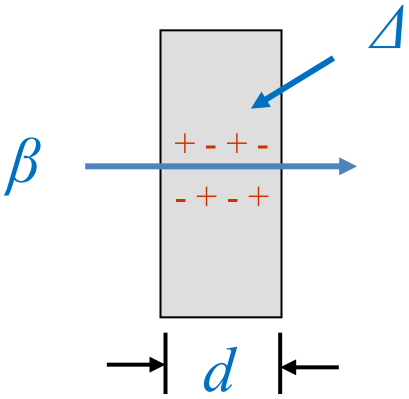
For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d, the energy loss Δ follows the Landau pdf:

$$f(\Delta;\beta) = \frac{1}{\xi}\phi(\lambda) \ ,$$

$$\phi(\lambda) = \frac{1}{\pi}\int_0^\infty \exp(-u\ln u - \lambda u)\sin \pi u \, du \ ,$$

$$\lambda = \frac{1}{\xi}\left[\Delta - \xi\left(\ln\frac{\xi}{\epsilon'} + 1 - \gamma_{\mathrm{E}}\right)\right] \ ,$$

$$\xi = \frac{2\pi N_{\mathrm{A}} e^4 z^2 \rho \sum Z}{m_{\mathrm{e}} c^2 \sum A}\frac{d}{\beta^2} \ , \qquad \epsilon' = \frac{I^2 \exp \beta^2}{2 m_{\mathrm{e}} c^2 \beta^2 \gamma^2} \ .$$



L. Landau, J. Phys. USSR **8** (1944) 201; see also
W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

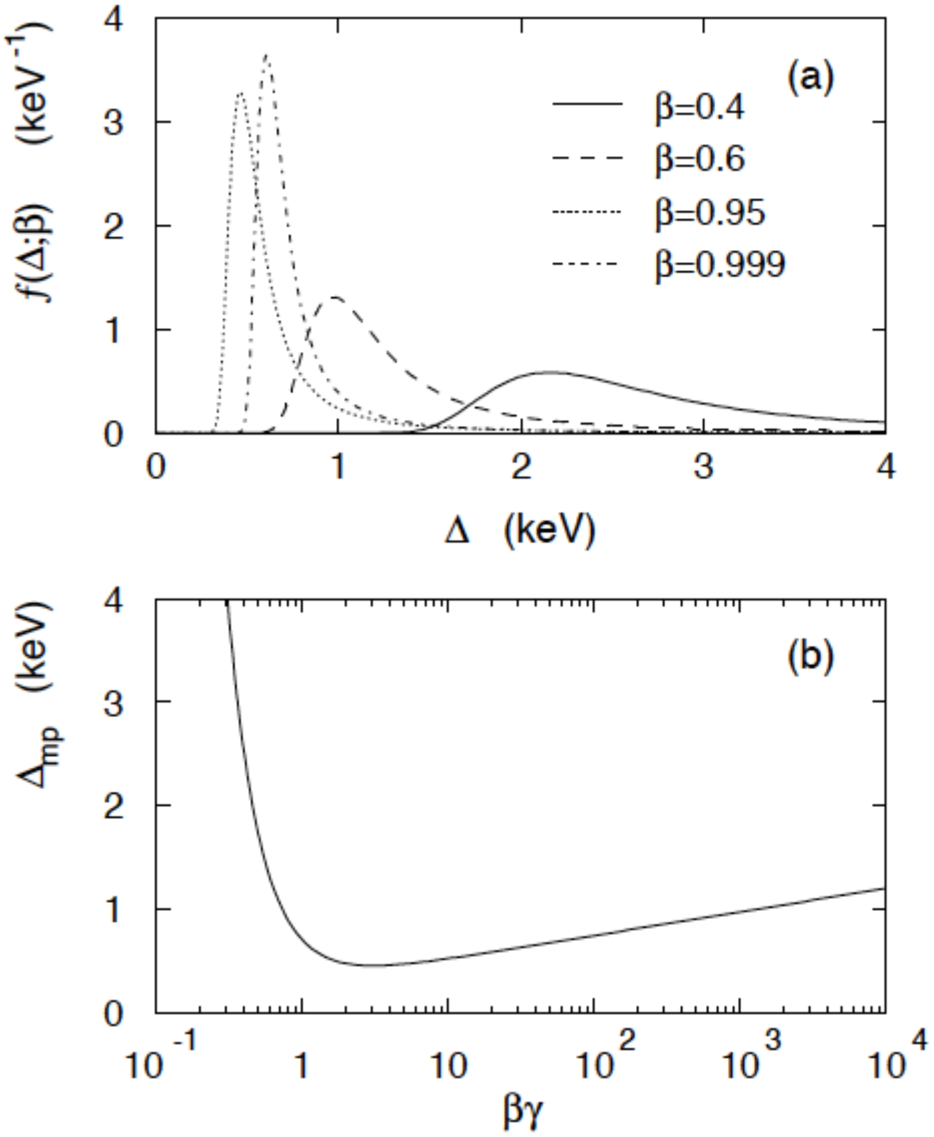
Landau distribution (2)

