

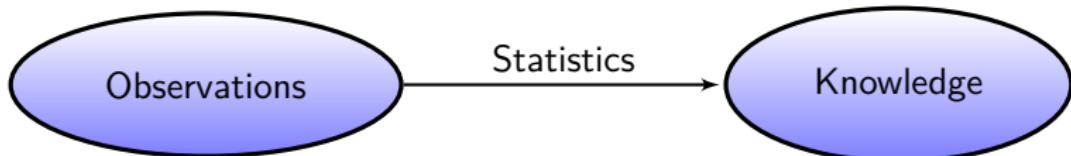
Statistics

E. Busato

October 7, 2019

Why this lecture ?

- Statistics omnipresent in our scientific (and everyday lives) activities
- **Purpose of this lecture:**
 - Fundamentals of statistical reasoning
 - Introduce basic concepts useful in specific fields such as machine learning
- Statistics = science of analyzing data and learning something useful out of it



Typical kind knowledge one is interested in

Value of parameters

- What is the mass of the electron ?
- What is the age of my clients ?

Correlation of variables

- How do the chances of developing some disease increase as people get older ?
- How do my benefits vary over the years ? What will they be in two years from now ?

Nature of underlying model describing my observations

- Are my observations best described by a linear or a quadratic function ?
- Is my physics model any good at describing the data I observe ?

Typical decisions one might want to take

- Should I buy a lotto ticket ?
- You're building a brand new product to sell on the market
 - Can you go and sell it (i.e. am I going to make some benefits) or is it too risky ?
- My car is old and showing some problems
 - Should I buy a new one now or wait ?

- In physics, one is interested in revealing the laws of nature, i.e. finding the best models describing the physical world
- **Examples of models:**
 - Newton's law of gravitation
 - General relativity
 - Quantum mechanics
 - Maxwell's equations
 - ...
- Experiments are carried in order to:
 - Decide whether these models are good
 - If yes, determine values of the model parameters

⇒ **Statistics crucial in physics**

General remarks about the learning of statistics

- Learning statistics is not an easy task
 - Scientific field on its own
 - Specific vocabulary and set of technics
 - Very large and active community producing new knowledge everyday
 - Not part of most french undergraduate and graduate programs
(except those for which statistics is the main topic of study)
- Easy to get lost
 - Large variety of methods
 - No unique solution to a given problem

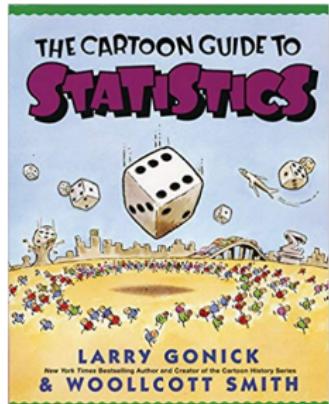
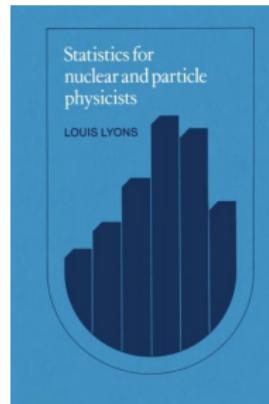
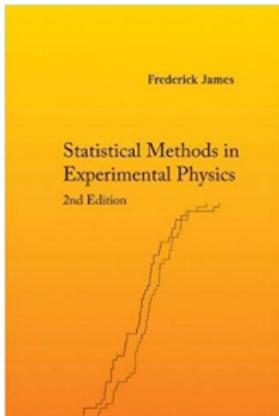
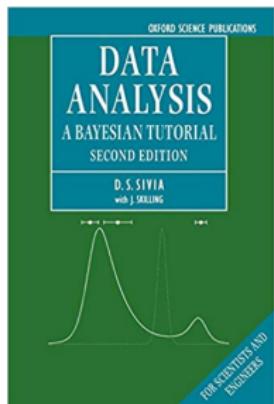
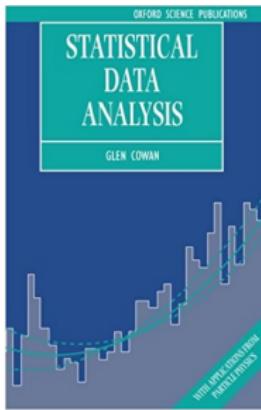
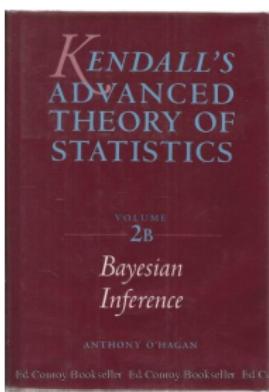
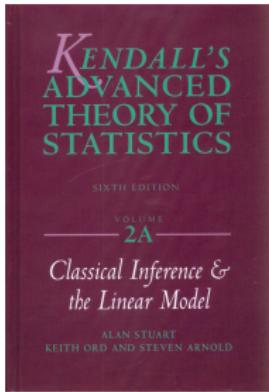
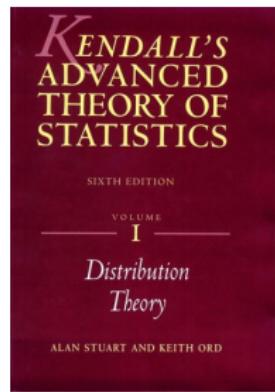
General remarks about the learning of statistics

- Can let statistics tell you almost anything if you don't pay attention
 - Applying statistical methods blindly, without understanding them a minimum, isn't a good idea
 - Doing statistics properly requires quite a lot of thinking and practice

**Get your mind in the world of statistics
focussing on fundamental concepts**

- **Competence and skills after this lecture:**
 - ☞ Formulate statistical problems using advanced modern concepts
 - ☞ Understand some very important statistical results and methods used everywhere

References



Basic blocks of statistical reasoning

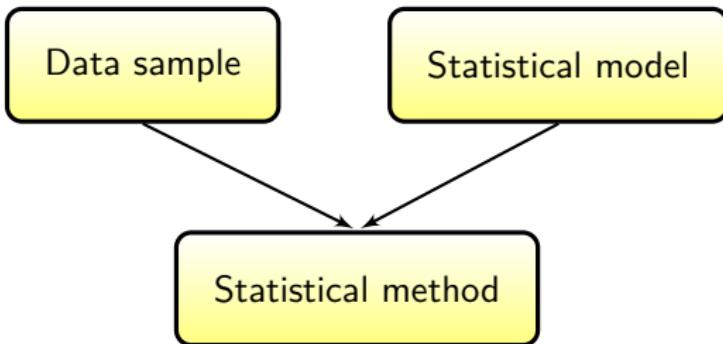
Data sample

Basic blocks of statistical reasoning

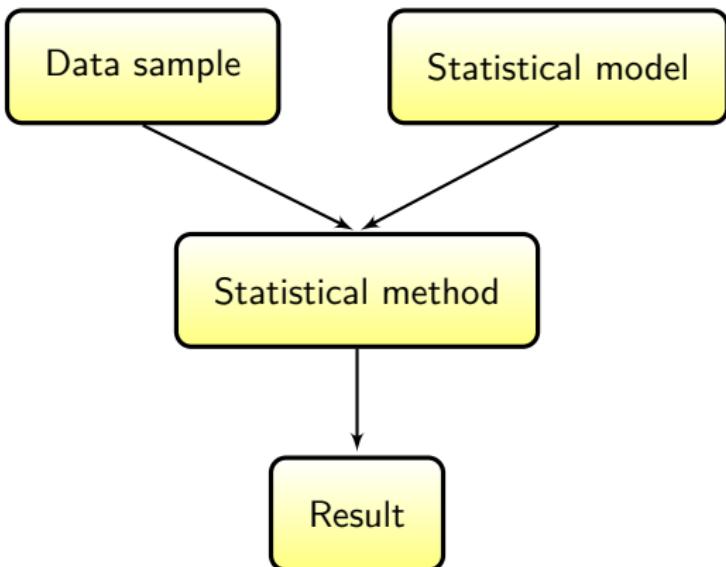
Data sample

Statistical model

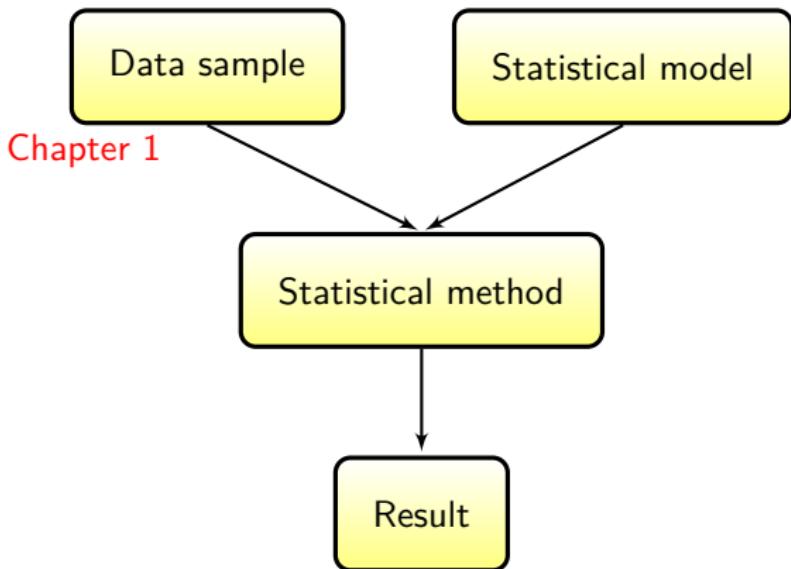
Basic blocks of statistical reasoning



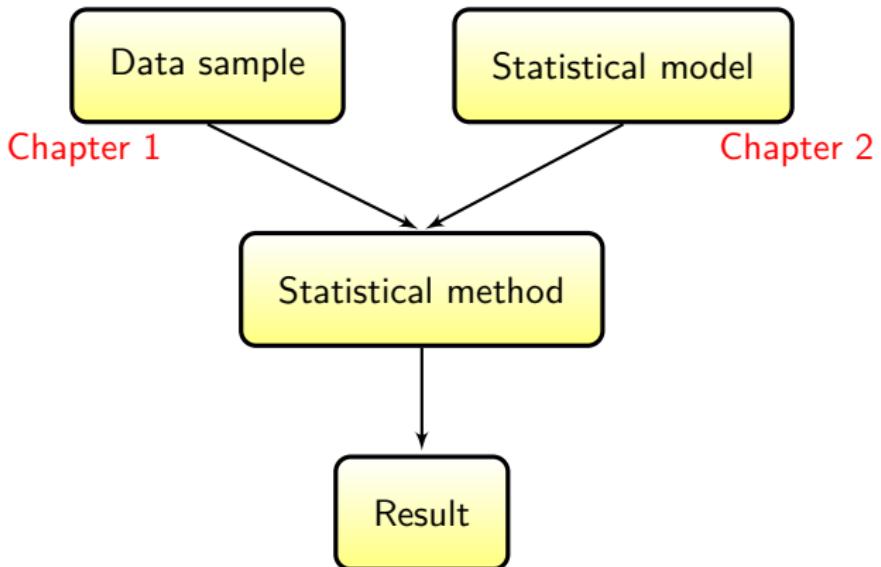
Basic blocks of statistical reasoning



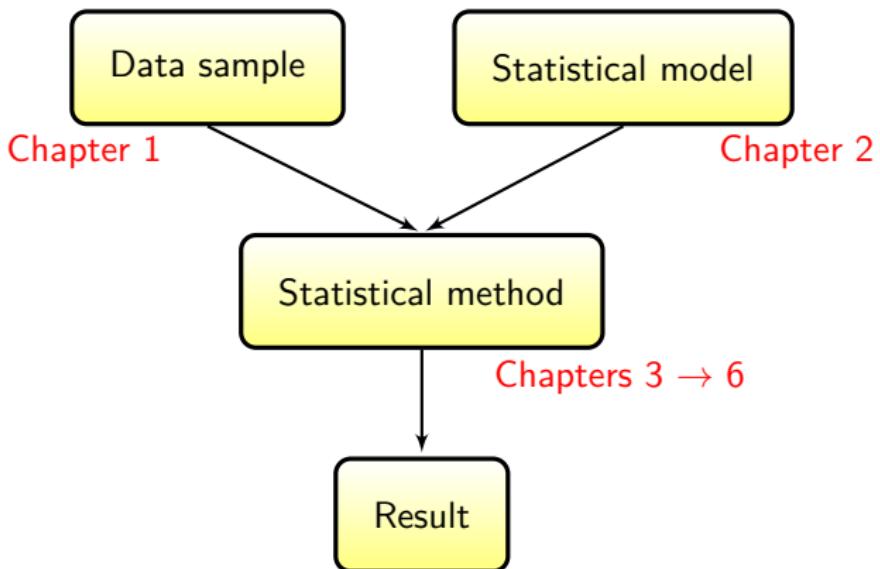
Basic blocks of statistical reasoning



Basic blocks of statistical reasoning

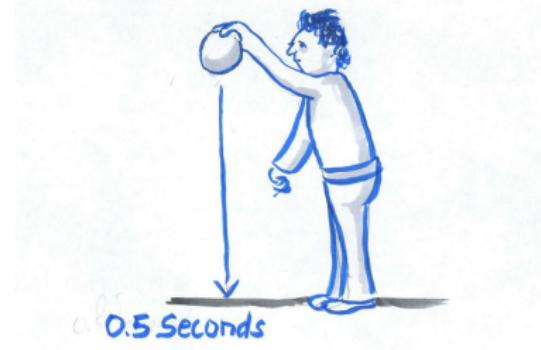


Basic blocks of statistical reasoning



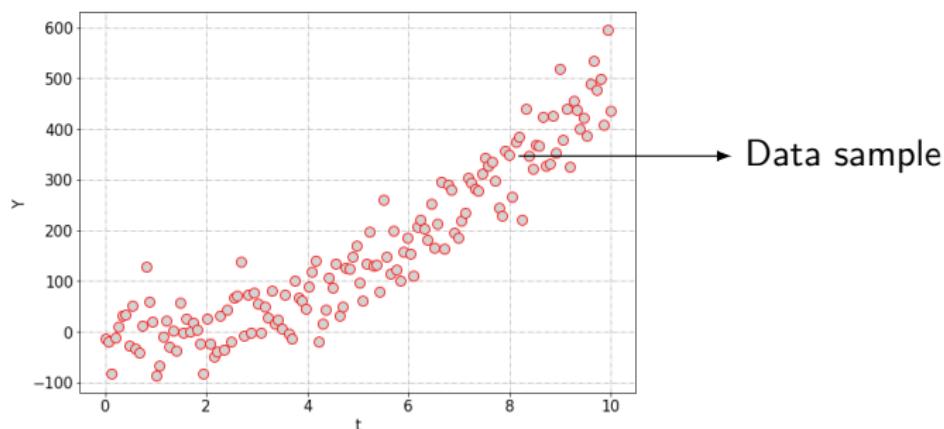
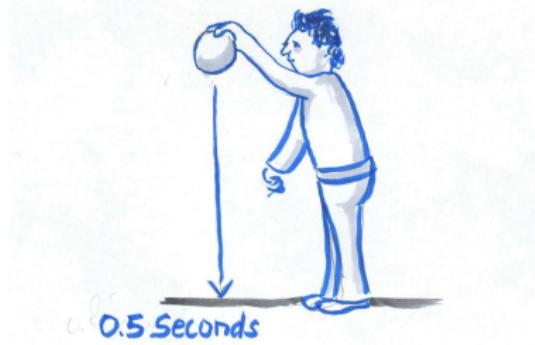
Preliminary example

- **Objective:** measure acceleration due to gravity
- **Experiment:** measure time t and position Y of free falling object

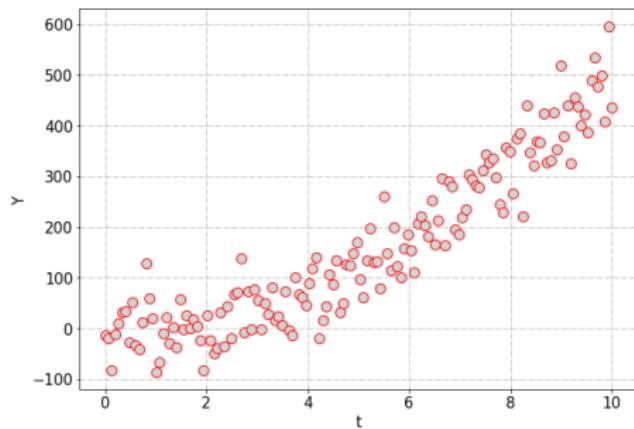


Preliminary example

- **Objective:** measure acceleration due to gravity
- **Experiment:** measure time t and position Y of free falling object

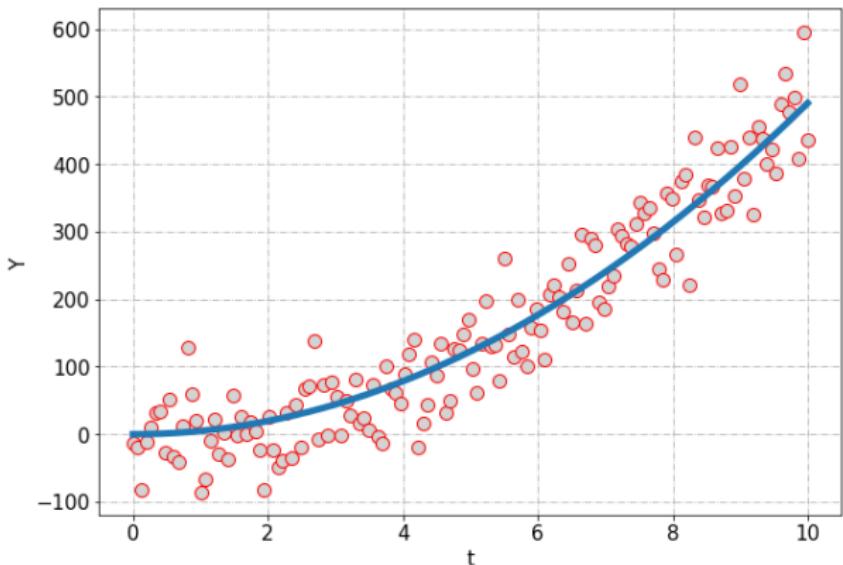


Preliminary example



- From physics we know
$$Y = \frac{1}{2}gt^2$$
- Measurement adds some fluctuations:
$$Y = \frac{1}{2}gt^2 + \varepsilon \rightarrow \text{stat. model}$$
- How do we measure g from these data and this stat. model ?
 - One possibility: linear regression

Preliminary example



Best fit: $\hat{g} = 9.86 \text{ m.s}^{-2}$

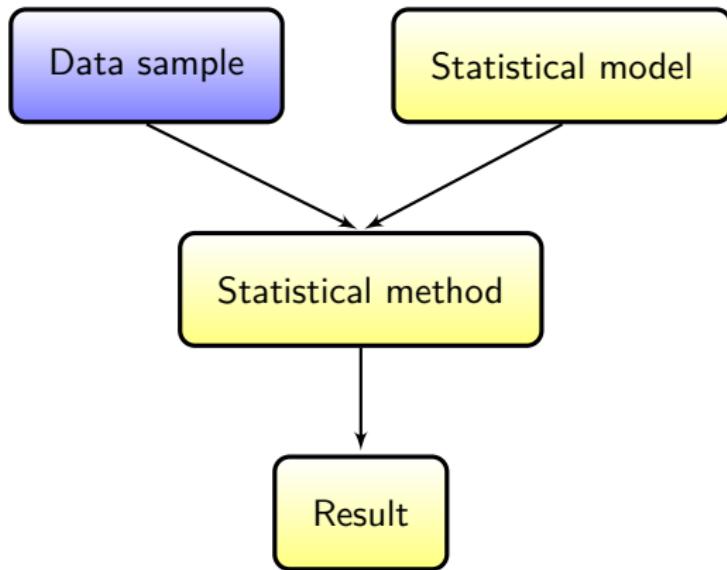
Sometimes you will see yellow background slides.

- They correspond to a review of some basic notions important for the topic under discussion

Slides in blue are exercices

Chap. 1 Sample

Basic blocks of statistical reasoning



What is a sample ?

- In a nutshell:

Sample = outcome of a measurement

- Measurement = survey, opinion poll, census, scientific experiment or any other data collection process
- Sample made of **random variables**
 - Sample = set of all random variables that are measured
- Synonyms: **sample = dataset (or data set)**

How is the sample chosen ?

- Choice of what random variables to measure depends on the question you're addressing
- Sample chosen so as to be a good representation of the population**



- If you're not the one who carried the measurement but your job is to analyze its data
 - Make sure you know as many details as possible about the sample and how it has been obtained

Example of samples

- Suppose you measure the temperature (T) 10 times to learn something about your system

$$\Rightarrow \text{Sample} = (T_0, T_1, \dots, T_9)$$

- Suppose you make a survey to learn something about the age (A) and height (H) of people in a certain population

$$\Rightarrow \text{Sample} = ((A_0, H_0), (A_1, H_1), \dots, (A_{n-1}, H_{n-1})) \quad (n = \text{number of people})$$

- Suppose you measure the 4-vect of a set of n particles

$$\Rightarrow \text{Sample} = ((E_0, \vec{p}_0), (E_1, \vec{p}_1), \dots, (E_{n-1}, \vec{p}_{n-1}))$$

Size

- Can be variable or not
- Usual notation: n

Number of dimensions

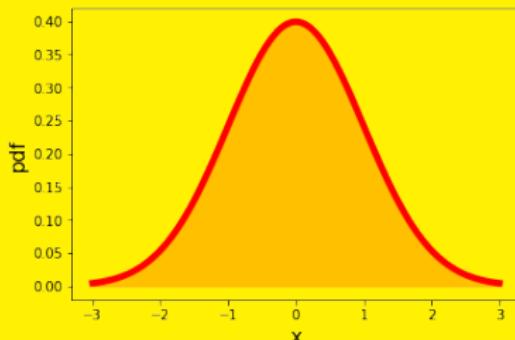
- Can be equal to 1 or greater
- Can be variable from one event to the other (e.g. if you measure for each event a different number of things)

Type of variables → See next slide

- Can be discrete or continuous or a mixture of both

Continuous random variable

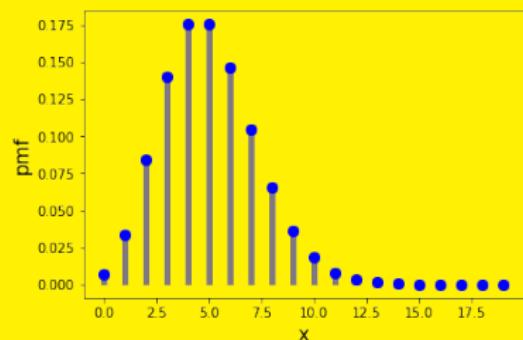
- Characterized by pdf: $f_X(x; \theta)$



- $P(x \leq X \leq x + dx) = f_X(x; \theta)dx$
- $\mathbb{E}[X] = \int xf_X(x; \theta)dx$

Discrete random variable

- Characterized by pmf: $p_X(x; \theta)$



- $p_X(x; \theta) = \text{prob. that } X = x$
- $\mathbb{E}[X] = \sum_i x_i p_X(x_i; \theta)$

Other important relations:

$$\rightarrow \text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$\rightarrow \text{cdf: } F(t; \theta) = P(X \leq t; \theta)$$

Sample notations

$$X = \{X_i\} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,m_1} \\ \vdots & & & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,m_n} \end{pmatrix}$$

For lazy people

Abbreviated notation

Short “ensemble” notation

Full
“matrix-like”
notation

- In the following, and unless stated otherwise, we will consider that the X'_i 's ($i \in [1, n]$) are iid random variables.
- iid = independent and identically distributed**

→ See next slide

Independence

- By definition, X and Y are independent if:

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

Or equivalently: $f_X(x|y) = f_X(x)$ and $f_Y(y|x) = f_Y(y)$

- **Remark:** Dependence \neq Correlation

- Correlation measured by covariance:

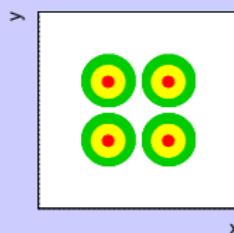
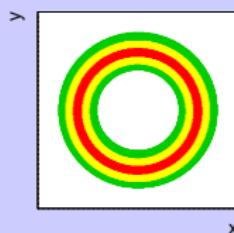
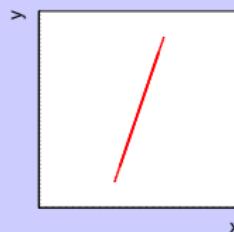
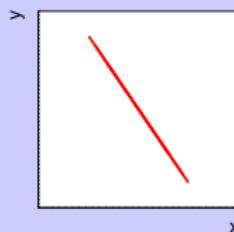
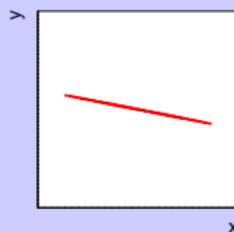
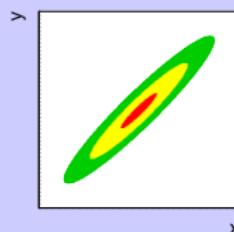
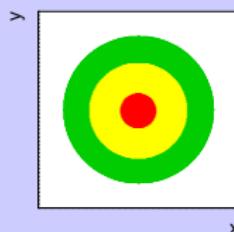
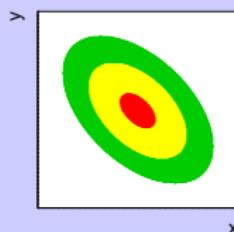
$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- Uncorrelated when $\text{cov}(X, Y) = 0$

- **Independence \Rightarrow Uncorrelation** (but not the contrary)

Exercice

Are the variables represented below independent ? uncorrelated ?



Characterization of samples: empirical quantities

- **Sample mean:**

$$M = \frac{1}{n} \sum_{i=1}^n X_i$$

- **Sample median:** assuming $X_1 < X_2 < \dots < X_n$

- median = $\frac{X_{n/2} + X_{1+n/2}}{2}$ (n even)
- median = $X_{(n+1)/2}$ (n odd)

- **Sample variance:**

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - M^2$$

or

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M)^2$$

- **Sample covariance matrix:** $Q = \frac{1}{n-1} \sum_{i=1}^n (\vec{X}_i - \vec{M})(\vec{X}_i - \vec{M})^T$

Characterization of samples: empirical quantities

- All empirical quantities are of the form:

$$Z = g(X_1, \dots, X_n)$$

Diagram illustrating the components of an empirical quantity:

- empirical quantity**: Z
- some function**: g
- sample**: (X_1, \dots, X_n)

- The X_i 's are random $\Rightarrow Z$ is random

- Empirical quantities fluctuate if we repeat the measurement
- Empirical quantities are characterized by a mean, a variance, ...

→ See next slide

Functions of random variables

□ **Mean:**

$$\begin{aligned}\mathbb{E}[Z] \simeq g(\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]) + \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 g}{\partial X_i^2} \text{var}[X_i] \\ + \sum_{\text{all pairs } (i,j)} \frac{\partial^2 g}{\partial X_i \partial X_j} \text{cov}(X_i, X_j)\end{aligned}$$

→ Particular case: $Z = X_1 + \dots + X_n \Rightarrow \mathbb{E}[Z] = \sum_{i=1}^n \mathbb{E}[X_i]$

Functions of random variables

□ Mean:

$$\begin{aligned}\mathbb{E}[Z] \simeq g(\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]) + \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 g}{\partial X_i^2} \text{var}[X_i] \\ + \sum_{\text{all pairs } (i,j)} \frac{\partial^2 g}{\partial X_i \partial X_j} \text{cov}(X_i, X_j)\end{aligned}$$

$$\rightarrow \text{Particular case: } Z = X_1 + \dots + X_n \Rightarrow \mathbb{E}[Z] = \sum_{i=1}^n \mathbb{E}[X_i]$$

□ Variance:

$$\text{var}[Z] \simeq \sum_{i=1}^n \sum_{j=1}^n \frac{\partial g}{\partial X_i} \frac{\partial g}{\partial X_j} \text{cov}(X_i, X_j)$$

$$\rightarrow \text{Particular case: iid} \Rightarrow \text{var}[Z] \simeq \sum_{i=1}^n \left(\frac{\partial g}{\partial X_i} \right)^2 \text{var}[X_i]$$

ps: all derivatives evaluated at $\{X_i\} = \{\mathbb{E}[X_i]\}$

Exercice

Prove the following equalities:

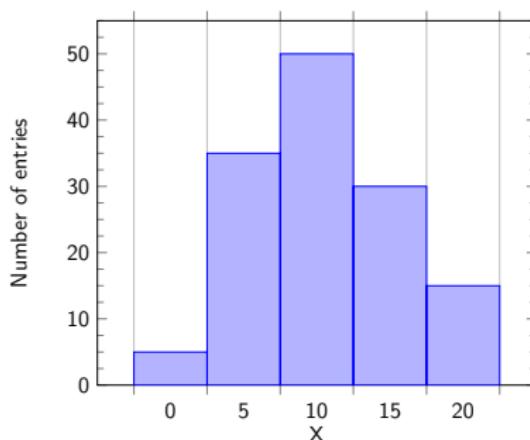
- $\mathbb{E}[M] = \mathbb{E}[X]$
- $\text{var}[M] = \frac{\text{var}[X]}{n}$
- $\mathbb{E}[S_{n-1}^2] = \text{var}[X]$

On the importance of the $\text{var}[M] = \frac{\text{var}[X]}{n}$ relation

- In terms of standard deviations: $\sigma[M] = \frac{\sigma[X]}{\sqrt{n}}$
- Sample mean = good estimate of expectation value when sample size large ($\sigma[M] \xrightarrow{n \rightarrow \infty} 0$)
 - Also true for higher order moments
- Monte Carlo method

Histograms

- Sometimes convenient to transform the sample into a **histogram**
- How to build a histogram ?**
 - ① Define intervals (**bins**) for the r.v.
 - ② Count the number of entries in each interval



- **Advantages:**
 - Good estimate of the pdf when n large
 - Nice way to visualize samples
 - Smaller than original sample

- **Drawback:** some information in the sample is lost

- **Terminology:**
 - Histogram = **binned sample**
 - Original sample = **unbinned sample**

A histogram is a statistical object → What is its law of probability ?

Binomial and multinomial distributions

□ **Binomial:**

$$P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Properties:

- $\mathbb{E}[k] = np$
- $\text{var}[k] = np(1-p)$

□ **Multinomial:**

$$P(n_1, \dots, n_m; n, p_1, \dots, p_m) = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}$$

where

- m : number of possible results in a trial
- n_i : number of results of type i ($i \in [1; m]$), $\sum n_i = n$
- p_i : probability that result in a trial is of type i

Properties:

- $\mathbb{E}[n_i] = np_i$
- $\text{var}[n_i] = np_i(1-p_i)$
- $\text{cov}(n_i, n_j) = -np_i p_j$

Histograms: law of probability (n fixed)

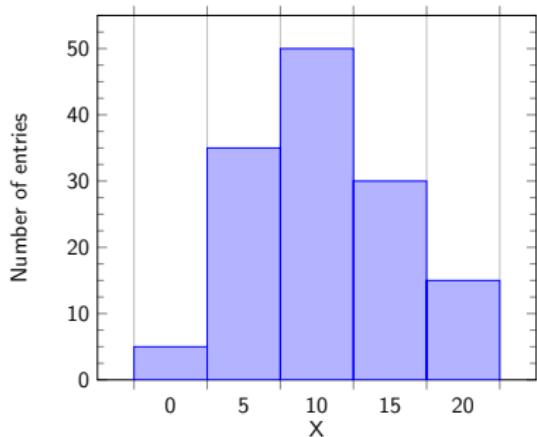
- Filling a histogram from a fixed size sample is a **multinomial process**

$$\Rightarrow P(\text{histogram}) = \frac{n!}{\prod_i n_i!} \prod_i p_i^{n_i}$$

where

- n_i = content of bin i
- p_i = probability to be in bin i

$$p_i = \int_{x \in \text{bin } i} f_X(x) dx$$



- **Note:** the bins are not independent (having more events in one bin necessarily implies having less in some other bins, as n is fixed)

→ What if n is not fixed ?

