Notions de statistiques:

validations des résultats et maîtrise de la qualité des analyses

R. Losno, Professeur Université Paris7 Denis Diderot, Laboratoire LISA

losno@lisa.univ.paris12.fr

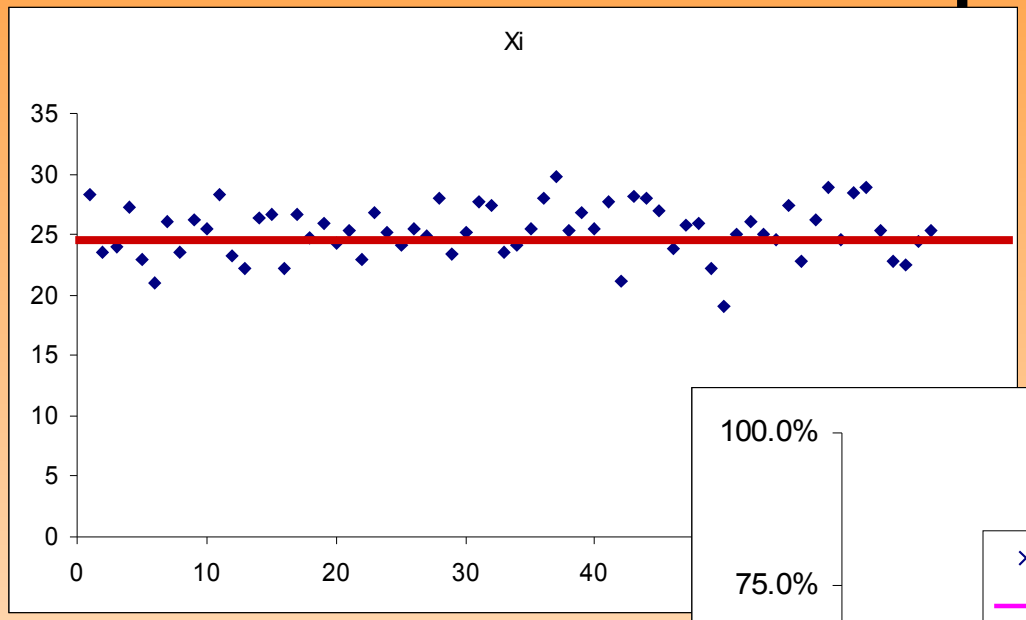
http://www.lisa.univ-paris12.fr

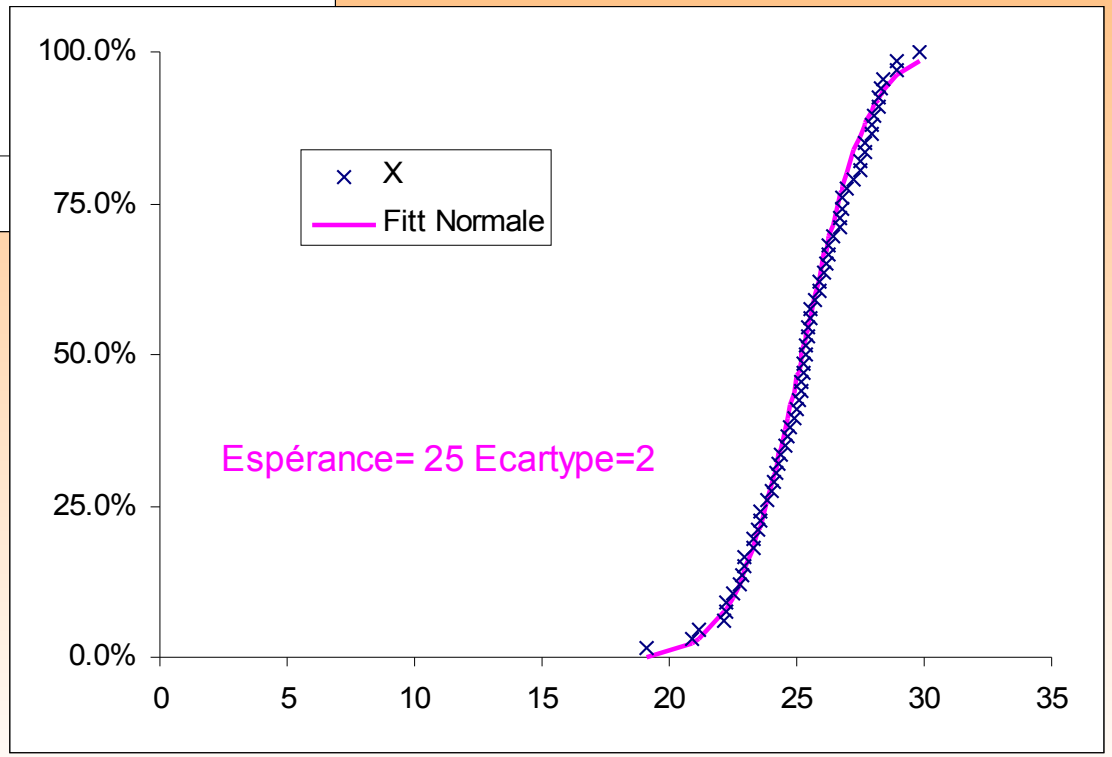
Apport des méthodes statistiques

- But d'une analyse:
  - En théorie: trouver la valeur vraie (μ)
  - En pratique: dispersion des résultats due aux sources de variations

$$x_i = \mu + \varepsilon$$

Exemples

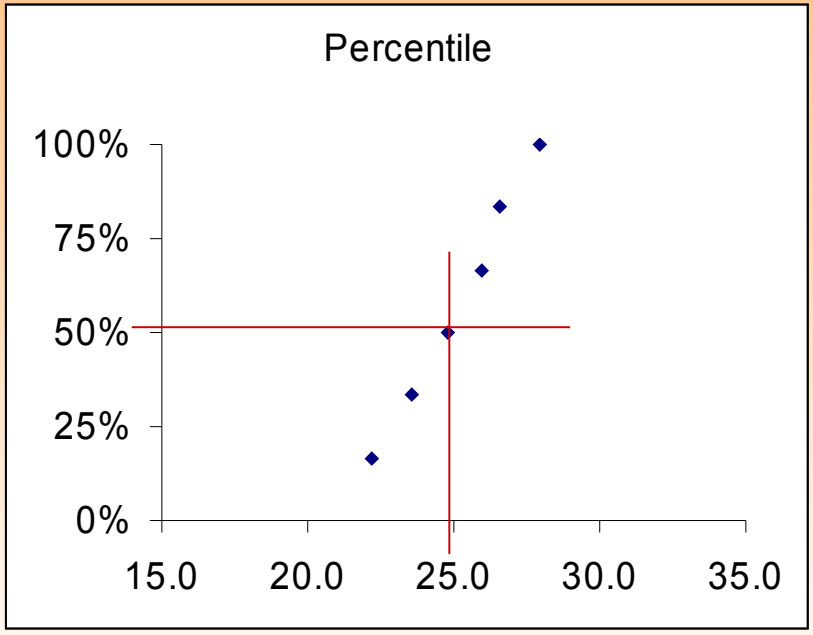




Courbe percentile

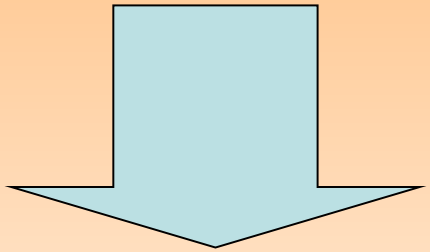
La représentation percentile

X	X trié	Rang	Percentile
25.9	22.2	1	17%
24.8	23.5	2	33%
27.9	24.8	3	50%
22.2	25.9	4	67%
26.6	26.6	5	83%
23.5	27.9	6	100%



Apport des méthodes statistiques

- Quelle est la valeur qu'on prend pour estimer la valeur vraie?
- Quelle est la qualité de cette valeur estimée?



Méthodes statistiques

« Statistique descriptive »

Apport des méthodes statistiques

- **Quelle est la valeur qu'on prend pour estimer la valeur vraie?**

  Détermination de la valeur la plus probable

  Calcul de valeur centrale

- **Quelle est la qualité de l'estimation?**
  - **Ecart type et variance**

Détermination de la fidélité de la mesure (precision) =

Répétabilité (repeatability): Étroitesse de l'accord entre les résultats des mesurages successifs réalisés dans les mêmes conditions de mesure

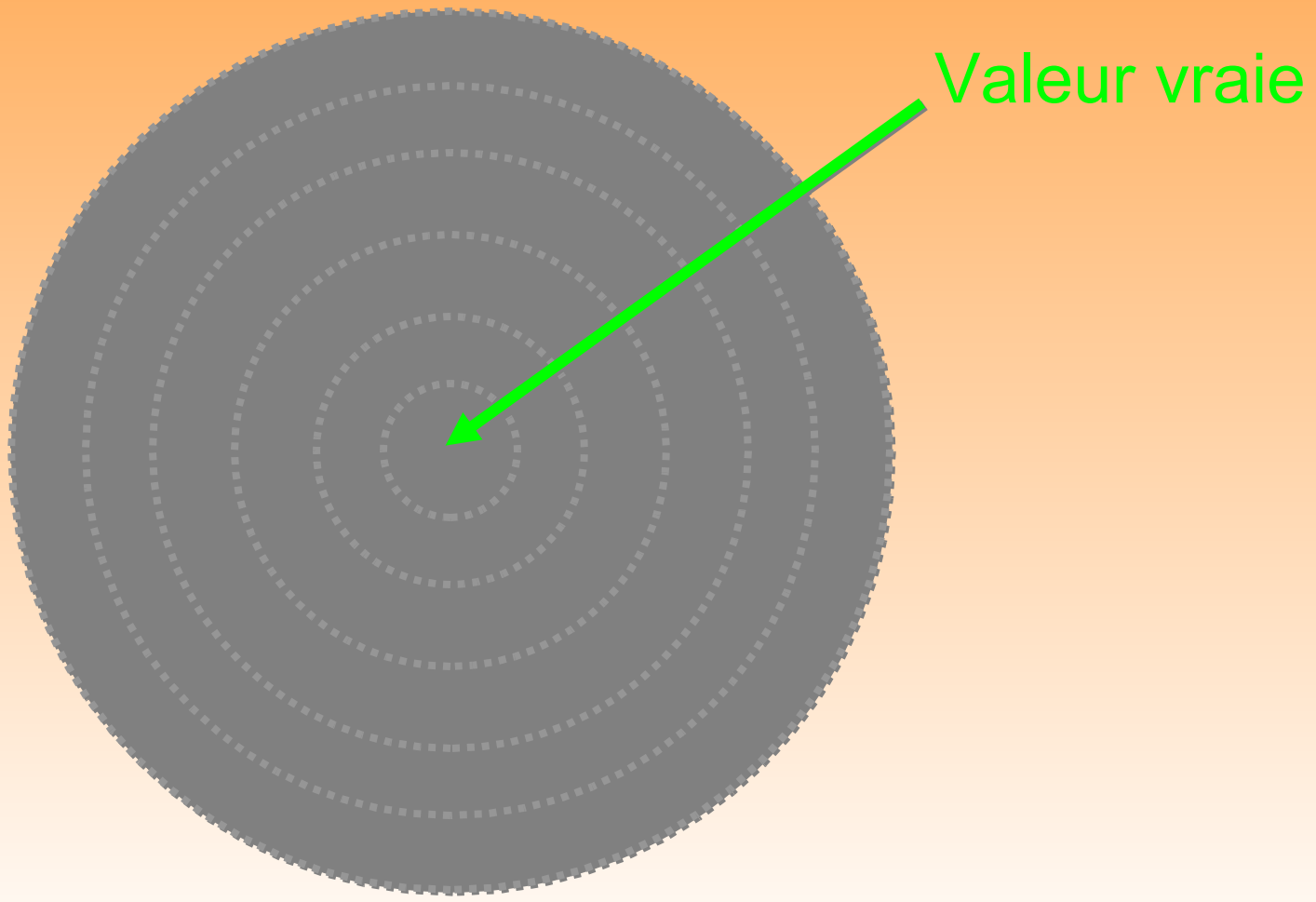
Reproductibilité (reproducibility): Étroitesse de l'accord entre les résultats des mesurages successifs réalisés en faisant varier les conditions de mesure

