GENERALITES

1.1 - TERMINOLOGIE.

- ensemble d'éléments, d'objets --------> **population (N)** (population finie ou infinie)

- sous-ensemble ==>"**échantillon**"

- **données ==>valeurs d'un caractère(** sexe, poids, longueur **)**sous forme de **variables**(continues, discrètes)

- regroupement des données en **suites** classées

- EX. : masses des étudiants d'une promotion de MP.

- **étendue du caractère** : 56 - 87 (en kg)

tailles	classes	effectifs de classes	effectifs cumulés	fréquences de classes	fréquences cumulées
56-59	k1	6	6	6,82	6,82
60-63	k2	10	16	11,36	18,18
64-67	k3	12	28	13,64	31,82
68-71	k4	18	46	20,45	52,27
72-75	k5	17	63	19,32	71,59
76-79	k6	13	76	14,77	86,36
80-83	k7	7	83	7,95	94,31
84-87	k8	5	88	5,68	99,99

- **classes (ki)** : (56-59) ; (60-63); (64-67), etc.. . ("catégories")

- **étendue des classes** : 4 kg

- **effectifs des classes** : (n_i)

- on rassemble des données en :

  - **tableaux** (tableaux des effectifs,....)
  - **graphes**

- répartition des effectifs par classes :**==>"distribution des effectifs"**

- effectifs relatifs ou fréquences relatives ou **fréquences==> distribution des fréquences** ou des pourcentages

- **effectifs cumulés** : effectifs des classes d'indice j inférieur ou égal à i.

1.2 - DISTRIBUTION A UN SEUL CARACTERE. REPRESENTATION GRAPHIQUE.

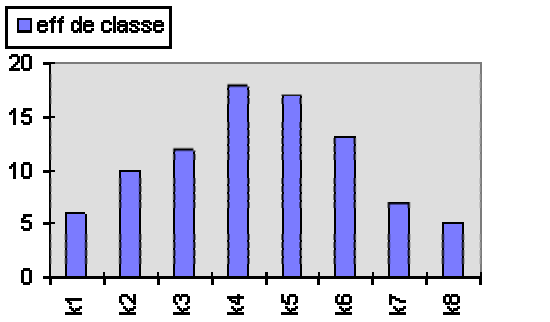
Mode opératoire:

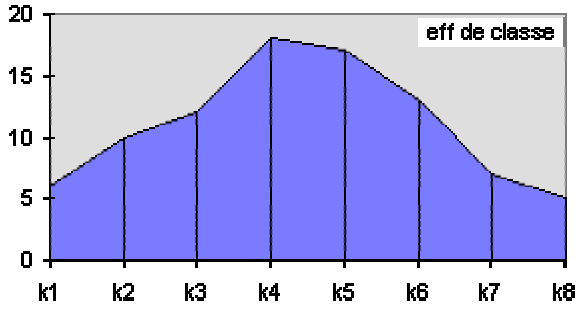
i) on définit l'étendue de la population

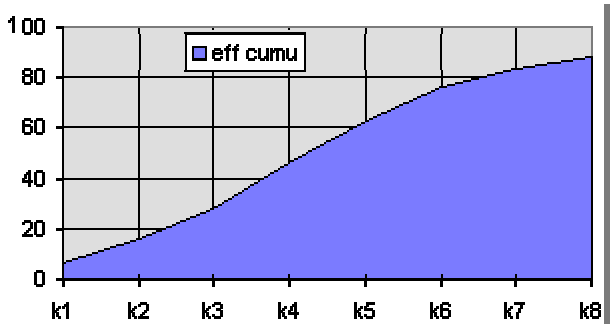
ii) on définit les classes

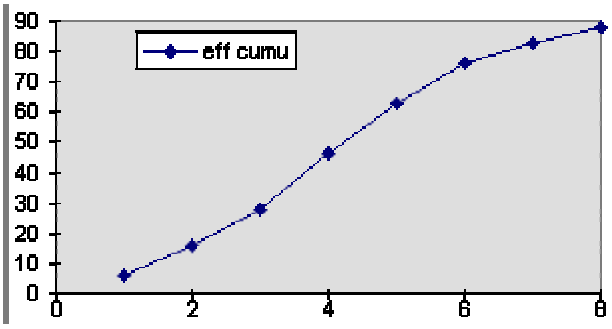
iii) on représente au mieux en fonction des renseignements requis

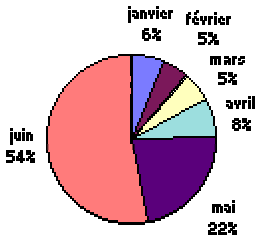
- HISTOGRAMMES -POLYGONES DES EFFECTIFS-DIAGRAMMES EN BATONS

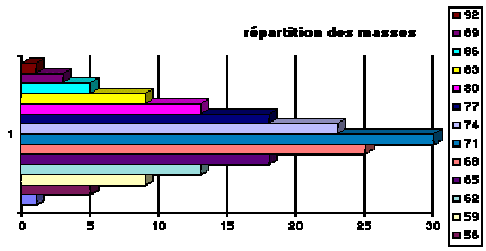












1.3 - CARACTERISTIQUES DES DISTRIBUTIONS

Variable X , valeurs (X_i) , population N , effectifs de classes (n_j).

1.3 .1 - Centrage

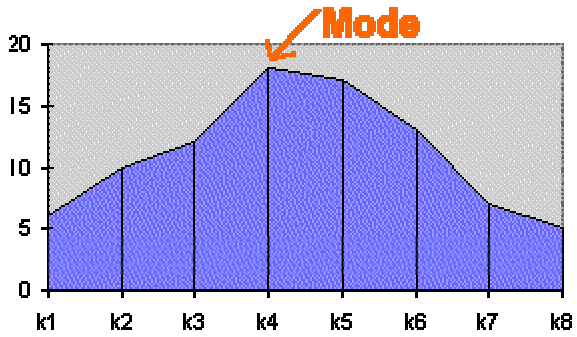
i) **moyenne arithmétique**

variables non classées	variables classées
$\mu = (\Sigma_i X_i) / N$	$\mu = (\Sigma j\ n_j\ X_j) / \Sigma j\ n_j$

C'est le paramètre de centrage le plus utilisé et le plus représentatif.

ii) **médiane** : c'est la valeur de X qui partage la population en 2 sous populations égales.

iii) **mode** : c'est la valeur de la variable correspondant à la fréquence maximale.



iv) **autres** : moyenne géométrique, harmonique, quadratique(voir en annexe).

1.3.2 - Dispersion

i) **étendue** : $W = X_{max} - X_{min}$

ii) variance : **v**

variables non classées	variables classées
$v = (\Sigma_i (X_i - \mu)^2) / N$	$v = (\Sigma_j n_j (X_j - \mu)^2) / \Sigma_j n_j$

iii) écart-type :

$$\sigma = v^{1/2}$$

iv) autres : écart moyen : c'est la moyenne des écarts arithmétiques ; coefficient de variation (C V).

REMARQUES :

i) considérons deux ensembles N_1 et N_2 de même moyenne et d'écarts types s_1 et s_2 ; la variance de l'ensemble $N_1 + N_2$ est : $s^2 = (N_1 s_1^2 + N_2 s_2^2)/(N_1 + N_2)$.

ii) deux ensembles 1 et 2 de même population N, de moyennes et écarts types respectifs : μ_1, μ_2, $s_1$et s_2, ont pour moyenne d'ensemble : $\mu = (\mu_1 + \mu_2)/2$ et pour variance commune : $s^2 = s_1^2 + s_2^2$.

LES PROBABILITES

DEFINITIONS

THEOREMES FONDAMENTAUX

RAPPELS D'ANALYSE COMBINATOIRE

LES DISTRIBUTIONS USUELLES

2.1 - DEFINITIONS

2.1.1 - Définition classique (mathématique)

p = Pr(E) = *Nombre de cas favorables / Nombre de cas possibles.*

* Pr(E*) probabilité de **non - réalisation** de E(E*) : **Pr(E*) = 1 - Pr(E)**

2.1.2 - Définition statistique

* Dans le calcul de p, on a supposé tous les cas possibles **équiprobables**. On aurait pu faire une étude empirique (c'est à dire réaliser "expérimentalement" un grand nombre de lancers de dés) et estimer**, à partir de l'observation,** la fréquence de divers événements favorables. Cette **fréquence de réalisation** correspondrait alors exactement à la **probabilité** de réalisation définie ci dessus, dès lors que l'hypothèse d'équiprobabilité des cas est vérifiée. D'où pour nous l'équivalence entre : fréquence relative de réalisation (expérimentale/constatée) et probabilité de réalisation (théorique /mathématique)

2.1.3 - Variable aléatoire - Fonction de répartition - Densité de probabilité

* X variable **aléatoire ;** à chaque valeur (Xi) on associe Pr (Xi).