Histograms: law of probability (n not fixed)

- The probability changes as follows:

$$P(n_1, \dots, n_m; n, p_1, \dots, p_m) \rightarrow P(n, n_1, \dots, n_m; v, p_1, \dots, p_m)$$

- Remarks:

- The sample size is $m+1$ (n is no more a parameter, it becomes a variable)
- v denotes the parameters of the distribution of n

- What is $P(n, n_1, \dots, n_m)$?

$$\rightarrow P(n, n_1, \dots, n_m) = \underbrace{P(n_1, \dots, n_m; n)}_{\text{multinomial}} \times P(n)$$

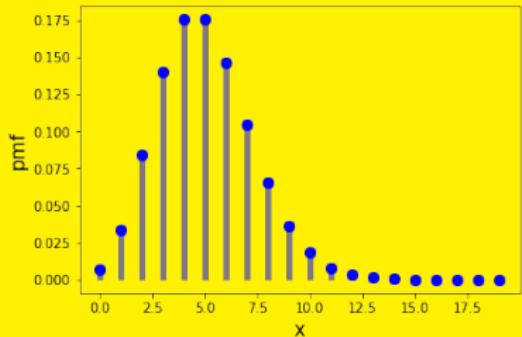
- Often n is a Poisson random variable → See next slide

Poisson distribution

$$P(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}$$

Properties:

- $\mathbb{E}[n] = \nu$
- $\text{var}[n] = \nu$



- Poisson = limit of binomial when $n \rightarrow \infty$ and $p \rightarrow 0$

$$\binom{n}{k} p^k (1-p)^{n-k} \xrightarrow[n \rightarrow \infty]{p \rightarrow 0} \frac{\nu^n}{n!} e^{-\nu} \quad \text{with} \quad \nu = np$$

- Poisson aka the "law of rare events"

Histograms: law of probability (n not fixed)

□ What is $P(n, n_1, \dots, n_m)$ when $n \sim \text{Poisson}$?

$$\begin{aligned} \rightarrow P(n, n_1, \dots, n_m) &= \underbrace{P(n_1, \dots, n_m; n)}_{\text{multinomial}} \times P(n) \\ &= \frac{n!}{\prod_i n_i!} \prod_i p_i^{n_i} \times \frac{v^n}{n!} e^{-v} \\ &= \dots \\ &= \dots \end{aligned}$$

$$\Rightarrow P(n, n_1, \dots, n_m) = \prod_{i=1}^m \frac{v_i^{n_i}}{n_i!} e^{-v_i}$$

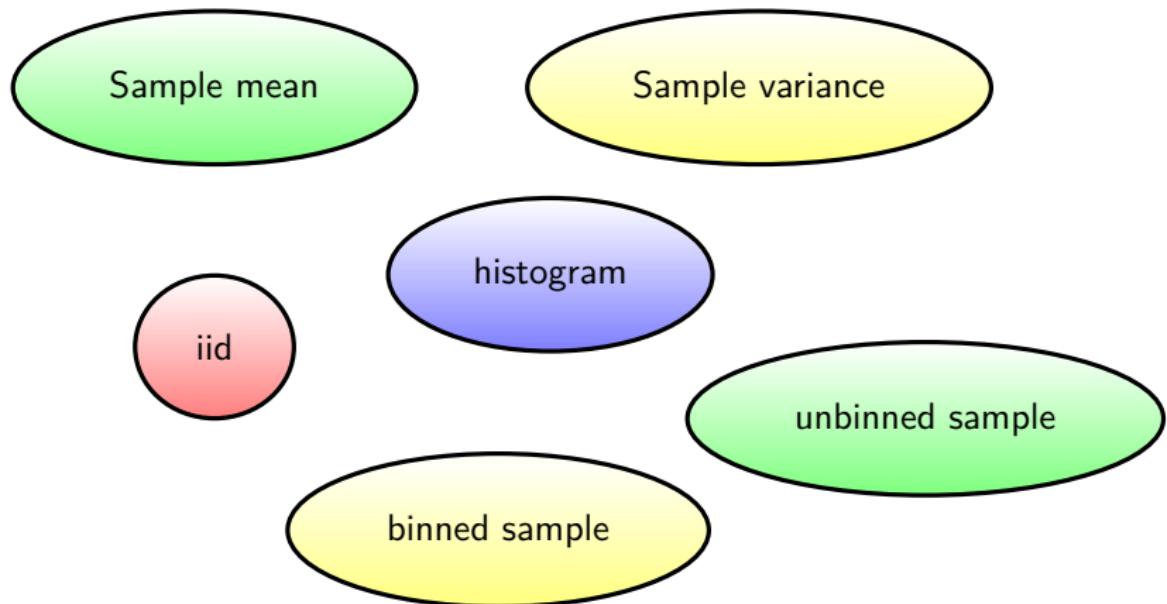
where

- $v_i = v \times p_i$
- $p_i = \int_{x \in \text{bin } i} f_X(x) dx$

Histograms: summary

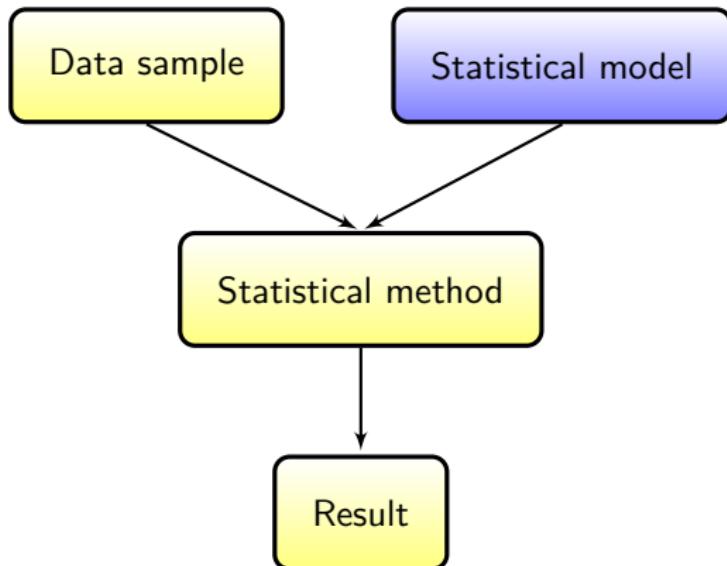
- **n fixed:** histogram \sim multinomial
- **n \sim Poisson:** histogram \sim product of poisson in each bin

Key words/concepts



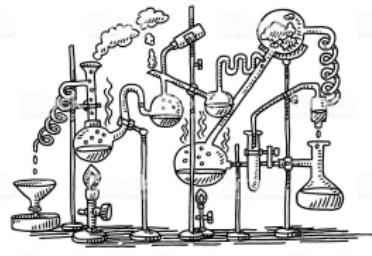
Chap. 2 Statistical model

Basic blocks of statistical reasoning

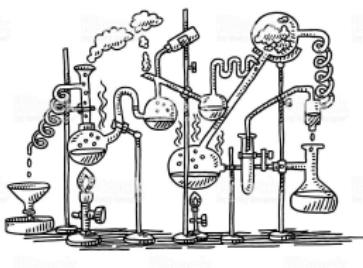


- **Statistical model** = mathematical expression describing your observations
- **Crucial object:**
 - Your measurement into equations

- **Statistical model** = mathematical expression describing your observations
- **Crucial object:**
 - Your measurement into equations



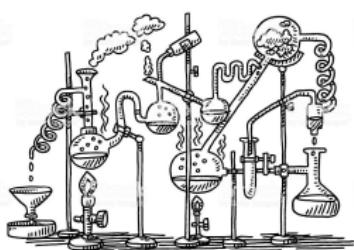
- Statistical model** = mathematical expression describing your observations
 - Crucial object:**
 - Your measurement into equations



Statistical model

- Statistical model** = mathematical expression describing your observations
- Crucial object:**

- Your measurement into equations

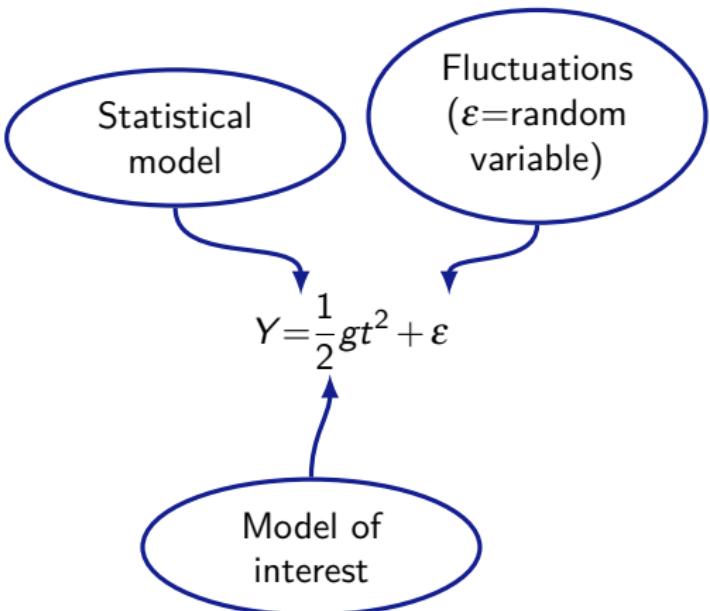
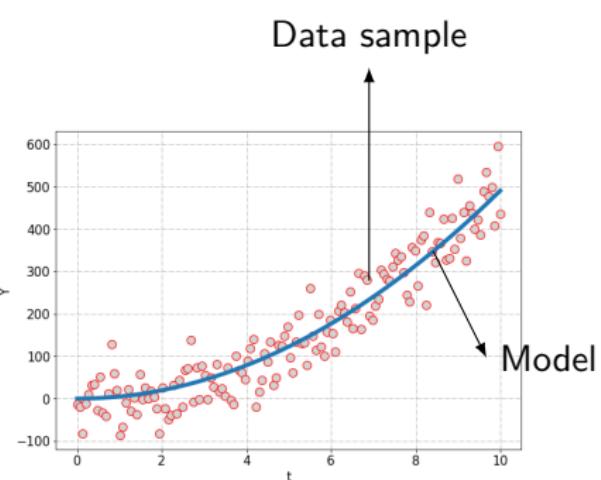


$$\begin{aligned} & \text{Pythagorean Theorem: } a^2 + b^2 = c^2 \\ & \text{Trigonometric Identity: } \sin^2 \theta + \cos^2 \theta = 1 \\ & \text{Quadratic Formula: } x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \\ & \text{Area of a Triangle: } P = \frac{1}{2} ab \sin Y \\ & \text{Area of a Circle: } P = \pi r^2 \\ & \text{Volume of a Sphere: } V = \frac{4}{3} \pi r^3 \\ & \text{Surface Area of a Sphere: } S = 4 \pi r^2 \\ & \text{Volume of a Cube: } V = a^3 \\ & \text{Surface Area of a Cube: } S = 6a^2 \\ & \text{Pythagorean Theorem: } a^2 + b^2 = c^2 \\ & \text{Trigonometric Identity: } \sin^2 \theta + \cos^2 \theta = 1 \\ & \text{Quadratic Formula: } x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \\ & \text{Area of a Triangle: } P = \frac{1}{2} ab \sin Y \\ & \text{Area of a Circle: } P = \pi r^2 \\ & \text{Volume of a Sphere: } V = \frac{4}{3} \pi r^3 \\ & \text{Surface Area of a Sphere: } S = 4 \pi r^2 \\ & \text{Volume of a Cube: } V = a^3 \\ & \text{Surface Area of a Cube: } S = 6a^2 \end{aligned}$$

- Used to derive your conclusions
 - If you mess up with your equations, you mess up with your conclusions

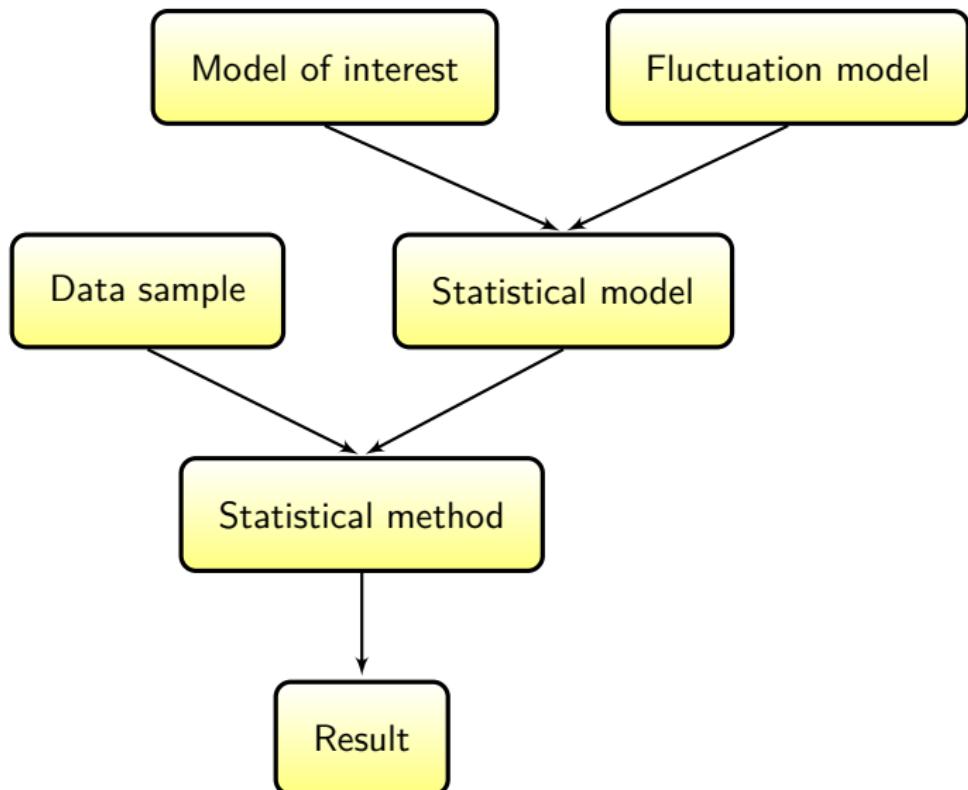
- In realistic cases, can be very difficult to write stat. model
 - **No unique “true” stat. model**
 - A stat. model is always to some extend approximate
- Should always spend some time thinking about what the best stat. model for your problem is
- Always good to **check the robustness** of your conclusions with respect to the stat. model

Example of statistical model



Stat. model = model of interest + fluctuation model (describes fluctuations inherent to measurement)

Basic blocks of statistical reasoning



Statistical model ingredients

□ Main ingredients:

- **Observables**: outcome of measurement (set of measured values of observables = sample)
- **Parameters**:
 - **Parameters of interest**: quantities you want to learn something about
 - **Nuisance parameters**: parameters you're not interested in

Statistical model ingredients

□ Main ingredients:

- **Observables**: outcome of measurement (set of measured values of observables = sample)
- **Parameters**:
 - **Parameters of interest**: quantities you want to learn something about
 - **Nuisance parameters**: parameters you're not interested in

□ For problems where correlation is under study:

- **Dependent variable**: one or more observables ("y-axis variable").
- **Explanatory variable**: variable on which dependent variable depends ("x-axis variable"). Can be an observable or not.

Simple examples

- Statistical model describing free fall measurement:

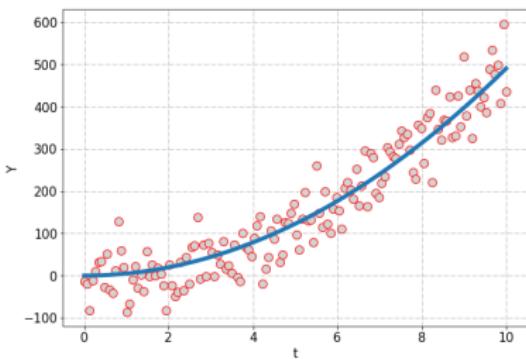
$$Y = \frac{1}{2}gt^2 + \varepsilon$$

parameter of interest

can depend
on some
nuisance
parameters

observable (dependent variable)

explanatory variable



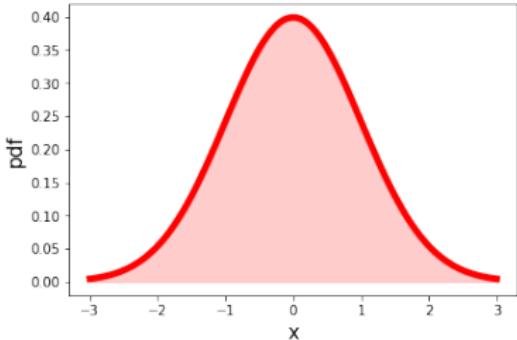
Simple examples

- Statistical model for simple gaussian measurement:

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Diagram illustrating the components of the Gaussian probability density function:

- The term $\exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$ is labeled "observable".
- The parameters μ and σ are labeled "parameter".



- Depending on what you want to do we can have:

- μ is the parameter of interest and σ is a nuisance parameter
- μ is a nuisance parameter and σ is the parameter of interest
- μ and σ are both parameters of interest (if you want to measure both simultaneously)

Write the statistical models for the two following measurements and say what are the observables and parameters

- Measurement 1:** Measurement of the number of successes in n Bernouilli trials
- Measurement 2:** Measurement of the number of Bernouilli trials until the k^{th} success is reached

Statistical model for iid samples

- Suppose you have a sample $X = (X_1, \dots, X_n)$
- The stat. model can be written

$$f_{X_1 \dots X_n}(x_1, \dots, x_n; \theta) = f_{X_1}(x_1 | x_2, \dots, x_n; \theta) \times f_{X_2}(x_2 | x_3, \dots, x_n; \theta) \times \dots \times f_{X_n}(x_n; \theta)$$

- For iid variables:

$$f_{X_1 \dots X_n}(x_1, \dots, x_n; \theta) = f_X(x_1; \theta) \times f_X(x_2; \theta) \times \dots \times f_X(x_n; \theta)$$

$$\Rightarrow f_{X_1 \dots X_n}(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_X(x_i; \theta)$$

Statistical model for iid samples: normal example

- If the X_i 's are iid normal variables:

$$\begin{aligned}\text{stat. model} &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right)\end{aligned}$$

The likelihood function

- When talking about statistical model, the term **likelihood** (or **likelihood function**) is often used
- Notation:**

$$\mathcal{L}(\theta; \mathbf{x})$$

parameters
(can have
multiple-
components)

observables
(can have
multiple-
components)

A diagram illustrating the notation $\mathcal{L}(\theta; \mathbf{x})$. Above the expression, there is a blue arrow pointing from the symbol \mathcal{L} to the word "parameters". Below the expression, there is a red arrow pointing from the symbol \mathbf{x} to the word "observables". To the left of the expression, the text "parameters (can have multiple-components)" is written in blue. To the right of the expression, the text "observables (can have multiple-components)" is written in red.

- Remarks:**
 - Parameters noted first, observables noted last
 - Sometimes notation simplified even more $\mathcal{L}(\theta; \mathbf{x}) = \mathcal{L}(\theta) = \mathcal{L}$
- In general **likelihood** and stat. model are used as synonyms:

$$\mathcal{L}(\theta; \mathbf{x}) = f(\mathbf{x}; \theta)$$

↑
stat. model

The likelihood function

- It's very common to not use the likelihood directly but either one of the following:
 - $\ln \mathcal{L}$
 - $-\ln \mathcal{L}$
 - $\ln \mathcal{L}$ (or $-\ln \mathcal{L}$) with "constant terms" removed
(constant terms = terms not depending on the parameters)



The term likelihood can sometimes be used to denote any of the above quantities (and not the stat. model directly)

Likelihood: gaussian example

- Stat. model:

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{(\sqrt{2\pi\sigma})^n} \exp\left(-\sum_{i=1}^n \frac{(\mathbf{x}_i - \boldsymbol{\mu})^2}{2\sigma^2}\right)$$

- We typically write down and use the following stuff:

Likelihood: gaussian example

- Stat. model:

$$f(\mathbf{x}; \mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}\right)$$

- We typically write down and use the following stuff:

$$\rightarrow \mathcal{L}(\mu, \sigma; \mathbf{x}) = f(\mathbf{x}; \mu, \sigma)$$

Likelihood: gaussian example

- Stat. model:

$$f(\mathbf{x}; \mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}\right)$$

- We typically write down and use the following stuff:

$$\rightarrow \mathcal{L}(\mu, \sigma; \mathbf{x}) = f(\mathbf{x}; \mu, \sigma)$$

$$\rightarrow \ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}$$

Likelihood: gaussian example

- Stat. model:

$$f(\mathbf{x}; \mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}\right)$$

- We typically write down and use the following stuff:

$$\rightarrow \mathcal{L}(\mu, \sigma; \mathbf{x}) = f(\mathbf{x}; \mu, \sigma)$$

$$\rightarrow \ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}$$

$$\rightarrow \ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = -n \ln \sigma - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}$$

Likelihood: gaussian example

- Stat. model:

$$f(\mathbf{x}; \mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}\right)$$

- We typically write down and use the following stuff:

$$\rightarrow \mathcal{L}(\mu, \sigma; \mathbf{x}) = f(\mathbf{x}; \mu, \sigma)$$

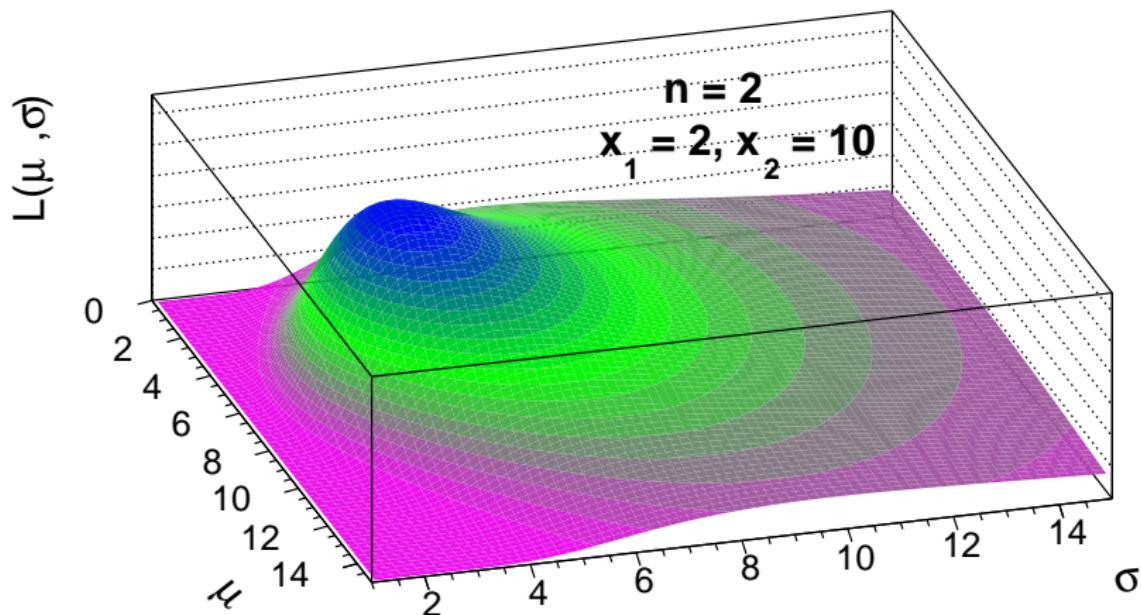
$$\rightarrow \ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}$$

$$\rightarrow \ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = -n \ln \sigma - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}$$

If σ is perfectly known prior to the measurement and the only parameter of interest is μ , you can even write

$$\rightarrow \ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}$$

Likelihood: gaussian example



Suppose you make m independent measurements of the number of successes in n Bernouilli trials (all having the same probability parameter p)

- Write the statistical model
- Suppose the parameter of interest is p
 - Write the log-likelihood with constant terms removed

The extended likelihood function

- Size of the sample n often not constant but follows Poisson distribution: $n \sim \text{Pois}(\nu)$

The extended likelihood function

- Size of the sample n often not constant but follows Poisson distribution: $n \sim \text{Pois}(\nu)$
- In such cases, the likelihood function has to be **extended**:

$$\mathcal{L}_{\text{ext}}(\theta, \nu; x, n) = \frac{\nu^n}{n!} e^{-\nu} \times \mathcal{L}(\theta; x)$$

For iid case:

$$\mathcal{L}_{\text{ext}}(\theta, \nu; x, n) = \frac{\nu^n}{n!} e^{-\nu} \times \prod_{i=1}^n \mathcal{L}(\theta; x_i)$$

Diagram illustrating the components of the extended likelihood function:

- extended likelihood (points to the entire formula)
- extended term (points to the term $\frac{\nu^n}{n!} e^{-\nu}$)
- likelihood (points to the term $\prod_{i=1}^n \mathcal{L}(\theta; x_i)$)



Term likelihood may be used to denote extended likelihood
→ Should be clear from context whether we're talking about likelihood or extended likelihood

Composite samples

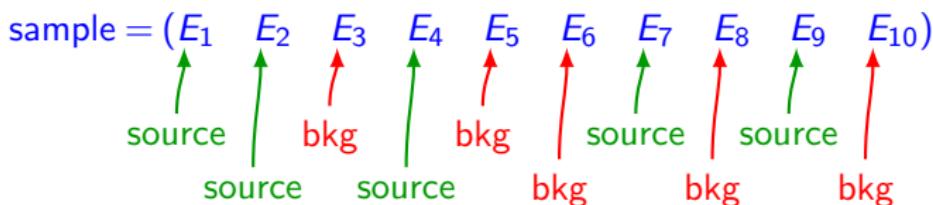
- In realistic cases, samples are often **composite**
- Composite sample** = sample in which events can come from different origins
- Examples:

Composite samples

- In realistic cases, samples are often **composite**
- Composite sample** = sample in which events can come from different origins
- Examples:**
 - Weights in a sample including men and women

Composite samples

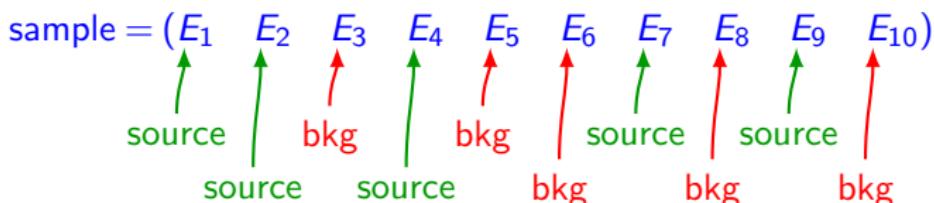
- In realistic cases, samples are often **composite**
- Composite sample** = sample in which events can come from different origins
- Examples:**
 - Weights in a sample including men and women
 - Measurement of radioactive source:



Composite samples

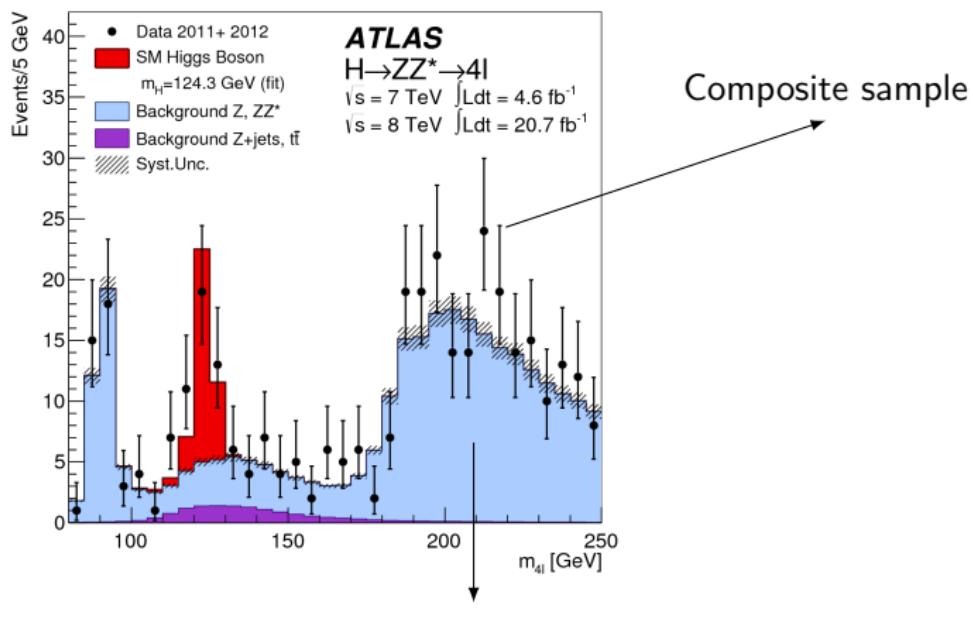
- In realistic cases, samples are often **composite**
- Composite sample** = sample in which events can come from different origins
- Examples:**

- Weights in a sample including men and women
- Measurement of radioactive source:



- Composite samples are said to be made of a **mixture of events**
- Composite samples must be described by **composite stat. models** (or **composite likelihoods**)

Composite sample and model: an realistic example in physics (the Higgs discovery)

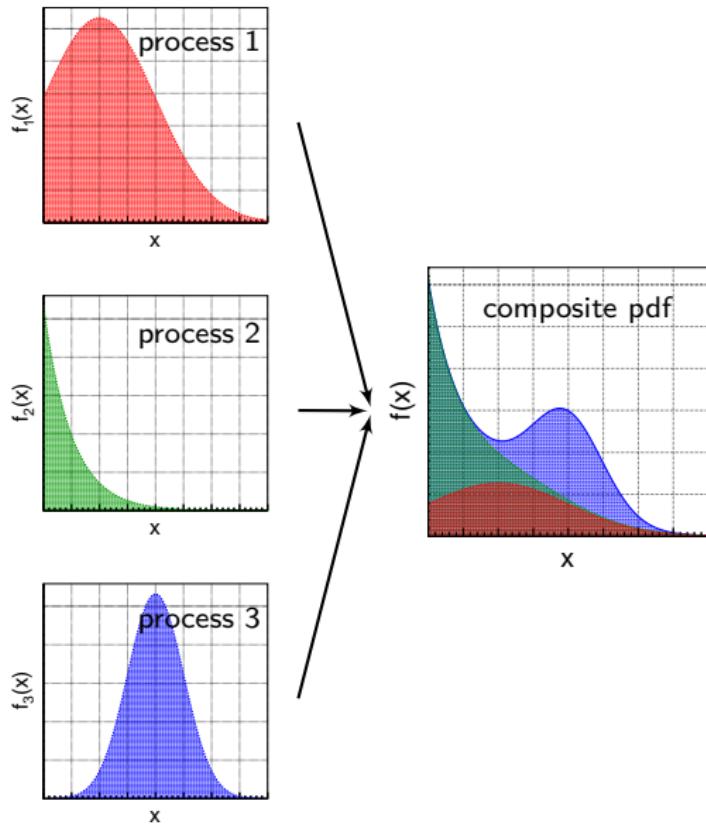


Composite sample

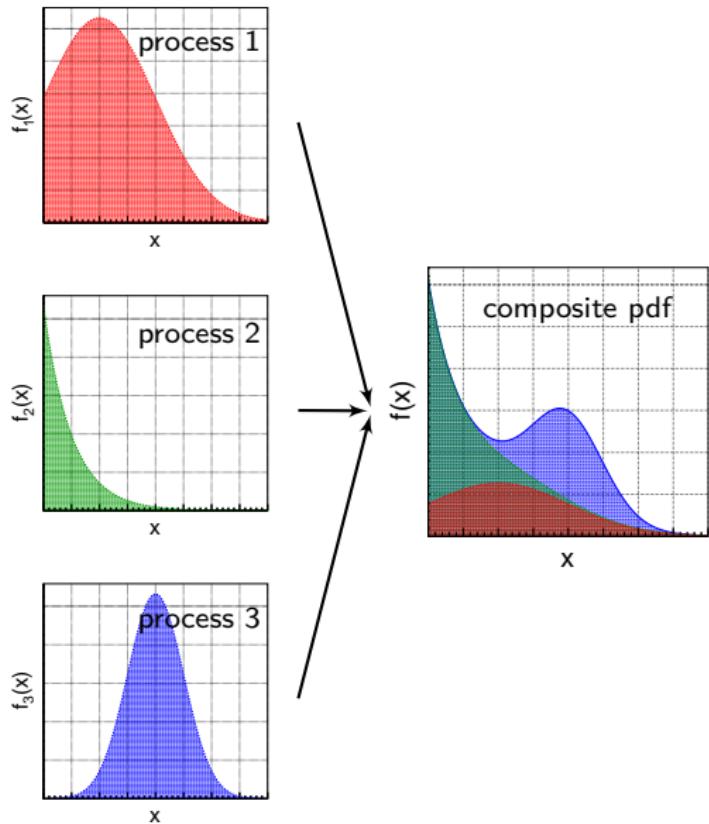
Composite model

(accounts for the various physics processes contributing to the dataset)

Composite model: formalism



Composite model: formalism



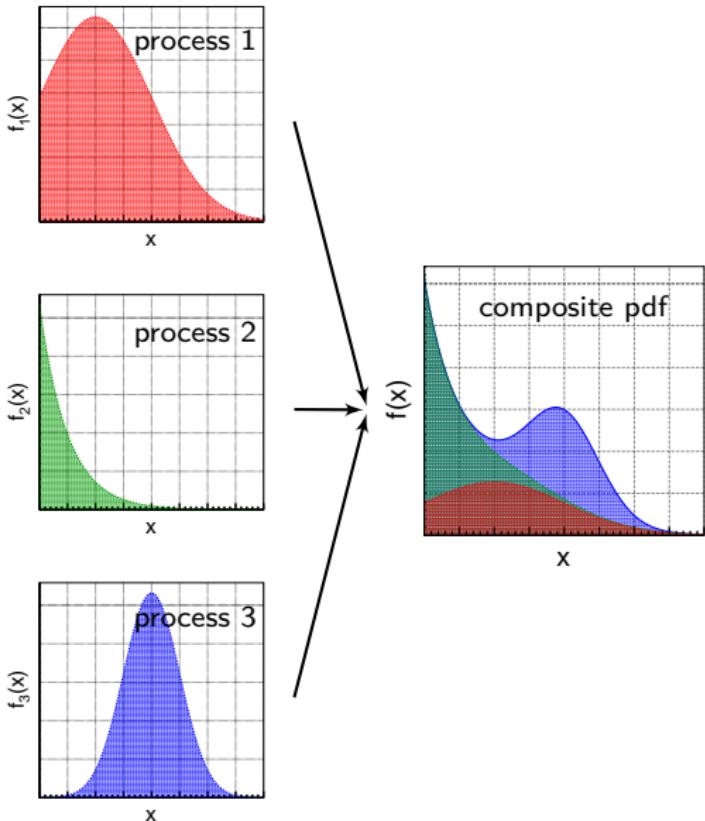
□ Composite pdf:

$$f(\textcolor{red}{x}; \{\mu_p\}, \theta) = \frac{\sum_{p=1}^P \mu_p f_p(\textcolor{red}{x}; \theta)}{\sum_{p=1}^P \mu_p}$$

where:

- p : process index
- μ_p : expected number of elements in sample from process p
- $f_p(x; \theta)$: pdf for process p

Composite model: formalism



□ Composite pdf:

$$f(\textcolor{red}{x}; \{\mu_p\}, \theta) = \frac{\sum_{p=1}^P \mu_p f_p(\textcolor{red}{x}; \theta)}{\sum_{p=1}^P \mu_p}$$

where:

- p : process index
- μ_p : expected number of elements in sample from process p
- $f_p(x; \theta)$: pdf for process p

□ Remark: the μ_p 's are often unknown

→ Determining them can be one of the objective

Composite models: formalism

□ General extended unbinned likelihood:

$$\mathcal{L}_{\text{ext}}(\{\mu_p\}, \theta; \{x_i\}, n) = \underbrace{\frac{(\sum \mu_p)^n}{n!} e^{-\sum \mu_p}}_{\text{extended term}} \times \prod_{i=1}^n \frac{\sum_{p=1}^P \mu_p f_p(x_i; \theta)}{\sum_{p=1}^P \mu_p}$$

↑
Assuming the
 x_i 's are iid

Composite models: formalism

- General extended unbinned likelihood:

$$\mathcal{L}_{\text{ext}}(\{\mu_p\}, \theta; \{x_i\}, n) = \underbrace{\frac{(\sum \mu_p)^n}{n!} e^{-\sum \mu_p}}_{\text{extended term}} \times \prod_{i=1}^n \frac{\sum_{p=1}^P \mu_p f_p(x_i; \theta)}{\sum_{p=1}^P \mu_p}$$

Assuming the
 x_i 's are iid

- Often written as (prove it !):

$$\ln \mathcal{L}_{\text{ext}} = \sum_i \ln \left[\sum_p \mu_p f_p(x_i; \theta) \right] - \sum_p \mu_p$$

Composite models: formalism

□ General extended unbinned likelihood:

$$\mathcal{L}_{\text{ext}}(\{\mu_p\}, \theta; \{x_i\}, n) = \underbrace{\frac{(\sum \mu_p)^n}{n!} e^{-\sum \mu_p}}_{\text{extended term}} \times \prod_{i=1}^n \frac{\sum_{p=1}^P \mu_p f_p(x_i; \theta)}{\sum_{p=1}^P \mu_p}$$

Assuming the
 x_i 's are iid

□ Often written as (prove it !):

$$\ln \mathcal{L}_{\text{ext}} = \sum_i \ln \left[\sum_p \mu_p f_p(x_i; \theta) \right] - \sum_p \mu_p$$

□ Remarks:

- If n fixed \rightarrow remove extended terms
- "Unbinned" because it is constructed from the unbinned dataset
- We can in a similar manner determine the binned likelihood (any guess what it looks like ?)

