Practical data analysis

Doru Constantin and Guillaume Tresset

doru.constantin@u-psud.fr
guillaume.tresset@u-psud.fr

Laboratoire de Physique des Solides, Orsay.

LPS ORSAY

References I

- Barlow, R. J. (1993).
  Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences.
  Chichester, England; New York: Wiley.
- Bevington, P. R. (1969).
  Data Reduction and Error Analysis for the Physical Sciences.
  New York: McGraw-Hill.
- Bevington, P. R. and K. Robinson (2003).
  Data Reduction and Error Analysis for the Physical Sciences (3 ed.).
  New York: McGraw-Hill.
- Bohm, G. and G. Zech (2010).
  Introduction to Statistics and Data Analysis for Physicists.
  Hamburg: Verlag Deutsches Elektronen-Synchrotron.
  Freely available online from
  `http://www-library.desy.de/preparch/books/vstatmp_engl.pdf`

LPS ORSAY

References II

- Drosg, M. (2009).
  Dealing with Uncertainties (2 ed.).
  Springer.
- Feller, W. (1968).
  An Introduction to Probability Theory and Its Applications (3rd edition ed.).
  New York: Wiley.
- Grinstead, C. M. and J. L. Snell (1997).
  Introduction to Probability (2 ed.).
  American Mathematical Society.
  Freely available online from
  `http://www.dartmouth.edu/~chance/`
- Hughes, I. G. and T. P. A. Hase (2010).
  Measurements and their Uncertainties.
  Oxford: Oxford University Press.
  Short and very legible introduction.

References III

- Jaynes, E. T. (2003).
  Probability Theory – The Logic of Science.
  Cambridge: Cambridge University Press.
- Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992).
  Numerical Recipes in C: The Art of Scientific Computing (2 ed.).
  Cambridge: Cambridge University Press.
- Taylor, J. R. (1997).
  An Introduction to Error Analysis (2 ed.).
  Sausalito: University Science Books.

Variability

1. When measuring the height of all adult males in a certain town, one finds 177 ± 5 cm.
2. The charge of the electron is $(1.602176565 \pm 0.000000035) \times 10^{-19}$ C.

The meaning of probability

Casting a die:

1. Out of a large number of trials, each face will come on top about 1 in 6 times.
2. Our state of knowledge gives us no reason to prefer one of the faces over the others.

Each face has a 1/6 probability of coming up.

Random variables

- A *random variable* "is simply an expression whose value is the outcome of a particular experiment" (Grinstead & Snell, 1997). It takes values in a certain domain Ω.
- This domain (or *sample space*) can be discrete, $\Omega = \{\omega_1, \omega_2, \ldots \omega_k, \ldots\} \subset \mathbb{Z}^n$ (finite or countably infinite) or continuous $\Omega \subset \mathbb{R}^n$
- The elements of the sample space (ω_k or $\mathbf{x} \in \mathbb{R}^n$) are called *outcomes*. Subsets of Ω are called *events*.
- We introduce a probability distribution, characterized by a *distribution function* m. In the discrete case, this function satisfies:

$$m(\omega) \geq 0, \quad \forall \omega \in \Omega$$
$$\sum_{\omega \in \Omega} m(\omega) = 1$$

The *probability* of an event E is defined as :
$P(E) = \sum_{\omega \in E} m(\omega)$.

Continuous distributions

Let X be a continuous real-valued random variable. A *density function* for X is a function $f : \Omega \to \mathbb{R}$ such that

$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x, \quad \forall\, a, b \in \mathbb{R}.$$

$$\forall E \subset \mathbb{R} \quad P(X \in E) = \int_E f(x)\mathrm{d}x.$$

$$P([x, x + \mathrm{d}x]) = f(x)\mathrm{d}x$$

$f(x)\mathrm{d}x$ is the probability of the outcome x

The *cumulative distribution function* of X is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t, \quad \text{with} \quad \frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$$