* L'ensemble des couples (Xi ; Pr(Xi)) définit la **distribution de probabilités de X.**

* La fonction : $Fr(Xi) = Pr(\propto < X < Xi) = \sum_{j<i} Pr(X_j)$ est appelée **fonction de répartition** ("distribution des fréquences cumulées"). Lorsque la variable Xi est continue, on définit aussi : **Pr(X) = f(X) dX** , où f(X) est appelée **densité de probabilité de X** (à rapprocher de la densité de fréquence).

2.1.4 - Evénements dépendants / indépendants. Probabilité conditionnelle.

* évènements **indépendants :** la réalisation d'un événement n'affecte pas la réalisation du/des autres (ex : apparition des chiffres lors de jets successifs d'un dé)

*événements **dépendants :** la réalisation d'un événement affecte celle du/des autres (ex : le tirage sans remise d'une carte d'un jeu de 32 cartes réduit à $1/31^{ème}$ la probabilité d'apparition d'une autre carte lors d'un second tirage).

* **Probabilité conditionnelle.** Elle est relative aux événements dits **dépendants** :

Pr (E_2 / E_1) est la probabilité de réalisation de E_2 , **sachant que E_1 est réalisé** (ou "si E_1 est réalisé").

2.2 - THEOREMES FONDAMENTAUX

2.2.1 - Probabilités composées : Pr(E_1 et E_2) ou Pr(E_1 U E_2) ou Pr(E_1,E_2).

* Pr(E_1 et E_2) probabilité pour que E_1 **et** E_2 soient réalisées (probabilité composée).

i) Si E_1 et E_2 sont **indépendants :**

$$\mathbf{Pr(E_1 \text{ et } E_2) = Pr(E_1,E_2) = Pr(E_1) \times Pr(E_2)}$$

ii) Si E_1 et E_2 **sont dépendants :**

$$\mathbf{Pr(E_1 \text{et } E_2) = Pr(E_1) \times Pr(E_2/E_1) = Pr(E_2) \times Pr(E_1/E_2)}$$

2.2.2 - Probabilités totales : Pr(E_1 ou E_2)

i) Si E_1 et E_2 sont **incompatibles** (ne peuvent être réalisés "en même temps" (donc Pr(E_1 et E_2) = 0):

$$\mathbf{Pr(E_1 \text{ou } E_2) = Pr(E_1) + Pr(E_2)}$$

ii) Si E_1 et E_2 sont **compatibles** (peuvent se réaliser en même temps):

$$\mathbf{Pr(E_1 \text{ou } E_2) = Pr(E_1) + Pr(E_2) + Pr(E_1,E_2)}$$

Remarque : **Pr(E1 ou E2) = 1 - Pr (E1 et E2).**

2.3 - RAPPELS D'ANALYSE COMBINATOIRE

2.3.1 - Combinaisons de n objets groupés p à p

$$C_n^p = n! / p! (n - p)!$$

2.3.2 - Arrangements de n objets, nombre de suites différentes de p objets

$$A_n^p = n!/(n-p)!$$ (**attention , l'ordre compte !)**

2.3.3 - Permutations de n objets ; nombre de suites différentes de ces n objets(l'ordre ne compte pas).

$$P_n = C_n^n = n!$$

2.4 - LES DISTRIBUTIONS USUELLES

Distributions de probabilités **théoriques** de variables continues ou discrètes

2.4.1 - Distribution binomiale : Bi (n,p)

* **Définition** :

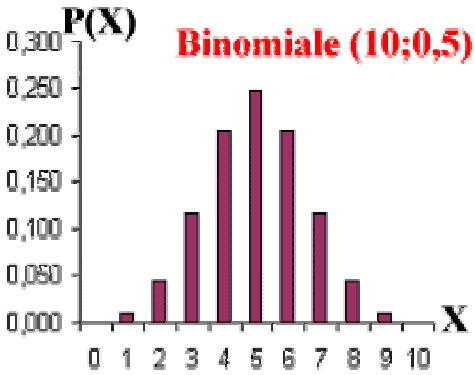
- on considère des épreuves à deux alternatives, aboutissant à la réalisation, ou la non réalisation, d'un évènement E

- p est la probabilité de réalisation de E et q = (1 - p) la probabilité de non réalisation

- variable aléatoire **discrète** X

- on définit la probabilité de X réalisations de l'événement E en n expériences? (X entier appartenant à l'intervalle fermé (0, n) :

$$P(X) = C_n^x p^x q^{n-x}$$



REMARQUE : les différentes valeurs de P(X) pour X = 0, .., n sont les termes du "développement du binôme " : $(p + q)^n = p^n + C_n^1\ p^{n-1}\ q + C_n^2\ p^{n-2}\ q^2 + .+ C_n^j\ p^{n-j}\ q^j + ...$ .Les C_n^j sont appelés **coefficients binomiaux.**

*** Propriétés**

centrage ----------> moyenne = np
dispersion ---------> variance = σ^2 = n.p.q

RAPPEL : on connaît ici le **nombre total n** d'événements, ce qui n'est pas le cas pour la distribution de Poisson .

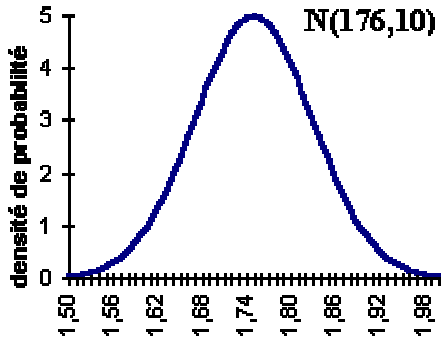
2.4.2 - Distribution de GAUSS ou "NORMALE" : $N(\mu, \sigma)$

*** Définition :**

- variable **continue X**

- valeur moyenne : μ ; écart type : σ

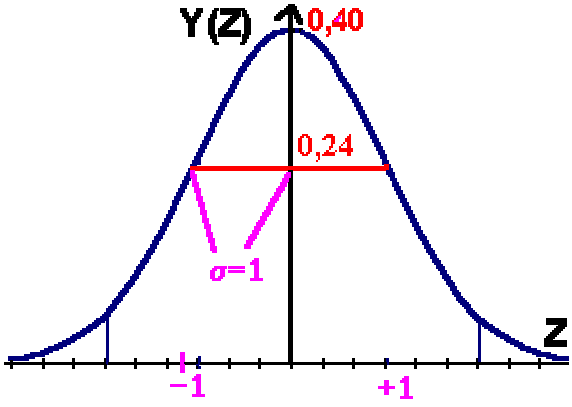
$$Y(X) = \exp(-(X-\mu)^2/2\sigma^2) / (2\pi)^{1/2}\ \sigma$$



- on introduit la variable **réduite centrée : Z = (X-μ)/σ**

- d'où la nouvelle "distribution **normale centrée":**

$$Y\ (\ Z\ ) = \exp(-z^2/2) / (2\pi)^{1/2}$$



- on a toujours : $\int_{-\infty}^{\infty} Y(X)\,dX = 1 = \int_{-\infty}^{\infty} Y(Z)\,dZ$

*** Propriétés**

centrage ----------> moyenne = μ
dispersion ----------> variance = σ^2

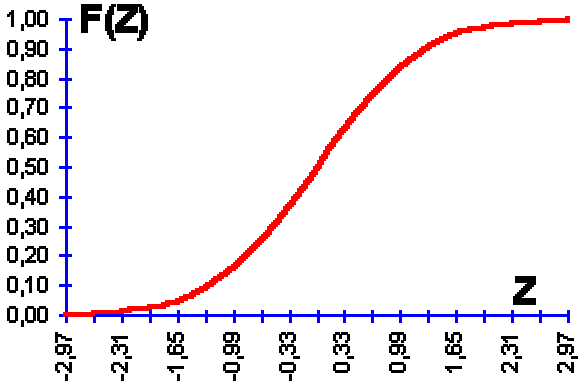
*** REMARQUES**

i)Changer X en (X - μ) a **centré** la distribution sur 0, elle est **symétrique.**

ii)Changer **(X - μ)** en **(X- μ)/σ** a **réduit** l'échelle sur l'axe des abscisses ; son écart type est maintenant égal à 1.

iii) Des tables donnent : Y(Z) , P(Z) et F(Z)

telles que : $P(Z) = \int_{-Z}^{Z} Y(Z')\,dZ'$ et $F(Z) = \int_{-Z}^{Z} Y(Z')\,dZ'$



iv) table succincte de Y(Z) et P(Z) :

Z	0,000	1,000	2,000	3,000	∞

Y(Z)	**0,399**	**0,242**	**0,054**	**0,004**	**0,000**
P(Z)	**0,000**	**0,682**	**0,954**	**0,997**	**1,000**

La probabilité de trouver Z :

- entre (-1 et +1) est de 68,2 %

- entre (-2 et +2) est de 95,4 %

- entre (-3 et +3) est de 99,7 %

à l'inverse, la probabilité d'avoir Z extérieur à [-1,+1] est de 31,8 % ; extérieur à Z [-2,+2] est de 4,6 % et extérieur à Z [-3,+3] est de 0,3 %.

iv) Considérons k échantillons de n éléments issus de la population N **normale** (μ,σ).

