



Statistiques de base

- **Importance de la probabilité et la statistiques**
- **Méthodes d'analyse**
- **Mesures statistiques**
 - Tendance centrale
 - Dispersion
- **La courbe normale**
- **Prévisions utilisant une distribution normale**
- **Théorème de la limite centrale**
- **Intervalle de confiance**
- **Exercices**

- On entend par *statistiques* la collecte, l'organisation, l'analyse, l'interprétation et la présentation des données.
- C'est un des nombreux outils permettant de résoudre les problèmes de qualité.
- Les *statistiques descriptives* nous donnent des informations sur les performances d'un procédé.
- Les *statistiques par inférence* nous permettent de prévoir les performances futures d'un procédé sur la base de mesures actuelles.
- L'objectif final est de *prévoir & prévenir* plutôt que de *contrôler & détecter*.
- Les statistiques peuvent avoir plusieurs formes : tabulaires, numériques, graphiques.

La probabilité est la base des prévisions

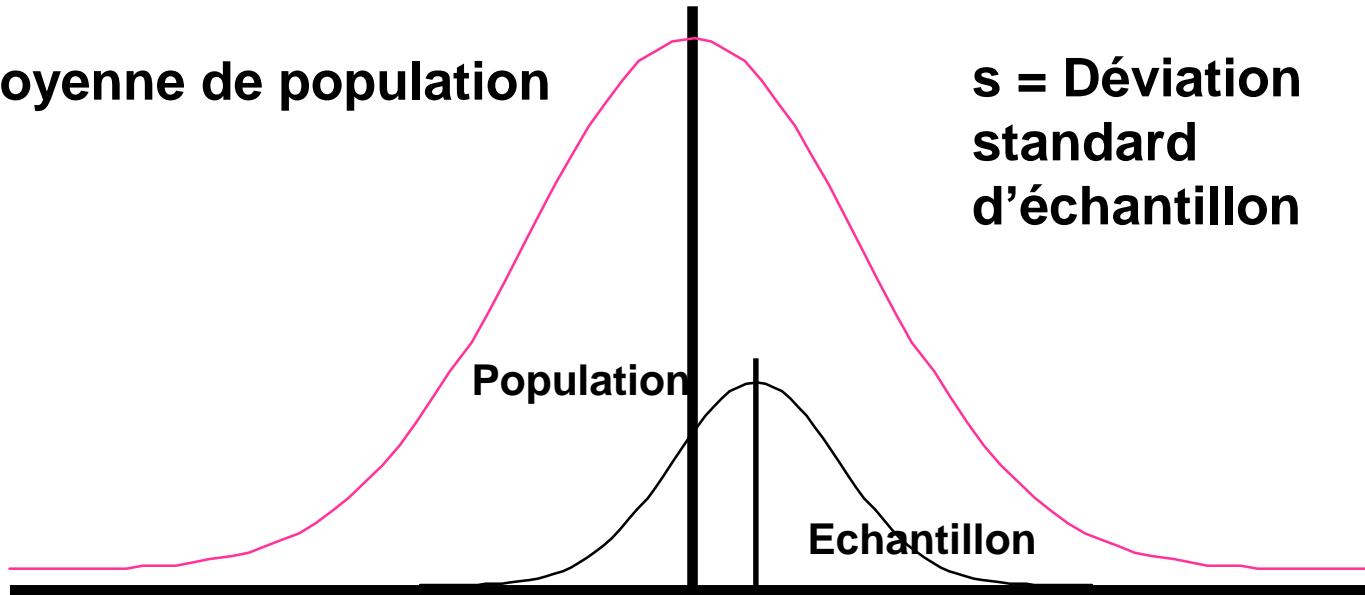
Terminologie

- **Un échantillon** est un nombre limité des éléments prise de la source des données.
 - le groupe d'objets véritablement mesuré dans une étude statistique
 - un échantillon est en général un sous-ensemble de la population à laquelle on s'intéresse.
- **Une population** la source des éléments ou les échantillons sont prises.
 - un groupe entier d'objets qui ont été ou vont être créés, présentant une caractéristique intéressante
 - il est probable que nous ne connaissons jamais les paramètres réels de la population
- **L'inférence statistique** implique la mesure sur un échantillon et les prévisions sur une population.
- Généralement, les symboles grecs représentent les paramètres de population (μ, σ) et les lettres romaines (x, s) sont utilisées pour représenter les valeurs d'échantillon.

“Paramètres de population”

σ = Déviation standard de population

μ = Moyenne de population



“Statistiques d'échantillons”

\bar{x} = Moyenne d 'échantillon

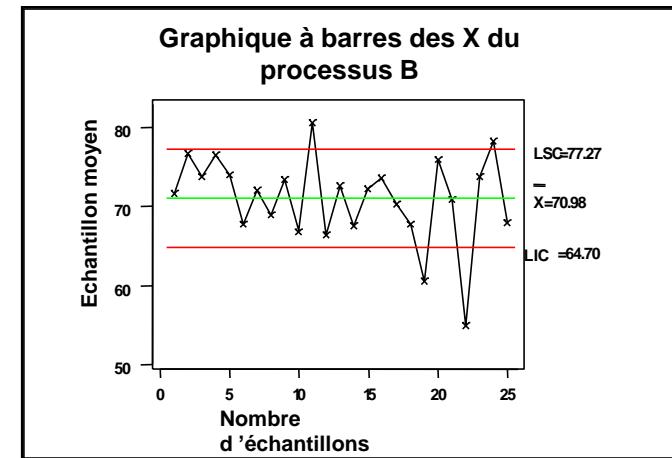
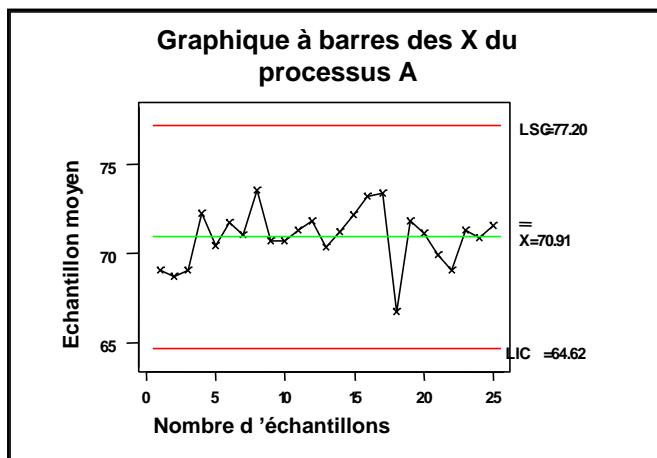
s = Déviation standard d'échantillon

➤ Variabilité

- Le processus atteint-il ses objectifs concernant la variabilité minimum ?
- On utilise la moyenne pour déterminer si le processus atteint son objectif, et la Déviation standard (σ), pour connaître la répartition.

➤ Stabilité

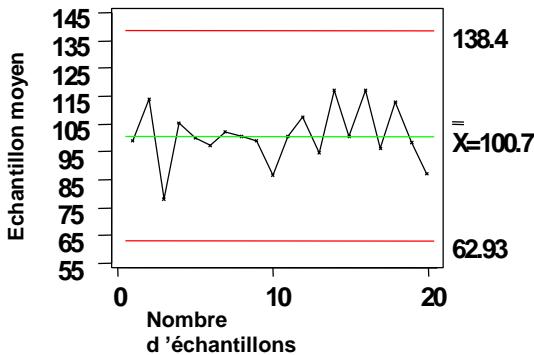
- Quelle est la performance du processus sur une durée donnée ?
- La stabilité est représentée par une variabilité moyenne constante et prévisible dans le temps.



- Supposons que les machines A, B, et C fabriquent des produits identiques
- Supposons que la valeur ciblée de chaque variable de produit est 100mm.
- Répondre aux questions suivantes:
 - Quelle(s) machin(es) présente(ent) une variation?
 - Sur quoi chaque machine est-elle centrée ?
 - Quelles machines sont prévisibles ?
 - Quelles machines présentent une variation ayant une cause spéciale ?
 - Quelle machine choisiriez-vous pour fabriquer votre produit ?
 - Quelle machine serait la plus facile à réparer ?

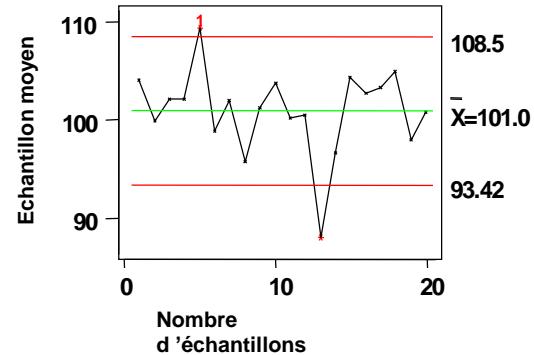
Graphique à barres des X
de la machine A

X-bar Chart for Machine A



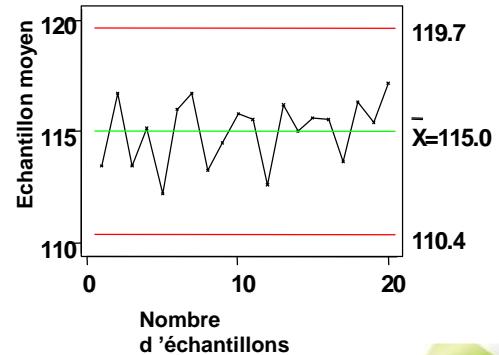
Graphique à barres des X
de la machine B

X-bar Chart for Machine B



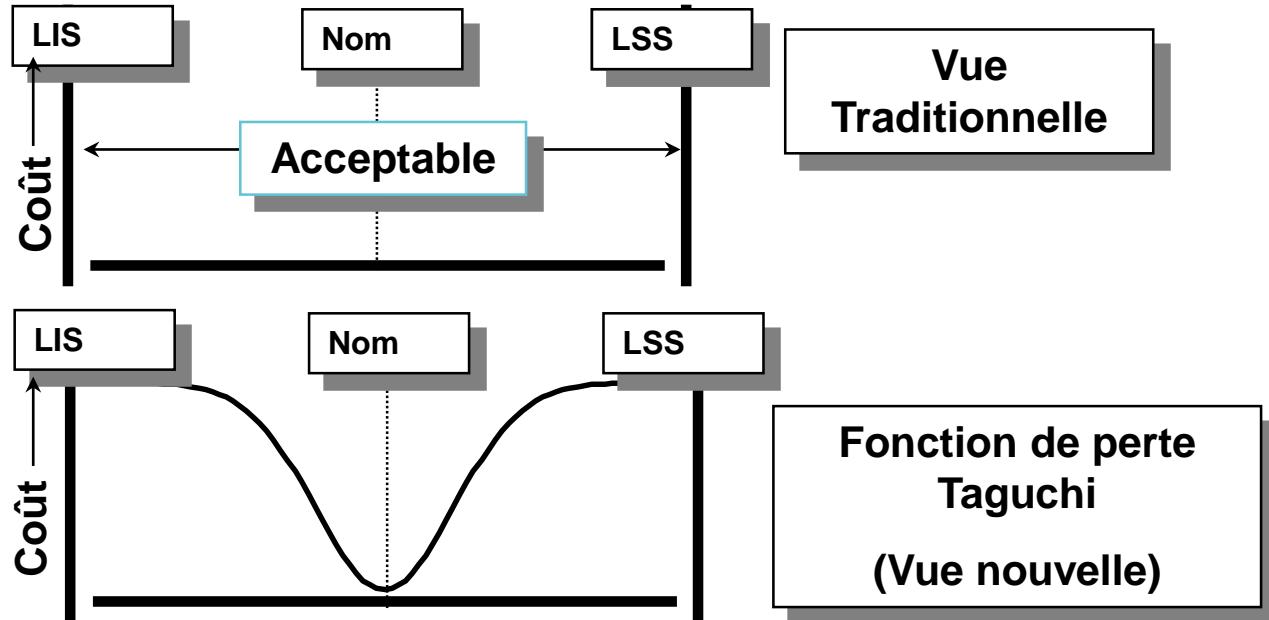
Graphique à barres des X de la
machine C

X-bar Chart for Machine C

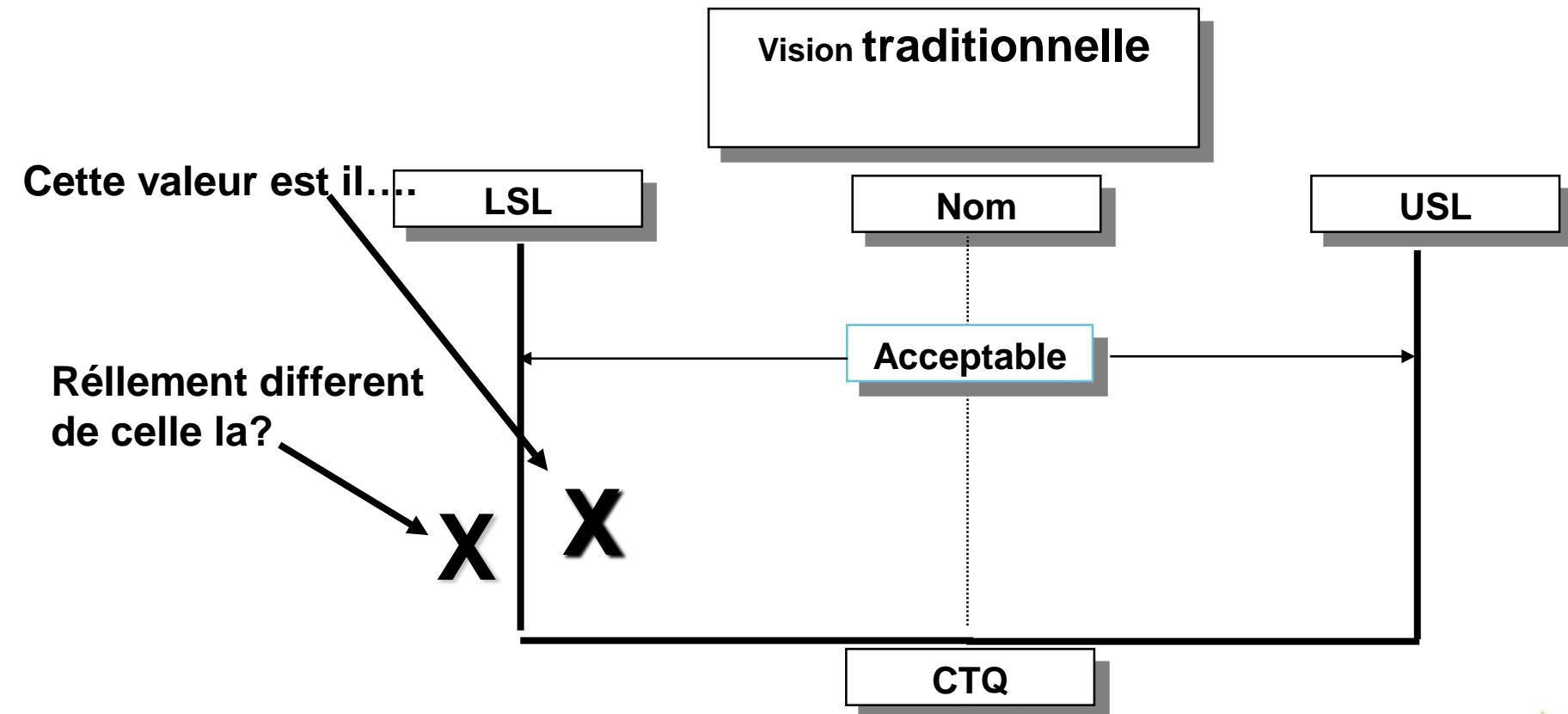


Pouvons-nous tolérer la variabilité?

- Tout processus présentera toujours une certaine variabilité
- Nous pouvons tolérer cette variabilité si:
 - le processus remplit ses objectifs;
 - la variabilité totale est relativement faible par rapport aux spécifications du processus;
 - le processus est stable dans le temps.



Notre souci ne sera plus “est ce que nous sommes dans les specs?”