LPS ORSAY

Central tendency

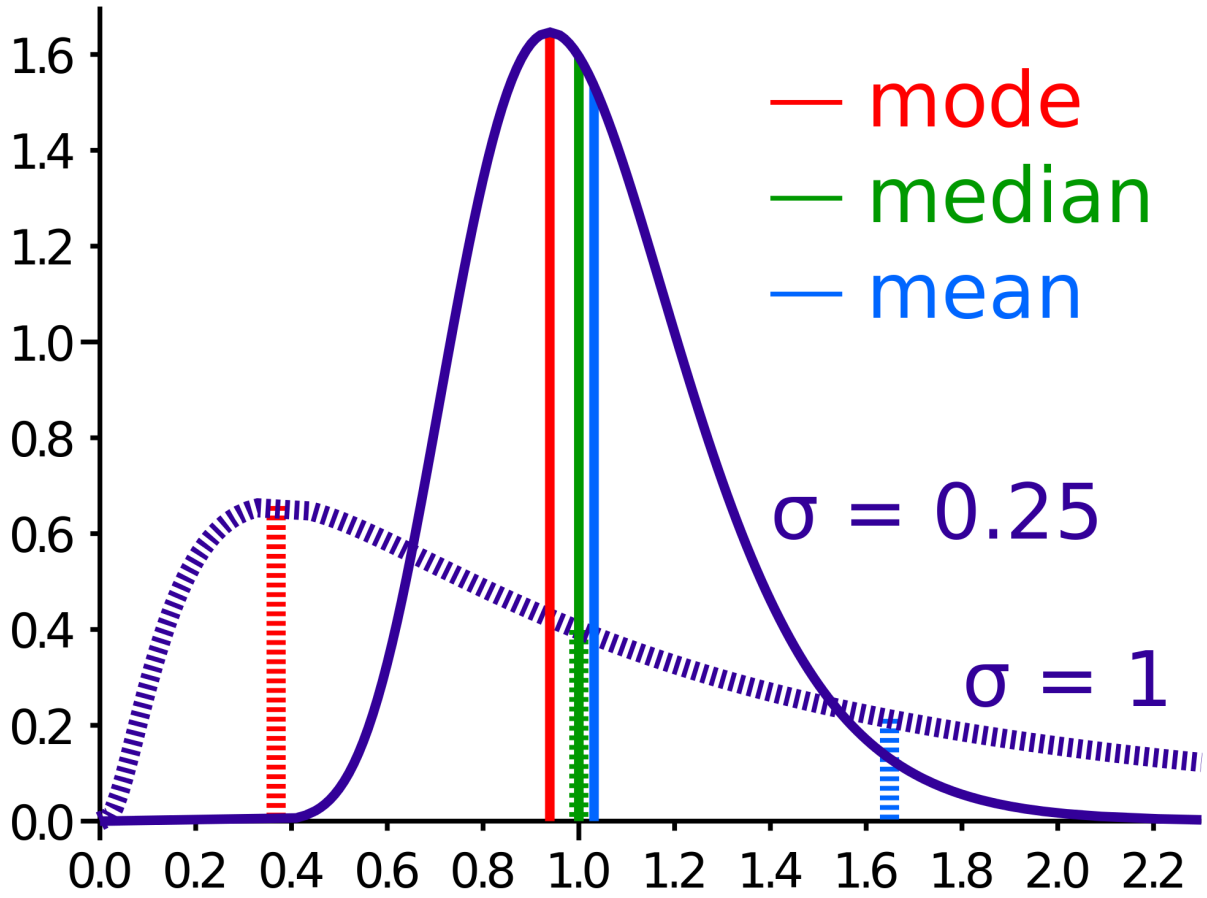


Figure: Log-normal distribution with parameters $\mu = 0$ and $\sigma = 0.25$ (solid line) and $\sigma = 1$ (dashed