Chaque échantillon i a pour moyenne et écart type : X_i et s_i

- l'ensemble des Xi a une répartition **normale**

* de moyenne : **moy(Xi) = μ = moy(X)**

* d'écart-type : **var(Xi) = σ^2 / n = var(X) / n**

v) si X appartient à la distribution normale N(μ,σ)alors **a**X appartient à la distribution normale : N (**a**μ,**a**σ)

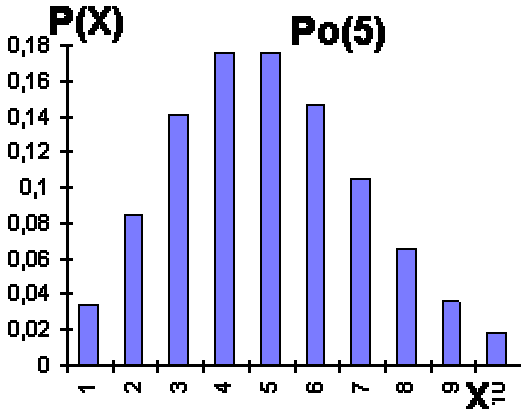
- si différentes variables Xi, Yi, Zi appartiennent à des distributions normales, alors la variable (Xi + Yi + Zi) appartient à la distribution normale :

$$N(\mu_X + \mu_Y + \mu_Z ; (v_X^2 + v_Y^2 + v_Z^2)^{1/2})$$

2.4.3 - Distribution de Poisson : $P_o(\lambda)$

* Distribution à **un seul paramètre** λ , relative à une **variable discrète**.

$$P(X) = \lambda^X e^{-\lambda} / X!$$



*** Propriétés :**

centrage ----------> moyenne = λ
dispersion ----------> écart-type : $\sigma = \lambda^{1/2}$

* P(X) ne dépend que du seul paramètre **λ.**

* La loi de Poisson ne s'applique qu'à **une seule alternative et avec une probabilité de réalisation faible (p << 1) ;** on a alors **λ = Np.** On notera également que le nombre total d'épreuves n'est pas connu et ceci contrairement à la distribution binomiale.

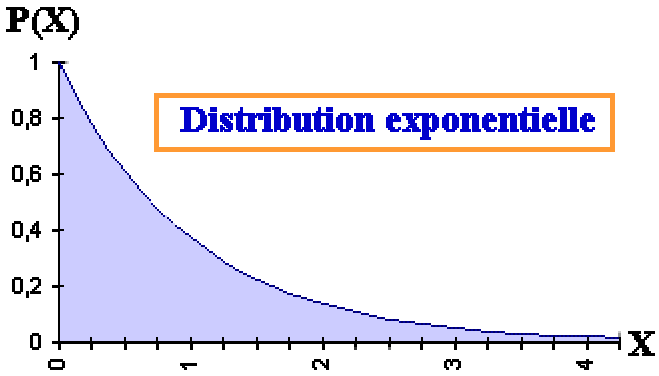
2.4.4 - Distribution exponentielle (Loi de Laplace)

* variable X **continue :** $0 < X < +\infty$

*** Définition :**

- un seul paramètre λ (moyenne = $1 / \lambda$)

- elle caractérise des processus temporels **sans mémoire**



$$P(X) = e^{-\lambda X}$$

*** Propriétés**

centrage----------> moyenne = 1 / λ
dispersion----------> écart-type = 1 / λ

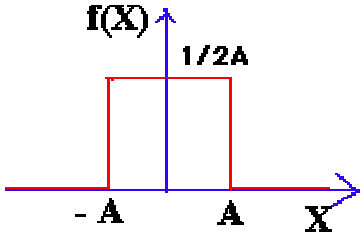
2.4.5 - Distribution uniforme.

* variable X **continue :** $-\infty < X < +\infty$

*** Définition :**

- un seul paramètre **A** (inverse de la moyenne)

- **f(X) = 1/2A** pour -A < X < + A
- **f(X) = 0** ailleurs



*** Propriétés**

centrage----------> moyenne = 0
dispersion----------> v(X) = $A^2/3$

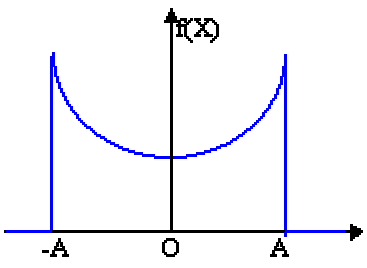
2.4.5 - Distribution en U.

* variable X **continue :** $-\infty < X < +\infty$

*** Définition :**

- un seul paramètre **A**

- **$f(X) = (1/2\pi A) / (1-X^2/A^2)^{1/2}$** pour -A < X < + A
- **f(X) = 0** ailleurs



*** Propriétés**

centrage----------> moyenne = 0
dispersion----------> v(X) = $A^2/2$

2.4.7 - Remarques. Relations entre distributions.

i) **N grand et X variable discrète :**

- Si p # 0 et N grand (> 50) tels que **Np > 5**

Loi binomiale ----------> Loi de Gauss

$$Bi(N, p) \text{----------} > N(\lambda , \lambda^{1/2}) ; (\lambda = Np)$$

- Si p >>0 et N > 50 tels que **Np < 5**

Loi binomiale ----------> Loi de Poisson

$$Bi(N, p) \text{----------} > Po\ (Np)$$

1. ii) λ croît indéfiniment :

Loi de Poisson ----------> Loi de Gauss

$$Po\ (\lambda) \text{----------} > N(\lambda , \lambda^{1/2})$$

Statistique Inférentielle

Généralités	Tests d'hypothèses :
Echantillonnage	- grands échantillons
Estimation	- petits échantillons
Décision statistique	- raccordement

Généralités

L'objectif de la statistique inférentielle est de fournir des résultats relatifs à une population à partir de mesures statistiques réalisées sur des échantillons.

- Ainsi estime-t-on la moyenne μ, ou l'écart type σ , ou la fréquence p d'une population, à partir de la moyenne X_i de l'écart type s_i , ou de la fréquence f_i d'un échantillon.
- Ainsi détermine-t-on si un échantillon étudié appartient à une population de paramètre caractéristique connu..
- Ainsi recherche-t-on si deux échantillons étudiés sont issus de deux populations de même paramètre caractéristique.
- Ainsi détermine-t-on la taille que doit avoir un échantillon si l'on désire qu'il fournisse une estimation d'un paramètre de population avec une "précision"définie à priori.

ECHANTILLONNAGE

- 1 - Définitions.

- échantillonnage **aléatoire.** Pour qu'un échantillon soit représentatif de la population, il faut que chaque élément de la population ait les mêmes chances d'appartenir à cet échantillon. On dit que dans ce cas on a **un échantillonnage aléatoire**.
- échantillonnage **exhaustif ou non exhaustif.** Si l'élément extrait de la population, pour effectuer l'échantillonnage, est remis dans cette population après relevé de ses caractéristiques ---> échantillonnage **non exhaustif** , sinon -->échantillonnage exhaustif.

- **REMARQUES IMPORTANTES :**
  - une population finie sur laquelle on effectue un échantillonnage non exhaustif peut être considérée comme infinie.
  - un échantillonnage exhaustif réalisé sur une population très grande (N >> n) est considérée comme non exhaustif.

- 2 - Distributions d'échantillonnage.

- Chaque échantillon donne des paramètres statistiques (moyenne, écart-type, fréquence,...) : $\mathbf{X_i}$, $\mathbf{s_i}$, $\mathbf{f_i}$.
- Les ensembles de paramètres d'échantillons constituent des : **distributions d'échantillonnage**.

(on parle de distribution d'échantillonnage de la statistique S, où S correspond à la moyenne d'échantillon Xi , ou à l'écart type d'échantillon s_i , ou à la fréquence d'échantillon f_i , etc.).