Calcul de dispersion statistique

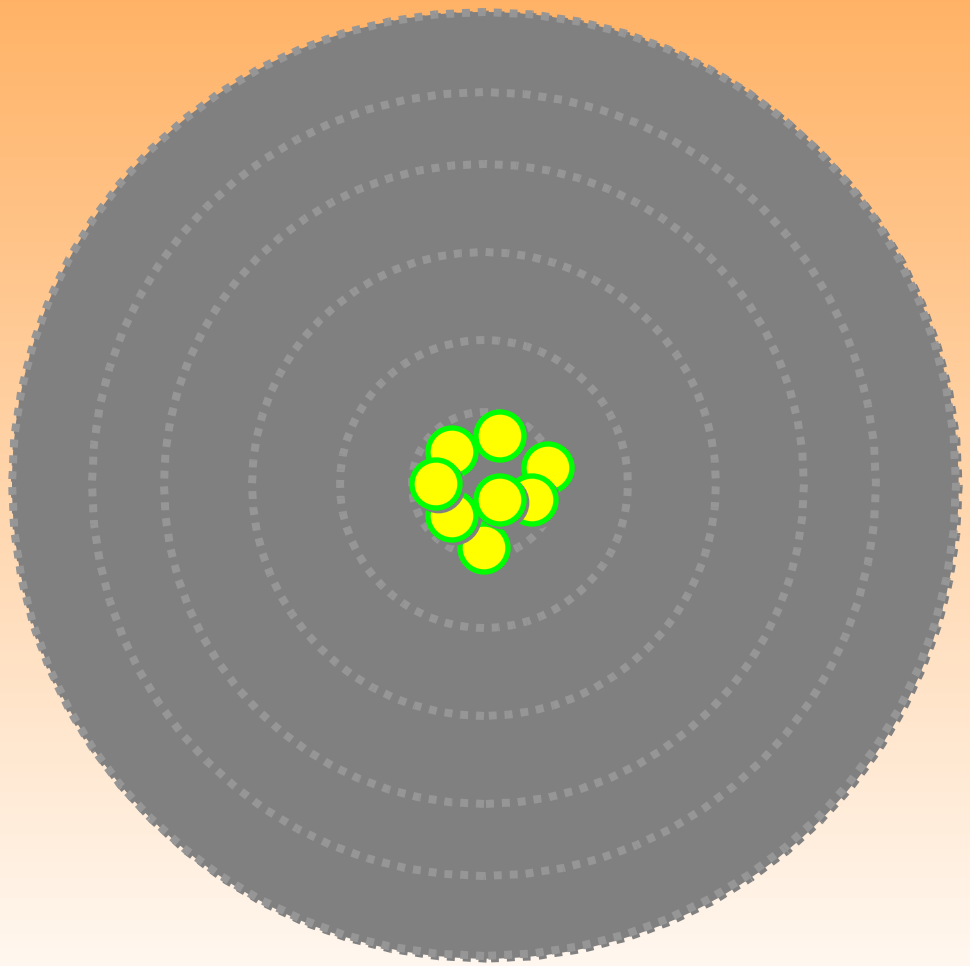
Détermination de la justesse de la mesure (accuracy) = étroitesse de l'accord entre valeur estimée et valeur vraie

