Dispersion statistique

Méthodes statistiques en métrologie

- Des méthodes statistiques sont utilisées en métrologie essentiellement pour évaluer la meilleure estimation et la dispersion des **erreurs aléatoires** (donc l'incertitude) à partir de séries de mesure.
- L'erreur aléatoire est le résultat d'un mesurage moins la moyenne d'un nombre infini de mesurages du même mesurande (grandeur physique) effectués dans des conditions de répétabilités (tout reste identique).
- Comme on ne peut faire qu'un nombre limité (fini) de mesurages, il est seulement possible de déterminer une **estimation** de la moyenne et de l'erreur aléatoire.
- Cela veut dire que non seulement la moyenne mais aussi l'erreur aléatoire ont elles-mêmes une incertitude associée.

Etude de séries de mesures d'une variable

Sommaire

- Estimation de l'espérance: moyenne, médiane, …
- Histogramme
- Effectifs cumulés et fonction de répartition (%)
- Quelques premiers éléments sur la représentation graphique des mesures
- Lois de probabilité
- La distribution normale
- Critères de normalité
- Quelques autres types de distributions utiles en métrologie et ingénierie

Meilleure estimation (**espérance**) d'une mesure

Quantités possibles pour avoir la meilleure estimation d'une mesure:

1. Moyenne arithmétique $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Médiane: la valeur qui départage le 50 %
3. Moyenne entre maximum et minimum
4. Autres types de moyenne (voir section 5.2 du polycopié)
   - Moyenne géométrique
   - Moyenne harmonique
   - Moyenne glissante

Quantification des écarts

- Ecart maximum par rapport à la meilleure estimation

- Ecart type empirique

$$\sigma_e = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

- l'**écart type empirique corrigé** σ_c d'une série finie de n mesures:

$$\sigma_c = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

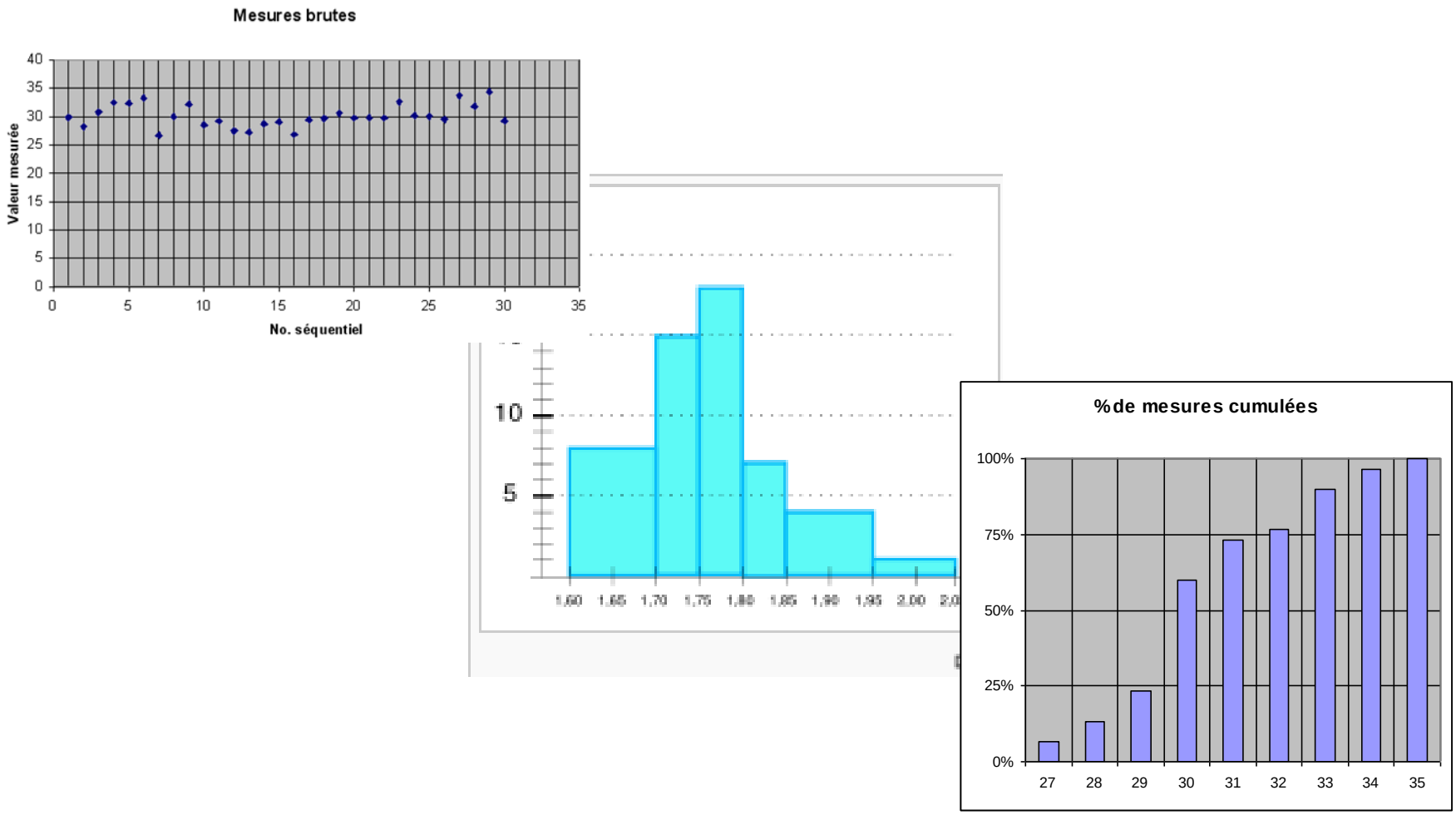
Pourquoi $(n-1)$?

- Le fait que l'estimateur de la variance doive être divisé par (n-1) - et donc dans un certain sens moins précis - pour être sans biais provient du fait que l'estimation de la variance implique l'estimation d'un paramètre en plus, la moyenne.
- Cette correction tient compte donc du fait que l'estimation de la moyenne (nécessaire pour calculer la variance) induit une incertitude supplémentaire.
- En effet si l'on suppose que la moyenne est parfaitement connue, l'estimateur

$$\sigma_e = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

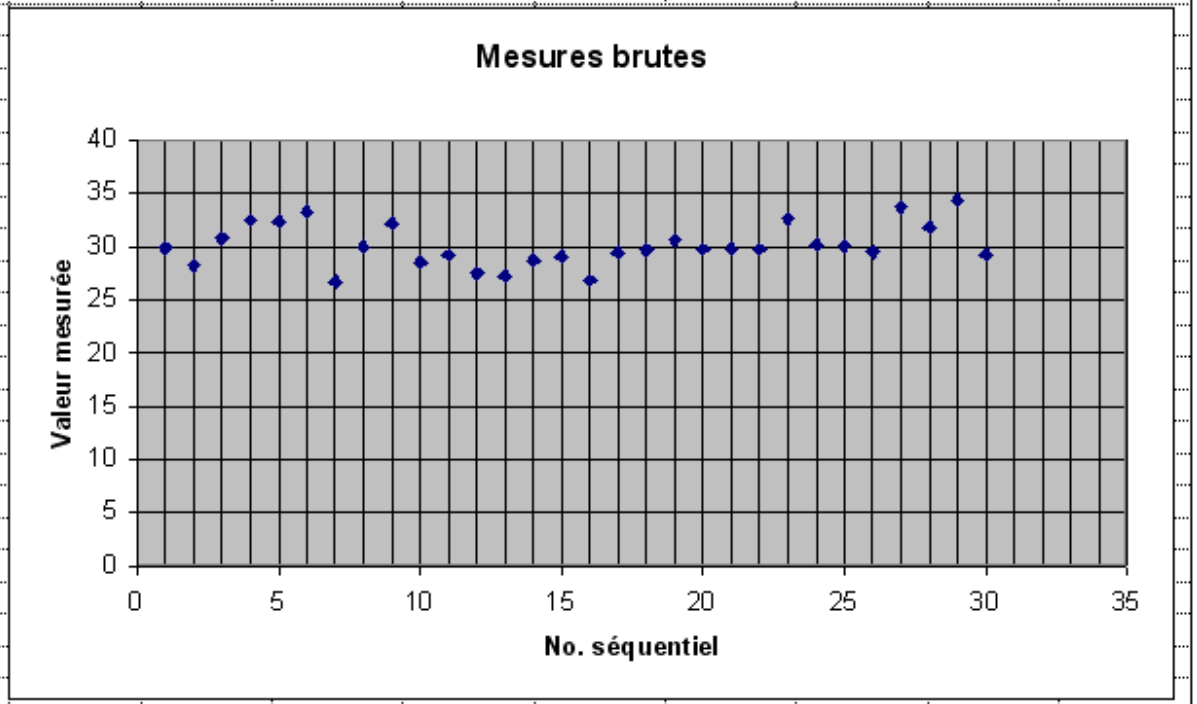
doit être utilisé.

Analyse de séries de mesures

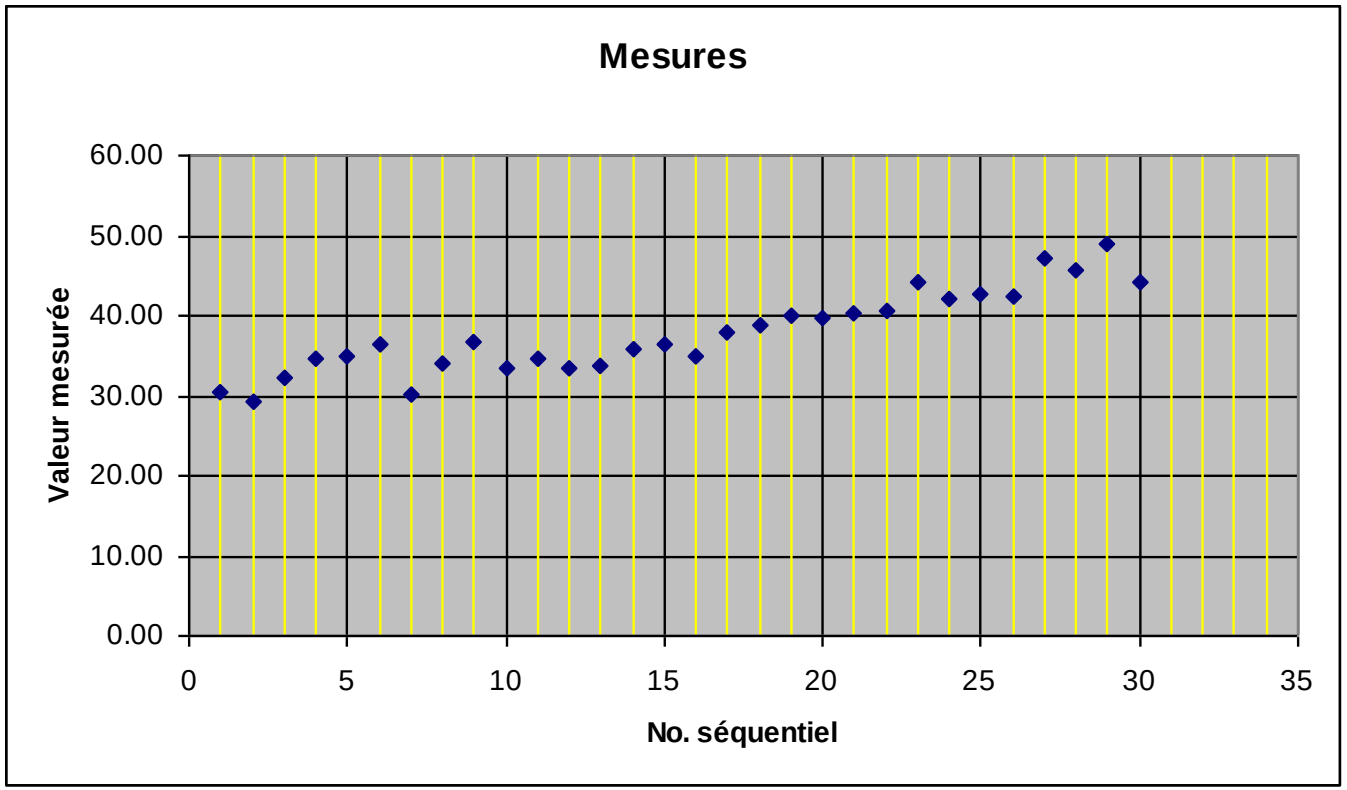


Représentation graphique simple d'une série de mesures

No.	Mesure
1	29.85
2	28.21
3	30.77
4	32.5
5	32.37
6	33.27
7	26.69
8	29.96
9	32.19
10	28.53
11	29.2
12	27.52
13	27.26
14	28.72
15	29.06
16	26.8
17	29.4
18	29.68
19	30.58
20	29.74
21	29.81
22	29.74
23	32.61
24	30.21
25	30.04
26	29.5
27	33.67
28	31.81
29	34.34
30	29.26

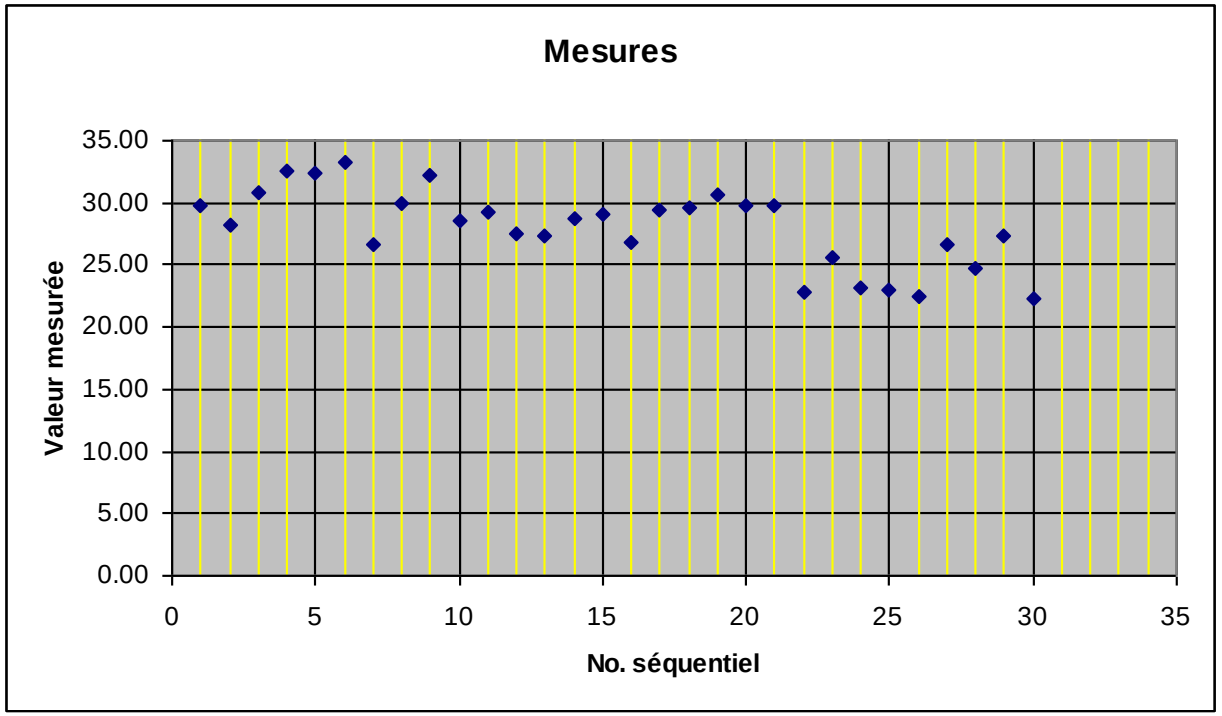


- Ce type de graphique simple peut (dans certains cas) permettre d'identifier d'éventuels facteurs systématiques …
(exemple: ici il apparaît qu'il y a une dérive des mesures dans le temps)