line). The mean (blue), median (green) and mode (red) are shown for both curves.

LPS ORSAY

Spread

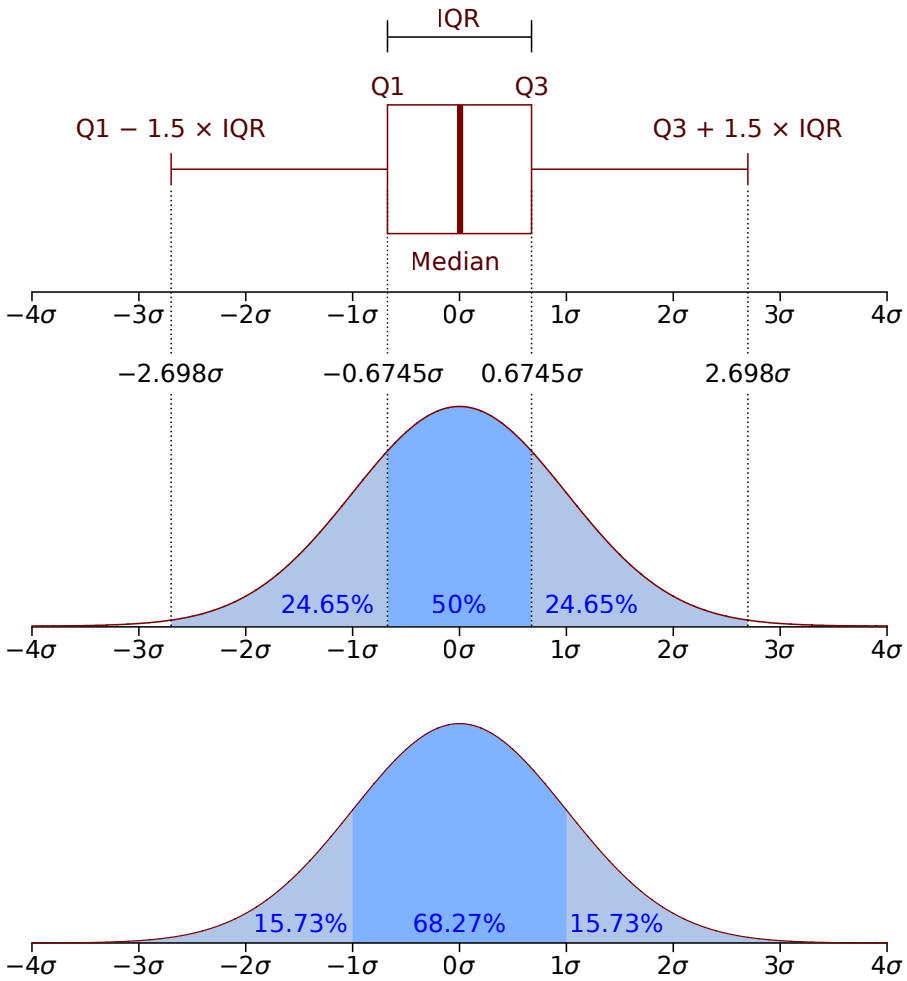
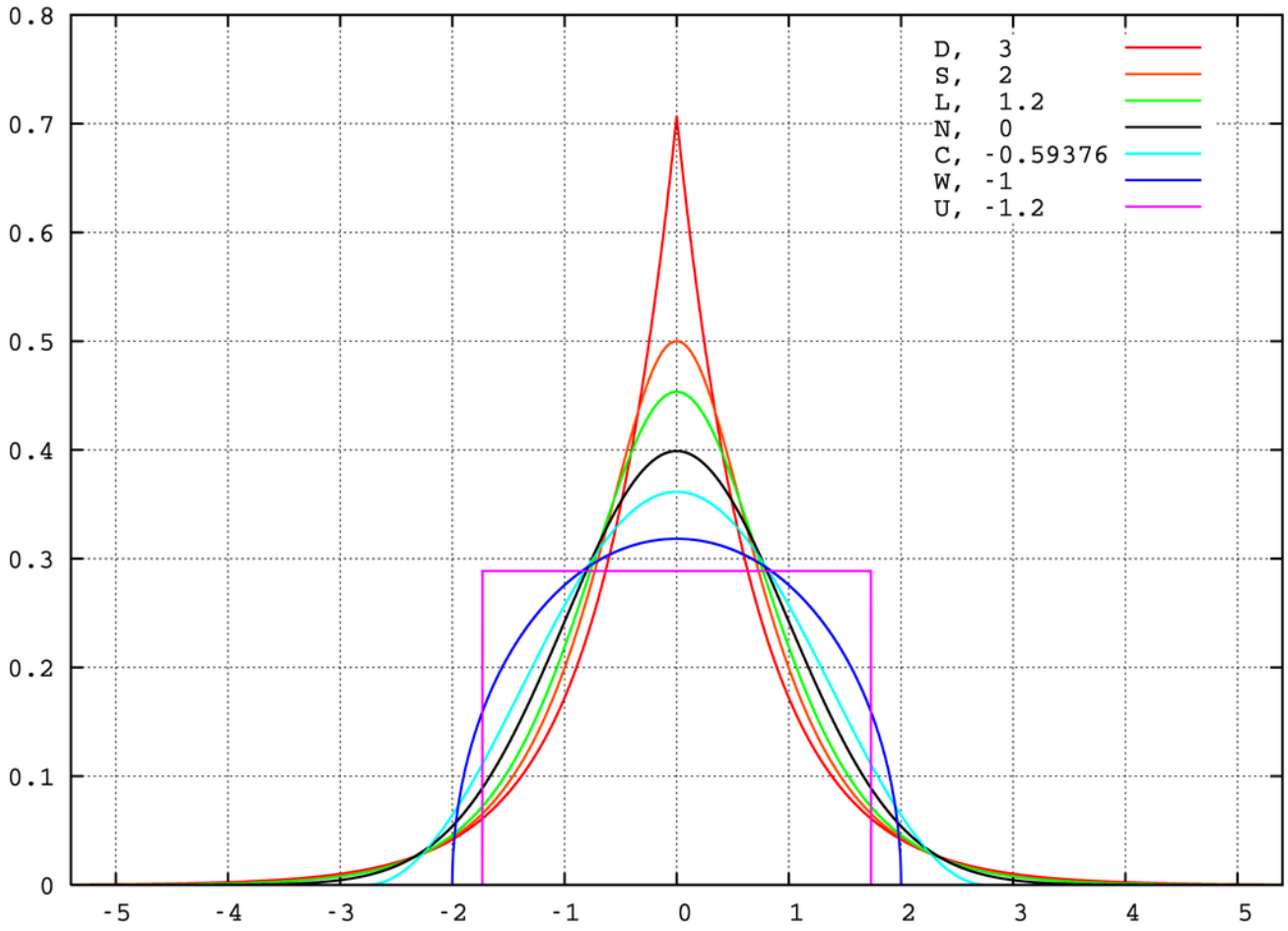