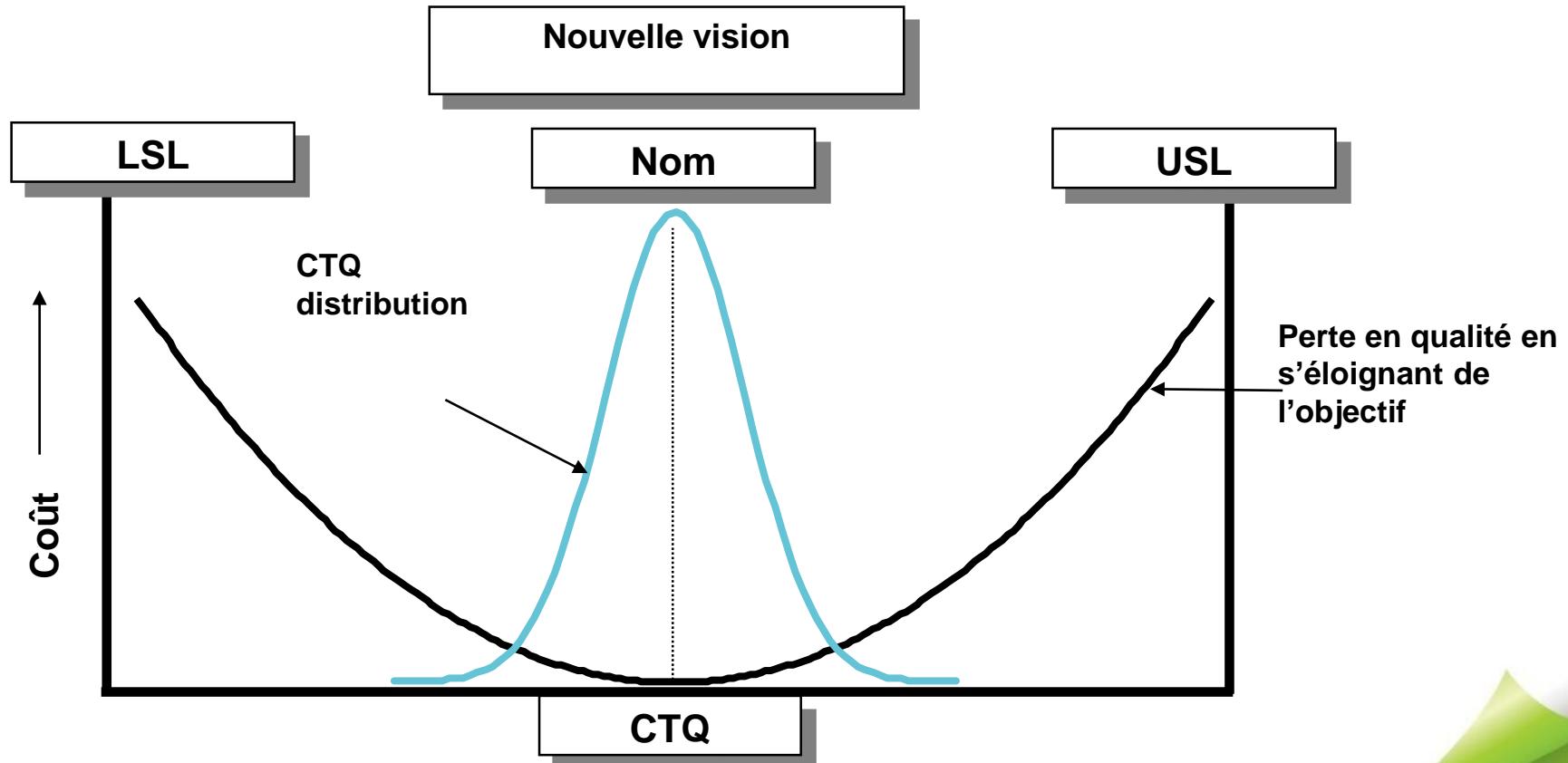


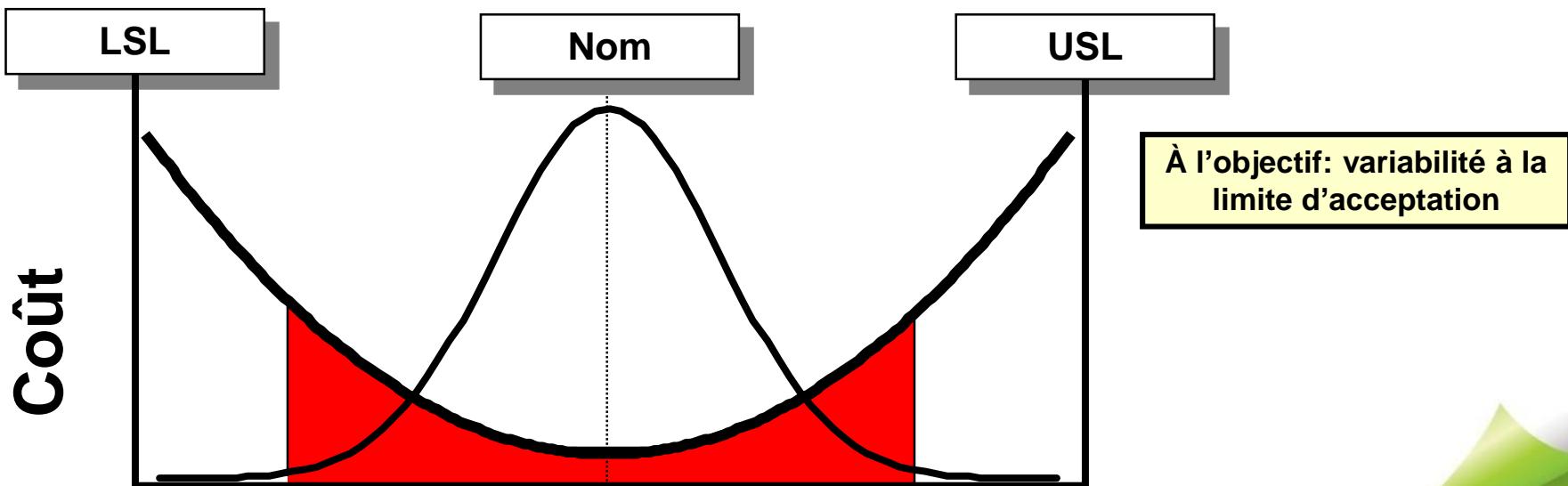
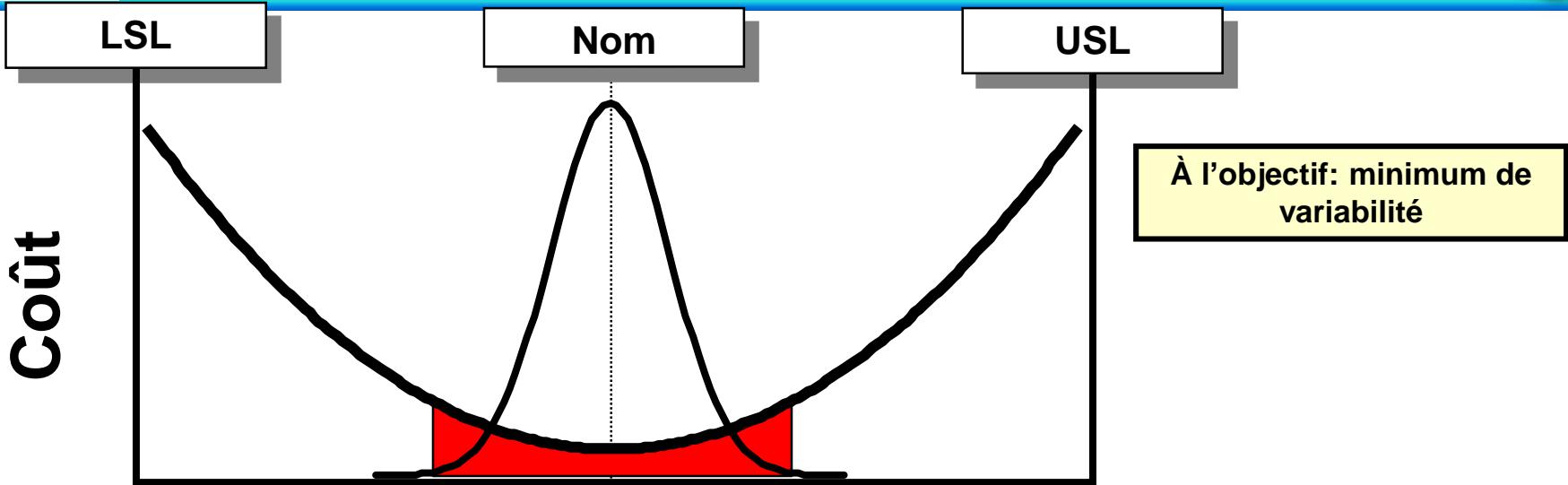
Nous sommes encore entraînés à utiliser l'une et de jeter l'autre !

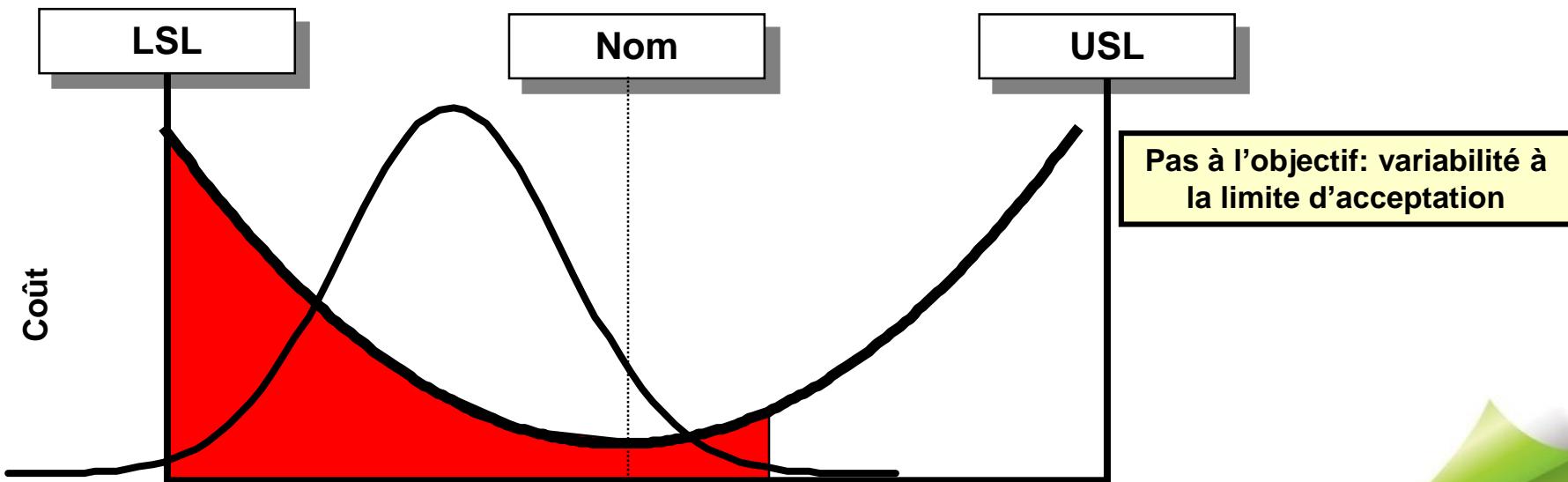
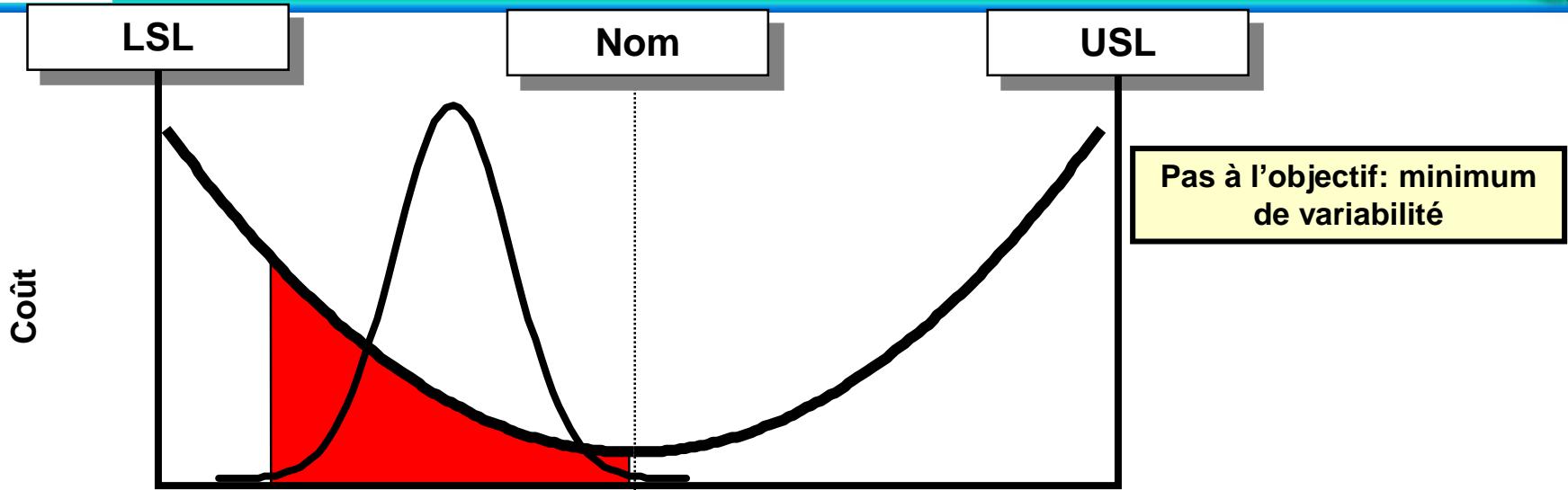
Nouvelle vision de la qualité

Notre souci devient :

“Est ce que nous sommes à l'objectif avec le minimum de variation?”







- **Déterminer si le processus est stable**
 - Si le processus n'est pas stable, identifier et supprimer les causes (X) d'instabilité (variation évidente et non aléatoire)
- **Situer la moyenne du processus. Répond-il à ses objectifs ?**
 - Si ce n'est pas le cas, identifier les variables (X) qui affectent la moyenne et déterminer les réglages optimaux pour atteindre les objectifs.
- **Estimer l'ampleur de la variabilité totale. Est-elle acceptable en ce qui concerne les exigences du client (limites de spécifications)?**
 - Si ce n'est pas le cas, identifier les sources de variabilité et supprimer ou réduire leur influence sur le processus.
- **Nous allons maintenant examiner des statistiques qui peuvent aider ce processus.**

- **On peut décrire le comportement de n'importe quel processus ou système en indiquant de multiples points de données pour la même variable**
 - sur une certaine durée
 - pour plusieurs produits
 - sur diverses machines, etc.
- **L'accumulation de ces données peut être considérée comme une répartition de valeurs représentée par:**
 - des graphiques à points
 - des histogrammes
 - des courbes normales ou autre répartition “arrondie”

- **Moyenne:** moyenne arithmétique d'un ensemble de valeurs
 - Reflète l'influence de toutes les valeurs
 - Fortement influencée par les valeurs extrêmes

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

- **Médiane:** reflète les 50% - le nombre central une fois qu'un ensemble de chiffres a été trié
 - Ne tient pas forcément compte de toutes les valeurs dans le calcul
 - Est “dure” avec les valeurs extrêmes
- **Mode:**
 - La valeur la plus fréquente dans les ensembles de données
- **Pourquoi utiliser la moyenne au lieu de la médiane dans nos efforts d'amélioration du processus ?**

Exercice

Nous allons calculer ensemble la moyenne et la médiane de la série de données n°1.

Moyenne =

Médiane =

Faites maintenant la série de données n°2 et 3 en groupes.

Série n°2 Moyenne =

Médiane =

Série n°3 Moyenne =

Médiane =

Set#1	Set#2	Set#3
98	105	107
102	109	169
108	116	131
105	76	84
89	148	81
92	87	67
114	86	81
90	70	122
97	137	52
100	99	233
104	119	46

Mesures de dispersion

■ Étendue

Distance numérique entre les valeurs les plus élevées et les valeurs les plus basses d'une série de données

- Très sensible aux valeurs extrêmes des données

$$Range = |\min - \max|$$

■ Étendue interquartile (IQR)

Distance extrême entre le 1er et le 3ème quartile d'une série de données divisée en 4 groupes égaux

- Utilisée pour générer des boîtes à moustaches

$$IQR = Q3 - Q1$$

■ Variance (σ^2 ; s^2)

Moyenne des carrés des écarts de chaque point de données individuel par rapport à la valeur moyenne

- Pas du tout sensible aux valeurs extrêmes des données

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

■ Écart type (σ ; s)

Racine carrée de la variance ; distance moyenne des données par rapport à la moyenne

- Mesure la plus communément utilisée pour quantifier une variation

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Exercice

Nous allons calculer ensemble l'étendue, la variance et l'écart type de la série de données n°1.