Calcul d'erreur systématique

Qualité de la mesure

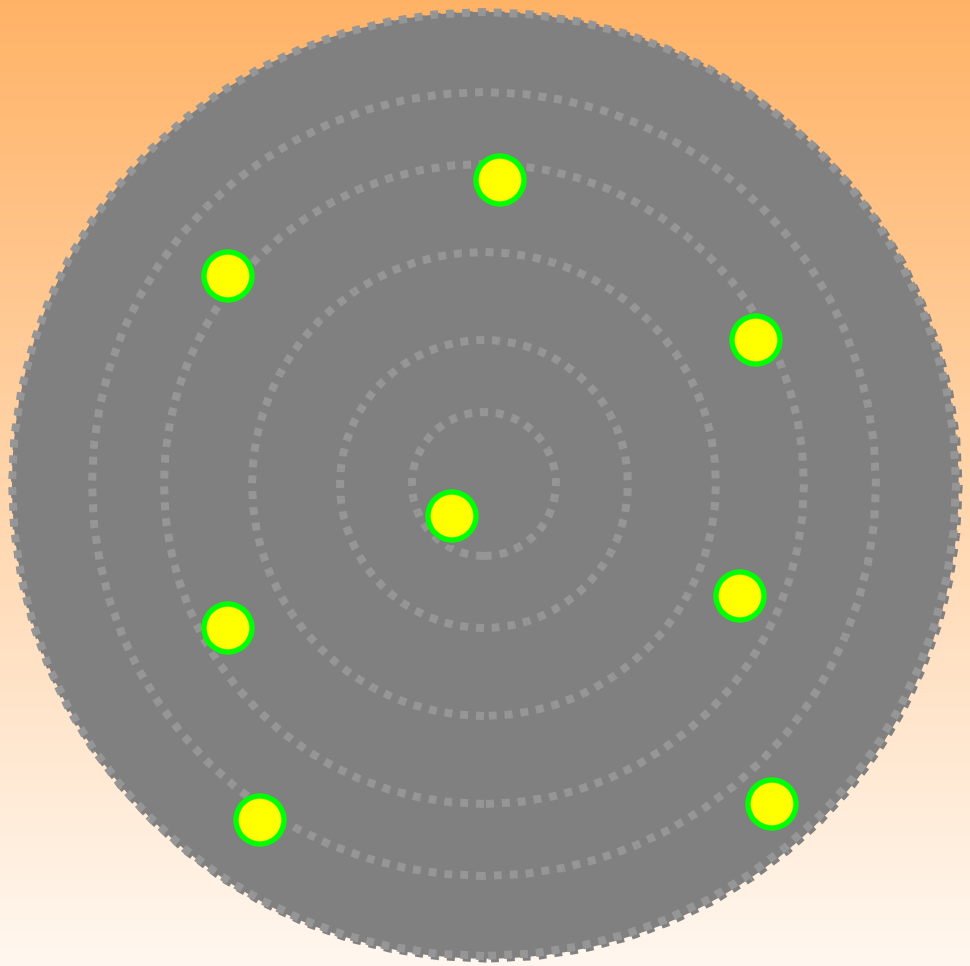


Qualité de la mesure



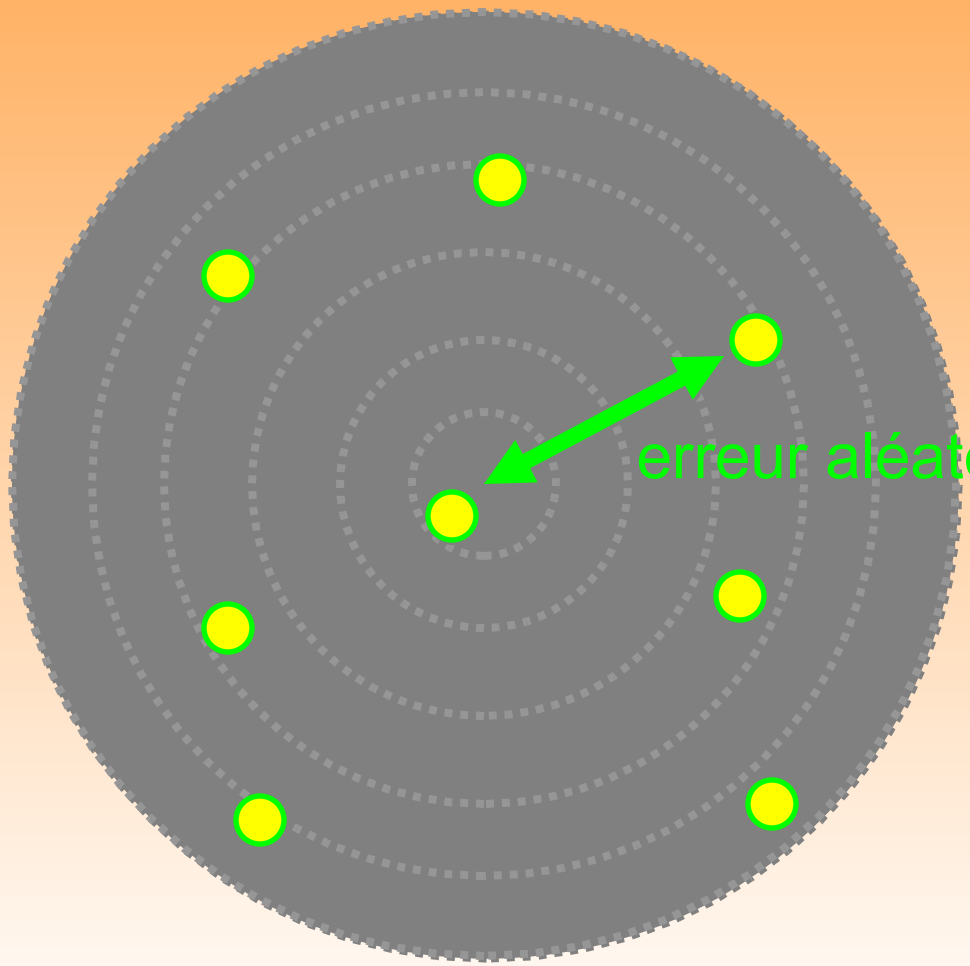
Juste, et
répétable

Qualité de la mesure



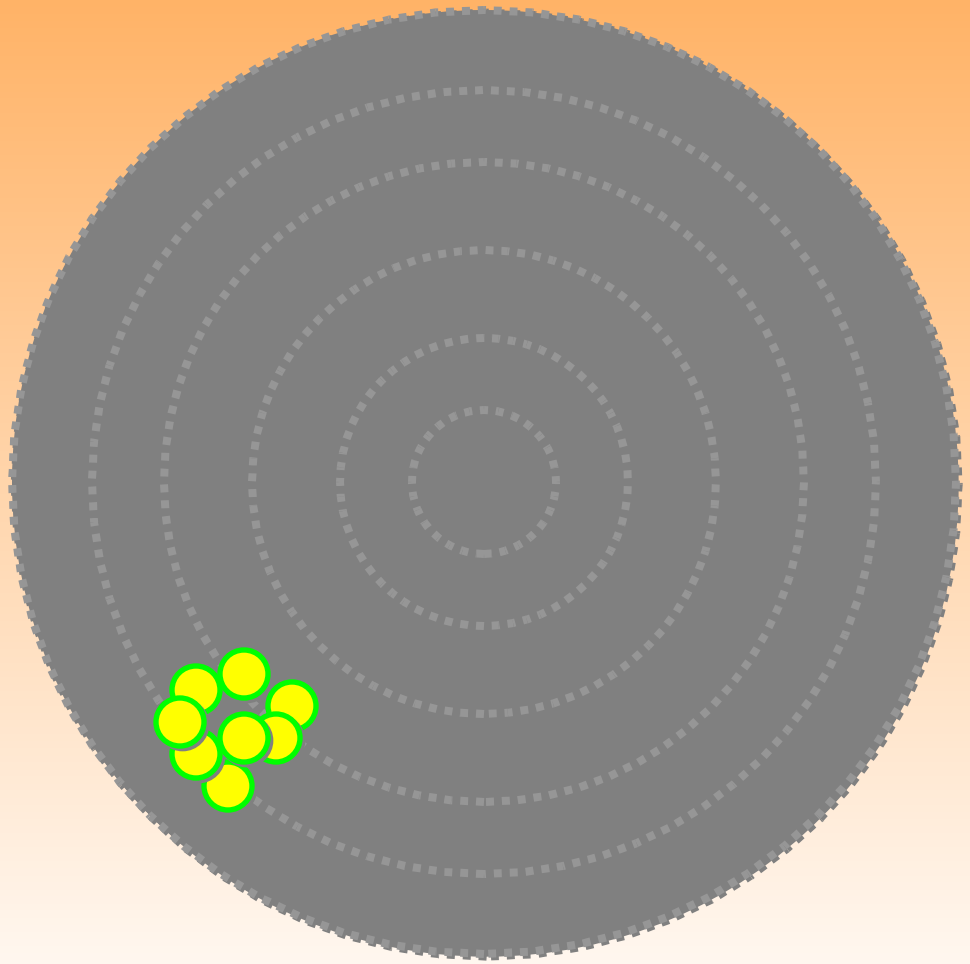
Peu répétable

Qualité de la mesure



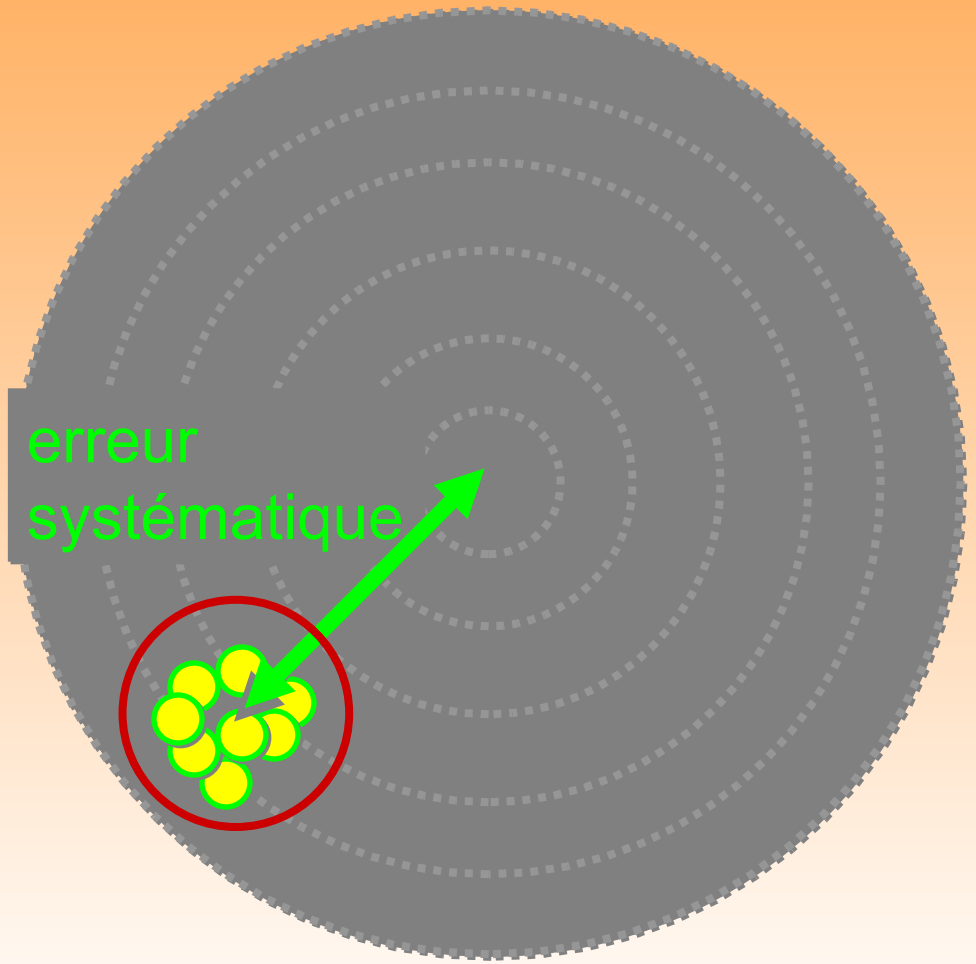
Peu répétable

Qualité de la mesure



Pas juste, mais répétable

Qualité de la mesure



Pas juste, mais
répétable

Apport des méthodes statistiques

- **Quelle est la valeur qu'on prend pour estimer la valeur vraie?**

  Détermination de la valeur la plus probable

  ➡ Calcul de valeur centrale

- **Quelle est la qualité de l'estimation?**

  – *Détermination de la fidélité de la mesure (precision)* =
    - Répétabilité (repeatability): Étroitesse de l'accord entre les résultats des mesurages successifs réalisés dans les mêmes conditions de mesure
    - Reproductibilité (reproducibility): Étroitesse de l'accord entre les résultats des mesurages successifs réalisés en faisant varier les conditions de mesure

  ➡ Calcul de dispersion statistique

  – *Détermination de la justesse de la mesure (accuracy) = étroitesse de l'accord entre valeur estimée et valeur vraie*

  ➡ Calcul d'erreur systématique

1. Valeur centrale

- Moyenne empirique ou arithmétique:

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

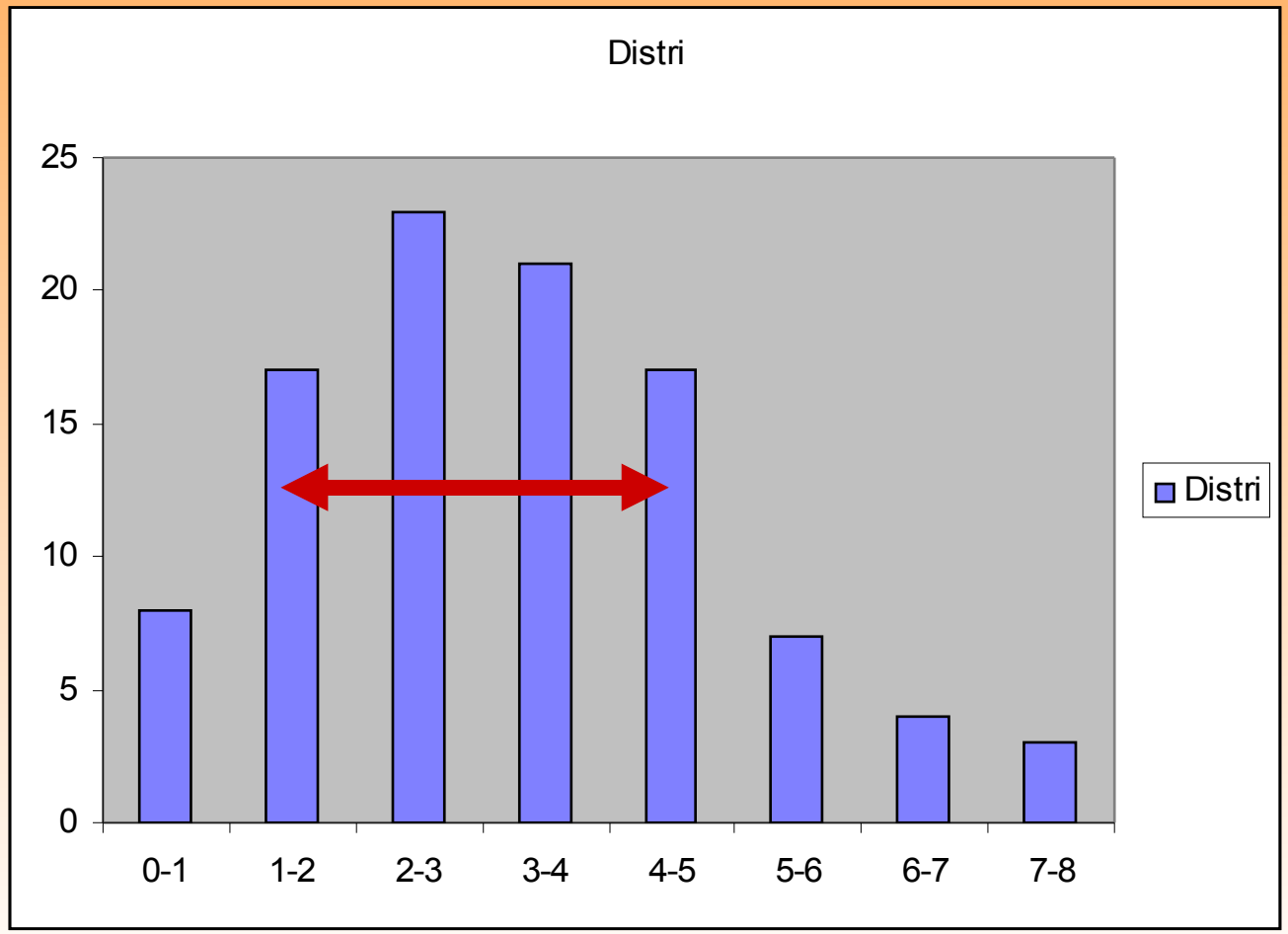
- Médiane = valeur qui définie l'échantillon en 2 partie égale

- Mode = valeur la plus fréquente

1. Valeur centrale

Médiane = 3

Moyenne = 3.4



1. Valeur centrale

- Quelle valeur centrale utiliser?
  - La moyenne est en générale implanté dans les systèmes d'acquisition
  - La médiane rarement implanté, mais l'intérêt est
    - quelle ne tient pas compte des valeurs aberrantes
      - 8.5, 8.3, 8.6,8.8, 8.4 a une médiane de 8.5 et une moyenne de 8.52
      - 8.5, 8.3, 8.6,12.2, 8.4 a une médiane de 8.5 et une moyenne de 9.2
    - quelle tient compte des valeurs sous le seuil de détection
      - <1, 1.2, 1.1, <1, 1.2 a une médiane de 1.1 et une moyenne de????

Le mode: valeur la plus probable

Le mode est la valeur recherchée: c'est la valeur la plus probable.

Pour une distribution normale:

