Statistical methods for Data Analysis

in particle physics

an introduction

Samuele Carli

September 5, 2013

Contents

Data analysis: Statistics and probability

- **Data analysis** is a *process* to transform raw data in usable information

$$\boxed{\text{RAW DATA}} \xrightarrow{\text{Data analysis}} \boxed{\text{INFORMATION}}$$

- **Statistics** is an instrument to perform presentation and interpretation of data
  - Descriptive statistics: describes main features of a collection of data
  - Inductive statistics: makes inference about a random process based on observation during a finite amount of time
- Probability theory is the mathematical foundation for statistics

Confirmatory and exploratory data analysis

- **Exploratory** data analysis: explores data to find new hypothesis to test
  - Suggest hypothesis about causes of observed phenomena
  - Asses assumptions on which statistical inference will be based
  - Select appropriate statistical tools and techniques
  - Eventually suggest further data collection
- **Confirmatory** data analysis: statistical hypothesis testing
  Used to make statistical decisions on top of experimental data
  - **Frequentist** hypothesis testing: Hypothesis is either true or false
  - **Bayesian** inference: degree of belief in truthfulness of hypothesis

Experimental vs observational studies

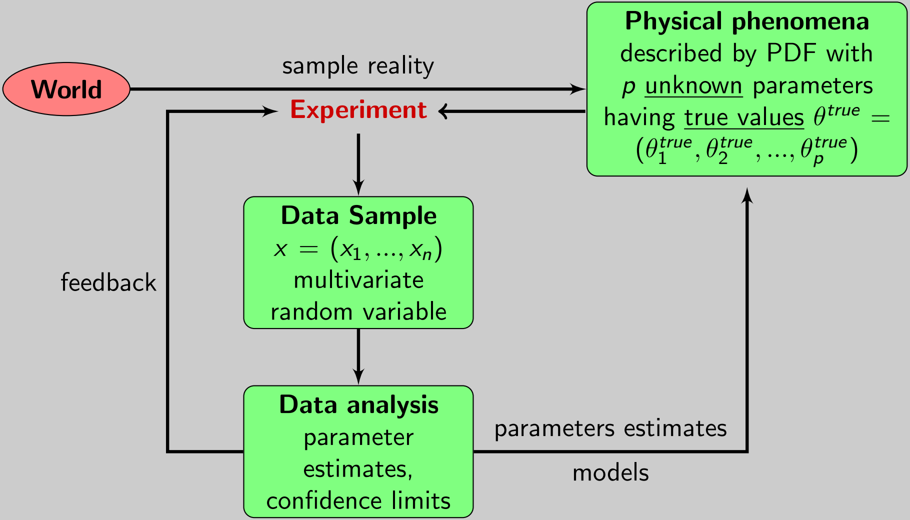
- **Experimental studies**

  Measure the system → Manipulate the system → Measure again, compare

  Example: study if and how free coffee will improve students' performance

- **Observational studies**: no experimental manipulation, *only gather and analyze data!*
  Example: Study correlation between number of beers drunk on Wednesday evening and performance on exam taken the day after
  - Be careful who pays! (expected results can be induced through inappropriate manipulation)

General picture



Data analysis in particle physics

- Observe events of a certain kind (particle collisions)
- Measure characteristics of each event
- Theory (SM) predicts distribution of this properties up to some free parameters

Hence one has to:

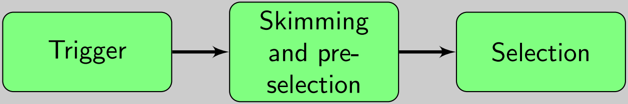
1. Estimate (measure) the parameters
2. Quantify the uncertainty of the parameters estimates
3. Test the extent to which a theory's predictions is in agreement with the data

Signal vs background(s)

- **Signal**: event coming from the physical process under study (for example $H \to ZZ \to e^+e^-e^+e^-$)
- **Background**: any other event
  - Trivial: any event which is not producing four electrons as final state
  - Dangerous: any process which can give four electrons in the final state; Any inaccuracy which results in the detection of four electrons (instead of three and a shower, for example)
    Example: signal $pp \to H \to ZZ \to 4e$, background $pp \to ZZ \to 4e$