Étendue =

Variance = s^2 =

Écart type = s =

Faites maintenant la série de données n°2 et 3 en groupes.

Série n°2

Étendue =

Variance = s^2 =

Écart type = s =

Série n° 3

Étendue =

Variance = s^2 =

Écart type = s =

	Set#1	Set#2	Set#3
	98	105	107
	102	109	169
	108	116	131
	105	76	84
	89	148	81
	92	87	67
	114	86	81
	90	70	122
	97	137	52
	100	99	233
	104	119	46

- Nous avons précédemment défini les valeurs d'échantillon et les paramètres de population.
- Une collection de valeurs de probabilité est appelée une *distribution*.
- Une *fonction de distribution de probabilité* est une formule mathématique qui rapproche les valeurs des caractéristiques avec leur probabilité d'occurrence dans la population.
- Lorsque la caractéristique mesurée peut prendre une valeur quelconque (dépendant de la finesse de la méthode de mesure), sa distribution de probabilité est *continue*.
 - Les distributions normales, exponentielles et de Weibull sont des exemples.
- Lorsque la caractéristique mesurée ne peut prendre qu'une valeur bien spécifique, sa distribution de probabilité est *discrète*.
 - Binôme et Poisson sont des exemples.

DISTRIBUTION BINOMIALE (*distribution discrète finie*)

1.1. VARIABLE DE BERNOULLI

Définition : Une variable aléatoire discrète qui ne prend que les valeurs 1 et 0 avec les probabilités respectives p et $q = 1 - p$ est appelée variable de BERNOULLI.

On affecte alors 1 à la variable en cas de succès et 0 en cas d'échec.

Distribution de probabilités

x	0	1
$f(x) = p(X = x)$	q	p

Paramètres de la distribution

$$E(X) = 0 \cdot q + 1 \cdot p = p,$$

$$E(X) = p \quad V(X) = pq \quad \sigma(X) = \sqrt{pq}$$

$$V(X) = E(X^2) - E(X)^2 = (0^2 q + 1^2 p) - p^2 = p - p^2 = pq,$$

Finalement, on obtient pour $0 \leq k \leq n$:

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

On dit que la variable aléatoire X suit une loi binomiale de paramètres n et p . On note : $X \sim B(n,p)$.

Nous savons que : $X = X_1 + \dots + X_n$ avec $E(X_i) = p$ pour $1 \leq i \leq n$.
 Donc : $E(X) = E(X_1) + \dots + E(X_n) = np$.

Les variables X_i sont indépendantes et $\text{Var}(X_i) = pq$ pour $1 \leq i \leq n$.
 Donc : $\text{Var}(X) = \text{Var}(X_1) + \dots + \text{Var}(X_n) = npq$.

$$E(X) = np \quad \text{Var}(X) = npq \quad \sigma(X) = \sqrt{npq}$$

Nous cherchons à déterminer la loi de probabilité de la variable $X = \text{« nombre de réalisations d'un événement donné pendant un intervalle de temps } t \text{ »}$, sachant que le nombre moyen de réalisations de cet événement par unité de temps est α .

Il s'agit d'une loi binomiale $B(n, p)$ où $p = \alpha \frac{t}{n}$. $B(n, \alpha \frac{t}{n})$ at = λ .

$$B(n, \frac{\lambda}{n})$$

Définition : La distribution $p(Y = k) = C_n^k \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$: λ est celle d'une variable discrète X qui prend ses valeurs dans N selon la fonction de densité :

On écrit : $X \sim P(\lambda)$.

$$f(k) = p(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad k \in N$$

$$E(X) = \lambda \quad \text{Var}(X) = \lambda \quad \sigma(X) = \sqrt{\lambda}$$

DISTRIBUTION NORMALE (distribution continue)

SITUATION CONCRÈTE

Une variable aléatoire continue suit une loi normale si l'expression de sa fonction de densité de probabilités est de la forme :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2} \quad x \in \mathbb{R}$$

La loi dépend des deux réels m et σ appelés paramètres de la loi normale.
On note : $X \sim N(m, \sigma)$.

$$E(X) = m$$

$$\text{Var}(X) = \sigma^2$$

$$\sigma(X) = \sigma$$

loi normale centrée réduite notée $N(0,1)$.

Donc si $X \sim N(m, \sigma)$, on pose $T = \frac{X - m}{\sigma}$ et $T \sim N(0,1)$.

On peut résumer la correspondance de la façon suivante:

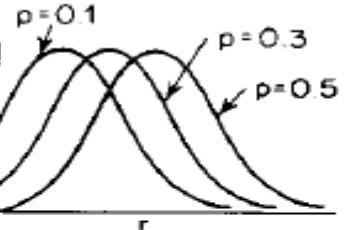
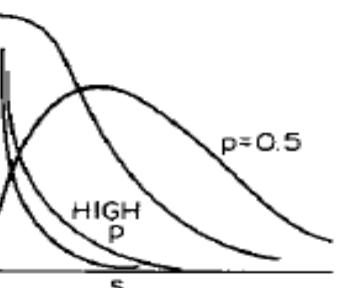
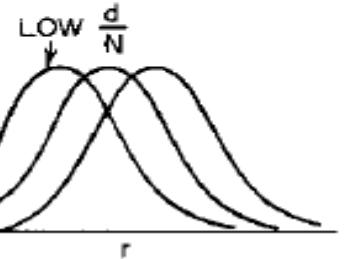
Variable normale		Variable normale centrée réduite
$X \sim N(m, \sigma)$	$T = \frac{X - m}{\sigma}$	$T \sim N(0,1)$
ensemble des valeurs prises : \mathfrak{N}		ensemble des valeurs prises : \mathfrak{N}
Paramètres : $E(X) = m$ $Var(X) = \sigma^2$		Paramètres : $E(T) = 0$ $Var(T) = 1$

APPROXIMATIONS PAR DES LOIS NORMALES

On approche la loi $B(n,p)$ par la loi $N(np, \sqrt{npq})$ dès que $\begin{cases} n \geq 30 \\ np \geq 15 \\ nq \geq 15 \end{cases}$

On approche la loi $P(\lambda)$ par la loi $N(\lambda, \sqrt{\lambda})$ dès que $\lambda \geq 16$.

Distributions de probabilité

BINOMIAL*	 <p>$p = 0.1$</p> <p>$p = 0.3$</p> <p>$p = 0.5$</p>	$y = \frac{n!}{r!(n-r)!} p^r q^{n-r}$ n = Number of trials r = Number of occurrences p = Probability of occurrence $q = 1-p$	Applicable in defining the probability of r occurrences in n trials of an event which has a probability of occurrence of p on each trial.
NEGATIVE BINOMIAL*	 <p>$p = 0.5$</p> <p>HIGH P</p> <p>s</p>	$y = \frac{(r+s-1)!}{(r-1)!(s!)}$ $p^r q^s$ r = Number of occurrences s = Difference between number of trials and number of occurrences p = probability of occurrence $q = 1-p$	Applicable in defining the probability that r occurrences will require a total of $r+s$ trials of an event which has a probability of occurrence of p on each trial. (Note that the total number of trials n is $r+s$.)
HYPERGEOMETRIC*	 <p>LOW $\frac{d}{N}$</p> <p>d</p> <p>N</p> <p>r</p>	$y = \frac{\binom{d}{r} \binom{N-d}{n-r}}{\binom{N}{n}}$	Applicable in defining the probability of r occurrences in n trials of an event when there are a total of d occurrences in a population of N .

Distributions de probabilité

DISTRIBUTION	FORM	PROBABILITY FUNCTION	COMMENTS ON APPLICATION
NORMAL		$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ <p>μ = Mean σ = Standard deviation</p>	Applicable when there is a concentration of observations about the average and it is equally likely that observations will occur above and below the average. Variation in observations is usually the result of many small causes.
EXPONENTIAL		$y = \frac{1}{\mu} e^{-\frac{x}{\mu}}$	Applicable when it is likely that more observations will occur below the average than above.
WEIBULL	<p>$\beta = 1/2$, $\alpha = 1$ $\beta = 1$, $\beta = 3$</p> <p>X</p>	$y = \alpha\beta(x-\gamma)^{\beta-1}e^{-\alpha(x-\gamma)^\beta}$ <p>α = Scale parameter β = Shape parameter γ = Location parameter</p>	Applicable in describing a wide variety of patterns of variation, including departures from the normal and exponential.
POISSON*	<p>$p = 0.1$, $p = 0.3$, $p = 0.5$</p> <p>r</p>	$y = \frac{(np)^r e^{-np}}{r!}$ <p>n = Number of trials r = Number of occurrences p = Probability of occurrence</p>	Same as binomial but particularly applicable when there are many opportunities for occurrence of an event, but a low probability (less than 0.10) on each trial.

- **La répartition “Normale” est une répartition des données qui possèdent certaines caractéristiques cohérentes**
- **Ces caractéristiques sont très utiles pour nous permettre de comprendre les propriétés du processus d'où viennent les données**
- **La plupart des phénomènes naturels et des processus créés par l'homme sont répartis normalement, ou peuvent être représentés comme tels.**

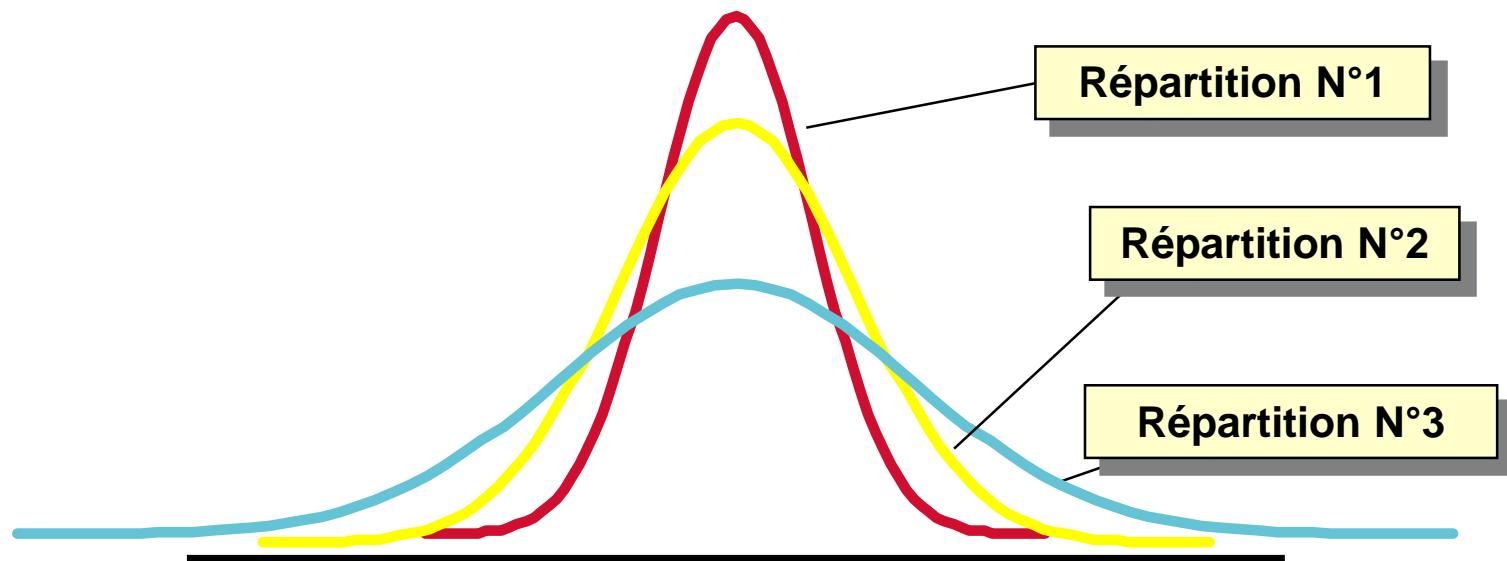
- La distribution utilisée le plus communément dans les analyses statistiques de procédés industriels est la distribution normale.
- La fonction de densité de probabilité (fdp) est :

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X - \mu)^2/2\sigma^2}$$

- Les propriétés significatives sont les suivantes :
 - La courbe est symétrique autour de la moyenne
 - Les valeurs de l'asymétrie et du kurtosis = 0
 - La moyenne et l'écart type sont indépendants
 - Moyenne = médiane
 - Le domaine en-dessous de la courbe de $-\infty$ à $+\infty$ = 1

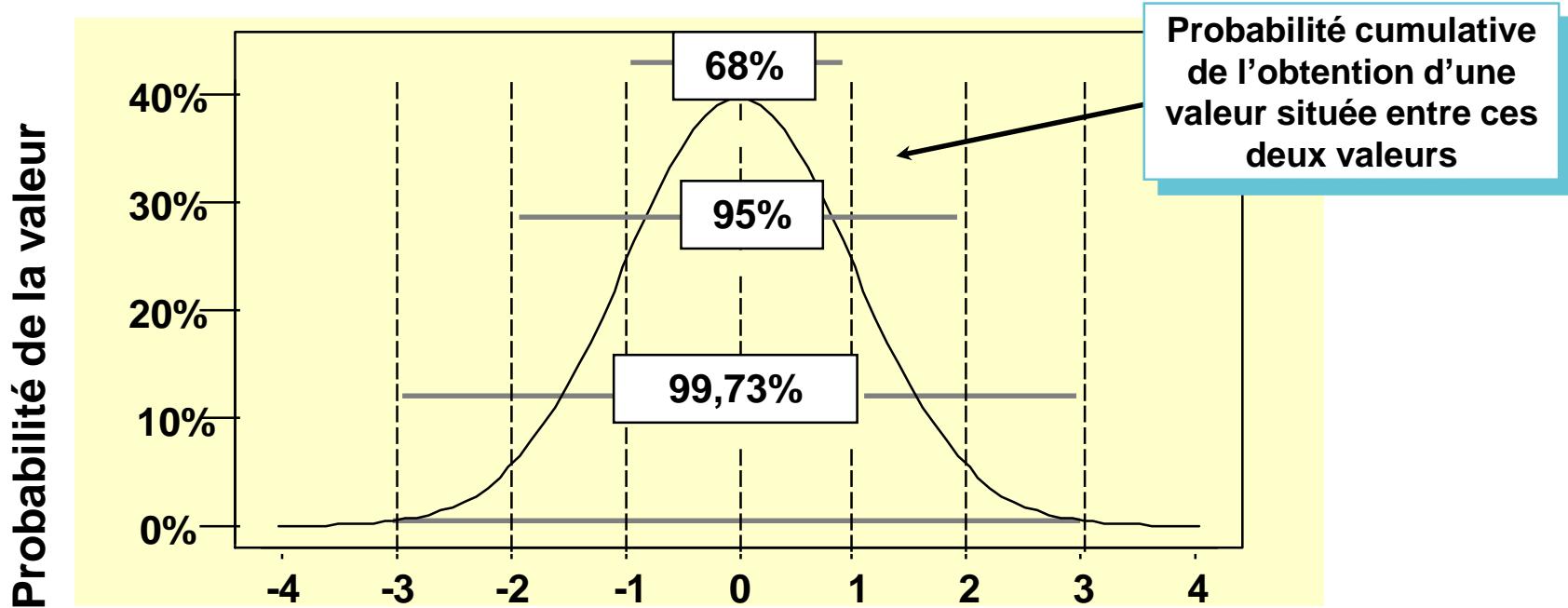
La répartition normale

- **Caractéristique 1:** on peut décrire une répartition normale en connaissant seulement:
 - la moyenne et
 - la déviation standard



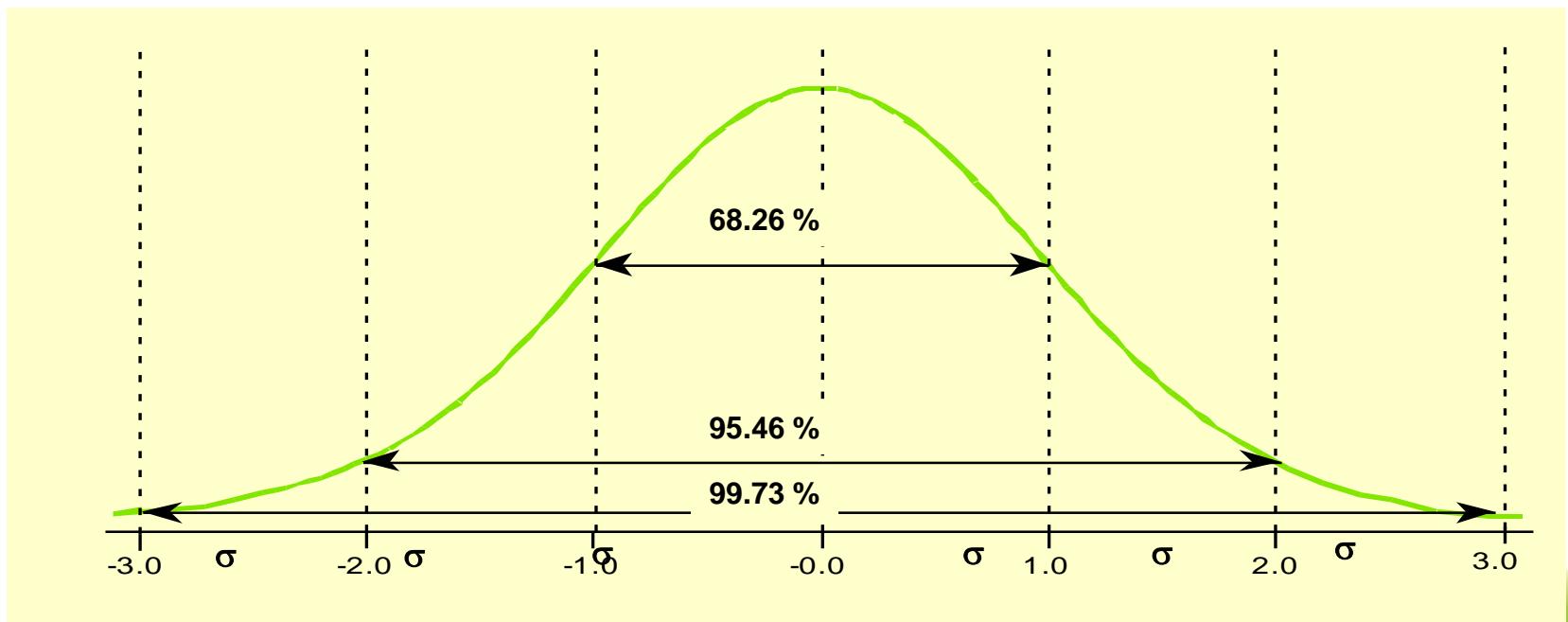
Qu'est-ce qui différencie ces trois répartitions normales ?