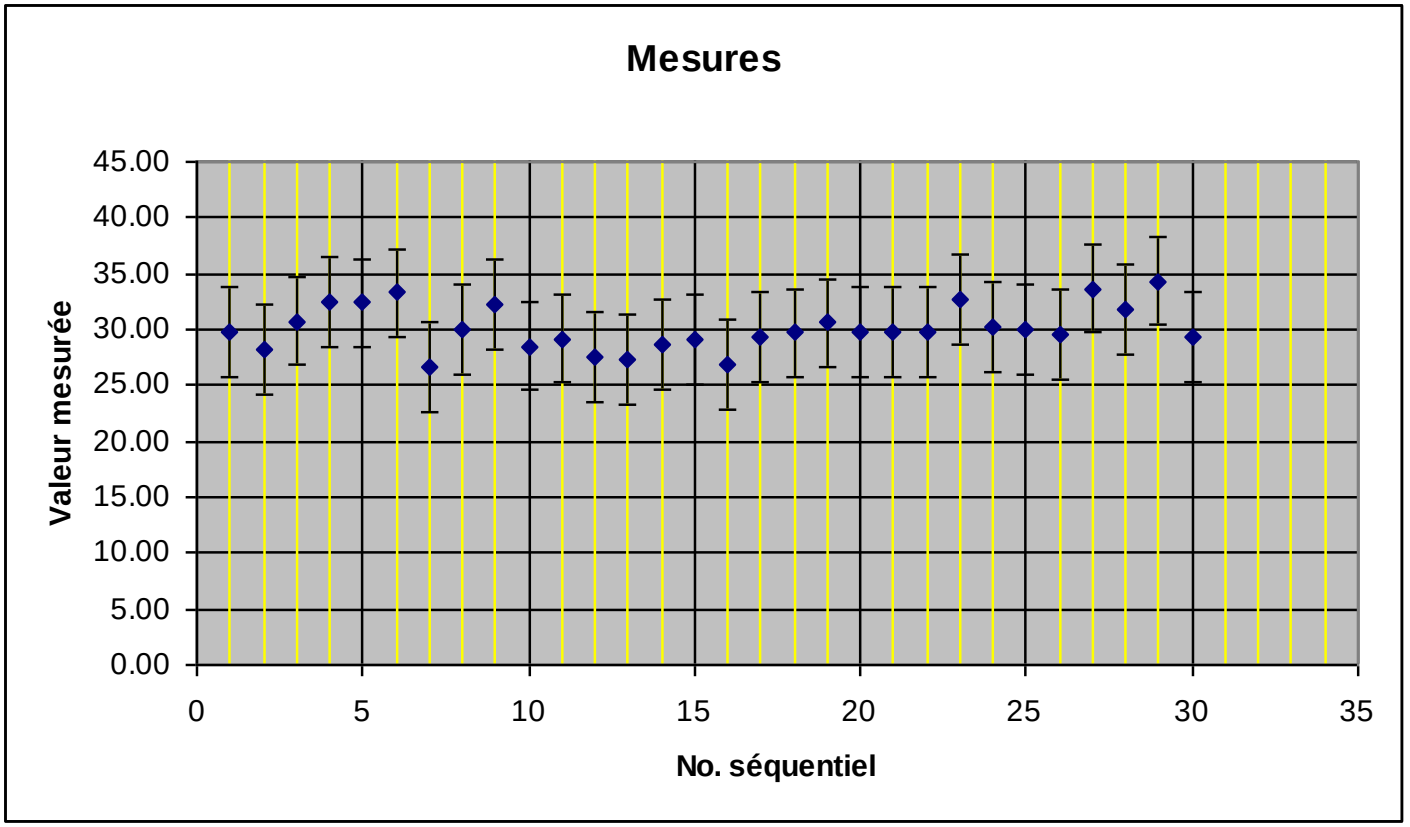
(autre exemple:

il semble ici que quelque chose d'abrupt se soit passé entre les mesures 6 et 7 ainsi qu'après la mesure no. 22 …)



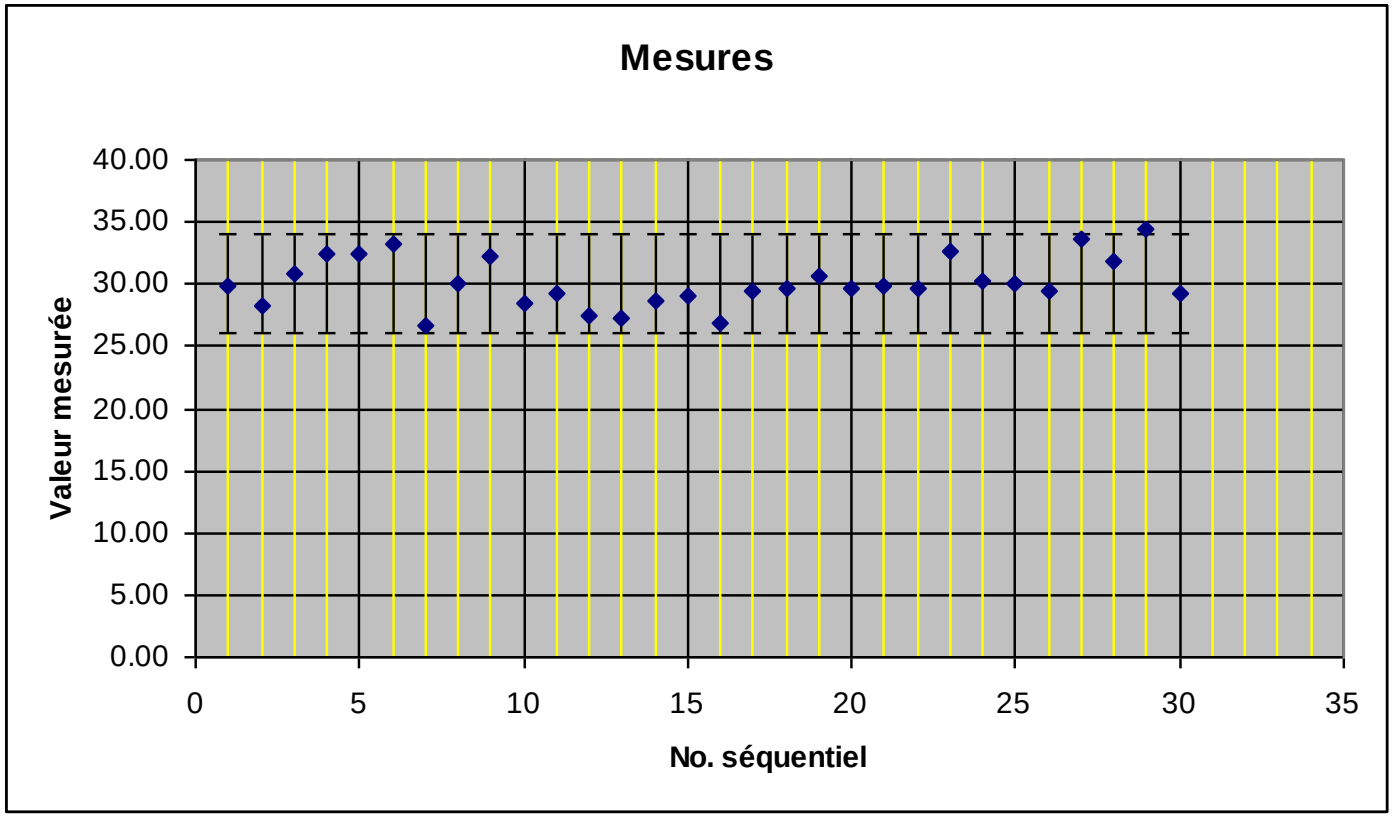
Graphique des mesures avec barres d'erreur indiquant l'incertitude de chaque mesure

Moyenne =	**30.11**	
Ecart-type =	**1.99**	
Incertitude estimée =	**3.99**	**(= 2 * sigma)**



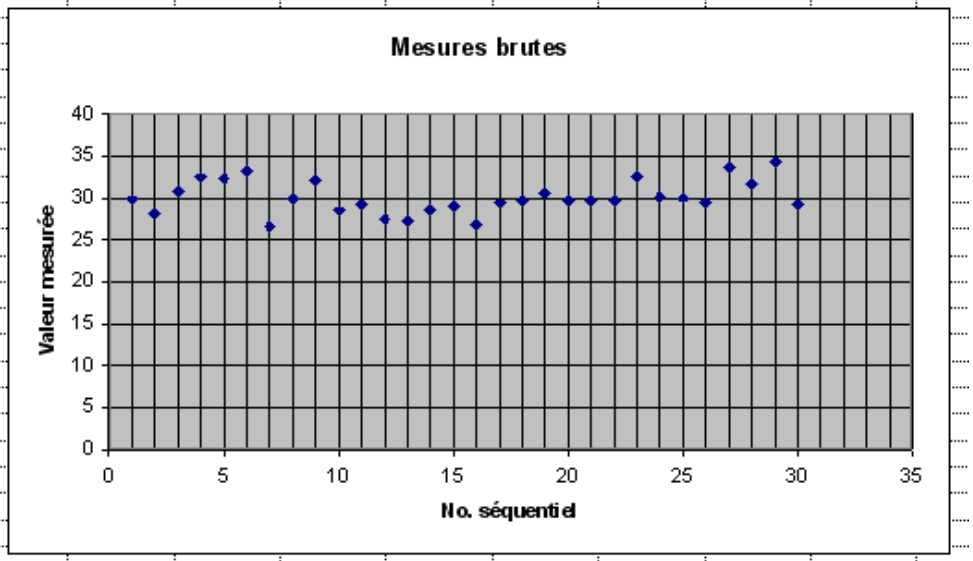
Graphique des mesures avec barre d'erreur indiquant l'estimation de l'incertitude basée sur l'écart-type de la série

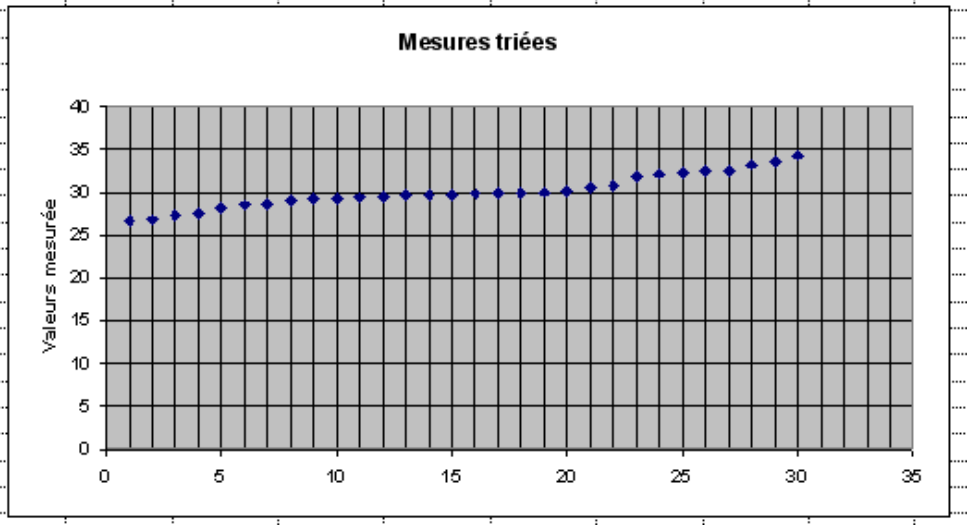
Moyenne =	**30.11**	
Ecart-type =	**1.99**	
Incertitude estimée =	**3.99**	**(= 2 * sigma)**



Triage des mesures

No.	Mesures	Mesures triées
1	29.85	26.69
2	28.21	26.8
3	30.77	27.26
4	32.5	27.52
5	32.37	28.21
6	33.27	28.53
7	26.69	28.72
8	29.96	29.06
9	32.19	29.2
10	28.53	29.26
11	29.2	29.4
12	27.52	29.5
13	27.26	29.68
14	28.72	29.74
15	29.06	29.74
16	26.8	29.81
17	29.4	29.85
18	29.68	29.96
19	30.58	30.04
20	29.74	30.21
21	29.81	30.58
22	29.74	30.77
23	32.61	31.81
24	30.21	32.19
25	30.04	32.37
26	29.5	32.5
27	33.67	32.61
28	31.81	33.27
29	34.34	33.67
30	29.26	34.34



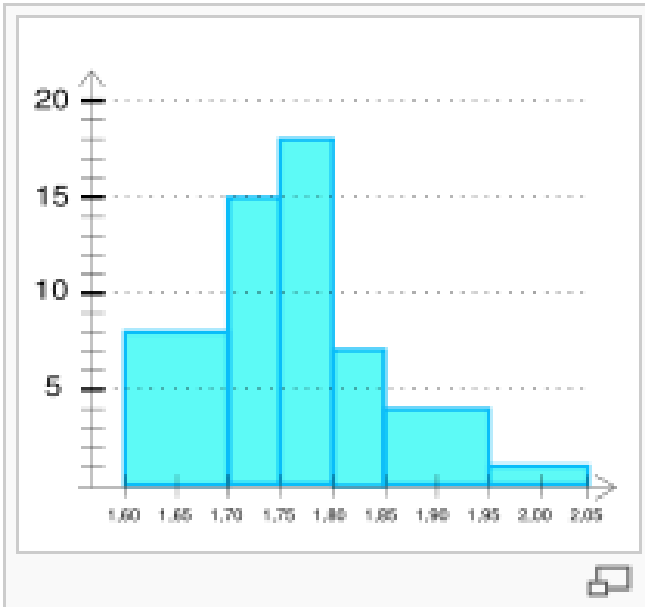


Histogramme

L'histogramme est un moyen simple et rapide pour représenter la distribution d'un paramètre.

