Revisiting some conventional statistical notions in the framework of possibility theory

Gilles MAURIS

LISTIC, Polytech' Annecy-Chambéry

Université Savoie Mont Blanc

Imprecise Probabilities Workshop May 2015 Toulouse

Plan

I- Generalities

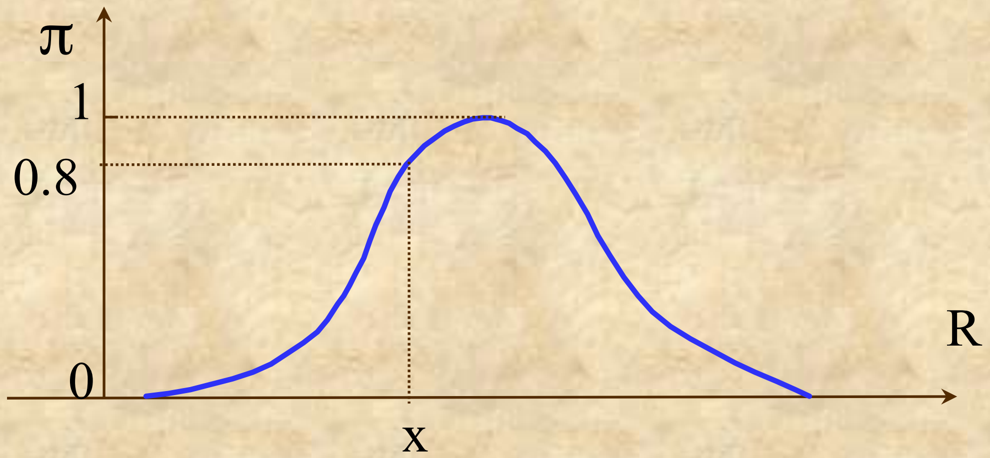
II- Probability possibility conversion

III- Relationships with descriptive statistical parameters

IV- Relationships with inferential statistical notions

V- Conclusion

Possibility distribution [Zadeh 78][Dubois-Prade 80]



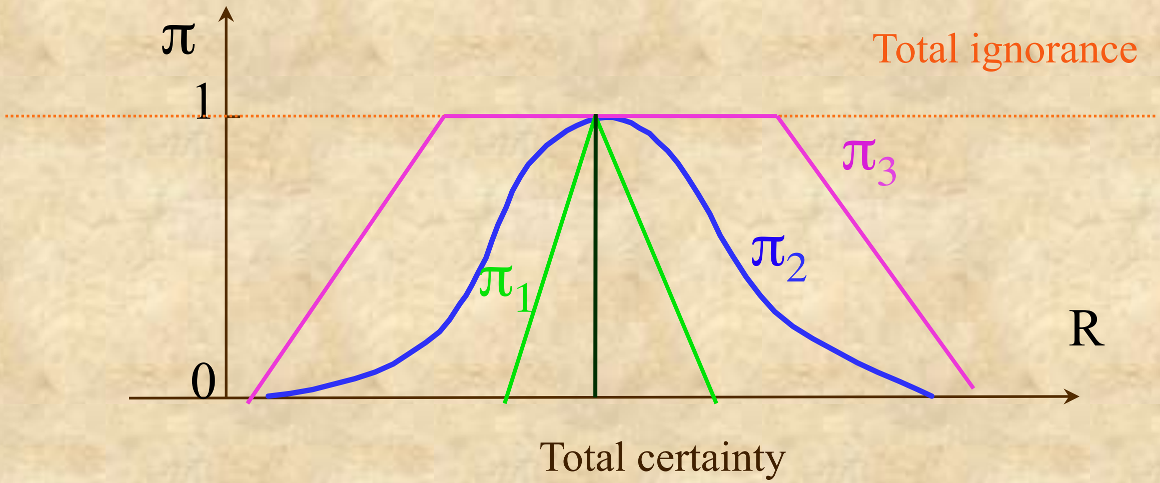
π(x) is a fuzzy set representing incomplete information in a gradual way (here the generalization of an interval)

sup (π(x)/x belongs to R)=1 (instead of int(p(x)dx)=1)

π(x) represents the possibility (instead of the probability) the value of the random variable X is equal to x

Specificity of a possibility distribution

$\pi(x)$ provides an intuitive expression of uncertainty



For all x : $\pi_1(x) < \pi_2(x) < \pi_3(x)$ (fuzzy subset inclusion)

π_1 is more specific π_2 that is more specific than π_3

(the more specific the less spread)

The specificity order reflects the informational content

Basics of Possibility theory

- Based on two non –additive set functions

the possibility measure Π and the necessity measure N)

- $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$
- $\Pi(A \cap B) > \min(\Pi(A), \Pi(B))$
- $N(A \cap B) = \min(N(A), N(B))$
- $N(A \cup B) > \max(N(A), N(B))$
- $\Pi(A) = 1 - N(A^c)$
- $N(A) > 0 \Rightarrow \Pi(A) = 1$

π is also a faithful representation of a family of probability distributions $P(\pi) = \{P / \forall A \subseteq \Omega, P(A) \leq \Pi(A)\}$

- $\Pi(A) = \sup\ (P(A) / P$ belongs to $P(\pi))$
- $N(A) = \inf\ (P(A) / P$ belongs to $P(\pi))$

Useful for cases of partial probability information

Plan

I- Generalities

II- Probability possibility conversion

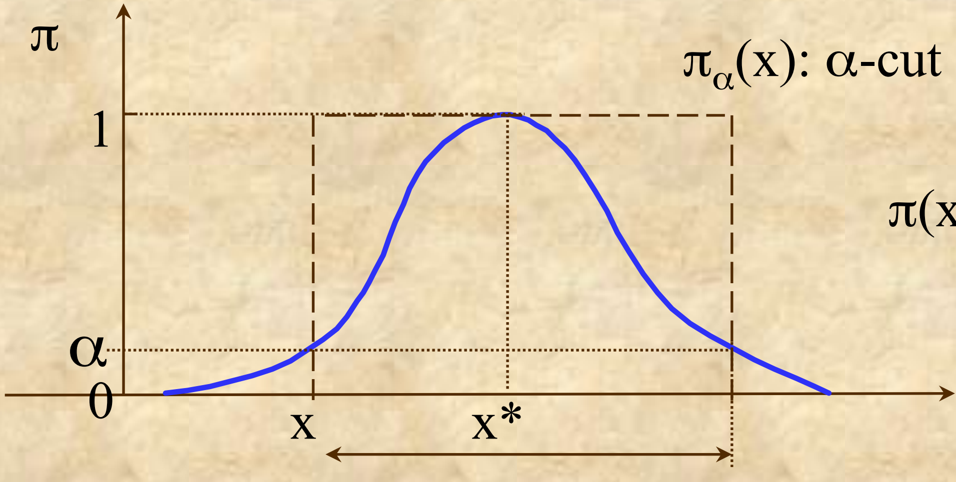
III- Relationships with descriptive statistical parameters

IV- Relationships with inferential statistical notions

V- Conclusion

Possibility/Interval Links

A possibility distribution can be viewed as gathering uncertainty intervals



$\pi_\alpha(x)$: α-cut of π (it is an interval)