- **Caractéristique 2:** la surface en-dessous de la courbe peut être utilisée pour estimer la probabilité cumulative de la survenance d'un certain "évènement".



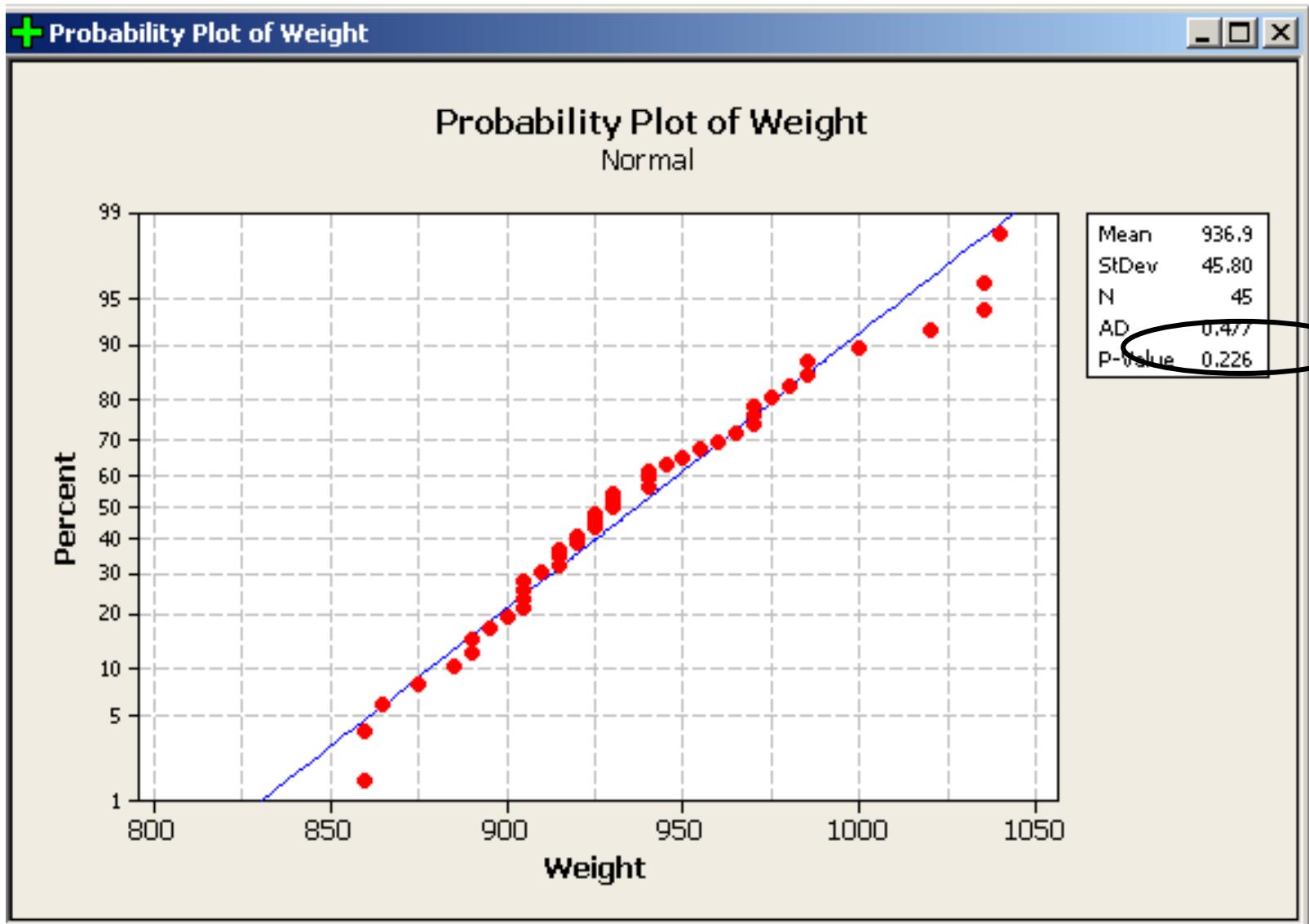
Nombre de déviations standard par rapport à la moyenne

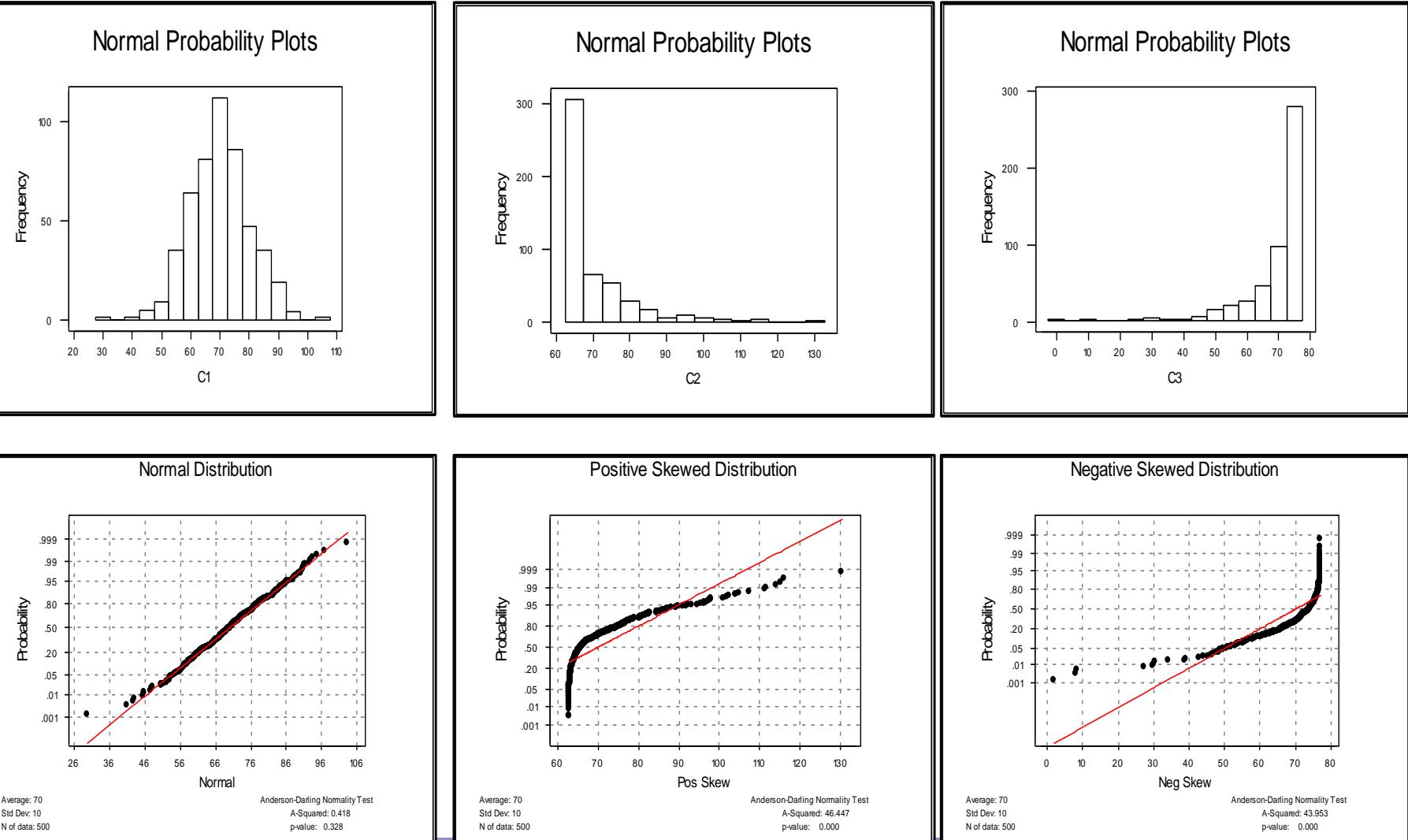
- 68,26 % des données tomberont dans les limites de + / - 1 σ par rapport à la moyenne**
- 95,46 % des données tomberont dans les limites de + / -2 σ par rapport à la moyenne**
- 99,73 % des données tomberont dans les limites de + / -3 σ par rapport à la moyenne**
- 99,9937 % des données tomberont dans les limites de + / -4 σ par rapport à la moyenne**
- 99,999943 % des données tomberont dans les limites de + / -5 σ par rapport à la moyenne**
- 99,999998 % des données tomberont dans les limites de + / -6 σ par rapport à la moyenne**



- Dans certaine circonstances si nécessaire, de savoir si les données sont normalement distribuées.
- Pour savoir si la distribution est normale on conduit un test de normalité, on peut utiliser Minitab
- Pour faire un test de normalité on a besoin de lister quelques concepts fondamentales concernant le test d'hypothèse.
- Pour conduire un test d'hypothèse, on a besoin de statuer notre hypothèse et décider le niveau de risque qu'on prend pour se tromper.

- Nous pouvons tester si un ensemble de données peut être décrit comme “normal” grâce à un test appelé le graphique de probabilité normale.
- Si la répartition est proche de la normale, le graphique de probabilité normale sera en ligne droite.
- Minitab permet de créer facilement un graphique de probabilité normale Anderson Darling





- Les règles précédentes de la probabilité cumulative s'appliquent même si un ensemble de données n'est pas réparti parfaitement normalement.
- Comparons les valeurs d'une répartition théorique (parfaite) et d'une répartition empirique (concrète).

Nombre de déviations Standard	Théorique Normale	Empirique Normale
+/- 1 σ	68%	60-75%
+/- 2 σ	95%	90-98%
+/- 3 σ	99,7%	99-100%

- **Les prévisions nécessitent 2 estimations et un tableau :**
 - Estimation de $\mu = X \text{ barre}$
 - Estimation de $\sigma = s$
 - Tableau Z
- **Transformation en Z :**
 - $Z = (X - \mu) / \sigma$

La transformation en Z :

$$Z = \frac{(x - \mu)}{\sigma} = \frac{(x - \bar{x})}{s}$$

- Cette transformation produit une “valeur” de distribution où : Moyenne = 0 et sigma = 1
- La valeur Z indique combien de déviations standard “entrent” dans la distance entre X (tout nombre intéressant, comme une limite de spécification) et μ (la moyenne d'une distribution donnée)
- Pour la prévision des niveaux de défauts (ou probabilité estimée), nous pouvons utiliser la moyenne actuelle et l'écart type du procédé et substituer la limite inférieure et la limite supérieure de la spécification (unes à unes) pour x

En utilisant cette méthode, nous pouvons calculer le score Z (ou Sigma) du procédé, les niveaux PPM et la probabilité de défaut

- Analysons un exemple.

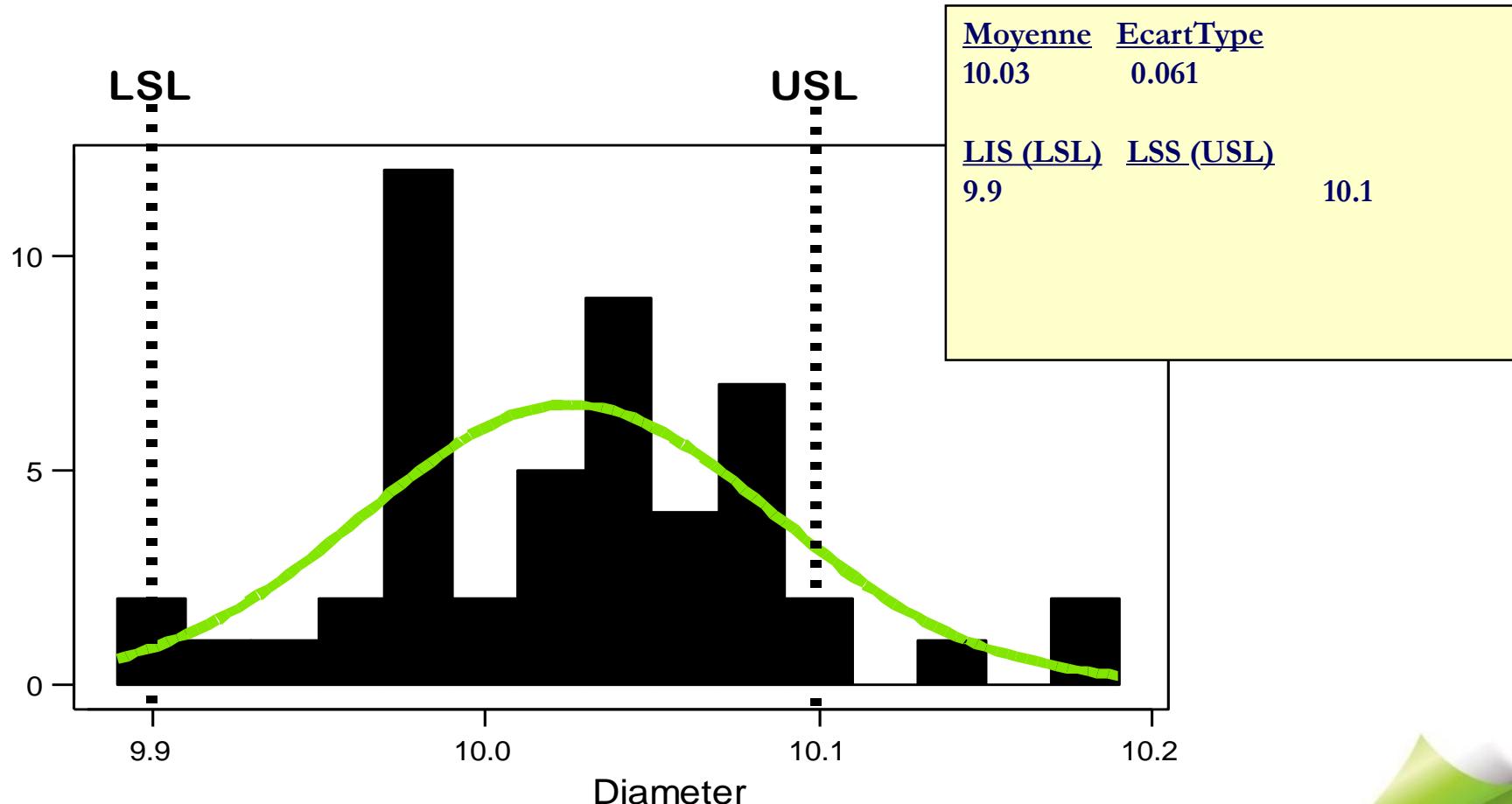
La conversion Z:

$$Z = \frac{(x - \mu)}{\sigma} = \frac{(x - \bar{x})}{s}$$

- Cette “transformation” convertit n’importe quelle distribution normale (avec une moyenne d’échantillon et un sigma d’échantillon) en une répartition standard qui a toujours une moyenne=0 et un sigma=1.
- Que l’on mesure en mm, en pouces, Volts, etc. la répartition transformée aura TOUJOURS une moyenne=0 & sigma=1. Toute distribution est transformée en distribution normale standard grâce au “transformation”Z .
- La valeur z (ou note z), indique l’éloignement d’un chiffre particulier, X, de la moyenne d’échantillon, en unités standard de déviation.
- Par exemple, si z = 2, le chiffre particulier X est à 2 unités standard de déviation de la moyenne d’échantillon.
- Pour prédire les niveaux d’échantillons, (ou le rendement estimé), nous substituons à X la limite inférieure de spécification (LIS) et la limite supérieure de spécification (LSS).
- Nous pouvons ainsi calculer la proportion de produit hors spécifications à partir d’une moyenne d’échantillon et de la déviation standard. Appliquons cette idée aux données du shampoing.