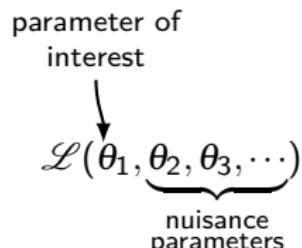
Treatment of nuisance parameters

- Suppose your statistical model has one parameter of interest and other parameters (nuisance parameters) you don't care about:

$$\begin{array}{c} \text{parameter of} \\ \text{interest} \\ \downarrow \\ \mathcal{L}(\theta_1, \underbrace{\theta_2, \theta_3, \dots}_{\text{nuisance}}) \\ \text{parameters} \end{array}$$

Treatment of nuisance parameters

- Suppose your statistical model has one parameter of interest and other parameters (nuisance parameters) you don't care about:

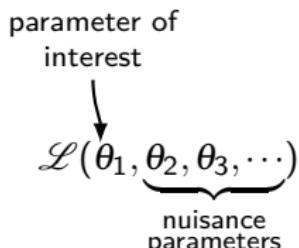


- Example:** gaussian case where we only care about the mean

$$\ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = -n \ln \sigma - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2} \quad \text{with} \quad \begin{cases} \mu \leftrightarrow \theta_1 \\ \sigma \leftrightarrow \theta_2 \end{cases}$$

Treatment of nuisance parameters

- Suppose your statistical model has one parameter of interest and other parameters (nuisance parameters) you don't care about:



- **Example:** gaussian case where we only care about the mean

$$\ln \mathcal{L}(\mu, \sigma; \mathbf{x}) = -n \ln \sigma - \sum_{i=1}^n \frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2} \quad \text{with} \quad \begin{cases} \mu \leftrightarrow \theta_1 \\ \sigma \leftrightarrow \theta_2 \end{cases}$$

- Two cases must be distinguished:
 - ① **Nuisance parameters perfectly known:** formalism presented previously works perfectly fine
 - ② **Nuisance parameters not perfectly known:** formalism presented previously should be extended

Realistic cases: accounting for uncertainties

- **Example:** normal case with μ =parameter of interest and $\sigma = 3 \pm 0.7$
 - How do you account for this unknown σ when determining the value of μ ?
 - Crucial question: depending on how you account for it you might get different values for the parameter of interest

Realistic cases: accounting for uncertainties

- **Example:** normal case with μ =parameter of interest and $\sigma = 3 \pm 0.7$
 - How do you account for this unknown σ when determining the value of μ ?
 - Crucial question: depending on how you account for it you might get different values for the parameter of interest

- **New terms in likelihood:**

$$\mathcal{L}_{\text{full}} = \underbrace{\mathcal{L}(\theta_1, \theta_2, \theta_3, \dots)}_{\text{likelihood w/o uncertainties}} \times \underbrace{g(\theta_2, \theta_3, \dots)}_{\text{new constraint term}}$$

- $\mathcal{L}(\theta_1, \theta_2, \theta_3, \dots)$ should be understood as $\mathcal{L}(\theta_1 | \theta_2, \theta_3, \dots)$
- The exact meaning of the "constraint term" and the way it is treated depends on the statistical approach you use (frequentist or bayesian)

Poissonian process with background (measurement of radioactivity, fraction of spam emails detected by your mailer, ...)

$$\text{Stat. model: } \mathcal{L}(s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

where

- n : number of measured events (observable)
- s : number of signal events (parameter of interest)
- b : number of background events (nuisance parameter)

Poissonian process with background (measurement of radioactivity, fraction of spam emails detected by your mailer, ...)

$$\text{Stat. model: } \mathcal{L}(s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

where

- n : number of measured events (observable)
- s : number of signal events (parameter of interest)
- b : number of background events (nuisance parameter)

□ Various choices for determining b :

- **Method 1:** make a measurement without the source of signal
- **Method 2:** use some discriminating variables to distinguish signal and background
- **Method 3:** make a theoretical calculation
- **Method 4:** look for previous estimates of b in the litterature
- ...

Uncertainties: detailed example

- In all cases you get an estimation of b with an uncertainty
 - This uncertainty impacts the parameter of interest and its uncertainty
(the higher the uncertainty on b , the higher the uncertainty on s)

Uncertainties: detailed example

- In all cases you get an estimation of b with an uncertainty
 - This uncertainty impacts the parameter of interest and its uncertainty (the higher the uncertainty on b , the higher the uncertainty on s)
- Depending on the method, uncertainty on b has different meanings
 - Can be of **statistical nature** if b estimated from a measurement (as in methods 1 or 2)
 - Can be of **systematic nature** if b estimated from theoretical calculations or some other method (as in methods 3 or 4)

Uncertainties: detailed example

- In all cases you get an estimation of b with an uncertainty
 - This uncertainty impacts the parameter of interest and its uncertainty (the higher the uncertainty on b , the higher the uncertainty on s)
- Depending on the method, uncertainty on b has different meanings
 - Can be of **statistical nature** if b estimated from a measurement (as in methods 1 or 2)
 - Can be of **systematic nature** if b estimated from theoretical calculations or some other method (as in methods 3 or 4)
- **In general:** always many ways to estimate things
 - Not equivalent, don't have the same statistical meaning
 - Often not easy to decide which one is best, but at the end you have to decide and move along (remember, what you're interested in is s , not b !)

Statistical uncertainty vs systematic uncertainty

- **Statistical uncertainties:** reflect error coming from fluctuations in some measurement

Statistical uncertainty vs systematic uncertainty

- **Statistical uncertainties:** reflect error coming from fluctuations in some measurement
 - Can be improved by making measurements more precise (e.g. by increasing the sample size)

Statistical uncertainty vs systematic uncertainty

- **Statistical uncertainties:** reflect error coming from fluctuations in some measurement
 - Can be improved by making measurements more precise (e.g. by increasing the sample size)
 - When considering statistical uncertainties, must distinguish 2 types of measurements:
 - **Main measurement:** aims at determining parameter of interest s (b is a nuisance parameter in this measurement)
 - **Auxiliary measurement:** aims at determining b (in this auxiliary measurement, b is the parameter of interest !)

Statistical uncertainty vs systematic uncertainty

- **Systematic uncertainties:** reflect bias we believe we make when we say that some parameter has such or such value

Statistical uncertainty vs systematic uncertainty

- **Systematic uncertainties:** reflect bias we believe we make when we say that some parameter has such or such value
 - Quantifies our belief on the value of the parameter prior to any measurement

Statistical uncertainty vs systematic uncertainty

- **Systematic uncertainties:** reflect bias we believe we make when we say that some parameter has such or such value
 - Quantifies our belief on the value of the parameter prior to any measurement
 - If at some point we can make measurements helping us determining the value of the parameter, the uncertainty stops being systematic and becomes statistical

What is the full likelihood for method 1 ?

What is the full likelihood for method 1 ?

- **Reminder:** the "incomplete" statistical model is

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

What is the full likelihood for method 1 ?

- **Reminder:** the "incomplete" statistical model is

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

- b determined in an auxiliary poissonian measurement → we can write its likelihood as

scaling parameter auxiliary observable

$$\mathcal{L}_{\text{aux}}(b) = \frac{(\alpha b)^{n_b}}{n_b!} e^{-(\alpha b)}$$

What is the full likelihood for method 1 ?

- **Reminder:** the "incomplete" statistical model is

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

- b determined in an auxiliary poissonian measurement → we can write its likelihood as

scaling parameter auxiliary observable

$$\mathcal{L}_{\text{aux}}(b) = \frac{(\alpha b)^{n_b}}{n_b!} e^{-(\alpha b)}$$

- Full likelihood is the joint stat. model of main and auxiliary measurements:

$$\begin{aligned}\mathcal{L}_{\text{full}} &= \mathcal{L}(s, b) \times \mathcal{L}_{\text{aux}}(b) \\ &= \frac{(s+b)^n}{n!} e^{-(s+b)} \times \frac{(\alpha b)^{n_b}}{n_b!} e^{-(\alpha b)}\end{aligned}$$

Uncertainties: detailed example

$$\mathcal{L}_{\text{full}} = \frac{(s+b)^n}{n!} e^{-(s+b)} \times \frac{(\alpha b)^{n_b}}{n_b!} e^{-(\alpha b)}$$

□ Remarks:

Uncertainties: detailed example

$$\mathcal{L}_{\text{full}} = \frac{(s+b)^n}{n!} e^{-(s+b)} \times \frac{(\alpha b)^{n_b}}{n_b!} e^{-(\alpha b)}$$

□ Remarks:

- This full likelihood has the general form (cf few slides back)

$$\mathcal{L}_{\text{full}} = \underbrace{\mathcal{L}(\theta_1, \theta_2, \theta_3, \dots)}_{\text{likelihood w/o uncertainties}} \times \underbrace{g(\theta_2, \theta_3, \dots)}_{\text{new constraint term}}$$

Uncertainties: detailed example

$$\mathcal{L}_{\text{full}} = \frac{(s+b)^n}{n!} e^{-(s+b)} \times \frac{(\alpha b)^{n_b}}{n_b!} e^{-(\alpha b)}$$

□ Remarks:

- This full likelihood has the general form (cf few slides back)

$$\mathcal{L}_{\text{full}} = \underbrace{\mathcal{L}(\theta_1, \theta_2, \theta_3, \dots)}_{\text{likelihood w/o uncertainties}} \times \underbrace{g(\theta_2, \theta_3, \dots)}_{\text{new constraint term}}$$

- \mathcal{L}_{aux} and thus $\mathcal{L}_{\text{full}}$ depend on an additional nuisance parameter: α
 - Is it known exactly ? If not should account for its uncertainty and write a "full-full" likelihood !

Uncertainties: detailed example

$$\mathcal{L}_{\text{full}} = \frac{(s+b)^n}{n!} e^{-(s+b)} \times \frac{(\alpha b)^{n_b}}{n_b!} e^{-(\alpha b)}$$

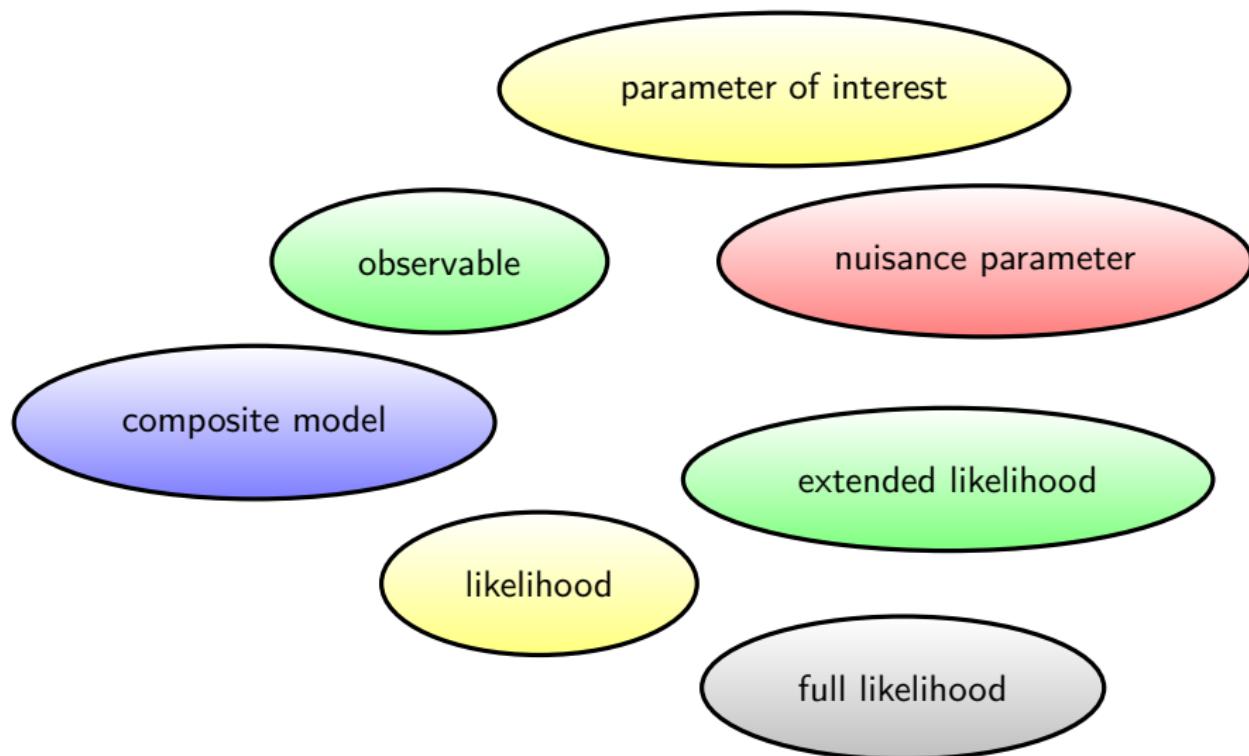
□ Remarks:

- This full likelihood has the general form (cf few slides back)

$$\mathcal{L}_{\text{full}} = \underbrace{\mathcal{L}(\theta_1, \theta_2, \theta_3, \dots)}_{\text{likelihood w/o uncertainties}} \times \underbrace{g(\theta_2, \theta_3, \dots)}_{\text{new constraint term}}$$

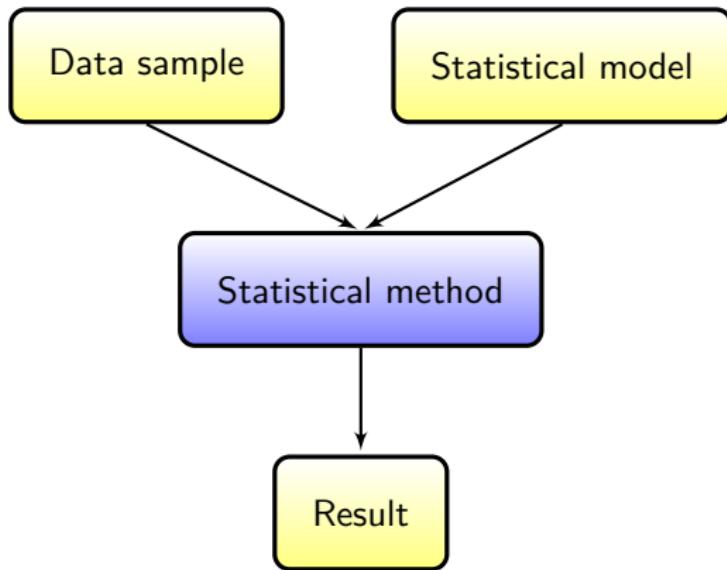
- \mathcal{L}_{aux} and thus $\mathcal{L}_{\text{full}}$ depend on an additional nuisance parameter: α
 - Is it known exactly ? If not should account for its uncertainty and write a "full-full" likelihood !
- We still haven't said anything on how to measure s
 - Now we need to decide which statistical method will, from the full likelihood above, help us measure s !

Key words/concepts



Chap. 3 Inference: generalities

Basic blocks of statistical reasoning



Statistical methods: generalities

- **Statistical methods** are what allow us to gain, from a **sample** and a **statistical model**, some knowledge about the parameter(s) of interest

Statistical methods: generalities

- **Statistical methods** are what allow us to gain, from a **sample** and a **statistical model**, some knowledge about the parameter(s) of interest
 - This knowledge acquisition process is called **inference**
 - The parameter of interest is said to be inferred from the data

Statistical methods: generalities

- **Statistical methods** are what allow us to gain, from a **sample** and a **statistical model**, some knowledge about the parameter(s) of interest
 - This knowledge acquisition process is called **inference**
 - The parameter of interest is said to be inferred from the data
- **Two big schools of thought:**

frequentist

bayesian

Statistical methods: generalities

- **Statistical methods** are what allow us to gain, from a **sample** and a **statistical model**, some knowledge about the parameter(s) of interest

- This knowledge acquisition process is called **inference**
- The parameter of interest is said to be inferred from the data

- **Two big schools of thought:**

frequentist

bayesian

- Depending on the field, one or the other may dominate
- You need to understand both to master the field of statistics
- Concrete methods can fall into one paradigm or the other or be somewhat hybrid

The different interpretations of probability

Probabilities can be interpreted in different ways:

- **Interpretation 1: probability = frequency of occurrence**

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n} \quad \text{where} \quad \left\{ \begin{array}{l} A: \text{some event} \\ n(A): \text{number of times event } A \text{ occurs} \\ n: \text{total number of events} \end{array} \right.$$

Example: dice roll

The different interpretations of probability

Probabilities can be interpreted in different ways:

□ **Interpretation 1: probability = frequency of occurrence**

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n} \quad \text{where} \quad \left\{ \begin{array}{l} A: \text{some event} \\ n(A): \text{number of times event } A \text{ occurs} \\ n: \text{total number of events} \end{array} \right.$$

Example: dice roll

□ **Interpretation 2: probability = degree of belief on occurrence**

- Often used for unique events (for which frequentist calculation impossible)
 - Example: probability that it'll rain tomorrow
- Bayesians use it always (even for non-unique events)

The two big paradigms

Frequentist

- Use frequentist interpretation of probability

Bayesian

- Use probabilities as degree of belief

The two big paradigms

Frequentist

- Use frequentist interpretation of probability
- Model parameters are fixed, the only thing that can have a probability associated to it is the data

$$P(\textcolor{red}{x}; \theta)$$

Bayesian

- Use probabilities as degree of belief

The two big paradigms

Frequentist

- Use frequentist interpretation of probability
- Model parameters are fixed, the only thing that can have a probability associated to it is the data

$$P(\textcolor{red}{x}; \theta)$$

Bayesian

- Use probabilities as degree of belief
- Model parameters have a probability

$$P(\theta; \textcolor{red}{x})$$

- They are not random variable as the x 's (better qualified as **uncertain variables**)
- Even if notation is the same, this probability doesn't have the same frequentist meaning

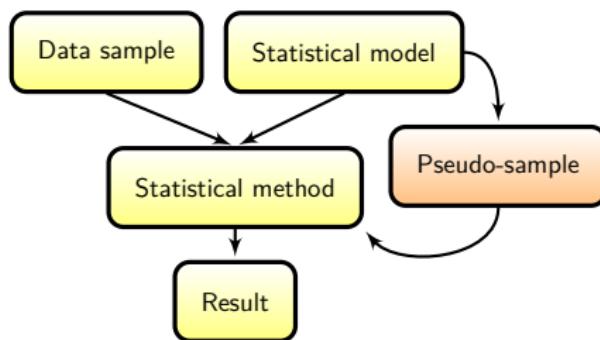
The two big paradigms

Frequentist

- Use frequentist interpretation of probability
- Model parameters are fixed, the only thing that can have a probability associated to it is the data

$$P(\textcolor{red}{x}; \theta)$$

- Inference based on notion of **pseudo-sample** or **pseudo-dataset**



Bayesian

- Use probabilities as degree of belief
- Model parameters have a probability

$$P(\theta; \textcolor{red}{x})$$

- They are not random variable as the x 's (better qualified as **uncertain variables**)
- Even if notation is the same, this probability doesn't have the same frequentist meaning

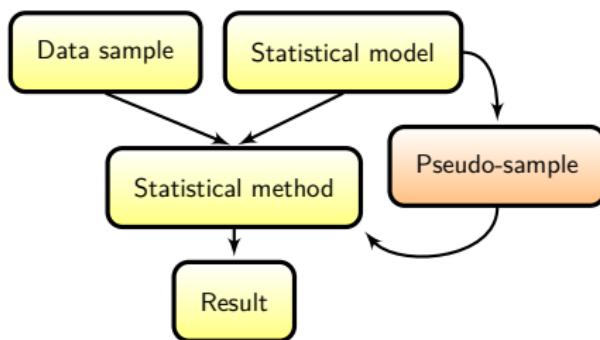
The two big paradigms

Frequentist

- Use frequentist interpretation of probability
- Model parameters are fixed, the only thing that can have a probability associated to it is the data

$$P(\textcolor{red}{x}; \theta)$$

- Inference based on notion of **pseudo-sample** or **pseudo-dataset**



Bayesian

- Use probabilities as degree of belief
- Model parameters have a probability

$$P(\theta; \textcolor{red}{x})$$

- They are not random variable as the x 's (better qualified as **uncertain variables**)
- Even if notation is the same, this probability doesn't have the same frequentist meaning

- Inference of parameter of interest based on this probability

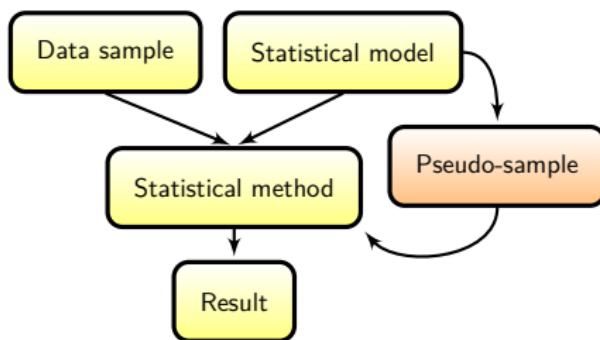
The two big paradigms

Frequentist

- Use frequentist interpretation of probability
- Model parameters are fixed, the only thing that can have a probability associated to it is the data

$$P(\textcolor{red}{x}; \theta)$$

- Inference based on notion of **pseudo-sample** or **pseudo-dataset**



Bayesian

- Use probabilities as degree of belief
- Model parameters have a probability

$$P(\theta; \textcolor{red}{x})$$

- They are not random variable as the x 's (better qualified as **uncertain variables**)
- Even if notation is the same, this probability doesn't have the same frequentist meaning

- Inference of parameter of interest based on this probability
- How is this probability computed ?

$$P(\theta; \textcolor{red}{x}) = \frac{P(\textcolor{red}{x}; \theta)P(\theta)}{P(\textcolor{red}{x})} \quad (\text{Bayes thm})$$

More on the frequentist approach

- ☐ Fundamental frequentist interrogation (underlying all frequentist approaches):

What data would one get if we repeat the same measurement many times ?

More on the frequentist approach

- Fundamental frequentist interrogation (underlying all frequentist approaches):

What data would one get if we repeat the same measurement many times ?

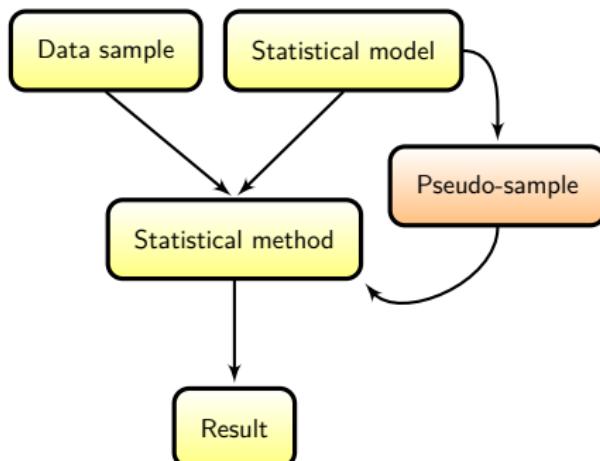
→ Pseudo-datasets are here to answer this interrogation

More on the frequentist approach

- ☐ Fundamental frequentist interrogation (underlying all frequentist approaches):

What data would one get if we repeat the same measurement many times ?

- Pseudo-datasets are here to answer this interrogation
- From the pseudo-datasets and the dataset actually observed one can make inference



Example: incidence of cancer

- From past experience, you know that there are on average 20 new cases of type X cancer each year
- During year 2017, you measure 35 new cases of type X cancer
→ Should we worry about this observation or is it "normal" ?

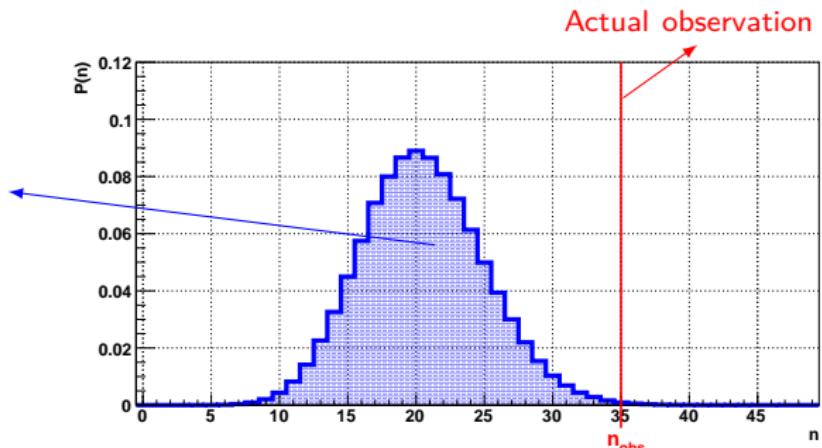
More on the frequentist approach: an example

□ Example: incidence of cancer

- From past experience, you know that there are on average 20 new cases of type X cancer each year
- During year 2017, you measure 35 new cases of type X cancer
→ Should we worry about this observation or is it "normal" ?

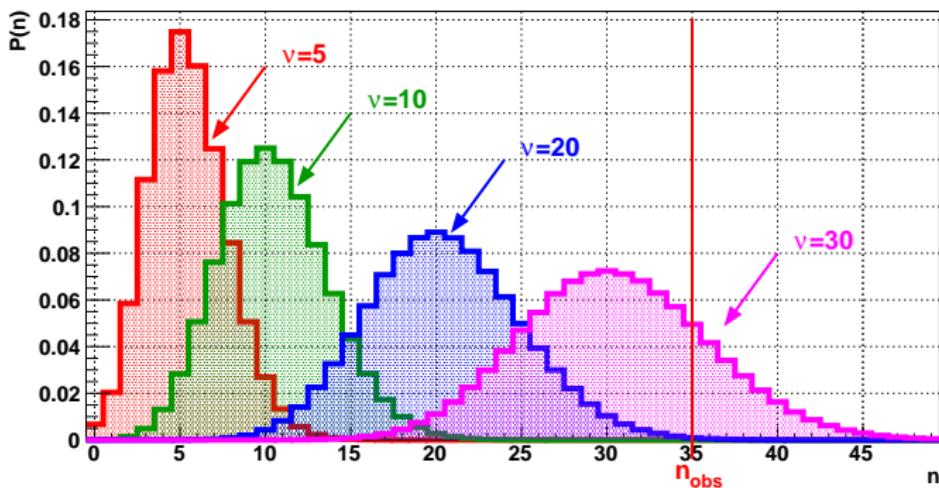
$$\text{Stat. model: } P(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}$$

Pseudo-dataset under the $\nu = 20$ hypothesis (data you would have observed if you repeated the measurement many times with $\nu = 20$)



More on the frequentist approach: an example

- Pseudo-dataset changes if you change the hypothesis (i.e. the value of v)
→ Your conclusions will change accordingly



More on the bayesian approach

□ Inference based on the probability $P(\theta; \mathbf{x})$

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})} \quad (P: \text{ pmf or pdf})$$

Diagram illustrating the components of the Bayesian posterior distribution formula:

- Posterior distribution: $P(\theta; \mathbf{x})$
- Stat. model (likelihood): $P(\mathbf{x}; \theta)$
- Prior distribution: $P(\theta)$
- Normalization constant: $P(\mathbf{x})$

More on the bayesian approach

□ Inference based on the probability $P(\theta; \mathbf{x})$

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})}$$

(P : pmf or pdf)

Posterior distribution

Stat. model (likelihood)

Prior distribution

Normalization constant

□ Getting rid of normalization constant: $P(\theta; \mathbf{x}) \propto P(\mathbf{x}; \theta)P(\theta)$

More on the bayesian approach

□ Inference based on the probability $P(\theta; \mathbf{x})$

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})}$$

(P: pmf or pdf)

The diagram illustrates the Bayesian formula $P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})}$. Above the equation, four labels are positioned: "Posterior distribution" points to $P(\theta; \mathbf{x})$, "Stat. model (likelihood)" points to $P(\mathbf{x}; \theta)$, "Prior distribution" points to $P(\theta)$, and "Normalization constant" points to $P(\mathbf{x})$.