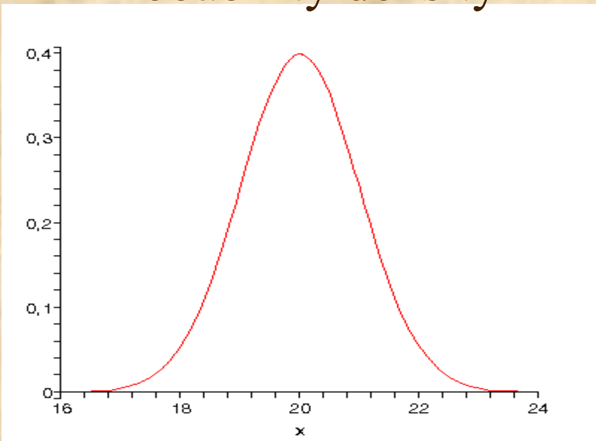
$\pi(x)=\sup_\alpha(\alpha.\pi_\alpha(x))$

Uncertainty interval

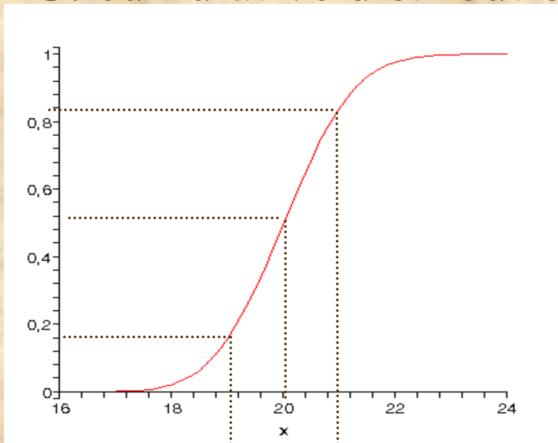
*The α-cuts of π can be identified to the $\beta=1-\alpha$ dispersion intervals of a probability density p around x**

Dispersion (or coverage) intervals

Probability density



G: cumulative distribution

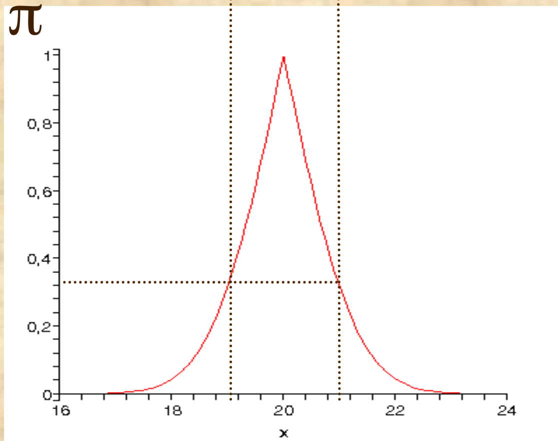


G^{-1}

A dispersion interval $I_{1-\alpha}$ of level $1\text{-}\alpha$ of a random variable X contains $(1\text{-}\alpha)\%$ of the population modeled by X, *it can be built around different centers (mode, median, mean).* For the median, it is defined by

$$\left[G_X^{-1}(\alpha/2), G_X^{-1}(1-\alpha/2)\right]$$

π



The set of dispersion intervals $I_{1-\alpha}$ for all the levels $1\text{-}\alpha$ constitutes a set of nested intervals, i.e. a possibility distribution

Probability/Possibility conversion

The equivalent possibility distribution π of the dispersion intervals is defined by identifying them to the *α-cuts of π*

$\Rightarrow$ **probability/possibility conversion** $\forall A, \Pi(A) \geq P(A)$

Normalizing the probability density does not satisfy this condition

Two main ways of building dispersion intervals

Type 1 conversion around a center c (two tailed)

$\pi_x^{1c}(x) = G(x) + 1 - G(g(x)) = \pi_x^{1c}(g(x))$ $\quad g:[-\infty, c] \to [c, +\infty]$ decreasing/ *g(c)=c*

$\forall x \in [-\infty, M], g(x) = \{y \geq M \mid p(x) = p(y)\}$ **gives the most specific**

Type 2 conversion about a center c (one tailed)

$\pi_x^{2c}(x) = \min(\frac{G(l(x))}{G(c)}, \frac{1 - G(r(x))}{1 - G(c)}, 1)$ *r* increasing, *l* decreasing ***r(c)=l(c)=c***

For continuous symmetric X about *c*: type1=type2 (*l=r=id*)

$$\pi_x^{1c}(x) = \pi_x^{2c}(x) = \min(2G(x), 2(1 - G(x)))$$

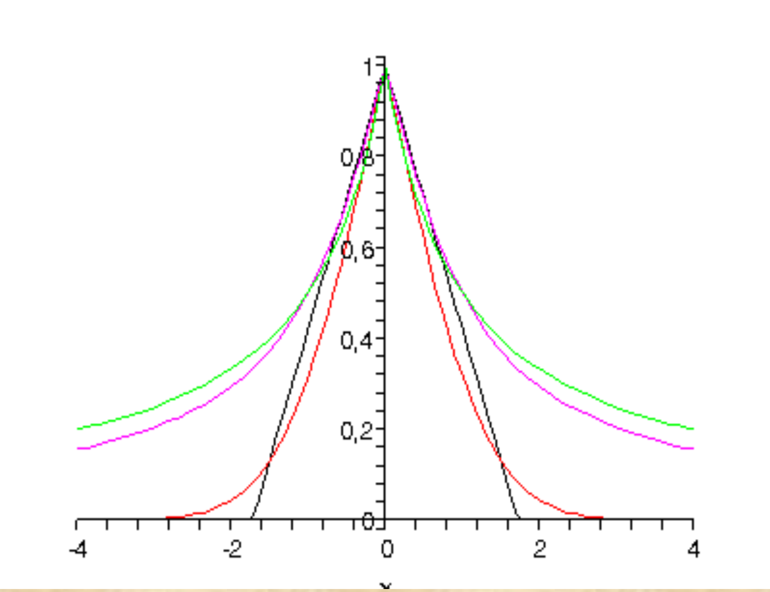
Conversion examples

Uniform (black)

Gauss (red)

Cauchy (magenta)

Pareto-Sym (green)