- 3 - Grands et petits échantillons.

- **On peut montrer que :**

si n > 30	**si n < 30**
les conclusions obtenues à partir de ces échantillons sont identiques à celles résultant d'échantillons issus d'une population normale de mêmes paramêtres (μ,σ,p,...) que la population étudiée (normale ou non) ; on parlera ici de **"grands échantillons"**	on parle de **"petits échantillons"** et les statistiques relatives aux paramètres d'échantillons ne correspondent plus à des distributions normales. On doit introduire ces nouvelles distributions.
n > 30 <==> "grands échantillons"	**n > 30 <==> "petits échantillons"**

- 4 - Théorèmes généraux.

- **Théorèmes sur les moyennes**

- Les moyennes $\{X_1,X_2 , X_3 ,.., X_k\}$ d'échantillons de taille n, prélevés au hasard dans une population (N, μ, σ), sont distribuées avec la **moyenne $m(X) = \mu$ et la variance $v_X = \sigma^2 / n$.**
- La distribution des moyennes d'échantillons(ou "distribution d'échantillonnage des moyennes") est :
  - NORMALE pour n > 10, si la population mère est normale.
  - NORMALE pour n > 30, quelle que soit la population mère ("grand échantillon").

- **Théorème sur les proportions**

- Les fréquences (f_1, f_2,..., f_k) d'une caractéristique donnée, dans des échantillons de taille n, prélevés au hasard dans une population où p est la proportion des éléments possédant cette caractéristique, sont distribuées suivant la **binomiale** de moyenne **m(f) = p** et de **variance v = pq/n.**
- pour n<30, on applique la loi normale **N**(p,$(pq/n)^{1/2}$)

ESTIMATION

* Dans la vie courante, on est souvent amené à **estimer** quelquechose : la valeur d'un produit, le poids d'un paquet, la route et le cap à suivre pour un navigateur, le temps de trajet d'une ville à une autre, etc.

* Cela peut consister tout simplement à fournir une valeur qui, à notre avis, est la valeur réelle.; on fait alors une **estimation ponctuelle**.

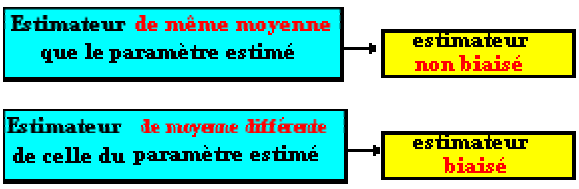
* Pour affiner l'estimation, on propose parfois, plutot qu'une valeur unique, un intervalle dans lequel la valeur estimée "a de grandes chances", de se trouver ; on fait une **estimation par intervalle**.

- 1 - Estimateur.

- 1-1 - Estimateur biaisé - efficace.

* un **estimateur** est un paramètre d'échantillon utilisé pour "estimer" la valeur d'un paramètre statistique de la population.

* si l'estimateur a **même moyenne** que le paramètre à estimer, on dit que cet estimateur est **non biaisé.** Dans le cas contraire, on dit qu'il est dit **biaisé**.



* exemple :

- si l'on prend la moyenne d'échantillon Xi comme estimateur de **μ,**on dira que c'est un estimateur non biaisé puisque la moyenne des Xi est égale à μ (cf théorème des moyennes).
- par contre, si on utilise l'écart type d'échantillon s_i comme estimateur de **σ**, le fait que la moyenne des s_i est **différente** de σ nous oblige à dire que s_i est un estimateur **biaisé**.

* de deux estimateurs non biaisés, le plus **efficace** est celui ayant la plus petite variance.

- 1-2 - Estimateur de l'écart type.

* On a déja vu que l'on peut prendre la moyenne X d'échantillon comme estimateur de la moyenne μ d'une population ; de plus cet estimateur est **non biaisé**.

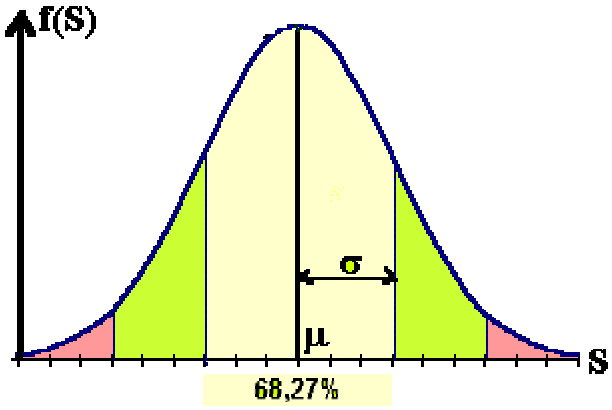
* En ce qui concerne l'estimation de l'écart type σ d'une population, on prendra aussi l'écart type s d'échantillon ; mais comme cet estimateur est **biaisé** (c'est à dire que sa moyenne n'est pas égale à σ), on doit le corriger en le multipliant par un terme correctif : $(n/(n-1))^{1/2}$ de telle sorte que **$s^* = s\, n^{1/2}/(n-1)^{1/2}$** ait pour moyenne σ.

* On notera que pour les échantillons de taille supérieure à 100, le terme correctif $n^{1/2}/(n-1)^{1/2} = 1{,}005$ est très proche de 1 ce qui peut justifier de prendre s* = s. En revanche , pour n = 30, limite basse prise plus haut pour la définition des "grands échantillons", le terme correctif vaut 1,017 et ne peut pas être **systématiquement** négligé.

- 2 - Intervalle de confiance

* considérons la statistique S (X_i ou s_i ou f_i) dont la distribution d'échantillonnage est **normale**. Le raisonnement développé ici restera cependant tout à fait valable pour toute autre distribution, dès lors qu'elle est tabulée.

* d'après la loi de distribution normale :



$\mu(S) - \sigma(S) < S < \mu(S) + \sigma(S)$ ----> **68,27%**
$\mu(S) - 2\sigma(S) < S < \mu(S) + 2\sigma\ S)$ ---> **95,45%**
$\mu(S) - 3\sigma(S) < S < \mu(S) + 3\sigma(S)$ --> **99,73%**

* Les intervalles symétriques par rapport au centre de symétrie(moyenne), sont appelés **"intervalles de confiance"** ; ainsi l'intervalle de confiance à 68,27 % est de largeur 2σ , celui à 95,45 % de largeur 4σ et celui à 99,73 % de largeur 6σ.

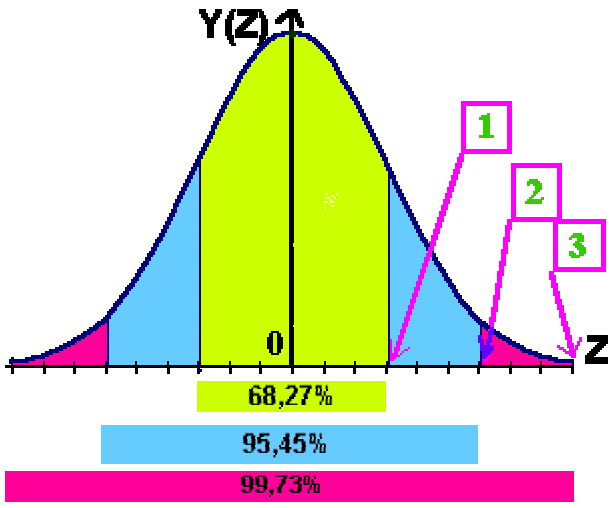
* Les valeurs limites de ces intervalles, appelées **"limites de confiance"**, sont respectivement : $\mu(S) \pm 1\sigma\ (S)$, $\mu(S) \pm 2\sigma\ (S)$, $\mu(S) \pm 3\sigma\ (S)$.

On les note : **$\mu(S) \pm Zc\ \sigma\ (S)$** , avec **Zc "coefficient de confiance"**.

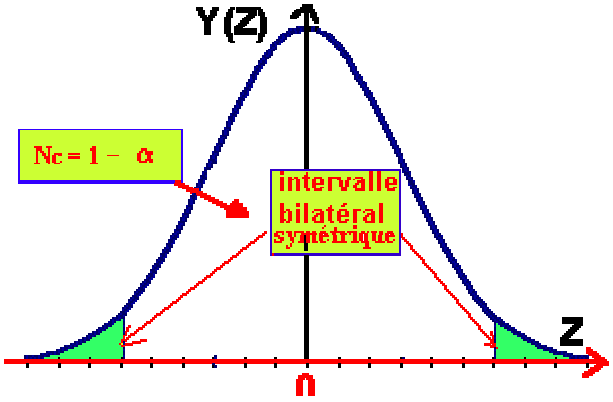
* Les valeurs : 68,3% , 95,4% ", 99,73% sont les **"seuils de confiance"** ou aussi **"niveaux de confiance"**(Nc).