- Getting rid of normalization constant: $P(\theta; \mathbf{x}) \propto P(\mathbf{x}; \theta)P(\theta)$
- In bayesian approach, inference proceeds as follows:

More on the bayesian approach

□ Inference based on the probability $P(\theta; \mathbf{x})$

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})} \quad (P: \text{pmf or pdf})$$

Diagram illustrating the components of the Bayesian formula:

- Posterior distribution: $P(\theta; \mathbf{x})$
- Stat. model (likelihood): $P(\mathbf{x}; \theta)$
- Prior distribution: $P(\theta)$
- Normalization constant: $P(\mathbf{x})$

```
graph TD; PD[P(\theta; x)] --> Ptheta["P(\theta)"]; SM["P(x; θ)"] --> Ptheta; PD --> NC["P(x)"]; NC --> NC
```

- Getting rid of normalization constant: $P(\theta; \mathbf{x}) \propto P(\mathbf{x}; \theta)P(\theta)$
- In bayesian approach, inference proceeds as follows:
 - ① Write stat. model

More on the bayesian approach

□ Inference based on the probability $P(\theta; \mathbf{x})$

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})}$$

(P: pmf or pdf)

The diagram illustrates the Bayesian formula $P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})}$. It shows the components: Posterior distribution, Stat. model (likelihood), Prior distribution, and Normalization constant. Arrows indicate the flow from the prior and likelihood into the numerator, and from the normalization constant into the denominator.

- Getting rid of normalization constant: $P(\theta; \mathbf{x}) \propto P(\mathbf{x}; \theta)P(\theta)$
- In bayesian approach, inference proceeds as follows:
 - ① Write stat. model
 - ② Find out what the best prior is

More on the bayesian approach

□ Inference based on the probability $P(\theta; \mathbf{x})$

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})} \quad (P: \text{ pmf or pdf})$$

Posterior distribution Stat. model (likelihood) Prior distribution
↓ ↓ ↓
Normalization constant

The diagram illustrates the Bayesian formula $P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})}$. It features four labels above the equation: "Posterior distribution" pointing to $P(\theta; \mathbf{x})$, "Stat. model (likelihood)" pointing to $P(\mathbf{x}; \theta)$, "Prior distribution" pointing to $P(\theta)$, and "Normalization constant" pointing to $P(\mathbf{x})$.

- Getting rid of normalization constant: $P(\theta; \mathbf{x}) \propto P(\mathbf{x}; \theta)P(\theta)$
- In bayesian approach, inference proceeds as follows:
 - ① Write stat. model
 - ② Find out what the best prior is
 - ③ Compute posterior (numerical computation for realistic cases in general)

More on the bayesian approach

□ Inference based on the probability $P(\theta; \mathbf{x})$

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})} \quad (P: \text{ pmf or pdf})$$

Posterior distribution Stat. model (likelihood) Prior distribution
↓ ↓ ↓
Normalization constant

- Getting rid of normalization constant: $P(\theta; \mathbf{x}) \propto P(\mathbf{x}; \theta)P(\theta)$
- In bayesian approach, inference proceeds as follows:
 - ① Write stat. model
 - ② Find out what the best prior is
 - ③ Compute posterior (numerical computation for realistic cases in general)
 - ④ Conclude from posterior

Bayes theorem

□ For two arbitrary events A and B :

- $P(A \cap B) = P(A|B)P(B)$
- $P(B \cap A) = P(B|A)P(A)$
- As $P(A \cap B) = P(B \cap A)$, one has $P(A|B)P(B) = P(B|A)P(A)$

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

□ Expliciting the denominator:

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

Exercice

3 suspects are tested with a lie detector to know who committed the murder. The lie detector triggered for only one of them.

- Is the suspect having triggered the detector the culprit ?
- What are the chances that you make a mistake ?

We know that:

- The lie detector detects 70% of the liars
- The lie detector triggers 3% of the time when the truth is said

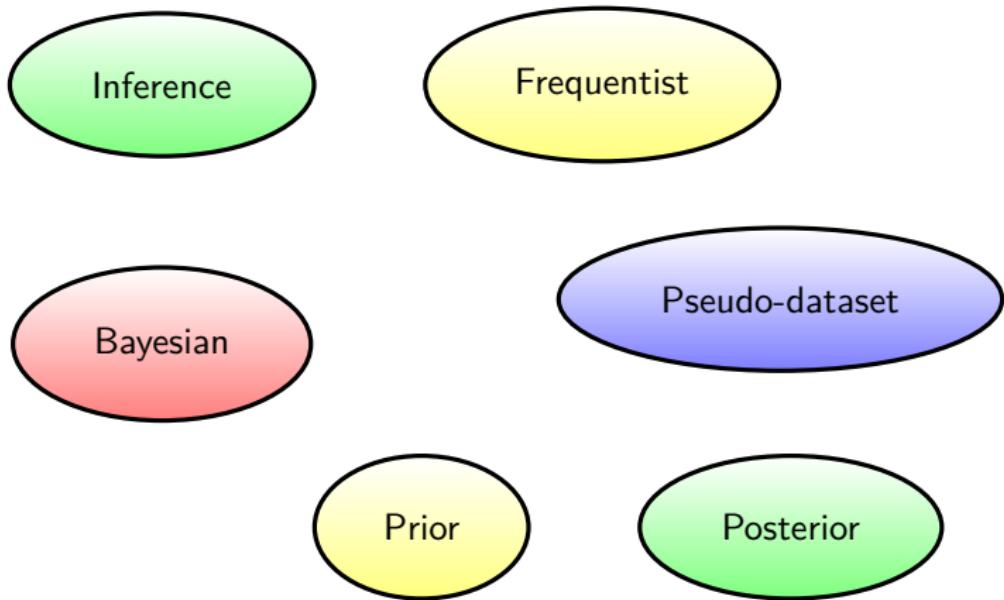
Our bayesian lives

- I'm not feeling so well → Am I sick ? Should I go to the doctor ?
- You are a doctor and you observe some big mass in the neck of your patient → Do you send him to surgery ?
- Tomorrow is my statistics exam, I only partly listened to the lecturer during the courses and didn't spend a minute at home working that topic
 - What should I study in priority in order to maximize the chances to have a good grade ?
- ...

All these are typical "bayesian" decisions:

- You unconsciously evaluate the probability of what will happen if you take or not the decision
- Your evaluation is based on **prior knowledge** (from past experience) and **actual observation**

Key words/concepts



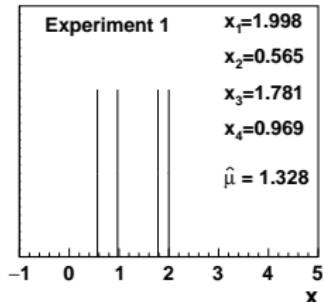
In the rest of this lecture, we will detail statistical methods for

- **Parameter estimation**
- **Confidence interval building**
- **Hypothesis testing**

Chap. 4 Parameter estimation

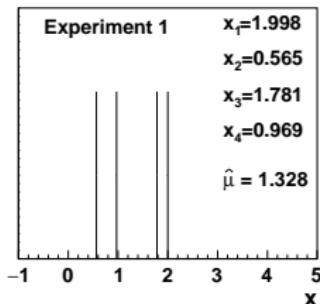
Frequentist approach: preliminary example

- Let X be a gaussian r.v. with $\sigma = 1$ and unknown mean μ
- Suppose we make a measurement providing a sample of 4 values:
 (X_1, X_2, X_3, X_4)



Frequentist approach: preliminary example

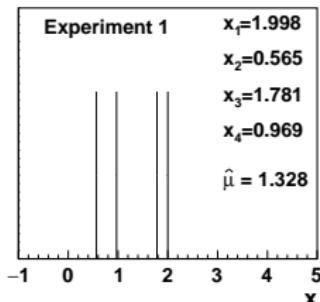
- Let X be a gaussian r.v. with $\sigma = 1$ and unknown mean μ
- Suppose we make a measurement providing a sample of 4 values:
 (X_1, X_2, X_3, X_4)



- Problem:** How do we determine μ from these X_i 's ?

Frequentist approach: preliminary example

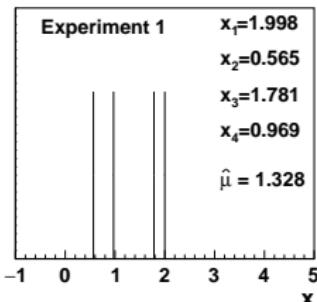
- Let X be a gaussian r.v. with $\sigma = 1$ and unknown mean μ
- Suppose we make a measurement providing a sample of 4 values: (X_1, X_2, X_3, X_4)



- **Problem:** How do we determine μ from these X_i 's ?
- **Possible solution:** take the sample mean of the X_i 's: $M = \frac{1}{4} \sum_{i=1}^4 X_i$

Frequentist approach: preliminary example

- Let X be a gaussian r.v. with $\sigma = 1$ and unknown mean μ
- Suppose we make a measurement providing a sample of 4 values: (X_1, X_2, X_3, X_4)



- Problem:** How do we determine μ from these X_i 's ?
- Possible solution:** take the sample mean of the X_i 's: $M = \frac{1}{4} \sum_{i=1}^4 X_i$
- Doing this, we hope we get a value that is close to the true value μ . If we believe it is the case we can use in future works

$$\mu \simeq M$$

Frequentist approach

- Previous example illustrates **frequentist workflow** for parameter estimation:

Frequentist approach

- Previous example illustrates **frequentist workflow** for parameter estimation:
 - ➊ Determine what parameter you want to measure (θ)

Frequentist approach

- Previous example illustrates **frequentist workflow** for parameter estimation:
 - ① Determine what parameter you want to measure (θ)
 - ② Make a measurement that is "sensitive" to this parameter

Frequentist approach

- Previous example illustrates **frequentist workflow** for parameter estimation:
 - ① Determine what parameter you want to measure (θ)
 - ② Make a measurement that is "sensitive" to this parameter
 - ③ Find an empirical quantity (i.e. a function of the sample) that you think gives a good estimate of θ
 - This function is called an **estimator**
 - Notation: $\hat{\theta}$ (in previous example: $\hat{\mu} = M$)

- Previous example illustrates **frequentist workflow** for parameter estimation:
 - ① Determine what parameter you want to measure (θ)
 - ② Make a measurement that is "sensitive" to this parameter
 - ③ Find an empirical quantity (i.e. a function of the sample) that you think gives a good estimate of θ
 - This function is called an **estimator**
 - Notation: $\hat{\theta}$ (in previous example: $\hat{\mu} = M$)
- **Remarks**
 - $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$
 - Parameter estimation often called "**fit**" (model parameters are said to be fitted to the data)

- Questions to be addressed:
 - **Question 1:** What does it mean when we say that an estimate is "close" to the true value ($\hat{\theta} \simeq \theta$) ?
 - **Question 2:** How are estimators determined in general ?

Question 1: What does $\hat{\theta} \simeq \theta$ mean ?

It means nothing !

- θ is not known
- Estimators are functions of the sample, so they are random variables and it means nothing to say

random variable \simeq constant

Question 1: What does $\hat{\theta} \simeq \theta$ mean ?

- It means nothing !**
 - θ is not known
 - Estimators are functions of the sample, so they are random variables and it means nothing to say
random variable \simeq constant
- In order to determine whether an estimator is good or not, need to get back to fundamental frequentist interrogation: **what would one get if we repeat the same measurement many times ?**

Question 1: What does $\hat{\theta} \simeq \theta$ mean ?

- It means nothing !**
 - θ is not known
 - Estimators are functions of the sample, so they are random variables and it means nothing to say
random variable \simeq constant
- In order to determine whether an estimator is good or not, need to get back to fundamental frequentist interrogation: **what would one get if we repeat the same measurement many times ?**
 - We would get a distribution of $\hat{\theta}$ characterized by a mean, a variance, ...

Question 1: What does $\hat{\theta} \simeq \theta$ mean ?

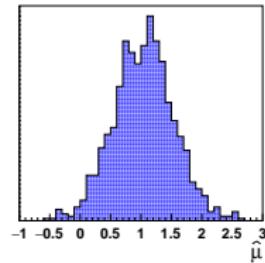
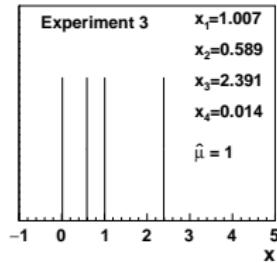
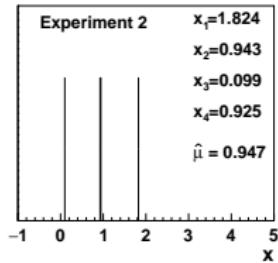
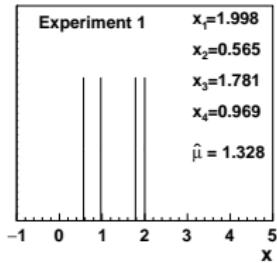
It means nothing !

- θ is not known
- Estimators are functions of the sample, so they are random variables and it means nothing to say

random variable \simeq constant

In order to determine whether an estimator is good or not, need to get back to fundamental frequentist interrogation: **what would one get if we repeat the same measurement many times ?**

- We would get a distribution of $\hat{\theta}$ characterized by a mean, a variance, ...



Question 1: What does $\hat{\theta} \simeq \theta$ mean ?

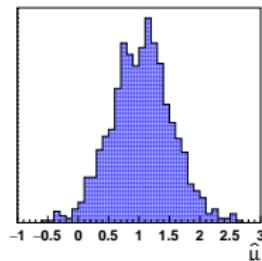
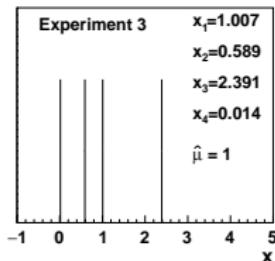
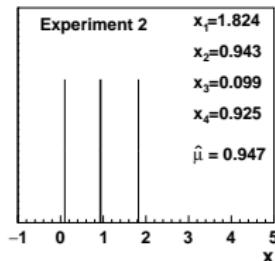
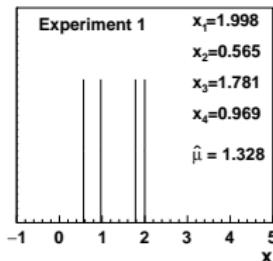
It means nothing !

- θ is not known
- Estimators are functions of the sample, so they are random variables and it means nothing to say

random variable \simeq constant

In order to determine whether an estimator is good or not, need to get back to fundamental frequentist interrogation: **what would one get if we repeat the same measurement many times ?**

- We would get a distribution of $\hat{\theta}$ characterized by a mean, a variance, ...



⇒ Quality of estimators assessed from this distribution

Three main properties:

- **Consistency**
- **Bias**
- **Efficiency**

- **Definition:** an estimator is consistent when it converges in probability towards the true value as $n \rightarrow \infty$

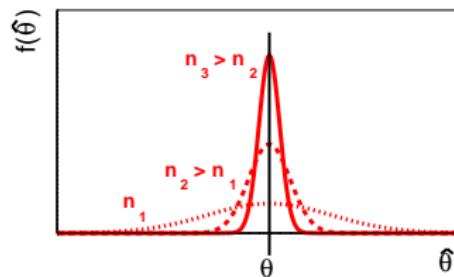
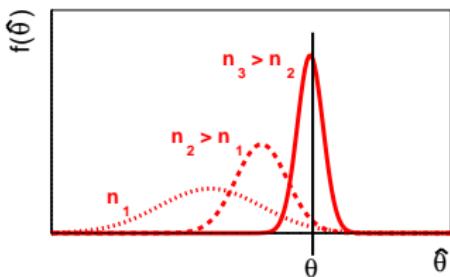
$$P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0$$

Consistency

- **Definition:** an estimator is consistent when it converges in probability towards the true value as $n \rightarrow \infty$

$$P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0$$

- Example of two consistent estimators:



- **Definition:** an estimator is unbiased when its expectation value is equal to the true value (for all n)

$$\mathbb{E} [\hat{\theta}] = \theta$$

- **Definition:** an estimator is unbiased when its expectation value is equal to the true value (for all n)

$$\mathbb{E} [\hat{\theta}] = \theta$$

- The bias of an estimator is defined as

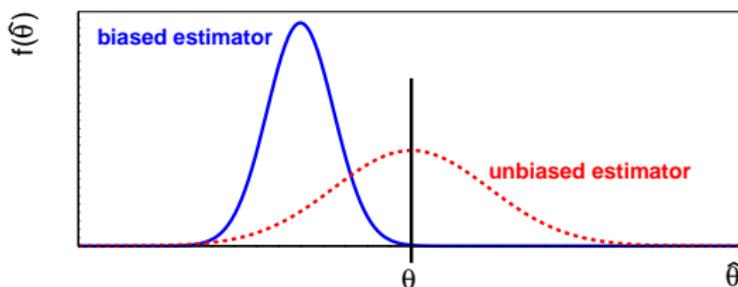
$$b = \mathbb{E} [\hat{\theta}] - \theta$$

Bias

- **Definition:** an estimator is unbiased when its expectation value is equal to the true value (for all n)

$$\mathbb{E} [\hat{\theta}] = \theta$$

- The bias of an estimator is defined as $b = \mathbb{E} [\hat{\theta}] - \theta$
- Example of biased and unbiased estimators:

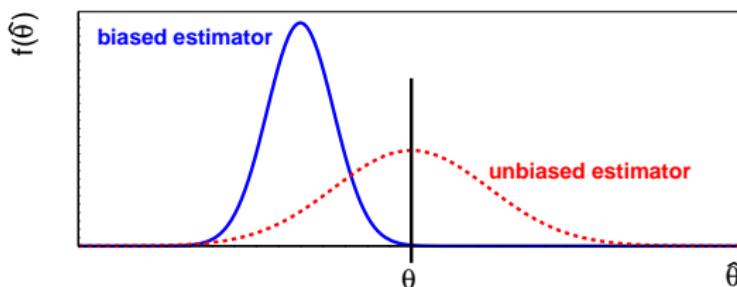


Bias

- **Definition:** an estimator is unbiased when its expectation value is equal to the true value (for all n)

$$\mathbb{E} [\hat{\theta}] = \theta$$

- The bias of an estimator is defined as $b = \mathbb{E} [\hat{\theta}] - \theta$
- Example of biased and unbiased estimators:



- **Remarks:**
 - An estimator can be unbiased and non-consistent
 - A consistent estimator is asymptotically unbiased

Exercice

- What is the sample mean an unbiased estimator of ?
- What is the sample variance S_{n-1}^2 an unbiased estimator of ?
- Show that $(k-1)/(n-1)$ is an unbiased estimator of the probability parameter of the negative binomial distribution where k is the number of successes and n the total number of Bernouilli experiments

- **Definition:** an estimator is efficient when its variance is equal to the RCF bound (RCF=Rao-Cramér-Fréchet)

- **Definition:** an estimator is efficient when its variance is equal to the RCF bound (RCF=Rao-Cramér-Fréchet)
- What is the RCF bound ?

$$\text{var} [\hat{\theta}] \geq \underbrace{\frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{\mathbb{E} \left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta} \right)^2 \right]}}_{\text{RCF bound}}$$

- Definition:** an estimator is efficient when its variance is equal to the RCF bound (RCF=Rao-Cramér-Fréchet)
- What is the RCF bound ?

$$\text{var} [\hat{\theta}] \geq \underbrace{\frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{\mathbb{E} \left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta} \right)^2 \right]}}_{\text{RCF bound}}$$

- Efficient when this inequality turns into an equality
- Also written as

$$\text{var} [\hat{\theta}] \geq -\frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{\mathbb{E} \left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right]}$$

Exercice

- Let k be a random variable distributed according to a binomial distribution with parameters p and n
 - Show that k/n is an efficient estimator of p
- Let $N_i \sim \text{Pois}(\mu)$ for $i \in [1; n]$
 - Show that the sample mean is an efficient estimator of μ

- Ideal estimators are consistent, unbiased and efficiency
- However, not always possible to fulfill all requirements simultaneously
- Should at the very least be consistent

Question 2: How are estimators determined in general ?

- Several statistical methods exist to determine estimators

Question 2: How are estimators determined in general ?

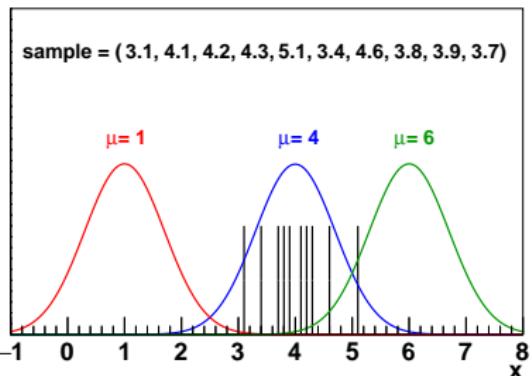
- Several statistical methods exist to determine estimators
- Requirements:**
 - Method should provide estimators with good properties
 - Method should be easy to implement and run fast on computers

Question 2: How are estimators determined in general ?

- Several statistical methods exist to determine estimators
- Requirements:
 - Method should provide estimators with good properties
 - Method should be easy to implement and run fast on computers
- Will describe two of them:
 - **Maximum likelihood method**
 - **Least squares method**

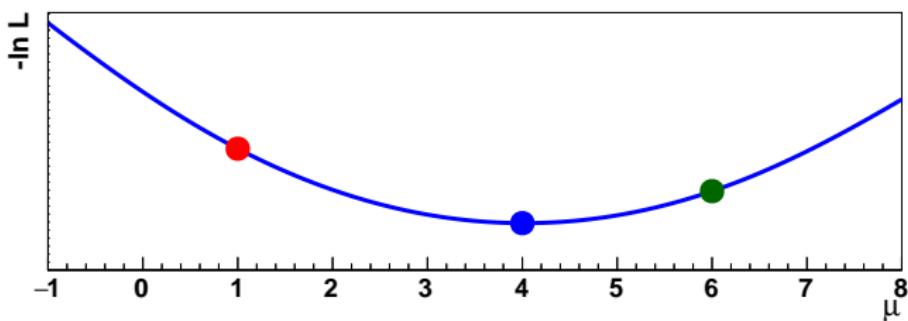
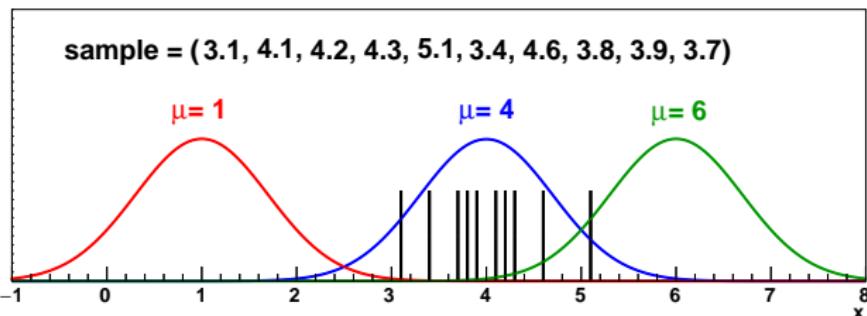
Maximum likelihood (ML) method: introduction

- Let X be a gaussian r.v. with known σ and unknown mean μ
- Suppose we make a measurement for the determination of μ providing 10 values (X_1, \dots, X_{10})



- Of the 3 gaussian distributions represented above, which one fits best the data ? Why ?

Maximum likelihood (ML) method: introduction



Maximum likelihood method: description

- The ML estimator $\hat{\theta}_{\text{ML}}$ is the function that maximizes the likelihood function:

$$\frac{\partial \mathcal{L}}{\partial \theta} \Big|_{\theta=\hat{\theta}_{\text{ML}}} = 0 \quad \text{et} \quad \frac{\partial^2 \mathcal{L}}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{\text{ML}}} < 0$$

Maximum likelihood method: description

- The ML estimator $\hat{\theta}_{\text{ML}}$ is the function that maximizes the likelihood function:

$$\frac{\partial \mathcal{L}}{\partial \theta} \Big|_{\theta=\hat{\theta}_{\text{ML}}} = 0 \quad \text{et} \quad \frac{\partial^2 \mathcal{L}}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{\text{ML}}} < 0$$

Equivalently

$$\frac{\partial(-\ln \mathcal{L})}{\partial \theta} \Big|_{\theta=\hat{\theta}_{\text{ML}}} = 0 \quad \text{et} \quad \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{\text{ML}}} > 0$$

Maximum likelihood method: description

- The ML estimator $\hat{\theta}_{\text{ML}}$ is the function that maximizes the likelihood function:

$$\frac{\partial \mathcal{L}}{\partial \theta} \Big|_{\theta=\hat{\theta}_{\text{ML}}} = 0 \quad \text{et} \quad \frac{\partial^2 \mathcal{L}}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{\text{ML}}} < 0$$

Equivalently

$$\boxed{\frac{\partial(-\ln \mathcal{L})}{\partial \theta} \Big|_{\theta=\hat{\theta}_{\text{ML}}} = 0 \quad \text{et} \quad \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{\text{ML}}} > 0}$$

- Remarks:**

- Often easier to make calculations with the log-likelihood than with the likelihood
- Constant terms in the likelihood do not participate to the inference

- Let X be a gaussian r.v. with known σ and unknown mean μ and suppose we have a sample $\{X_i\}$
 - What is the ML estimator of μ ?
- Suppose you make m independent measurements of the number of successes in n Bernouilli trials (all having the same probability parameter p)
 - What is the ML estimator of p ?

Maximum likelihood method: properties of estimators

The ML method provides estimators with nice properties:

- Functional invariance:**

$$\hat{\tau}_{\text{ML}} = \tau(\hat{\theta}_{\text{ML}})$$

Maximum likelihood method: properties of estimators

The ML method provides estimators with nice properties:

- Functional invariance:**

$$\hat{\tau}_{\text{ML}} = \tau(\hat{\theta}_{\text{ML}})$$

- ML estimators are consistent**

Maximum likelihood method: properties of estimators

The ML method provides estimators with nice properties:

- Functional invariance:**

$$\hat{\tau}_{\text{ML}} = \tau(\hat{\theta}_{\text{ML}})$$

- ML estimators are consistent**
- If an unbiased and efficient estimator exists, then
 - It is found by the ML method
 - It is unique
 - Its variance is

$$\text{var}[\hat{\tau}_{\text{ML}}] = -\frac{(\partial \tau / \partial \theta)^2}{\left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{\text{ML}}}}$$

Maximum likelihood method: properties of estimators

The ML method provides estimators with nice properties:

- Functional invariance:**

$$\hat{\tau}_{\text{ML}} = \tau(\hat{\theta}_{\text{ML}})$$

- ML estimators are consistent**
- If an unbiased and efficient estimator exists, then
 - It is found by the ML method
 - It is unique
 - Its variance is

$$\text{var}[\hat{\tau}_{\text{ML}}] = -\frac{(\partial \tau / \partial \theta)^2}{\left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{\text{ML}}}}$$

- ML estimators are **asymptotically normal**

Maximum likelihood method: properties of estimators

The ML method provides estimators with nice properties:

- Functional invariance:**

$$\hat{\tau}_{\text{ML}} = \tau(\hat{\theta}_{\text{ML}})$$

- ML estimators are consistent**
- If an unbiased and efficient estimator exists, then
 - It is found by the ML method
 - It is unique
 - Its variance is

$$\text{var}[\hat{\tau}_{\text{ML}}] = -\frac{(\partial \tau / \partial \theta)^2}{\left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{\text{ML}}}}$$

- ML estimators are **asymptotically normal**
- ML estimators are **asymptotically efficient**

Maximum likelihood method: confidence interval

- ☐ ML method as described above provides **point estimation**
→ **Is it possible to have an error bar in addition to a single value ?**

Maximum likelihood method: confidence interval

- ☐ ML method as described above provides **point estimation**
→ **Is it possible to have an error bar in addition to a single value ?**

Answer: yes, in some cases, thanks to a "graphical method"

Maximum likelihood method: confidence interval

- ML method as described above provides **point estimation**
→ Is it possible to have an error bar in addition to a single value ?