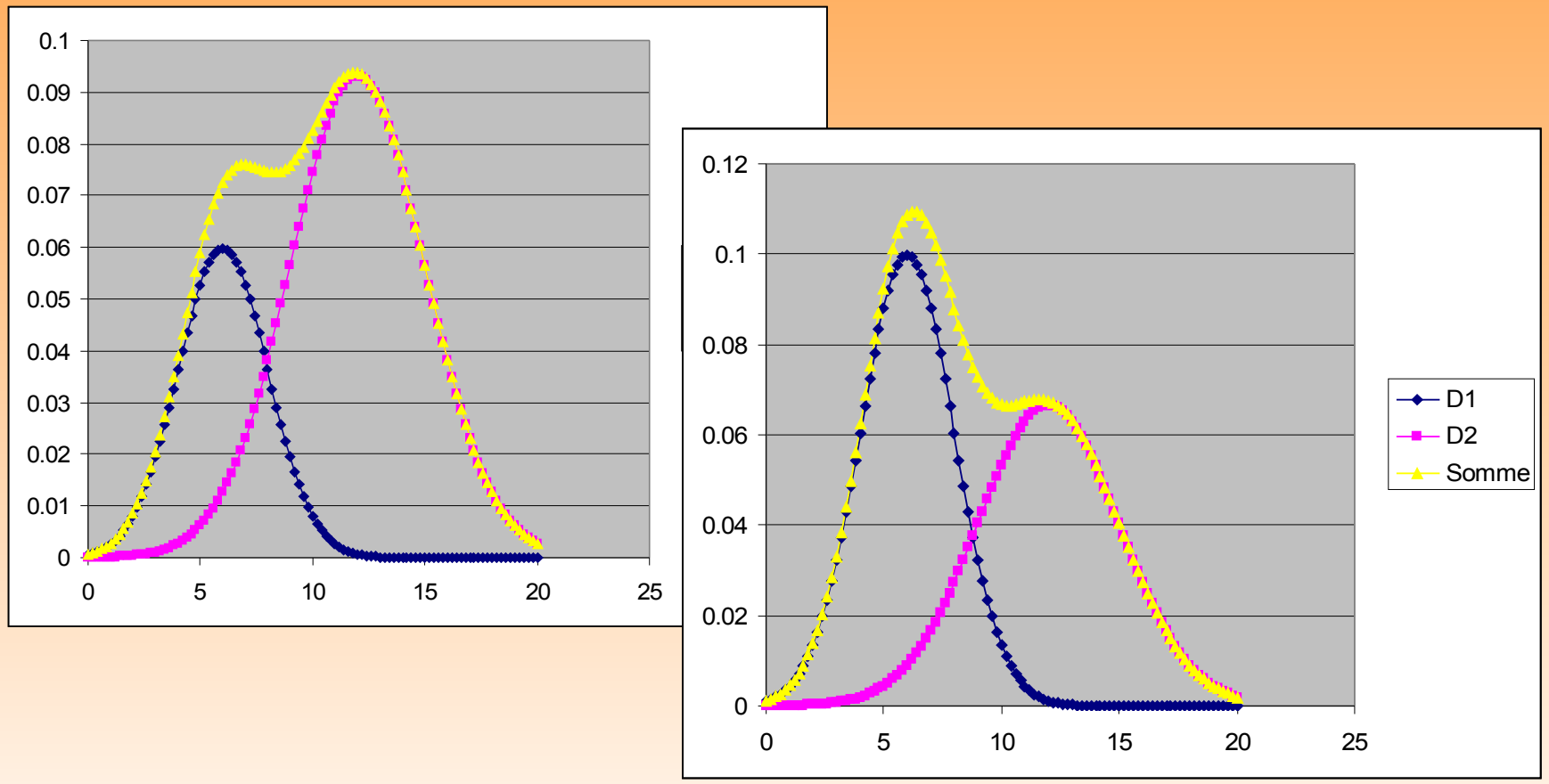
mode = moyenne arithmétique = médiane

Pour une log-normale

mode= moyenne géométrique = médiane

souvent: $M_{géo} <= mode <= M_{arith}$

Distribution bimodale



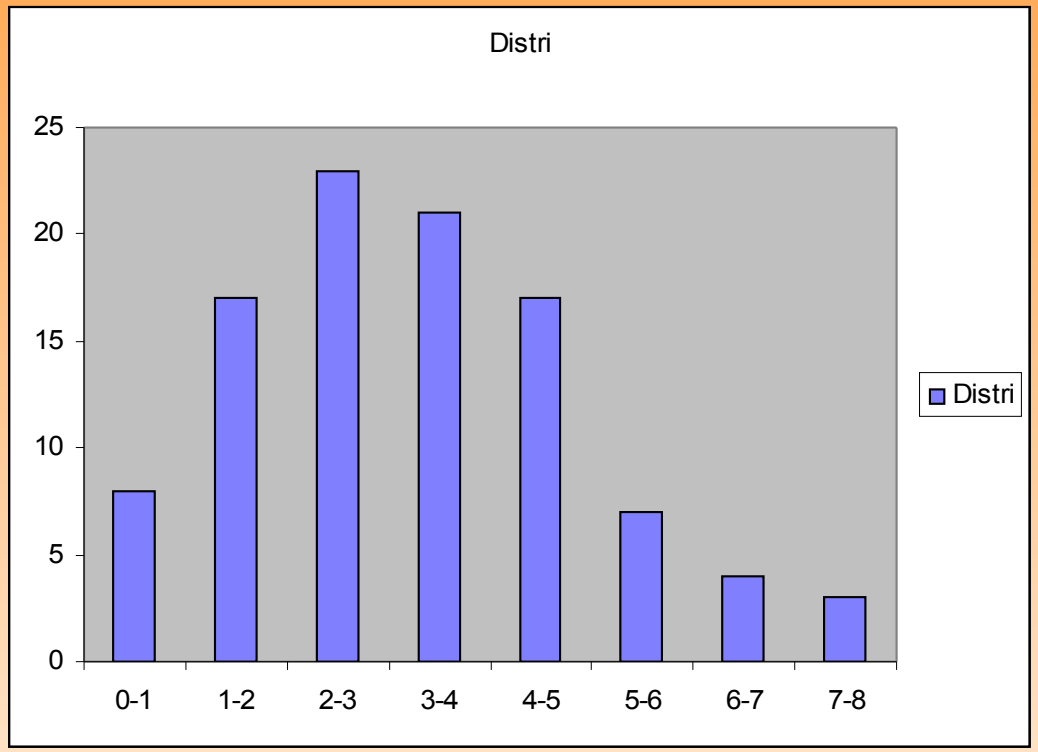
2. Valeur de dispersion

- La dispersion des résultats est lié aux erreurs aléatoires sur les mesures qui sont dues:
  - l'erreur d'échantillonnage : si échantillon hétérogène, le résultat dépend alors de la manière dont on choisit l'échantillon ;
  - l'erreur de préparation : c'est lorsque la préparation de l'échantillon introduit un biais ; l'échantillon s'altère pendant le transport, le stockage ou la manipulation
  - la stabilité de l'appareil : celui-ci peut être sensible aux variations de température, de tension d'alimentation électrique, aux vibrations, aux perturbations électromagnétiques des appareils environnants… ou bien présenter un défaut de conception ou une usure

2. Valeur de dispersion

- Etendue: écart entre la plus petite et la plus grande valeur
- Ecart à la moyenne: $e_i = x_i - \bar{x}$
  - Variance $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$ $\sigma = \sqrt{\sigma^2}$
  - Ecart-type: racine carré de la variance
  - Coefficient de variation $C.V. = \frac{\sigma}{\bar{x}}$

2. Valeur de dispersion

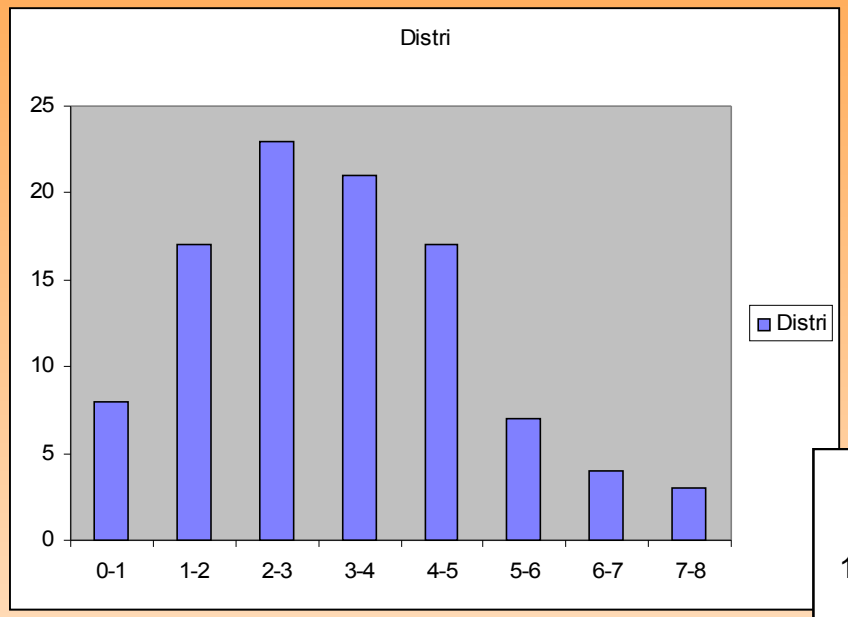


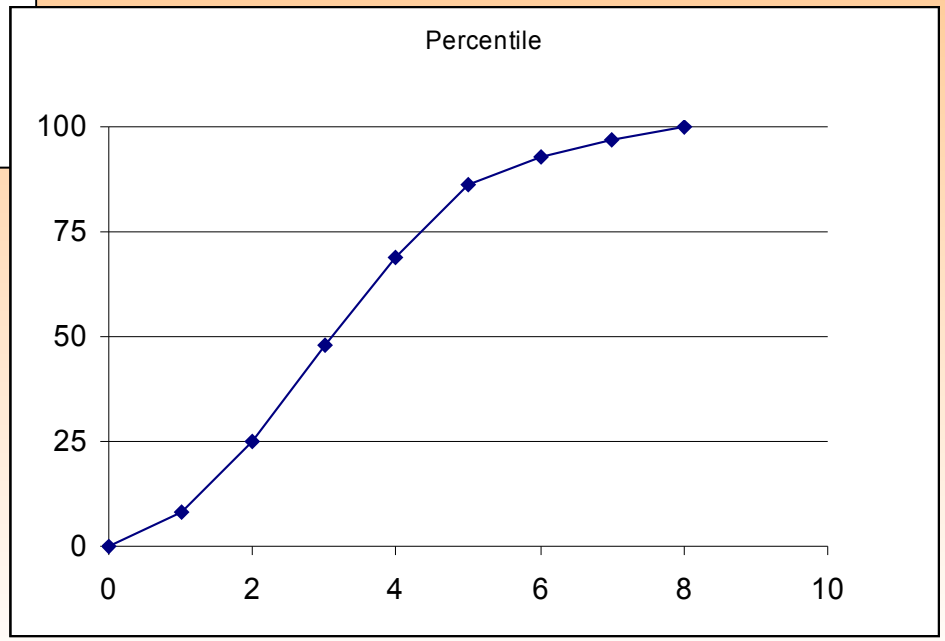
Pour apprécier la qualité de l'estimation, il faut considérer la

Loi de distribution attachée à l'échantillon

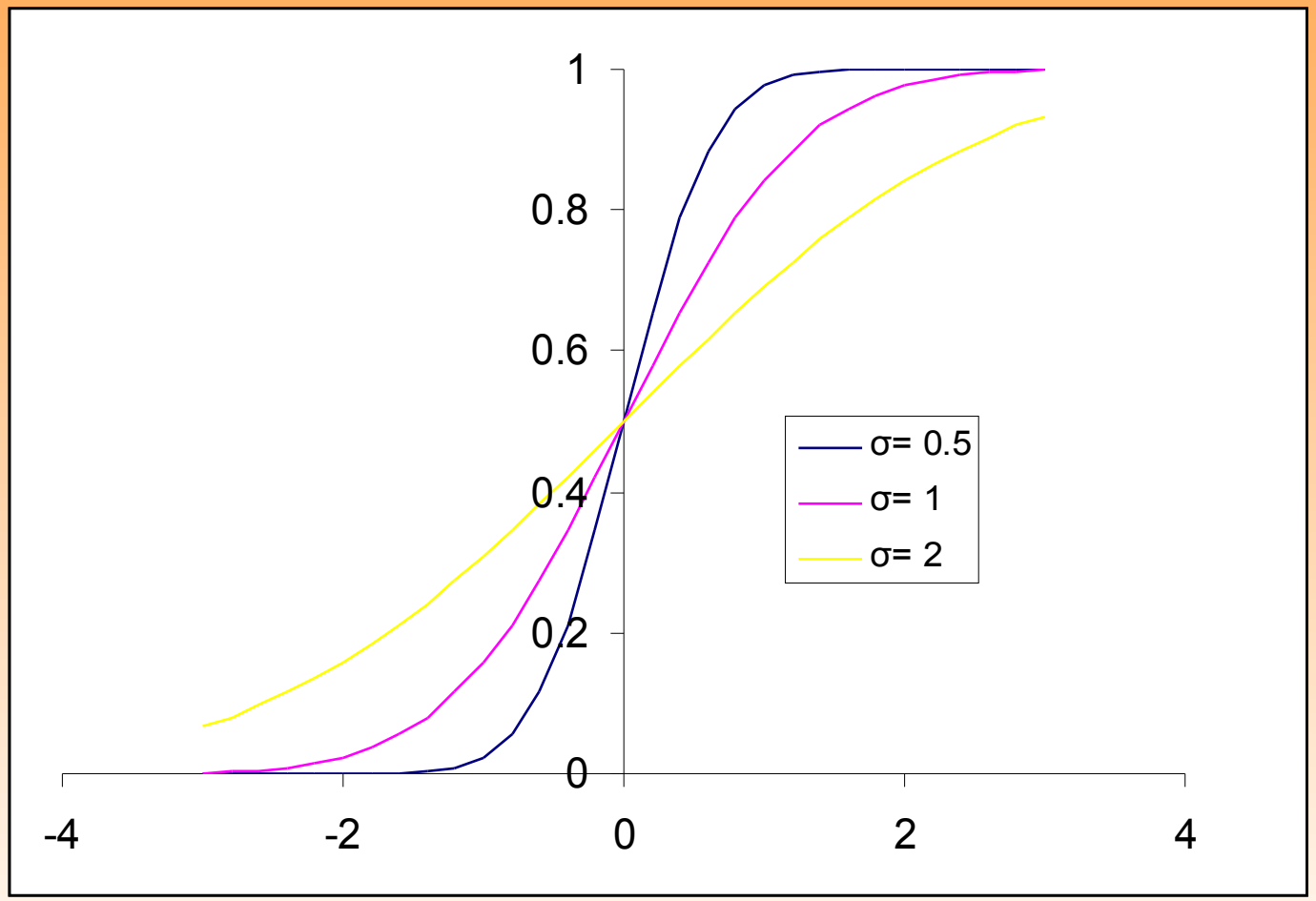
= Modèle mathématique introduit pour calculer les probabilités d'une variable aléatoire continue

Modélisation de la distribution





Percentile loi normale



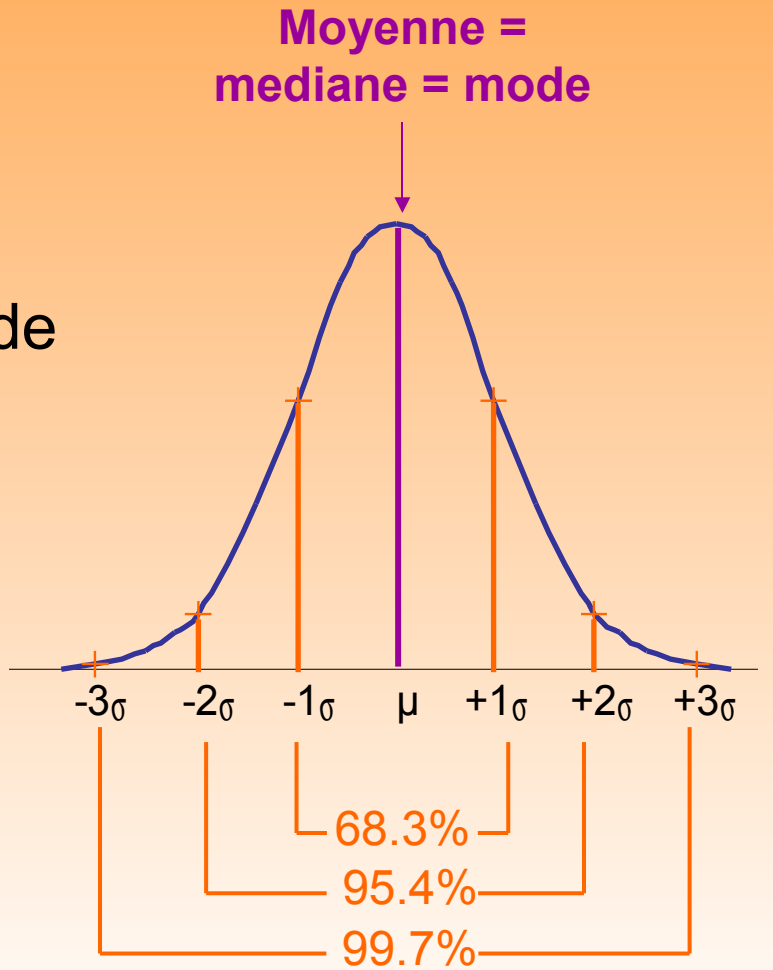
Equation de distribution

Loi Normale

- C'est la plus utilisée
- Elle est définie par la densité de probabilité :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}$$