* Tableau de correspondance entre les seuils de confiance(Nc) et les coeff. de confiance(Zc), dans le cas de **distributions normales et pour des intervalles de confiance bilatéraux symétriques.**

Seuil de confiance Nc	99,73%	99,00%	98.00%	96,00%	95,45%	95.00%	90,00%	68,27%	50,00%
Zc	3,00	2,58	2,33	2,05	2,00	1,96	1,64	1,00	0,67

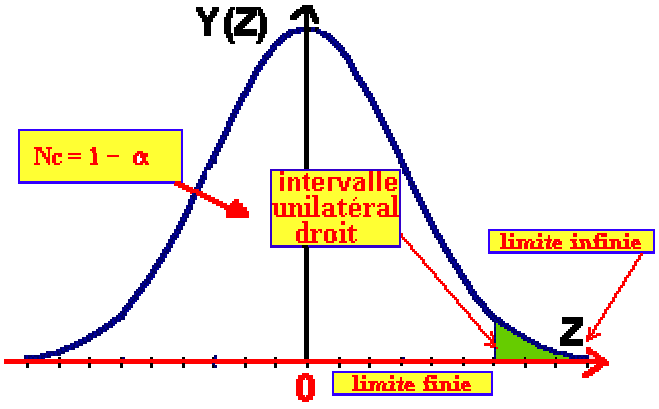


* On a défini ci dessus des intervalles de confiance **bilatéraux,** c'est à dire tels que la probabilité pour que la variable soit **hors de cet intervalle** est $\alpha = 1 - Nc$; de plus ces intervalles sont symétriques et ont des limites finies "centrées" sur la moyenne de la distribution.



* Si, "pour simplifier", on considère que toutes les valeurs de X sont comprises dans l'intervalle de confiance à Nc, on néglige d'office celles situées à l'extérieur de cet intervalle, dont la probabilité est égale à $\alpha = 1 - Nc$. **Cette probabilité α est le risque** pris en restreignant les valeurs de X à celles comprises dans l'intervalle de confiance.

* Il arrive aussi parfois que, pour un niveau de confiance donné Nc, on veuille restreindre la distribution des valeurs **d'un seul coté** de la distribution (à droite ou à gauche). On définit alors un intervalle **"unilatéral"**, limité d'un côté et comportant donc de l'autre côté une borne infinie.



* Le choix entre un intervalle bilatéral et un intervalle unilatéral dépend du problème

- 3 -Distributions se rapportant au centrage et à la dispersion des échantillons.

Centrage :

	Population mère **quelconque**	Population mère **normale**	**Intervalle** de confiance
GRANDS ECHANTILLONS	Loi de distribution des X **normale** $m(X) = \mu$ $\sigma_X = \sigma/n^{1/2}$	**IDEM**	$X = \mu \pm Z_c\sigma / n^{1/2}$
PETITS ECHANTILLONS	////////	**σ connu** σ appartient à une loi normale de moyenne $m(X) = \mu$ et d'écart type $\sigma_X = \sigma / n^{1/2}$	$X = \mu \pm Z_c\sigma/n^{1/2}$ Z_c paramètre de confiance associé au niveau de confiance Nc=1-α
	////////	**σ inconnu** $t = (X-\mu)(n-1)^{1/2}/ s$ suit la loi de STUDENT tabulée.	$\mu = X \pm t\ s /(n-1)^{1/2}$ t est fourni par une table (il dépend de α et de la population n de l'échantillon)

Dispersion

	Population mère **quelconque**	Population mère **normale**	**Intervalle** de confiance
GRANDS ECHANTILLONS	La loi de distribution des σ **est normale** : $N_s(\sigma ; \sigma /(2n)^{1/2})$ $m(s) = \sigma$; $\sigma_s = \sigma/(2n)^{1/2}$	**IDEM**	$s = \sigma +/- Z_c\sigma /(2n)^{1/2}$

PETITS ECHANTILLONS	///////	$\chi^2 = ns^2/\sigma^2$ **suit la loi de Pearson** Ceci conduit à un intervalle de confiance pour s dans lequel $\chi^2_{\alpha/2}$ et $\chi^2_{1-\alpha/2}$ sont fournis par une table numérique et dépendent de α et de n	$\sigma\ \chi_{\alpha/2}\ n^{-1/2}$ **< s <** $\sigma\ \chi_{1-\alpha/2}\ n^{-1/2}$

DECISION STATISTIQUE

Une "décision statistique" est prise après traitement de l'information brute fournie par l'échantillon.

MAIS :

De quel genre de décision s'agit-il ?

Comment est prise cette décision ?

- 1 - Quelle décision ?

* Par exemple :

- estimation de paramètres d'un lot de pièces livrées par un fournisseur à partir des paramètres fournis par des échantillons analysés (moyenne, écart-type, proportion...) et confirmation, ou non, des valeurs données par ce fournisseur.
- recherche si deux échantillons proviennent d'une même production de paramètres connus.
- détermination de la nature d'une population (apparentement à une distribution théorique donnée, nature aléatoire de la variable étudiée, ..).
- taille minimale d'un échantillon, si l'on désire une estimation plus ou moins serrée du paramètre de population.

3.3.2 Comment est prise la décision ?

- On met en place des **tests statistiques.** Ces tests sont des aides à la **prise de décision,** ils permettant de juger si une hypothèse avancée est **vraie** ou **fausse.**
- On procède, en trois phases, de la façon suivante :
    - on définit une **hypothèse à contrôler** appelée aussi **hypothèse nulle Ho** ; il lui correspond une **hypothèse alternative H1(**exemple : deux échantillons testés appartiennent ils à la même population ?). Cette hypothèse nulle Ho est destinée à être testée contre l'hypothèse alternative. On dira, après le test, que l'hypothèse Ho est "rejetée" ou "acceptée".
    - on détermine un critère **C** sur lequel on se basera pour juger. Plus précisément, on prendra un paramètre d'échantillon, ou une fonction plus ou moins simple des résultats expérimentaux, dont la valeur numérique servira à établir une comparaison.
    - on exprime un **jugement**.
        - pour cela on considère un **risque d'erreur** α , puis une répartition de ce risque (c'est à dire test bilatéral ou unilatéral). A ce risque d'erreur, donc au niveau de confiance 1- α , correspondra, dans la cas où l'hypothèse Ho est vraie, un intervalle de confiance de C, aux limites numériques connues, par exemple C_1 et C_2.
        - on juge alors, en fonction de la valeur expérimentale trouvée pour C, et des limites de cet intervalle de confiance :
            - si C est dans $[C_1,C_2]$, alors **on ne peut** refuser l'hypothèse Ho, et donc **on accepte Ho** (C est alors réputé être dans le" domaine d'acceptation").
            - si C est en dehors de $[C_1,C_2]$, alors **on rejette Ho** (C est dans le" domaine de rejet").
- Les risques d'erreur de jugement ; deux cas distincts d'erreur de jugement se présentent :
    - ou bien **on rejette Ho alors qu'elle est acceptable** ; c'est une "erreur de première espèce", dont le risque est α.
    - ou bien **on accepte Ho alors qu'elle doit être rejetée** ; c'est une" erreur de deuxième espèce", dont le risque est β.
        - β caractérise le **"risque client"** puisqu'il amène celui ci à accepter quelquechose qu'il devrait refuser (un élément hors norme par exemple).
        - par opposition α caractérise le **"risque fournisseur"** puisqu'il traduit le refus par le client de quelquechose qu'il aurait dû accepter .

LES TESTS STATISTIQUES

On verra successivement :

- **GRANDS ECHANTILLONS**
  - estimation de la moyenne et de la proportion de la population mère
  - appartenance d'un échantillon à une population de moyenne ou de proportion connue.
  - appartenance de deux échantillons à deux populations de même moyenne ou de même proportion
  - taille minimale d'un échantillon
- **PETITS ECHANTILLONS**
  - estimation de la moyenne d'une distribution mère normale d'écart type connu, ou inconnu
  - appartenance d'un petit échantillon à une population normale de moyenne connue
  - estimation de l'écart type d'une population mère normale
- **RACCORDEMENT** (comparaison d'une distribution expérimentale à une distribution théorique
  - ------->Test du KHI 2

- 1 - Grands échantillons (n > 30)

1.1 - Estimation d'une moyenne de population mère μ, à partir des paramètres d'un échantillon.

Exemple : on utilise un sondage sur les salaires, dans une entreprise de 1000 salariés. Pour cela on effectue une "mesure" sur un échantillon de 100 personnes. Cette étude donne les résultats suivants : X = 300F et s = 20F . Donner une estimation, au niveau de confiance de 95%, du salaire moyen dans l'entreprise.