Long 'Landau tail'

→ all moments ∞

Mode (most probable value) sensitive to β ,

→ particle i.d.

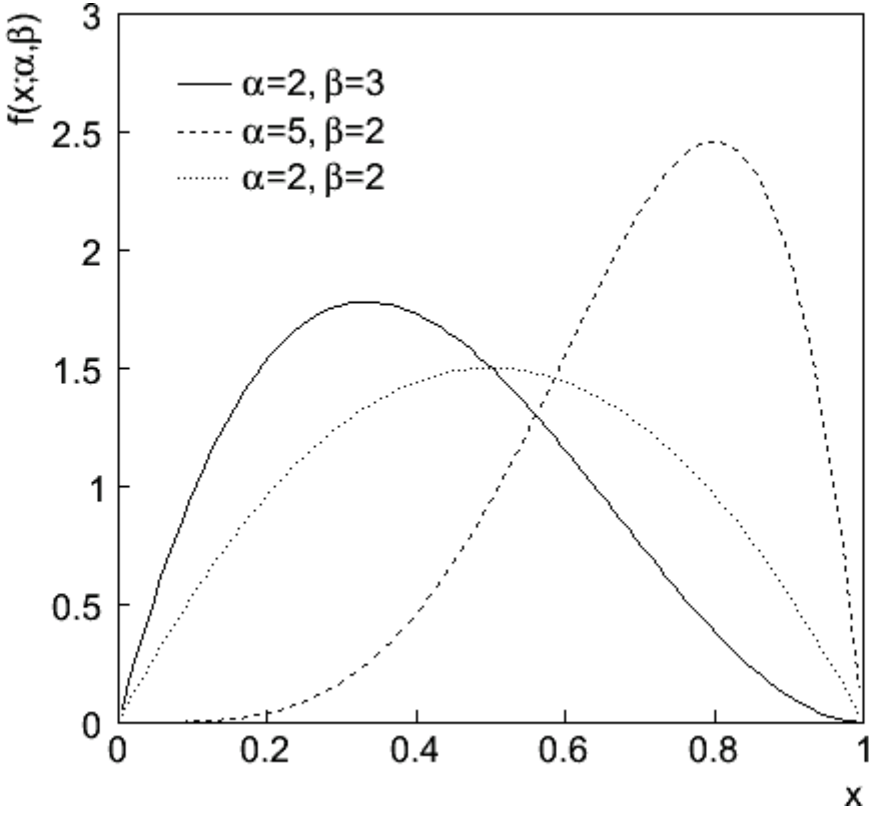


Beta distribution

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$$

$$E[x] = \frac{\alpha}{\alpha+\beta}$$

$$V[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



Often used to represent pdf of continuous r.v. nonzero only between finite limits, e.g.,