Separating signal and background

- Be aware:
    - Nature is probabilistic: for a given event it's not possible to tell whether it's signal or backgroud
    - We can only make educated guess:
      $$p(event|signal),\ p(event|background)$$
- Separate as much as possible signal from background events $\rightarrow$ clean **reduced sample**



- Often we have to find **maximum reduction of background for given signal acceptance**

Exploring the data

- After data is collected $\rightarrow$ exploratory data analysis
- Example: **data reduction** (skimming and preselection)
  - Goal: get rid of unuseful events
  - Unuseful is not uniquely defined: some background events are interesting for control and measurement (detector calibration, etc.)
  - LHC-CMS example:
    - $\sim 10^9$ events/year (after trigger!)
    - $\sim 1MB$ per event
    - $\Rightarrow \sim 1PB$/year
  - Interesting physical processes are rare
    - 10 H$\rightarrow$ZZ$\rightarrow$4e events/year
    - Difficult not to lose too many signal events when skimming!

Exploring the data (cont.)

Skimming and preselection are quite different processes depending on purpose:

- Measure properties of a particle
- Measure frequency of decay
- Explore possible hypotheses
- Test existing hypotheses

Skimming and preselection are post-mortem processes that can be corrected and reprocessed.

Trigger is critical

- Hardware system which decides which event to store
- If event not stored, there is no way back! It can influence the whole analysis!
- Multi-level decision based on small subset of inaccurate data from very fast detectors
- ATLAS trigger LV1 decides in $\approx 2\mu s$ including cable delays
- Reduces stored events from 15 MHz to 70 KHz, whole trigger goes down to 500Hz

Contents

Mathematical probability

- Define Ω as an **exclusive** set of all possible elementary events x_i
  Exclusiveness: the occurrence of x_i implies none of the others occurs
- $P(x_i)$ probability of occurrence of x_i, such that:
  a. $P(x_i) \geq 0 \quad \forall i$
  b. $P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$
  c. $\sum_\Omega P(x_i) = 1$
- This is the base for more complex expressions:
  - non-elementary events (i.e. sets of elementary events)
  - non-exclusive events (i.e. overlapping sets of elementary events)

Frequentist probability

- Experiment:
  - N events observed
  - Out of them n is of type x
- **Frequentist probability** that an event will be of type x:

$$P(x) = \lim_{N \to \infty} \frac{n}{N}$$

- Important restriction: can only be applied to repeatable experiments
  - Ex. cannot define probability that it will snow tomorrow
  - Note that the job of a scientist is to try to get as close as possible to repeatable experiments

...more on Frequentist probability

- Probabilities are only associated with data: outcomes of repeatable observations
- P(Higgs boson exists) or $P(0.1 < x < 0.2)$ are either 0 or 1, but we don't know which.
  (Frequentist statistics tools not suitable for this)
- Tools of frequentist statistics tell what to expect, under the assumption of certain probabilities, about hypothetical repeated observations:
  Preferred theories are those for which observatios would be considered "usual"

Bayesian probability

- Based on the concept of "degree of belief"
- Operational definition (by Finneti): "What amount of money one is willing to bet based on her belief on the future occurrence of the event?"
- Bayesian inference:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

  where in this case H is an **hypothesis**, D is **data**
- $P(H)$ is **prior probability** of H: probability that H is correct before D is seen
- $P(D|H)$ is **conditional probability** of seeing data D knowing that the hypothesis H is true (**likelihood**)
- $P(D)$ is **marginal probability** of D: probability of D to happen under all possible hypotheses
- $P(H|D)$ is **posterior probability**: probability that hypothesis is true, given the data and the previous state of belief about the hypothesis

...more on Bayesian probability

- Provide natural treatment of non-repeatable fenomena:
  P(Higgs boson exists) or $P(0.1 < \alpha_s < 0.2)$

- No golden-rule for priors, it's a subjective opinion

Example: Who will pay the next round?

Drinking with a friend, next round payed by who extracts lower valued card

Probability that friend is cheating if you pay *losts* consecutive times?