- Rappelons les relations entre les moyennes d'échantillon X_i de taille n issus de la population et la moyenne μ de cette dernière (théorème des moyennes). Les X_i sont répartis selon : $\mathbf{N}(\mu,\sigma/n^{1/2})$(**).
- D'après les résultats énoncés au chapitre "Estimation", on peut définir un intervalle de confiance bilatéral pour les X_i, au niveau de confiance de Nc =1- α, par :

$$\mu - Z_c\sigma/n^{1/2} < X_i < \mu - Z_c\sigma/n^{1/2} \text{ ; } Z_c \text{ paramètre de confiance.}$$

d'où, pour μ :

$$X_i - Z_c\sigma/n^{1/2} < \mu < X_i - Z_c\sigma/n^{1/2}$$

- Malheureusement, dans cette expression μ est exprimé en fonction de σ, généralement non connu, qu'il faut alors estimer à partir du paramètre de dispersion fourni par l'échantillon, à savoir s. L'estimateur à utiliser est $s^* = s(n/n-1)^{1/2}$. Dans le cas des "grands échantillons", et plus particulièrement lorsque ceux ci sont de taille n supérieure à 100, on peut prendre $s^* \# s$.
- D'où :

$$\mathbf{X_i - Z_c s/n^{1/2} < \mu < X_i - Z_c s/n^{1/2}}$$

- Dans le cas d'étude proposé ci dessus, on a : X = 300, s = 20, n = 100, Zc = 1,96.

1.2 - Estimation d'une proportion de population mère p, à partir de la fréquence f d'un échantillon .

Exemple : on souhaite déterminer le taux de panne d'une machine. La machine a été observée 200 fois. On a noté qu'elle était 150 fois en état de marche et 50 fois en panne. Quel est, au niveau de confiance de 90%, l'estimation du taux moyen de "panne" de cette machine ?

- Les relations entre les fréquences f_i d'échantillons de taille n issus de la population et la fréquence(proportion) p de celle ci, sont telles que f_i appartient à une distribution binomiale de moyenne p et d'écart type $(pq/n)^{1/2}$(q=1-p).
- Pour les grands échantillons, on peut considérer que cette distribution se ramène à une distribution normale :

$$\mathbf{N}(p, (pq/n)^{1/2})$$

- On peut alors définir un intervalle de confiance pour f_i , au niveau de confiance de Nc =1- α, par :

$$p - Z_c(pq/n)^{1/2} < f_i < p + Z_c(pq/n)^{1/2}$$; avec Zc paramètre de confiance et q = 1- p.

- Soit pour p, l'intervalle de confiance suivant :

$$f_i - Z_c(pq/n)^{1/2} < p < f_i + Z_c(pq/n)^{1/2}$$

et, si dans $(pq / n)^{1/2}$ on estime p par f_i :

$$\mathbf{f_i - Z_c(f_i(1-f_i)/n)^{1/2} < p < f_i + Z_c(f_i(1-f_i)/n)^{1/2}}$$

1.3 - Un échantillon (n ,X) appartient-il à une population donnée (μ ; σ) ?

Exemple : Un fabricant d'appareils électroménagers prétend fournir à un grand magasin des ampoules électriques de durée de vie moyenne 1250 heures et d'écart type 150 heures. On réalise un test sur un lot de 50 de cette fabrication qui donne une durée de vie

moyenne de 1200 heures. Peut-on dire, au niveau de confiance de 95%, que l'affirmation du fabricant est vraie ?

- On émet **l'hypothèse nulle Ho** comme quoi l'échantillon étudié appartient à la population de moyenne μ.. On peut alors utiliser les résultats statistiques relatifs aux moyennes X_i des échantillons de taille n extraits d'une population de moyenne μ.
  - Tous les échantillons de taille n issus de la population appartiennent à une distribution normale $\mathbf{N}(\mu,\sigma/n^{1/2})$.
  - On sait donc établir, pour les moyennes d'échantillons X_i et au niveau de confiance Nc, un intervalle de confiance dépendant de μ et de l'écart type σ de la population.
  - Ce dernier étant inconnu, on l'estime par $s^* = s(n/(n-1))^{1/2}$ et on obtient alors l'intervalle de confiance pour X_i :

$$\mu - Z_c s^*/n^{1/2} < Xi < \mu + Z_c s^*/n^{1/2}$$

- On réalise alors le test suivant :
  - si la valeur trouvée expérimentalement pour X_i **appartient à cet intervalle**, alors on considère que l'**on ne peut rejeter l'hypothèse Ho.** On admet que l'échantillon étudié peut être issu de la population définie par sa moyenne μ.
  - si la valeur mesurée de X_i est **en dehors de l'intervalle de confiance**, alors on rejette l'hypothèse Ho ; on rejette par la même occasion, les affirmations du fabricant concernant les caractéristiques de la population d'où est issu l'échantillon étudié. Deux cas alors :
    - l'échantillon est bien issu de la population mais celle ci n'a pas les caractéristiques prétendues par le fabricant.
    - l'échantillon n'est pas issu de la fabrication (donc est issu d'une autre).
- Remarque : on n'a considéré ici que le test d'appartenance d'un échantillon à une **population de paramètre connu**. On peut aussi, et ceci sera fait ci dessous pour les fréquences, orienter le test pour analyser si l'échantillon appartient à une population **de paramètre supérieur ou inférieur à une valeur donnée**(test unilatéral).

1.4 - Un échantillon, de population n et de proportion f , appartient-il à une population de proportion p ?

Exemple : la fabrication d'un produit industriel comporte un taux de rejet de 10% considéré comme trop élevé. On essaie une nouvelle matière première afin de réduire ce taux et on fait alors un test sur un échantillon de taille 100. On trouve un taux de 9% . Au niveau de confiance de 95%, peut-on considérer que la nouvelle matière utilisée a amélioré la qualité de la fabrication c'est à dire réduit le taux de rejet au dessous de 10% ?

- Cet exemple précis va au delà de l'étude de l'appartenance ou non d'un échantillon à une population de paramètre connu ; il s'agit plutôt de **l'appartenance à une population de paramètres supérieurs ou inférieurs à une valeur donnée**. Ceci introduit la notion de t**ests unilatéraux.**
- **Le problème ici est de bien poser l'hypothèse nulle Ho** et d'exploiter correctement les résultats du test statistique.
  - On considérera qu'il est plus facile de prendre comme hypothèse Ho le fait que l'échantillon appartient à la production initiale dont on connaît le taux de rejet ;

on espère bien entendu être conduit à rejeter Ho, l'hypothèse alternative H_1 est alors à définir en fonction du test utilisé.
    - Si l'on prenait comme hypothèse Ho que la production est améliorée, le test statistique devrait montrer que l'échantillon étudié appartient à une distribution dont le taux de rejet est inférieur à 10% avec une valeur pour autant inconnue! La première solution est la plus facile à mettre en oeuvre.
    - Par ailleurs, il est intéressant de discuter du caractère unilatéral ou bilatéral du test.
- **Hypothèse nulle Ho** : l'échantillon appartient à la population initiale de fréquence p c'est à dire qu'il n'y a pas eu d'amélioration . L'hypothèse alternative H_1, c'est à dire celle prise en compte en cas de rejet de Ho, dépend de la façon de définir l'intervalle de confiance.
    - Les fréquences d'échantillon f_i, pour les grands échantillons, appartiennent à une distribution normale :

$\mathbf{N}(p, (pq/n)^{1/2})$ à partir de laquelle on établit, d'habitude, un intervalle de confiance bilatéral pour f_i :

$$p - Z_c(pq/n)^{1/2} < f_i < p + Z_c(pq/n)^{1/2}$$

    - Cet intervalle symétrique est celui où sont situés Nc% des valeurs des fréquences des échantillons de taille n issus de la population ; on remarque que le risque $\alpha = 1\text{-Nc}$ est pris en « éliminant » de l'ensemble de toutes les valeurs prises par les f_i, les plus écartées de p, **de part et d'autre de celle ci.**
- On pourrait, au contraire, éliminer les valeurs situées soit au dessus de p soit au dessous de p. Dans le cas présent, par exemple, les valeurs basses c'est à dire celles correspondant à p<10% nous intéressent plus particulièrement.
    - Nous définirons donc un intervalle de confiance **limité à gauche** ; on considérera ainsi que Nc% des fréquences d'échantillon sont contenues dans **un intervalle de valeurs supérieures à la valeur fg** définie pour un intervalle de confiance unilatéral à gauche. Le risque pris ici est de rejeter des valeurs de f_i inférieures à fg .
    - Le test consiste alors à comparer fg à la fréquence f de l'échantillon .
        - si f > fg ; f appartient à l'intervalle de confiance. On peut alors considérer que l'échantillon peut être issu d'une population de fréquence p c'est à dire que l'**on ne peut rejeter l'hypothèse Ho** . On en conclut qu'au niveau de confiance Nc il n'y a pas lieu de considérer qu'il y a amélioration de la fabrication.
        - si f < fg ; f est en dehors de l'intervalle de confiance ; **on rejette l'hypothèse Ho** : l'échantillon n'appartient pas à une population de fréquence p mais est donc issu d'une population de fréquence inférieure à p, pour laquelle correspondent, au niveau de confiance Nc, des valeurs des f_i , inférieures à fg.