Figure: Boxplot details

LPS ORSAY

Higher-order moments

$$\gamma_1 = \left\langle \left(\frac{X - \mu}{\sigma} \right)^3 \right\rangle \quad \textit{skewness;} \quad \gamma_2 = \left\langle \left(\frac{X - \mu}{\sigma} \right)^4 \right\rangle - 3 \quad \textit{kurtosis}$$

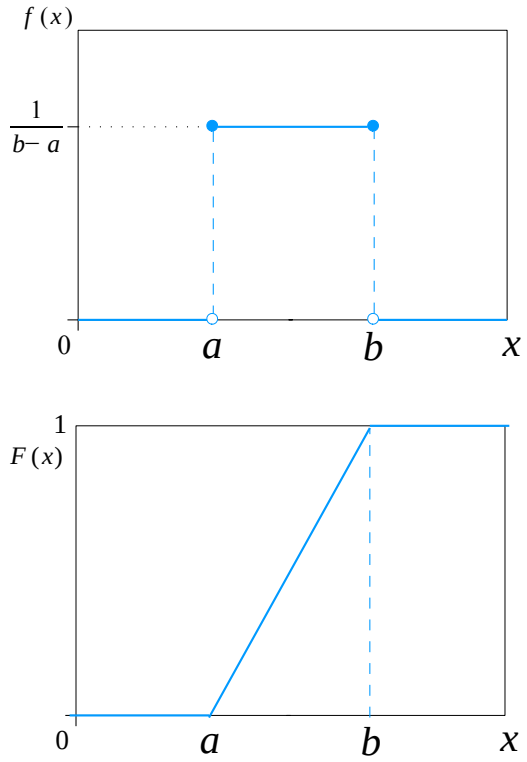


Graphics by MarkSweep. Licensed under Public domain via Wikimedia Commons

LPS ORSAY

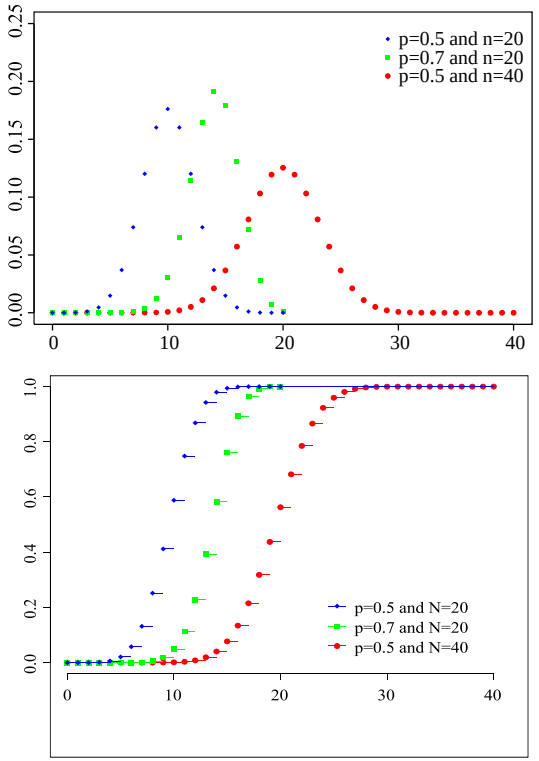
Uniform

- All outcomes have equal probability
- $\mathcal{U}(x;a,b) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
- $\mu = \frac{1}{2}(a+b),\ m = \frac{1}{2}(a+b)$
  $M =$ any value in $[a,b]$.
- $\sigma^2 = \frac{1}{12}(b-a)^2,\ \gamma_1 = 0,\ \gamma_2 = -6/5$
- One cannot have a uniform distribution over an infinite domain (discrete or continuous)!



Binomial

- Number k of successes in a sequence of n independent yes/no experiments (Bernoulli trials), each of which yields success with probability p.
- $\mathcal{B}(k; n, p) = C_n^k \, p^k (1-p)^{n-k}$; $k \in \{0, 1, \ldots, n\}$
- $\mu = np$, $m = \lfloor np \rfloor$ or $\lceil np \rceil$ $M = \lfloor (n+1)p \rfloor$ or $\lfloor (n+1)p \rfloor - 1$.
- $\sigma^2 = np(1-p)$, $\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}$, $\gamma_2 = \frac{1-6p(1-p)}{np(1-p)}$
- k is **the variable**, n and p are **parameters**.

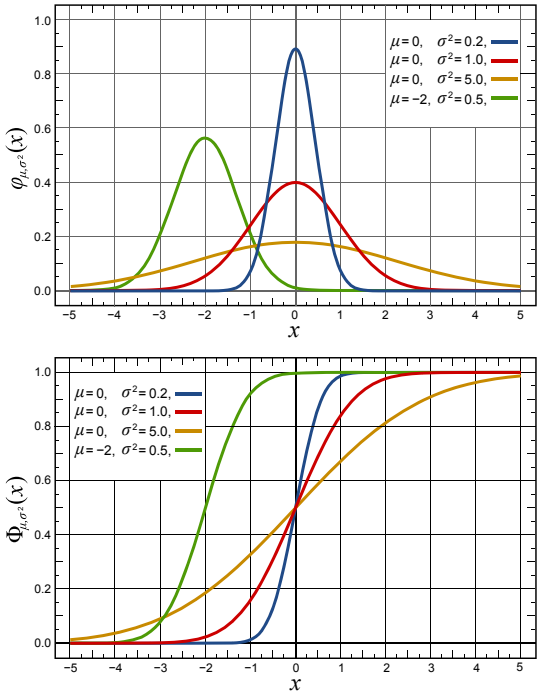


Graphics by Tayste. Licensed under Public domain via Wikimedia Commons

LPS ORSAY

Normal

- Very widely encountered.
- $\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \quad x \in \mathbb{R}$
- $\langle X \rangle = m = M = \mu$
  $\left\langle X^2 \right\rangle = \sigma^2, \quad \gamma_1 = 0, \quad \gamma_2 = 0$



Graphics by Inductiveload. Licensed under Public domain via Wikimedia Commons

LPS ORSAY

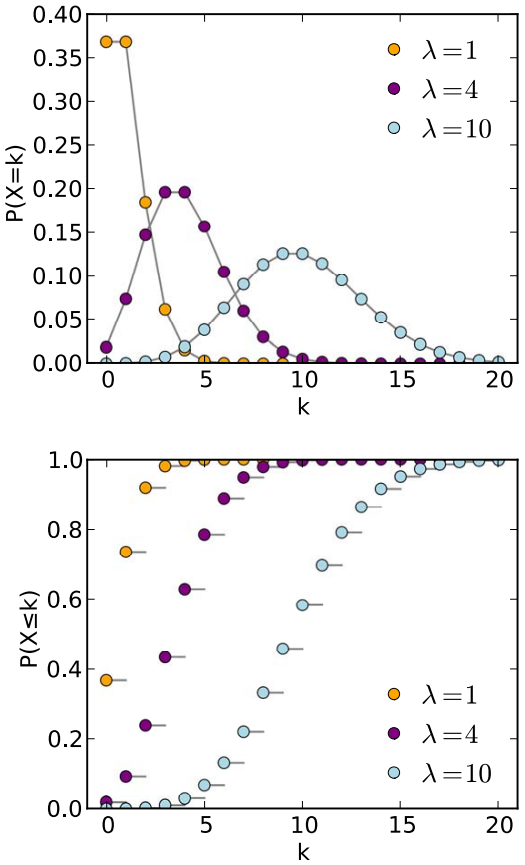
Poisson

- Probability of a given number of independent events k occurring in a fixed interval with a known average rate.
- $\mathcal{P}(k;\lambda) = \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda};\ k \in \mathbb{N},\ \lambda \in \mathbb{R}^+$
- $\mu = \lambda,\ m \simeq \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
  $M = \lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
- $\sigma^2 = \lambda, \quad \gamma_1 = \lambda^{-1/2}, \quad \gamma_2 = \lambda^{-1}$
- Can be seen as the limit of a binomial distribution for large n:

$$\mathcal{P}(k;\lambda = np) \simeq \mathcal{B}(k;n,p)$$

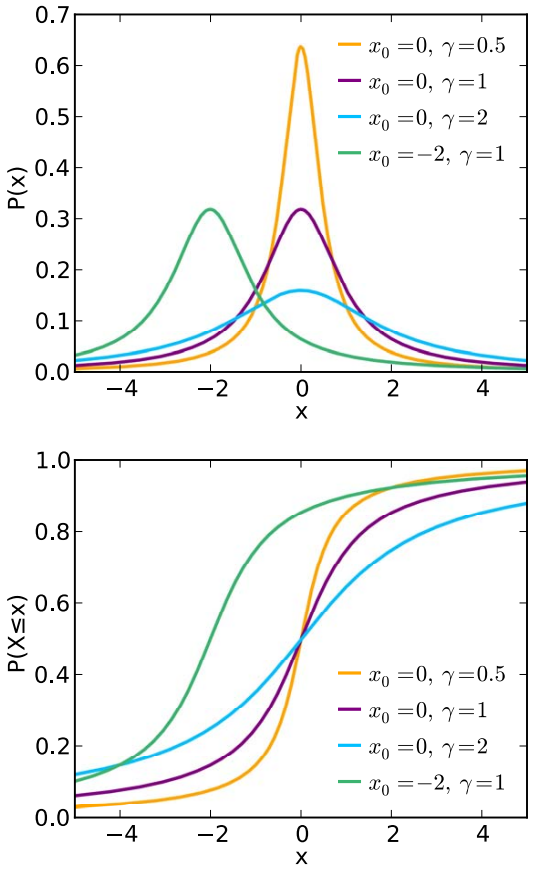
- Approaches $\mathcal{N}$ for large λ:

$$\mathcal{P}(k;\lambda) \simeq \mathcal{N}(x = k;\mu = \lambda,\sigma^2 = \lambda)$$



Lorentzian

- Shape of resonance peaks. Also named after Cauchy (in mathematics) and Breit and Wigner (in spectroscopy)
- $\mathcal{L}(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$;
  $x \in \mathbb{R},\ x_0 \in \mathbb{R},\ \gamma \in \mathbb{R}^+$
- $m = M = x_0$
- No μ or higher moments!



Graphics by Skbkekas. Licensed under CCA 3.0 via Wikimedia Commons

LPS ORSAY

The law of large numbers

Statement [Feller 1968, Vol. I, Chapter X, Eq. (1.2)]
Let X_k be a sequence of mutually independent random variables with a common distribution. If the expectation $\mu = E(X_k)$ exists, then for every $\epsilon > 0$, as $n \to \infty$.

$$P\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| > \epsilon\right\} \to 0$$

The central limit theorem

Statement [Feller 1968, Vol. I, Chapter X, Eq. (1.3)]
Let X_k be a sequence of mutually independent random variables with a common distribution. Suppose that $\mu = E(X_k)$ and $\sigma^2 = Var(X_k)$ exist and let $S_n = X_1 + \ldots + X_n$. Then for every fixed β

$$P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} < \beta\right\} \to \mathcal{N}(\beta)$$

where the *normal (cumulative) distribution function*

$$\mathcal{N}(\beta) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right].$$

Weak convergence: $\forall x \in \mathbb{R},\ \lim_{n\to\infty} F_n(x) = F(x)$

LPS ORSAY

Sum of variables and convolution

The sum of two random variables:

$$Z = X + Y$$

$$f_Z(z) = \int_{-\infty}^{\infty} \mathrm{d}x\, f_X(x) f_Y(z - x) = f_X * f_Y$$

Convolution theorem:

$$\tilde{f}_Z(q) = \tilde{f}_X(q)\tilde{f}_Y(q)$$

LPS ORSAY

Standard deviation and width

We are usually interested in σ as a measure of the HWHM!

- Gaussian: $\sqrt{\langle X^2 \rangle} = \sigma \quad \text{HWHM} = \sigma\sqrt{2\ln 2}$
- Lorentzian: $\sqrt{\langle X^2 \rangle} = ? \quad \text{HWHM} = \gamma$

$$\mathcal{N}(x;0,\sigma) \overset{\mathcal{F}}{\Longrightarrow} \mathcal{N}(q;0,1/\sigma) \sim \exp[-(q\sigma)^2]$$

$$\mathcal{L}(x;0,\gamma) \overset{\mathcal{F}}{\Longrightarrow} \exp(-|q|\gamma)$$

For the sum:

$$\sigma_{\text{sum}} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

$$\gamma_{\text{sum}} = \gamma_1 + \gamma_2$$

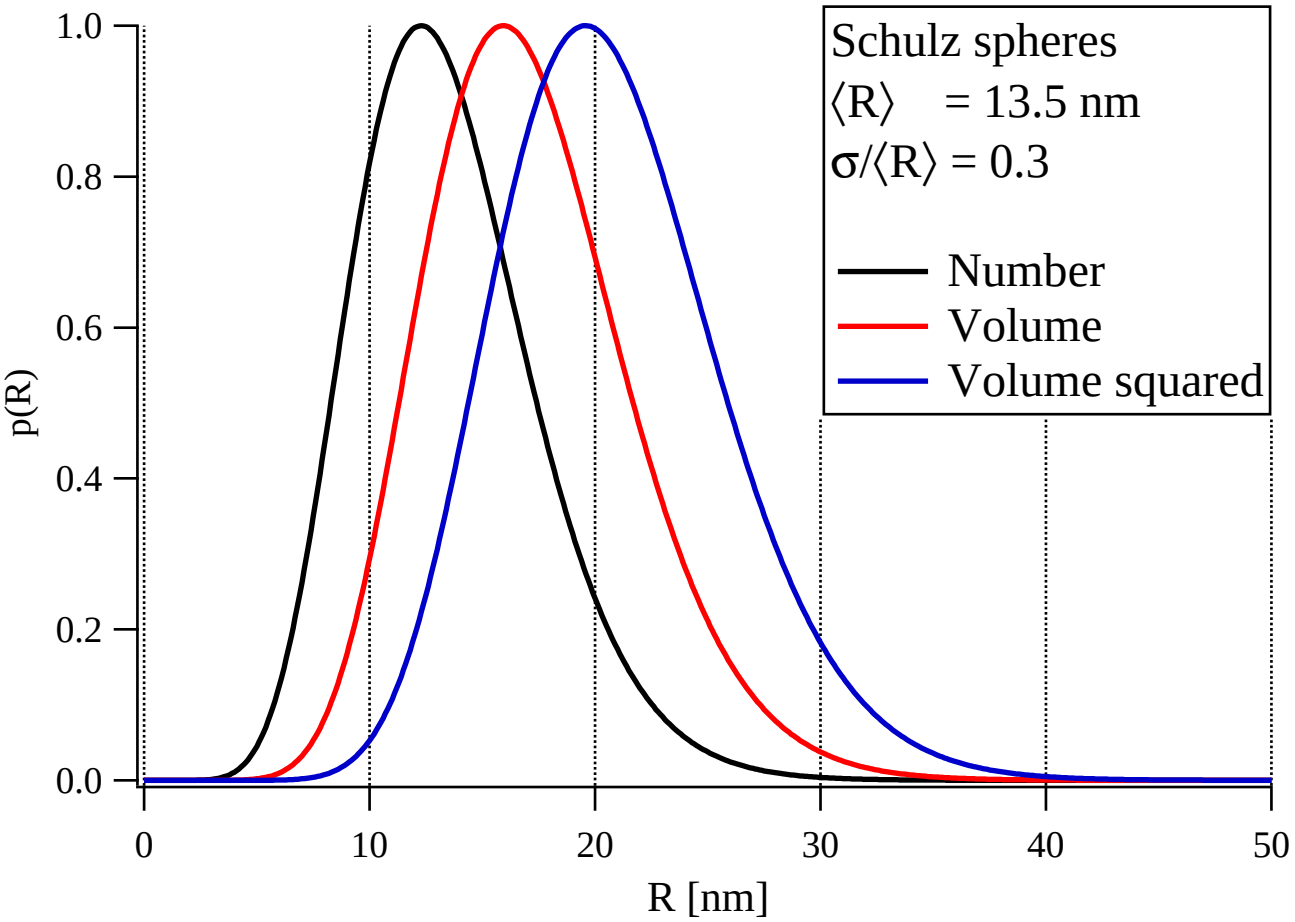
Sampling

Estimate the population parameters μ and σ by taking a *sample* of n measurements $x_1, x_2, \ldots, x_n$ followed by computing the *sample mean* $\bar{x}$ and *sample variance* s^2:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The relevant distribution



LPS ORSAY

Sample statistics – the mean

$$\langle \bar{x} \rangle = \mu$$

$$\left\langle (\bar{x} - \mu)^2 \right\rangle = \left\langle \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \left[\sum_{i=1}^{n} (x_i - \mu) \right]^2 \right\rangle$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \left\langle (x_i - \mu)^2 \right\rangle + \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left\langle (x_i - \mu)(x_j - \mu) \right\rangle = \frac{\sigma^2}{n}$$

Standard error of the mean

$$\text{SEM} = \sqrt{\langle (\bar{x} - \mu)^2 \rangle} = \frac{\sigma}{\sqrt{n}}$$

Sample statistics – the variance

$$\left\langle s^2 \right\rangle = \left\langle \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \right\rangle = \frac{1}{n}\sum_{i=1}^{n}\left\langle [(x_i - \mu) - (\bar{x} - \mu)]^2 \right\rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\langle (x_i - \mu)^2 \right\rangle - \frac{2}{n}\sum_{i=1}^{n}\left\langle (x_i - \mu)(\bar{x} - \mu) \right\rangle + \frac{1}{n} n \underbrace{\left\langle (\bar{x} - \mu)^2 \right\rangle}_{\sigma^2/n}$$

$$= \sigma^2 - \frac{2}{n}\underbrace{\left\langle (\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \mu) \right\rangle}_{n(\bar{x}-\mu)} + \frac{\sigma^2}{n} = \sigma^2 - \frac{2\sigma^2}{n} + \frac{\sigma^2}{n} \Longrightarrow$$

$$\left\langle s^2 \right\rangle = \frac{n-1}{n}\sigma^2$$

Estimators

- **Definition:** Estimator $\hat{a}$ of the property a of the parent population.
- **Consistency:**