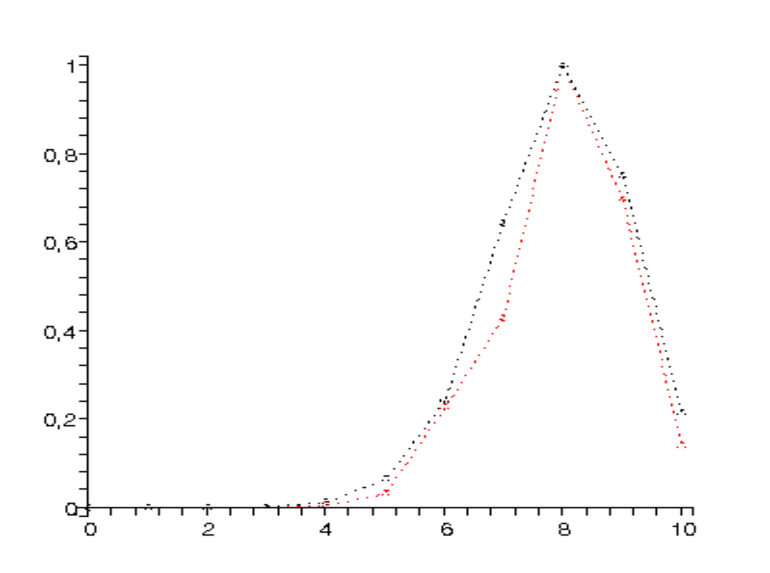


Binomial

$$P(S_n = k) = C_n^k p^k (1-p)^{n-k}$$

$$E = np$$

The mode and the median are varying according to *n* et p



Possibility conversion of a probability family

Not easy to identify a single probability distribution [Gauss 1823]

=> Probability inequalities

Gauss inequality: family of unimodal symmetric distributions having the same variance and the same mode

Bienaymé-Chebyshev [1853]: family of distributions having the same mean and the same variance

The possibility distribution is obtained by taking the envelop of the dispersion intervals of all the probability distributions

$$\pi(t) = \max_{X \in P} \Pr(|X - m| \geq t)$$

This maximum specificity principle is better founded that the maximum entropy principle

Infinite support family conversion examples

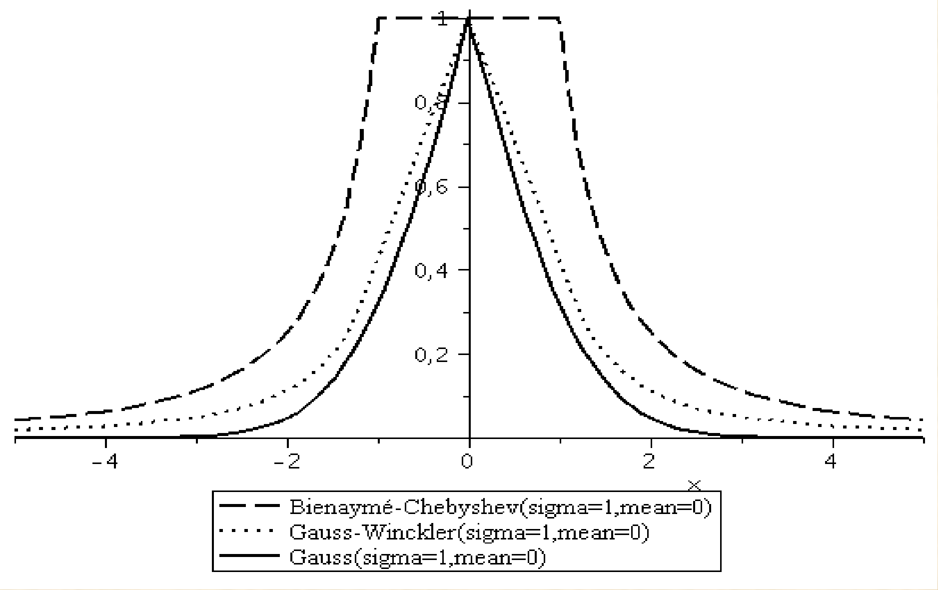
Mean and standard deviation ->BC $\pi(m-t)=\pi(m+t)=P\left[\left|X-m\right|>t\right]=\min(1,\frac{\sigma^2}{t^2})$

+

Mode + standard deviation -> GW $\pi(m-t)=\pi(m+t)=\min(1,\max(1-\frac{t}{\sqrt{3}\sigma},\frac{4\sigma^2}{9t^2}))$

+

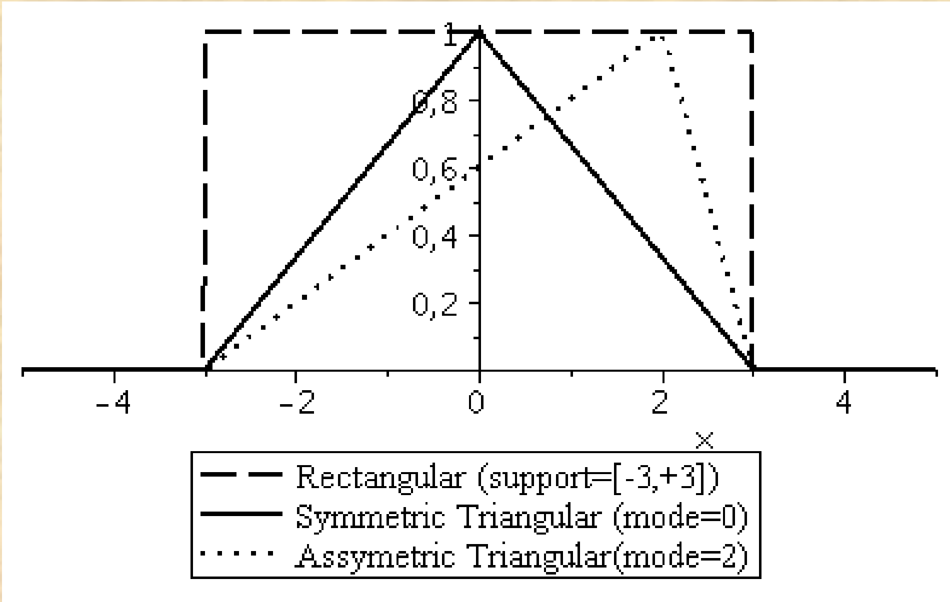
known distribution e.g. Gauss $\pi(m-t)=\pi(m+t)=1-2\left|F_{G(0,\sigma)}(t)-0.5\right|$



Bounded support family conversion examples

Only the range is known ->rectangular possibility distribution

Unimodality and symmetry -> triangular distribution



Plan

I- Generalities

II- Probability possibility conversion

III- Relationships with descriptive statistical parameters

IV- Relationships with inferential statistical notions

V- Conclusion

Specificity/Loss functions

$$\pi_{T_\theta(X)}(x) \le \pi_{U_\theta(X)}(x) \Leftrightarrow E_\vartheta(L(T_\theta(X)) \le E_\vartheta(L(U_\theta(X))$$

L is any loss (or risk) function, e.g. :

$L(x,\theta) = |x-\theta|$ Laplace

$L(x,\theta) = |x-\theta|^2$ Gauss

Rem: for a symmetric continuous variable $T_\theta(X)$

$$spindex(\pi_{T_\theta(X)}) = \int_{-\infty}^{+\infty} \pi_{T_\theta(X)}(x)dx = 2.E\left|T_\theta(X) - \theta\right|$$

Specificity/Entropies

H is any generalized entropy : $H(f) = -\int_{-\infty}^{\infty} \varphi(f(x))dx$

φ convex and continuous

For the Shannon entropy $H(X) = -\int_{-\infty}^{\infty} f(x)Ln(f(x))dx$

For two continuous unimodal symmetric probability densities *f* and *g* and

$$\pi^{f}(x) \le \pi^{g}(x), \forall x \Leftrightarrow H_{\varphi}(f) \le H_{\varphi}(g)$$