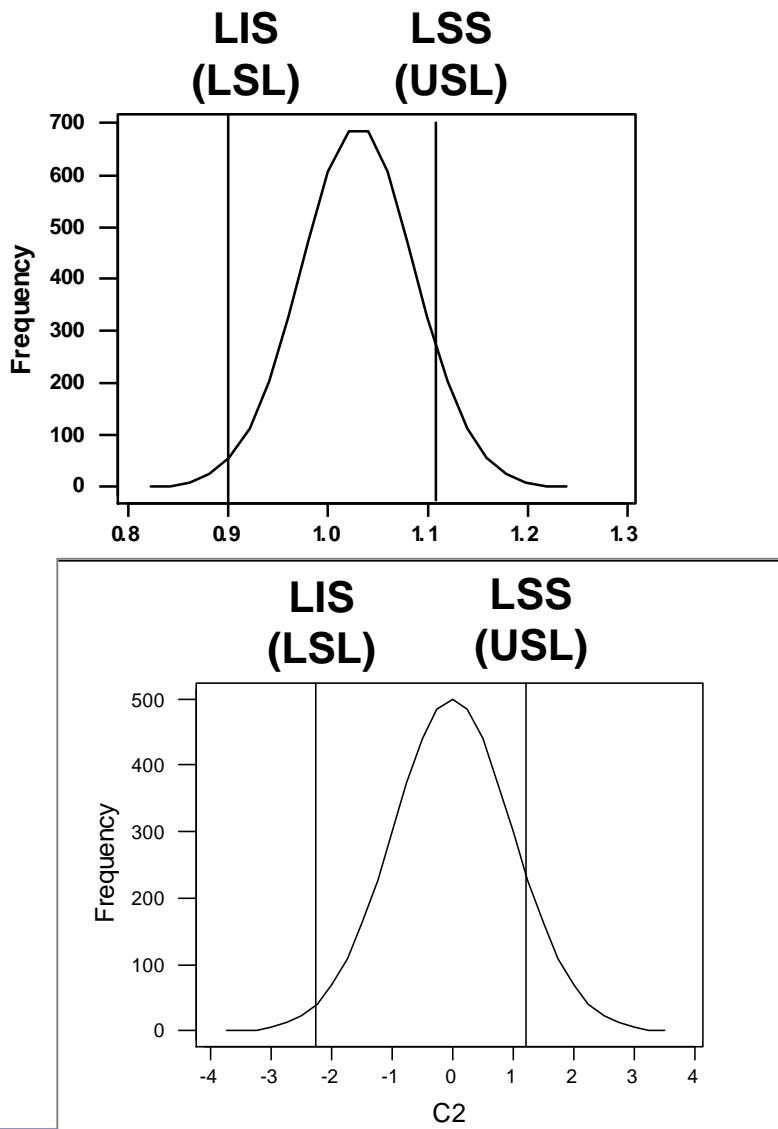
Exemple de transformation en Z

Les données à long terme récapitulées ici ont été collectées à partir du processus d'un tour automatique produisant des broches pour garnitures nues (blank armature shafts)



Problème pratique : Déterminer le % de produits hors spécifications.

Problème statistique : Évaluer la proportion de la courbe normale en-dehors des limites supérieure et inférieure de la spécification. Nous y arrivons en “transformant” les données en une distribution normale standard et en calculant une valeur Z pour chaque limite de la spécification.



La fraction hors des limites de spéc. peut être estimée de la façon suivante :

$$Z_U = \frac{(USL - \bar{x})}{\sigma}$$

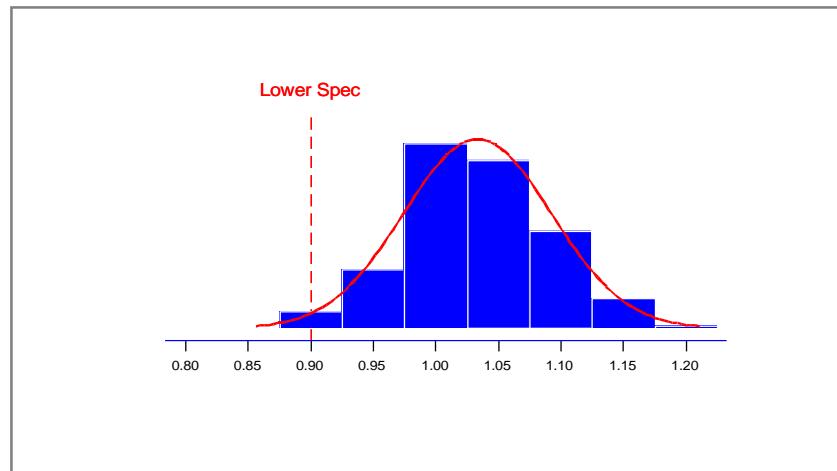
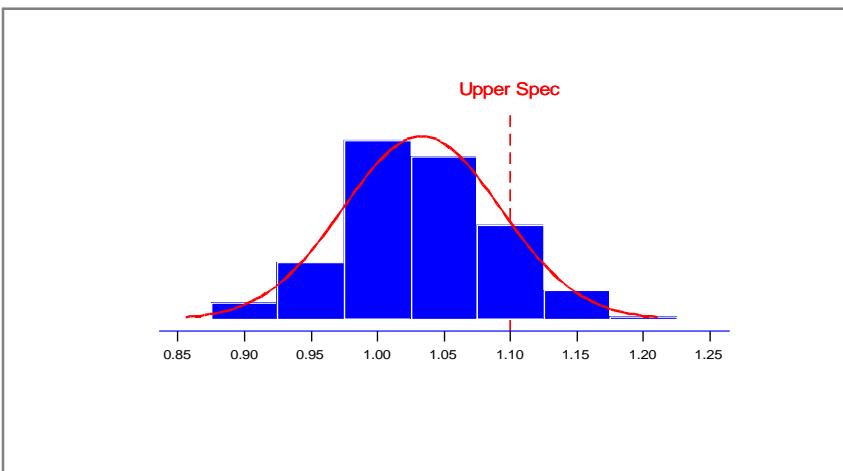
$$= \frac{(10.1 - 10.03)}{.061}$$

$$= 1.15$$

$$Z_L = \frac{(LSL - \bar{x})}{\sigma}$$

$$= \frac{(9.9 - 10.03)}{.061}$$

$$= -2.13$$



$$\Pr(x \leq 0.9) + \Pr(x \geq 1.1) = \Pr(Z \leq -2.13) + \Pr(Z \geq 1.15)$$

$$= 1.7\% + 12.5\%$$

$$\cong 14.2\%$$

Transformation Z

Où allons-nous trouver ces probabilités ?

Méthode 3 : utiliser les fonctions de répartition des probabilités de Minitab.

Z_{LIS}

Fonction de répartition cumulative
Normale avec moyenne = 0 et dév std
= 1,00

x	$P(X \leq x)$
-2,3560	0,0092

Z_{LSS}

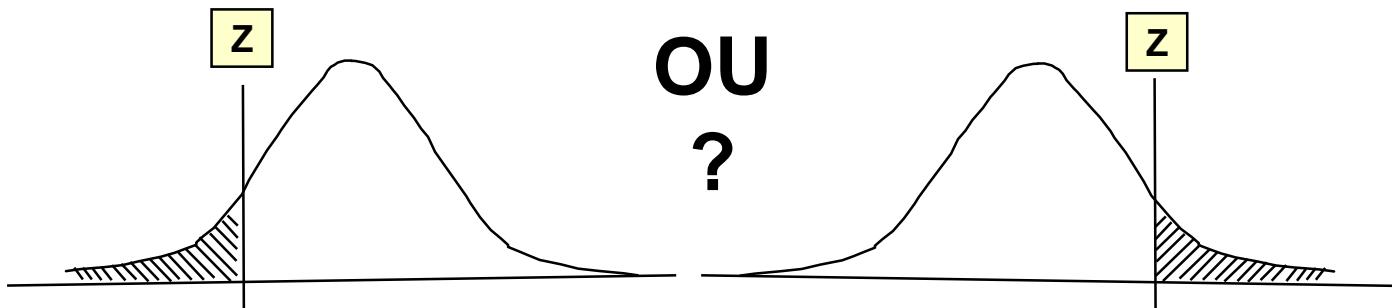
Fonction de répartition cumulative
Normale avec moyenne = 0 et dév std
= 1,00

x	$P(X \leq x)$
1,0340	0,8494

$$\text{Probabilité } (Z < -2,356) = 0,0092$$

$$\text{Probabilité } (Z < 1,034) = 0,8494$$

Question: Minitab donne-t-il un biseautage à droite ou à gauche de la répartition ?



OU
?

Table de conversion

Comment décider s'il faut additionner ou soustraire 1.5 d'une évaluation Sigma

Convertir DE

	Z court terme	Z long terme
Z court terme	Pas d'action	+ 1.5 σ
<u>Convertir EN</u>		
Z long terme	- 1.5 σ	Pas d'action

- Les données à court terme sont sans causes attribuables. Par conséquent, elles ne représentent que l'effet des causes aléatoires.
- Les données à long terme reflètent l'influence des causes aléatoires aussi bien que celle des phénomènes attribuables.
- Si les données de probabilité ou de défaut ont été collectées sur un grand intervalle de production, considérez la situation comme étant à long terme. Dans les autres cas, partez du principe qu'il s'agit de court terme.

- **Qu'est-ce qui peut causer un décalage de 1.5σ**
 - Équipe de production
 - Opérateur
 - Machine
 - Usure de l'outil
 - Arrêt pour réparation
 - Étalonnage
 - Température
 - Humidité
 - Nouvelle matière
- **LES DERIVES EXISTENT ! Nous devons donc en tenir compte.**

- **Les Statistiques nous permettent de comprendre les processus d'une façon à nous permettre de prédire le future performance au lieu de seulement détecter les problèmes actuels**
- **Nous donne une idée de ce qui se passe**
- **Nous permet de comprendre le comportement de la population entière via des échantillons représentative du processus**
- **Nous permet de prendre des décisions avec un certain niveau de confiance**
- **Savoir la capacité actuelle du processus et prendre ultierierement les décisions pour améliorer**

Intervalle de Confiance

- Les statistiques telles que les déviations standard et la moyenne ne sont que des estimations des valeurs Mu et Sigma et se basent sur des échantillons.
- Etant donné qu'il y existe une variabilité d'un échantillon à l'autre, nous pouvons quantifier cette incertitude à l'aide des Intervalles de Confiance basés sur les statistiques.
- La plupart du temps, nous calculons des Intervalles de Confiance de 95% (IC).
- Ces derniers sont interprétés comme suit:
 - Environ 95 sur 100 IC contiennent le paramètre de population, ou
 - nous sommes certains à 95% que le paramètre de population se situe à l'intérieur de l'intervalle.

- Si l'on revient sur ce que nous venons de dire, nous avons vu qu'environ 95% de toutes les moyennes d'échantillons sont à deux Erreurs Standard de la Moyenne de population.

Nous pouvons donc dire que si nous prélevons un échantillon au hasard dans un processus et en calculons la moyenne, nous serons sûrs à 95% d'être à deux Erreurs Standard du paramètre de population.

➤ Pratiquement:

L'intervalle de confiance (C.I.) est une fourchette de valeurs qui inclue, avec une probabilité pré définit nommé niveau de confiance, la valeur réelle des paramètres de la population

➤ Statistiquement:

A [100 (1 - α)]% intervalle de confiance des paramètres de la population, mu ou sigma, est un intervalle aléatoire:

- Probabilité [$\text{Inf C.I.} < \mu < \text{Sup C.I.}$] = $1 - \alpha$
- Probabilité [$\text{Inf C.I.} < \sigma < \text{Sup C.I.}$] = $1 - \alpha$

➤ C'est quoi Alpha (α)?

Le risque maximum ou probabilité de **rejet** de l'hypothèse nulle quand il est **vraie** (connu aussi comme erreur type I ou niveau de signification). Cette probabilité est toujours supérieur à 0, et souvent établit à 5%.

Les intervalles de confiance paramétriques prennent cette forme générale:

$$C.I. = \text{Statistique } e + / - K * \frac{s}{\sqrt{n}}$$

quand :

$$C.I. = \text{Statistique } e + / - K * (\text{erreur standard})$$

statistique e = moyenne des échantillons (\bar{x}) or Ecart type des échantillons (s)

K = Constante varie selon le type de la distribution

Les intervalles de confiance reflètent la variation de nos estimations ponctuelles d'un échantillon à l'autre.

Nous pouvons observer les Intervalles de Confiance pour:

\bar{x}, σ_x, Cp et Proportion de défauts

- Pour la moyenne, k est une valeur-t
- Pour la déviation standard, k est une fonction de la distribution Chi au carré

- Les intervalles de confiance paramétriques supposent une distribution-t des moyennes d'échantillons et utilisent ceci pour calculer les Intervalles de Confiance.
- La formule générale des Intervalles de Confiance Paramétriques pour la moyenne est:

$$\boxed{\bar{x} - t_{\alpha/2,n-1} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1} \frac{\sigma}{\sqrt{n}}}$$

\bar{x} = Moyenne des échantillons

σ = standard deviation des échantillons

n = Nombre des échantillons

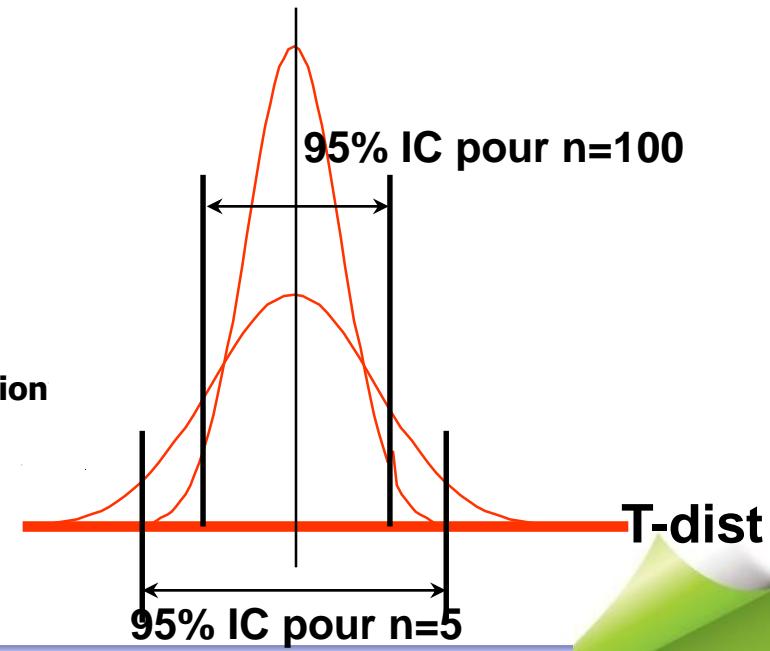
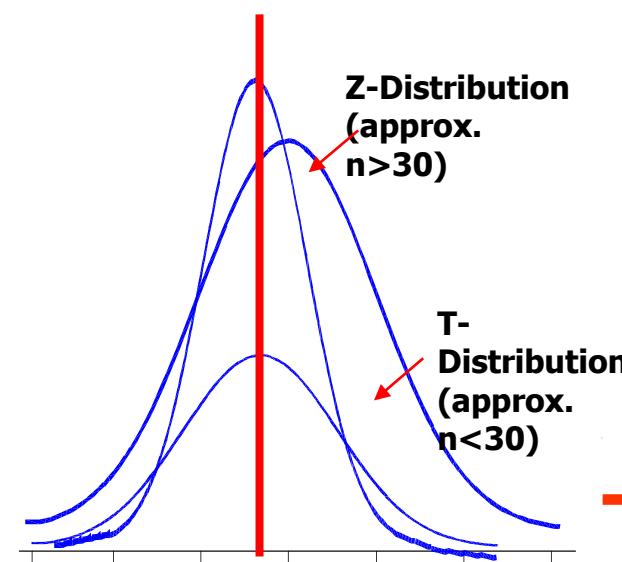
$t_{\alpha/2,n-1}$ = t - valeur de la probabilité $\alpha/2$ et $n - 1$ degrés de liberté

La distribution de référence ici est la distribution-t. Les distributions-t représentent une famille de distributions en forme de cloche caractérisées par la taille de l'échantillon.