- Elle suit une distribution Gaussienne

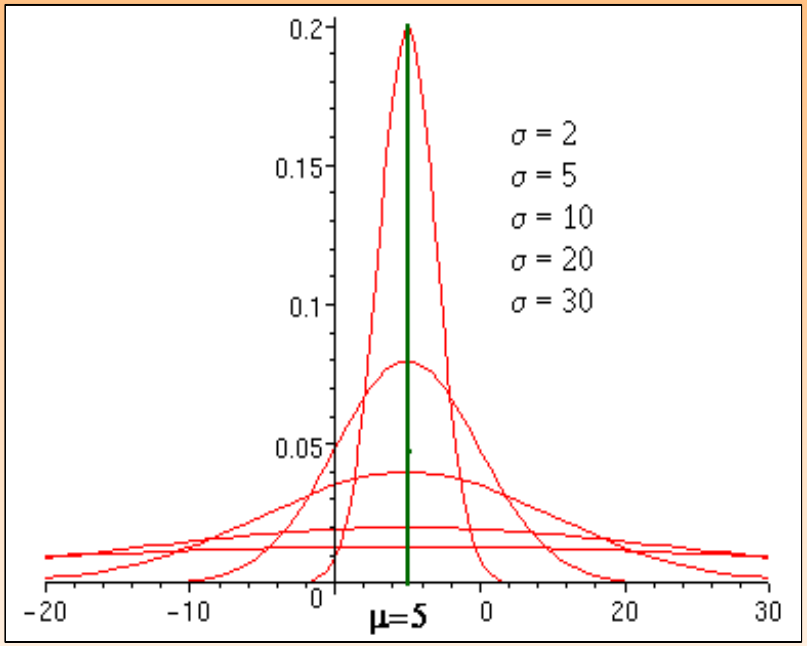


L'écartype

Loi Normale

Aplatissement:

- Surface toujours identique
- Moyenne = médiane = mode
- Ecart-type augmente avec aplatissement et inversement

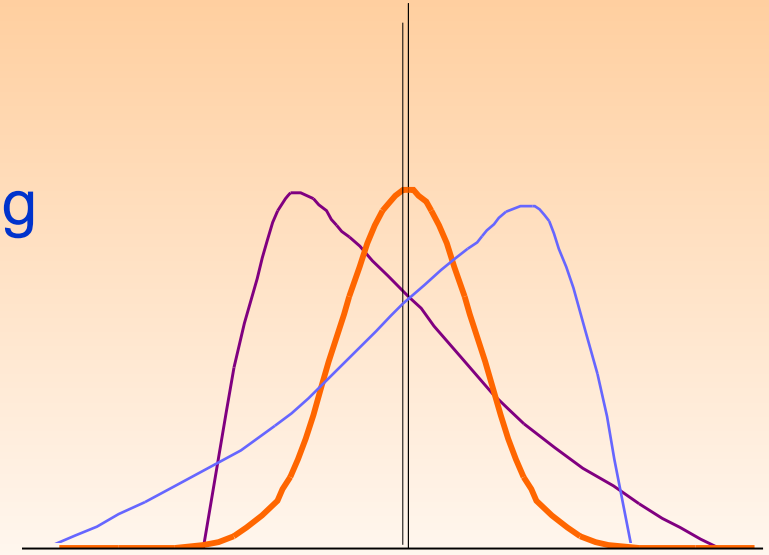


Loi Lognormale

$$f(x) = \frac{1}{x.\sigma\sqrt{2\pi}} e^{\frac{-(\log x - \log \bar{x})^2}{2\sigma^2}}$$

Excentricité:

- Courbe symétrique en échelle log
- Moyenne $\neq$ médiane = mode
- Ecart-type augmente avec étalement de l'excentricité



Estimation de l'écartype

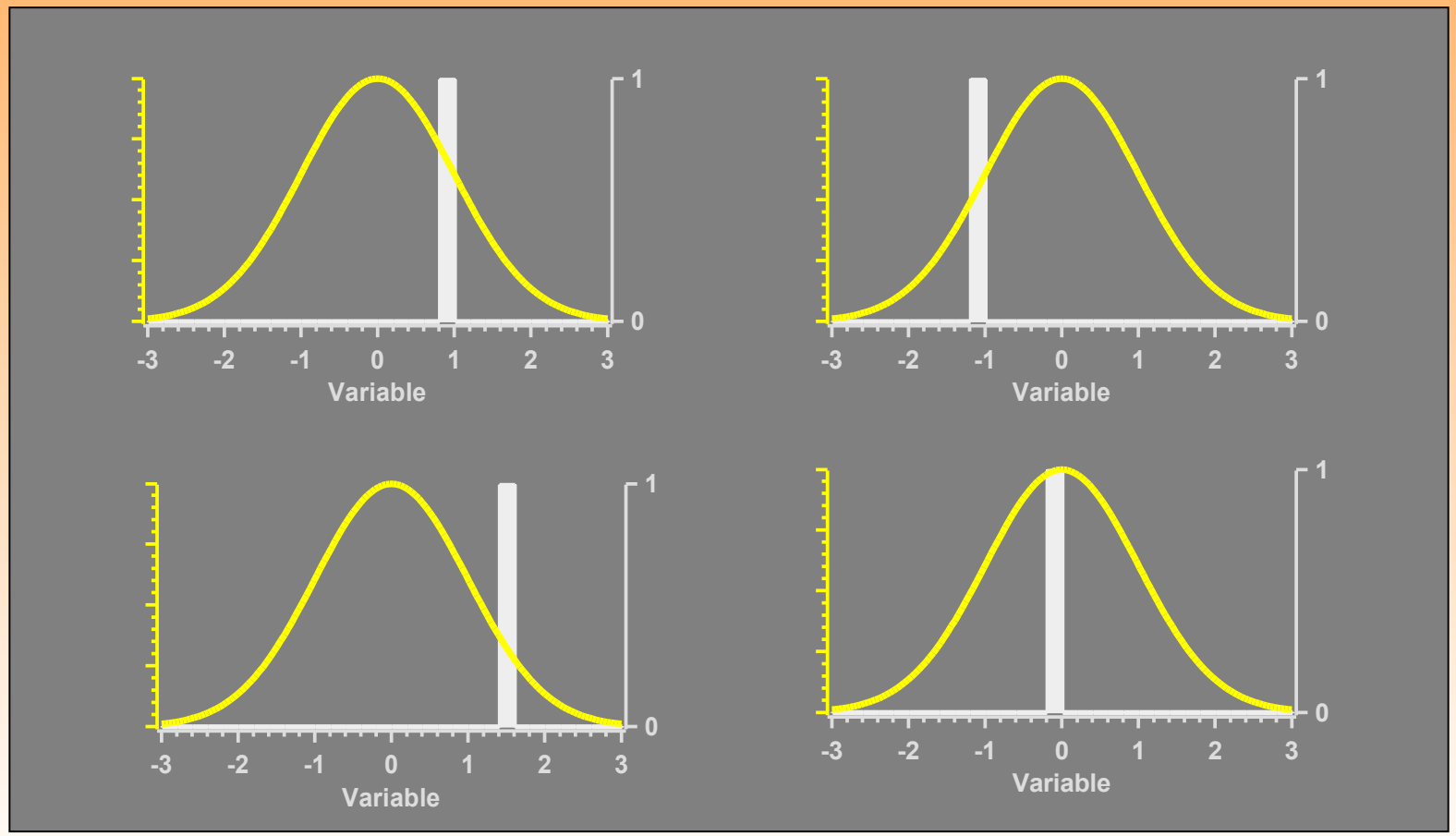
- Si on considère une loi normale, l'estimation de l'erreur aléatoire est définie comme: $\varepsilon = 3.\sigma$

  - le chiffre 3 correspondant à la prise en compte de 99,73 % des mesures

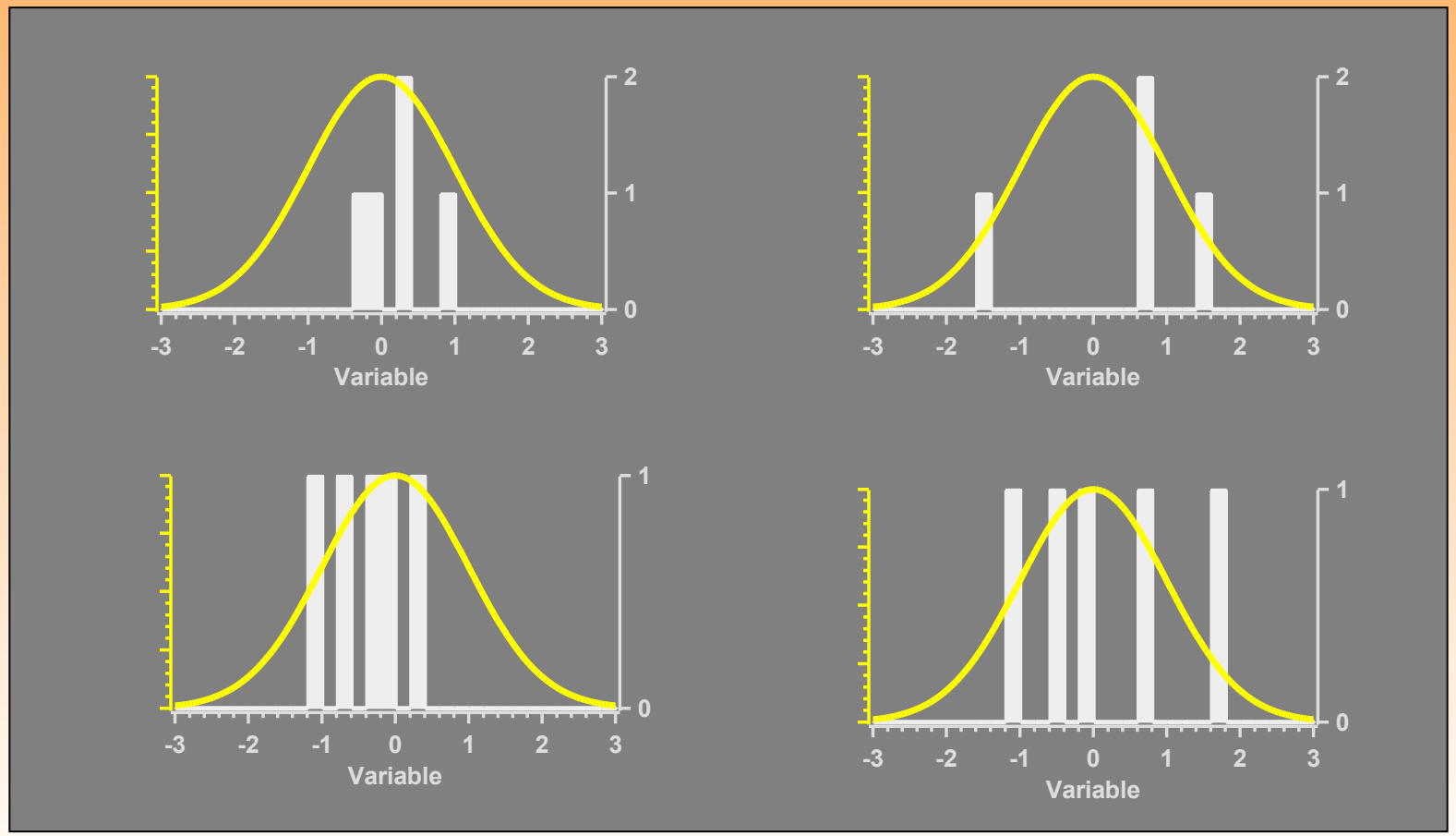
- Si l'on a peu d'échantillons, la formule suivante n'est plus applicable, il faut trouver un estimateur de sigma.