Answer: yes, in some cases, thanks to a "graphical method"

- **Graphical method:**

$$\begin{aligned}\ln \mathcal{L}(\theta) = & -\ln \mathcal{L}(\hat{\theta}_{ML}) + (\theta - \hat{\theta}_{ML}) \frac{\partial(-\ln \mathcal{L})}{\partial \theta} \Big|_{\theta=\hat{\theta}_{ML}} \\ & + \frac{1}{2}(\theta - \hat{\theta}_{ML})^2 \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{ML}} + \dots\end{aligned}$$

But $\frac{\partial(-\ln \mathcal{L})}{\partial \theta} \Big|_{\theta=\hat{\theta}_{ML}} = 0$, thus

$$\ln \mathcal{L}(\theta) = -\ln \mathcal{L}(\hat{\theta}_{ML}) + \frac{1}{2}(\theta - \hat{\theta}_{ML})^2 \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{ML}} + \dots$$

Maximum likelihood method: graphical method

$$\ln \mathcal{L}(\theta) = -\ln \mathcal{L}(\hat{\theta}_{ML}) + \frac{1}{2}(\theta - \hat{\theta}_{ML})^2 \left. \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{ML}} + \dots$$

Maximum likelihood method: graphical method

$$\ln \mathcal{L}(\theta) = -\ln \mathcal{L}(\hat{\theta}_{ML}) + \frac{1}{2}(\theta - \hat{\theta}_{ML})^2 \left. \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{ML}} + \dots$$

- If estimator unbiased and efficient: $\text{var}[\hat{\theta}_{ML}]^{-1} = -\left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{ML}}$

Maximum likelihood method: graphical method

$$\ln \mathcal{L}(\theta) = -\ln \mathcal{L}(\hat{\theta}_{ML}) + \frac{1}{2}(\theta - \hat{\theta}_{ML})^2 \left. \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{ML}} + \dots$$

- If estimator unbiased and efficient: $\text{var}[\hat{\theta}_{ML}]^{-1} = -\left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{ML}}$

- Thus

$$-\ln \mathcal{L}(\theta) \simeq -\ln \mathcal{L}(\hat{\theta}_{ML}) + \frac{(\theta - \hat{\theta}_{ML})^2}{2\text{var}[\hat{\theta}_{ML}]}$$

Maximum likelihood method: graphical method

$$\ln \mathcal{L}(\theta) = -\ln \mathcal{L}(\hat{\theta}_{ML}) + \frac{1}{2}(\theta - \hat{\theta}_{ML})^2 \left. \frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{ML}} + \dots$$

- If estimator unbiased and efficient: $\text{var}[\hat{\theta}_{ML}]^{-1} = -\left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{ML}}$

- Thus

$$-\ln \mathcal{L}(\theta) \simeq -\ln \mathcal{L}(\hat{\theta}_{ML}) + \frac{(\theta - \hat{\theta}_{ML})^2}{2\text{var}[\hat{\theta}_{ML}]}$$

- Conclusion:

$$\text{if } \theta = \hat{\theta}_{ML} \pm \sqrt{\text{var}[\hat{\theta}_{ML}]}$$

$$\text{then } -\ln \mathcal{L}(\theta) \simeq -\ln \mathcal{L}(\hat{\theta}_{ML}) + 1/2$$

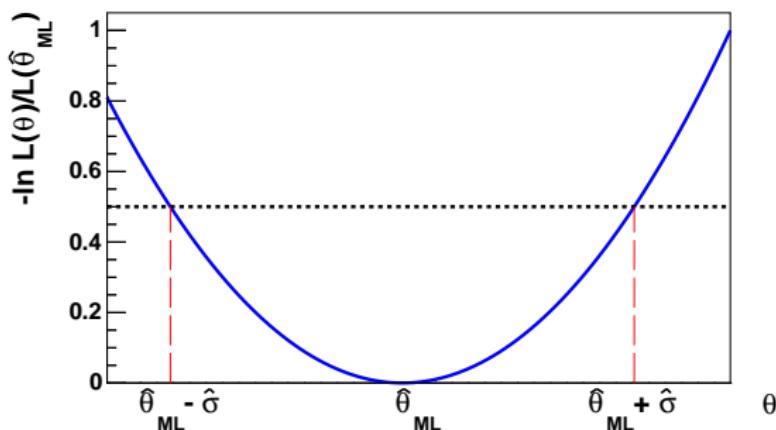
Maximum likelihood method: graphical method

- Conclusion:

$$\text{if } \theta = \hat{\theta}_{\text{ML}} \pm \sqrt{\text{var}[\hat{\theta}_{\text{ML}}]}$$

$$\text{then } -\ln \mathcal{L}(\theta) \simeq -\ln \mathcal{L}(\hat{\theta}_{\text{ML}}) + 1/2$$

- This leads to the following graphical method:



Maximum likelihood method: graphical method

- **Important remark:** method in general approximate
 - Taylor expansion stopped at order 2
 - Hypotheses: unbiasedness and efficiency

Maximum likelihood method: graphical method

- **Important remark:** method in general approximate
 - Taylor expansion stopped at order 2
 - Hypotheses: unbiasedness and efficiency
- **Particular case where it is exact:** normal likelihood with known variance

$$-\ln \mathcal{L}(\mu) = \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2}$$

Maximum likelihood method: graphical method

- **Important remark:** method in general approximate
 - Taylor expansion stopped at order 2
 - Hypotheses: unbiasedness and efficiency
- **Particular case where it is exact:** normal likelihood with known variance

$$-\ln \mathcal{L}(\mu) = \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2}$$

We can show that in this case the following equality holds exactly:

$$-\ln \mathcal{L}(\mu) = -\ln \mathcal{L}(M) + \frac{(\mu - M)^2}{2(\sigma/\sqrt{n})^2} \quad (M = \text{sample mean})$$

Maximum likelihood method: graphical method

- **Important remark:** method in general approximate
 - Taylor expansion stopped at order 2
 - Hypotheses: unbiasedness and efficiency
- **Particular case where it is exact:** normal likelihood with known variance

$$-\ln \mathcal{L}(\mu) = \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2}$$

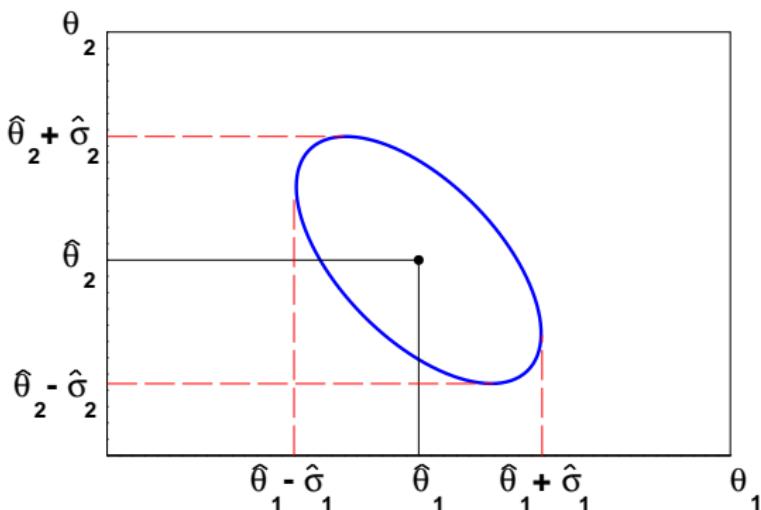
We can show that in this case the following equality holds exactly:

$$-\ln \mathcal{L}(\mu) = -\ln \mathcal{L}(M) + \frac{(\mu - M)^2}{2(\sigma/\sqrt{n})^2} \quad (M = \text{sample mean})$$

- **Note:** doing this we build a so-called "**confidence interval**" for the parameter of interest
 - We'll come back to this notion later

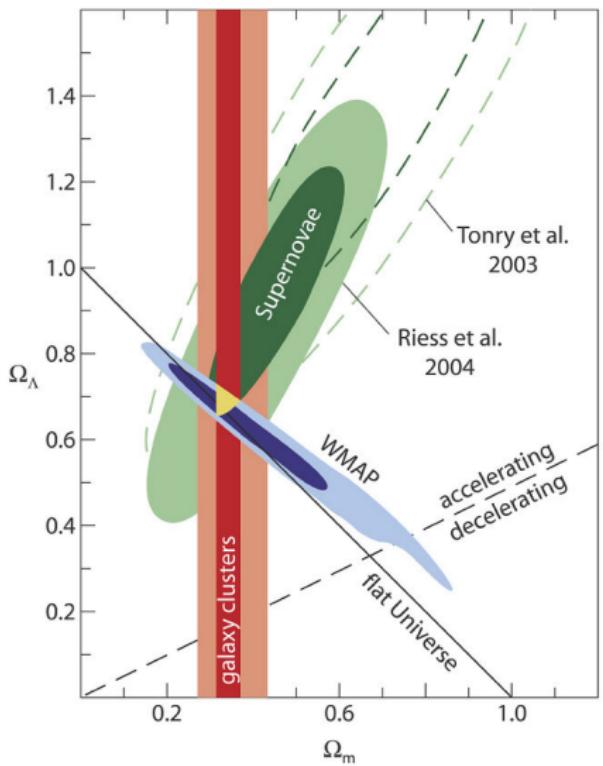
Maximum likelihood method: graphical method in 2D case

- The graphical method also works with 2 parameters of interest



- Using this method we can:
 - Estimate uncertainties on the 2 parameters
 - Estimate correlation between the two estimates

Example



Maximum likelihood method: composite samples

- For composite samples we know that

$$\ln \mathcal{L}_{\text{ext}} = \sum_i \ln \left[\sum_p \mu_p f_p(x_i; \theta) \right] - \sum_p \mu_p$$

Maximum likelihood method: composite samples

- For composite samples we know that

$$\ln \mathcal{L}_{\text{ext}} = \sum_i \ln \left[\sum_p \mu_p f_p(x_i; \theta) \right] - \sum_p \mu_p$$

- Let v_p be the expected fraction of events from process p :

$$v_p = \frac{\mu_p}{\mu} \quad (\mu = \sum \mu_p)$$

Maximum likelihood method: composite samples

- For composite samples we know that

$$\ln \mathcal{L}_{\text{ext}} = \sum_i \ln \left[\sum_p \mu_p f_p(x_i; \theta) \right] - \sum_p \mu_p$$

- Let v_p be the expected fraction of events from process p :

$$v_p = \frac{\mu_p}{\mu} \quad (\mu = \sum \mu_p)$$

- The likelihood thus writes:

$$\ln \mathcal{L}_{\text{ext}} = \sum_i \ln \left[\mu \sum_p v_p f_p(x_i; \theta) \right] - \mu$$

and leads to

$$\hat{\mu}_{\text{ML}} = n$$

Exercice

Let's consider a Bernouilli experiment with probability parameter p . Suppose that when repeating the experiment one gets the following results:

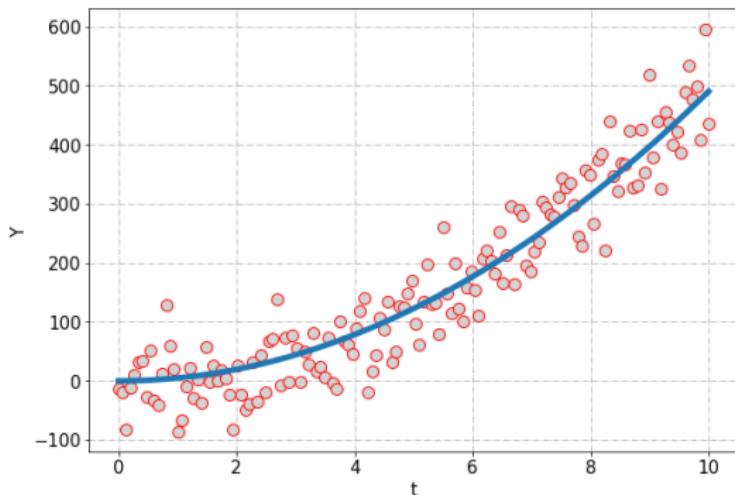
S S S F F S F S F S

where S is success and F is fail.

→ Estimate p

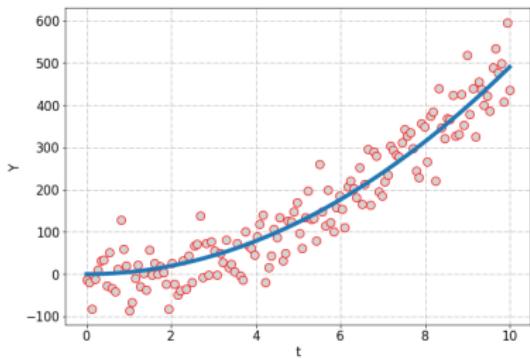
Least squares method

- Least squares method useful when one is interested in dependence of one variable (dependent variable) with one or more other variables (explanatory variables)



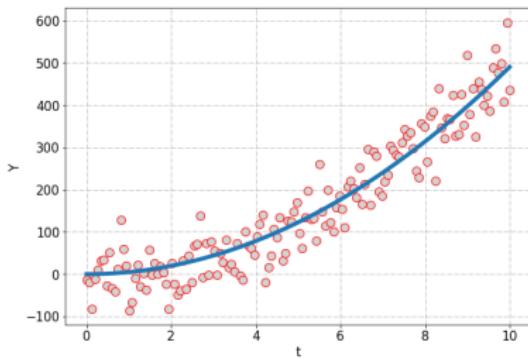
Least squares method: notations

- \mathbf{Y}_i = observation i



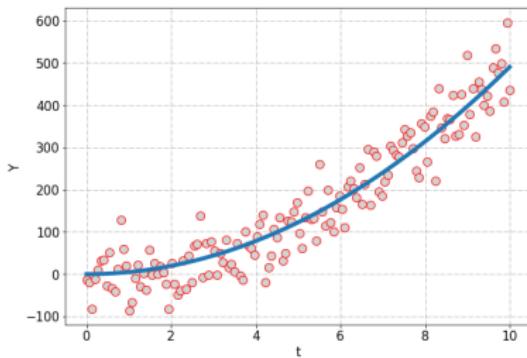
Least squares method: notations

- \mathbf{Y}_i = observation i
- $m_i(\theta) = \mathbb{E}[\mathbf{Y}_i]$ (expectation value of observation i)
→ unknown (depends on the parameters θ to be estimated)



Least squares method: notations

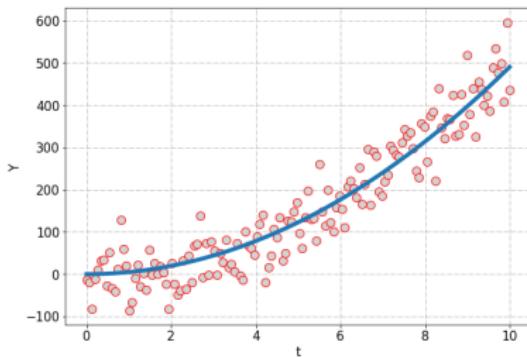
- \mathbf{Y}_i = observation i
- $\mathbf{m}_i(\theta) = \mathbb{E}[\mathbf{Y}_i]$ (expectation value of observation i)
→ unknown (depends on the parameters θ to be estimated)
- \mathbf{V} = covariance matrix of the Y_i 's
(either known or estimated from sample):



$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & & \dots & & \text{cov}(Y_1, Y_n) \\ & \ddots & & & \\ \vdots & & \text{cov}(Y_i, Y_j) & & \vdots \\ & & & \ddots & \\ \text{cov}(Y_n, Y_1) & & \dots & & \sigma_n^2 \end{pmatrix}$$

Least squares method: notations

- \mathbf{Y}_i = observation i
- $\mathbf{m}_i(\theta) = \mathbb{E}[\mathbf{Y}_i]$ (expectation value of observation i)
→ unknown (depends on the parameters θ to be estimated)
- \mathbf{V} = covariance matrix of the Y_i 's
(either known or estimated from sample):



$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \dots & \text{cov}(Y_1, Y_n) \\ \vdots & \ddots & \vdots \\ \text{cov}(Y_n, Y_1) & \dots & \sigma_n^2 \end{pmatrix}$$

- **Advantage w.r.t ML method:** fluctuation model doesn't need to be known

Least squares method

- LS estimators are values that minimize

$$\chi^2(\theta) = (Y - m)^T V^{-1} (Y - m)$$

- Thus

$$\frac{\partial \chi^2}{\partial \theta} \Big|_{\theta=\hat{\theta}_{LS}} = 0 \text{ et } \frac{\partial^2 \chi^2}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{LS}} > 0$$

Least squares method

- LS estimators are values that minimize

$$\chi^2(\theta) = (Y - m)^T V^{-1} (Y - m)$$

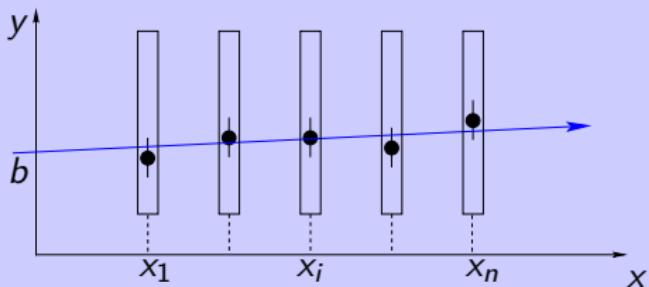
- Thus

$$\frac{\partial \chi^2}{\partial \theta} \Big|_{\theta=\hat{\theta}_{LS}} = 0 \text{ et } \frac{\partial^2 \chi^2}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_{LS}} > 0$$

- In the following, we'll consider the case where the observations are independent:

$$\boxed{\chi^2(\theta) = \sum_{i=1}^n \frac{(Y_i - m_i(\theta))^2}{\sigma_i^2}}$$

Straight line fit



- Show that the LS estimators of the slope and y -intercept are respectively (we assume the observations to be independent and subject to identical fluctuations for all x):

$$\hat{a}_{LS} = \frac{n \sum_i x_i y_i - \sum_i \sum_j x_i y_j}{n \sum_i x_i^2 - \sum_i \sum_j x_i x_j} \quad \text{and} \quad \hat{b}_{LS} = \frac{\sum_i \sum_k x_i^2 y_k - \sum_i \sum_k x_i y_i x_k}{n \sum_i x_i^2 - \sum_i \sum_j x_i x_j}$$

Least squares method: link with the ML method

- For gaussian observations, the likelihood is

$$\mathcal{L} = e^{-\frac{1}{2}(y-m)^T V^{-1}(y-m)}$$

- For independent observations:

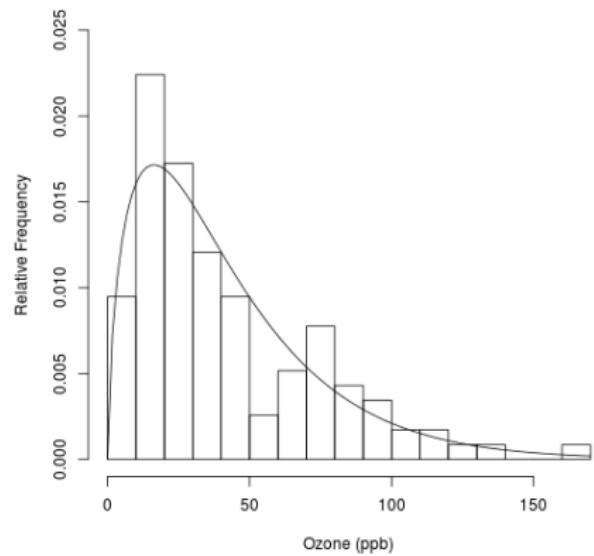
$$\mathcal{L} = \prod_{i=1}^n e^{-\frac{(y_i - m(x_i, \theta))^2}{2\sigma_i^2}}$$

- We see that $-2 \ln \mathcal{L} = \chi^2$
- **Conclusion:** minimizing the χ^2 is equivalent to maximizing the likelihood

Least squares method: binned sample

- ☐ LS method often used with histograms: $\{C_i; Y_i\} (i \in [1, n])$

Histogram of Ozone Pollution Data with Gamma Density Curve

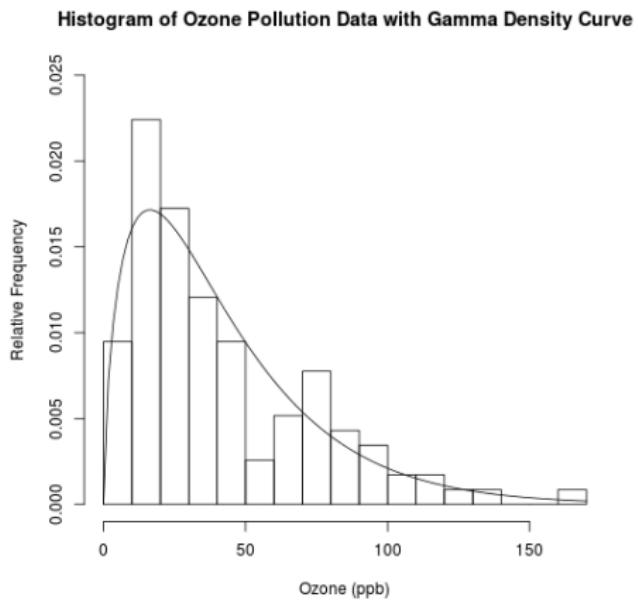


Least squares method: binned sample

- ☐ LS method often used with histograms: $\{C_i; Y_i\} (i \in [1, n])$

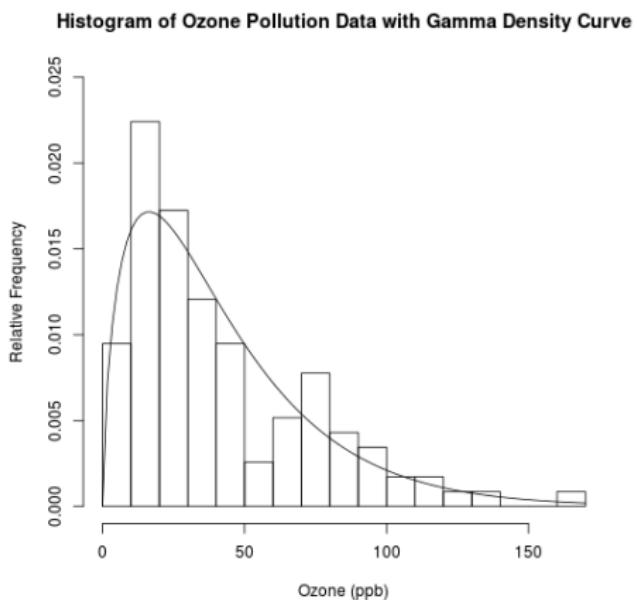
- ☐ **Hypothesis:** the Y_i 's are independent and $Y_i \sim \text{Pois}(m_i)$

$$\Rightarrow \sigma_i = \sqrt{m_i}$$



Least squares method: binned sample

- LS method often used with histograms: $\{C_i; Y_i\} (i \in [1, n])$



- **Hypothesis:** the Y_i 's are independent and $Y_i \sim \text{Pois}(m_i)$

$$\Rightarrow \sigma_i = \sqrt{m_i}$$

- **Remark:** m_i depends on θ

$$m_i(\theta) = v p_i(\theta) = v \int_{C_i} f_X(x; \theta) dx$$

(v =total expected number of events= $\mathbb{E} [\sum Y_i]$)

$$\Rightarrow \sigma_i(\theta) = \sqrt{m_i(\theta)}$$

Least squares method: binned sample

- The χ^2 writes:

$$\chi^2 = \sum_{i=1}^n \frac{(Y_i - m_i(\theta))^2}{m_i(\theta)} \quad (\text{Pearson's } \chi^2)$$

Least squares method: binned sample

- The χ^2 writes:

$$\chi^2 = \sum_{i=1}^n \frac{(Y_i - m_i(\theta))^2}{m_i(\theta)} \quad (\text{Pearson's } \chi^2)$$

which is often simplified to ("modified LS method")

$$\chi^2 = \sum_{i=1}^n \frac{(Y_i - m_i(\theta))^2}{Y_i} \quad (\text{Neyman's } \chi^2)$$

Least squares method: binned sample

- The χ^2 writes:

$$\chi^2 = \sum_{i=1}^n \frac{(Y_i - m_i(\theta))^2}{m_i(\theta)} \quad (\text{Pearson's } \chi^2)$$

which is often simplified to ("modified LS method")

$$\chi^2 = \sum_{i=1}^n \frac{(Y_i - m_i(\theta))^2}{Y_i} \quad (\text{Neyman's } \chi^2)$$

- **Remark:** approximation $\sigma_i(\theta) = \sqrt{m_i(\theta)} \simeq \sqrt{Y_i}$ often done in various contexts
 - Equivalent to using for $\sigma(\theta)$ not the true value (unknown) but an estimate:

$$\hat{\sigma}_i(\theta) = \sqrt{Y_i}$$

Exercice

Show that, noting $N = \sum_i Y_i$, one has

- $\hat{v}_{LS} = N + \frac{\chi^2}{2}$ with Pearson's χ^2
- $\hat{v}_{LS} = N - \chi^2$ with Neyman's χ^2

Least squares method: linear case

- LS method particularly convenient in linear case:

$$m(x_i, \theta) = \sum_{j=1}^q a_j(x_i) \theta_j$$

or, in matrix form,

$$m(x_i, \theta) = A\theta \quad (A_{ij} = a_j(x_i))$$

- χ^2 writes:

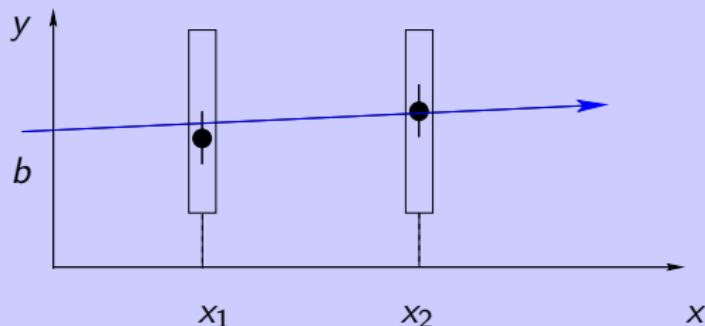
$$\chi^2(\theta) = (Y - A\theta)^T V^{-1} (Y - A\theta)$$

- We can prove that this linear problem has analytical solutions:

$$\boxed{\hat{\theta}_{LS} = (A^T V^{-1} A)^{-1} A^T V^{-1} y = B y}$$

Exercice

We consider a straight line fit with two points $y = ax + b$:



- ① Calculate \hat{a} and \hat{b} , the LS estimators of a and b (the uncertainty σ is the same in both planes)
- ② Calculate the covariance matrix of \hat{a} and \hat{b} as a function of σ , x_1 and x_2
 - directly
 - using the fact that the estimators are efficient
- ③ Under which condition on x_1 and x_2 are the estimators \hat{a} and \hat{b} uncorrelated ?

The bayesian way of estimating parameters

- Based on the posterior distribution:

$$P(\theta; \mathbf{x}) = \frac{P(\mathbf{x}; \theta)P(\theta)}{P(\mathbf{x})}$$

(P: pmf or pdf)

Diagram illustrating the Bayesian formula:

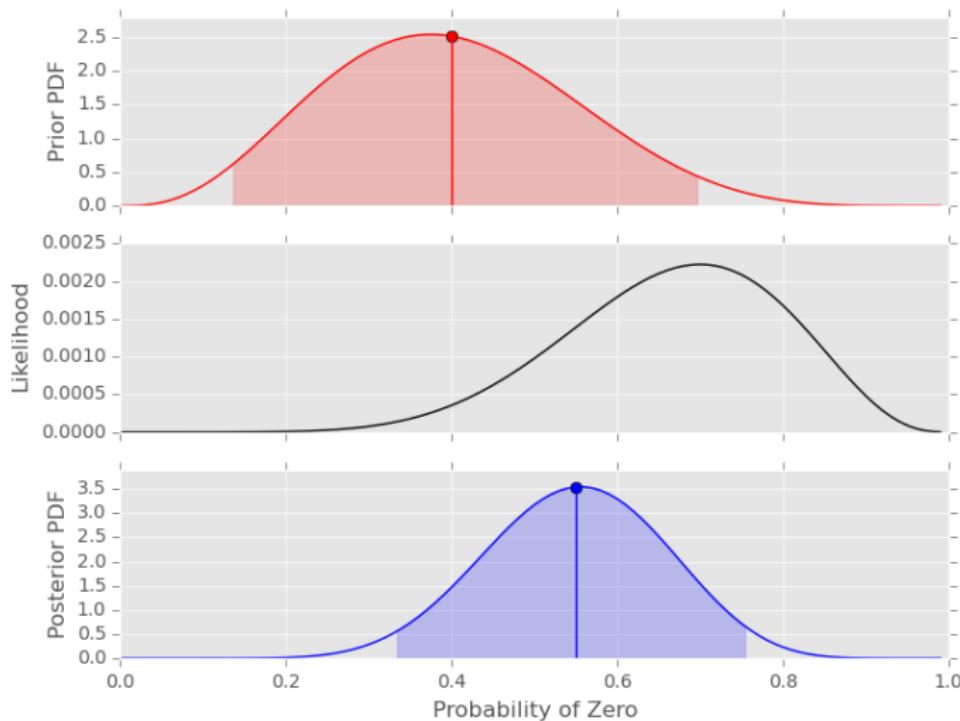
- Posterior distribution: $P(\theta; \mathbf{x})$
- Stat. model (likelihood): $P(\mathbf{x}; \theta)$
- Prior distribution: $P(\theta)$
- Normalization constant: $P(\mathbf{x})$

Arrows point from the prior and likelihood terms to the numerator, and from the normalization constant to the denominator.

- In less rigorous terms:

$$P(model|data) = \frac{P(data|model)P(model)}{P(data)}$$

Illustration



The bayesian way of estimating parameters

- Once the posterior $f(\theta; x)$ is found, parameters can be estimated using either the

The bayesian way of estimating parameters

- Once the posterior $f(\theta|x)$ is found, parameters can be estimated using either the
 - **Mean:** $\hat{\theta} = \int \theta f(\theta|x) d\theta$

The bayesian way of estimating parameters

- Once the posterior $f(\theta|x)$ is found, parameters can be estimated using either the
 - **Mean:** $\hat{\theta} = \int \theta f(\theta|x) d\theta$
 - **Mode:** $\hat{\theta} = \max_{\theta} f(\theta|x)$

The bayesian way of estimating parameters

- Once the posterior $f(\theta|x)$ is found, parameters can be estimated using either the

- Mean:** $\hat{\theta} = \int \theta f(\theta|x) d\theta$

- Mode:** $\hat{\theta} = \max_{\theta} f(\theta|x)$

- Median:** $F_{\theta}(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}} f(\theta|x) d\theta = 1/2$

The bayesian way of estimating parameters

- Once the posterior $f(\theta|x)$ is found, parameters can be estimated using either the
 - **Mean:** $\hat{\theta} = \int \theta f(\theta|x) d\theta$
 - **Mode:** $\hat{\theta} = \max_{\theta} f(\theta|x)$
 - **Median:** $F_{\theta}(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}} f(\theta|x) d\theta = 1/2$
 - Any other location measurement quantity

Choice of prior

- Once the stat. model is known, bayesian inference "only" requires finding a proper prior distribution
- Choice of prior always arbitrary to some extend
- Final result can depend a lot on choice of prior
- Often not easy to decide what the "best" prior is

→ All these issues related to the choice and impact of priors are what make many people dislike bayesian approach

- **Important fact:** impact of prior on final result decreases as sample size increases (or equivalently as the number of measurement increases)

- **Important fact:** impact of prior on final result decreases as sample size increases (or equivalently as the number of measurement increases)
- **Example:** n counting experiments

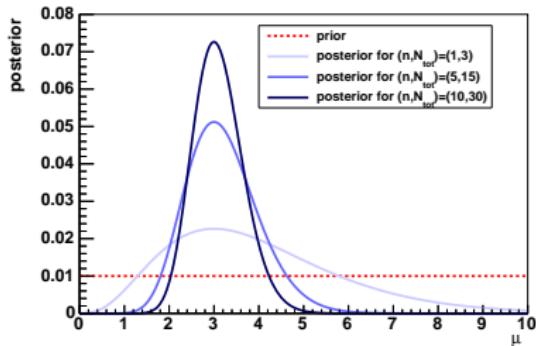
$$f(\mu | \{N_i\}) \propto \prod_{i=1}^n \frac{\mu^{N_i}}{N_i!} e^{-\mu} \times g(\mu) \propto e^{-n\mu} \mu^{N_{\text{tot}}} \times g(\mu)$$

Trade-off likelihood-prior

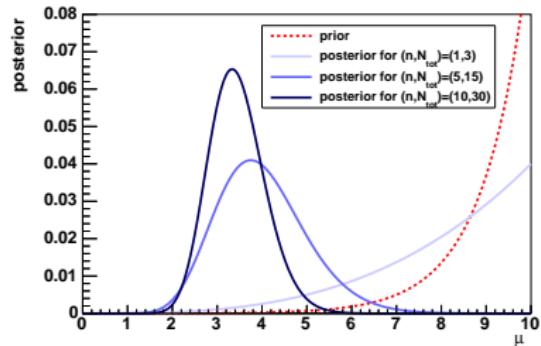
- **Important fact:** impact of prior on final result decreases as sample size increases (or equivalently as the number of measurement increases)
- **Example:** n counting experiments

$$f(\mu | \{N_i\}) \propto \prod_{i=1}^n \frac{\mu^{N_i}}{N_i!} e^{-\mu} \times g(\mu) \propto e^{-n\mu} \mu^{N_{\text{tot}}} \times g(\mu)$$

uniform prior ($g(\mu) \propto \text{constant}$)



exponential prior ($g(\mu) \propto e^\mu$)



The ML method: frequentist or bayesian ?

- ML method provides estimators with good frequentist properties
 - It can be considered a frequentist method

The ML method: frequentist or bayesian ?

- ML method provides estimators with good frequentist properties
 - It can be considered a frequentist method
- ML estimators are also found by taking the mode of the posterior distribution when a uniform prior is used
 - It can be considered a bayesian method

The ML method: frequentist or bayesian ?