[Mauris,2010][Couso and Dubois, 2010]

Specificity/SOSD, VaR and Gini

For continuous symmetric random variables

SOSD $$F \leq_{SOSD} G \Leftrightarrow \int_{-\infty}^{t} F(x)dx \leq \int_{-\infty}^{t} G(x)dx, \forall t$$

$$\pi_X^{\theta}(x) \leq \pi_Y^{\theta}(x) \Rightarrow X \leq_{SOSD} Y$$

VaR $$P(X > VaR_{\alpha}(X)) = 1 - \alpha$$

$$\pi_X^{\theta}(x) \leq \pi_Y^{\theta}(x) \Rightarrow VaR_{\alpha}(X) \geq VaR_{\alpha}(Y), \forall \alpha \in [0,1]$$

Gini $$L_X(t) = \frac{1}{E(X)} \int_O^t F(x)dx \qquad G(X) = 1 - 2\int_O^1 L_X(t)dt$$

$$\pi_X^{\theta}(x) \leq \pi_Y^{\theta}(x) \Rightarrow G(X) \geq G(Y)$$

Specificity/Peakedness [Birnbaum, 1948]

$$X \geq_{\theta}^{peaked} Y \Leftrightarrow \Pr(|X-\theta| \geq t) \leq \Pr(|Y-\theta| \geq t), \forall t$$

Peakedness is related to conventional stochastic ordering

$$X \geq^{peaked} Y \Leftrightarrow |X-\theta| \leq^{sto} |Y-\theta| \overset{def}{\Leftrightarrow} F_{|X-\theta|}(x) \leq F_{|Y-\theta|}(x)$$

For unimodal continuous symmetric random variables

$$\pi_X^{\theta}(x) \leq \pi_Y^{\theta}(x) \Leftrightarrow X \geq^{peaked} Y \Leftrightarrow X \leq^{maj} Y$$

The same holds for discrete random variables (Dubois and Hùllermeier 2007)

Specificity/Lévy concentration [1935)]

$$\forall x \geq 0, Q_{X'}(x) = \sup_{x_0}\left[F(x_0 + x) - F(x_0 - x)\right]$$

Introduced by Lévy for overcoming the limitation of using one dispersion parameter, e.g. the standard deviation