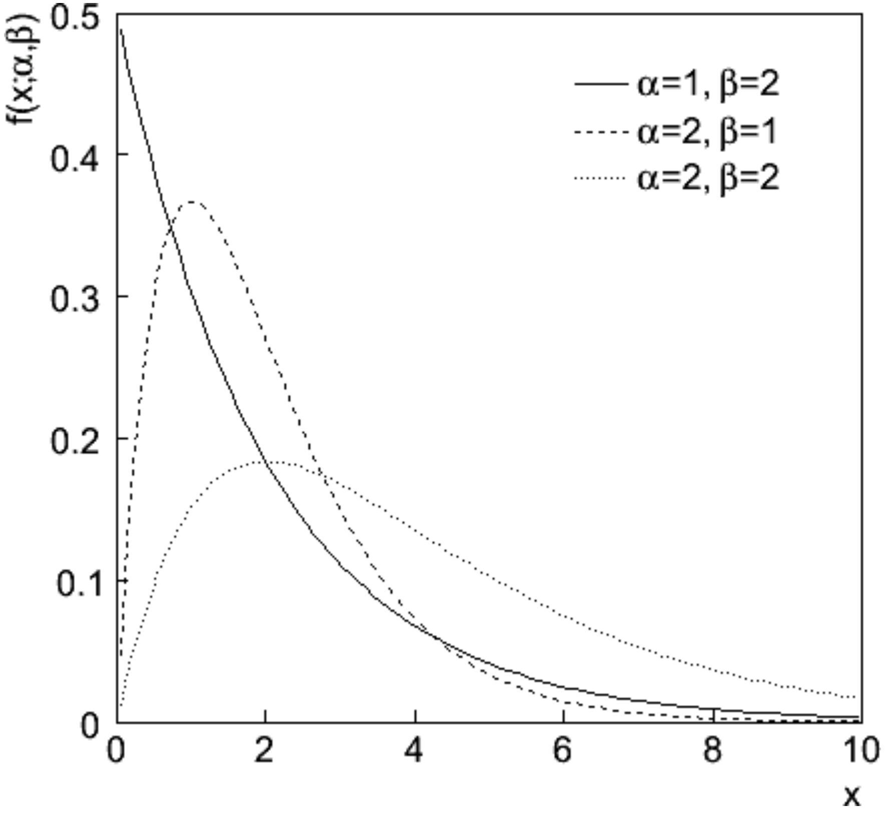
$y = a_0 + a_1 x, \quad a_0 \le y \le a_0 + a_1$

Gamma distribution

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\, x^{\alpha-1}\, e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$



Often used to represent pdf of continuous r.v. nonzero only in $[0,\infty]$.