$$\lim_{n\to\infty} \hat{a} = a$$

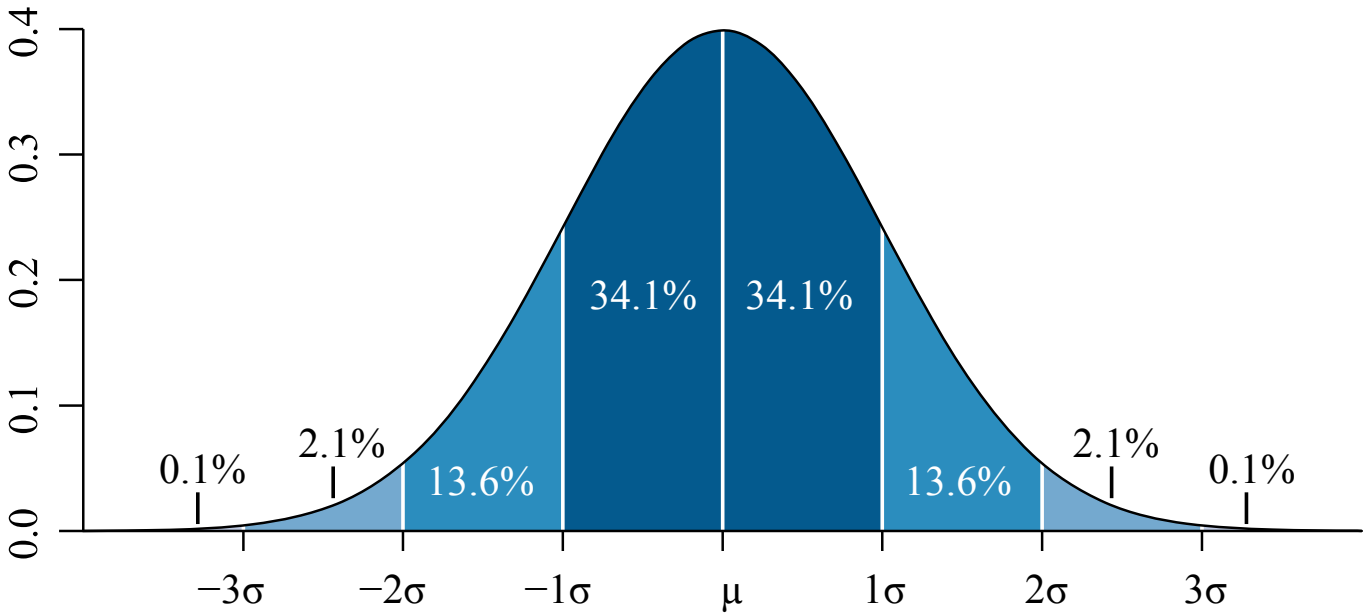
- **Lack of bias:**

$$\langle \hat{a} \rangle = a$$

- **Efficiency:**

$$\left\langle (a - \langle \hat{a} \rangle)^2 \right\rangle \quad \text{is small.}$$

Confidence intervals



Two-sided confidence intervals for the normal distribution:

- $[\mu - \sigma, \mu + \sigma]$ $C \simeq 68\%$
- $[\mu - 2\sigma, \mu + 2\sigma]$ $C \simeq 95\%$
- $[\mu - 3\sigma, \mu + 3\sigma]$ $C \simeq 99\%$

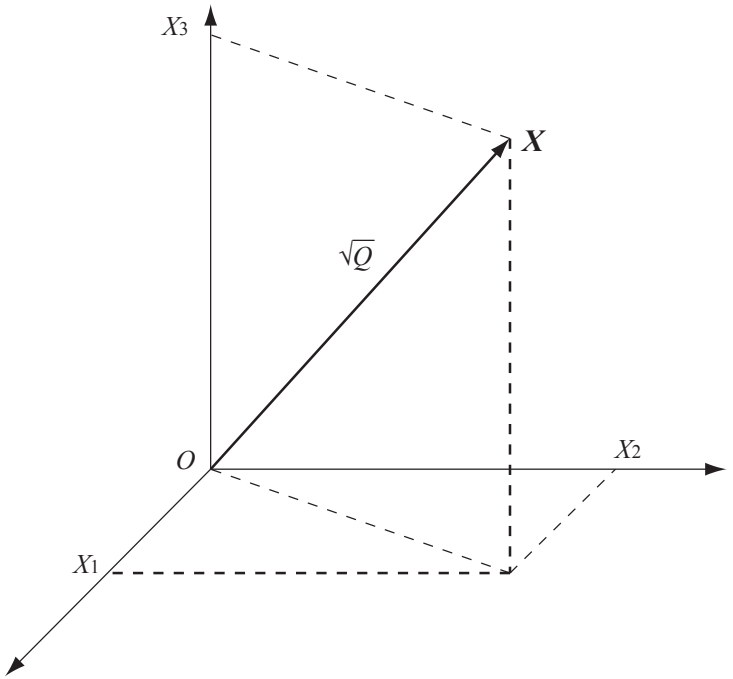
LPS ORSAY

Chi-squared distribution

The *chi-squared distribution* with ν degrees of freedom is the distribution of a sum of the squares of ν independent standard normal random variables:

$$Q = \sum_{j=1}^{\nu} X_j^2 \quad \text{with} \quad X_j \sim \mathcal{N}(0,1)$$

$$f_Q(x;\nu) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \exp(-x/2) & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



LPS ORSAY

Constraints

The constraint $\sum_j \delta_j = 0$ imposes a linear relation on the normalized variables: $\sum_j \sigma_j X_j = 0$, so that $\mathbf{X}$ is contained in the $\nu - 1$ hyperplane perpendicular to the vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_\nu)$