Assumptions:

- $P(cheat) = 0.05$ and $P(honest) = 0.95$ (old friend unlikely to cheat)
- $P(lose|cheat) = 1$ and $P(lose|honest) = 2^{-N}$ (50% probab. each turn)

Bayesian solution:

$$\begin{aligned}
P(cheat|losts) &= \frac{P(losts|cheat)P(cheats)}{P(losts|cheat)P(cheat) + P(losts|honest)P(honest)} \\
P(cheat|0) &= \frac{0.05}{0.05 + 0.95} = 0.05 \\
P(cheat|5) &= \frac{0.05}{0.05 + 2^{-5}0.95} = 0.63
\end{aligned}$$

Random variables

- **Random event**: event having more than one possible outcome
  - Each outcome may have associated a probability
  - Outcome not predictable, only probabilities are known
- Different outcomes may take different numerical values:
  $x_1, x_2, \ldots \rightarrow$ **random variable** x
  $P(x_1), P(x_2), \ldots$ form a **probability distribution**
- If observations are **independent** the distribution of each random variable is unaffected by knowledge of any other observation
- At experiment consisting of N repeated observations of the same random variable x can be considered as a single observation of a random vector $\mathbf{x}$ with coponents $x_1, \ldots, x_n$

Discrete and continuos random variables

- Discrete:
  - "Roll a dice": limited and discrete sample space
  - Discrete probability distribution (one value for each possible outcome)
- Continuos:
  - "Spin a spinner": real number in $[0, 2\pi]$
  - x = an outcome
  - $P(x) = 0 \ \forall \ x$
  - $P(x \in [i, j]) > 0$
    $P(x \in [0, \pi]) = \frac{1}{2}$ (for the spinner)
  - In general: $P(A < x < B) = \int_A^B p(x)dx$

Probability density function

Let m be a possible outcome of an observation with possible values $x \in [a, b]$
We define the p.d.f. as:

$$F(x;\theta)dx = P(m \in [x, x + dx])$$

where θ represents one or more parameters for f

- $\displaystyle\int_a^b f(x)dx = 1 \quad \left(\sum_a^b f(x) = 1 \text{ if discrete}\right)$
- x,θ may be vectors
- Usually in physics θ unknown, we want to estimate its value from a set of measurements of x (discussed later)

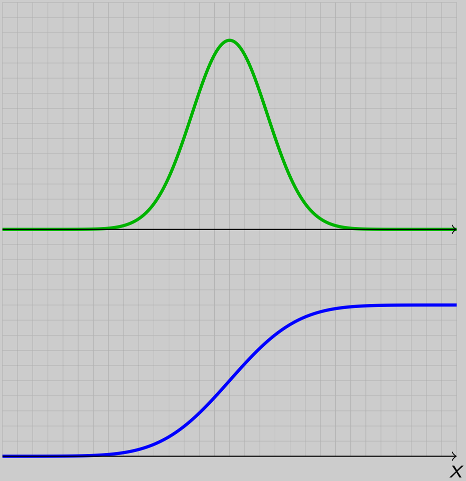
Cumulative and marginal distribution

Cumulative distribution function

- CDF: $\forall Y \in \mathbb{R}$,
  $$F(Y) = P(x \leq Y) = \int_{x_{min}}^{Y} f(x)dx$$
- $x \in [x_{min}^{(\neq -\inf)}, x_{max}^{(\neq \inf)}]$
  $\Rightarrow F(x_{min}) = 0, F(x_{max}) = 1$
- $F(Y)$ is monotonic

Marginal density function

- Is the projection of multidimensional density
  Ex: given $f(x, y)$,
  $$F_x(X) = \int_{y_{min}}^{y_{max}} f(x, y)dy$$
  $$F_y(Y) = \int_{x_{min}}^{x_{max}} f(x, y)dx$$



Main distribution's properties

Let $f(x)$ be a probability density function.