For unimodal symmetric distributions

$$\forall x \geq 0, Q_{X'}(x) = 1 - \pi_{X'}^{\theta}(x + \theta)$$

The concentration is the complement of dispersion

Plan

I- Generalities

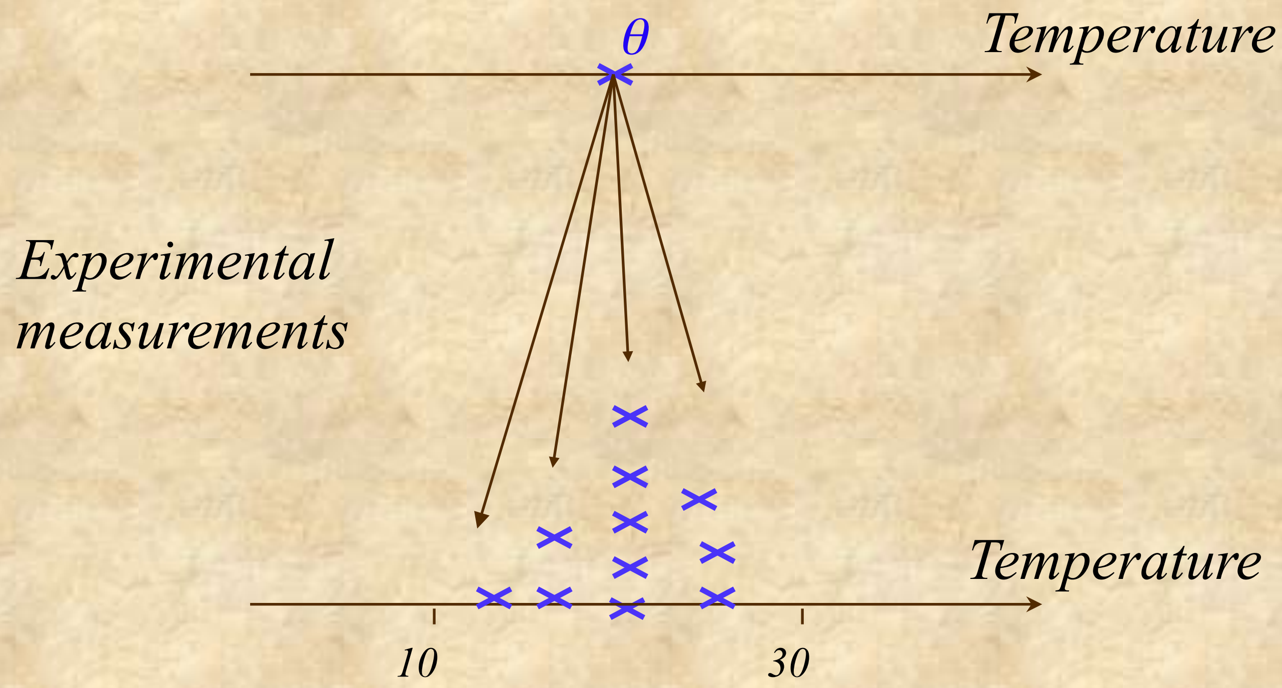
II- Probability possibility conversion

III- Relationships with descriptive statistical parameters

IV- Relationships with inferential statistical notions

V- Conclusion

Parameter inference example

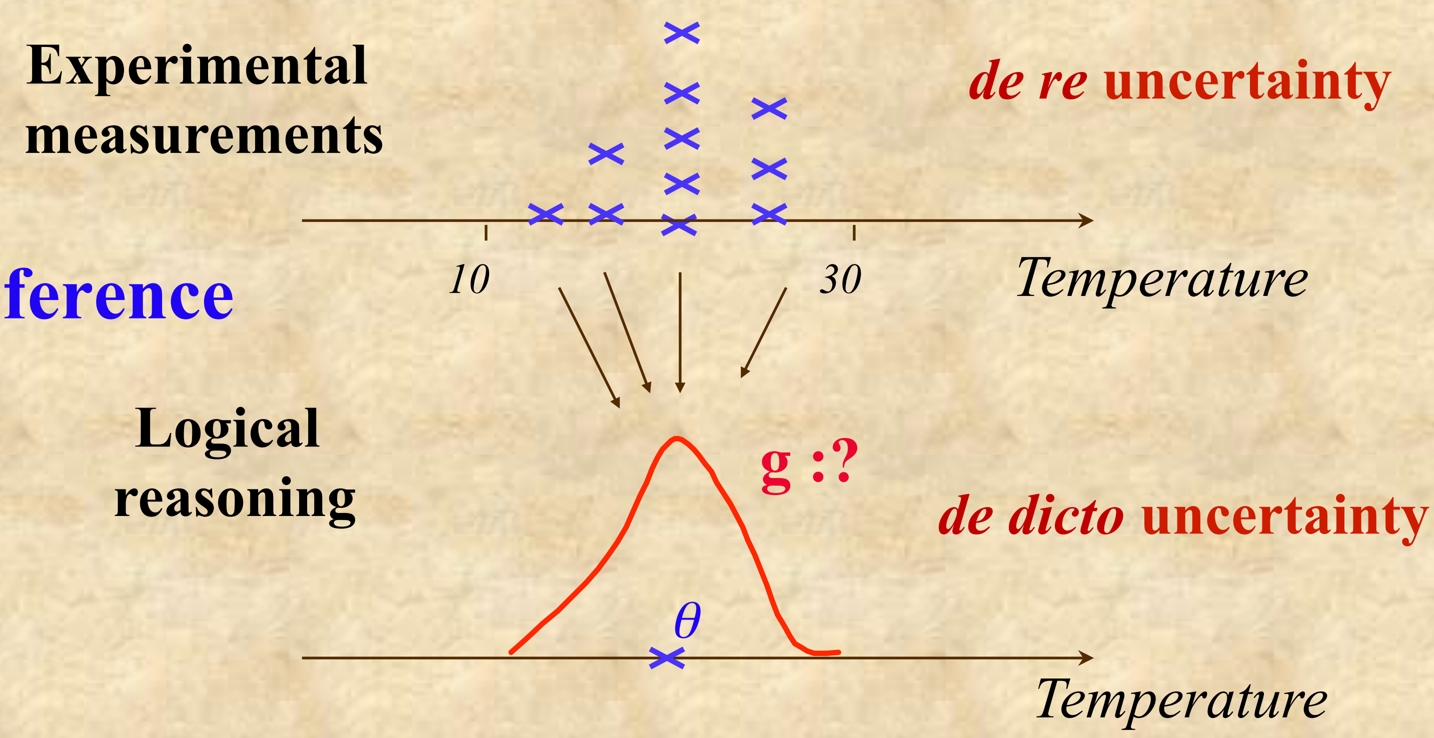


θ is a fixed unknown parameter, the variability of the measurements is due to the measurement process

The measurements are realizations of a random variable X

Parameter inference problem

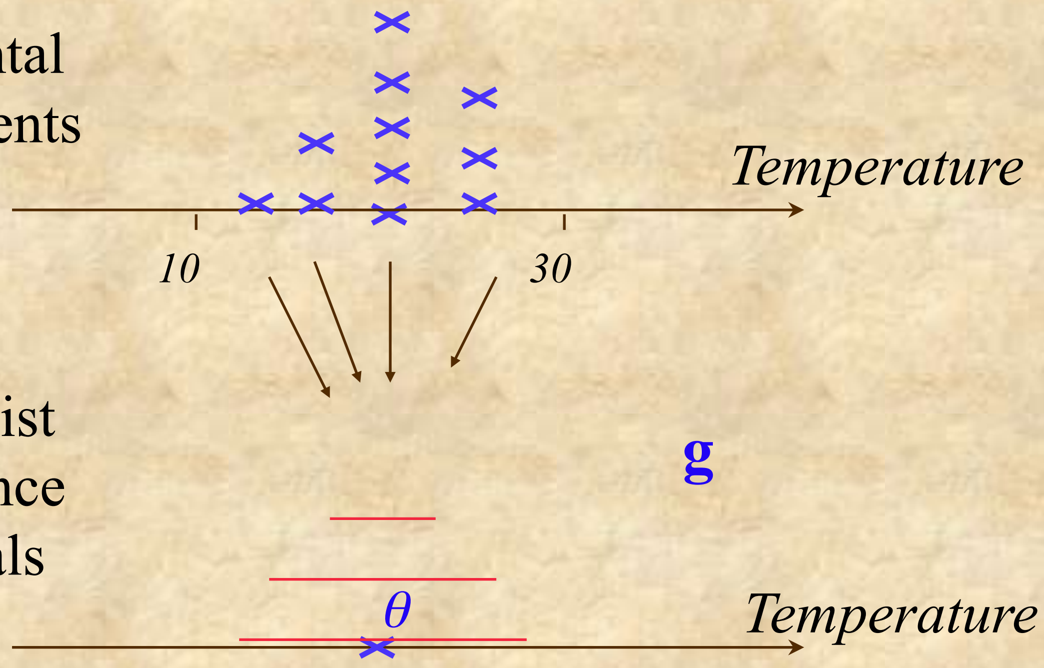
Inference



Representing θ by $g(\theta / x)$ from the measurements x_i's and from an inference method: distribution of probability, possibility, plausibility,..

Conventional Probability Inference

Experimental measurements



frequentist confidence intervals

g: sets of confidence intervals , i.e. random intervals

$$\Pr\left[u(X) \le \theta \le v(X) \right] = 1 - \alpha$$

u, v statistics *derived from the measurements*

Case of Proportion Estimation

At the root of the justification of the estimation of an unknown probability by the observed realized frequency on a large sample

Weak law of large numbers (J. Bernoulli ~1700)

Involved in a lot of practical problems concerning estimation from samples

e.g. number of defective parts in a production

Weak law of large numbers

Jacques Bernoulli *Ars Conjectandi* 1713

$$\forall\varepsilon > 0 \qquad P(|F_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

The sampled proportion F_n converges to the true one p with probability *1*

$$\forall\varepsilon > 0, P(p-\varepsilon \leq F_n \leq p+\varepsilon) \leq \frac{1}{4n\varepsilon^2} \qquad \forall\varepsilon > 0, P(F_n-\varepsilon \leq p \leq F_n+\varepsilon) \leq \frac{1}{4n\varepsilon^2}$$

The first probability inequality!

Knowing p allows to deduce the sampled dispersion of F_n with a definite probability

***de re* dispersion intervals**

Observing the sampled F_n allows to induce the proportion p with a definite confidence

***de dicto* confidence intervals**

Confidence interval issues

When the random variable X is replaced by its realization we obtain usual **numerical confidence intervals or realized confidence intervals**

$$\left[L(f_n),U(f_n)\right]$$ f_n sampled proportion

$$p\in\left[L(f_n),U(f_n)\right]$$ **is either true or false and is not subject to a probability statement in a frequentist sense**

The theoretical confidence interval is a procedure which once reiterated satisfies a success ratio equal to the confidence level

e.g.: for 100 realized confidence intervals of level 90%, 90 contain the parameter

By transfer of the confidence level of the theoretical confidence interval to the realized confidence interval a *de dicto* uncertainty level is obtained for *p*