Représentativité de l'échantillonage

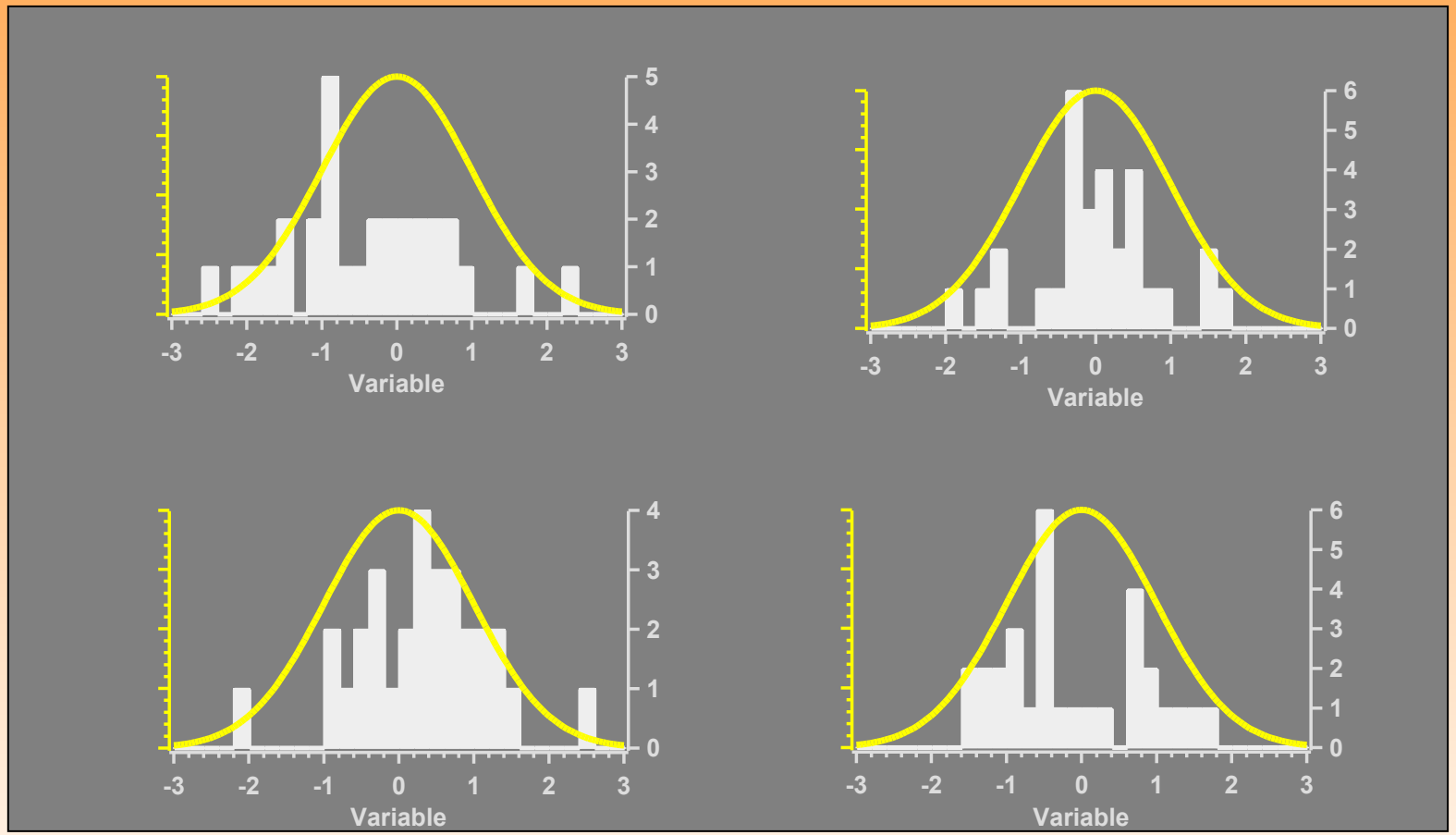
- Effet de la taille de l'échantillon: **n=1**



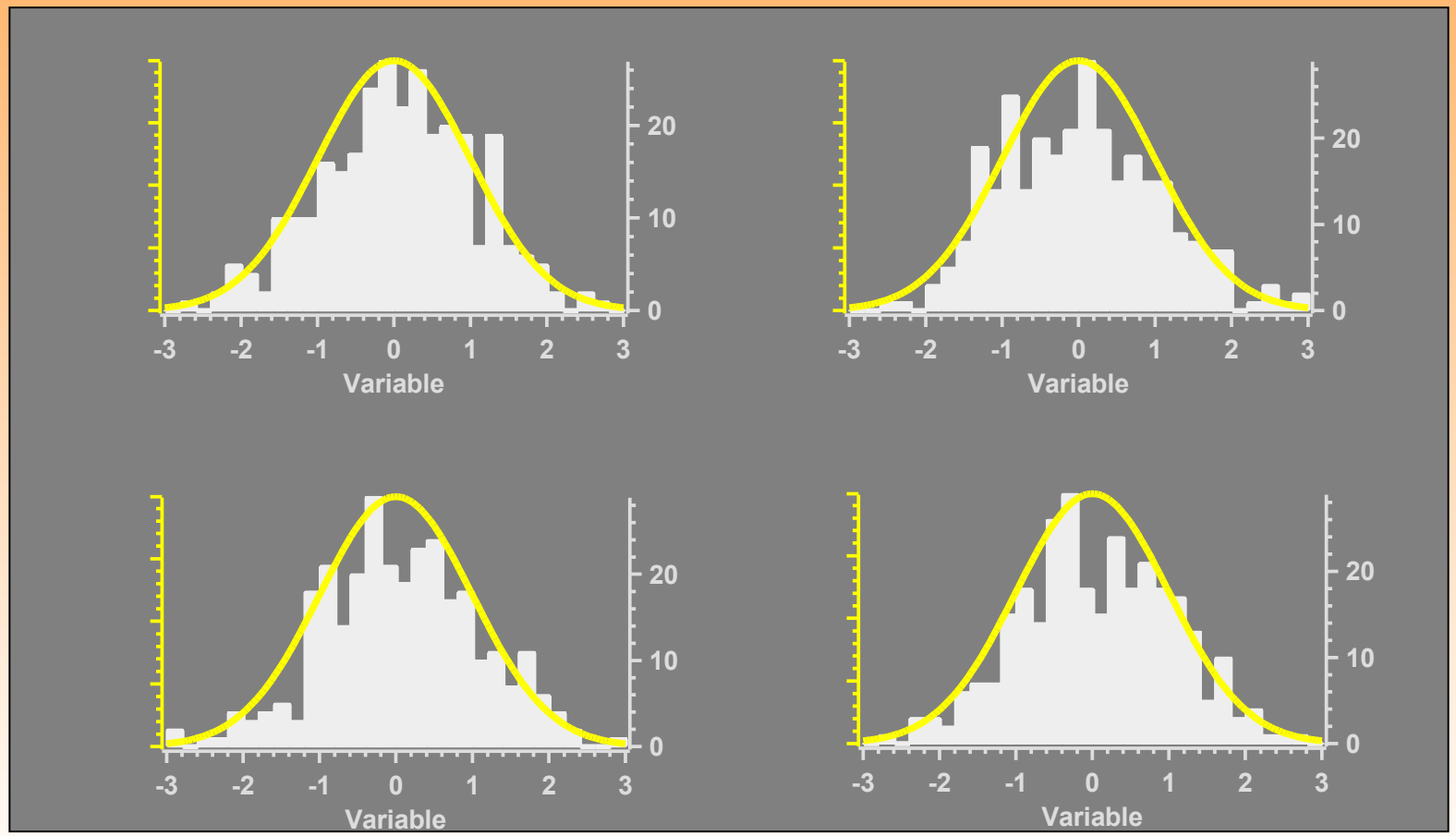
- Effet de la taille de l'échantillon: **n=5**



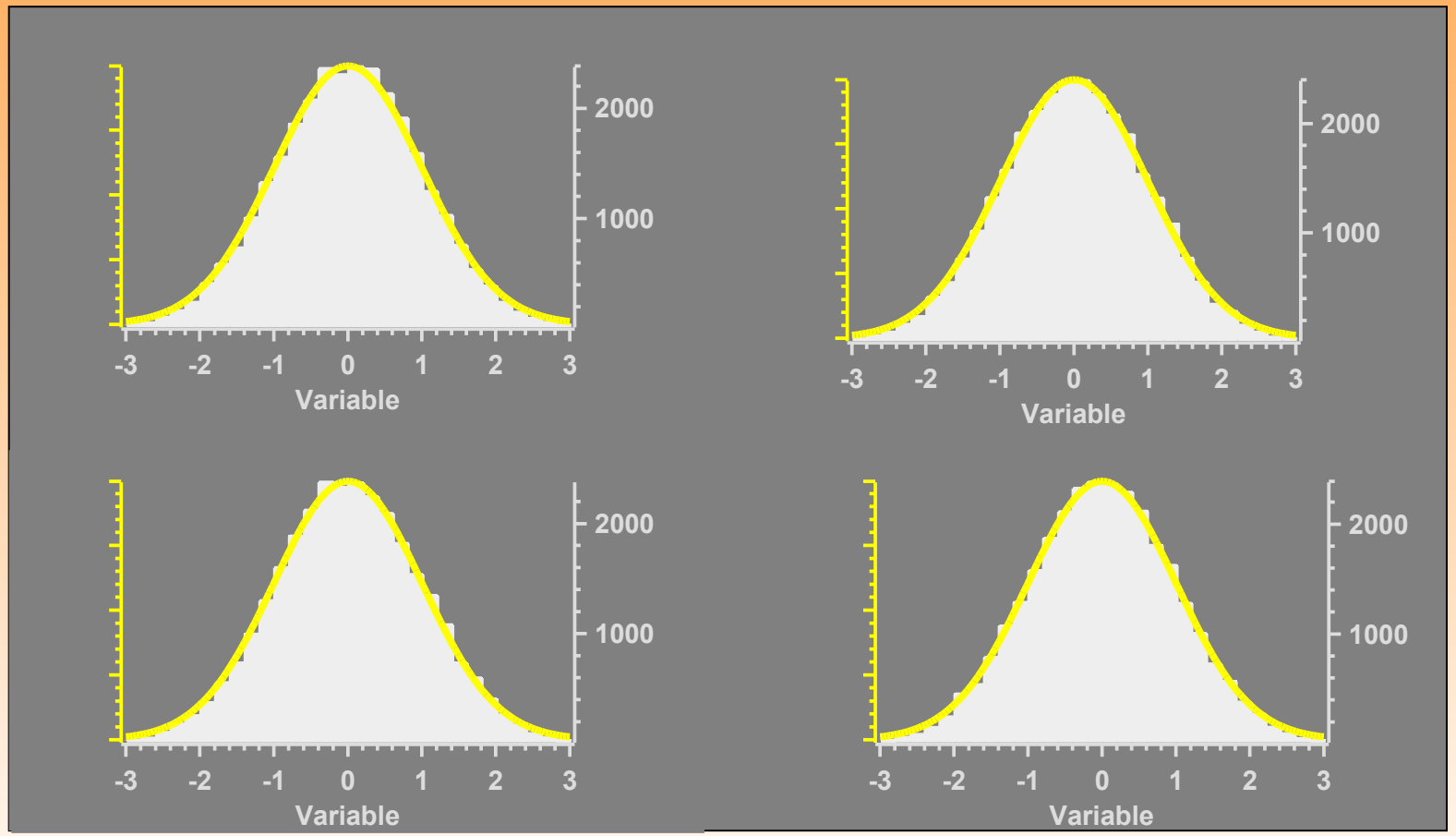
- Effet de la taille de l'échantillon: **n=30**