- **Expectation**:
  - Expectation of x (expected value, mean value, measure of the distribution's **location**):

$$E(x) = \mu = \bar{x} = \langle x \rangle = \int x \, f(x) dx$$

- **Variance** (measure of the distribution's **spread**):

$$V(x) = \sigma^2 = E[(x-\mu)^2] = E(x^2) - \mu^2 = \int (x-\mu)^2 f(x) dx$$

  - σ is called **standard deviation**

Reminder of most important distributions

$$
\begin{tikzcd}
\begin{array}{c}\text{Binomial}\\ P(r;N,p)=\binom{N}{r}p^r(1-p)^{N-r}\end{array} \arrow[rr, "p\to 0;\ Np=\mu"] \arrow[dr, "N\to\infty"] & & \begin{array}{c}\text{Poisson}\\ P(r;\mu)=\frac{\mu^r e^{-\mu}}{r!}\end{array} \arrow[dl, "\mu\to\infty"'] \\
& \begin{array}{c}\text{Normal}\\ P(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}\end{array} & \\
\begin{array}{c}\text{Multinomial}\\ P(r1,...,r_k;N,p_1,..,p_k)=\\ \frac{N!}{r_1!...r_k!}p_1^{r_1}...p_k^{r_k}\end{array} \arrow[uu, "i=2"] \arrow[ur, "N\to\infty"'] & & \begin{array}{c}\text{Chi-square}\\ P(x;N)=\frac{x^{\frac{N}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{N}{2}}\Gamma\frac{N}{2}}\end{array} \arrow[ul, "N\to\infty"]
\end{tikzcd}
$$

Contents

Prediction: Two general classes of problems

1. Probabilistic model assumed to be **known**: want to make predictions about future observations
   Ex: we know the distribution of a random variable x, we wish to predict the average $\bar{x}$ of next n future outcomes
2. Probabilistic model **not known**: one or more parameters θ_i unknow
   - **Estimate** parameters values (parameter estimation)
   - **Decide** if the θ_is form a set of known constants (hypothesis testing)

   Ex: after tossing a coin 1000 times, decide if coin is fair
   Ex: after a finite number of observation of a random variable x, estimate its average value $\bar{x}$

Known model: possible predictions

x random variable with known distribution, predict its value at a future trial.

- **Point prediction**: determine a constant c which minimizes error $x - c$ in some sense (future outcome cannod be predicted but only estimated).
  If error defined as $E_{rr}(x) = (x - c)^2$ then $c = E(x)$.
- **Interval prediction**: determine two constants c_1, c_2 such that

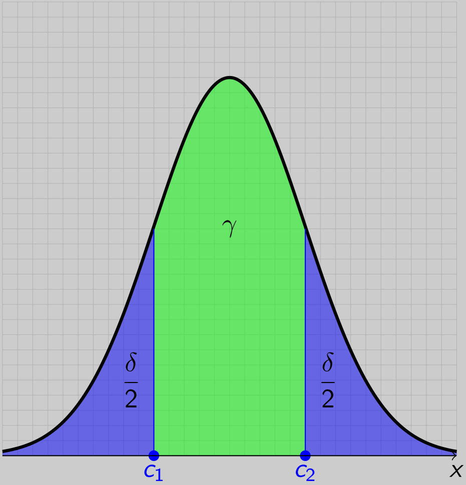
$$P(c_1 < x < c_2) = \gamma = 1 - \delta$$

where γ is an arbitrary constant called **confidence level**

Known model: Interval prediction

$$P(c_1 < x < c_2) = \gamma = 1 - \delta$$

- Bigger γ means prediction $x \in (c_1, c_2)$ reliable but $c_1 - c_2$ big.
- Usually γ fixed and c_1, c_2 chosen to minimize distance (symmetric choice not given to be the best one).
- Many methods to determine c_1, c_2 available depending on random variable distribution.



Unknown model: parameter estimation

- Distribution of a random variable x is a known function $f(x, \theta)$
- θ is an unknown parameter, scalar or vector
- We want to estimate θ after n repetitions of an experiment (x_i outcome of i-th experiment, $X = [x_1, ..., x_n]$ is called observation vector)

Point estimate

- Function $\hat{\theta} = g(X)$
- $\hat{\theta}$ is the **point estimator** of θ.
- $\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta$
- If error limit $\lim_{n \to \infty} \hat{\theta} - \theta = 0$ $\bar{\theta}$ is called a **consistent** estimator

Interval estimate

- **interval estimator** $(\theta_1, \theta_2) = (g_1(X), g_2(X))$
- (θ_1, θ_2) is a γ **confidence interval** of θ if $P(\theta_1 < \theta < \theta_2) = \gamma$
- $g_1(X), g_2(X)$ are to be chosen to minimize $\theta_2 - \theta_1$

Contents

Hypothesis testing

Statistical Hypothesis

Assertion or conjecture concerning one or more populations.