Exemple :

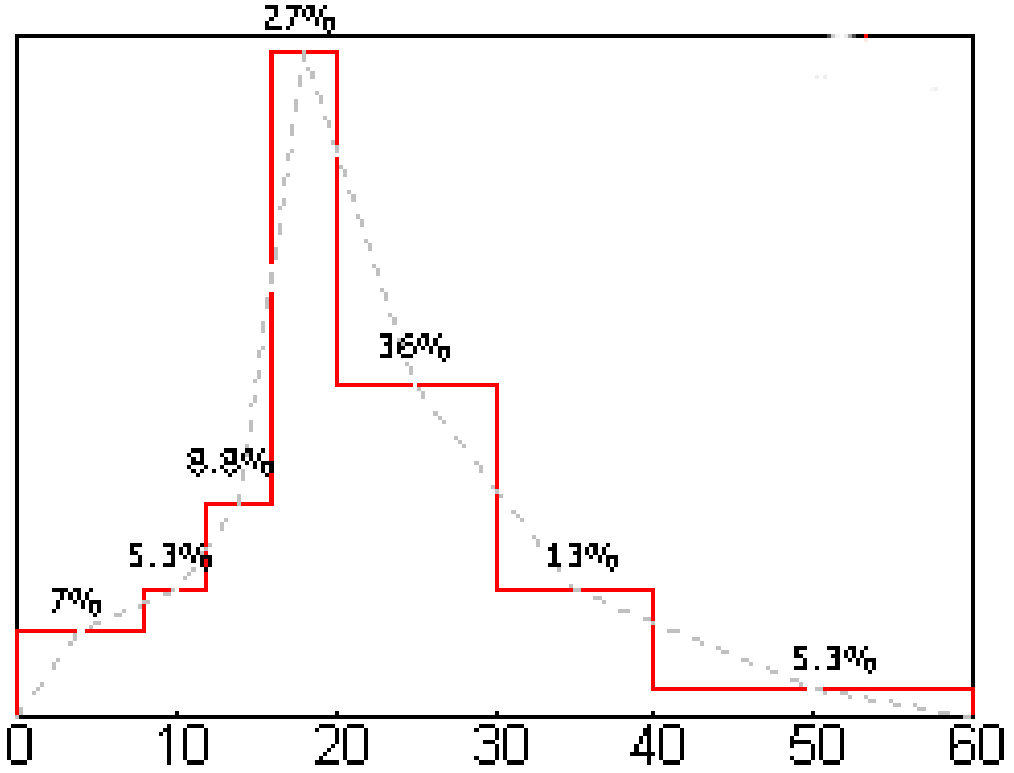
- diamètre d'un arbre après usinage,
- dureté d'une série de pièces après un traitement thermique,
- concentration d'un élément dans la composition d'alliages produit par une fonderie,
- masse de préparation alimentaire dans une boîte de conserve



- On utilise l'**histogramme** en respectant la règle des aires.
- Pour éviter toute ambiguïté, il est préférable de travailler avec des classes d'amplitude constante.
- Dans ce cas, les hauteur des rectangles sont proportionnelles aux effectifs (ou aux fréquences).
- L'histogramme peut aussi être normalisé pour représenter les **pourcentages** des effectifs ou des fréquences

Pour pouvoir bien mener l'étude de la dispersion d'un paramètre à l'aide d'un ou de plusieurs histogrammes, il faut connaître les conditions de collecte des données:

- fréquence de mesure,
- outil de mesure utilisé,
- possibilité de mélange de lots,
- possibilité de tri,
- etc.



Histogramme – 1. Collecte des données

- La première phase est la collecte des données en cours de fabrication.
- Cette collecte peut être réalisée soit de façon exceptionnelle à l'occasion de l'étude du paramètre soit en utilisant un relevé automatique ou manuel fait lors d'un contrôle réalisé dans le cadre de la surveillance du procédé de fabrication (**contrôle de la qualité**).
- Sans qu'il soit réellement possible de donner un nombre minimum, il faut que le nombre de valeurs relevées soit suffisant.
- Plus l'on dispose d'un nombre élevé de valeurs, plus l'interprétation sera aisée.

- Généralement on utilise des classes de largeur identique.
- Le nombre de classes dépend du nombre de valeurs *N* dont on dispose.
- Le nombre de classe K peut être déterminé par la formule suivante :

$$K = 1 + \frac{10 \log(N)}{3}$$

ou plus simplement:

$$K = \sqrt{N}$$

- **Ces formules ne sont que des indications**, l'histogramme étant un outil visuel, il est possible de faire varier le nombre de classes.

Ce qui importe est la meilleure distribution qui facilitera l'interprétation.

Histogramme – 2. Définir les intervalles de classe

- L'amplitude *w* de l'histogramme est

  w = *valeur maximale* ─ *valeur minimale*,

- L'amplitude *h* théorique de chaque classe est alors :

  $h = w / K$

- Il faut arrondir cette valeur à un multiple de résolution de l'instrument de mesure (arrondi à l'excès).

Histogramme – 3. Calculer et tracer l’histogramme

Exemple (tiré de l'article sur Wikipédia)

Soit la fabrication de rations alimentaires, la pesée des rations avant emballage donne la série de mesures suivantes en kg :

0,547	0,563	0,532	0,521	0,514	0,547	0,578	0,532	0,552	0,526	0,534	0,560	0,502	0,503	0,516	0,565
0,532	0,574	0,521	0,523	0,542	0,539	0,543	0,548	0,565	0,569	0,574	0,596	0,547	0,578	0,532	0,552
0,554	0,596	0,529	0,555	0,559	0,503	0,499	0,526	0,551	0,589	0,588	0,568	0,564	0,568	0,556	0,523
0,526	0,579	0,551	0,584	0,551	0,512	0,536	0,567	0,512	0,553	0,534	0,559	0,498	0,567	0,589	0,579

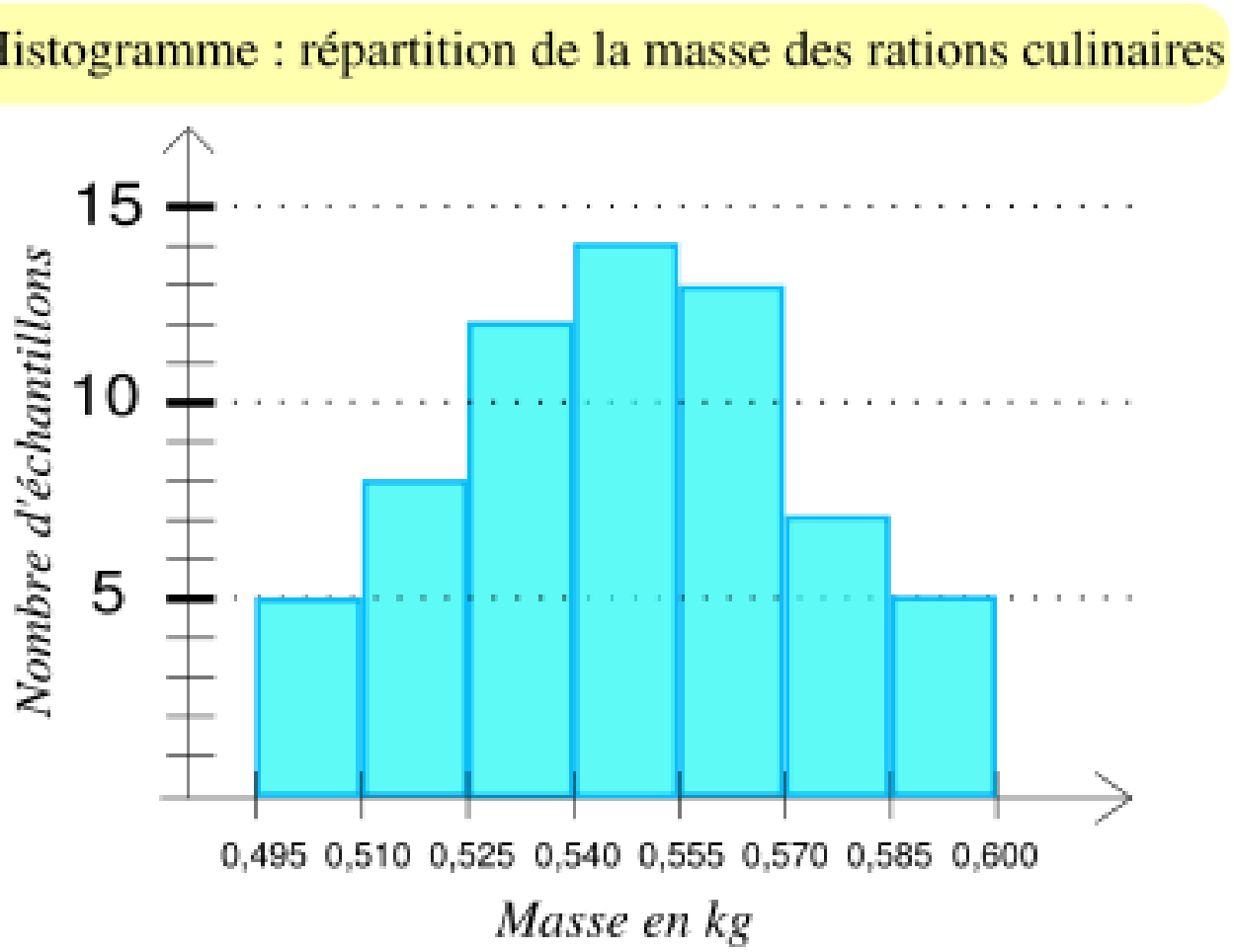
Les caractéristiques du relevé sont les suivantes :

- Le nombre d'échantillons : N=64
- L'étendue : w=0,098 kg
- Valeur minimale : 0,498 kg
- Valeur maximale : 0,596 kg

On en déduit les paramètres suivants pour l'histogramme :

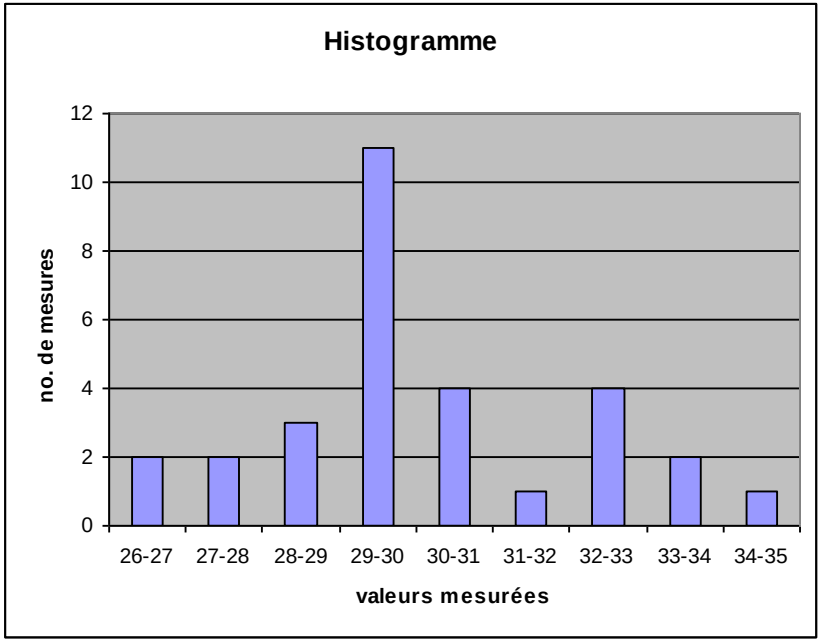
- Le nombre de classes est de 7 (en utilisant la formule avec le logarithme)
- L'amplitude de classe est 0,098/7 = 0,014 kg que l'on arrondit à 0,015 kg (résolution de la balance : 0,001 kg)
- La valeur minimale de la première classe est de 0,498 – (0,001/2) = 0,4975. Par souci de facilité pour l'interprétation, on peut arrondir cette valeur à 0,495 kg.

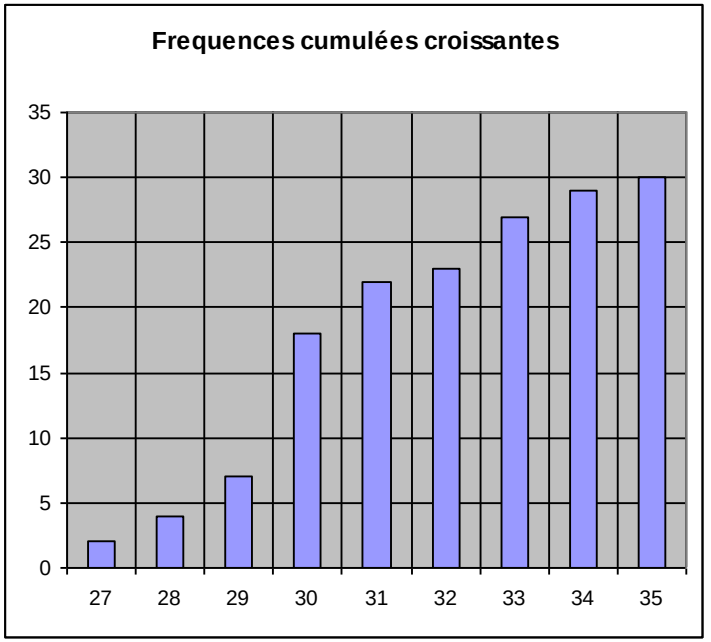
On obtient l'histogramme suivant



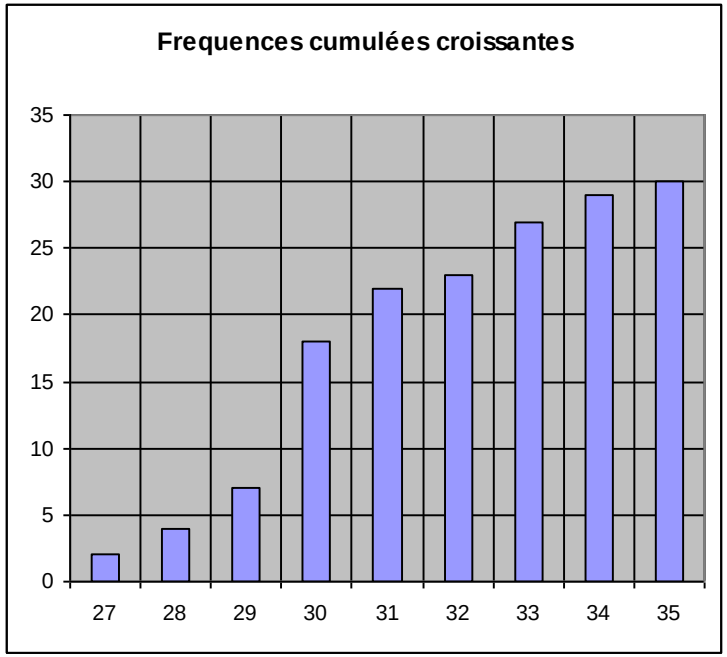
Histogramme – 4. Diagramme des effectifs cumulés (fonction de répartition)

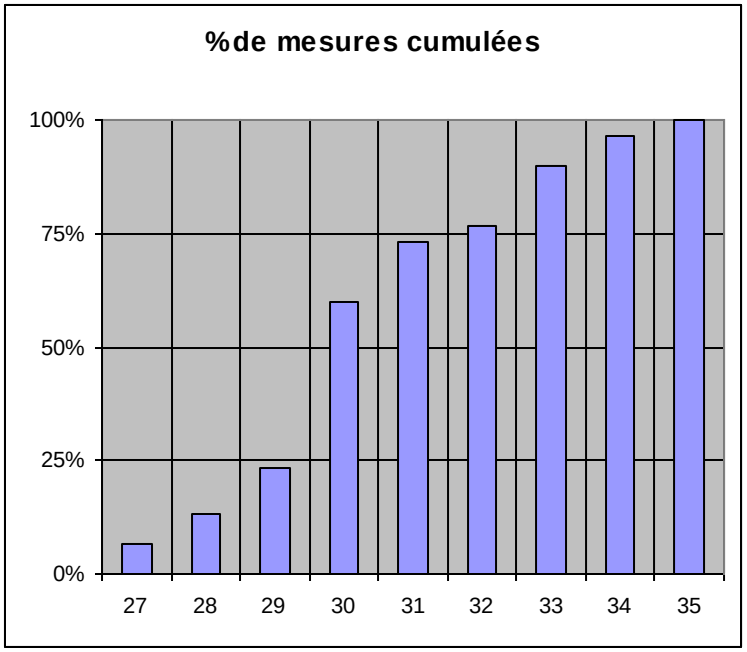
- Ce diagramme peut s'obtenir facilement depuis l'histogramme.
- Ce diagramme permet de lire l'effectif d'un **intervalle entre 0 et x** et , par différence, l'effectif de tout intervalle.
- Cette représentation préfigure le tracé de la fonction de répartition en probabilité.