- Effet de la taille de l'échantillon: **n=300**



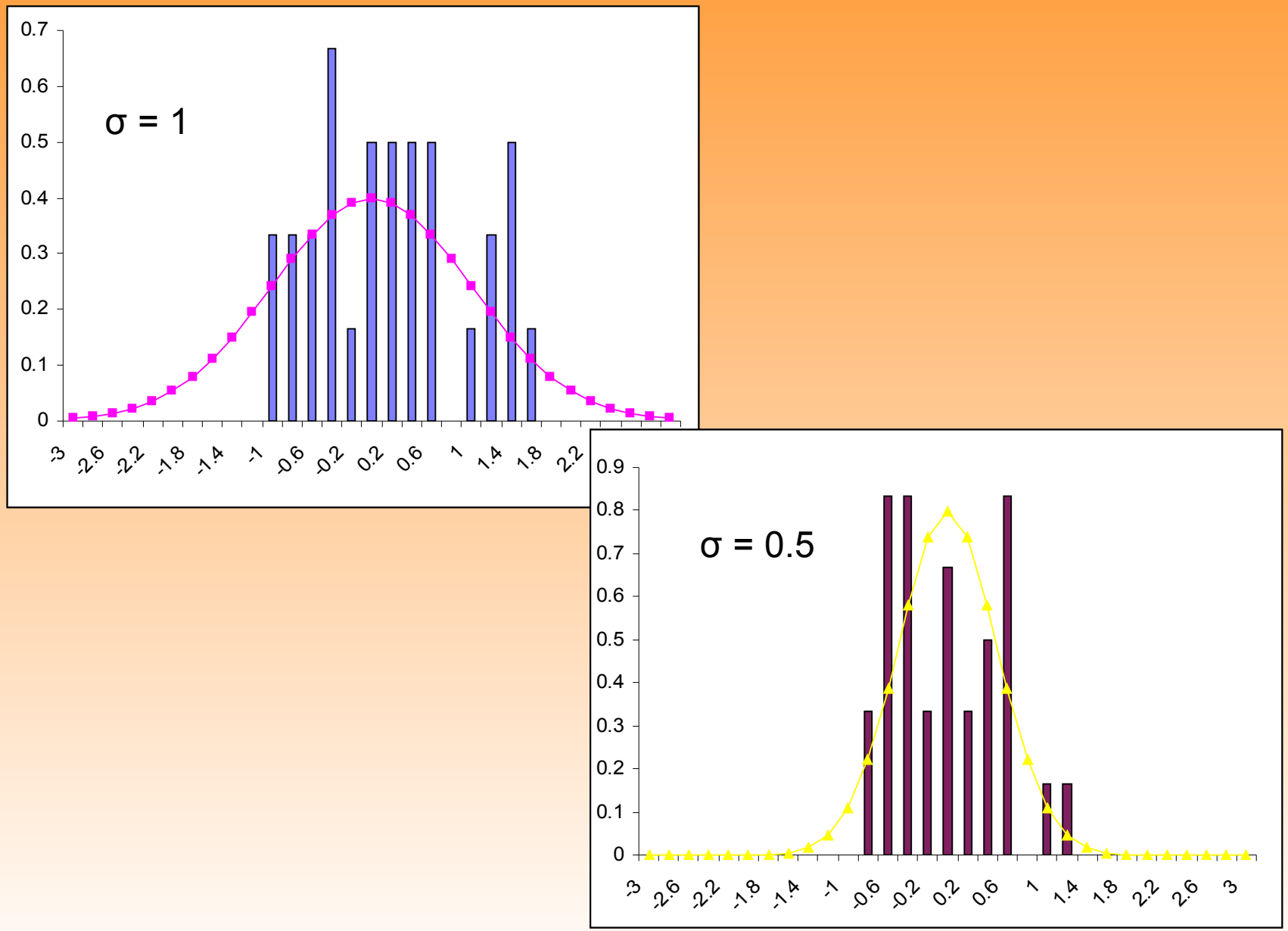
- Effet de la taille de l'échantillon: **n=30000**



Sensibilité à l'écart type

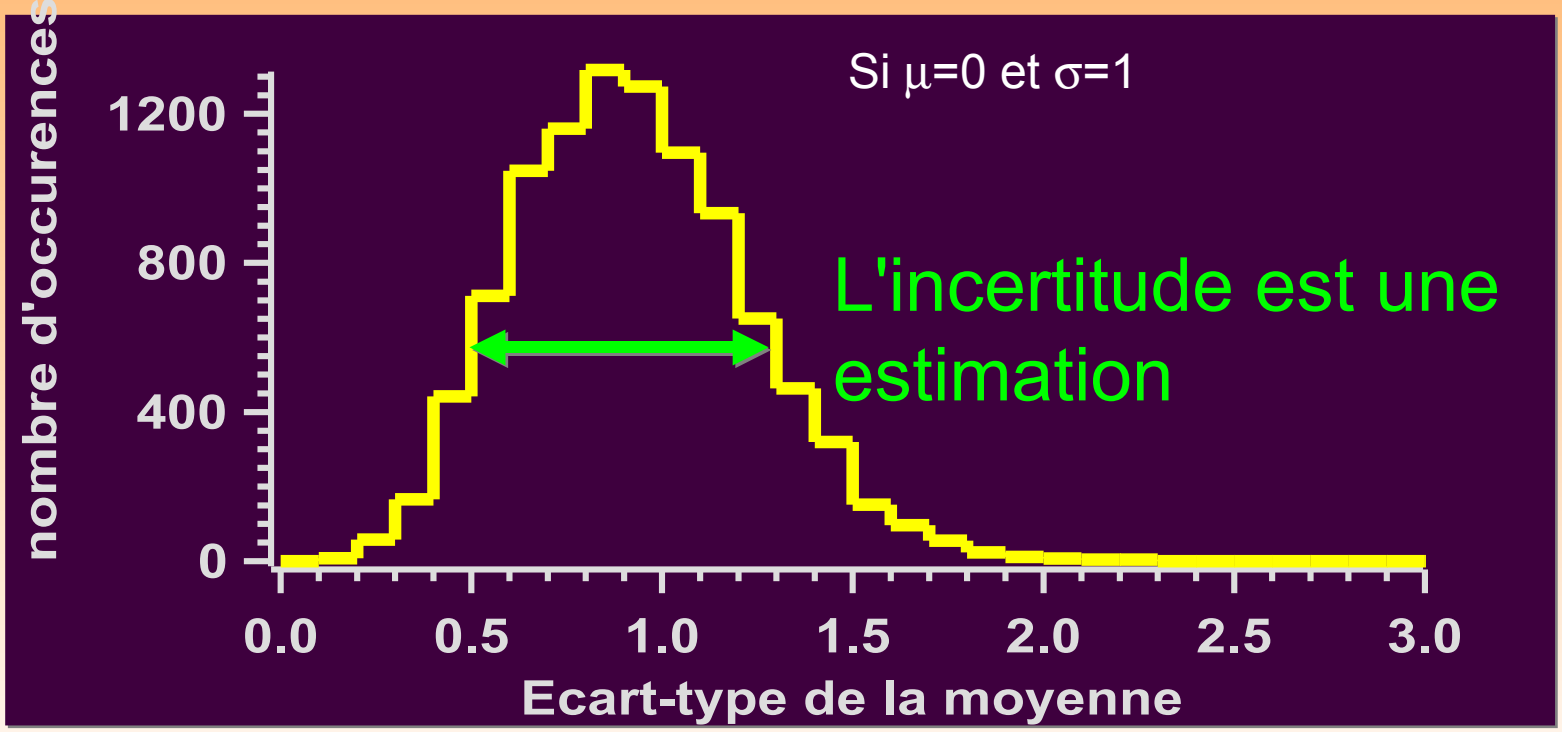
- Plus *n* est élevé, plus la distribution observée est proche d'une distribution modèle.
- σ ne caractérise pas directement l'incertitude de la moyenne pour un *n* quelconque mais plutôt le nombre d'essai à faire pour obtenir une image de la distribution

Influence du σ sur l'échantillonage



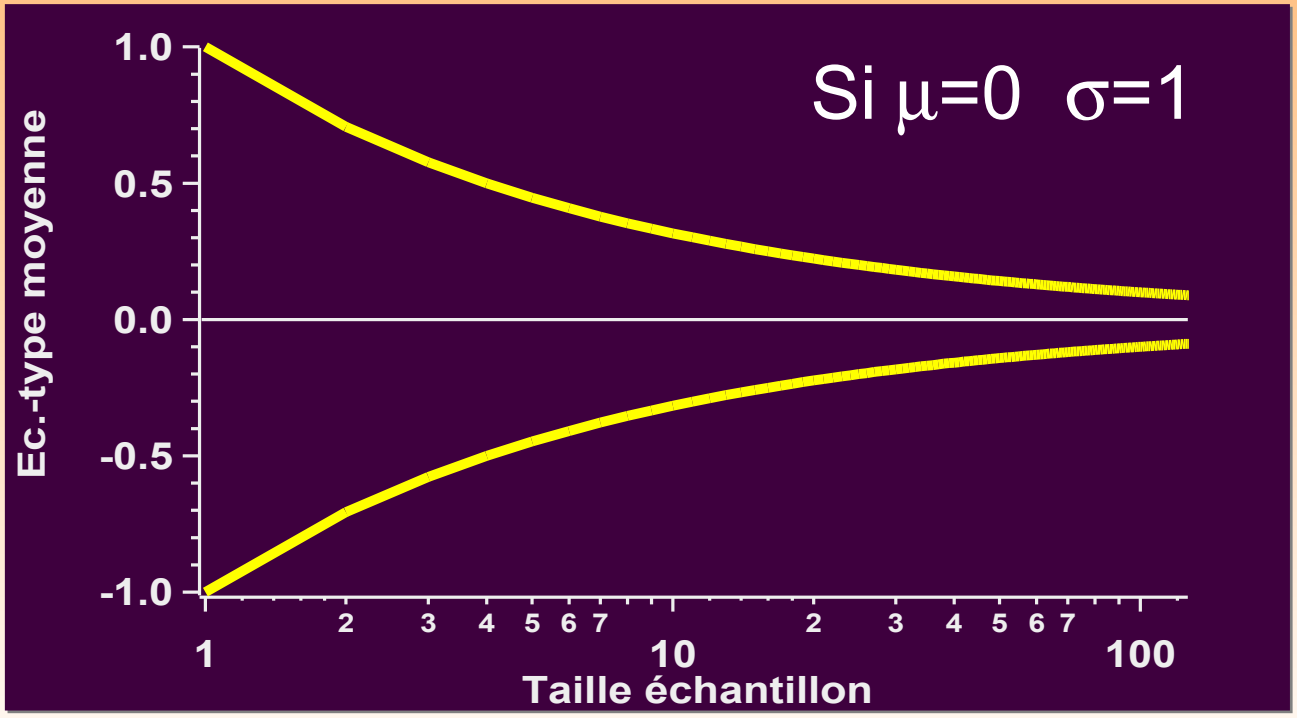
2. Valeur de dispersion

Distribution de l'écart-type expérimental de la moyenne



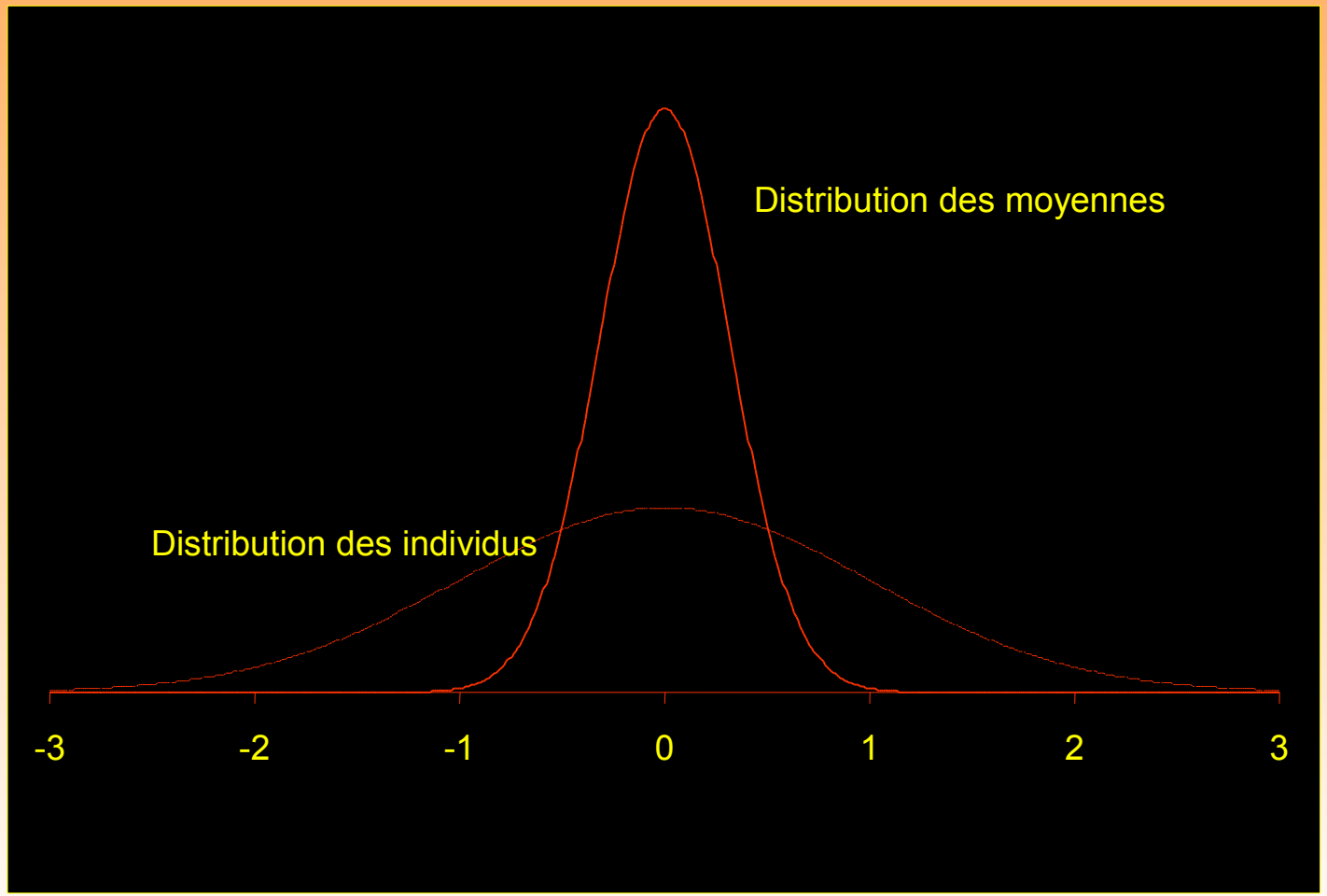
2. Valeur de dispersion

- Ecart-type à la moyenne $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$