- Prove with certainty: absolute knowlegde, examine *entire population*. Not physically possible.
- How to use a random sample as evindence in support or against the hypothesis?

Hypothesis testing...

...is formulated in terms of two hypotheses:

- H_0: the **null** hypothesis
- H_1: an alternate hypothesis (the one we want to test)

We reduce the problem to two possible outcomes:

- **Reject** H_0 **and accept** H_1: sample provides sufficient evidence in favor of H_1
- **Not reject** H_0: sample does *not* provide sufficient evidence in favor of H_1

Warning!

Failure to reject H_0 does not imply H_0 true. There just is no sufficient evidence in favor of H_1 to assert it true.

Example jury trial: H_0 (innocent) is rejected if H_1 (guilty) is supported by evidence **beyond reasonable doubt**. Failure to reject H_0 does not imply innocence, just lack of evidence.

Hypothesis testing (example)

- Distribution of a random variable is a function $f(x, \theta)$
- Test $\theta = \theta_0$ (H_0) versus $\theta \neq \theta_0$ (H_1)
- Possible values of θ in H_1 form a set Θ_1
- If $|\Theta_1| = 1$ H_1 is called **simple**, otherwise **composite**
- The null hypothesis H_0 is usually simple.

Basic idea

- Under H_0, $f(x, \theta)$ is negligible in a certain region D_c of sample space
- If $x \in D_c$, it is reasonable to reject H_0
- If $x \in \bar{D}_c$, it is reasonable not to reject H_0

D_c is called **critical region** of the test, $\bar{D}_c$ the **region of acceptance**

Hypothesis testing: purpose

- Purpose of hypothesis testing is NOT to determine whether H_0 or H_1 true!
  ... but to establish whether the evidence supports the rejection of H_0

Example

Establish if a coin is fair (H_0)

- Toss the coin 100 times: head shows up k times
- If $k \leq 15$, we reject H_0: evidence shows that coin is not fair
- If $k \geq 40$, we fail to reject H_0: evidence does not support the rejection of hypothesis that coin is fair
- But this does not mean that the coin is fair, it could be $p = 0.48$

Case study: RAM chip manufacture

Claim: rate of defective chips is 5%

Let p_d be the true defective probability, we want to test wether:

- H_0: $p_d = 0.05$
- H_1: $p_d \geq 0.05$

based on a sample of 100 chips from the production line

Test statistic

Is a function of the sample $f : S \to \mathbb{R}$.

It is used to reduce the data (multiple data in a multidimensional space) to a number that can be used to perform an hypothesis test.

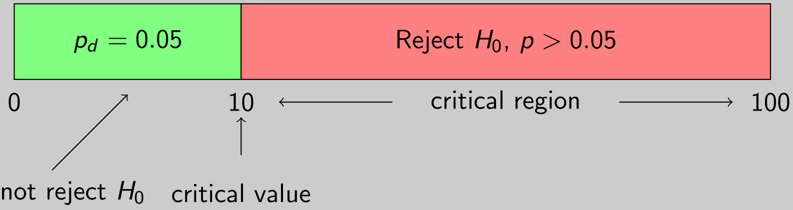
Generally chosen such that it can quantify behaviours that would distinguish the null hypothesis from the alternative one.

Case study: RAM chip manufacture (1)

Test statistic: X denotes the number of defective pieces in the sample of 100.

This is a Bernoulli process (defectiveness of each chip is independent), in a sample of size S_s we expect $S_s p_d$ (in the example $100 \cdot 0.05 = 5$) defective pieces.

An example of a good test is to reject H_0 if $X \geq 10$, which gives a strong indication that $p_d \geq 0.05$.



Types of errors

Decision is based on a finite sample: may be wrong!