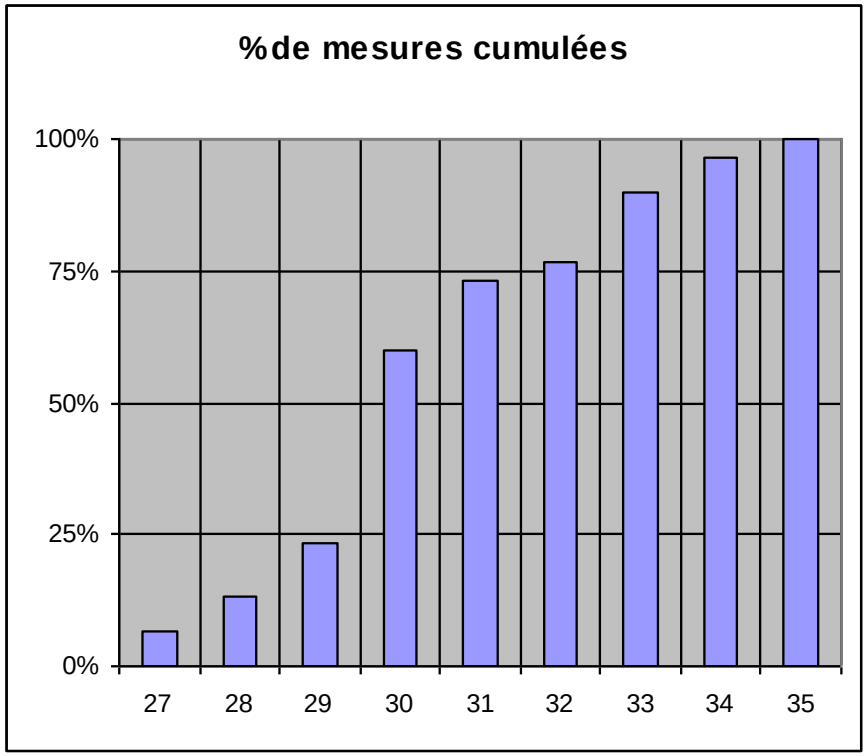


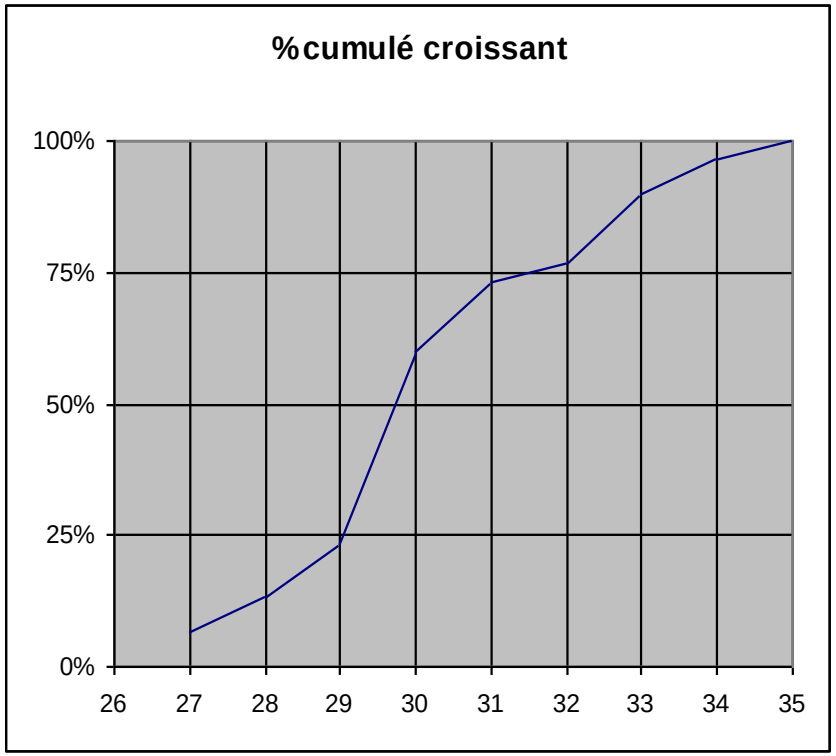
- Le diagramme des effectifs cumulés peut indiquer soit le nombre absolu de mesures, soit le pourcentage.





- Le diagramme des effectifs cumulés peut aussi être mis en forme de **polygone** des effectifs cumulés, aussi pour des intervalles discrets.





Les quantiles

- Les quantiles sont des points essentiels pris à des intervalles réguliers verticaux d'une fonction de distribution cumulative d'une variable aléatoire.

- Diviser des données ordonnées en q sous-jeux de données de dimension essentiellement égale est la motivation des q-quantiles ; les quantiles sont les valeurs de données marquant les limites entre deux sous-jeux consécutifs.

- Certains quantiles ont des noms spéciaux :
  - Les 100-quantiles sont appelés **centiles ou percentiles** selon un *anglicisme* fréquent,
  - Les 10-quantiles sont appelés **déciles**,
  - Les 4-quantiles sont appelés **quartiles**.

- Le diagramme des effectifs cumulés ou fonction de répartition permet de lire facilement les quantiles intéressants (en général des quartiles et les déciles).

Comment faire un histogramme avec Excel 2003

- Supposons la **série suivante:**
  - **5, 7, 8, 3, 7, 7, 1, 9, 6, 8, 5, 6, 7, 8, 7, 9, 6, 8, 6, 6**.
  - Inserer les chiffres dans la colonne A, en commençant par A2, par exemple.
- On veut construire un histogramme avec **5 barres**, avec intervalles**:**
  - 0 – 2
  - 2 – 4
  - 4 – 6
  - 6 – 8
  - 8 – 10
- Dans les colonnes B et C, commençant par B2-C2 on insert respectivement 0, 2, ensuite à la ligne suivante 2, 4, etc..
- **Selectionnez 5 cellules contigues,** par exemple E2:E6.
- Entrez la formule
  = frequence (A2:A22; C2:C6).
  Le premier vecteur contient les données, le deuxième contient les **limites supérieurs des intervalles (colonne C2:C6)**.
  Presser **Control-Shift-Enter** simultanément.
  Les valeurs de l'histogramme apparaissent dans les cases E2:E6.
- Afin d'obtenir un graphique illustratif on peut insérer à la colonne D les plages d'intervalles, par exemple pour D2: **=B2&"-"&C2**, ce qui donne le texte **0-2** .
- Ensuite on sélectionne ensemble les colonne D et E et on produit un graphique de type Histogramme.

Exercice 1

- Télécharger le fichier Excel avec les séries de mesure (30, 70 et 100 valeurs).

- Calculer:
  - Nombre de mesures - fonction NB()
  - Moyenne
  - Médiane
  - Ecart-type
  - Minimum
  - Maximum
  - Histogramme

Exercice 2

- Compléter l'exercice statistique précédent avec:
  - Le diagramme des effectifs cumulés
    - en escalier, escalier en pourcentages,
    - polygone en pourcentages = fonction de répartition
  - Le calcul des quartiles

Exercice 3

- Une série de mesures a été pré-traitée et publiée en forme d'histogramme (mais sans accès aux données de bases).

  Télécharger le fichier Excel.

  Calculer les estimations de:

  - La moyenne
  - L'écart-type
  - Le polygone des effectifs cumulés en pourcentages (fonction de répartition)
  - Evaluer la médiane et les quartiles à partir du diagramme

Probabilité: quelques notions de base

- Définitions
- Lois de probabilité ou distributions
- Densité de probabilité
- Fonction de répartition

Définitions

- La probabilité (du latin probare, « prouver », « tester ») est une évaluation du caractère probable d'un événement.
  Un événement est probable « s'il peut se produire » (dans le cas de futures éventualités), ou s'il est « vraisemblable » (dans le cas d'inférences de l'évidence).

- **La notion de probabilité est inévitablement associée à celle d'incertitude**: L'incertitude peut naître de notre ignorance, être due à un embrouillement ou une incompréhension, ou provoquée par l'aspect aléatoire essentiel de la nature.

- Dans tous les cas, nous mesurons l'incertitude des évènements sur une échelle de **zéro** (pour les évènements impossibles) à **un** (pour les évènements certains).

L'idée de probabilité est le plus souvent séparée en deux concepts:

1. la probabilité de l'**aléatoire**, qui représente la probabilité d'évènements futurs dont la réalisation dépend de quelques phénomènes physiques aléatoires, comme obtenir un as en lançant un dé ou obtenir un certain nombre en tournant une roue;

2. la probabilité de l'*épistémé* *), qui représente l'incertitude que nous avons devant des affirmations, lorsque nous ne disposons pas de la connaissance complète des circonstances et des causalités.

***) L'*épistémé* est pour Platon le savoir authentique, qu'il oppose à la *doxa*, l'opinion.**

Lois (ou distributions) de probabilité

- Une des notions les plus importantes en probabilité est celle de **variable aléatoire**.
- Une variable aléatoire est une application qui à un résultat possible de l'expérience associe une valeur. Une variable aléatoire va donc prendre telle ou telle valeur suivant le résultat obtenu; et ce ne sont pas les valeurs possibles de la variable, ni la valeur qu'elle prend une fois que l'on connaît le résultat de l'expérience qui sont aléatoires, mais la valeur qu'elle va prendre avant d'avoir effectué l'expérience.
- La somme des probabilités de toutes les valeurs possibles d'une variable aléatoire valant un, ces probabilités sont en quelque sorte réparties sur ces différentes valeurs.

- Toute relation qui établit correspondance entre les valeurs prises par une variable et leur probabilité s'appelle une **loi (ou distribution) de probabilité.**

- Une loi de probabilité ou distribution décrit les répartitions typiques des fréquences d'apparition des résultats d'un phénomène aléatoire.

- Dans le dernier quart du XXe siècle, on a largement étendu le concept à des domaines où il n'était plus question de fréquences, mais aussi de représentation d'états de connaissance.

- On associe généralement une loi de probabilité à une variable aléatoire pour décrire la répartition des valeurs qu'elle peut prendre.

Lois discrètes et continues

Une loi de probabilité peut concerner:

1. des événements ou résultats discrets,
2. des résultats continus.

Selon si la variable aléatoire sous-jacente est discrète ou continue

Densité de probabilité

Wikipédia :

En mathématiques statistiques, on appelle **densité de probabilité** d'une variable aléatoire X réelle continue une fonction f

- positive ou nulle sur $\mathbb{R}$;
- intégrable sur $\mathbb{R}$;
- vérifiant $\int_{\mathbb{R}} f(t)\,dt = 1$

La probabilité $P(a < X \leq b)$ se calcule alors par la relation suivante :

$$P\left(a < X \leq b\right) = \int_a^b f\left(t\right)\,dt$$

En traçant la représentation graphique de la densité de probabilité, la probabilité $P(a < X < b)$ se lit comme l'aire sous la courbe sur l'intervalle [a ; b].

Fonction de répartition (*probabilités cumulées*)

Wikipédia :

En probabilité, la **fonction de répartition** d'une variable aléatoire X est la fonction F_X qui à tout réel x associe

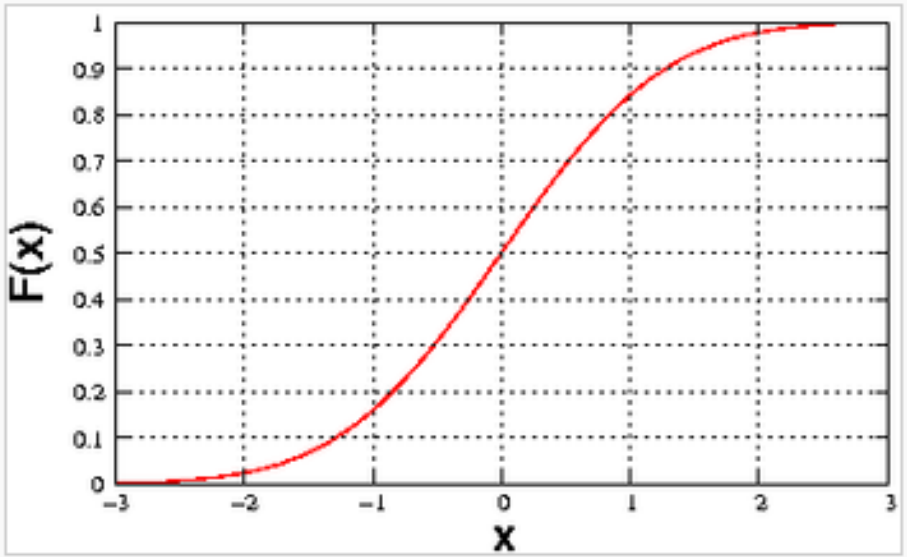
$$F_X(x) = P[X \leq x].$$

La fonction de répartition d'une variable aléatoire continue est la primitive de la densité de probabilité f_X.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

La fonction de répartition a les propriétés suivantes :

- F_X est croissante.
- Elle est partout continue à droite. et admet en tout point x_0 une limite à gauche, égale à $P[X < x_0]$.
- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to +\infty} F_X(x) = 1$



Fonction de répartition de la loi normale centrée réduite

Travail personnel

- Etudier les articles de Wikipedia
  - *Statistiques*
  - *Probabilité*
  - *Variable aléatoire réelle*

(noter d'éventuelles questions …)

Les plus importante lois (distributions) de probabilité

- Il y a beaucoup de lois de probabilité qui ont été formulé pour représenter divers phénomènes aléatoires.

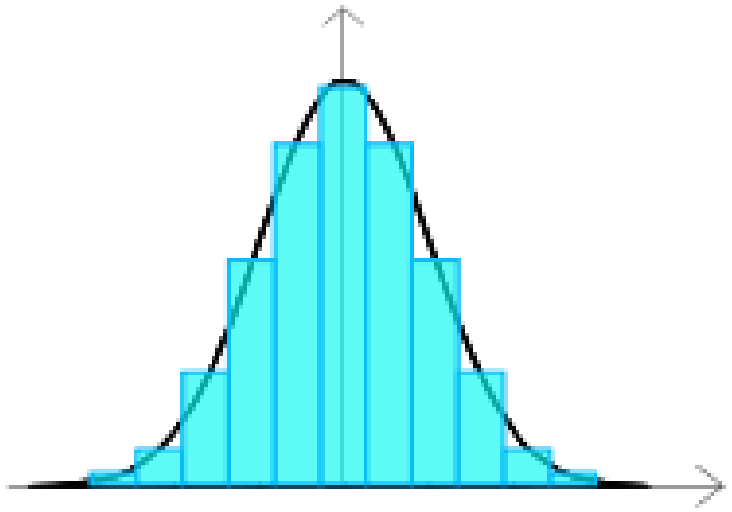
- Parmi l'ensemble des **lois de probabilités possibles**, on distingue un certain nombre de familles usuelles qui correspondent à des phénomènes aléatoires simples: lancer de dés, jeu de pile ou face, **erreurs de mesures**, etc.

- Combinées entre elles, elles permettent d'élaborer des modélisations de phénomènes aléatoires plus complexes.

La loi ou distribution normale ou gaussienne

- La distribution de beaucoup de paramètres industriels correspond souvent à une loi normale, avec son profil « en cloche » .

- Typiquement ce sera la première distribution avec la quelle nous allons comparer nos histogrammes de mesure.