1.5 - Comparaison de deux moyennes d'échantillons. Deux échantillons sont-ils issus de deux populations de même moyenne ?

Exemple . Une enquête auprès des étudiants pratiquant deux sports différents (handball et badminton) a donné les résultats suivants concernant leur masse :

	effectifs	moyenne	écart type
handball	80	78kg	8kg
badminton	500	83kg	5kg

Peut on en déduire que les joueurs de badminton sont plus lourds que les joueurs de handball?

- On est ici ramenés à répondre à la question générale suivante : deux échantillons de moyennes et écarts types donnés appartiennent ils à deux populations de même moyenne ?
- On posera comme **hypothèse nulle Ho** que les deux échantillons appartiennent à deux populations de même moyenne. L'hypothèse alternative H_1 étant ici simplement que les deux échantillons "n'appartiennent pas " à deux populations de même moyenne.
- **Supposant Ho vraie**, c'est à dire que l'on a deux populations distinctes 1 et 2 de même moyenne, on va étudier les propriétés de la variable $(X_1 - X_2)$, écart entre les moyenne d'échantillons issus des populations 1 et 2.
    - On connaît les propriétés des X_1 et X_2 . Leurs différences appartiennent à une distribution normale :

$$N((\mu_1 - \mu_2 = 0);\sigma_d), \text{ avec : } \sigma_d^2 = \sigma^2(X_1 - X_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2 \text{ (additivité des variances).}$$

    - Comme ni σ_1 ni σ_2 ne sont connus, on utilise comme estimateurs $s_1^* = s_1$ et $s_2^* = s_2$ (grands échantillons).
    - Au niveau de confiance Nc, les écarts $(X_1 - X_2)$ appartiennent alors à un intervalle de confiance :

$$-Z_c\,\sigma_d < (X_1 - X_2) < Z_c\,\sigma_d$$

    - Le test consiste donc à calculer **l'écart mesuré t** = $(X_1 - X_2)$ entre les deux échantillons étudiés et de constater si oui ou non cet écart est compris dans l'intervalle de confiance.
        - si **t** est à l'intérieur de l'intervalle de confiance, l'hypothèse **Ho ne peut être rejetée.** Les deux échantillons étudiés peuvent appartenir à deux populations ayant même moyenne.
        - si **t** est en dehors de l'intervalle de confiance, alors **l'hypothèse Ho est rejetée** : les deux échantillons sont issus de deux populations de moyennes différentes.
- Remarques :
    - la valeur de la moyenne commune des deux populations n'intervient pas.
    - s'il s'agit d'une même population, alors $\sigma_1 = \sigma_2$; on prendra quand même comme estimateurs , dans σ_d , les valeurs de s_1 et de s_2 .
    - l'estimateur s* de l'écart type de la population (1 ou 2) est l'écart type de l'échantillon s corrigé, à savoir $s^* = s(n/n-1)^{1/2}$; lorsque l'échantillon atteint ne taille supérieure à 100, la correction est négligeable.

1.6 - Comparaison de deux proportions.

Exemple. L'étude du taux annuel d'accidents du travail dans deux entreprises de tailles différentes a donné les résultats suivants : entreprise A (taille 800 salariés ; taux annuel 90) ; entreprise B(taille 300 salariés ; taux annuel 50). Peut admettre, au niveau de confiance Nc, que le taux d'accident est le même dans les deux entreprises ?

- L'étude se fait de la même façon qu'au chapitre précédent. On fait **l'hypothèse Ho d'appartenance des deux échantillons à deux populations de même fréquence**. Puis on étudie la variable $t = f_1 - f_2$.
- Pour les grands échantillons, la variable t appartient à une distribution normale : $\mathbf{N((f_1 - f_2 = 0);\sigma_d)}$ avec $\sigma_d = (pq/n_1 + pq/n_2)^{1/2}$. On remplace, dans σ_d , p par son estimateur dans les deux échantillons, à savoir, f_1 et f_2 . D'où : $\sigma_d = (f_1(1- f_1)/n_1 + f_2(1- f_2)/n_2)^{1/2}$.
- Le test consiste alors à vérifier si, pour les deux échantillons étudiés, l'écart observé ($f_1 - f_2$) appartient ou non à líintervalle de confiance :

$$- Zc\ \sigma_d < (f_1 - f_2) < Zc\ \sigma_d$$

- Si $f_1 - f_2$ appartient à l'intervalle, l'hypothèse **Ho ne peut être rejetée** .
- Si $f_1 - f_2$ n'appartient pas à l'intervalle, l'hypothèse **Ho est rejetée**.

1.7 - Quelle doit être la taille d'un échantillon si l'on définit la valeur maximale de l'intervalle d'estimation (moyenne ou proportion de la population) ?

- Dans tout ce que nous avons vu jusqu'ici, la statistique inférentielle fournissait des résultats relatifs à une population à partir de valeurs statistiques relatives à un échantillon (taille, moyenne, écart type ou fréquence). La taille de l'échantillon, en particulier, conditionnait la largeur de l'intervalle de confiance. Mais dans certains cas on voudrait, à **priori**, pouvoir définir une taille minimale d'échantillon permettant une largeur de l''intervalle fixée à l'avance.
- On rappellera, ci dessous,les relations entre paramètres d'échantillon et de population dans le cas le plus simple des fréquences. Le raisonnement reste cependant valable dans le cas de la moyenne au prix de quelques modifications d'ordre mineur.
- D'après le théorème des fréquences, pour les grands échantillons, les fréquences f_i d'échantillons appartiennent à une distribution normale de moyenne p et d'écart type $(pq/n)^{1/2}$.
    - on définit pour les f_i des intervalles de confiance : $p - Z_c(pq/n)^{1/2} < f_i < p + Z_c(pq/n)^{1/2}$ dont la demi largeur : $Z_c(pq/n)^{1/2}$ dépend de Z_c , de la valeur à estimer p et de la taille de l'échantillon n.
    - p n'intervient que par le produit pq = p(1-p) majoré par 1/4 ; il s'ensuit que la demi largeur de l'intervalle de confiance, qui représente l'écart arithmétique maximal entre f_i et la valeur recherchée p, ne dépend plus que du niveau de confiance, prédéfini, et la taille n de l'échantillon.
- D'où la relation entre Zc, la taille de l'échantillon n et l'écart maximal souhaité Δ :

$$n > Z_c^2 / 4\Delta^2$$

- La méthode ci dessus peut être adaptée au cas où on veut estimer par intervalle de confiance la moyenne d'une population. Il faut dans ce cas avoir une valeur

approximative de la moyenne recherchée. Celle ci peut être fournie par une lesure préalable sur un petit échantillon

Petits échantillons (n < 30)

2.1 - Estimation de la moyenne d'une population normale d'écart type connu.

- Pour une population **normale** de moyenne μ et **d'écart type connu σ**, les moyennes X_i d'échantillons de taille n < 30 appartiennent à une distribution normale $\mathbf{N}(\mu, \sigma/n^{1/2})$.
- On procède exactement comme pour les grands échantillons(chap11). Il s'ensuit pour μ :

$$X_i - Z_c\sigma /n^{1/2} < \mu < X_i - Z_c\sigma /n^{1/2}$$

ou : $\mu = X_i +/- Z_c\sigma /n^{1/2}$

2.2 - Estimation de la moyenne d'une population *normale d'écart-type inconnu*.

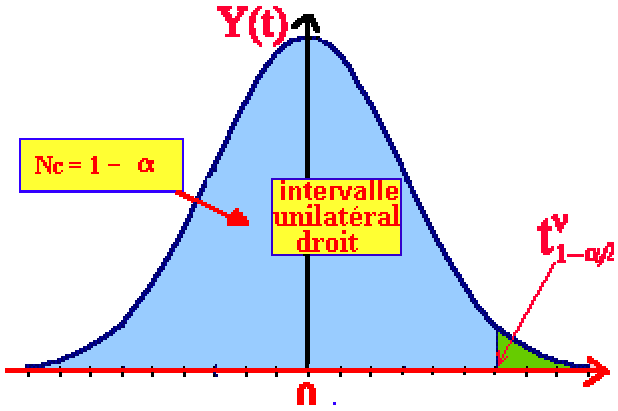
TEST DE STUDENT.