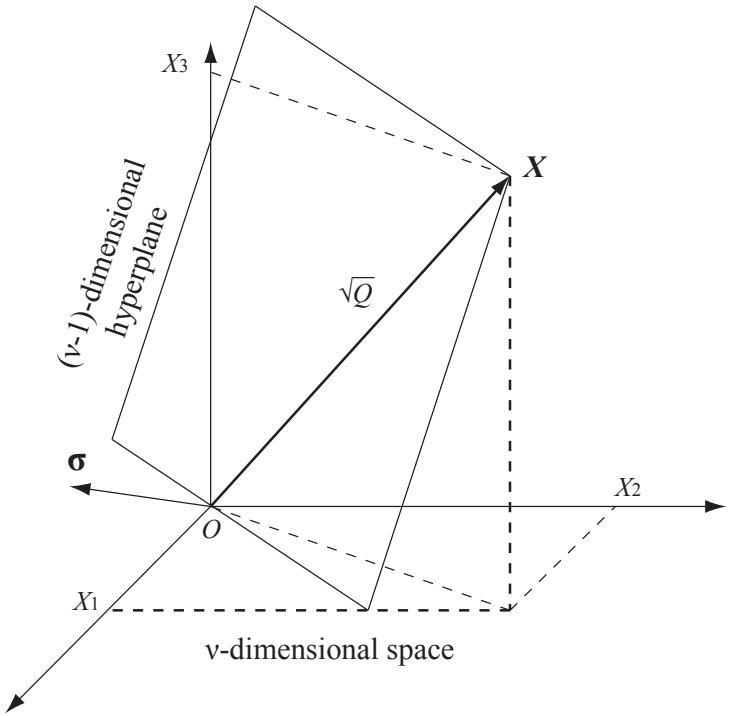


Figure: Geometrical visualization of the linear constraint $\mathbf{X} \cdot \sigma = 0$.

LPS ORSAY

Definitions

Random and systematic errors (this lecture)

- *Random (or statistical) errors* are those that can be reduced by increasing the sample size.
- *Systematic errors* are those that are not random.

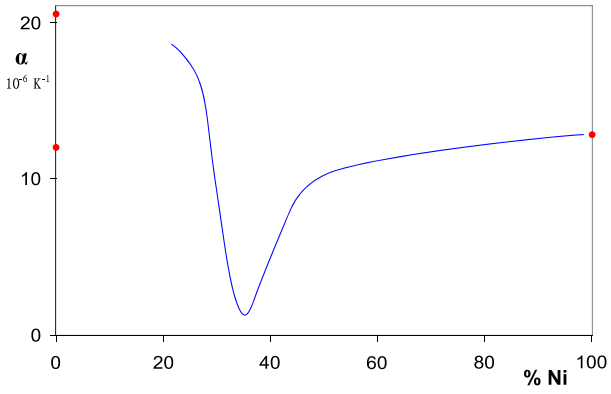
Type A and B components of uncertainty (GUM 2008, § 0.7)

- *type A components of uncertainty* are those evaluated by statistical methods (analysis of series of observations).
- *type B* are those evaluated by other means.

Length measurements

Invar and elinvar

- Discovered by Charles-Édouard Guillaume
- Nobel Prize in Physics in 1920
- Invar: low temperature expansion coefficient
- Elinvar: low temperature variation of the elastic coefficient



Portrait photo by A. B. Lagrelius & Westphal. Licensed under Public domain via Wikimedia Commons

Graph by RicHard-59. Licensed under CC BY-SA 3.0 via Wikimedia Commons

LPS ORSAY

Photomultiplier tube

Response as a function of light flux intensity

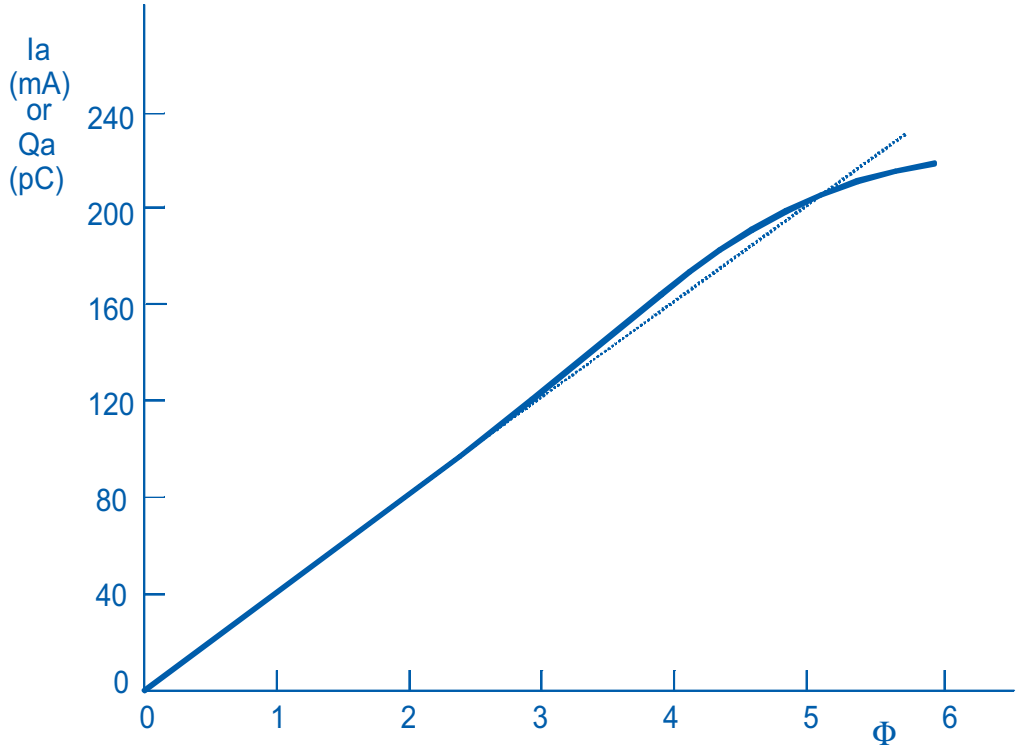


Fig.23 Typical current or charge linearity characteristics of a PMT operating from a supply with type B voltage division (photon flux Φ in arbitrary units).

Graphics from *Photomultiplier tubes basics* by Photonis.