Le théorème de la limite centrale

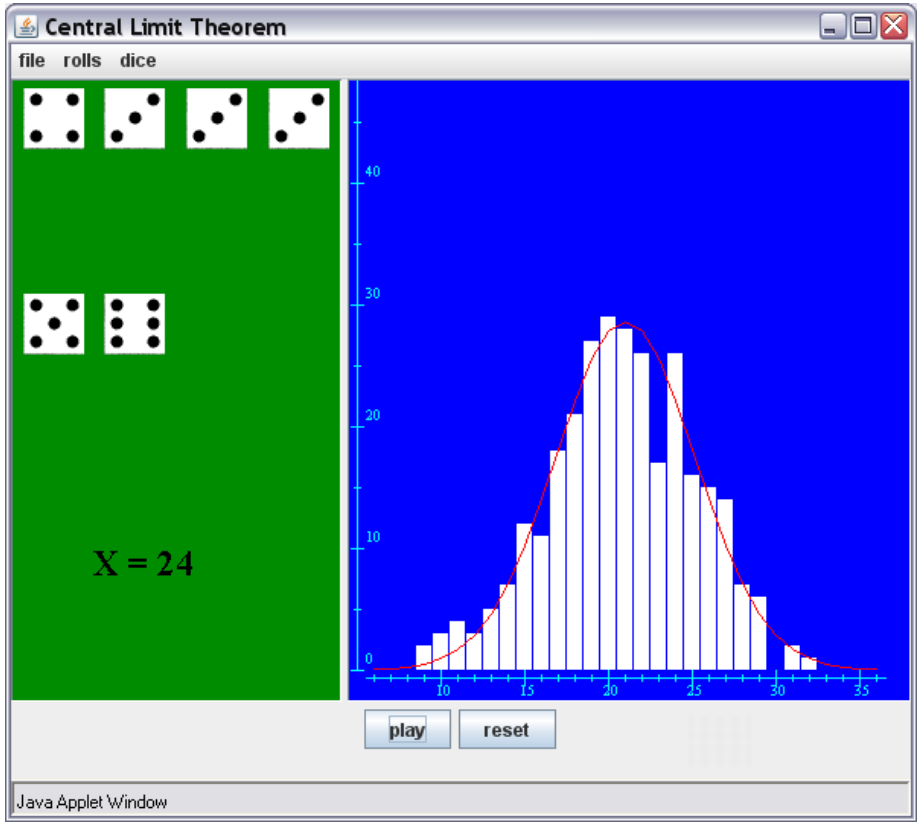
- Le **théorème de la limite centrale** (ou : de la limite centrée ; on trouve en franglais l'appellation fréquente de théorème central limite) est un ensemble de résultats sur la convergence faible d'une suite de variables aléatoires en probabilité.
- Intuitivement, d'après ces résultats, toute somme de variables aléatoires indépendantes et identiquement distribuées tend vers une certaine variable aléatoire.
- Le résultat le plus connu et le plus important est simplement appelé « théorème de la limite centrale » qui concerne une somme de variables aléatoires dont le nombre tend vers l'infini.

- Dans le cas le plus simple, considéré ci-dessous pour la démonstration du théorème, ces variables sont indépendantes et possèdent la même moyenne et la même variance. En général, la somme croît indéfiniment en même temps que le nombre de termes.
- Pour tenter d'obtenir un résultat fini, il faut centrer cette somme en lui soustrayant sa moyenne et la réduire en la divisant par son écart-type.

- **Sous des conditions assez larges, la loi de probabilité converge alors vers une loi normale unitaire (dont l'écart-type est 1).**
- **L'omniprésence de la loi normale s'explique par le fait que de nombreux phénomènes considérés comme aléatoires sont dus à la superposition de causes nombreuses.**

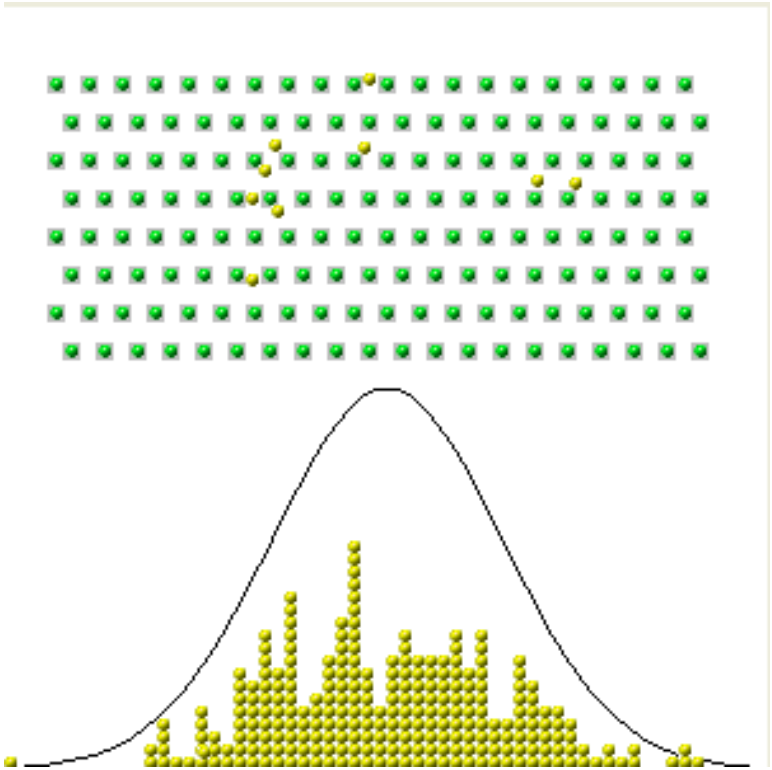
Le théorème de la limite centrale – vérification expérimentale

Applet:
http://www.math.csusb.edu/faculty/stanton/probstat/clt.html



Une autre applet qui montre comment une distribution normale résulte de la superposition de plusieurs nombreuses causes aléatoires qui sont – elles – distribuées uniformement.

http://php.iai.heig-vd.ch/~lzo/applets/metrologie/BallDrop/



Distribution normale ou gaussienne

La distribution de beaucoup de paramètres industriels correspond souvent à une **loi normale**, avec son profil « en cloche » .

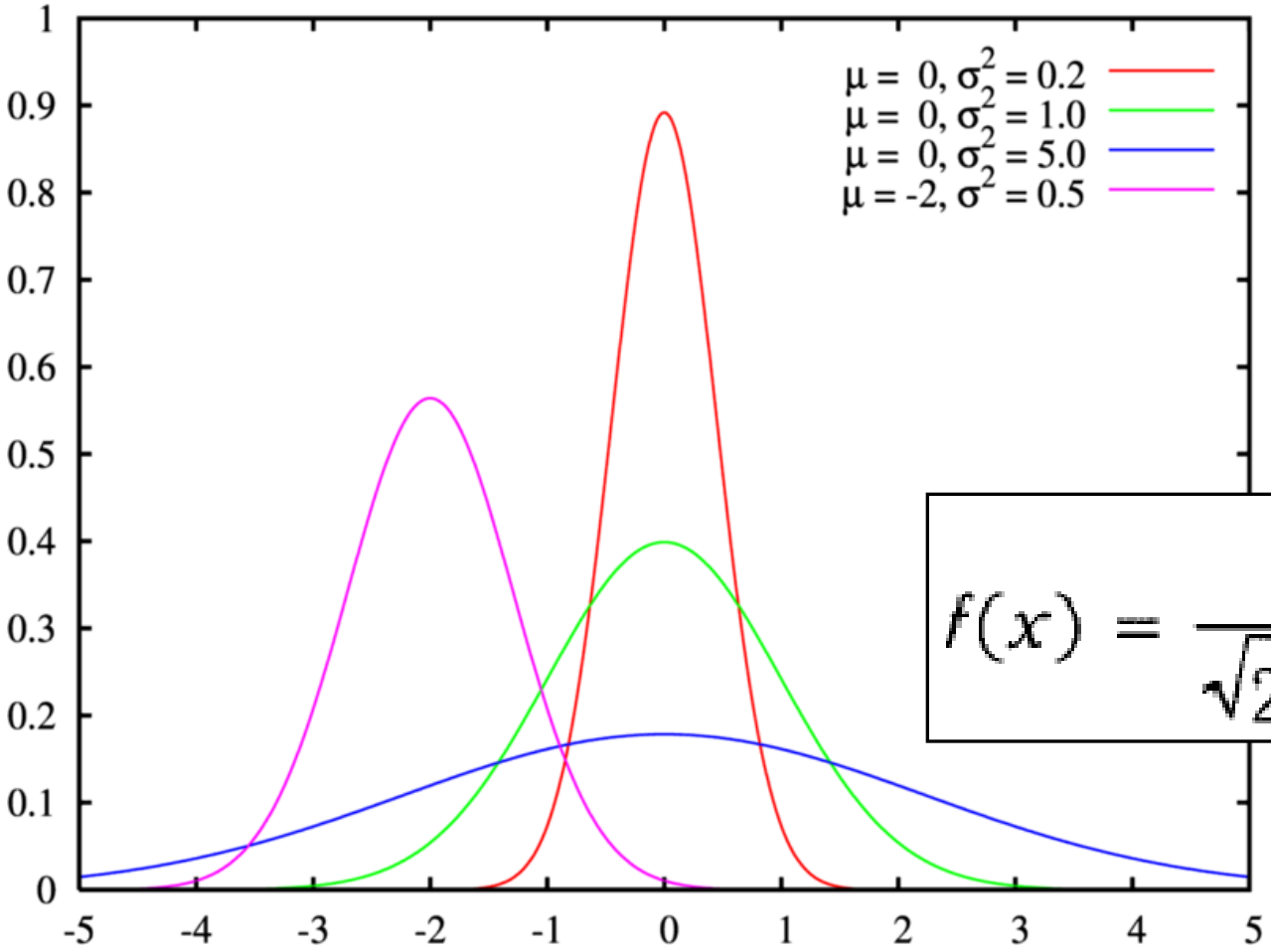
Typiquement ce sera la première distribution avec la quelle nous allons comparer nos histogrammes de mesure.

Une variable aléatoire suit une loi normale (ou loi normale gaussienne, loi de Laplace-Gauss) de moyenne μ et d'écart type σ (donc de variance σ^2) si elle admet une densité de probabilité f telle que:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

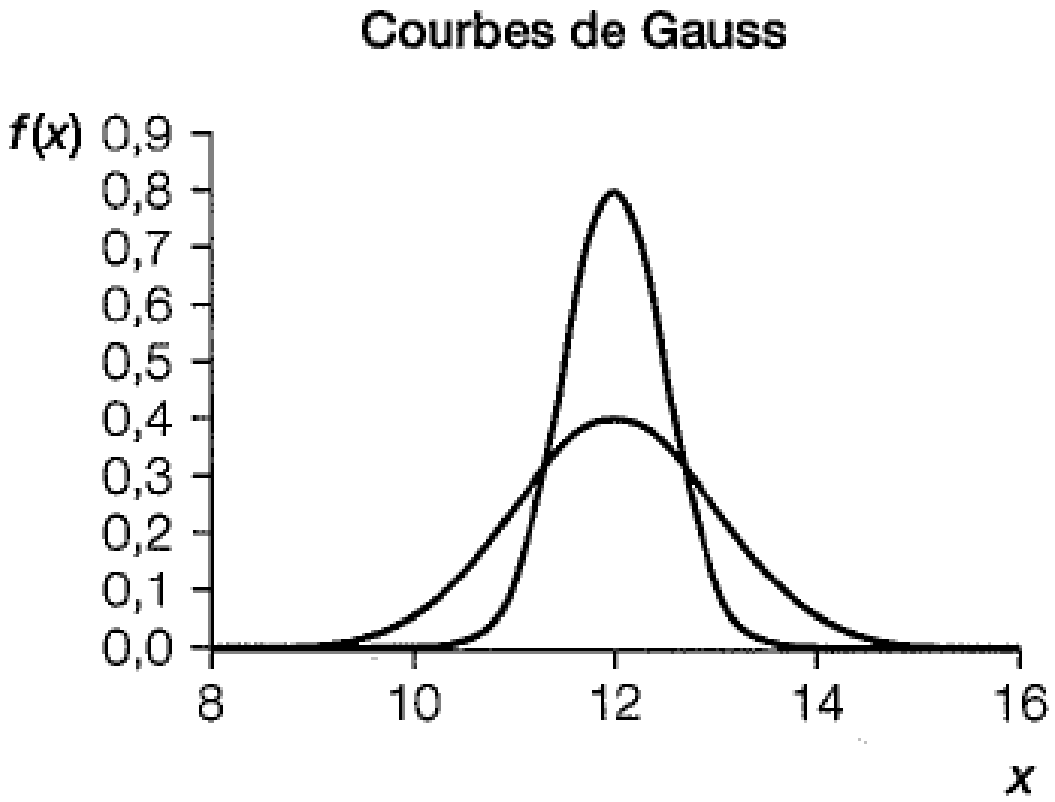
Une telle variable aléatoire est dite **variable gaussienne**.

La distribution normale est une courbe avec deux paramètres

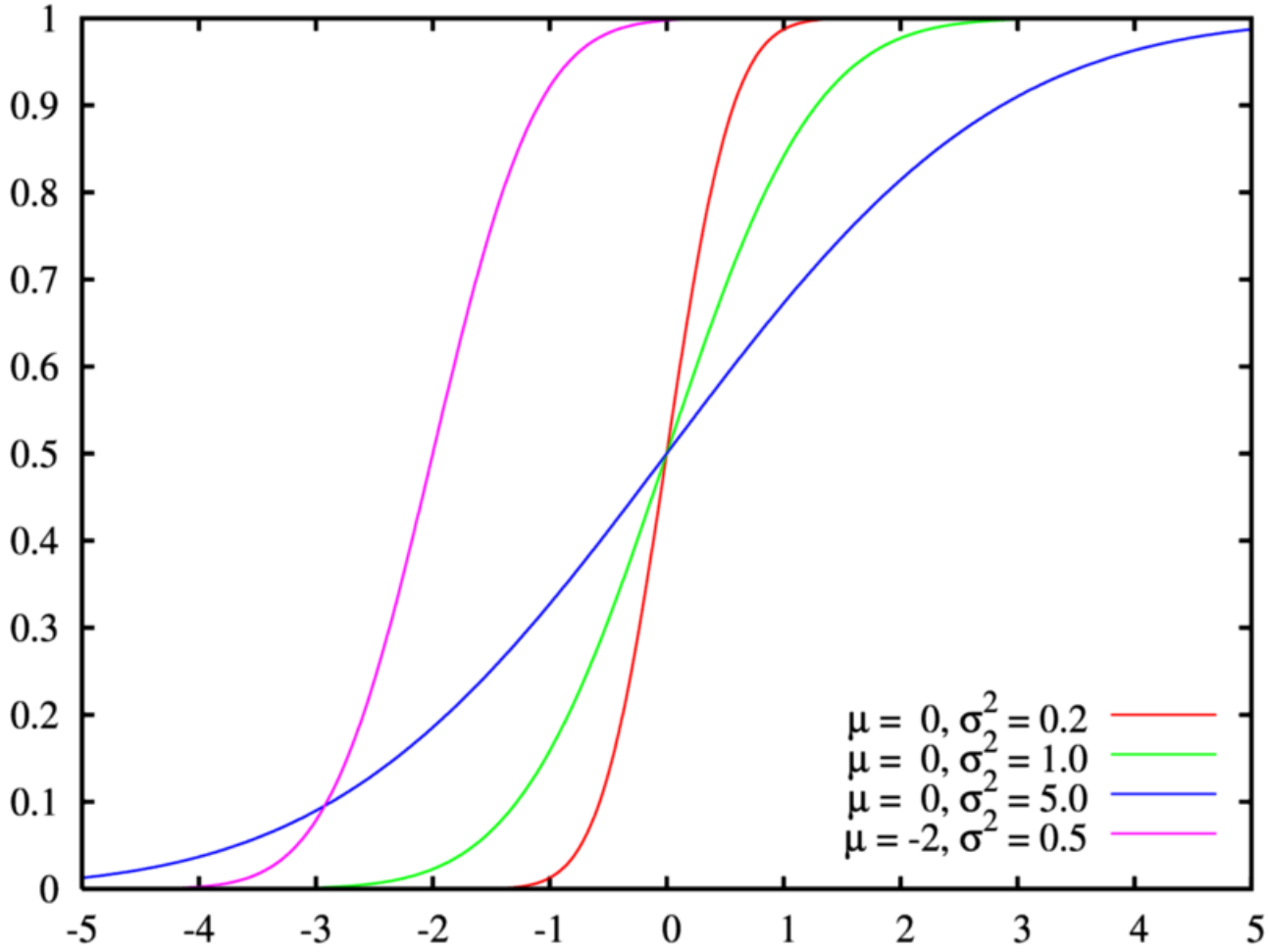


$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

On voit ici deux courbes de Gauss ayant la même moyenne, c'est-à-dire $\mu = 12$. La courbe arrondie a un écart type σ de 1 ; la courbe pointue, un écart type σ de 0,5.