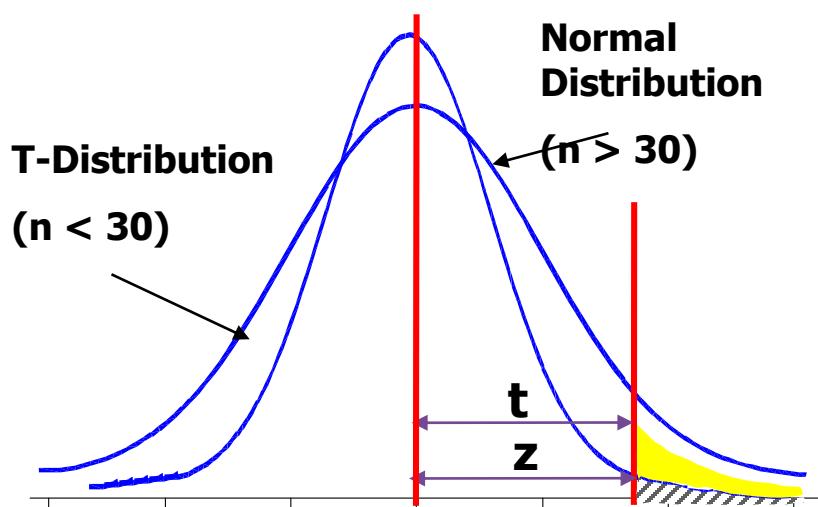
Qu'est-ce que la distribution-t ?

- La distribution-t est une famille de distributions (normales) en forme de cloches qui dépendent de la taille de l'échantillon.
- Plus l'échantillon est petit, plus la distribution est large et plate.
- Pour avoir une idée des valeurs de t pour des intervalles de confiance de 95% pour diverses tailles d'échantillons, regardons le tableau ci-dessous:

Echantillon	Valeur-t
5	2.78
10	2.26
20	2.09
30	2.05
100	1.98
1000	1.96



- Utiliser la table t au lieu du table z pour calculer l'air du queue quand $n < 30$
- Note: L'air sous queue du distribution "t" est plus grand que celle distribution "z"
- Si σ est inconnu, s est la meilleures estimation



$$T_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Si $n > 30$
Si $n < 30$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Souvenez-vous, pour un IC de 95%, on peut en général utiliser +/- 2 Sigma autour d'une moyenne. Si nous connaissons le Sigma de la population, la formule précédente :

$$\bar{x} - t_{\alpha/2,n-1} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1} \frac{\sigma}{\sqrt{n}}$$

serait raccourcie comme ceci:

$$\bar{x} - 2 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2 \frac{\sigma}{\sqrt{n}}$$

car la valeur Mu de la population est à 2 déviations standard de la moyenne d'échantillon.

- Quand nous estimons la performance du processus, nous le faisons sur la base d'un échantillon relativement restreint.
- Voyons d'abord l'Intervalle de Confiance pour la Moyenne. Supposons que nous voulions déterminer l'Intervalle de Confiance de 95% pour la moyenne à partir de 10 échantillons d'un réacteur. Prélevons des échantillons dans le réacteur et nous obtenons:

Moyenne = 249,6

Sigma = 14,15

n = 10

à partir des données suivantes:

263.1 249.2 247.4 263.7 262.4 255.6 252.4 251.5 227.3 223

$$\bar{X} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

- Exemple: Supposons qu'on veut déterminer les paramètre à 95% comme IC pour la moyenne de la population à partir de 10 échantillons (n=10),

La moyenne des échantillons est de 249,56

l'écart type est de = 14.15

$$\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

$$249.56 - 2.262 * \frac{14.15}{\sqrt{10}} \leq \mu \leq 249.56 + 2.262 * \frac{14.15}{\sqrt{10}}$$

$$249.56 - 10.11 \leq \mu \leq 249.56 + 10.11$$

$$239.43 \leq \mu \leq 259.69$$

- Solution: Nous sommes confiants à 95% que la moyenne est entre deux valeurs 239.43 et 259.69

- Exemple: Supposant on prend un échantillons de 16 valeurs et une déviation standard de 1.66. Le degrés de liberté ($n-1$) est 16-1 ou 15. Avec ces données, on peut utiliser la formule de l'intervalle de confiance pour estimer le sigma de la population.

$$s \sqrt{\frac{n-1}{\chi_{\alpha/2}^2}} \leq \sigma \leq s \sqrt{\frac{n-1}{\chi_{1-\alpha/2}^2}}$$

Ou :

$\alpha = 1 - \% \text{ confiance}$

$n = \text{taille 'échantillon}$

➤ Example:

- Note that the χ^2 distribution is not symmetrical, so the value on either side of the confidence interval is different.

$$1.66 \sqrt{\frac{16 - 1}{\chi_{.05/2}^2}} \leq \sigma \leq 1.66 \sqrt{\frac{16 - 1}{\chi_{1-.05/2}^2}}$$

$$1.66 \sqrt{\frac{16 - 1}{\chi_{.025}^2}} \leq \sigma \leq 1.66 \sqrt{\frac{16 - 1}{\chi_{.975}^2}}$$

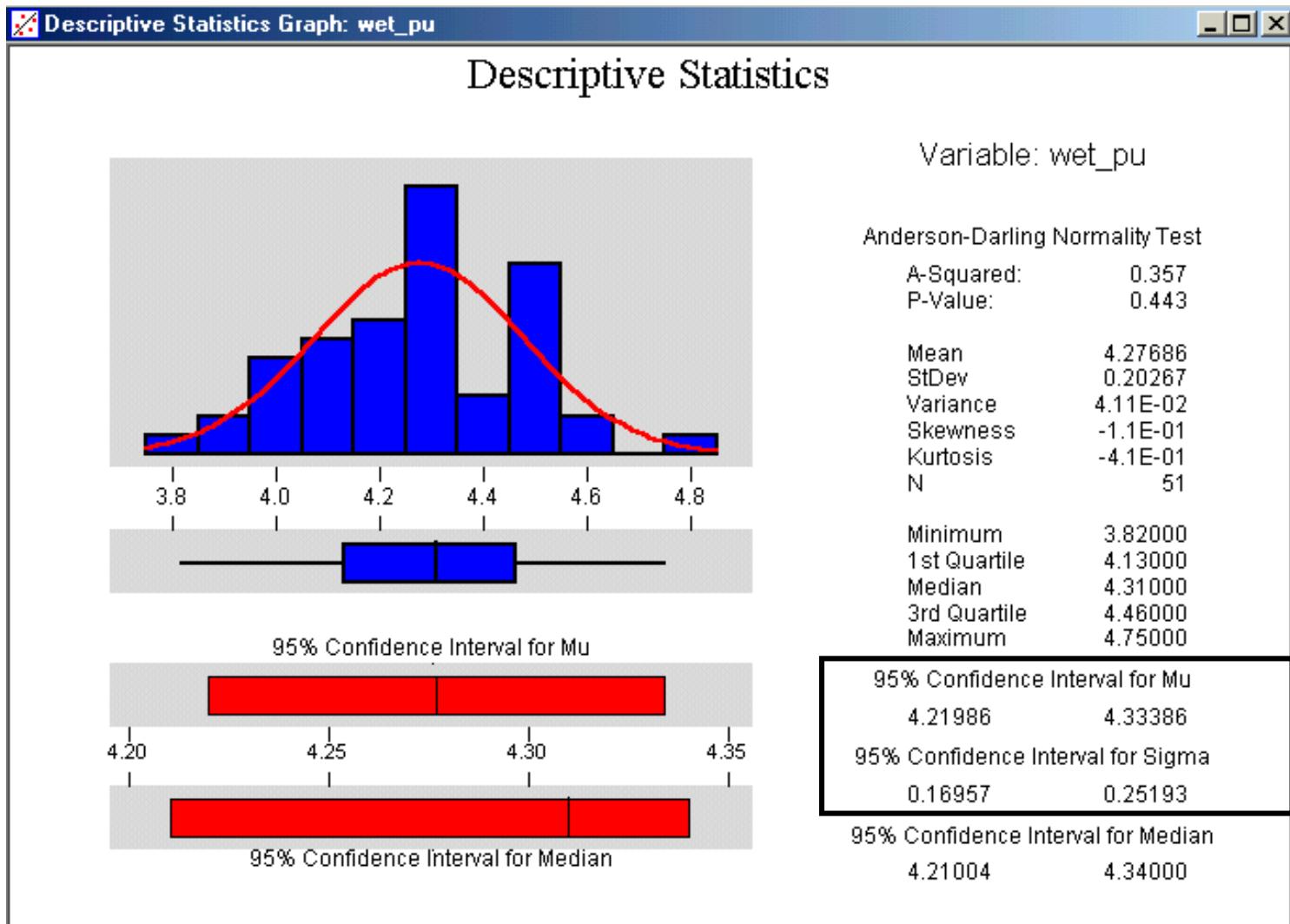
$$1.66 \sqrt{\frac{15}{27.49}} \leq \sigma \leq 1.66 \sqrt{\frac{15}{6.26}}$$

$$1.23 \leq \sigma \leq 2.57$$

- Question: On suppose qu'on collecte six échantillons du processus et on a calculé sigma à 3.6. quel est CI de 95% ?

$$s \sqrt{\frac{n-1}{\chi^2_{\alpha/2}}} \leq \sigma \leq s \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2}}}$$
$$\underline{s} \sqrt{\frac{n-1}{\chi^2_{\alpha/2}}} \leq \sigma \leq \overline{s} \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2}}}$$
$$2.23 \leq \sigma \leq 8.82$$

- Conclusion: On a prédit que 95% des échantillons vont avoir une déviation standard entre 2.23 and 8.82, ou nous sommes confiant que 95% que la déviation standard du processus va être entre 2.23 et 8.82.



Que veut dire ces valeurs ?

- Les défauts sont d'habitude plus faciles à mesurer et à enregistrer que la probabilité de chaque opération d'un procédé
- Si les défauts sont utilisés pour prévoir les niveaux de qualité, il faut collecter des données à long terme.
- Le calcul des occurrences est un outil utile permettant de définir les priorités des activités d'amélioration complémentaires lorsqu'une entreprise a atteint un niveau de qualité globalement élevé (5.0 Sigma).
- En général, les défauts augmentent les temps de cycle, les coûts, les travaux en cours et le stock des produits finis. De plus, ils restreignent les capacités.

- **Les Statistiques nous permet de comprendre nos processus de manière à prévoir le futur au lieu de détecter les problèmes**
- **Nous donne une image de ce qui se passe dans nos processus**
- **Nous permet de comprendre le comportement de la population à travers des échantillons**
- **Nous permet de prendre des décision à un certain niveau de confiance**
- **Comprendre la capacité actuelle du processus et se prendre les bonnes décision et choix pour l'améliorer ultérieurement**

Le Théorème de la Limite Centrée

- Pourquoi apprendre ce concept ?
 - C'est le concept fondamental des statistiques par déduction et la base des outils que nous allons apprendre à utiliser.
 - Le « théorème de la limite centrée » est rarement appliqué sous sa forme la plus pure bien qu'il puisse être utilisé par les inspecteurs pour réduire les erreurs de jauge (que nous aborderons plus loin).
 - Les intervalles de confiance sont dérivés du théorème de limite centrée. Ils sont utilisés pour quantifier un niveau de certitude ou d'incertitude qui concerne un paramètre de population basé sur un échantillon.

Théorème de la limite centrale

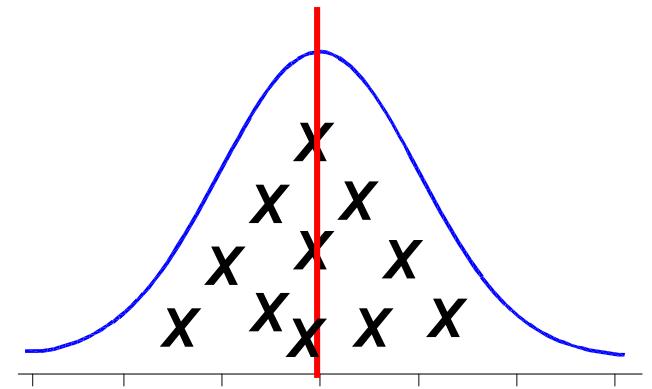
- Consiste en 3 règles pertinentes pour l'outil que nous allons utiliser

La distribution sous-jacente mise à part,

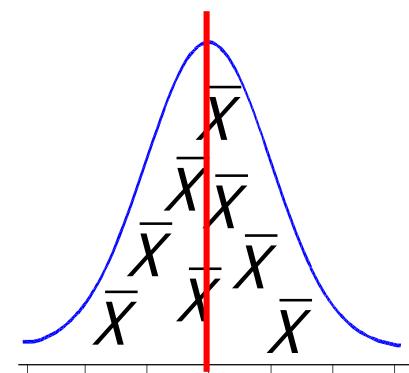
- les moyennes des échantillons de la distribution tendront à être normalement distribués pour un échantillon suffisamment grand

$$\overline{X}_{\text{individus}} = \overline{\overline{X}}$$

$$\sigma_{\overline{x}} = \frac{\sigma_{x_i}}{\sqrt{n}}$$

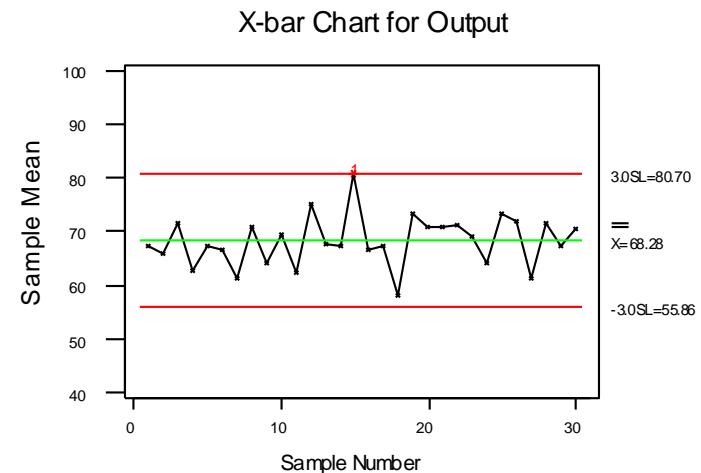
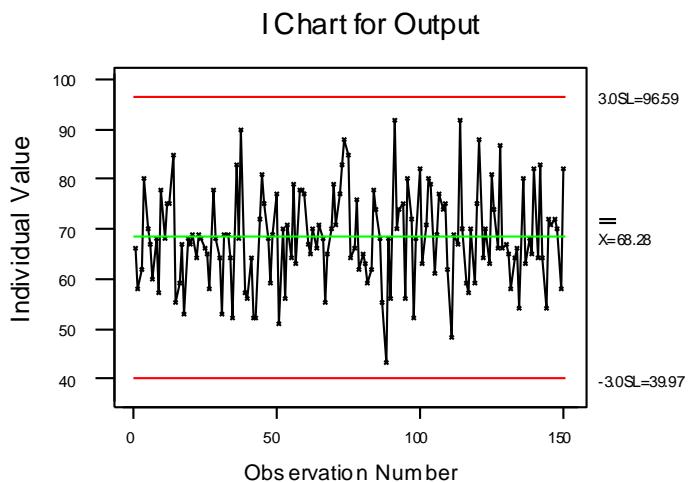


Distribution of Individuals



Distribution of Means

- Après tout, quelle est la différence des limites de contrôle, il s'agit des mêmes données ?
- Que peut-on dire des limites des contrôles ?