- ML method provides estimators with good frequentist properties
 - It can be considered a frequentist method
- ML estimators are also found by taking the mode of the posterior distribution when a uniform prior is used
 - It can be considered a bayesian method
- **Conclusion:**
 - ML method (as many other stat. methods) is not frequentist or bayesian
 - Makes sense under both approaches

2D example

- Bayesian approach works similarly in multiple dimensions
- **Example: poissonian measurement with signal and background**

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

- Parameters: s and b
- Suppose b is measured in some auxiliary measurement to be $b = b_0 \pm \sigma$ and you take the following normal prior:

$$g(b|b_0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b-b_0)^2}{2\sigma^2}}$$

- Suppose s is totally unknown prior to the measurement and you take a uniform prior: $g(s) \propto \text{constant}$
- **2D posterior:**

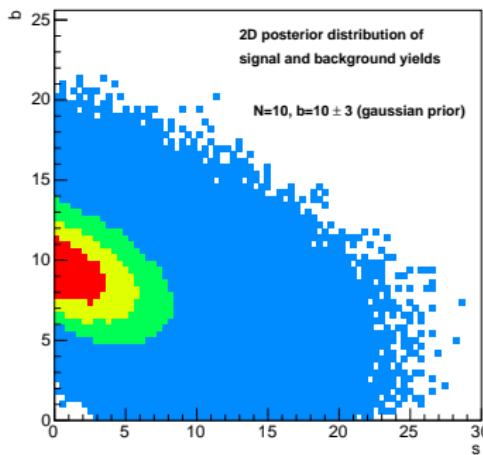
$$f(s, b|N) \propto P(N|s, b) \underbrace{g(b|b_0, \sigma) g(s)}_{\text{priors}}$$

2D example

□ 2D posterior:

$$f(s, b|N) \propto P(N|s, b) \underbrace{g(b|b_0, \sigma)}_{\text{priors}} g(s)$$

$$\propto \frac{(s+b)^N}{N!} e^{-(s+b)} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b-b_0)^2}{2\sigma^2}}$$



Bayesian increase of knowledge

- ☐ Bayesian inference offers an easy way to describe how knowledge increases as measurements are performed

prior knowledge $\xrightarrow{\text{measurement}}$ posterior knowledge

More formally:

$g(\theta) \xrightarrow{x} f(\theta|x)$

- ☐ Increase of knowledge measured by comparing prior and posterior distributions
 - If very different, measurement was very informative
 - If not very different, measurement didn't contain much information about the parameter to be estimated

Bayesian increase of knowledge

- Suppose you make two measurements with likelihoods \mathcal{L}_1 and \mathcal{L}_2
 - After first measurement: $f_1(\theta|x_1) \propto \mathcal{L}_1 \times g(\theta)$

Bayesian increase of knowledge

- Suppose you make two measurements with likelihoods \mathcal{L}_1 and \mathcal{L}_2
 - After first measurement: $f_1(\theta|x_1) \propto \mathcal{L}_1 \times g(\theta)$
 - After second measurement: we could use

$$f_2(\theta|x_2) \propto \mathcal{L}_2 \times g(\theta)$$

Bayesian increase of knowledge

- Suppose you make two measurements with likelihoods \mathcal{L}_1 and \mathcal{L}_2
 - After first measurement: $f_1(\theta|x_1) \propto \mathcal{L}_1 \times g(\theta)$
 - After second measurement: we could use

$$f_2(\theta|x_2) \propto \mathcal{L}_2 \times g(\theta)$$

but better to use knowledge acquired from first measurement:

$$f_2(\theta|x_2, x_1) \propto \mathcal{L}_2 \times f_1(\theta|x_1) \propto \mathcal{L}_2 \times \mathcal{L}_1 \times g(\theta)$$

Bayesian increase of knowledge

- Suppose you make two measurements with likelihoods \mathcal{L}_1 and \mathcal{L}_2
 - After first measurement: $f_1(\theta|x_1) \propto \mathcal{L}_1 \times g(\theta)$
 - After second measurement: we could use

$$f_2(\theta|x_2) \propto \mathcal{L}_2 \times g(\theta)$$

but better to use knowledge acquired from first measurement:

$$f_2(\theta|x_2, x_1) \propto \mathcal{L}_2 \times f_1(\theta|x_1) \propto \mathcal{L}_2 \times \mathcal{L}_1 \times g(\theta)$$

- Posterior of first measurement used as prior in second one

Bayesian increase of knowledge

- Suppose you make two measurements with likelihoods \mathcal{L}_1 and \mathcal{L}_2
 - After first measurement: $f_1(\theta|x_1) \propto \mathcal{L}_1 \times g(\theta)$
 - After second measurement: we could use

$$f_2(\theta|x_2) \propto \mathcal{L}_2 \times g(\theta)$$

but better to use knowledge acquired from first measurement:

$$f_2(\theta|x_2, x_1) \propto \mathcal{L}_2 \times f_1(\theta|x_1) \propto \mathcal{L}_2 \times \mathcal{L}_1 \times g(\theta)$$

→ Posterior of first measurement used as prior in second one

- **Generalization to n measurements:**

$$f_n(\theta|x_n, \dots, x_1) \propto \mathcal{L}_n \times \mathcal{L}_{n-1} \times \dots \times \mathcal{L}_1 \times g(\theta)$$

→ As the number of measurement increases, the likelihood term "wins" and the prior becomes more and more subdominant in final result

Uncertainties on parameters: credible intervals

- In addition to providing point estimation, bayesian approach provides with no additional effort error bars for the parameter of interest

$$\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$$

Uncertainties on parameters: credible intervals

- In addition to providing point estimation, bayesian approach provides with no additional effort error bars for the parameter of interest

$$\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$$

- The interval

$$[\hat{\theta} - \sigma_{\hat{\theta}}; \hat{\theta} + \sigma_{\hat{\theta}}]$$

is called "**credible interval**" for the parameter of interest

Uncertainties on parameters: credible intervals

- In addition to providing point estimation, bayesian approach provides with no additional effort error bars for the parameter of interest

$$\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$$

- The interval

$$[\hat{\theta} - \sigma_{\hat{\theta}}; \hat{\theta} + \sigma_{\hat{\theta}}]$$

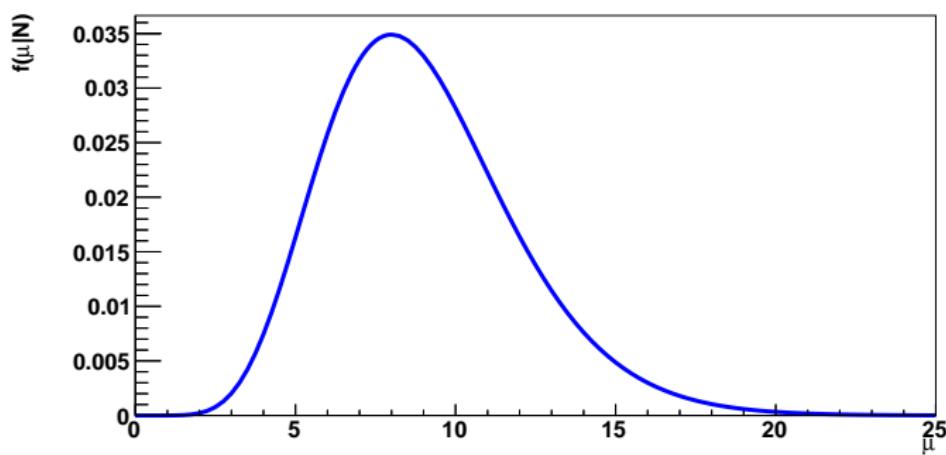
is called "**credible interval**" for the parameter of interest

- **Remark:**

- Credible interval → bayesian
- Confidence interval → frequentist (will be addressed later)

Uncertainties on parameters: credible intervals

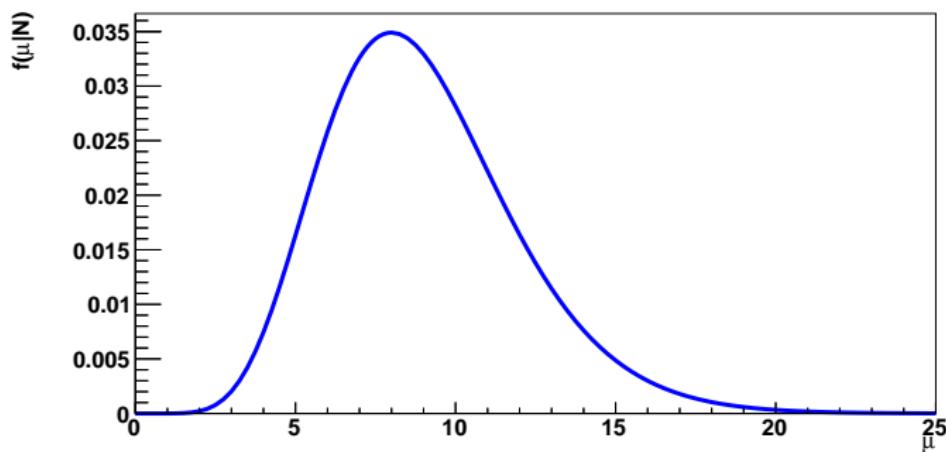
- How are credible intervals determined ?



Uncertainties on parameters: credible intervals

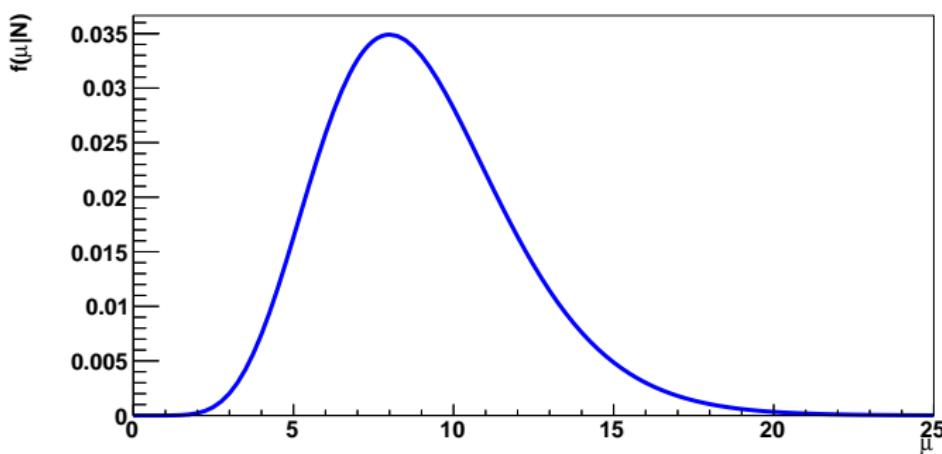
- How are credible intervals determined ?

- ➊ Take posterior distribution



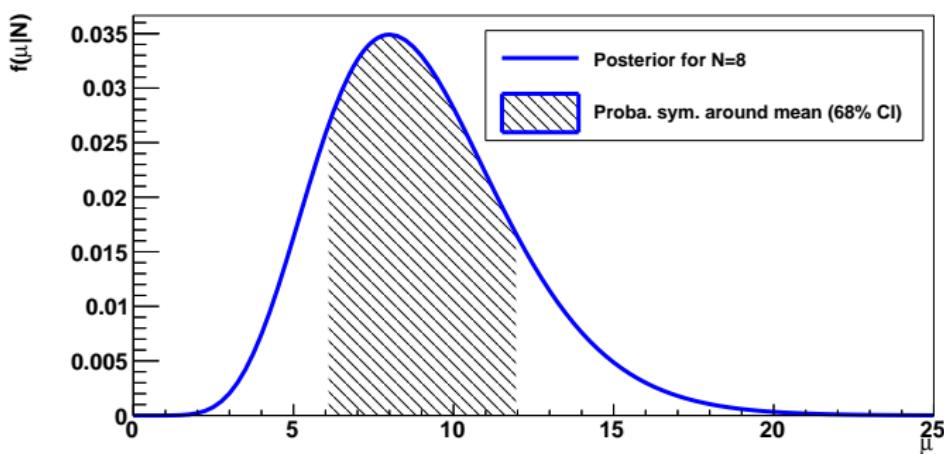
Uncertainties on parameters: credible intervals

- How are credible intervals determined ?
 - ① Take posterior distribution
 - ② Choose degree of credibility α (typical values: 68%, 90%, 95%)



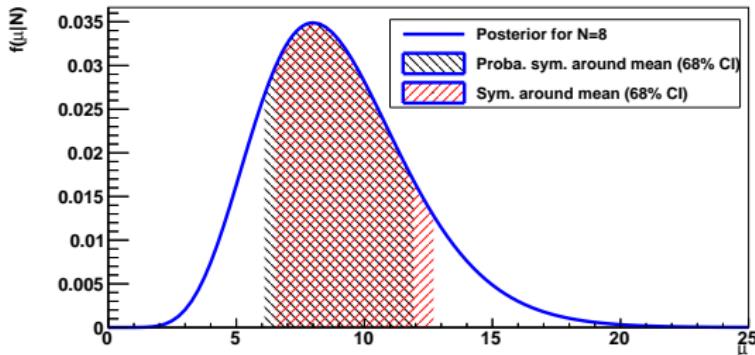
Uncertainties on parameters: credible intervals

- How are credible intervals determined ?
 - ➊ Take posterior distribution
 - ➋ Choose degree of credibility α (typical values: 68%, 90%, 95%)
 - ➌ Build interval corresponding to the chosen degree of credibility



Uncertainties on parameters: credible intervals

- Multiple ways of building the interval:



- **Symmetric around mean:**

$$[\mathbb{E}[\theta] - a; \mathbb{E}[\theta] + a] \quad \text{with} \quad \int_{\mathbb{E}[\theta] - a}^{\mathbb{E}[\theta] + a} f(\theta) d\theta = \alpha$$

- **Probability symmetric around mean:**

$$[a; b] \quad \text{with} \quad \int_a^{\mathbb{E}[\theta]} f(\theta) d\theta = \int_{\mathbb{E}[\theta]}^b f(\theta) d\theta = \alpha/2$$

- **Highest probability density interval:**

$$[a; b] \quad \text{with} \quad \int_a^b f(\theta) d\theta = \alpha \text{ and } f(\theta_1) > f(\theta_2) \text{ for } \theta_1 \in [a; b] \text{ and } \theta_2 \notin [a; b]$$

The likelihood principle

- The bayesian approach is said to obey the **likelihood principle**
- Here likelihood="statistical model with constant terms removed"
- Likelihood principle:**

Two measurements with the same likelihoods should lead to the same inference

The likelihood principle

- The bayesian approach is said to obey the **likelihood principle**
- Here likelihood="statistical model with constant terms removed"
- Likelihood principle:**

Two measurements with the same likelihoods should lead to the same inference

In other words: constant terms in the statistical model should not participate to the inference

The likelihood principle

- The bayesian approach is said to obey the **likelihood principle**
- Here likelihood="statistical model with constant terms removed"
- Likelihood principle:**

Two measurements with the same likelihoods should lead to the same inference

In other words: constant terms in the statistical model should not participate to the inference

- Remark:** frequentist inference does not in general obey the likelihood principle (remember the exercise where you have to estimate p from the following result: S S S F F S F S F S)

Bayesian treatment of nuisance parameters

- Reminder:** nuisance parameter = parameters other than the parameters of interest
 - Example:** poissonian measurement with signal and background

param. of
interest

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

$$\Rightarrow f(s, b | N) \propto P(N|s, b) \underbrace{g(b)g(s)}_{\text{priors}}$$

Bayesian treatment of nuisance parameters

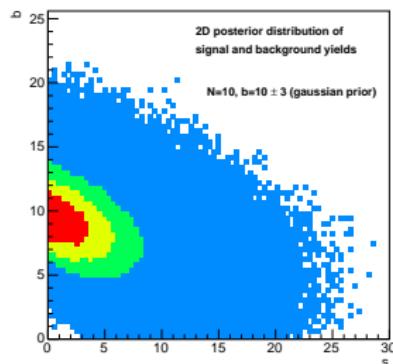
- Reminder: nuisance parameter = parameters other than the parameters of interest
 - Example: poissonian measurement with signal and background

param. of
interest

nuisance
param.

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

$$\Rightarrow f(s, b | N) \propto P(N | s, b) \underbrace{g(b)g(s)}_{\text{priors}}$$

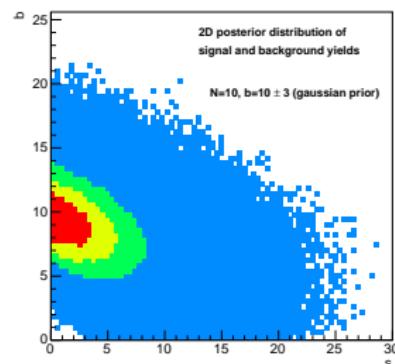


Bayesian treatment of nuisance parameters

- Reminder: **nuisance parameter** = parameters other than the parameters of interest
- **Example:** poissonian measurement with signal and background

param. of interest nuisance param.

$$P(N|s, b) = \frac{(s+b)^N}{N!} e^{-(s+b)}$$
$$\Rightarrow f(s, b|N) \propto P(N|s, b) \underbrace{g(b)g(s)}_{\text{priors}}$$



→ We have $f(s, b|N)$ but we want to make inference on s only

$$f(s, b|N) \xrightarrow{??} f(s|N)$$

Bayesian treatment of nuisance parameters

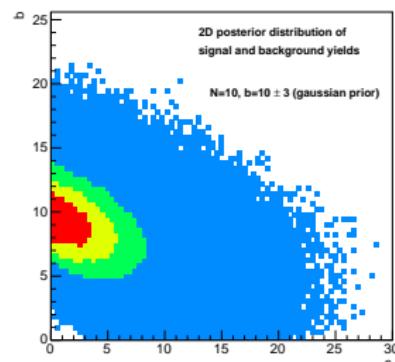
- Reminder: **nuisance parameter** = parameters other than the parameters of interest
- **Example:** poissonian measurement with signal and background

param. of
interest

nuisance
param.

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

$$\Rightarrow f(s, b|N) \propto P(N|s, b) \underbrace{g(b)g(s)}_{\text{priors}}$$



→ We have $f(s, b|N)$ but we want to make inference on s only

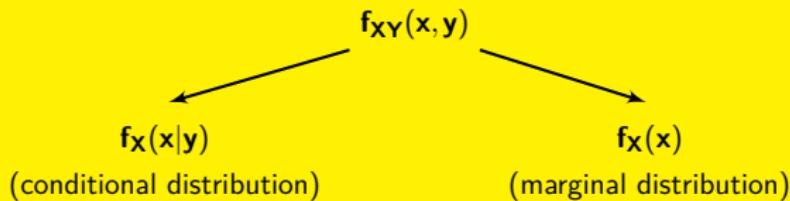
$$f(s, b|N) \xrightarrow{??} f(s|N)$$

→ Solution: marginalization

Marginalization

- Consider a 2D problem with two random variables: X and Y
- Let $f_{XY}(x,y)$ be their joint distribution
- Suppose you're not interested in the 2D distribution but in the distribution of X only

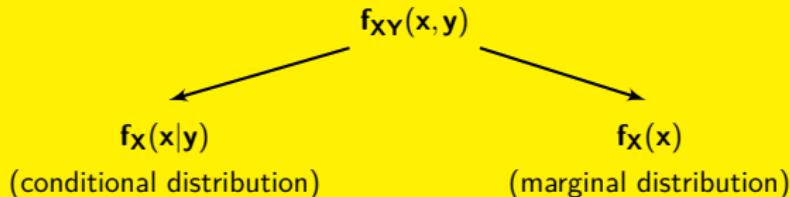
→ 2 types of distributions



Marginalization

- Consider a 2D problem with two random variables: X and Y
- Let $f_{XY}(x,y)$ be their joint distribution
- Suppose you're not interested in the 2D distribution but in the distribution of X only

→ 2 types of distributions



- Marginal distribution of X given by:

$$f_X(x) = \int f_{XY}(x,y)dy$$

- As a function of conditional distribution:

$$f_X(x) = \int f_X(x|y)f_Y(y)dy = \mathbb{E}_Y [f_X(x|y)]$$

Exercice

Let X and Y be two random variables taking the following values:
 $X = \{1, 2, 3\}$ et $Y = \{-1, 1\}$. Their joint distribution is

		X	1	2	3
		Y			
		-1	0,1	0,3	0,1
		1	0,2	0,1	0,2

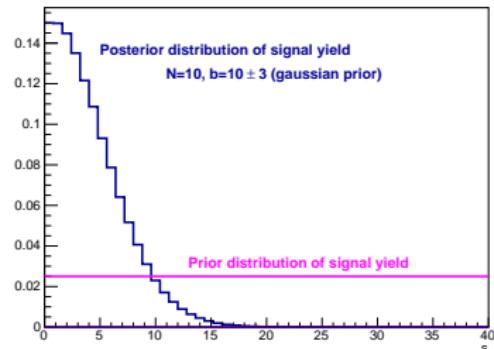
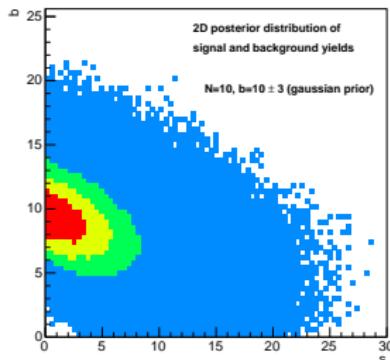
- ① What are the marginal distributions of X and Y ?
- ② Calculate the expectation values of X and Y
- ③ What are the conditional distributions of X and Y ?
- ④ Calculate $\mathbb{E}[X|Y = 1]$, $\mathbb{E}[X|Y = -1]$, $\mathbb{E}[Y|X = 1]$, $\mathbb{E}[Y|X = 2]$ et $\mathbb{E}[Y|X = 3]$
- ⑤ Calculate the covariance matrix of X and Y
- ⑥ Are X and Y independent ?

Bayesian treatment of nuisance parameters

$$f(s, b|N) \xrightarrow{\text{marginalization}} f(s|N)$$

- Marginalization of 2D posterior to get 1D posterior for parameter of interest:

$$f(s|N) = \int f(s, b|N) db$$



□ Other formulation

$$\begin{aligned}f(s|N) &= \int f(s, b|N)db \\&\propto \int P(N|s, b) \underbrace{g(b)g(s)}_{\text{priors}} db \\&\propto \underbrace{\left[\int P(N|s, b)g(b)db \right]}_{\text{marginal likelihood}} \times g(s)\end{aligned}$$

- Other formulation

$$\begin{aligned}f(s|N) &= \int f(s, b|N) db \\&\propto \int P(N|s, b) \underbrace{g(b)g(s)}_{\text{priors}} db \\&\propto \underbrace{\left[\int P(N|s, b) g(b) db \right]}_{\text{marginal likelihood}} \times g(s)\end{aligned}$$

- Denoting \mathcal{L}_m the marginal likelihood, we have:

$$f(s|N) \propto \mathcal{L}_m(s) \times g(s)$$

- Other formulation

$$\begin{aligned}f(s|N) &= \int f(s, b|N)db \\&\propto \int P(N|s, b) \underbrace{g(b)g(s)}_{\text{priors}} db \\&\propto \underbrace{\left[\int P(N|s, b)g(b)db \right]}_{\text{marginal likelihood}} \times g(s)\end{aligned}$$

- Denoting \mathcal{L}_m the marginal likelihood, we have:

$$f(s|N) \propto \mathcal{L}_m(s) \times g(s)$$

- **Conclusion:** we can write directly Bayes' theorem for the parameter of interest using not the full likelihood but the marginalized one (where marginalization done over nuisance parameters)

Marginal likelihood

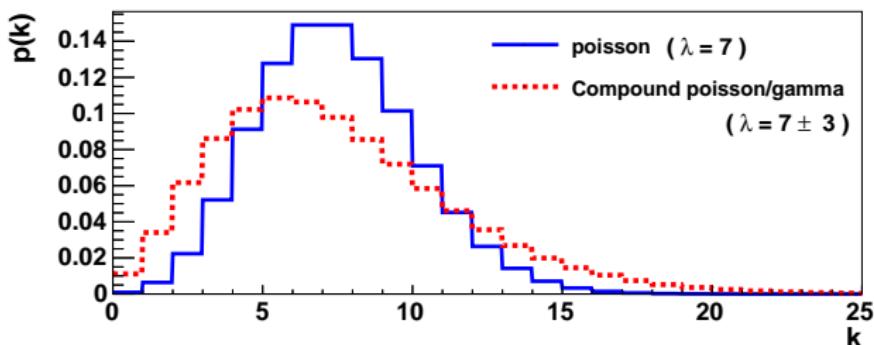
- Marginal likelihood is the full likelihood where nuisance parameters have been "eliminated"
 - Marginal likelihoods depend only on the parameter of interest

Marginal likelihood

- Marginal likelihood is the full likelihood where nuisance parameters have been "eliminated"
 - Marginal likelihoods depend only on the parameter of interest
- Marginal likelihood is a so-called **compound distribution** or **mixture distribution**

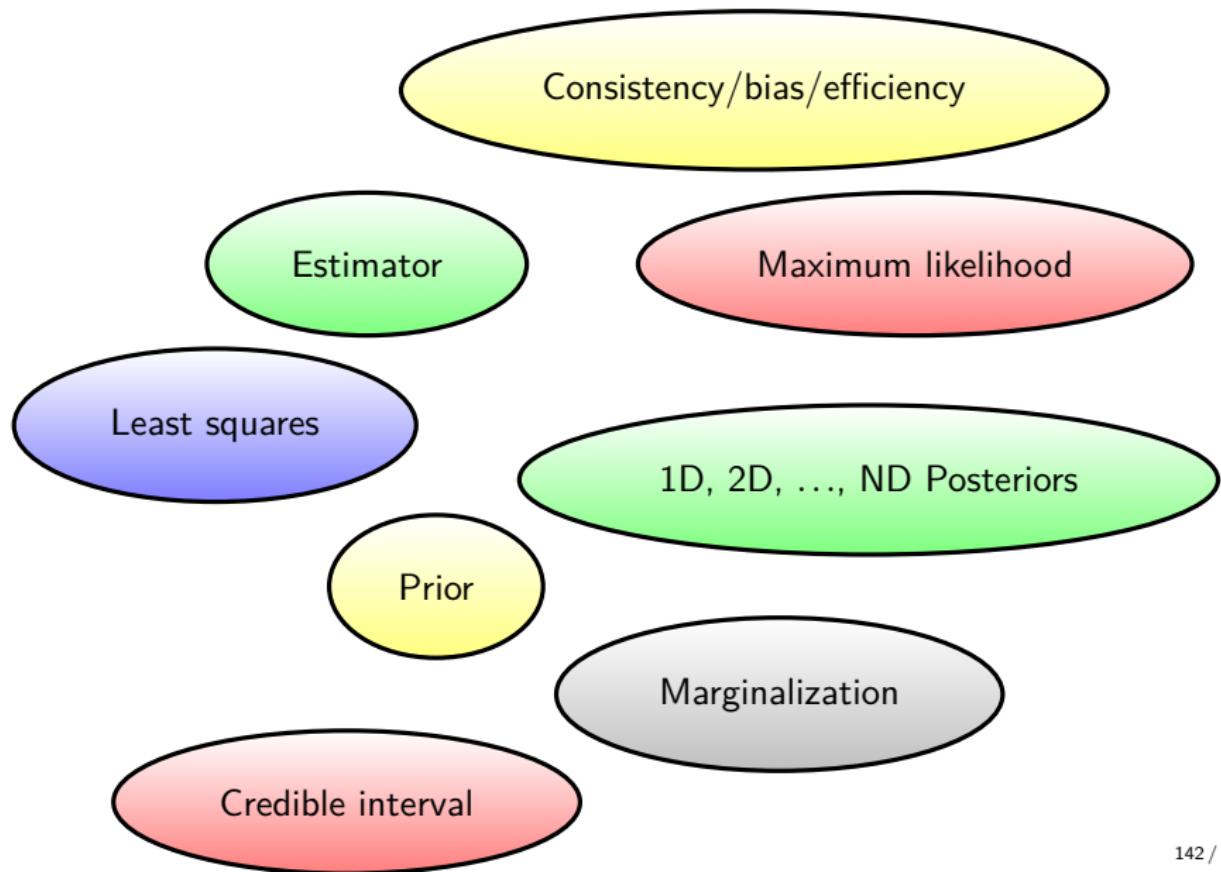
Marginal likelihood

- Marginal likelihood is the full likelihood where nuisance parameters have been "eliminated"
 - Marginal likelihoods depend only on the parameter of interest
- Marginal likelihood is a so-called **compound distribution** or **mixture distribution**
- Example: Poisson-Gamma mixture



- Effect of marginalization: smearing of distribution of observable
 - Extreme values are more likely to occur

Key words/concepts



Chap. 5 Confidence intervals

□ **Definition:**

Confidence interval (CI) = interval supposed to contain true value of the parameter of interest with high probability

□ **Definition:**

Confidence interval (CI) = interval supposed to contain true value of the parameter of interest with high probability

- This definition is very vague
- Purposes of this chapter:
 - ① Understand what this means exactly
 - ② Learn methods to find confidence intervals

□ Definition:

Confidence interval (CI) = interval supposed to contain true value of the parameter of interest with high probability

- This definition is very vague
- Purposes of this chapter:
 - ① Understand what this means exactly
 - ② Learn methods to find confidence intervals

□ Examples of CI:

- Fraction of left-handed people $\in [9.1\%; 10.2\%]$ @ 90% CL
- Fraction of people with no access to clean water $\in [0.10; 0.12]$ @ 99% CL
- Higgs mass $\in [124.94; 125.36]$ GeV @ 95% CL

Introduction

- CI always have the following structure:

$$\theta \in [\underbrace{\theta_{\min}; \theta_{\max}}_{\text{interval}}] @ \underbrace{\alpha}_{\substack{\text{value of} \\ \text{confidence level}}} CL$$

- θ_{\min} and θ_{\max} called **lower** and **upper bound**

Introduction

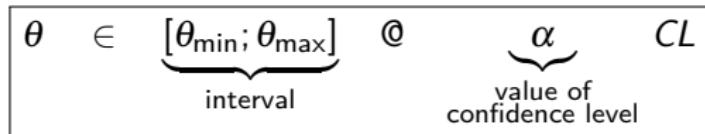
- CI always have the following structure:

$$\theta \in [\underbrace{\theta_{\min}; \theta_{\max}}_{\text{interval}}] @ \underbrace{\alpha}_{\substack{\text{value of} \\ \text{confidence level}}} CL$$

- θ_{\min} and θ_{\max} called **lower** and **upper bound**
- If either of the two bounds is equal to the parameter limit, interval said to be **one-sided**
- Otherwise, interval said to be **two-sided**

Introduction

- CI always have the following structure:



- θ_{\min} and θ_{\max} called **lower** and **upper bound**
- If either of the two bounds is equal to the parameter limit, interval said to be **one-sided**
- Otherwise, interval said to be **two-sided**
- **Confidence level** α reflects how confident we are that the true parameter is in the quoted interval

Introduction

- CI always have the following structure:



- θ_{\min} and θ_{\max} called **lower** and **upper bound**
- If either of the two bounds is equal to the parameter limit, interval said to be **one-sided**
- Otherwise, interval said to be **two-sided**
- **Confidence level** α reflects how confident we are that the true parameter is in the quoted interval
- Quoting a result as follows makes no sense

$$\theta \in [\theta_{\min}; \theta_{\max}]$$

Introduction

- **CI are random objects** (bounds are functions of the data $\textcolor{blue}{x}$):

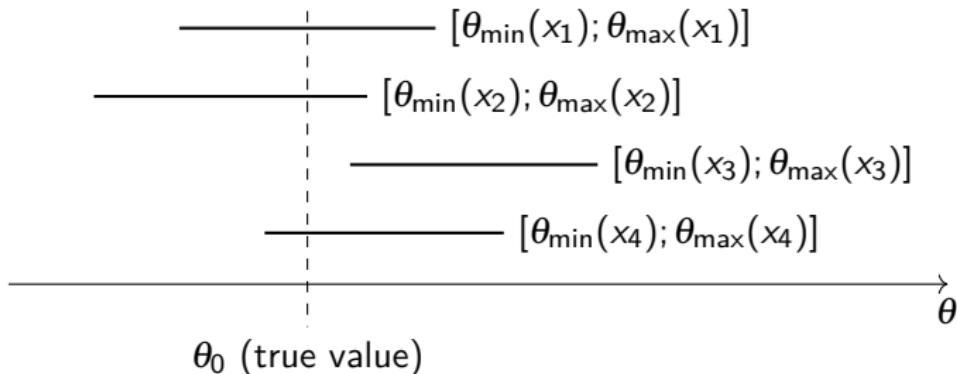
$$[\theta_{\min}(\textcolor{blue}{x}); \theta_{\max}(\textcolor{blue}{x})]$$

Introduction

- CI are random objects (bounds are functions of the data x):

$$[\theta_{\min}(x); \theta_{\max}(x)]$$

- If you repeat the measurement, you'll get different intervals



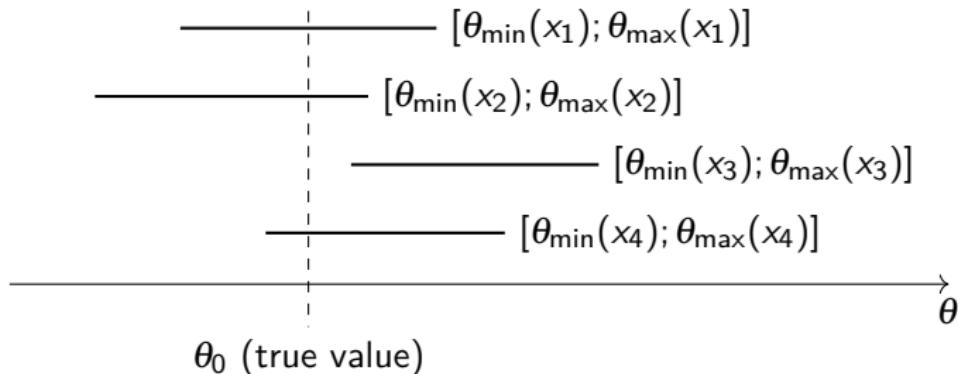
- CI can contain the true value or not
- You're never 100% sure that the CI contains the true value

Introduction

- CI are random objects (bounds are functions of the data x):

$$[\theta_{\min}(x); \theta_{\max}(x)]$$

- If you repeat the measurement, you'll get different intervals



- CI can contain the true value or not
- You're never 100% sure that the CI contains the true value

- Bayesian approach already described → Will focus on **frequentist approach**

Frequentist approach

- In frequentist approach, building of CI based on **coverage** notion
- Coverage = probability that CI contains true value

$$\text{coverage} = P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)])$$

- In frequentist approach, building of CI based on **coverage** notion
- Coverage = probability that CI contains true value

$$\text{coverage} = P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)])$$



This probability should not be misunderstood:

- **It is not** the probability that θ_0 belongs to the interval you compute from your measurement
- **It is** the probability that, if you repeat the measurement many many times, the intervals you get contain θ_0

Frequentist approach

- CI built in order to have coverage as close as possible to a predefined value
 - This predefined value is called **confidence level (α)**

- CI built in order to have coverage as close as possible to a predefined value
 - This predefined value is called **confidence level (α)**
- Typical values of α : 68%, 90%, 95%, 99%

- CI built in order to have coverage as close as possible to a predefined value
 - This predefined value is called **confidence level (α)**
- Typical values of α : 68%, 90%, 95%, 99%
- Confidence level and coverage are not the same thing:
 - Confidence level = objective we try to attain
 - Coverage = what we actually reach when trying to attain the objective

Goal: coverage $\simeq \alpha$

More on coverage

- coverage = α can be difficult to achieve in realistic cases
- 3 cases must be distinguished:

More on coverage

- coverage = α can be difficult to achieve in realistic cases
- 3 cases must be distinguished:
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) = \alpha$: ideal case (perfect coverage)

More on coverage

- coverage = α can be difficult to achieve in realistic cases
- 3 cases must be distinguished:
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) = \alpha$: ideal case (perfect coverage)
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) > \alpha$: not ideal but acceptable (**overcoverage**)

More on coverage

- coverage = α can be difficult to achieve in realistic cases
- 3 cases must be distinguished:
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) = \alpha$: ideal case (perfect coverage)
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) > \alpha$: not ideal but acceptable (**overcoverage**)
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) < \alpha$: should be avoided (**undercoverage**)

More on coverage

- coverage = α can be difficult to achieve in realistic cases
- 3 cases must be distinguished:
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) = \alpha$: ideal case (perfect coverage)
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) > \alpha$: not ideal but acceptable (**overcoverage**)
 - $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) < \alpha$: should be avoided (**undercoverage**)
- A good frequentist method is a method that has no undercoverage and minimal overcoverage



Methods described here in general don't have known coverage

→ Can undercover

→ Use with caution !