Confidence intervals / possibility distribution

The function $\inf_{p \in [0,1]} P_p$ **defined by** $A \mapsto \inf_{p \in [0,1]} P_p(A)$

does not define a probability but indeed a necessity

$$\inf_{p \in [0,1]} P_p(L(x) \le p \le U(x)) \ge 1 - \alpha$$

defines a probability lower bound

By stacking up the realized confidence intervals for all the levels, a possibility distribution is obtained

$$\pi_{X=x}(x) = \min_{i=1,\ldots,m} \max(1-\alpha_i, I_i(x))$$

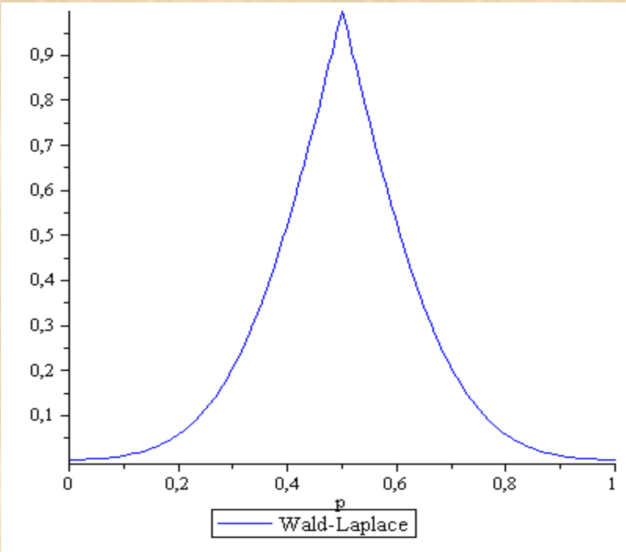
Expresses the conjunction of the possibility distributions issued from each level realized confidence interval and it corresponds to the most specific distribution versus the available data [Dubois-Prade 1992]

Conventional approach (Wald)

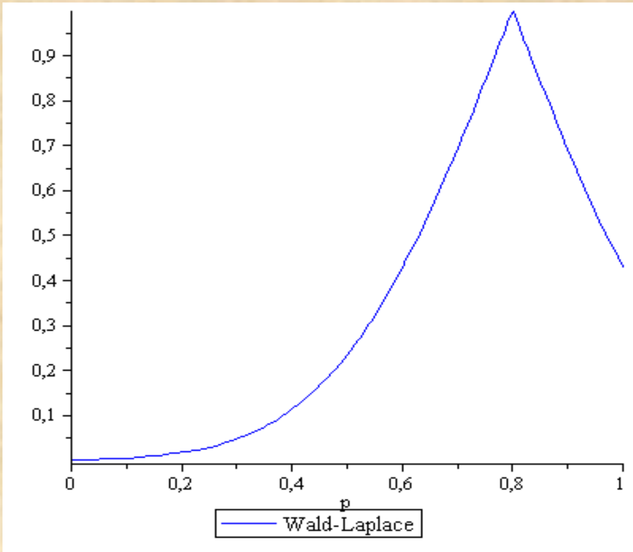
based on the approximation of the binomial law by a Gaussian one proposed by De-Moivre and Laplace

F_n : the random variable associated to the sampled proportion and f_n one of its realizations; p the unknown fixed proportion

$$F_n - p \simeq N(0, \sqrt{\frac{f_n(1-f_n)}{n}})$$



n=10 f_n=0.5



n=10 f_n=0.8

Laplace most advantageous method

Simon Laplace *Essai Philosophique* 1814

"Le procédé d'estimation le plus avantageux est évidemment celui dans lequel une même erreur dans les résultats est moins probable que suivant tout autre procédé"

"the most advantageous" method is the one in which the error of the results is less probable as with any other method

This principle is equivalent to say that the estimator T is better than U if

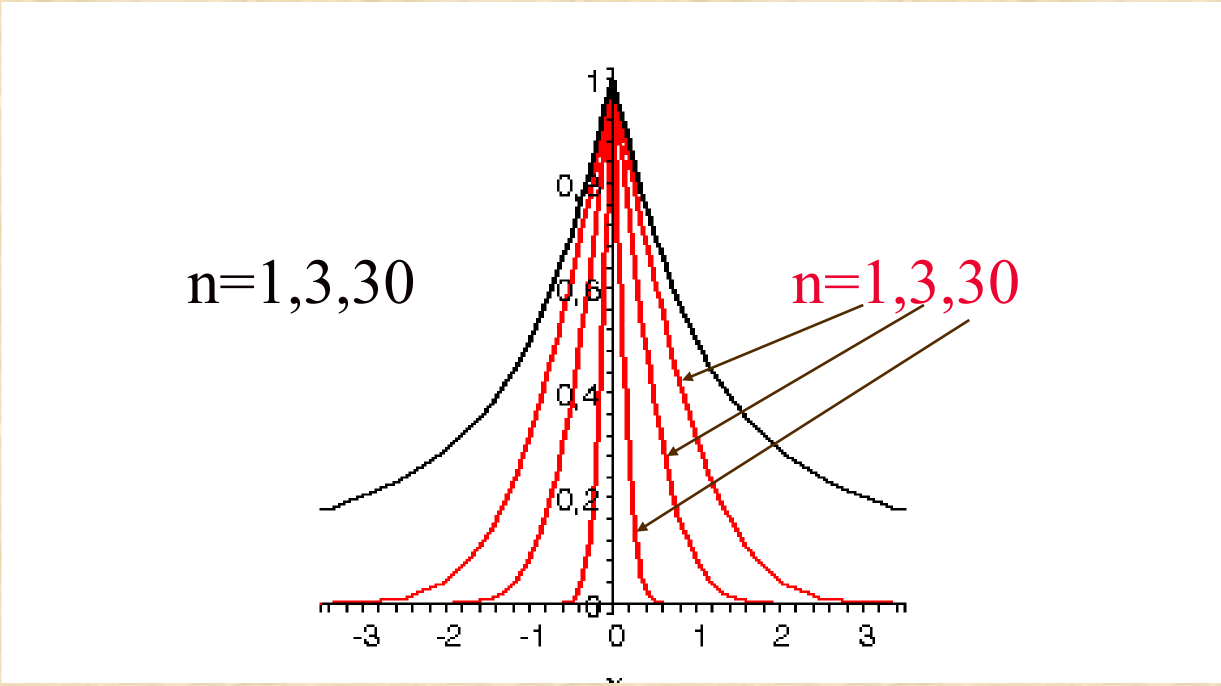
$$\Pr(\left|T(\vec{X}_\theta) - \theta\right| \geq t) \leq \Pr(\left|U(\vec{X}_\theta) - \theta\right| \geq t), \forall t \geq 0$$

This is equivalent to the maximum specificity possibility principle that is more general than the minimum variance principle

Laplace has proved that for the Gaussian distribution the most advantageous method has minimal variance (Least square)