Valeurs individuelles

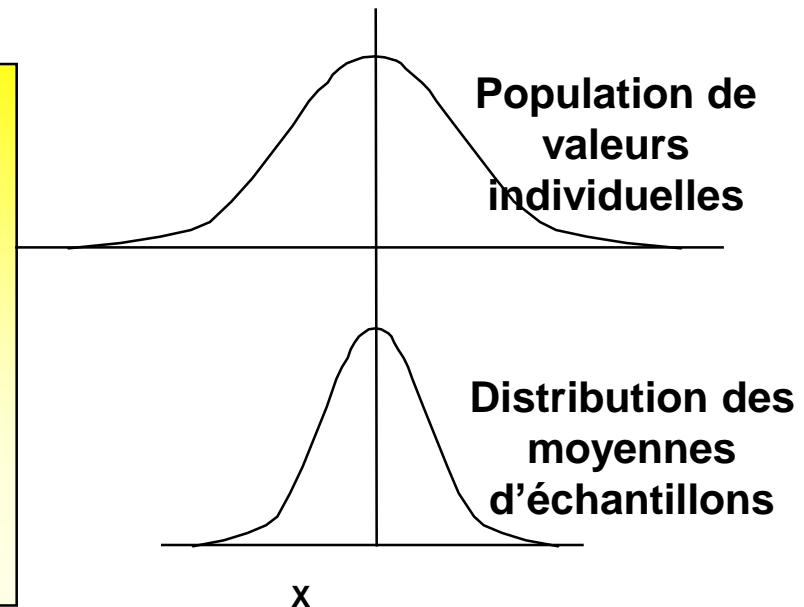
Moyennes d 'échantillon (barres X)

- Si l'on prélève tous les échantillons possibles au hasard d'une taille « n » dans une population de valeurs individuelles ayant une moyenne (μ) et une déviation standard (σ) connues, la moyenne des moyennes d'échantillons sera la moyenne de la population:
- En outre, la déviation standard des moyennes d'échantillons sera calculée approximativement:

$$\bar{X}_{\text{Moy.}} \quad \bar{\bar{X}} = \mu_{\text{individuelles}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma_{\text{individuelles}}}{\sqrt{\text{Echantillon}}}$$

$$Z_{\bar{x}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \quad \text{ou} \quad Z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma_x}$$



- Note importante: ces concepts sont valables lorsqu'on prélève des échantillons sur des populations de valeurs individuelles normalement distribuées ou non.

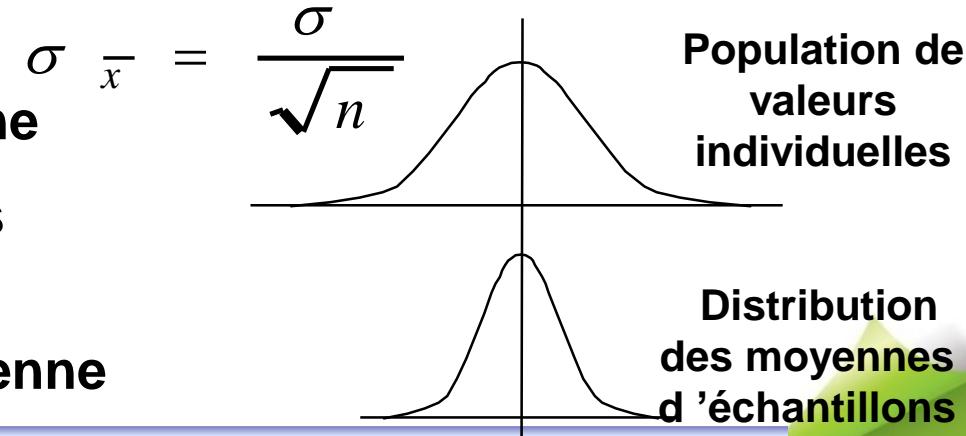
- L'erreur standard de la moyenne (SE Mean) est la déviation standard pour la distribution des moyennes d'échantillons.
- Cette formule montre qu'une « moyenne d'échantillons » est plus stable qu'une observation individuelle par un facteur égal à la racine carrée de la taille de l'échantillon (n).
- La moyenne SE Mean nous dit que la distribution des moyennes d'échantillons a une moindre variance que la population d'origine pour tout $n > 1$
- Si $n = 1$, nous prélevons un échantillon de toute la population une unité à la fois. Dans ce cas, la moyenne $SE = \sigma_x$. Ceci est vrai car les moyennes d'échantillons et la moyenne de population sont les mêmes dans ce cas particulier.

σ_x = Erreur standard de la moyenne

σ = Déviation Standard des notes

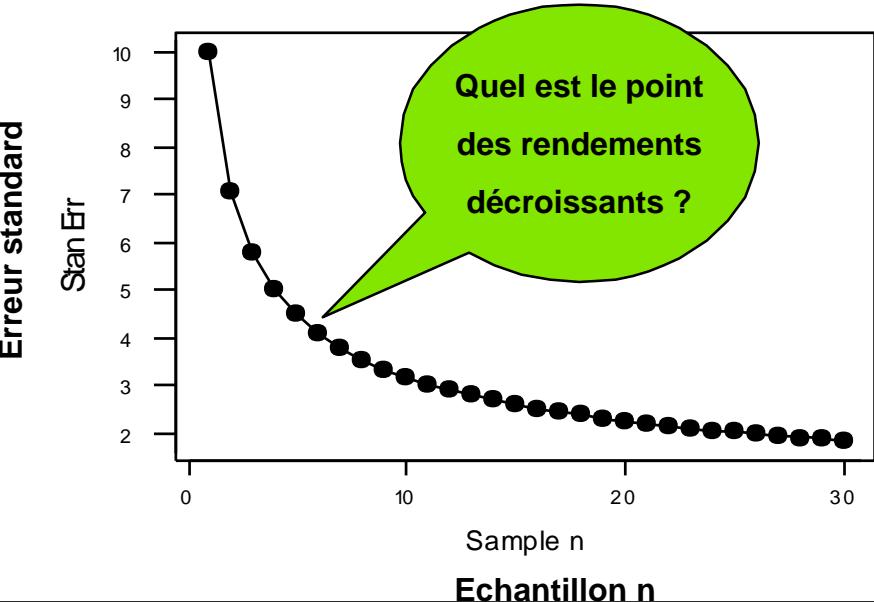
n = individuelles

Taille de l'échantillon pour la moyenne



Let $\sigma_x = 10$ therefore $\sigma_{\bar{x}} = \frac{10}{\sqrt{n}}$
Si alors

Relation entre l'Erreur Standard de la moyenne et la taille de l'échantillon
Relation between Standard Error of the Mean and Sample Size



Pour une déviation standard de population, la moyenne SE décroît quand la taille de l'échantillon augmente.

- On s'appuie en général sur une valeur par pièce fournie par un système de mesure. Cette valeur sert à estimer la « véritable » qualité de notre caractéristique.
- Si nous identifions un problème de répétabilité avec notre jauge, nous pouvons réduire l'erreur du système de mesure à l'aide du TLC en prenant les moyennes de deux valeurs ou plus sur la même pièce.
- La précision de notre système de mesure augmente automatiquement d'un facteur égal à la racine carrée de la taille de l'échantillon (nombre de mesures répétées).
- Autrement dit, nous pouvons réduire considérablement la variation que nous observons dans nos valeurs (due à des erreurs de mesure) en augmentant légèrement la taille de l'échantillon pour les faibles niveaux de n .
- Ceci n'est pas une excuse pour éviter d'avoir à réparer la jauge !

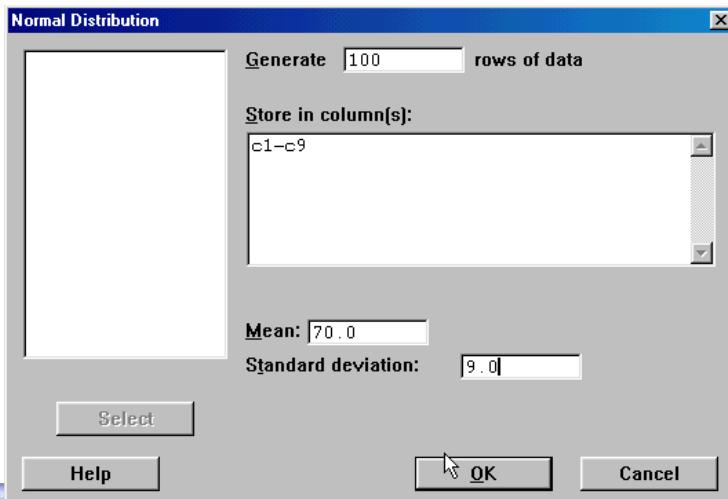
Rappelons: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ aussi $\sigma_{SM(moy.)} = \frac{\sigma_{SM}}{\sqrt{n}}$

$$\sigma_{Total}^2 = \sigma_{Pièces}^2 + \sigma_{SM}^2$$

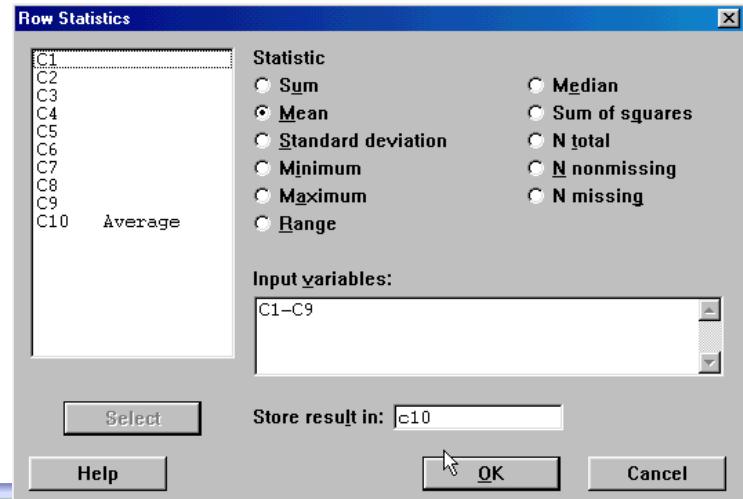
$$\% \text{ Contribution du SM} = \frac{\sigma_{SM}^2}{\sigma_{Total}^2}$$

- Créons des données pour tester le TLC
- Créer 9 colonnes de données ayant une distribution normale
 - Moyenne = 70 et déviation standard = 9
 - Incrire la moyenne des 9 premières colonnes dans C10
 - Question: quelle est la déviation standard attendue pour la colonne C10?

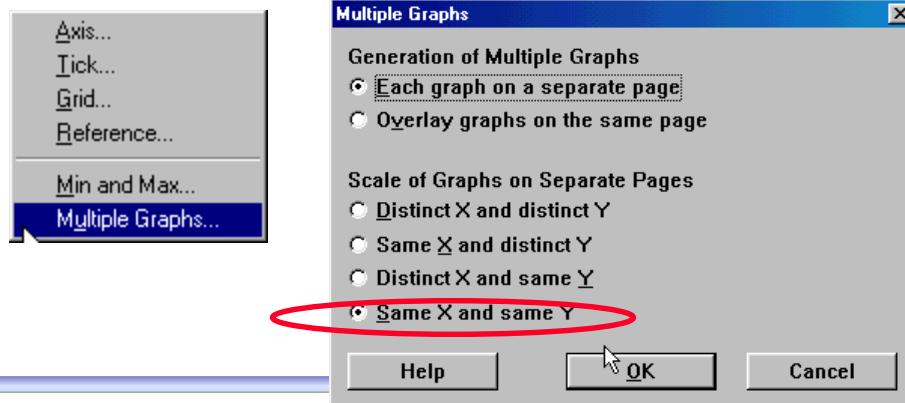
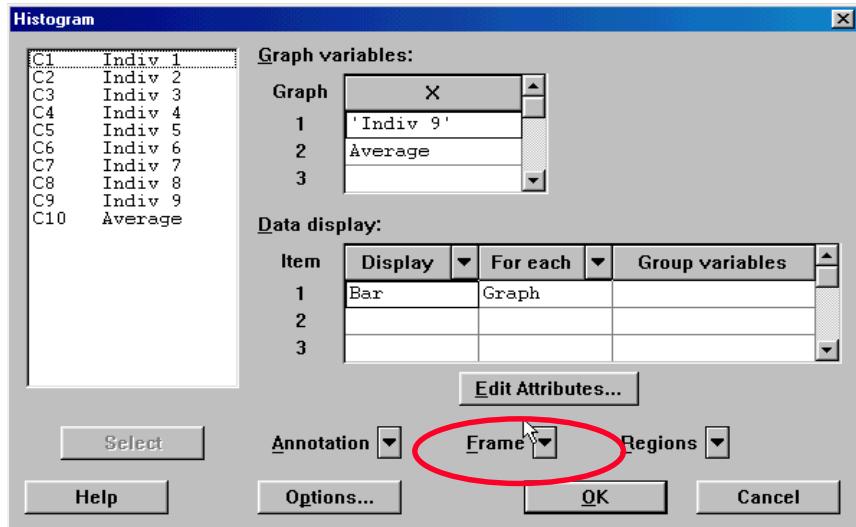
Calc > Random Data > Normal



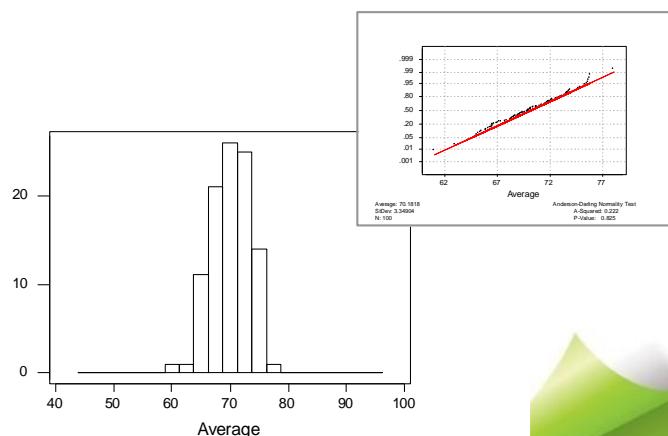
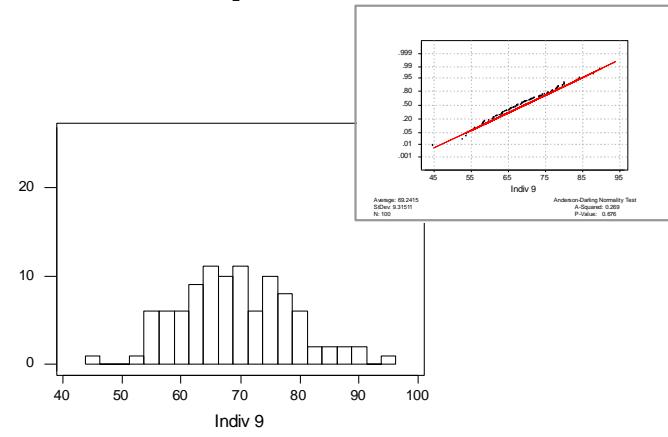
Calc > Row Statistics



Graph > Histogram



- Question: quelle est la déviation standard attendue pour la colonne C10?



Démonstration: Résultats numériques

- La moyenne de la distribution des moyennes d'échantillons est très proche de la moyenne de la population.
- La déviation standard de la distribution des moyennes d'échantillon est la déviation standard de la population réduite par la racine carrée de la taille de l'échantillon.
- La distribution des moyennes d'échantillons approche d'une distribution normale.