Approximate methods



Methods described here in general don't have known coverage

→ Can undercover

→ Use with caution !

- Simplest way to build CI is to start from an estimator $\hat{\theta}$:

$$\text{CI} = \left[\hat{\theta} - d \sqrt{\text{var}[\hat{\theta}]} ; \hat{\theta} + d \sqrt{\text{var}[\hat{\theta}]} \right]$$

where

- $\sqrt{\text{var}[\hat{\theta}]}$ is the "uncertainty" on the estimate
- d is a real number used to adjust the size of the interval (and thus the coverage)
 - If you want high confidence level, use large d (example: $d = 3$)
 - If you want low confidence level, use small d (example: $d = 1$)

Unbiased normal case

- In this case, possible to achieve perfect coverage

$$\hat{\theta} \sim \mathcal{N}\left(\theta_0, \sqrt{\text{var}[\hat{\theta}]}\right)$$

$$\begin{aligned}\Rightarrow \text{coverage} &= P\left(\theta_0 \geq \hat{\theta} - d\sqrt{\text{var}[\hat{\theta}]} \cap \theta_0 \leq \hat{\theta} + d\sqrt{\text{var}[\hat{\theta}]}\right) \\ &= P\left(\theta_0 \geq \hat{\theta} - d\sqrt{\text{var}[\hat{\theta}]}\right) + P\left(\theta_0 \leq \hat{\theta} + d\sqrt{\text{var}[\hat{\theta}]}\right) - 1 \\ &= \underbrace{P\left(\hat{\theta} \leq \theta_0 + d\sqrt{\text{var}[\hat{\theta}]}\right)}_{\Phi(d)} + \underbrace{P\left(\hat{\theta} \geq \theta_0 - d\sqrt{\text{var}[\hat{\theta}]}\right)}_{1 - \Phi(-d) = \Phi(d)} - 1 \\ \Rightarrow \text{coverage} &= 2\Phi(d) - 1 \quad \text{or} \quad d = \Phi^{-1}\left(\frac{1 + \text{coverage}}{2}\right)\end{aligned}$$

- **Conclusion:** coverage known once d fixed (we can thus achieve coverage = α)

Unbiased normal case

- **Reminder:** maximum likelihood method leads, in asymptotic limit, to estimators that are normal, unbiased and efficient
⇒ Previous result applies to ML estimators when sample size large

Unbiased normal case

- **Reminder:** maximum likelihood method leads, in asymptotic limit, to estimators that are normal, unbiased and efficient
 - ⇒ Previous result applies to ML estimators when sample size large
- What value of d for what confidence level ?

d	α
1	0.68
1.64	0.90
1.96	0.95
2	0.9545
3	0.997
5	0.9999994

The "number of sigma" way of speaking

- d called in jargon the "**number of sigma**"
 - Example: a 2σ confidence interval is an interval with a confidence level of 95.45%

The "number of sigma" way of speaking

- d called in jargon the "**number of sigma**"
 - Example: a 2σ confidence interval is an interval with a confidence level of 95.45%
- **Note:** "number of sigma" terminology used even when problem is not normal
 - Quoting an interval with its corresponding "number of sigma" doesn't mean that underlying estimator is gaussian
 - Underlying estimator can have any distribution (e.g. gamma distribution)
 - The number of sigma is just a number that tells what the confidence level is

Approximate methods

- In cases other than the unbiased normal one, coverage in general not known
- However, the following holds in unbiased case:

$$\text{coverage} \geq 1 - \frac{1}{d^2}$$

Leading to:

d	1	2	3	4	5
coverage \geq	0	0,75	0,88	0,9375	0,96

- Note: **this is true in general for unbiased estimators !**

Bienaymé-Tchebichev inequality

- Bienaymé-Tchebichev inequality:

$$P(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\text{var}[X]}{\varepsilon^2} \quad \forall \varepsilon > 0$$

or equivalently

$$P(|X - \mathbb{E}[X]| \geq \varepsilon \sigma[X]) \leq \frac{1}{\varepsilon^2}$$

- 2 important consequences:

- ➊ Coverage bounded from below (result under discussion)
- ➋ Law of large numbers:

$$\mathbb{E}[M] = \mathbb{E}[X] \quad \text{and} \quad \text{var}[M] = \frac{\text{var}[X]}{n}$$

$$\Rightarrow P(|M - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\text{var}[X]}{n\varepsilon^2}$$

- Bienaymé-Tchebichev inequality:

$$P(|X - \mathbb{E}[X]| \geq \varepsilon \sigma[X]) \leq \frac{1}{\varepsilon^2}$$

- In unbiased case, this leads to:

$$P\left(\left|\hat{\theta} - \theta_0\right| \geq \varepsilon \sqrt{\text{var}[\hat{\theta}]}\right) \leq \frac{1}{\varepsilon^2}$$

$$\Leftrightarrow P\left(\left|\hat{\theta} - \theta_0\right| \leq \varepsilon \sqrt{\text{var}[\hat{\theta}]}\right) \geq 1 - \frac{1}{\varepsilon^2}$$

$$\Leftrightarrow P\left(\hat{\theta} - \varepsilon \sqrt{\text{var}[\hat{\theta}]} \leq \theta_0 \leq \hat{\theta} + \varepsilon \sqrt{\text{var}[\hat{\theta}]}\right) \geq 1 - \frac{1}{\varepsilon^2}$$

Approximate methods

- ☐ If variance not known, can use an estimator:

$$\theta \in \left[\hat{\theta} - d \sqrt{\widehat{\text{var}} [\hat{\theta}]} ; \hat{\theta} + d \sqrt{\widehat{\text{var}} [\hat{\theta}]} \right]$$

Approximate methods

- If variance not known, can use an estimator:

$$\theta \in \left[\hat{\theta} - d \sqrt{\widehat{\text{var}}[\hat{\theta}]} ; \hat{\theta} + d \sqrt{\widehat{\text{var}}[\hat{\theta}]} \right]$$

- Can be **very risky** to do so:
 - Coverage in general not known
 - Can lead to large undercoverage

Approximate methods

- If variance not known, can use an estimator:

$$\theta \in \left[\hat{\theta} - d \sqrt{\widehat{\text{var}}[\hat{\theta}]} ; \hat{\theta} + d \sqrt{\widehat{\text{var}}[\hat{\theta}]} \right]$$

- Can be **very risky** to do so:
 - Coverage in general not known
 - Can lead to large undercoverage
- Typical situations where such intervals are computed: calculation of proportions and efficiencies
 - What is the fraction of left-handed people in population ?
 - What percentage will such or such candidate get at the next presidential election ?
 - What is the detection efficiency of your device ?

Confidence interval for the proportion

- **Statistical model:** binomial law of probability

$$P(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

Confidence interval for the proportion

- **Statistical model:** binomial law of probability

$$P(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

- **Goal:** find confidence interval for p

Confidence interval for the proportion

- **Statistical model:** binomial law of probability

$$P(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

- **Goal:** find confidence interval for p
- **Possible solution:** start from ML estimator of p

$$\hat{p} = \frac{k}{N}$$

- Variance of \hat{p} is:

$$\text{var}[\hat{p}] = \frac{p(1-p)}{N}$$

Confidence interval for the proportion

- **Statistical model:** binomial law of probability

$$P(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

- **Goal:** find confidence interval for p
- **Possible solution:** start from ML estimator of p

$$\hat{p} = \frac{k}{N}$$

- Variance of \hat{p} is:

$$\text{var}[\hat{p}] = \frac{p(1-p)}{N}$$

- Problem: variance depends on unknown parameter of interest p
→ Rather than true variance, use an estimate:

$$\widehat{\text{var}}[\hat{p}] = \frac{\hat{p}(1-\hat{p})}{N}$$

Confidence interval for the proportion

- Following previous reasoning, the CI is:

$$p \in \left[\hat{p} - d \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}; \hat{p} + d \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \right] \quad (\text{Wald interval})$$

Confidence interval for the proportion

- Following previous reasoning, the CI is:

$$p \in \left[\hat{p} - d \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}; \hat{p} + d \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \right] \quad (\text{Wald interval})$$

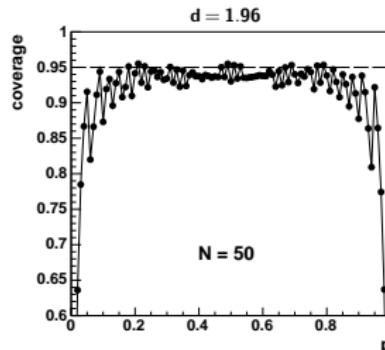
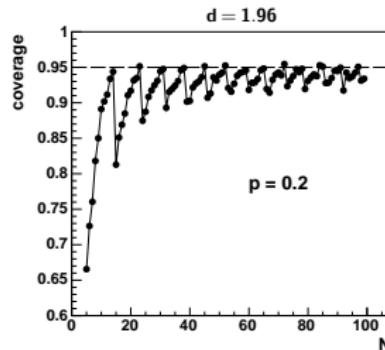
- What is the coverage of this CI ?

Confidence interval for the proportion

- Following previous reasoning, the CI is:

$$p \in \left[\hat{p} - d \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}; \hat{p} + d \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \right] \quad (\text{Wald interval})$$

- What is the coverage of this CI ?



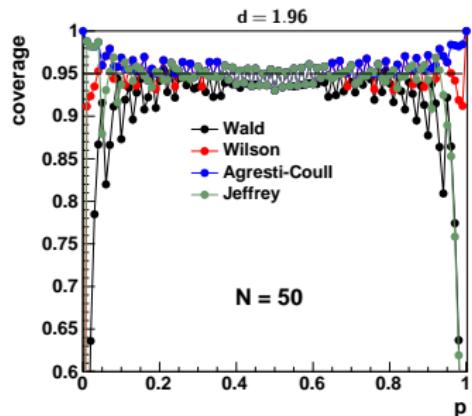
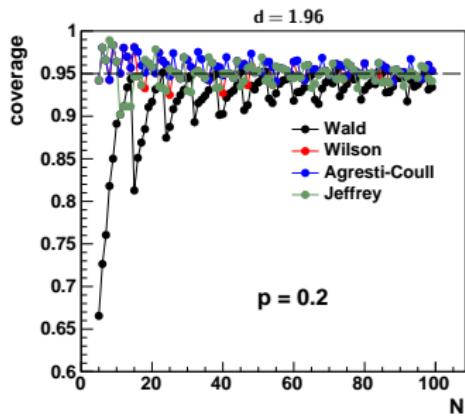
- Very large variations (in particular as a function of p)
- As you don't know p , you don't know coverage

Confidence interval for the proportion

- **Other issue with Wald interval:** leads to empty set when $\hat{p} \rightarrow 0$ or 1
 - **Example:** Suppose you ask $N = 2$ people whether they are left-handed or not
 - If both say no ($k = 0$), then the CI is: 0 ± 0
 - If both say yes ($k = 2$), then the CI is: 1 ± 0
- ⇒ Even though sample size very small ($N = 2$), we arrive at certain conclusions (uncertainty on estimated proportion is 0)

Confidence interval for the proportion

- Because of these issues, better intervals have been proposed over the years:



Exercice

Let's consider the following sample:

$$(20.4, 25.4, 25.6, 25.6, 26.6, 28.6, 28.7, 29, 29.8, 30.5, 30.9, 31.1)$$

We assume that these values are realizations of a normal random variable with mean μ and standard deviation σ (both unknown)

→ Find a confidence interval for μ at 95% CL

Exercice

For the exercice of the previous slide, is it possible to find a better interval than the one you found ?

In order to address this question, consider the fact that

$$t = \frac{\sqrt{n-1}(M - \mathbb{E}[X])}{s}, \quad \text{with} \quad s^2 = \frac{\sum_i (X_i - M)^2}{n}$$

follows a Student distribution with $n-1$ degrees of freedom.

Quantiles of Student distribution

k	γ										
	0.25	0.20	0.15	0.10	0.05	0.025	0.010	0.005	0.0025	0.0010	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073

Neyman construction

- ☐ Method for building CI with good frequentist properties (i.e. no undercoverage, minimal overcoverage): **Neyman construction**

Neyman construction

- Method for building CI with good frequentist properties (i.e. no undercoverage, minimal overcoverage): **Neyman construction**
- **Principle:**

Neyman construction

- Method for building CI with good frequentist properties (i.e. no undercoverage, minimal overcoverage): **Neyman construction**
- **Principle:**
 - ① For each θ , build acceptance region with probability α :

$$[x_{\min}(\theta); x_{\max}(\theta)]$$

Neyman construction

- Method for building CI with good frequentist properties (i.e. no undercoverage, minimal overcoverage): **Neyman construction**
- **Principle:**
 - ① For each θ , build acceptance region with probability α :

$$[x_{\min}(\theta); x_{\max}(\theta)]$$

- ② From all acceptance regions, build confidence belt

- Method for building CI with good frequentist properties (i.e. no undercoverage, minimal overcoverage): **Neyman construction**
- **Principle:**
 - ① For each θ , build acceptance region with probability α :

$$[x_{\min}(\theta); x_{\max}(\theta)]$$

- ② From all acceptance regions, build confidence belt
- ③ Determine CI for θ from observed value x_{obs}

$$[\theta_{\min}(x_{\text{obs}}); \theta_{\max}(x_{\text{obs}})]$$

Neyman construction

- Method for building CI with good frequentist properties (i.e. no undercoverage, minimal overcoverage): **Neyman construction**
- **Principle:**
 - ① For each θ , build acceptance region with probability α :

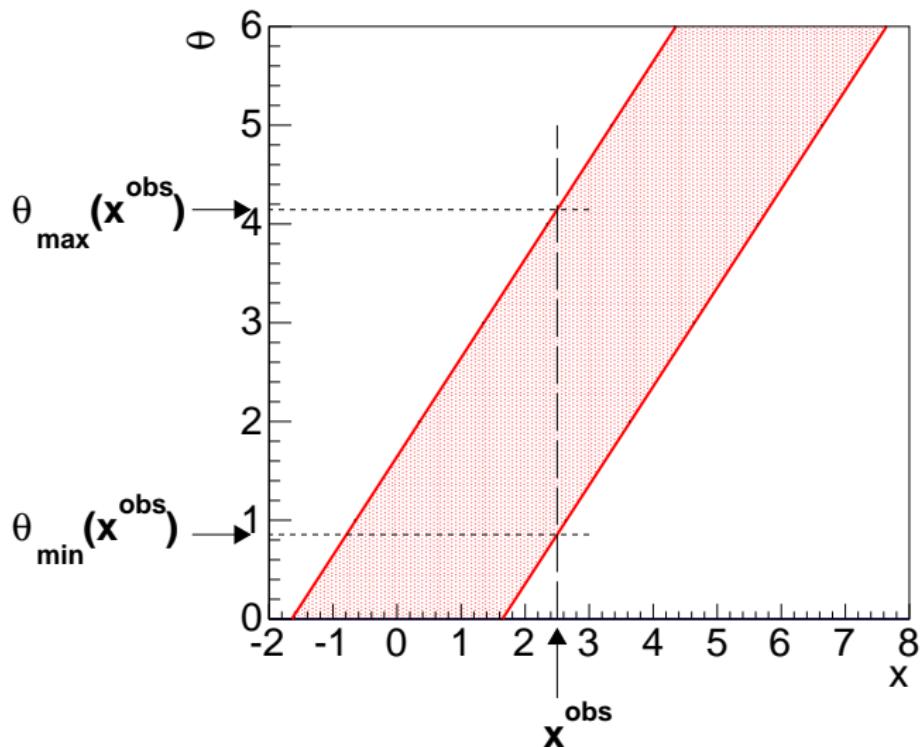
$$[x_{\min}(\theta); x_{\max}(\theta)]$$

- ② From all acceptance regions, build confidence belt
- ③ Determine CI for θ from observed value x_{obs}

$$[\theta_{\min}(x_{\text{obs}}); \theta_{\max}(x_{\text{obs}})]$$

→ Resulting interval has confidence level = α

Neyman construction



Remarks:

□ Remarks:

- Perfect coverage in continuous case

□ Remarks:

- Perfect coverage in continuous case
- Impossible to achieve perfect coverage in discrete case
 - Choose narrower CI that has coverage $> \alpha$ (i.e. minimal overcoverage)

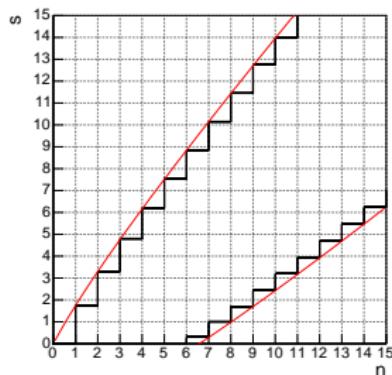
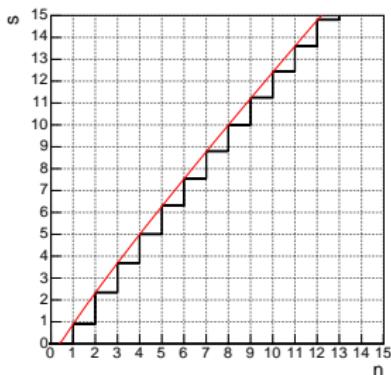
□ Remarks:

- Perfect coverage in continuous case
- Impossible to achieve perfect coverage in discrete case
 - Choose narrower CI that has coverage $> \alpha$ (i.e. minimal overcoverage)
- Choice of how to build acceptance region free
 - If $x_{\min}(\theta) = -\infty$ or $x_{\max}(\theta) = +\infty$: **one-sided** interval
 - If $x_{\min}(\theta) \neq -\infty$ and $x_{\max}(\theta) \neq +\infty$: **two-sided** interval
 - If $P(x < x_{\min}(\theta); \theta) = P(x > x_{\max}(\theta); \theta) = (1 - \alpha)/2$: **central interval**

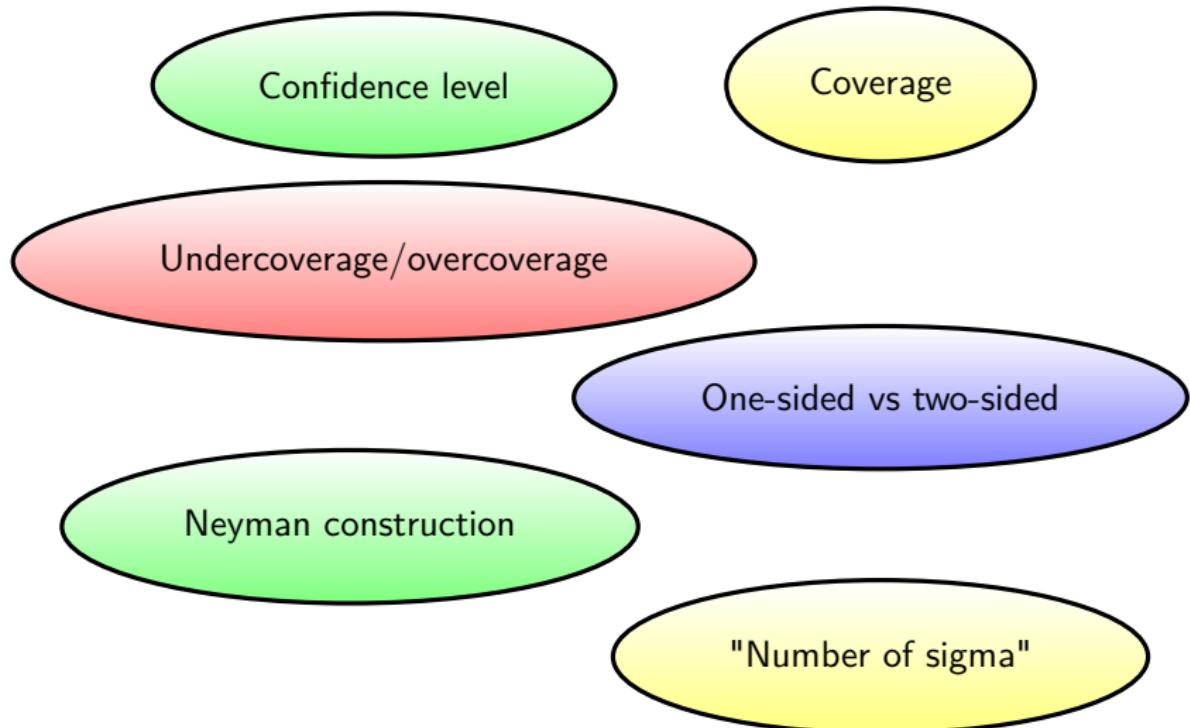
Neyman construction: discrete example

- ☐ Neyman construction for one-sided and two-sided intervals in Poisson case:

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



Key words/concepts



Chap. 6 Hypothesis testing

Introduction

- Often interested in questions such as:
 - Are my data well described by a linear function ?
 - Is the average age of people in europe the same as the one in south america ?
 - Is this patient affected by cancer ?
 - Are my data in favor of the standard model of particle physics or some other model ?
 - Is the particle I detect a photon or an electron ?
 - ...

Introduction

- Often interested in questions such as:
 - Are my data well described by a linear function ?
 - Is the average age of people in europe the same as the one in south america ?
 - Is this patient affected by cancer ?
 - Are my data in favor of the standard model of particle physics or some other model ?
 - Is the particle I detect a photon or an electron ?
 - ...

→ In order to answer, one has to invoke statistical techniques gathered under the name "**hypothesis testing**"

- Possible outcomes of hypothesis testing

		The truth	
		1	0
Your conclusion	1	true positive	false positive
	0	false negative	true negative

- Possible outcomes of hypothesis testing

		The truth	
		1	0
Your conclusion	1	true positive	false positive
	0	false negative	true negative

- General objective in hypothesis testing: derive conclusions that are as certain as possible
- Of course, impossible to be 100% sure of your conclusions
 - Always a chance that you're wrong
 - You need to decide what risk you're ready to take

Workflow of hypothesis testing

Workflow of hypothesis testing

- ① Clearly state what hypothesis you want to test (considered as true by default)

→ "null" hypothesis H_0

And, if needed, other alternative hypothesis you also want to confront the data to

→ "alternative" hypothesis H_1

Workflow of hypothesis testing

- ➊ Clearly state what hypothesis you want to test (considered as true by default)

→ "null" hypothesis H_0

And, if needed, other alternative hypothesis you also want to confront the data to

→ "alternative" hypothesis H_1

- ➋ Choose one (or more) variable on which the hypothesis test will be based

→ so-called "test statistic" (often written t)

Workflow of hypothesis testing

- ① Clearly state what hypothesis you want to test (considered as true by default)

→ "null" hypothesis H_0

And, if needed, other alternative hypothesis you also want to confront the data to

→ "alternative" hypothesis H_1

- ② Choose one (or more) variable on which the hypothesis test will be based

→ so-called "test statistic" (often written t)

- ③ Determine what values the test statistic typically get under the null and alternative hypotheses

Workflow of hypothesis testing

- ➊ Clearly state what hypothesis you want to test (considered as true by default)

→ "null" hypothesis H_0

And, if needed, other alternative hypothesis you also want to confront the data to

→ "alternative" hypothesis H_1

- ➋ Choose one (or more) variable on which the hypothesis test will be based

→ so-called "test statistic" (often written t)

- ➌ Determine what values the test statistic typically get under the null and alternative hypotheses
- ➍ Make your measurement and determine the observed value of the test statistic

Workflow of hypothesis testing

- ➊ Clearly state what hypothesis you want to test (considered as true by default)

→ "null" hypothesis H_0

And, if needed, other alternative hypothesis you also want to confront the data to

→ "alternative" hypothesis H_1

- ➋ Choose one (or more) variable on which the hypothesis test will be based

→ so-called "test statistic" (often written t)

- ➌ Determine what values the test statistic typically get under the null and alternative hypotheses
- ➍ Make your measurement and determine the observed value of the test statistic
- ➎ Compare observed value to typical values under null and alternative hypotheses and conclude

Workflow of hypothesis testing

- ➊ Clearly state what hypothesis you want to test (considered as true by default)

→ "null" hypothesis H_0

And, if needed, other alternative hypothesis you also want to confront the data to

→ "alternative" hypothesis H_1

- ➋ Choose one (or more) variable on which the hypothesis test will be based

→ so-called "test statistic" (often written t)

- ➌ Determine what values the test statistic typically get under the null and alternative hypotheses
- ➍ Make your measurement and determine the observed value of the test statistic
- ➎ Compare observed value to typical values under null and alternative hypotheses and conclude

→ Purpose of this chapter: detail all these steps !

Central notion: **test statistic**

- Word "**test statistic**" may seem strange and difficult to understand at first glance

Central notion: **test statistic**

- Word "**test statistic**" may seem strange and difficult to understand at first glance
- Not as difficult as it seems
 - As first approximation: "**test statistic**" *synonym* "**random variable**"
 - But not any random variable: one that is suited for hypothesis testing

Central notion: **test statistic**

- Word "**test statistic**" may seem strange and difficult to understand at first glance
 - Not as difficult as it seems
 - As first approximation: "**test statistic**" synonym "**random variable**"
 - But not any random variable: one that is suited for hypothesis testing
 - **What is a r.v. suited for hyp. testing ?**
 - It is a r.v. sensitive to the hypotheses under test (H_0 and H_1)
 - It behaves differently under H_0 and H_1
- Test statistic = **discriminating random variable**
- Test statistics chosen so as to offer **discrimination power** as good as possible

Test statistic: examples

- **Example 1:** Are my data distributed according to the normal distribution with mean = 0 ?
 - Use sample mean as test statistic

Test statistic: examples

- **Example 1:** Are my data distributed according to the normal distribution with mean = 0 ?
 - Use sample mean as test statistic

- **Example 2:** Is the activity of a given radioactive source greater than 10 MBq ?
 - Use number of counts as test statistic

Test statistic: examples

- **Example 1:** Are my data distributed according to the normal distribution with mean = 0 ?
 - Use sample mean as test statistic
- **Example 2:** Is the activity of a given radioactive source greater than 10 MBq ?
 - Use number of counts as test statistic
- Test statistic can be either:
 - An observable (as in example 2)
 - A function of the observables (as in example 1)

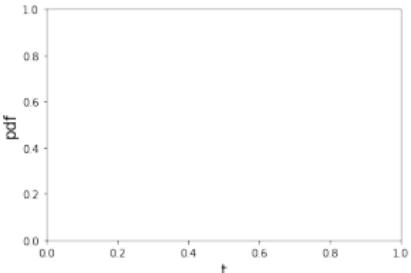
Workflow of hyp. testing in pictures

- ➊ Clearly state what H_0 and H_1 are

Workflow of hyp. testing in pictures

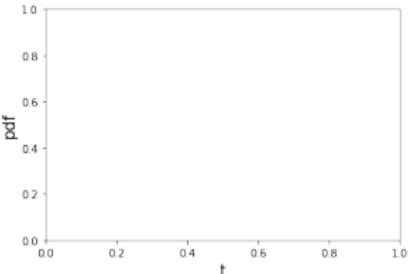
- ➊ Clearly state what H_0 and H_1 are
- ➋ Choose test statistics t

$$t(x_1, \dots, x_n)$$



Workflow of hyp. testing in pictures

- 1 Clearly state what H_0 and H_1 are

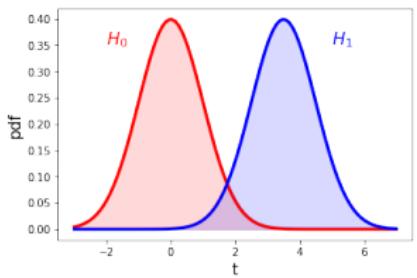


- 2 Choose test statistics t

$$t(x_1, \dots, x_n)$$

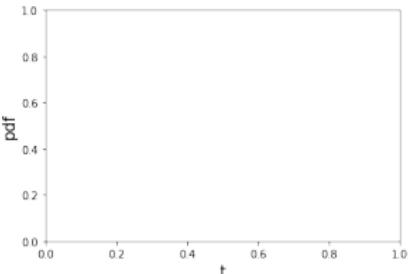
- 3 Determine distributions of t under H_0 and H_1

$$f(t|H_0) \quad \text{and} \quad f(t|H_1)$$



Workflow of hyp. testing in pictures

- ➊ Clearly state what H_0 and H_1 are

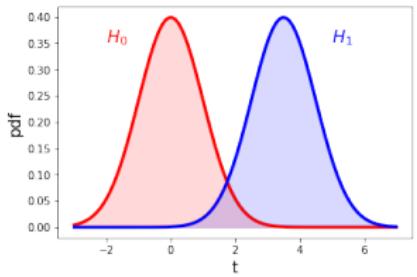


- ➋ Choose test statistics t

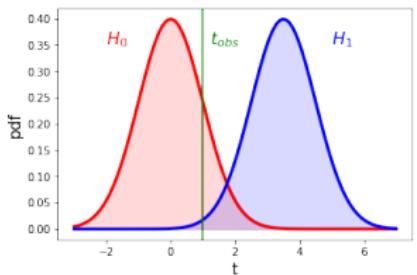
$$t(x_1, \dots, x_n)$$

- ➌ Determine distributions of t under H_0 and H_1

$$f(t|H_0) \quad \text{and} \quad f(t|H_1)$$

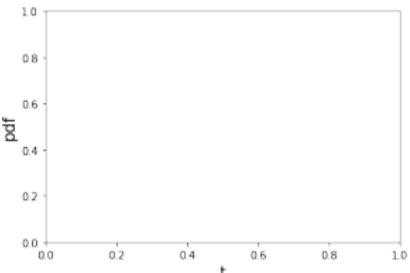


- ➍ Make measurement and calculate t_{obs}



Workflow of hyp. testing in pictures

- ➊ Clearly state what H_0 and H_1 are

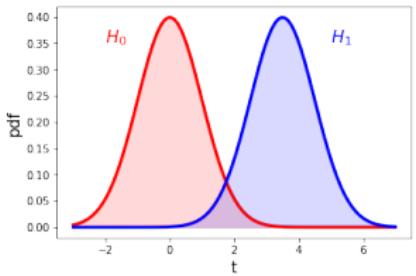


- ➋ Choose test statistics t

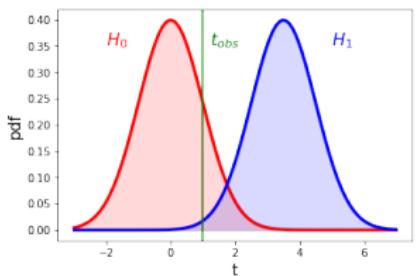
$$t(x_1, \dots, x_n)$$

- ➌ Determine distributions of t under H_0 and H_1

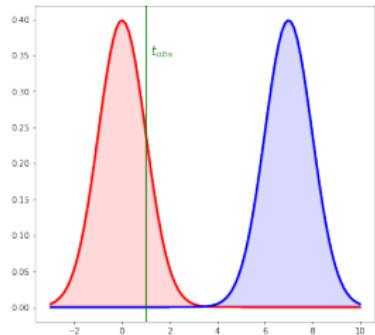
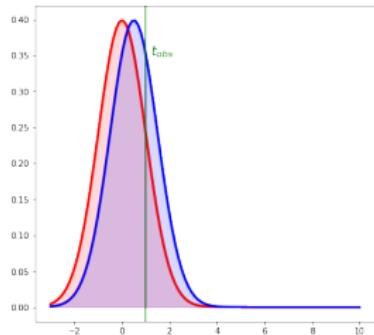
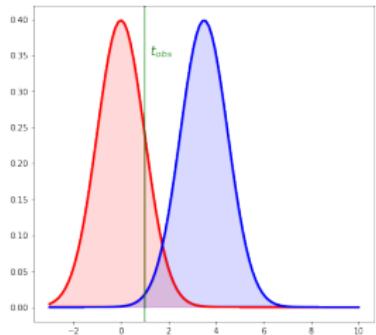
$$f(t|H_0) \quad \text{and} \quad f(t|H_1)$$



- ➍ Make measurement and calculate t_{obs}
- ➎ Conclude



Discrimination power: illustration



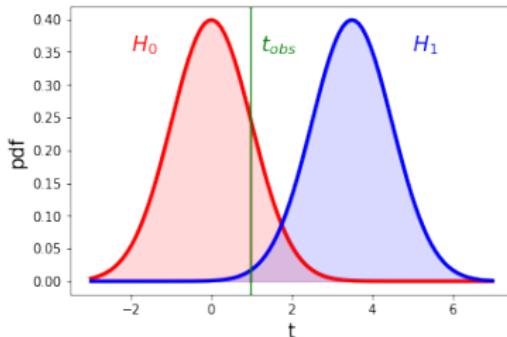
A word of caution

- Word "test statistic" often shortened to "test" or "statistic"
- You may hear/read formulations such as:
 - For my test I use the sample mean as test.
 - For my test I use the sample mean as statistic.
 - The statistic is the sample mean.
 - The distribution of the test is ...
 - The distribution of the statistic is ...