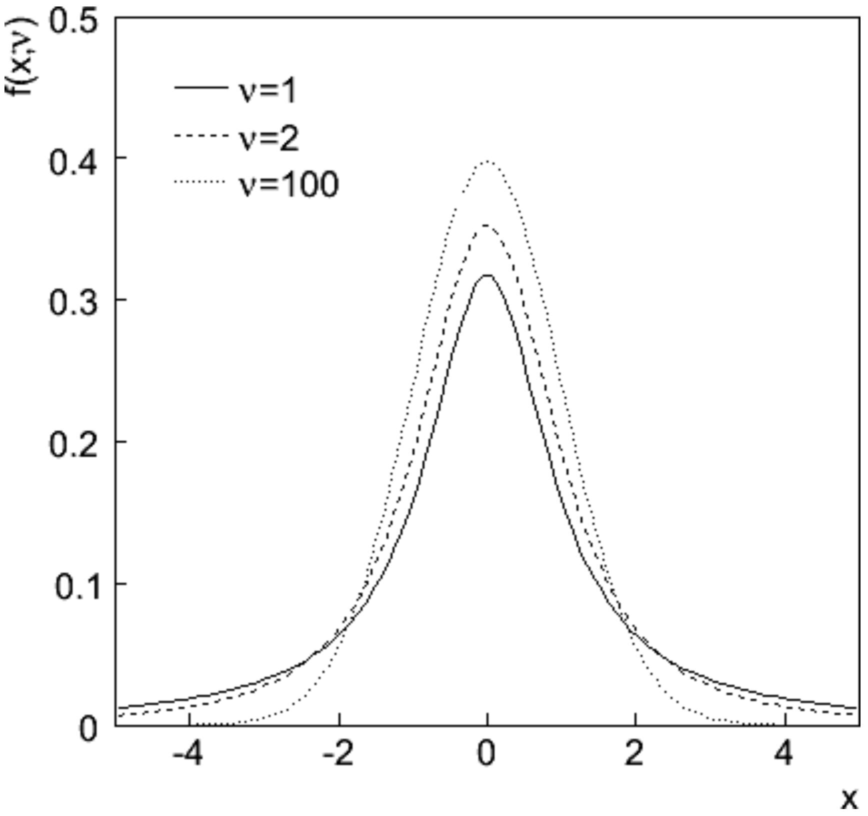
Also e.g. sum of n exponential r.v.s or time until nth event in Poisson process $\sim$ Gamma

Student's t distribution

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

$$E[x] = 0 \quad (\nu > 1)$$

$$V[x] = \frac{\nu}{\nu-2} \quad (\nu > 2)$$



ν = number of degrees of freedom (not necessarily integer)

$\nu = 1$ gives Cauchy,

$\nu \rightarrow \infty$ gives Gaussian.

Student's *t* distribution (2)

If $x \sim$ Gaussian with $\mu = 0$, $\sigma^2 = 1$, and

$z \sim \chi^2$ with n degrees of freedom, then

$t = x \,/\, (z/n)^{1/2}$ follows Student's t with $\nu = n$.

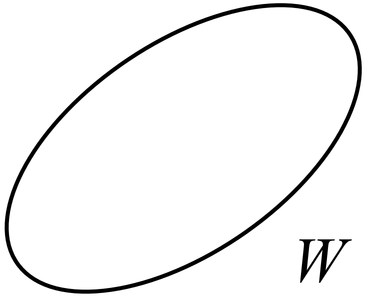
This arises in problems where one forms the ratio of a sample mean to the sample standard deviation of Gaussian r.v.s.

The Student's t provides a bell-shaped pdf with adjustable tails, ranging from those of a Gaussian, which fall off very quickly, ($\nu \rightarrow \infty$, but in fact already very Gauss-like for $\nu =$ two dozen), to the very long-tailed Cauchy ($\nu = 1$).

Developed in 1908 by William Gosset, who worked under the pseudonym "Student" for the Guinness Brewery.

Proof of Neyman-Pearson Lemma

Consider a critical region W and suppose the LR satisfies the criterion of the Neyman-Pearson lemma:

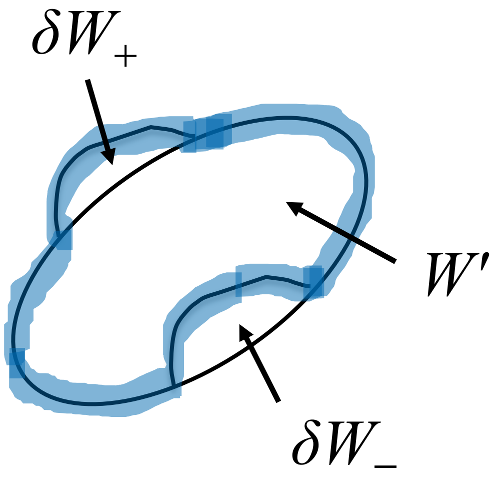


$P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0) \geq c_\alpha$ for all $\boldsymbol{x}$ in W,

$P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0) \leq c_\alpha$ for all $\boldsymbol{x}$ not in W.