Fonction de répartition *(probabilité cumulée)* gaussienne



- On note Φ la fonction de répartition de la loi normale centrée réduite.
- Elle est définie, pour tout réel *x*, par :

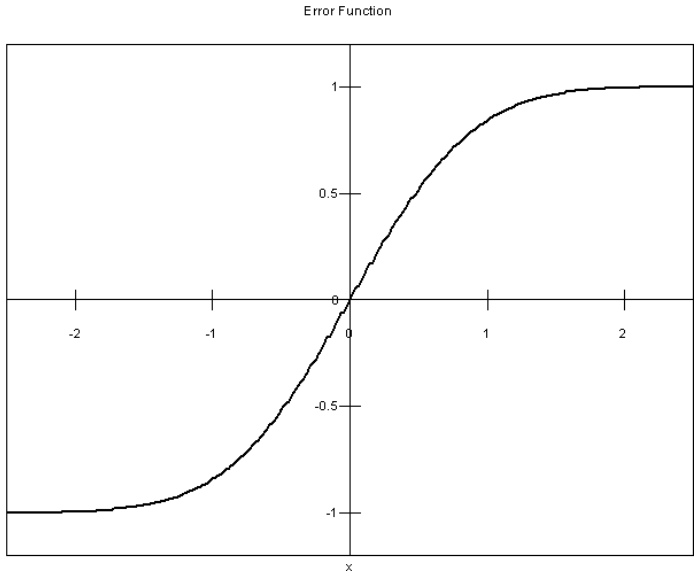
$$\begin{aligned}\Phi_{\mu,\sigma^2}(x) &= \int_{-\infty}^{x} \varphi_{\mu,\sigma^2}(u)\,du \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du \\ &= \frac{1}{2}\left(1 + \operatorname{erf}\frac{x-\mu}{\sigma\sqrt{2}}\right)\end{aligned}$$

Fonction d'erreur

- En mathématiques, la fonction d'erreur (aussi appelée fonction d'erreur de Gauss) est une fonction utilisée en analyse.

- Cette fonction se note ***erf*** et fait partie des fonctions spéciales.

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-\zeta^2}\, d\zeta$$



Approximation de la fonction de répartition

- Il n'existe pas d'expression pour Φ mais on peut exploiter avec profit son aspect régulier pour en donner une approximation grâce à un développement en série de Taylor.

- Par exemple, voici une approximation (à l'ordre 5) autour de 0:

$$\Phi(x) \approx \frac{1}{2} + 0,3989423 \left[x - \frac{x^3}{6} + \frac{x^5}{40} \right]$$

- Cette approximation est performante pour

$$| x | < 2$$

Intervalle et niveau de confiance

- Lorsque le caractère statistique a une distribution normale gaussienne, donc au moins grossièrement en forme de cloche, l'écart type prend tout son sens:
- Dans l'intervalle $[\overline{x} - \sigma; \overline{x} + \sigma]$, on trouve **68%** de la population.
- Dans l'intervalle $[\overline{x} - 2\sigma; \overline{x} + 2\sigma]$, on trouve **95%** de la population.
- Dans l'intervalle $[\overline{x} - 3\sigma; \overline{x} + 3\sigma]$, on trouve **99,7%** de la population.

On appelle ces intervalles
les **plages de normalité au niveau de confiance** de

1. 68% ou 1-sigma,
2. 95% ou 2-sigma,
3. 99,7% ou 3-sigma.

Largeur à mi-hauteur

- Lorsque l'on travaille sur une représentation graphique, on estime fréquemment la largeur de la gaussienne par sa largeur à mi-hauteur H (en anglais *full width at half maximum*, *FWHM*), qui est la largeur de la courbe à une altitude qui vaut la moitié de l'altitude du sommet.

- La largeur à mi-hauteur est proportionnelle à l'écart type :

$$H = 2\sqrt{2\ ln(2)}\ \sigma \simeq 2,3548\sigma$$

- 76% de la population est dans l'intervalle

$$[\overline{x} - 0,5H\ ;\ \overline{x} + 0,5H]$$

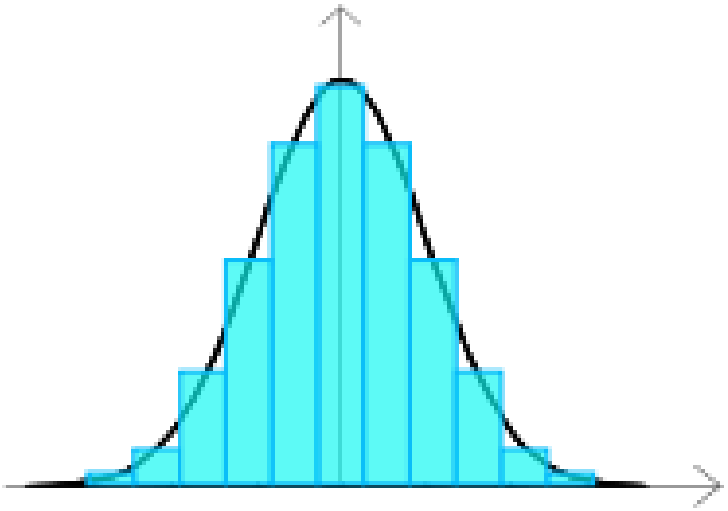
Exercices

- Exploration numérique de la fonction de Gauss et de la fonction de répartition

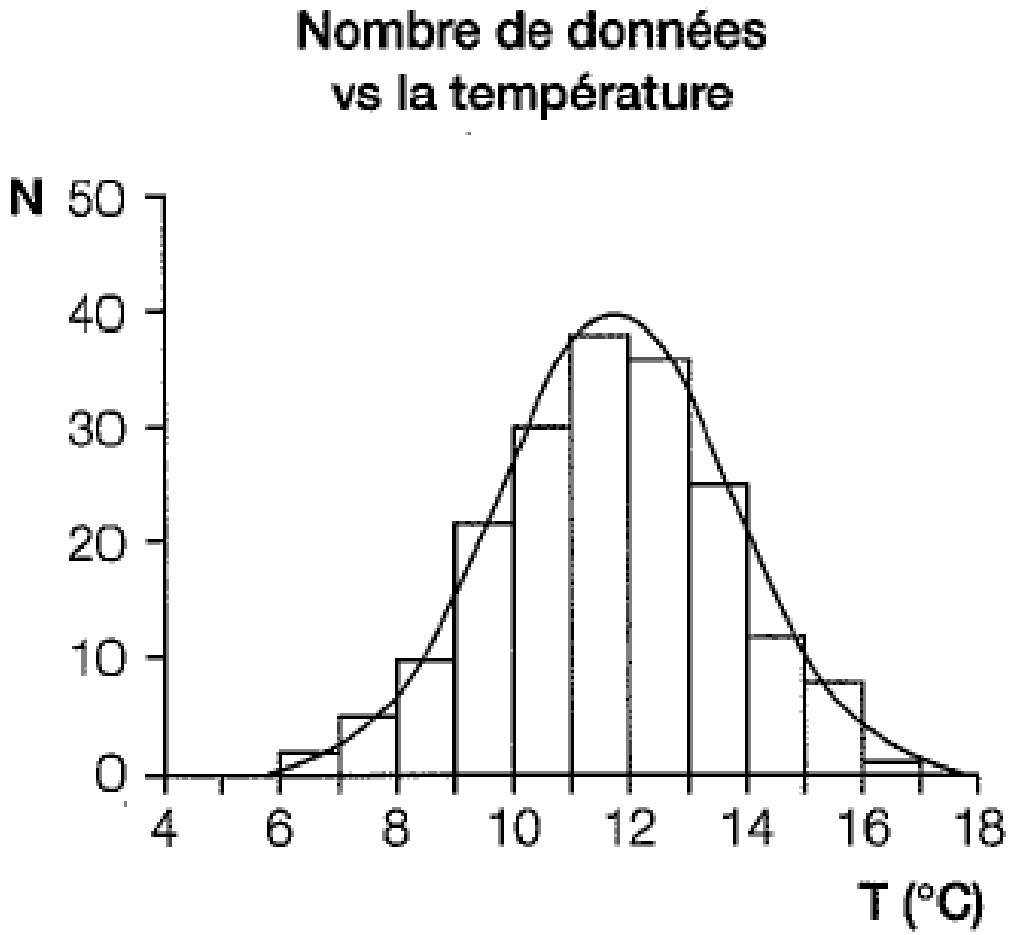
- Le temps requis par un étudiant pour se rendre à l'école suit une loi normale de moyenne 47 minutes et écart type 8 minutes. Quelle est la probabilité qu'il prenne plus d'une heure pour arriver ?
- Selon la météo locale les précipitations journalière pour le mois de juillet se distribuent normalement avec moyenne de 75 mm et écart type de 5 mm.
  - Combien de jours estimons-nous avoir durant ce mois avec 80 mm et plus de pluie ?
  - Un intervalle de 95% correspond à combien de mm de pluie ?

Interprétation d'un histogramme

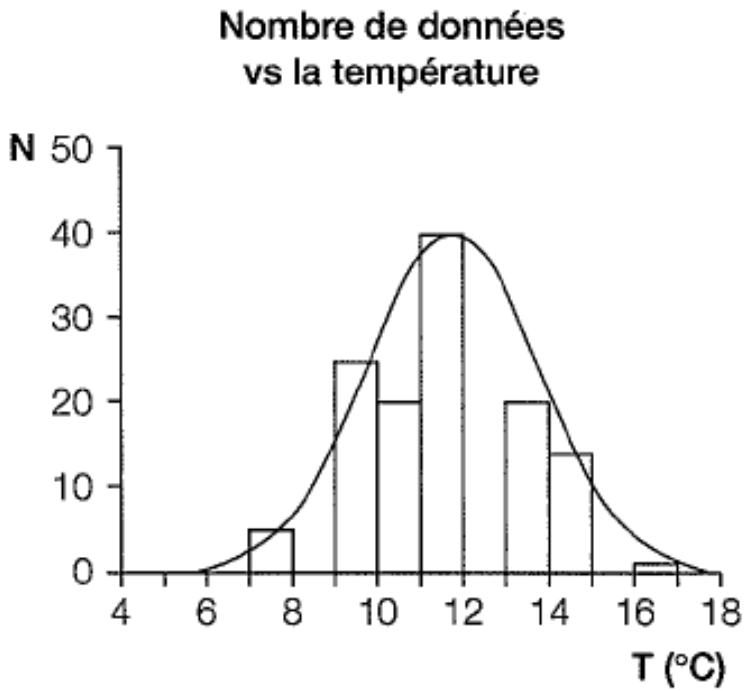
- La distribution de beaucoup de paramètres industriels correspond souvent à une loi normale. On compare souvent l'histogramme obtenu au profil « en cloche » de la loi normale.
- Cette comparaison est visuelle et même si elle peut être une première approche, elle ne constitue pas un test de «normalité».



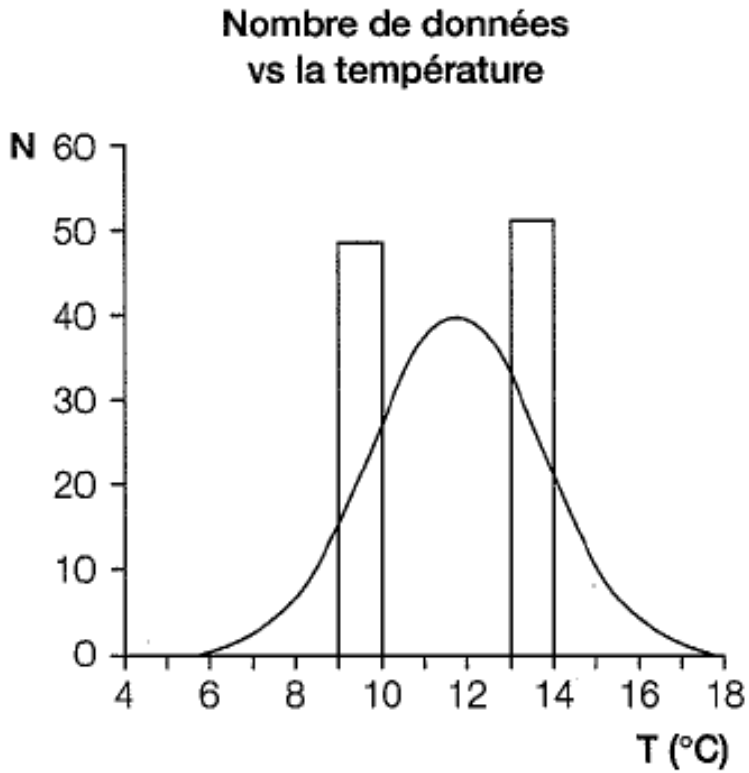
On voit ici un cas idéal : l'histogramme s'insère parfaitement dans une courbe de Gauss.
Au laboratoire, on rencontre rarement un tel cas.



Très souvent, les histogrammes qu'on obtient ne s'ajustent pas parfaitement à une courbe de Gauss. Dans ces cas, on utilise tout de même les outils statistiques d'une distribution gaussienne.



Parfois, l'histogramme ne s'insère pas du tout dans une courbe de Gauss. Dans ce cas, on ne doit pas utiliser les outils statistiques d'une distribution gaussienne.



Critères de normalité

- Le recours à une distribution gaussienne est si fréquent qu'il peut finir par être abusif. Il faut alors rechercher des critères de normalité.

1. Le **premier critère**, le plus simple, consiste à tracer l'histogramme ou le diagramme en bâtons de la distribution et à vérifier si le diagramme est en forme de « cloche ». Ce critère, subjectif, permet cependant d'éliminer une partie des distributions jugées alors non gaussiennes.

2. Le **critère suivant** consiste à utiliser les plages de normalité ou intervalles de confiance. On a vu que si une distribution est gaussienne :
    1. 68% de la population est dans l'intervalle +/- 1 sigma
    2. 95% de la population est dans l'intervalle +/- 2 sigma
    3. 99,7% de la population est dans l'intervalle +/- 3 sigma

    Lorsque ces pourcentages ne sont pas (pus ou moins bien) respectés, il est fort à parier que la distribution n'est pas gaussienne.

Exercice

Vérifions – tout d’abord par les critères les plus simples - si les histogrammes des exercices précédents sont **compatibles** avec une distribution normale.

Méthode de la **droite de Henry**

- La droite de Henry est une méthode pour visualiser les chances qu'a une distribution d'être gaussienne.

- Elle permet de lire rapidement la **moyenne** et **l'écart type** d'une telle distribution.

Droite de Henry

Wikipédia:
http://fr.wikipedia.org/wiki/Droite_de_henry

- Principe
- Exemple

Exercice

1. Vérifions par la méthode la droite de Henry
   si les histogrammes des exercices précédents sont **compatibles** avec une distribution normale.

2. Calculons la moyenne et l'écart-type dérivés pour cette distribution normale.

Travail personnel (préparation au TE)

Tout le chapitre 5 du polycopié

Exercices

Vous trouverez à l'adresse:
http://php.iai.heig-vd.ch/~lzo/metrologie/exercices/mesures.xls
un fichier Excel avec plusieurs séries de mesures.