	H_0 true	H_1 true
Not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Type I error

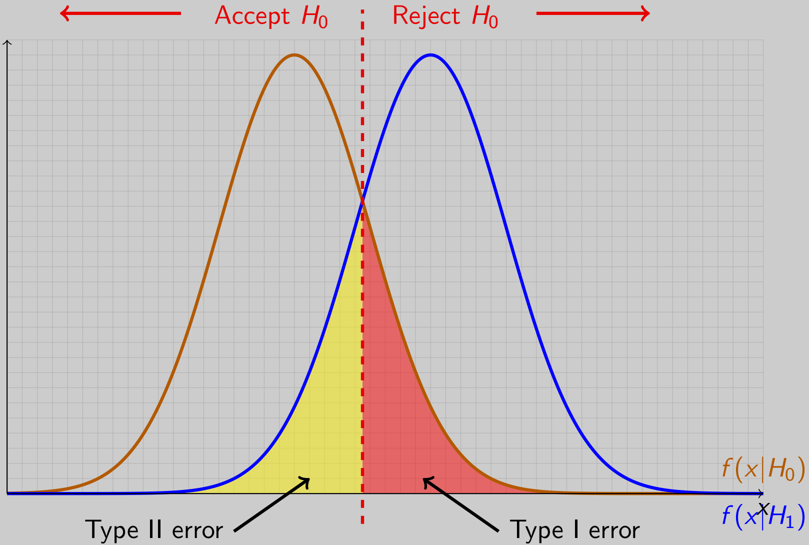
Acceptance of H_1 when H_0 is true. The probability α of committing this error is called **significance level** or **size** of the test

Type II error

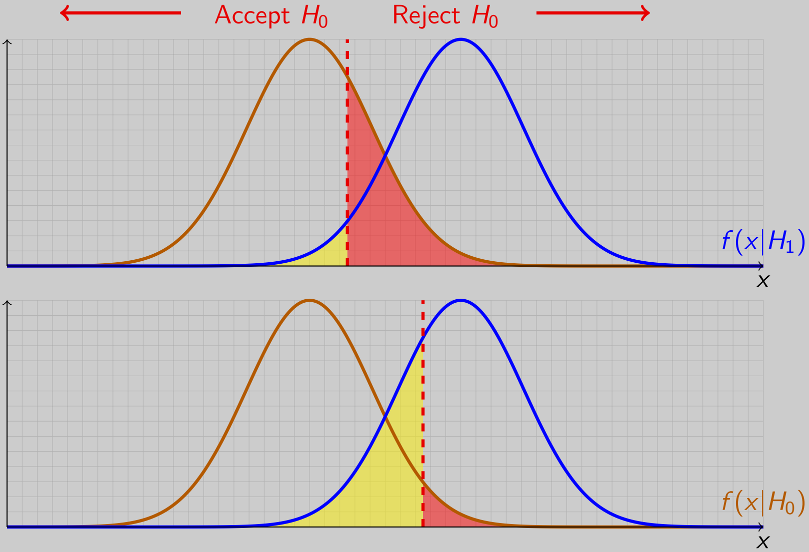
Failure to reject H_0 where H_1 is true. The probability $1 - \beta$ of **not** committing this error is called **power** of the test (with respect to the alternative H_1)

The objective is to reduce both α and β as much as possible.