Try to change this into a different critical region W' retaining the same size α, i.e.,

$$P(\mathbf{x} \in W'|H_0) = P(\mathbf{x} \in W|H_0) = \alpha$$



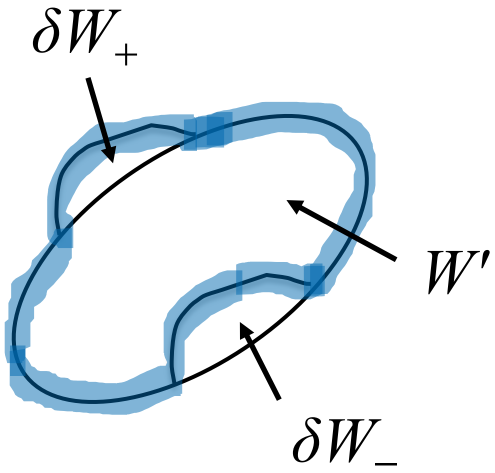
To do so add a part δW_+, but to keep the size α, we need to remove a part δW_-, i.e.,

$$W \rightarrow W' = W + \delta W_+ - \delta W_-$$

$$P(\mathbf{x} \in \delta W_+|H_0) = P(\mathbf{x} \in \delta W_-|H_0)$$

Proof of Neyman-Pearson Lemma (2)

But we are supposing the LR is higher for all $\boldsymbol{x}$ in δW_{-} removed than for the $\boldsymbol{x}$ in δW_{+} added, and therefore



$$P(\mathbf{x} \in \delta W_{+} | H_1) \leq P(\mathbf{x} \in \delta W_{+} | H_0) c_{\alpha}$$

$$P(\mathbf{x} \in \delta W_{-} | H_1) \geq P(\mathbf{x} \in \delta W_{-} | H_0) c_{\alpha}$$

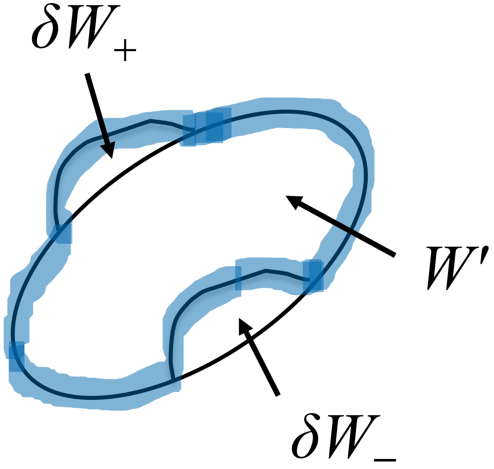
The right-hand sides are equal and therefore

$$P(\mathbf{x} \in \delta W_{+} | H_1) \leq P(\mathbf{x} \in \delta W_{-} | H_1)$$

Proof of Neyman-Pearson Lemma (3)

We have

$$W \cup W' = W \cup \delta W_+ = W' \cup \delta W_-$$



Note W and δW_+ are disjoint, and W' and δW_- are disjoint, so by Kolmogorov's 3rd axiom,

$$P(\mathbf{x} \in W') + P(\mathbf{x} \in \delta W_-) = P(\mathbf{x} \in W) + P(\mathbf{x} \in \delta W_+)$$

Therefore

$$P(\mathbf{x} \in W'|H_1) = P(\mathbf{x} \in W|H_1) + \underbrace{P(\mathbf{x} \in \delta W_+|H_1) - P(\mathbf{x} \in \delta W_-|H_1)}_{\leq 0}$$

Proof of Neyman-Pearson Lemma (4)

And therefore

$$P(\mathbf{x} \in W'|H_1) \leq P(\mathbf{x} \in W|H_1)$$

i.e. the deformed critical region W' cannot have higher power than the original one that satisfied the LR criterion of the Neyman-Pearson lemma.