Pour chaque série de mesure, créez une nouvelle feuille de calcul et déterminez les éléments suivants de l'échantillon de mesure:

- Minimum
- Maximum
- Moyenne
- Ecart type corrigé
- Incertitude nominale à 95% (2-sigma)

Calculer et tracez:

- Histogramme
- Répartition en %
- Droite de Henry

Calculer la moyenne et l'écart type estimé de la population.
Ajoutez la courbe de la densité de probabilité normale sur l'histogramme.

Compléments sur la loi normale

Stabilité de la loi normale par la somme

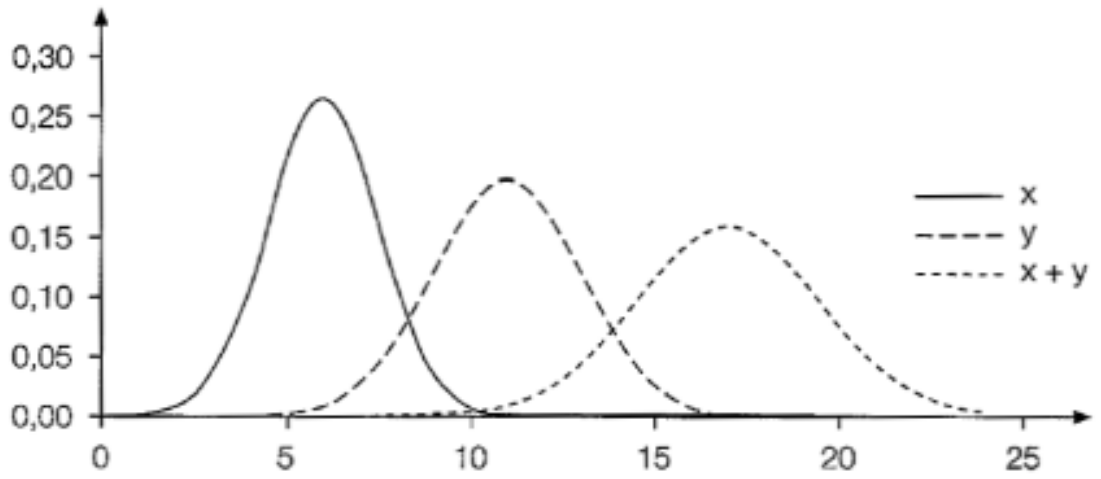
La somme de deux variables gaussiennes **indépendantes** est elle-même une variable gaussienne. Plus explicitement :

Soient X_1, X_2 deux variables aléatoires indépendantes suivant respectivement les lois $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$.

Alors, la variable aléatoire $X_1 + X_2$ suit la loi normale $\mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.

Le contraire est aussi vrai.
Si deux variables aléatoires ont une somme qui a une distribution gaussienne, elles sont aussi gaussiennes.

La largeur σ_1 de la distribution des x_i se combine avec la largeur σ_2 des y_i selon une somme quadratique telle que : $\sigma_T = \sqrt{\sigma_1^2 + \sigma_2^2}$.



Stabilité de la loi normale

Stabilité de la loi normale également par

- la différence
- la moyenne

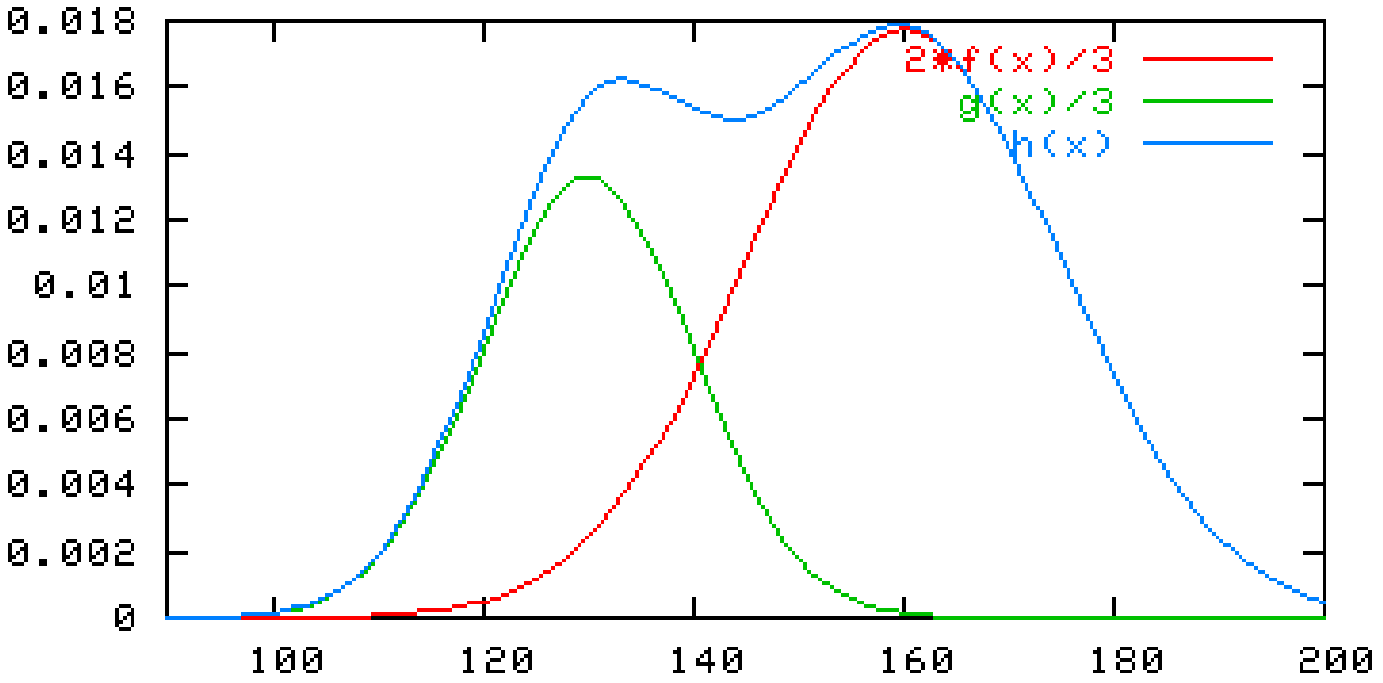
Mélange de populations

- Il ne faut pas confondre la somme de deux variables gaussiennes indépendantes, qui reste une variable gaussienne, et le **mélange de deux populations gaussiennes**, qui n'est pas une population gaussienne.

- Un mélange constitué de
  - 2/3 d'individus dont la taille suit une loi normale de moyenne 160 cm et d'écart type 15 cm, de densité *f*
  - 1/3 d'individus dont la taille suit une loi normale de moyenne 130 cm et d'écart type 10 cm, de densité *g*

- suit une loi de moyenne (2/3)×160+(1/3)×130 = 150 cm,
  mais non gaussienne, de densité
  $h = (2/3)\, f + (1/3)\, g$.

- Sur la représentation graphique de la densité *h*, on peut apercevoir une double bosse : la distribution est bimodale.



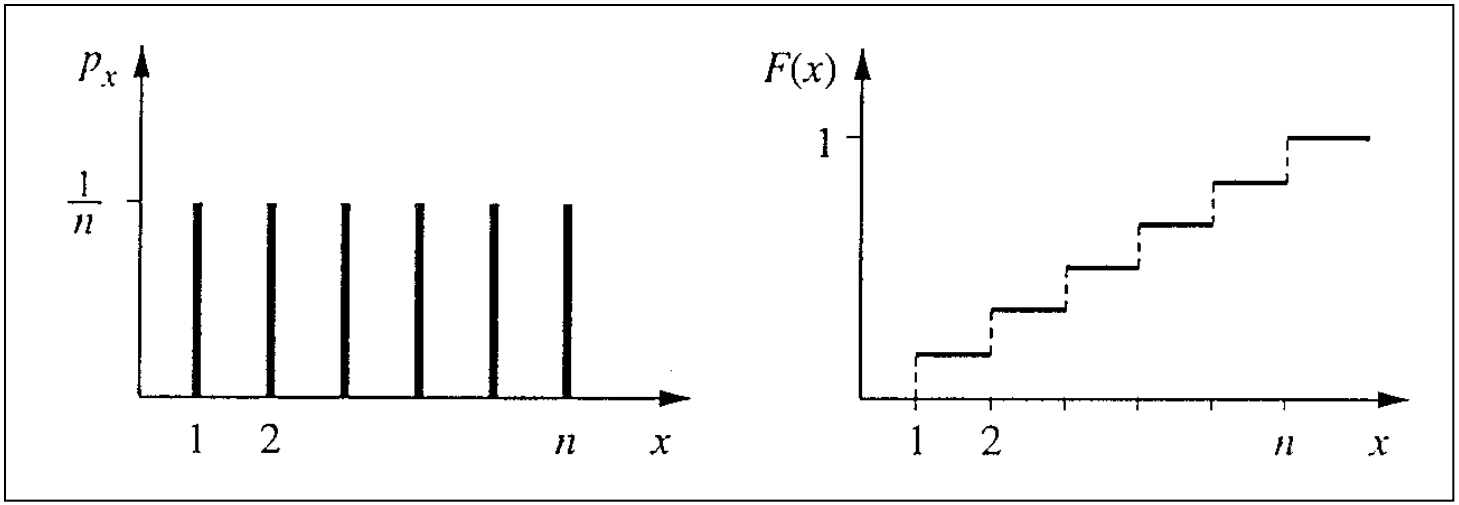
Sur la représentation graphique de la densité *h*, on peut apercevoir une double bosse: on dit que la distribution est **bimodale**.

Quelques autres distributions:

- distribution uniforme
- distribution exponentielle
- distribution log-normale
- distribution de Weibull

La distribution uniforme

- Exemple: point obtenu en lançant un dé:
- Définition: $x \in \{1, 2, \ldots, n\} \quad p_x = \frac{1}{n}$
- Ces distributions sont en principe **bornées**.



La distribution uniforme

La *distribution uniforme dans l'intervalle* $[a, b]$ est définie par sa densité

$$f(x) = 1/(b-a) \qquad \text{si } x \in [a, b],$$
$$= 0 \qquad\qquad \text{dans le cas contraire.}$$

La moyenne de cette distribution est $(a+b)/2$ et sa variance $(b-a)^2/12$.

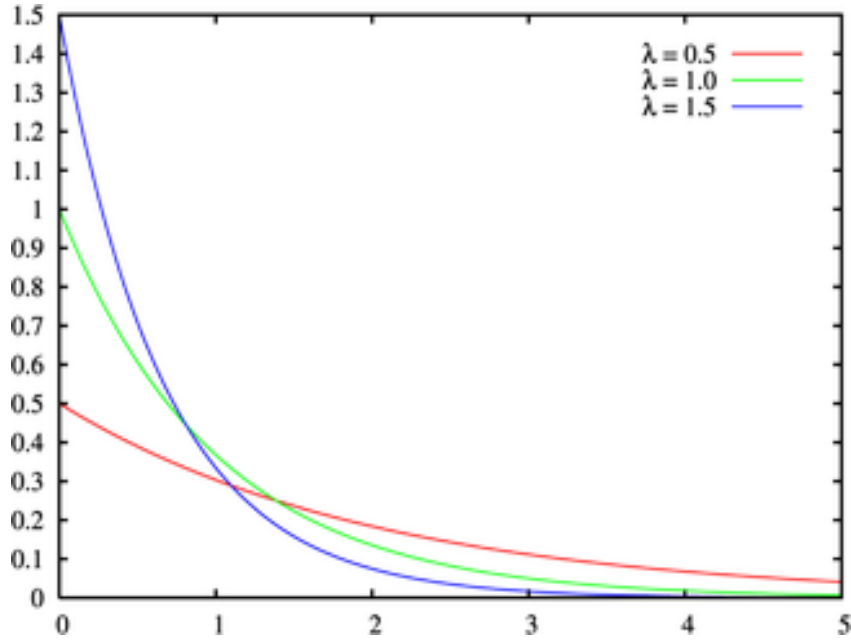
La distribution exponentielle

Une loi exponentielle correspond au modèle suivant:

X est une variable aléatoire définissant la durée de vie d'un phénomène.
Si l'espérance de vie du phénomène est *E*(*X*) et si la durée de vie est sans vieillissement, c'est-à-dire si la durée de vie au-delà de l'instant T est indépendante de l'instant T, alors *X* a pour densité de probabilité :

$$f(t) = 0 \text{ si } t < 0$$

$$f(t) = \frac{1}{E(X)} e^{-\frac{t}{E(X)}} \quad \text{pour tout } t \geq 0.$$



X suit donc une **loi exponentielle de paramètre**

$$\lambda = \frac{1}{E(X)}$$

Cette propriété traduit l'absence de mémoire de la loi exponentielle.

Par exemple, la probabilité qu'un phénomène se produise entre les temps t et $t+s$ s'il ne s'est pas produit avant est la même que la probabilité qu'il se produise entre les temps 0 et s.

On peut oublier l'instant de départ pour modéliser la probabilité.

Cette caractérisation est importante car elle permet de montrer que certains phénomènes peuvent être modélisés par une distribution exponentielle.

Cette loi permet entre autres de modéliser la durée de vie de la radioactivité ou d'un composant électronique.

La durée de vie **moyenne** $\frac{1}{\lambda}$ s'appelle le **temps caractéristique** et

$\frac{1}{\lambda}$ est aussi égal à l'**écart-type**.

Distributions de type log-normal

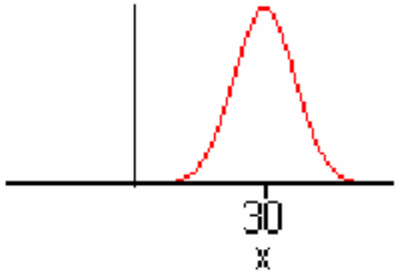
Une distribution normale s'étale en principe de $-\infty$ à $+\infty$

On a pu voir que les valeurs possibles d'une variable aléatoire normale étaient l'ensemble des nombres réels. Pour une situation réelle ne pouvant prendre des valeurs négatives, on peut malgré tout utiliser une loi normale lorsque la moyenne et l'écart type sont tels que la probabilité théorique d'avoir une valeur négative est à toute fin pratique nulle.

Exemple :

Prenons une loi normale X de moyenne $\mu = 30$ et d'écart type $\sigma = 10$.

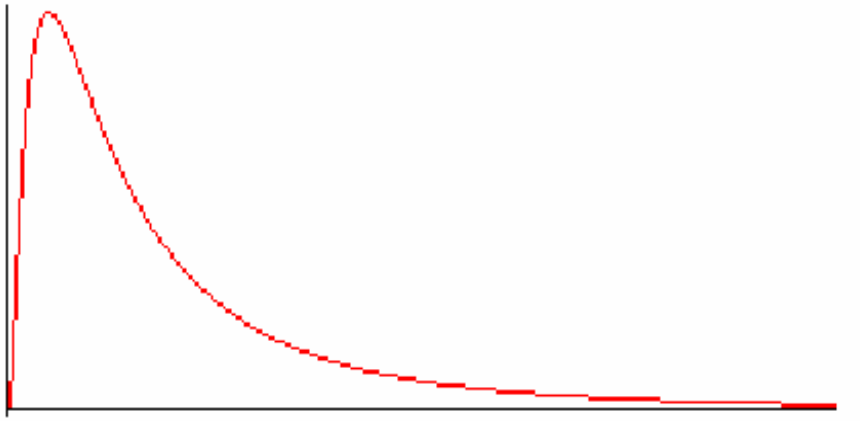
$$P(X<0)=P\left(\frac{X-\mu}{\sigma}<\frac{0-30}{10}\right)=P(Z<-3)=0{,}0013$$



La distribution log-normale

Par contre, pour une situation réelle ayant une moyenne très faible (ce qui est souvent le cas d'analyses en météorologie ou d'analyses sur la concentration de produits chimiques), il est très difficile d'oublier ce fait. De plus, les situations ne pouvant prendre des valeurs négatives et ayant une moyenne très faible ont tendance à présenter une distribution dissymétrique, donnant ainsi une proportion d'événements extrêmes plus grande que celle prévue par la loi normale.