Possibility view of mean estimator

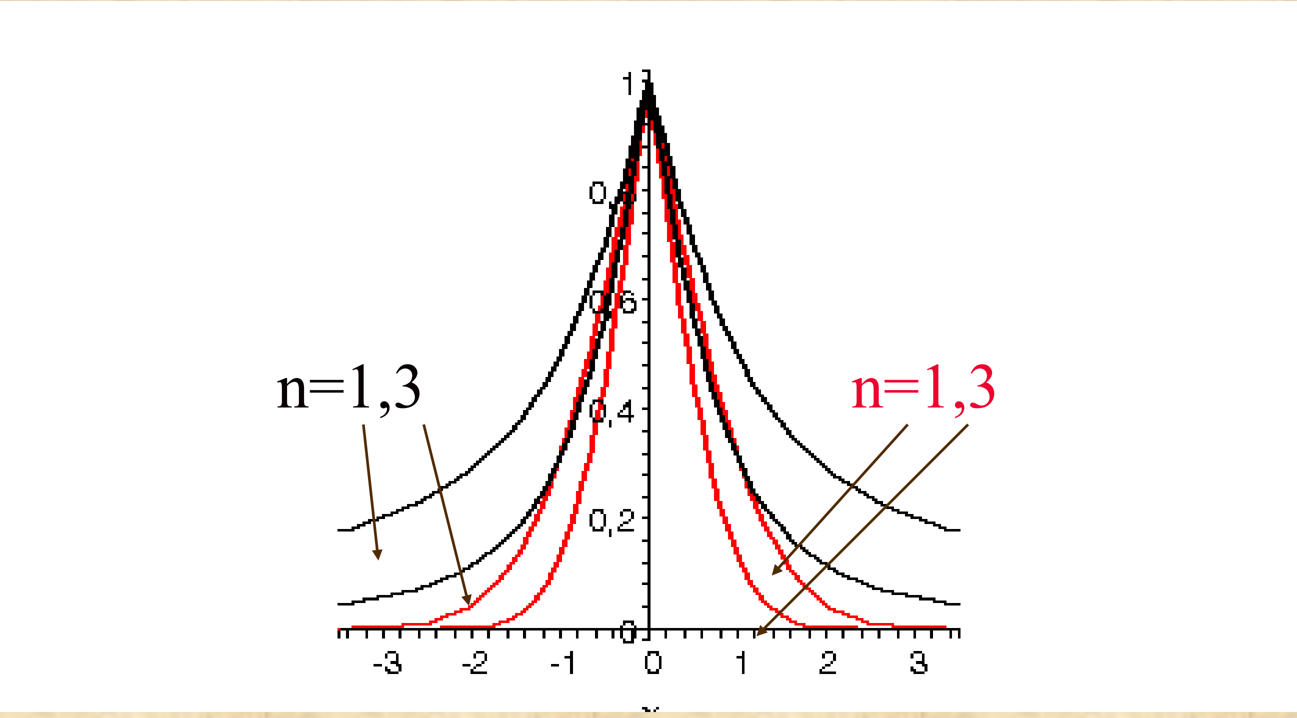
Let us consider again the temperature example and that the measurements are an *iid* sample from a continuous symmetric distribution, e.g. Gauss and Cauchy with dispersion=1 mean=0



For the Gauss distribution, the specificity increases with the sample size, but not for the Cauchy distribution

Possibility view of median estimator

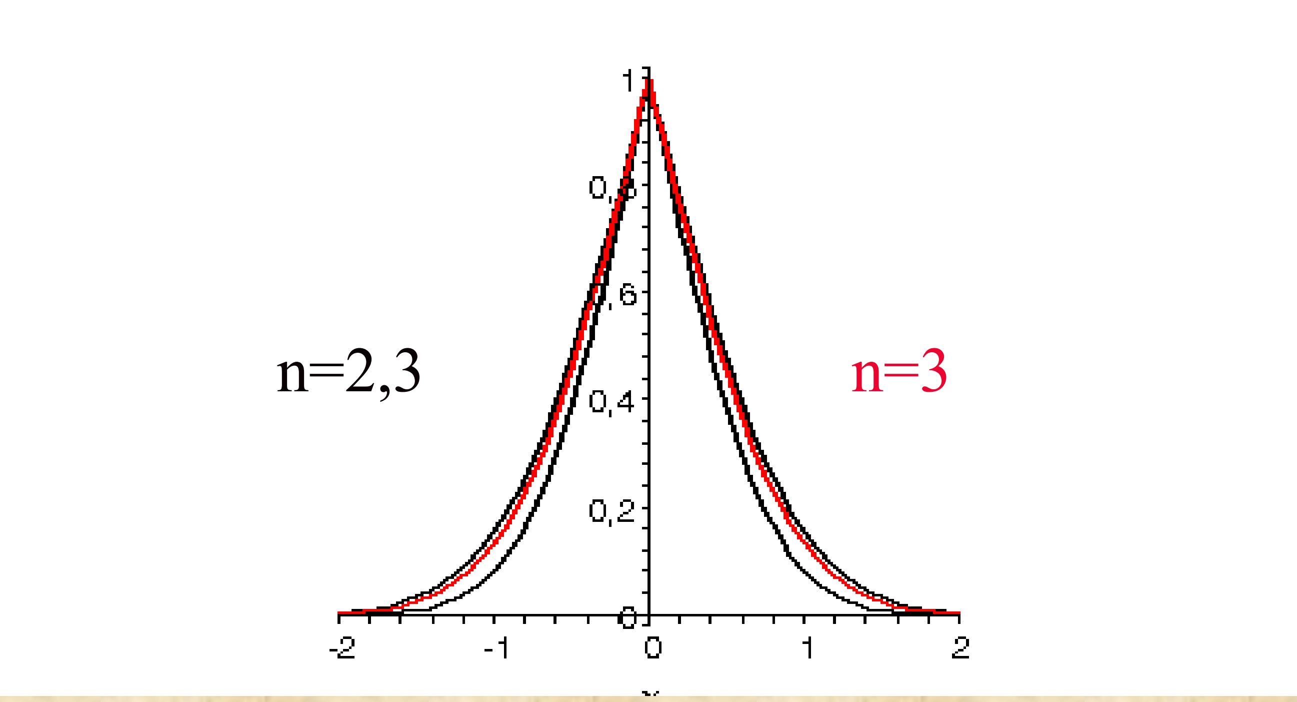
Let us consider an *iid* sample from a continuous symmetric distribution, e.g. Gauss and Cauchy with dispersion=1 mean=0



For the Gauss and also the Cauchy distributions, the specificity increases with the sample size

Comparison median versus mean

Let us consider an *iid* sample from a Gauss distribution with standard deviation=1 mean=0



It seems that the possibility median estimator for 2n+1 data is more specific than the mean estimator for 2n data

Inference with poor knowledge

- Case of very few measurements (Gauss approximation not applicable)
- Limited knowledge about X (no single probability)

It can be modeled by a family F of probability distributions, rather than selecting a single one

Again probability inequalities can be used to define a possibility distribution dominating all probability distributions in the family

Illustration with one measurement

The distribution, the support and the variance are unknown

P*(θ ε *[X – k* |*X*| , *X* +*k* +|*X* |]) ≥ *1 – 2/(k+1) (k>1)(unimodal)

P*(θ ε *[X – k* |*X*| , *X* +*k* +|*X* |]) ≥ *1 – 1/(k+1) (k>1) (+ symmetric)