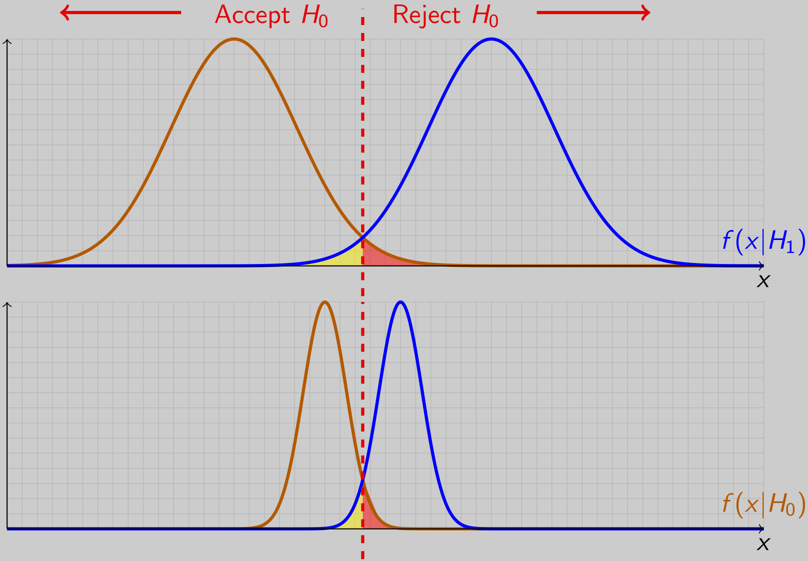
Visual interpretation



Visual interpretation: critical region



Visual interpretation: overlap



Case study: RAM chip manufacture (2)

Test statistic: $X > 10$
Size of the test: probability of type I error: reject H_0 when true
We assume a binomial distribution: $f(k; n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

$$
\begin{aligned}
\alpha &= P(X \geq 10 | p_d = 0.05) \\
&= \sum_{i=10}^{100} P(X = i | p_d = 0.05) = \sum_{i=10}^{100} b(i; n = 100, p = 0.05) \\
&= \sum_{i=10}^{100} \binom{100}{i} 0.05^i (1 - 0.05)^{100-i} = 0.0282
\end{aligned}
$$

Case study: RAM chip manufacture (3)

The power of the test $1-\beta$ (probability of not rejecting H_0 when H_1 true) for $H_1 : p_d > 0.05$ cannot be computed because the true p_d is unknown.
H_1 can be reformulated to be for example $H_1 : p_d = 0.1$ or $H_2 : p_d = 0.15$

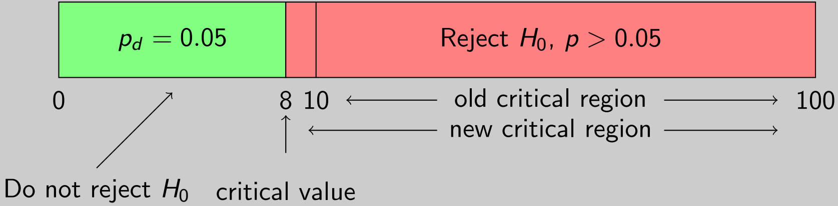
$$\begin{aligned}\beta_{H_1} &= P(X < 10 | p_d = 0.1) \\ &= \sum_{i=0}^{9} b(i, n = 100, p = 0.1) = 0.4513\end{aligned}$$

and

$$\begin{aligned}\beta_{H_2} &= P(X < 10 | p_d = 0.15) \\ &= \sum_{i=0}^{9} b(i, n = 100, p = 0.15) = 0.0551\end{aligned}$$

Case study: RAM chip manufacture (4): critical value

A bigger critical region reduces β but enlarges α, and viceversa.
What happens by reducing the critical value?



$$\alpha = \sum_{i=8}^{100} b(i, n = 100, p = 0.05) = 0.128 \quad \text{(was 0.0282)}$$

$$\beta_{H_1} = \sum_{i=0}^{7} b(i, n = 100, p = 0.1) = 0.206 \quad \text{(was 0.4513)}$$

Case study: RAM chip manufacture (5): sample size

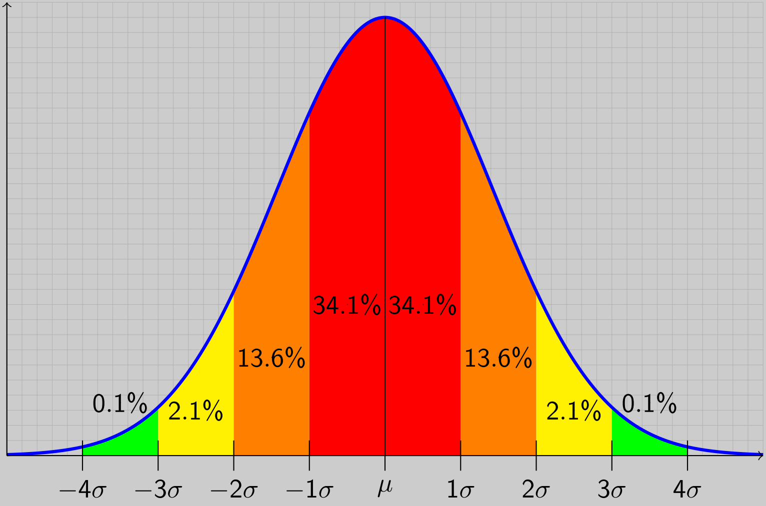
Both α and β can be reduced simultaneously increasing the sample size.
For example, increasing the sample size to 150 and setting the critical value to 12 yields:

$$\begin{aligned} \alpha &= \sum_{i=12}^{150} b(i, n = 150, p = 0.05) = 0.074 \quad \text{(was 0.128)} \\ \beta_{H_1} &= \sum_{i=0}^{7} b(i, n = 150, p = 0.1) = 0.171 \quad \text{(was 0.206)} \end{aligned}$$

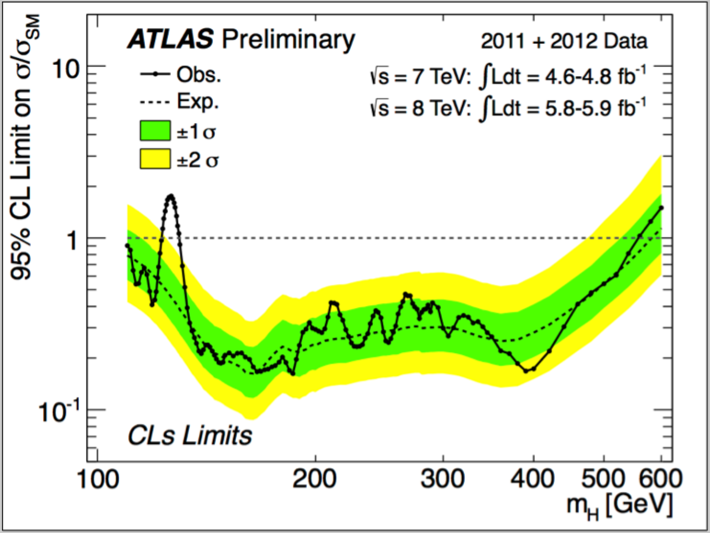
Contents

Standard deviations

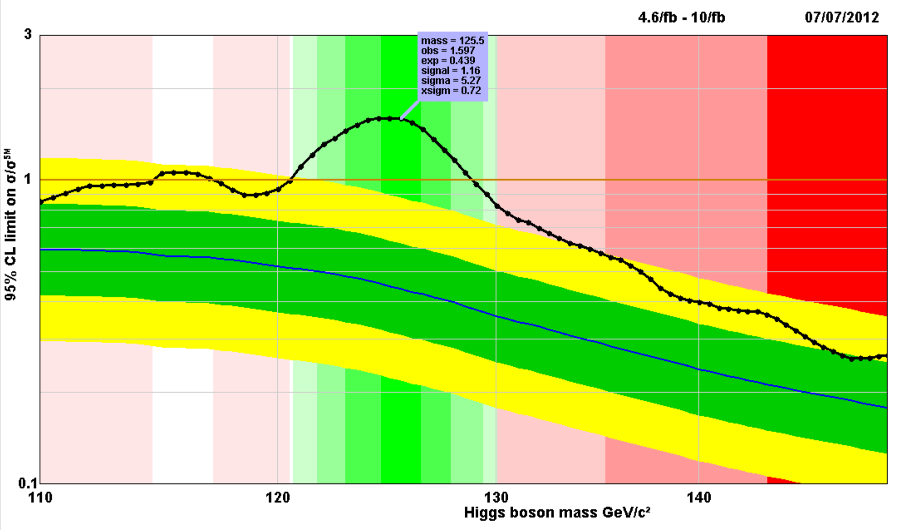
$$\sigma = \sqrt{(E[X]-\mu)^2} = \sqrt{E[X^2]-E[X]^2}$$



… and here is the Higgs!



... more in details, combined plot



Want to know more?

Q&A session!

For further information call 1-800-scarli-help or:

Samuele Carli

E-mail: scarli@cern.ch
Web: www.csspace.net