$\pm \frac{1}{\sqrt{n}}$

2. Valeur de dispersion



2. Valeur de dispersion

Nbre d'observations *n*	$\sigma(\sigma_x)/\sigma_x$ [%]
2	76
3	52
4	42
5	36
10	24
20	16
30	13
50	10

- Plus n est grand, plus l'incertitude sur l'incertitude diminue
- Pas la peine d'indiquer les incertitudes avec beaucoup de nombres significatifs (Typiquement 2)

2. Valeur de dispersion

- Comme toutes les distributions ne suivent pas la loi normale, au lieu d'utiliser l'estimation ponctuelle de l'erreur comme étant la variance, on est obligé d'utiliser la notion d'intervalle de confiance…

- Risque d'erreur que la valeur estimée est différente de valeur vraie
- Pour quantifier ce risque: Intervalle de confiance autour de la moyenne
  - % de chance de ne pas trouver la valeur vraie dans intervalle de confiance = α
  - (1- α) = niveau de confiance de l'intervalle

2. Valeur de dispersion

- On définie l'intervalle de confiance à partir de la loi de Student (densité de probabilité)

$$\bar{x} - t_{(1-\alpha)/2}^{x}\ \sigma_{\bar{x}} \leq \mu \leq \bar{x} + t_{(1-\alpha)/2}^{x}\ \sigma_{\bar{x}}$$

intervalle de confiance	5 mesures	10 mesures	20 mesures	> 100 mesures
50 %	0.92	0.88	0.86	0.84
90 %	1.48	1.37	1.06	1.29
95 %	2.57	2.22	1.72	1.2
99 %	4.03	3.17	2.53	2.6

2. Valeur de dispersion

- En pratique, l'évaluation de la dispersion statistique se fait par des mesures de :
  - répétabilité : caractérise la dispersion intralaboratoire sur une même série d'essais
  - reproductibilité : caractérise la dispersion intralaboratoires dans des conditions de travail différentes (changement d'opérateurs) ou interlaboratoires pour un même protocole d'analyse

2. Valeur de dispersion

- Exemple:

µg/g				
1.095	1.232	1.135	1.21	0.975
1.18	1.165	1.342	0.956	1.242

Moyenne = 1.153
Variance = 0.014
écart type = 0.119

Intervalle de Confiance

95% $0.89 \leq \mu \leq 1.42$

99% $0.77 \leq \mu \leq 1.53$

3. Erreur systématique

- L'erreur systématique comprend des phénomènes comme les erreurs d'échantillonnage, de préparation, d’étalonnage

- Ces problèmes peuvent introduire une dispersion statistique (cf. ci-dessus) ou bien un décalage des résultats si l'erreur commise est toujours la même.

3. Erreur systématique

Matériaux de référence

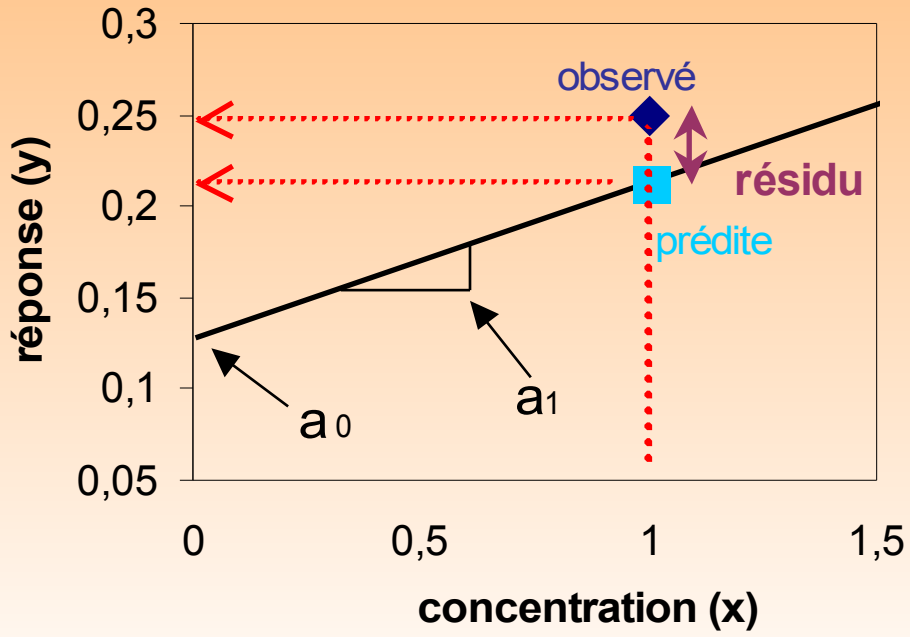
- **MRC : Matériaux de référence certifiés**
- **MRI : Matériaux de référence internes**
- **Solutions étalons de vérification**
- **Ajouts dosés**

- **BNM : Bureau Nationale de Métrologie**
- **NIST : National Institute of Standards and Technology (USA)**
- **BCR : Bureau communautaire de référence (Bruxelles)**
- **AIEA : Agence Internationale pour l'Energie Atomique (Vienne)**
- **NRC CNRC: Bureau de certification canadien**

3. Erreur systématique

Etalonnage

$$y_i = a_0 + a_1 x_i$$



3. Erreur systématique

Intervalle de confiance

$$a_0 - t_{1-a/2,\,n} \cdot \sigma_{a0} \leq a_0 \leq a_0 + t_{1-a/2,\,n} \cdot \sigma_{a0}$$

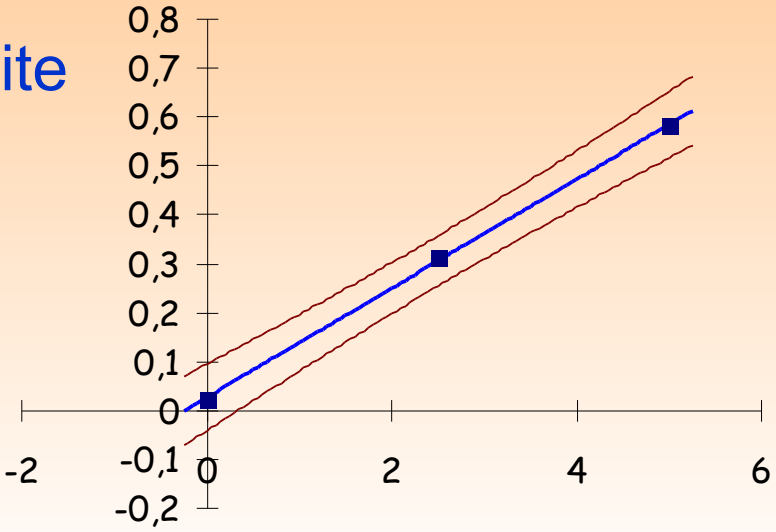
$$a_1 - t_{1-a/2,\,n} \cdot \sigma_{a1} \leq a_1 \leq a_1 + t_{1-a/2,\,n} \cdot \sigma_{a1}$$

Intervalle de confiance de la droite prédite :

valeurs supérieures et inférieures:

$$\hat{y}_+ = a_0 + a_1 x + t_{1-a/2}\, \sigma_{\hat{y}}$$

$$\hat{y}_- = a_0 + a_1 x - t_{1-a/2}\, \sigma_{\hat{y}}$$



Régression linéaire

$Y = a.X + b$

$a = Ymoy - b.Xmoy$

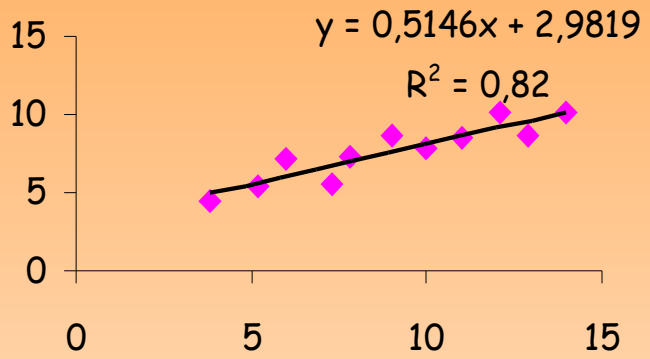
$$b = \frac{\sum XiYi - n.Xmoy.Ymoy}{\sum Xi^2 - n.Xmoy^2}$$

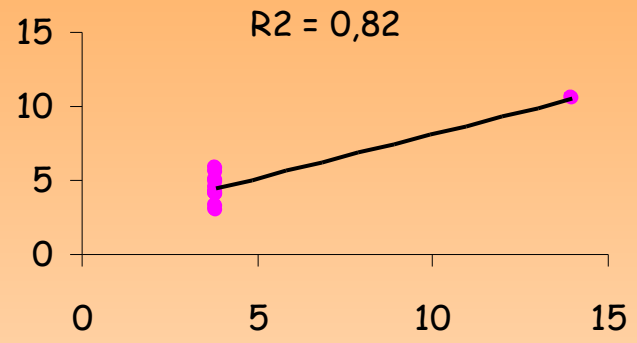
$$r = \frac{\sum (Xi - Xmoy).(Yi - Ymoy)}{\sqrt{\sum (Xi - Xmoy)^2.\sum (Yi - Ymoy)^2}}$$

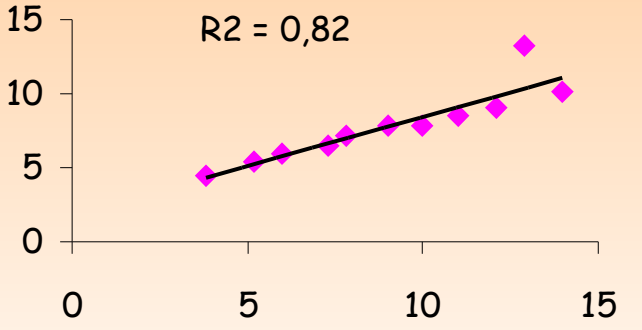
$$\sigma = \sqrt{\frac{\sum (Yi - a.Xi - b)^2}{n - 2}} = \sqrt{\frac{1 - r^2}{n - 2}.\sum (Yi - Ymoy)^2}$$

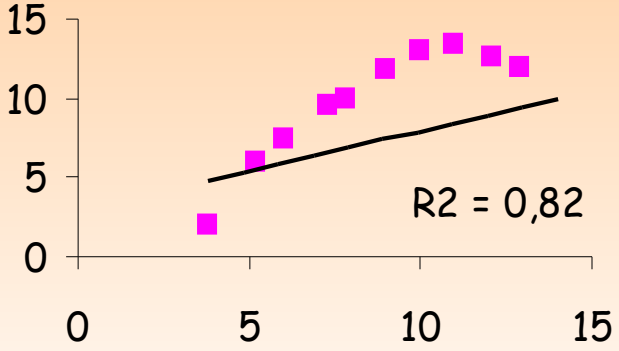
$$\sigma a^2 = \left(\frac{1}{n} + \frac{Xmoy^2}{\sum (Xi - Xmoy)^2}\right).\sigma^2 \quad \text{et} \quad \sigma b^2 = \frac{\sigma^2}{\sum (Xi - Xmoy)^2}$$

3. Erreur systématique









Expression du résultat

Moyenne ou médiane

$$\mu = \bar{x} + \varepsilon$$

$t_{(1-\alpha/2,n)} \cdot \sigma_x$ **+** **%erreur relative**

Qualification et validation

- Qualifier appareil, opérateur, méthode.
  - SST (System Sustainability Test).
- Vérifier la conformité en cours d'analyse (QC):
  - CRM (MRC)
  - Etalon "maison"
- Une méthode doit être validée dans sa totalité y compris le prélèvement.

FIN