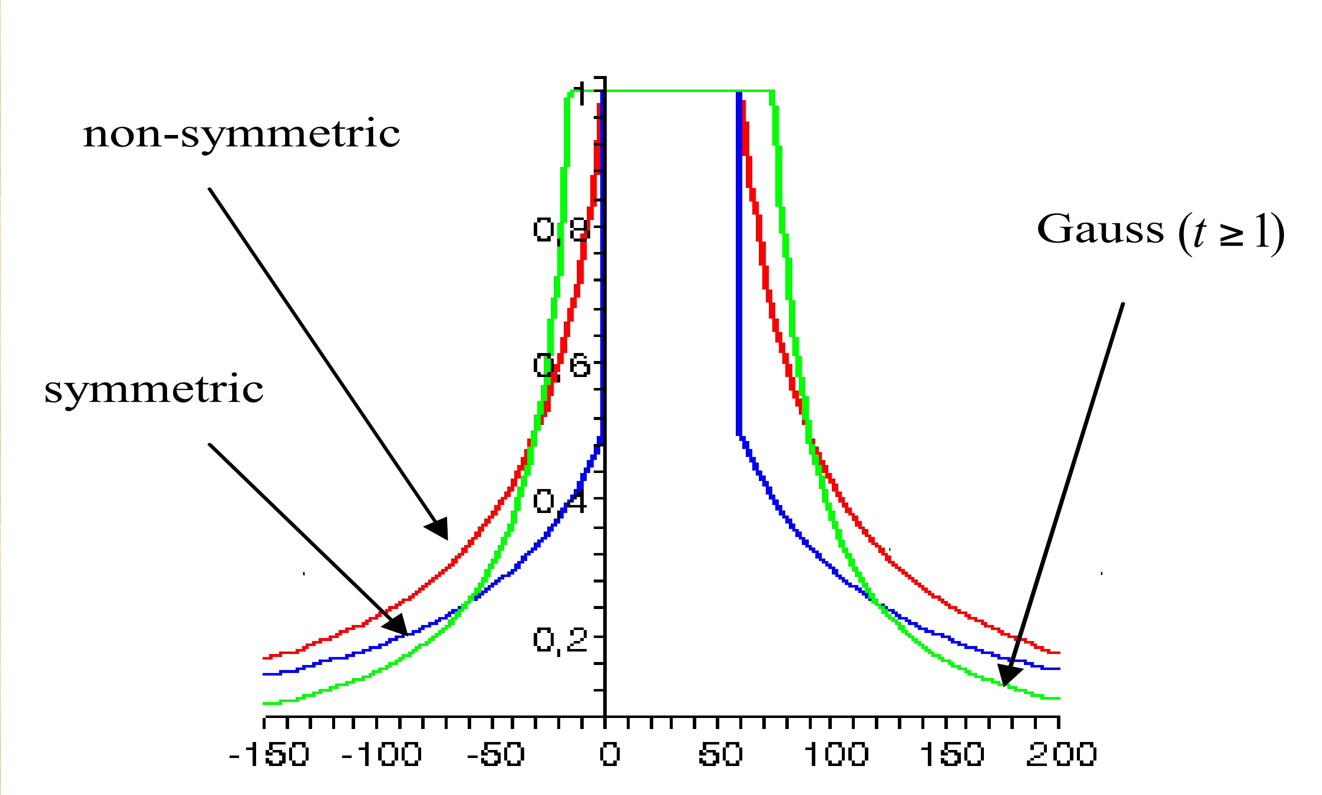


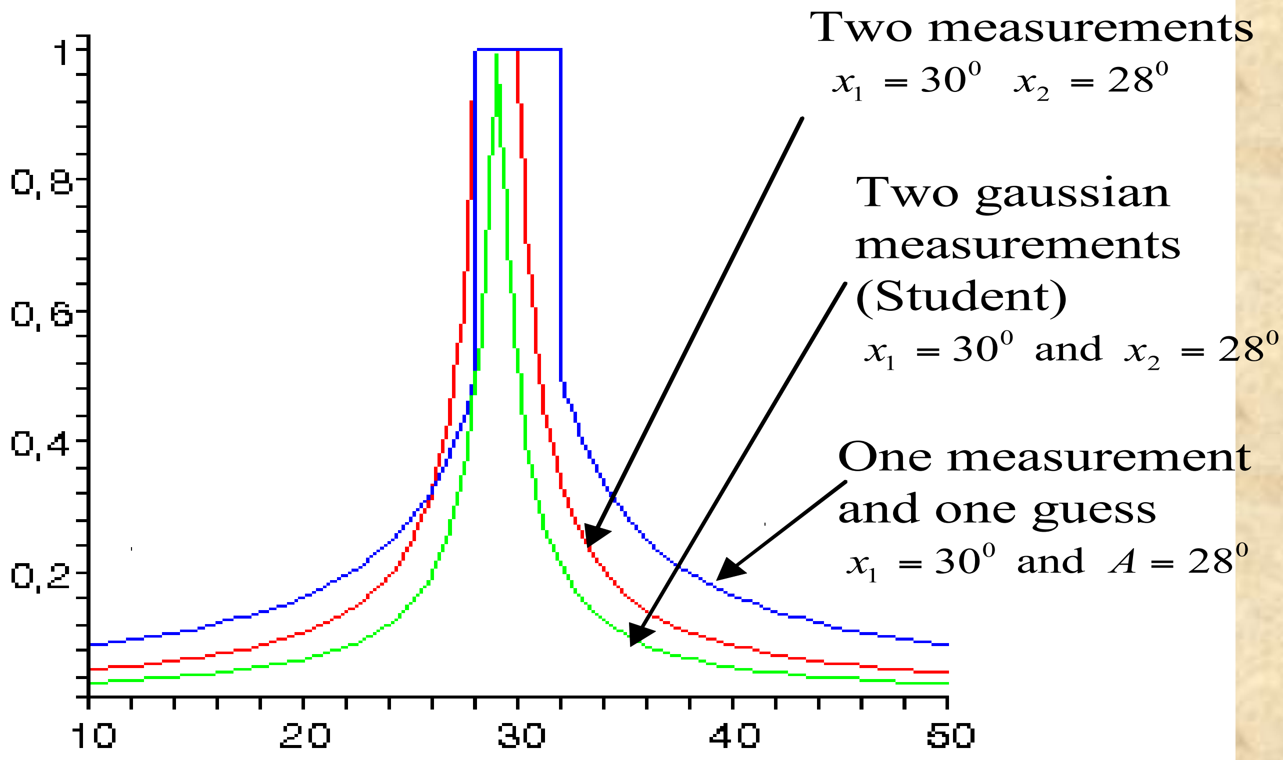
Illustration with two measurements X, Y

The distribution, the support and the variance are unknown

$$P(\theta \in [(X+Y)/2 - k\,|X-Y|/2\,,\ (X+Y)/2 + k\,|X-Y|/2\,]) \geq 1 - 2/(k+1) \quad (k>1)$$

$$P(\theta \in [(X+Y)/2 - k\,|X-Y|/2\,,\ (X+Y)/2 + k\,|X-Y|/2\,]) \geq 1 - 1/(k+1) \quad (k>1)$$

(symmetric)



Conclusion/Perspectives

A possibility distribution can provide a useful uncertainty representation related to many conventional descriptive and inferential statistical notions

The maximum specificity principle (i.e. fuzzy subset inclusion) is a strong general principle for statistical inference

Casting the Fisher and Bayesian approaches (credible fiducial and intervals) in the possibility framework?

Thank you for your attention