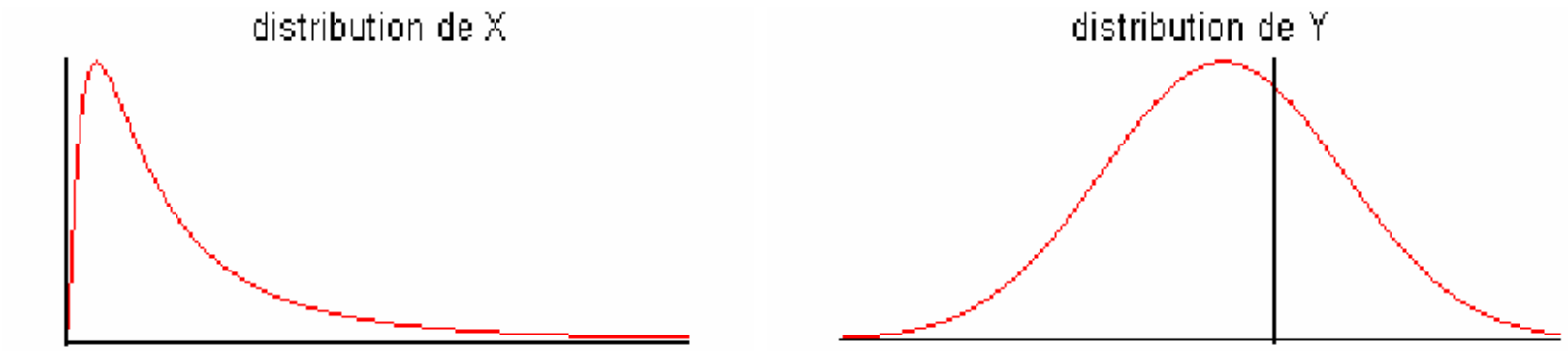
Voici une courbe typique d'une telle situation :



La distribution log-normale

Puisque nous avons affaire à une situation ne prenant pas de valeurs négatives, il est possible de calculer le logarithme de ces valeurs dans une base quelconque et ces nouvelles valeurs se distribuent sur tous les réels, les valeurs comprises entre 0 et 1 ayant des valeurs logarithmiques négatives et les valeurs supérieures à 1 ayant des valeurs logarithmiques positives. Ce raisonnement est

Une distribution de probabilité d'une variable aléatoire **X est** dite **log-normale si** la distribution de probabilité de la variable aléatoire **Y = ln X est normale**.



Soit X la variable aléatoire log-normale

Si Y = ln X est une variable aléatoire normale ayant

moyenne $\mu_Y = \mu$ connue,

variance $\sigma_Y^2 = \sigma^2$ connue,

écart type $\sigma_Y = \sigma$ connu,

alors

la fonction de densité de la variable aléatoire X est

$$f(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \dfrac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} & \text{si } x > 0 \end{cases}$$

La distribution log-normale

Il est possible algébriquement d'inverser les calculs.

Soit Y une variable aléatoire normale ayant

moyenne $\mu_Y = \mu$,

variance $\sigma_Y^2 = \sigma^2$,

écart type $\sigma_Y = \sigma$,

Si $X = e^Y$ est une variable aléatoire log-normale ayant

moyenne μ_X connue ,

variance σ_X^2 connue ,

écart type σ_X connu,

alors

$$\mu = 2\ln\mu_X - \frac{1}{2}\ln(\mu_X^2 + \sigma_X^2) ,$$

$$\sigma^2 = \ln(\mu_X^2 + \sigma_X^2) - 2\ln\mu_X ,$$

$$\sigma = \sqrt{\ln(\mu_X^2 + \sigma_X^2) - 2\ln\mu_X} ,$$

Exercice sur la distribution log-normale:

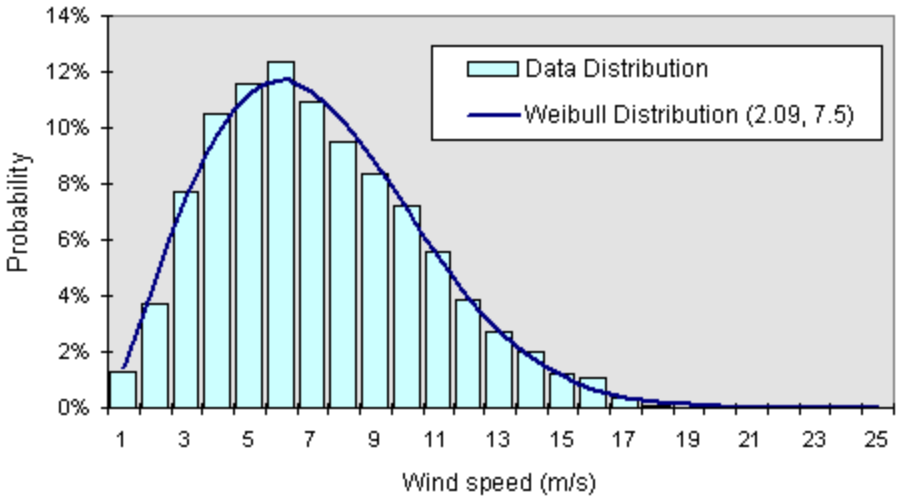
Soit X une variable aléatoire **log-normale** de

- moyenne arithmétique **10** et
- d'écart type **1,8**.

Calculer la moyenne et l'écart type de la variable aléatoire $\boldsymbol{Y = ln\,(X)}$ correspondante.

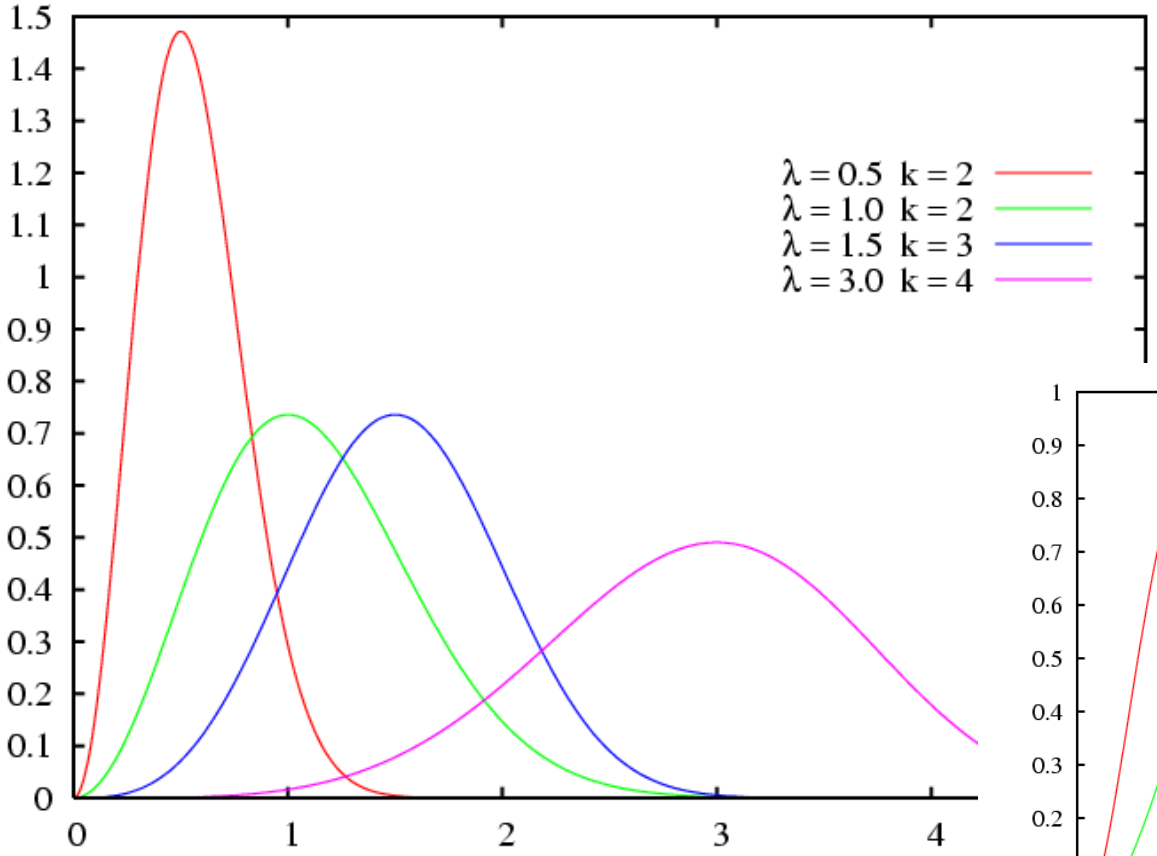
Calculer et faire le graphique de la densité de probabilité et de la fonction de répartition.

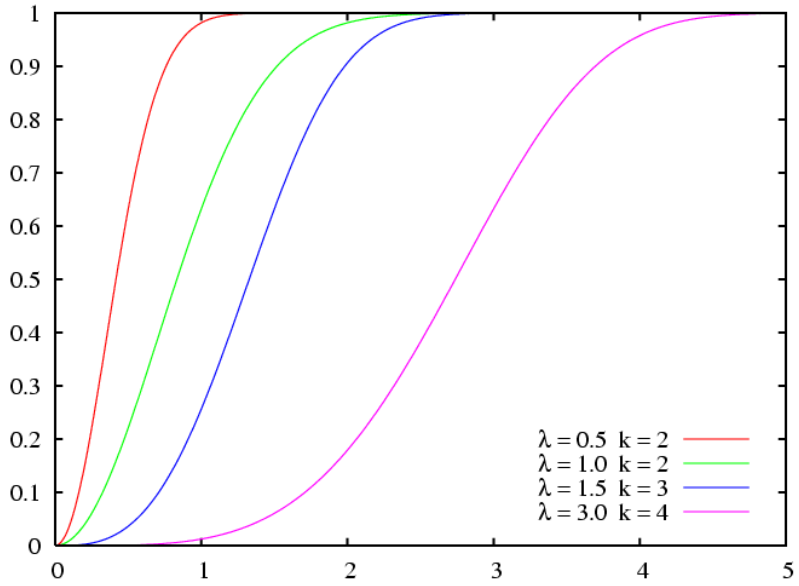
Distribution de Weibull



- La loi de Weibull recouvre en fait toute une famille de lois avec des distribution à valeurs positives (ou, plus généralement mais moins fréquemment, à valeurs supérieures à une valeur donnée), plus générale que la formulation log-normale.
- Certaines d'entre elles apparaissant en physique comme conséquence de certaines hypothèses.
- Dans d'autres cas une loi de Weibull constitue surtout une **approximation particulièrement utile et flexible**, alors qu'il serait très difficile et sans grand intérêt de justifier une forme particulière de loi.
- La distribution de Weibull peut aussi reproduire le comportement d'autres lois de probabilités, telle les lois normale, log-normale et exponentielle.

Il est alors possible de trouver dans la famille de Weibull une loi qui ne s'éloigne pas trop des données disponibles en calculant deux constantes k et λ, à partir de la moyenne et la variance (carré de l'écart-type) observées.





Distribution de Weibull

- Densité de probabilité $f(x;k,\lambda) = (k/\lambda)(x/\lambda)^{(k-1)} e^{-(x/\lambda)^k}$

  Où $k > 0$ est le paramètre de forme et

  $\lambda > 0$ le paramètre d'échelle de la distribution

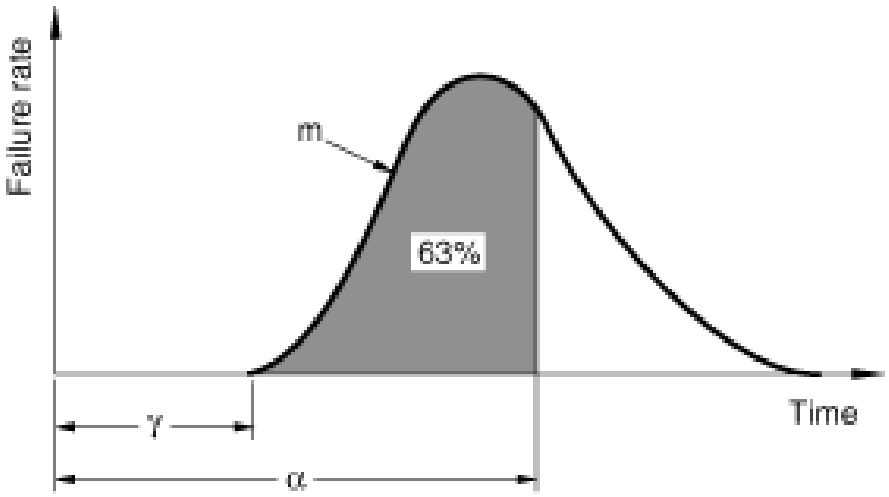
- Fonction de répartition $F(x;k,\lambda) = 1 - e^{-(x/\lambda)^k}$
- Moyenne (espérance) $\mu = \lambda\Gamma\left(1+\frac{1}{k}\right)$
- Variance $\sigma^2 = \lambda^2\Gamma\left(1+\frac{2}{k}\right) - \mu^2$

La distribution de Weibull est souvent utilisée dans le domaine de l'analyse de la durée de vie

- Dans ce cas l'axe des x dans la densité de probabilité et la fonction de répartition représente une échelle de temps.

- Le taux de panne $\boldsymbol{h}$ est donné par:

$$h(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}.$$



- Si le taux de panne diminue au cours du temps alors: $k < 1$
- Si le taux de panne est constant dans le temps alors: $k = 1$
- Si le taux de panne augmente avec le temps alors: $k > 1$

La compréhension du taux de panne peut fournir une indication au sujet de la cause des pannes.

- Un **taux de panne décroissant** relève d'une "mortalité infantile". Ainsi, les éléments défectueux tombent en panne rapidement, et le taux de panne diminue au cours du temps, quand les éléments fragiles sortent de la population.

- Un **taux de panne constant** suggère que les éléments tombent en panne à cause d'évènements aléatoires.

- Un **taux de panne croissant** suggère une "usure" : les éléments ont de plus en plus de chances de tomber en panne quand le temps passe.

Rappel des principales définitions

- **Population**: ensemble des objets de l'étude (ex. pièces mécaniques)
- **Individu**: élément de la population (ex. chaque pièce mécanique)
- **Échantillon**: partie de la population
- **Taille**: nombre d'individus dans la population/échantillon
- **Variable**: application associant à chaque individu un caractère (ex. valeur d'une cote mesurée).
- **Espérance**: meilleure estimation d'une quantité, en générale la moyenne d'une série.

- **Ecart-type**: valeur indiquant la dispersion des mesures
- **Variance**: carré de l'écart-type.
- **Histogramme**: graphique décrivant la probabilité qu'un individu aie une valeur à l'intérieur de plages (fourchettes) données.
- **Distribution de probabilité**: fonction continue décrivant la probabilité qu'un individu aie une valeur donnée.
- **Fonction de répartition**: fonction décrivant la probabilité qu'un individu aie une valeur **inférieure à** un chiffre donné.
- **Intervalle** ou **niveau de confiance**: probabilité que la mesure se trouve à l'intérieur d'une fourchette d'incertitude donnée, par exemple $\pm 2\sigma$.