Don't be confused ! Make sure you understand what this means !

- How do we conclude once we have the following plot ?



→ Need to have a quantitative measure of agreement between observation (t_{obs}) and hypotheses

Measure of agreement between observation and hypotheses

- Agreement measured with ***p-value***

- **Definition:**

p-value = probability to observe what you observed
in measurement or "more extreme" values

- Meaning of "more extreme" depends on context:

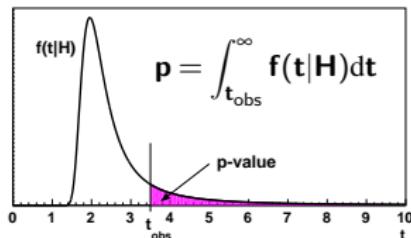
Measure of agreement between observation and hypotheses

- Agreement measured with **p-value**

- **Definition:**

p-value = probability to observe what you observed in measurement or "more extreme" values

- Meaning of "more extreme" depends on context:
 - If only large values are considered a sign of disagreement:



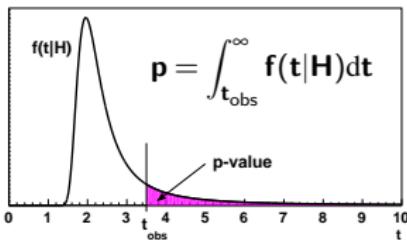
Measure of agreement between observation and hypotheses

- Agreement measured with **p-value**

- **Definition:**

p-value = probability to observe what you observed in measurement or "more extreme" values

- Meaning of "more extreme" depends on context:
 - If only large values are considered a sign of disagreement:



- If only low values: $p = \int_{-\infty}^{t_{\text{obs}}} f(t|H) dt$

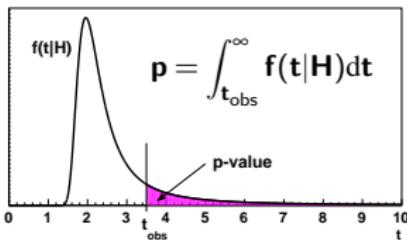
Measure of agreement between observation and hypotheses

- Agreement measured with **p-value**

- **Definition:**

p-value = probability to observe what you observed in measurement or "more extreme" values

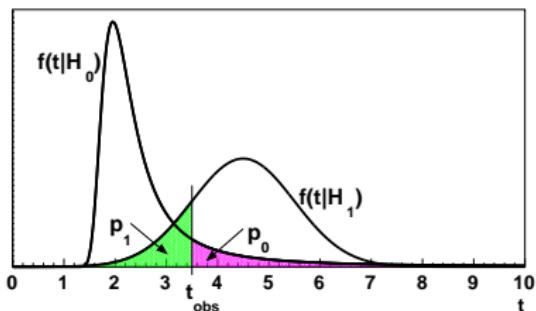
- Meaning of "more extreme" depends on context:
 - If only large values are considered a sign of disagreement:



- If only low values: $p = \int_{-\infty}^{t_{\text{obs}}} f(t|H)dt$
- If both: two-sided definition of *p-value*

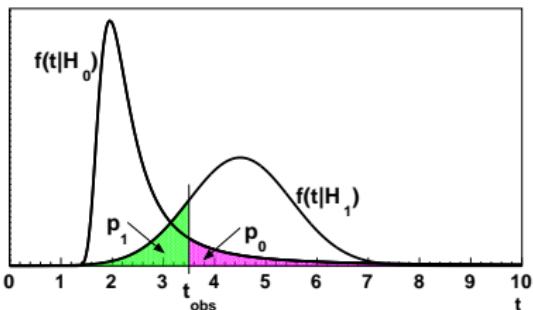
More on *p*-values

- With two hypotheses, the *p*-values look like



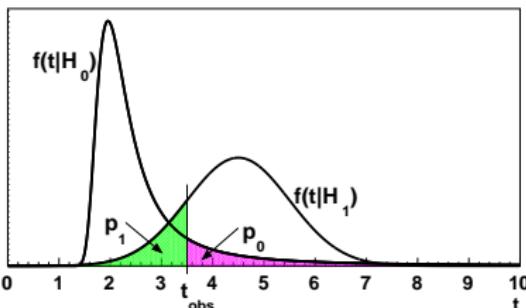
More on *p*-values

- With two hypotheses, the *p*-values look like



- p*-values are random variables
→ How are they distributed ?

Distribution of p -values



- Suppose $t \sim H_0$ and let $g(p_0|H_0)$ be the distribution of p_0 under H_0 :

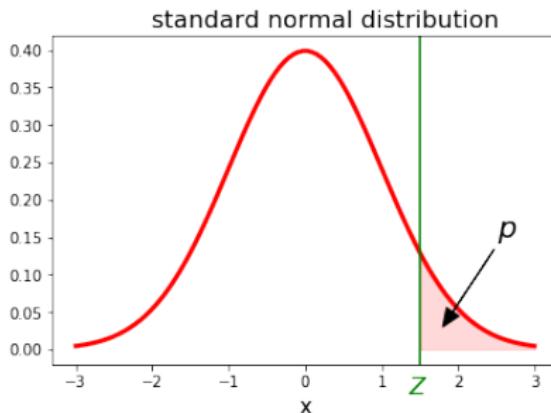
$$g(p_0|H_0) = f(t_{\text{obs}}|H_0) \frac{1}{|dp_0/dt_{\text{obs}}|}$$

- $p_0 = \int_{t_{\text{obs}}}^{\infty} f(t|H_0) dt = 1 - F(t_{\text{obs}}|H_0)$
 $\Rightarrow dp_0 = -dF(t_{\text{obs}}|H_0) = -f(t_{\text{obs}}|H_0) dt_{\text{obs}}$

- **Conclusion:** $g(p_0|H_0) = 1 !!!$

Significance

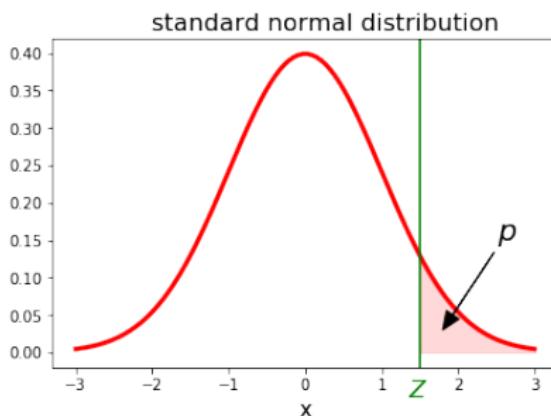
- Rather than *p-value*, one sometimes uses the **significance Z**



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

Significance

- Rather than *p-value*, one sometimes uses the **significance Z**

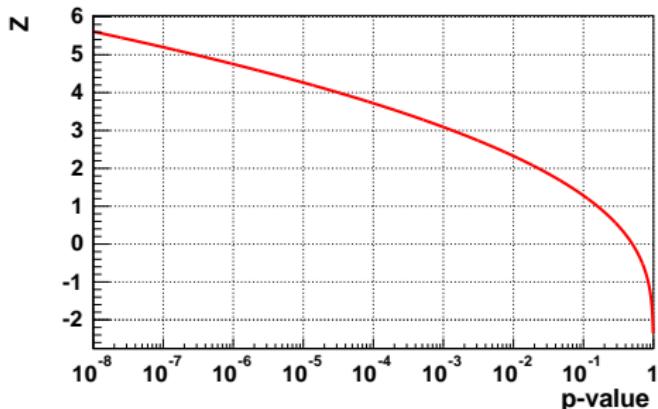


$$Z = \Phi^{-1}(1 - p)$$

- **Important remark:** using Z rather than *p-value* to report values doesn't mean that the test follows normal distribution
 - Test distribution can be anything
 - Using Z just means performing a simple change of variable

Significance

- Why is this change of variable interesting ?



p-value	Z
0.05	1,64
0.00135	3
$2,87 \times 10^{-7}$	5

8 orders of magnitude in $p \Leftrightarrow$ 1 order of magnitude in Z

- Terminology: when $Z = X$, we say that the "significance is $X\sigma$ "

Exercice

Let's consider a Poisson counting experiment with one signal and one or more background processes. Let b be the total number of expected background events and N_{obs} the observed number of events. We consider the case $N_{\text{obs}} > b$. Show that, in the asymptotic limit and under the background only hypothesis, the significance of the observation is

$$Z \simeq \frac{\hat{s}}{\sqrt{b}}$$

where $\hat{s} = N_{\text{obs}} - b$ is the estimator of the number of signal events.

How do we conclude eventually ?

- We choose *a priori* a threshold value for the *p-value*: α

How do we conclude eventually ?

- We choose *a priori* a threshold value for the *p-value*: α
 - If $p < \alpha$: observation considered too extreme to be compatible with hypothesis
 - **Hypothesis is rejected**

How do we conclude eventually ?

- We choose *a priori* a threshold value for the *p-value*: α
 - If $p < \alpha$: observation considered too extreme to be compatible with hypothesis
 - **Hypothesis is rejected**
 - If $p > \alpha$: observation considered compatible with hypothesis
 - **Hypothesis is accepted**

How do we conclude eventually ?

- We choose *a priori* a threshold value for the *p-value*: α
 - If $p < \alpha$: observation considered too extreme to be compatible with hypothesis
 - **Hypothesis is rejected**
 - If $p > \alpha$: observation considered compatible with hypothesis
 - **Hypothesis is accepted**
- Terminology: α called the **size** or **significance level** of the test

Exercice

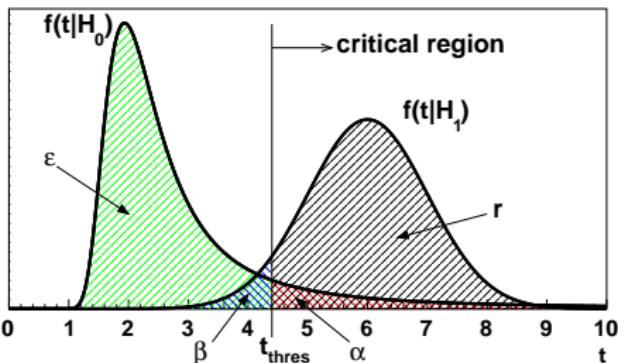
Suppose that, during a certain period of time in a certain population, 49581 boys and 48870 girls were born. Are these observations in favor of the hypothesis according to which the fractions of male birth and female birth are equal to 50% at the 5% level ?

Standard normal distribution cdf values

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

Equivalent approach

- Rather than using the *p-value* to conclude, we can use the value of the test t_{obs} itself



- If observation in critical region ($p < \alpha$): null hypothesis rejected
- If observation not in critical region ($p > \alpha$): null hypothesis accepted

How is the size of the test α chosen ?

- Complex question:

No general answer → very problem dependent

- It depends on:
 - The nature of the test you're carrying, of the null and alternative hypothesis
 - The consequences of the different possible conclusions
 - In particular the consequences of deriving false conclusions (what happens in case of false negative or false positive ?)
 - The risk you're ready to take to suffer the consequences

Choice of α : example 1

- Testing well established hypothesis** (e.g. Einstein's theory of relativity):

- *A priori* very confident in hypothesis
 - Rejecting it would be a major event
 - Need very strong evidence to reject hypothesis
 - ⇒ Choose small size
 - Typical value is $\alpha = 2.87 \times 10^{-7}$
 - Requires "5 σ observation" to reject hypothesis

Cancer diagnosis:

- If you conclude that your patient is not affected by cancer while he is (false negative)
→ Enormous consequences
- If you conclude that your patient is affected by cancer while he isn't (false positive)
→ Important consequences too

⇒ **Goal:** minimize false negative and false positive probabilities

Problem: not possible to lower false negative and false positive probabilities to arbitrary low values at the same time

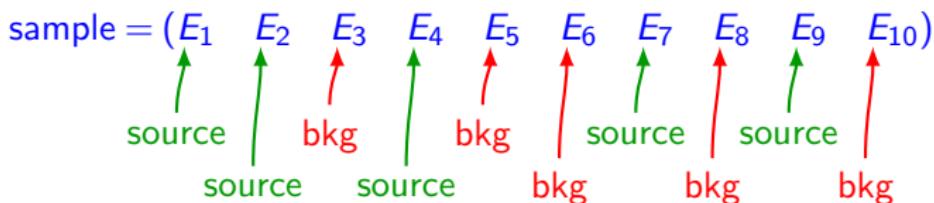
→ Always a trade-off between the two

→ Do you prefer minimizing false negatives or false positives ?

Choice of α : example 3

Selection of pure samples:

- Occurs when you have composite samples and want to measure some properties of only one type of event appearing in the mixture
- Example: measurement of radioactive source:

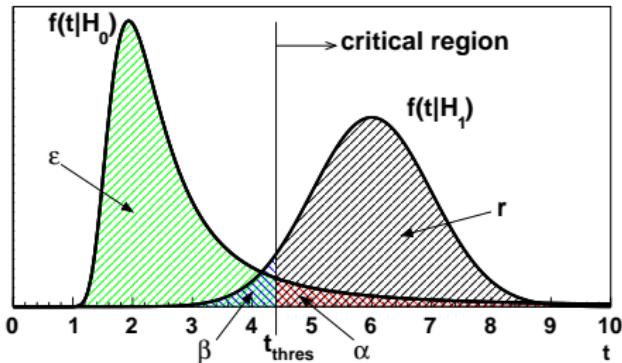


→ From this sample, how to build a sample enriched in events from the source ?

In other words: How to reject the background as much as possible while keeping signal events ?

- ## In next slides, will describe how to maximize purity

Type-I and type-II errors (and related quantities)

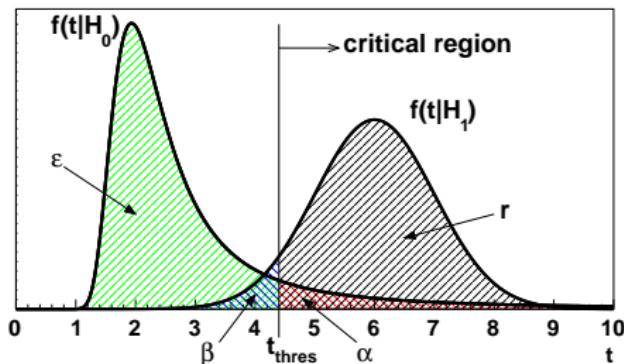


- Type-I error (error of the first kind)** = reject of H_0 if it's true
 - Probability of type-I error = α
- Type-II error (error of the second kind)** = accept H_0 if it's false
 - Probability of type-II error = β
- Related quantities:**
 - Efficiency: $\varepsilon = 1 - \alpha$
 - Power (or rejection): $r = 1 - \beta$

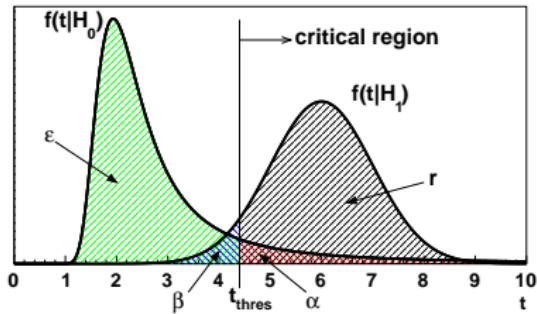
Purity

- By definition:

purity = fraction of events from H_0 in selected sample
 $= P(H_0 | t \notin \mathcal{C}) \quad (\mathcal{C} = \text{critical region})$



Purity calculation

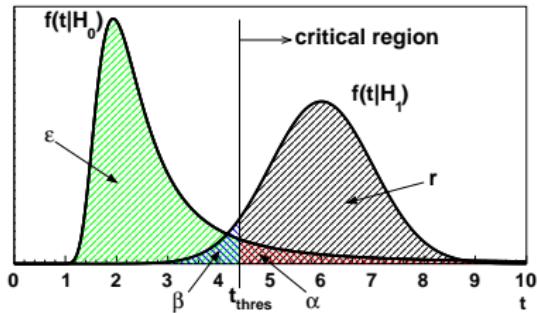


□ Notations:

- $c = P(H_0)$
- $1 - c = P(H_1)$

$$f(t) = c \times f(t|H_0) + (1 - c) \times f(t|H_1)$$

Purity calculation



□ Notations:

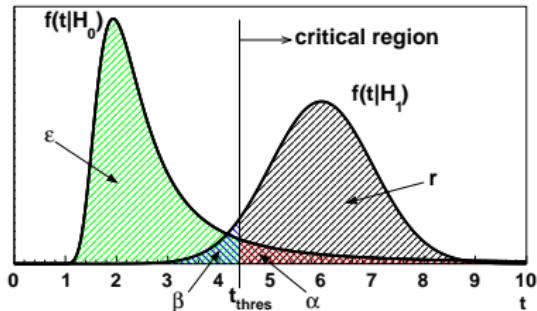
- $c = P(H_0)$
- $1 - c = P(H_1)$

$$f(t) = c \times f(t|H_0) + (1 - c) \times f(t|H_1)$$

□ Purity:

$$P(H_0 | t \notin \mathcal{C}) = \frac{P(t \notin \mathcal{C} | H_0) P(H_0)}{P(t \notin \mathcal{C})} = \dots = \frac{1}{1 + \frac{\beta}{\varepsilon} \frac{1-c}{c}}$$

Purity calculation



□ Notations:

- $c = P(H_0)$
- $1 - c = P(H_1)$

$$f(t) = c \times f(t|H_0) + (1 - c) \times f(t|H_1)$$

□ Purity:

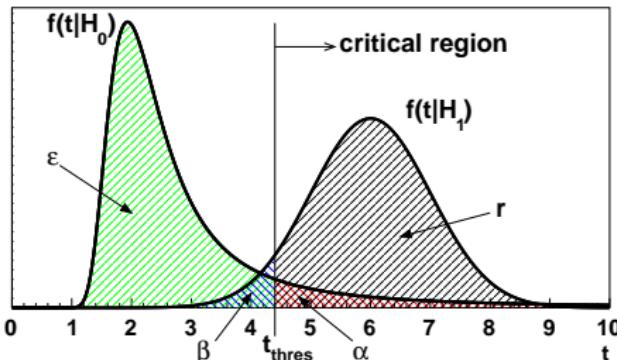
$$P(H_0 | t \notin \mathcal{C}) = \frac{P(t \notin \mathcal{C} | H_0)P(H_0)}{P(t \notin \mathcal{C})} = \dots = \frac{1}{1 + \frac{\beta}{\varepsilon} \frac{1-c}{c}}$$

□ Conclusion: purity maximized by maximizing the ratio ε/β

→ For fixed ε (α), must maximize power

Purity

- **Conclusion:** for fixed ε (α), must maximize power



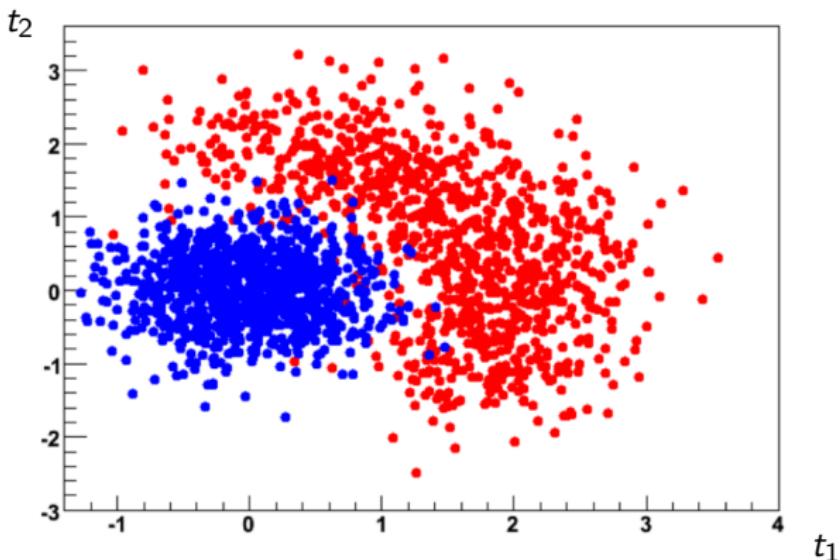
- **1D case:**

- No degrees of freedom once size fixed
- If varying the size is acceptable, maximize ratio ε/β

- **2D (or higher) case:**

- More complex (see next slide)

Purity: 2D (or higher) case



- No unique critical region for a given size α
- Must select critical region that maximizes power

→ How do we do that ?

Neyman-Pearson lemma

- Critical region that maximizes power for fixed α given by:

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

- Critical region that maximizes power for fixed α given by:

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

- Notes:

- Critical region that maximizes power for fixed α given by:

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

- Notes:

- t is a multidimensional object: $t = (t_1, t_2, \dots)$

- Critical region that maximizes power for fixed α given by:

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

- Notes:

- t is a multidimensional object: $t = (t_1, t_2, \dots)$
- $\frac{f(t|H_0)}{f(t|H_1)}$ called **likelihood ratio**
 - Traditionally noted Λ

- Critical region that maximizes power for fixed α given by:

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

- Notes:

- t is a multidimensional object: $t = (t_1, t_2, \dots)$
- $\frac{f(t|H_0)}{f(t|H_1)}$ called **likelihood ratio**
 - Traditionally noted Λ
- k_α is a function of α

Neyman-Pearson lemma

- Critical region that maximizes power for fixed α given by:

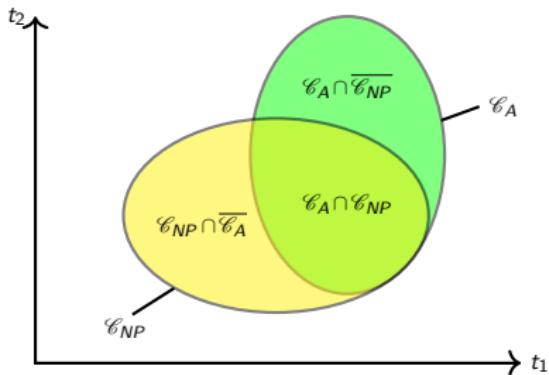
$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

- Notes:

- t is a multidimensional object: $t = (t_1, t_2, \dots)$
- $\frac{f(t|H_0)}{f(t|H_1)}$ called **likelihood ratio**
 - Traditionally noted Λ
- k_α is a function of α

- Neyman-Pearson critical region denoted as \mathcal{C}_{NP} and referred to as the **Best Critical Region (BCR)**

Neyman-Pearson lemma: proof



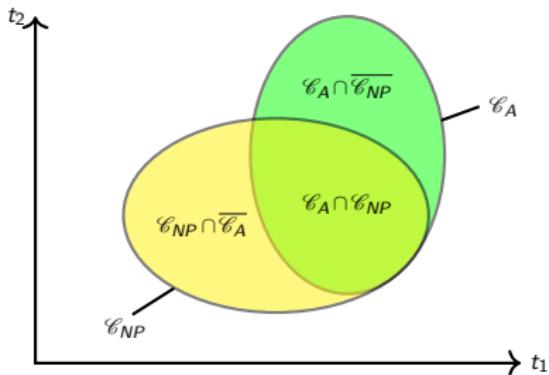
□ \mathcal{C}_A and \mathcal{C}_{NP} have same size

$$P(C_A|H_0) = P(C_{NP}|H_0) = \alpha$$

□ Must prove that

$$P(C_A|H_1) \leq P(C_{NP}|H_1) \quad \text{when } \Lambda \leq k_\alpha$$

Neyman-Pearson lemma: proof



□ \mathcal{C}_A and \mathcal{C}_{NP} have same size

$$P(C_A|H_0) = P(C_{NP}|H_0) = \alpha$$

□ Must prove that

$$P(C_A|H_1) \leq P(C_{NP}|H_1) \quad \text{when } \Lambda \leq k_\alpha$$

□ Proof:

$$P(C_A|H_1) \leq P(C_{NP}|H_1) \Leftrightarrow P(C_A \cap \overline{C_{NP}}|H_1) \leq P(C_{NP} \cap \overline{C_A}|H_1)$$

If lemma true:

$$P(C_{NP} \cap \overline{C_A}|H_1) \geq \frac{1}{k_\alpha} P(C_{NP} \cap \overline{C_A}|H_0) = \frac{1}{k_\alpha} P(C_A \cap \overline{C_{NP}}|H_0) \geq P(C_A \cap \overline{C_{NP}}|H_1)$$

Thus $P(C_A|H_1) \leq P(C_{NP}|H_1)$

Neyman-Pearson lemma

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

Neyman-Pearson lemma

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

- Neyman-Pearson lemma allows reformulation of multidimensional problems in simpler terms

$$ND \longrightarrow 1D$$
$$t = (t_1, t_2, \dots) \longrightarrow \Lambda = \frac{f(t|H_0)}{f(t|H_1)}$$

Neyman-Pearson lemma

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha$$

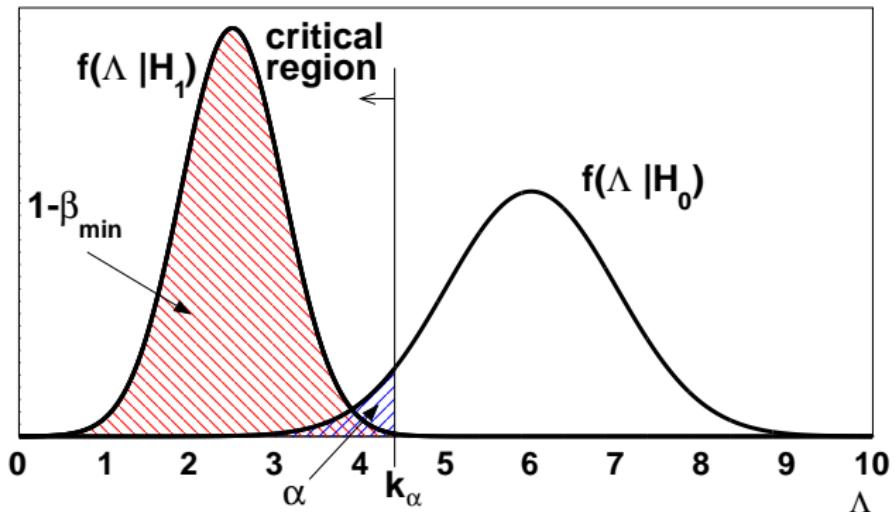
- Neyman-Pearson lemma allows reformulation of multidimensional problems in simpler terms

$ND \longrightarrow 1D$

$$t = (t_1, t_2, \dots) \longrightarrow \Lambda = \frac{f(t|H_0)}{f(t|H_1)}$$

- $1D$ machinery described previously can be employed using Λ as test statistics
 - Automatically leads to "optimal" hypothesis tests
 - **Λ is the most discriminating variable**

Neyman-Pearson lemma



Exercice

Let (X_1, X_2, \dots, X_n) be a set of n i.i.d. variables distributed according to a gaussian distribution with mean equal to μ and variance equal to 1. We consider the two following hypotheses:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu = \mu_1 > \mu_0$

Show that the BCR region is given by

$$\bar{x} \geq \frac{\mu_0 + \mu_1}{2} + \frac{1}{n} \frac{\ln k_\alpha}{\mu_0 - \mu_1}$$

where \bar{x} is the sample mean and k_α a constant depending only of the size of the test α .

Link between confidence interval building and hypothesis testing

- Confidence interval building can be done using hypothesis testing language

Link between confidence interval building and hypothesis testing

- Confidence interval building can be done using hypothesis testing language
- Example:** CI for mean of normal distribution with known variance

Link between confidence interval building and hypothesis testing

- Confidence interval building can be done using hypothesis testing language
- **Example:** CI for mean of normal distribution with known variance
 - Sample: (x_1, \dots, x_n)

Link between confidence interval building and hypothesis testing

- Confidence interval building can be done using hypothesis testing language
- **Example:** CI for mean of normal distribution with known variance
 - Sample: (x_1, \dots, x_n)
 - Test statistic: sample mean $M = \frac{1}{n} \sum x_i$

Link between confidence interval building and hypothesis testing

- Confidence interval building can be done using hypothesis testing language
- **Example:** CI for mean of normal distribution with known variance
 - Sample: (x_1, \dots, x_n)
 - Test statistic: sample mean $M = \frac{1}{n} \sum x_i$
 - $M \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$

Link between confidence interval building and hypothesis testing

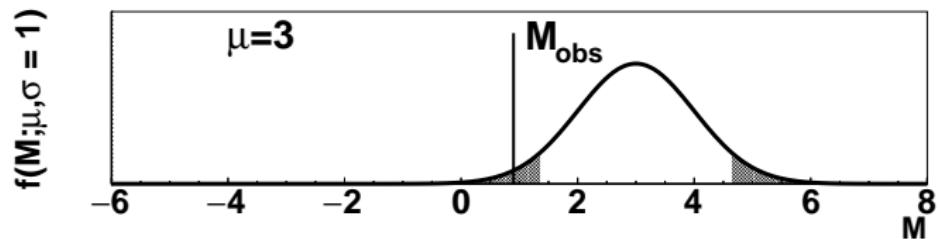
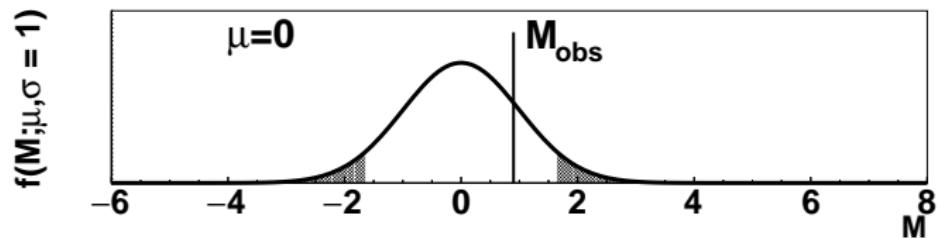
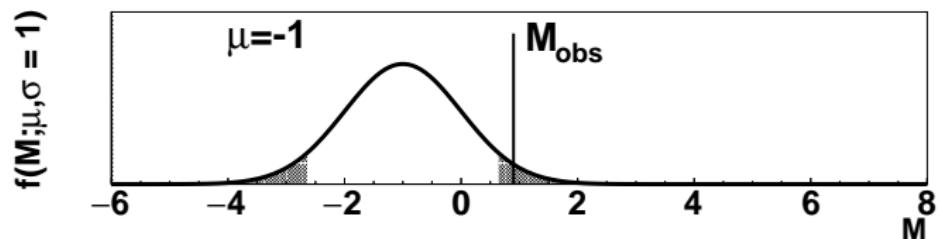
- Confidence interval building can be done using hypothesis testing language
- **Example:** CI for mean of normal distribution with known variance
 - Sample: (x_1, \dots, x_n)
 - Test statistic: sample mean $M = \frac{1}{n} \sum x_i$
 - $M \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
 - Perform hypothesis test for each value of μ and, based on the observed value of M , accept or reject these values

Link between confidence interval building and hypothesis testing

- Confidence interval building can be done using hypothesis testing language
- **Example:** CI for mean of normal distribution with known variance
 - Sample: (x_1, \dots, x_n)
 - Test statistic: sample mean $M = \frac{1}{n} \sum x_i$
 - $M \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
 - Perform hypothesis test for each value of μ and, based on the observed value of M , accept or reject these values

CI made of all values of μ not rejected in hypothesis test

Link between confidence interval building and hypothesis testing



Link between confidence interval building and hypothesis testing

- This way of building confidence intervals sometimes called "**hypothesis test inversion**"
- It is strictly equivalent to Neyman construction with

$$\text{confidence level} = 1 - \text{size of test}$$

→ Hyp. test inversion thus leads to good coverage properties

- Note that notations can be confusing
 - In confidence interval chapter: α =confidence level
 - In this chapter: α =size of test

Key words/concepts