Exemple.Une machine produit des tôles d'épaisseur 0,050cm. Désirant déterminer si cette machine fonctionne normalement, on analyse un échantillon de 10 tôles. On trouve X=0,053cm et s=0,003cm. Peut-on en déduire, au seuil de signification de 0,05, que la machine fonctionne correctement ?

- L'écart-type de la population est inconnu, en revanche on connaît celui de l'échantillon s ; s* estimateur de s est égal à $s(n/(n-1))^{1/2}$.
- Aux paramètres X et s de tous les échantillons de taille n issus de la population, on associe le paramètre :

$$\mathbf{t = (X-\mu)(n-1)^{1/2}/s = (X-\mu)/(s^*/n^{1/2})}$$

- ce paramètre est distribué selon la loi : $\mathbf{Y(t) = Yo/(1 + t^2/(n-1))^{n/2}}$ appelée "**distribution de Student** ".
  - Yo est tel que l' intégrale de $-f$ à $+f$ est égale à 1.
  - Y(t) dépend de la taille **n** de l'échantillon. Cependant, lorsque n--> 30, cette distribution tend vers la distribution normale réduite centrée $Y(Z) = \exp(-t^2/2)/(2\pi)^{1/2}$.



- la fonction Y(t) est tabulée. Les tables fournissent les limites d'intervalles de confiance unilatéraux à droite : $t^{n}_{1-\alpha}$ qui dépendent du paramètre $\nu = n-1$ (n "degrés de liberté) et du seuil de signification $\alpha = 1-Nc$.

- On peut, comme pour les grands échantillons, définir pour t des intervalles de confiance unilatéraux ou bilatéraux selon les besoins.
  - pour une estimation par intervalle de confiance, on utilise généralement un test bilatéral qui fournit un intervalle fini, centré sur la mesure de l'échantillon (X).
  - pour un test de comparaison, on utilise des tests unilatéraux à droite ou à gauche.
- Considérons le cas d'une estimation par intervalle de confiance (test bilatéral).
  - le "risque" $\alpha = 1- Nc$ étant réparti à droite et à gauche, le paramètre de confiance fourni par la table est défini par : $t^{n}_{1-\alpha/2}$; d'où l'intervalle de confiance pour t :

$$-t^{n}_{1-\alpha/2} < t < t^{n}_{1-\alpha/2}$$

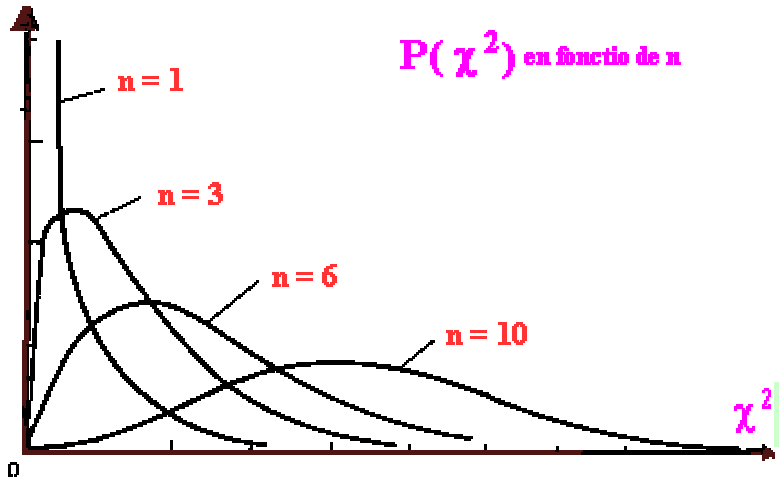
- D'où : $\mu = X +/- t^{n}_{1-\alpha/2}\, s/(n-1)^{1/2}$
- **REMARQUES :**
  1. la méthode d'estimation de m est **la même que pour les grands échantillons excepté** qu'il faut utiliser la distribution de STUDENT au lieu de la distribution normale. Le paramètre t joue le rôle du Zc de la loi normale.
  2. attention, la règle de définition de t dans les tables est différente de celle des Zc définis, dans le cas de la loi normale, pour des intervalles de confiance bilatéraux.
  3. lors de tests de comparaison, on pourra, comme pour les grands échantillons, mettre en oeuvre des tests bilatéraux.

2.3 - Estimation de l'écart-type d'une population normale. TEST DU KHI DEUX (χ^2)

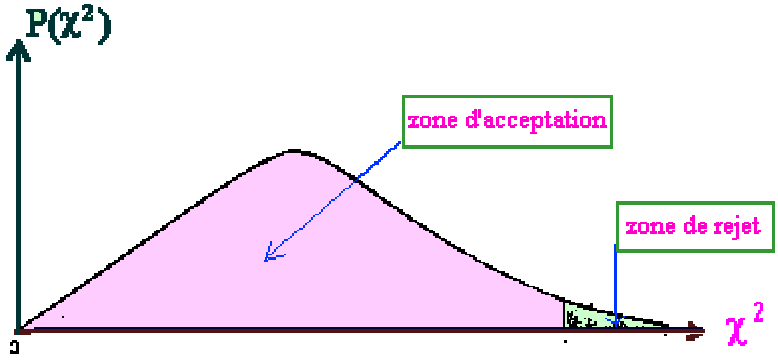
Exemple. L'écart type de durée de vie d'un échantillon de 20 d'appareils ménagers est de 100 heures. Donner un intervalle de confiance, au niveau de confiance de 95%, pour l'écart type de la production de ces ampoules.

- Considérons l'ensemble des écarts types des échantillons de taille n issus d'une **population mère normale**.
- On définit la fonction : $\chi^2 = ns^2/\sigma^2 = \left(\sum_i (X_i-\mu)^2\right)/\sigma^2$

- Cete fonction de distribution, appelée aussi "distribution de Pearson ou du KHI2", paramètrée en n (ou ν=n-1), normée, est tabulée.



On peut donc, comme ce fut le cas pour les distributions normale et Student, définir des intervalles de confiance bilatéraux ou unilatéraux, selon les besoins.



- Les tables utilisées fournissent des limites à droite d'intervalles paramètrés en ν = n-1 (n degrés de liberté).
  - au niveau de confiance de 1-α , on a :

$$\chi^2_{\alpha/2} < \chi^2 = ns^2/\sigma^2 < \chi^2_{1-\alpha/2}$$

  - soit pour σ :

$$sn^{1/2}/\chi_{1-\alpha/2} < \sigma < sn^{1/2}/\chi_{\alpha/2}$$

CM

3 - Raccordement d'une distribution expérimentale avec une distribution théorique. TEST DU KHI DEUX (χ^2)

On désire vérifier la perfection d'un dé ; sur un essai de 300 jets on obtient :

X_i	1	2	3	4	5	6
n_i	40	49	72	50	42	47

Au seuil de signification de 0,01 peut on dire que le dé comporte une malfaçon (est truqué)?

- On dispose d'une distribution expérimentale que l'on veut la comparer à un distribution théorique ayant pour paramètres ceux de la distribution expérimentale (μ, s, f, etc.). On utilisera pour cela le test dit "du KHI DEUX".
- Pour une approche graphique de la normalité de la distribution expérimentale, on peut ausi utiliser "la droite de Henry".
- Méthode :
  - on définit la distribution théorique à laquelle on veut "comparer" la distribution expérimentale(cf chapitre 2).
  - on répartit les données expérimentales par classes ; n_i est l'effectif de la classe i .
  - on calcule les effectifs théoriques de classes n_{ith} c'est à dire les effectifs des classes .i pour la distribution théorique
  - on exprime le "KHI2 observé" : $\chi^2_{obs} = \sum_i (n_i - n_{ith})^2 / n_{ith}$
  - on recherche, dans la table du χ^2, la valeur du χ^2_{tab} associée au paramètre ν = n-1(n nombre de classes) et au niveau de confiance Nc. On se base sur un intervalle unilatéral à droite, dont la valeur précédente est la limite finie.
  - on compare χ^2_{obs} et χ^2_{tab} :
    - si $\chi^2_{obs} < \chi^2_{tab}$ on admet le raccordement entre les deux distributions.
    - si $\chi^2_{obs} > \chi^2_{tab}$ on rejette raccordement entre les deux distributions.