Stat > Basic Statistics > Display > Descriptive Statistics

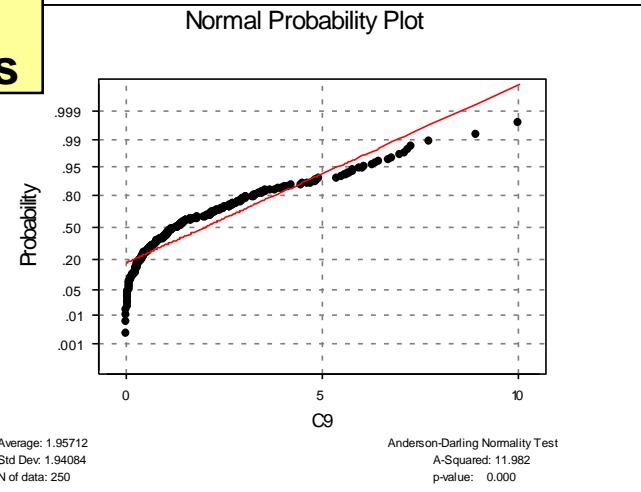
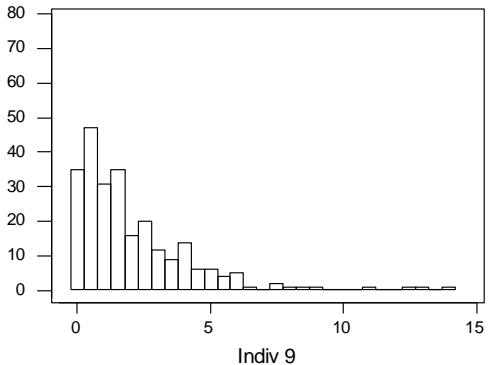
Variable	N	Mean	Median	TrMean	StDev	SE Mean
Indiv 1	100	70.624	70.383	70.640	9.154	0.915
Indiv 2	100	70.684	70.387	70.723	8.324	0.832
Indiv 3	100	69.774	71.068	69.762	9.081	0.908
Indiv 4	100	69.604	69.984	69.548	7.960	0.796
Indiv 5	100	70.905	71.007	71.072	9.145	0.915
Indiv 6	100	69.859	69.726	69.996	8.839	0.884
Indiv 7	100	72.00	72.76	72.10	10.17	1.02
Indiv 8	100	68.95	68.27	69.00	10.14	1.01
Indiv 9	100	69.242	68.757	69.074	9.315	0.932
Average	100	70.182	70.117	70.215	3.349	0.335

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

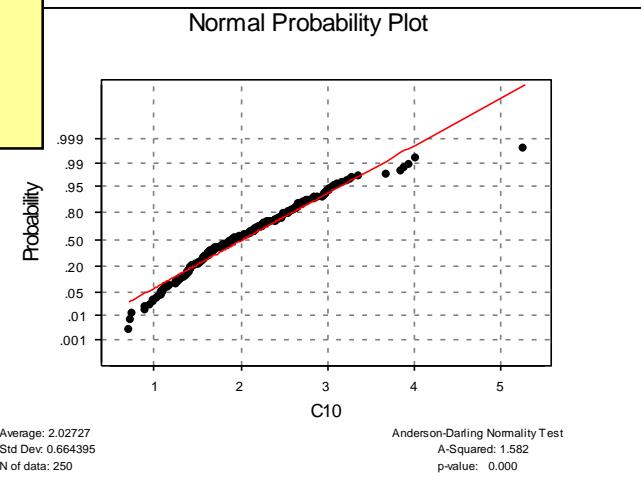
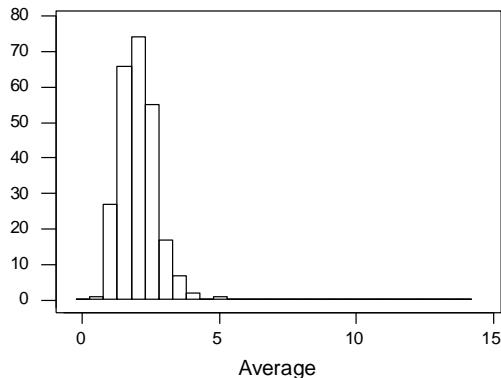
$$\sigma_{\bar{x}} = \frac{9}{\sqrt{9}} = \frac{9}{3} = 3$$

Distribution Non-normale

Distribution des notes individuelles



Distribution des moyennes d'échantillons



- **Le Théorème de Limite Centrée nous permet de supposer que la distribution des moyennes d'échantillons approchera de la distribution normale si le « n » est suffisamment élevé ($n > 30$ pour les distributions inconnues).**
- **Le Théorème de Limite Centrée nous permet aussi de supposer que les distributions des moyennes d'échantillons d'une population normale sont elles-mêmes normales quelle que soit la taille de l'échantillon.**
- **La moyenne SE montre que lorsque la taille de l'échantillon augmente, la déviation standard de la moyenne d'échantillon diminue. L'erreur standard nous aide à calculer les intervalles de confiance.**

Introduction aux tests d'Hypothèses

- On veut savoir si une pièce de monnaie n'est pas truquée, pour répondre à cette question, on lance la pièce plusieurs fois et on note le nombre de pile. Par chance si la pièce est juste quel est le % attendue?
- Si on lance 10 fois la pièce et on a 10 fois piles donc on sera confiant que la pièce est truquée. Mais il y a une chance sur 1000 'avoir 10 piles avec une pièce non truquée. Donc on peut conclure qu'on peut accepter 0.1% chance d'être incorrecte.
- On suppose qu'on accepte qu'on peut confirmer si la pièce n'est pas truquée avec seulement 10 lancées. On lance la pièce 10 fois et on obtient 08 piles quel sera la décision?
- On peut utiliser la probabilité d'une distribution binomiale pour répondre à la question.

“Est-ce qu'il y a une différence significative entre cette pièce et une pièce non truquée?”

Exemple

- Notre supposition avant le test avec une pièce non truquée on attend devoir le pile à 50% des lancées.
- Rappel:

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

- r = le nombre des piles lancées
- n = 10 (total lancées)
- p = .5
- Quel probabilité d'avoir 07 piles ou moins avec une pièce non truquée? Ou quel est la probabilité d'avoir 08 piles ou plus sur 10 lancées avec une pièce non truquée?

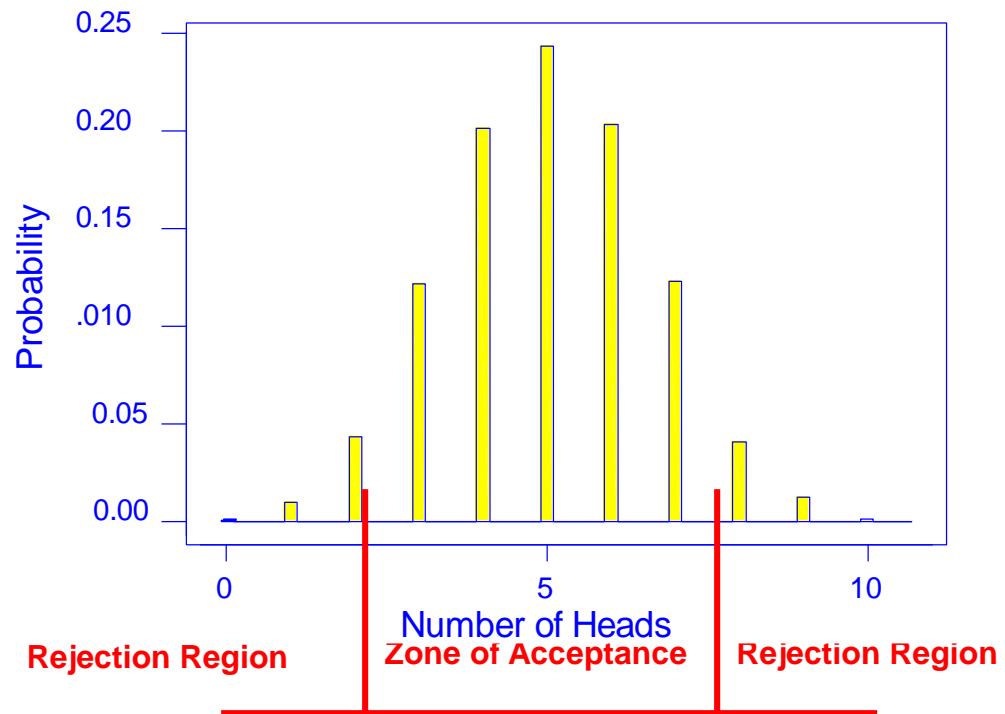
Cumulative Probability Density Function

Heads	Prob	Cum
Prob		
0	.0010	0.0010
1	.0098	0.0108
2	.0439	0.0547
3	.1172	0.1719
4	.2051	0.3770
5	.2461	0.6231
6	.2051	0.8282
7	.1172	0.9454
8	.0439	0.9893
9	.0098	0.9991
10	.0010	1.0000

“Est-ce qu'il y a une différence significative entre cette pièce et une pièce non truquée?”

Exemple

- Ici la distribution binomiale avec la supposition que la probabilité d'avoir l'évènement (d'avoir un pile) est 0.50 pour chaque essai.
- Si on conclu que cette pièce est truquée on prend un risque de 11,7% de se tromper de décision.



- Les statistiques communiquent des informations à partir des données.
- Les statistiques ne peuvent se substituer au jugement d'un professionnel.
- La vérification d 'hypothèse répond à une question pratique:
 - "Y-a-t-il vraiment une différence entre _____ et _____ ?"
- Un problème pratique de processus est traduit en une hypothèse statistique afin de répondre à cette question.
- Dans la vérification d'hypothèse, nous utilisons des échantillons relativement petits pour répondre aux questions sur les paramètres de population.
- Il est toujours possible que nous choisissons un échantillon qui n'est pas représentatif de la population. Par conséquent, il y a toujours un risque que la conclusion soit erronée.
- Avec quelques suppositions, les statistiques par déduction nous permettent d'estimer la probabilité d'avoir un échantillon « bizarre ». Ceci nous permet de quantifier la probabilité (valeur-P) d'une conclusion erronée.

- **La vérification d'hypothèse emploie des tests basés sur des données qui aident à déterminer les quelques X vitaux.**
- **On utilise cet outil pour identifier des sources de variabilité et établir des rapports entre les X et les Y.**
- **Pour aider à identifier les quelques X vitaux, on peut prélever un échantillon de données passées ou actuelles.**
 - Passives: soit vous avez prélevé directement des données dans votre processus, soit vous avez obtenu des données d'échantillons passés.
 - Actives: vous avez modifié votre processus, puis vous avez prélevé des données.
- **Les essais statistiques apportent des solutions objectives à des questions qui ont la plupart du temps une réponse subjective.**

Les termes dont vous devez vous souvenir

1. **Hypothèse nulle (Ho)** - énonce pas de changement ou de différence. Cet énoncé est supposé vrai jusqu'à preuve du contraire.
2. **Erreur Type I** - le risque du « faux positif » qui dit qu'il y a quelque chose de significatif qui se passe alors que ce n'est pas vrai.
3. **Risque Alpha** - le risque ou la probabilité maximum de faire une Erreur de Type I. Cette probabilité est toujours supérieure à zéro, et se situe en général autour de 5%. Le chercheur prend des décisions au plus haut niveau de risque qui est acceptable pour un faux positif.
4. **Degré de signification** - identique au risque Alpha.
5. **Hypothèse alternative (Ha)** - énoncé d'un changement ou d'une différence. On déduit que cet énoncé est vrai si Ho est rejetée.
6. **Erreur de Type II** - le risque du « faux négatif » qui dit qu'il n'y a rien de significatif qui se passe alors que ce n'est pas vrai.

7. **Risque Beta** - le risque ou la probabilité de faire une erreur du type II ou de ne pas remarquer une solution efficace au problème.
8. **Déférence significative** - le terme utilisé pour décrire les résultats d'une vérification d'hypothèse statistique où la différence est trop large pour être raisonnablement attribuée au hasard. Il se passe probablement quelque chose.
9. **Puissance** - l'aptitude d'un test statistique à détecter quelque chose de significatif lorsqu'il y a vraiment quelque chose de significatif. Utilisé en général pour déterminer si les tailles des échantillons suffisantes pour détecter une différence entre les traitements s'il en existe une.
10. **Statistique d 'essai** - une valeur standard (z, t, F, etc.) qui représente la faisabilité d'un faux positif et qui est distribuée d'une manière connue de telle sorte que la probabilité de cette valeur observée puisse être déterminée. En général, plus le faux positif est faisable, moindre est la valeur absolue de la statistique d'essai et plus la probabilité d'observer cette valeur dans la distribution est grande.

	Paramètres de la population	Statistiques de l'échantillon
Moy.	μ	\bar{x}
Déviation Standard	σ	s
Proportion (%)	P	p

1. Les paramètres de population (valeurs) sont fixes, mais inconnus.
2. Les statistiques d'échantillon servent à estimer les valeurs de la population.

Les hypothèses sont des énoncés sur les paramètres de la population, pas des statistiques d'échantillons.

Hypothèse nulle (H₀)

- la supposition
- Interprétation statistique: il n'y a pas de différence entre les moyennes de population A et B.
- Interprétation pratique: il n'y a aucune différence entre les moyennes des rendements des A &B. (vos modifications n'ont servi à rien)

$$H_0 : \mu_a = \mu_b$$

$$H_a : \mu_a \neq \mu_b$$

$$H_0 : \sigma_a = \sigma_b$$

$$H_a : \sigma_a \neq \sigma_b$$

$$H_0 : p_a = p_b$$

$$H_a : p_a \neq p_b$$

Hypothèse alternative (H_a)

- ce que vous voulez déduire
- Interprétation statistique: les moyennes de population pour A et B viennent de distributions différentes;
- Interprétation pratique: le rendement moyen du B diffère de celui du A.

But: Nous devons montrer qu'il est tellement peu probable que les valeurs observées proviennent de la même population, que H_a doit être erronée. (Nous devons rejeter H₀, et par déduction, accepter H_a.)

... Si peu probable ...

Quel degré d'improbabilité ?

(C'est le niveau de signification (α))

Nous aimerais qu'il y ait moins de 10% de risque que ces observations surviennent au hasard ($\alpha = .10$).

Cinq pour cent est bien plus confortable ($\alpha = .05$).

Industry
Standard

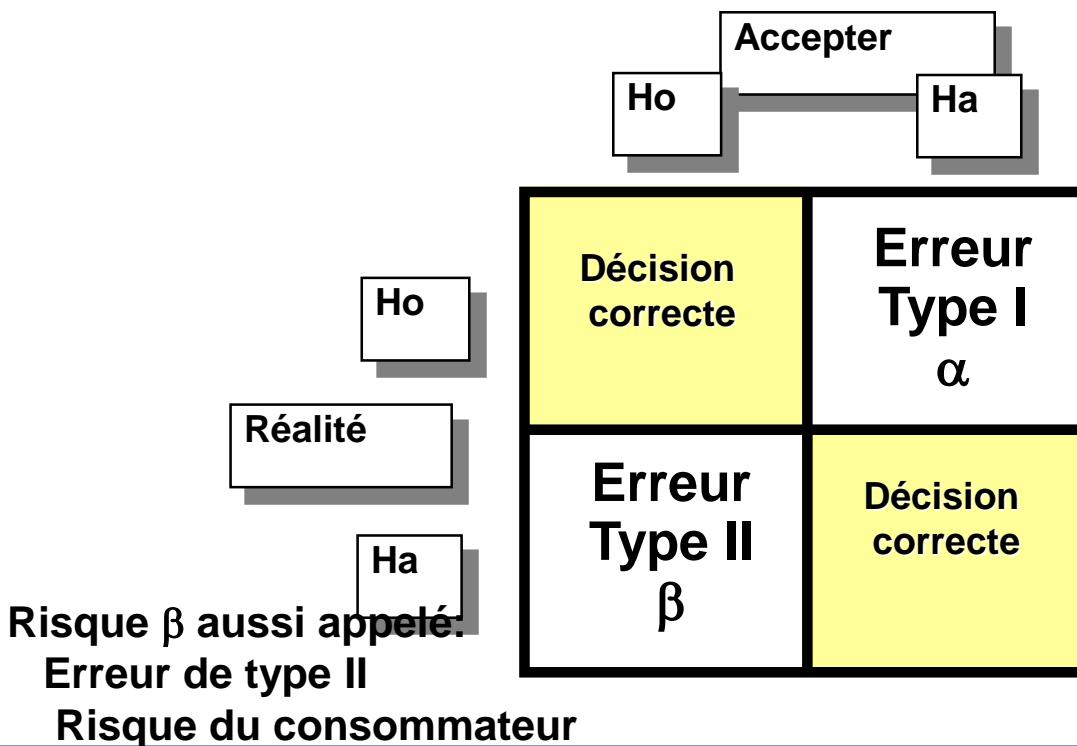
Un pour cent, c'est très bien ($\alpha = .01$).

Ce niveau alpha est basé sur notre supposition d'aucune différence et d'une certaine distribution de référence.

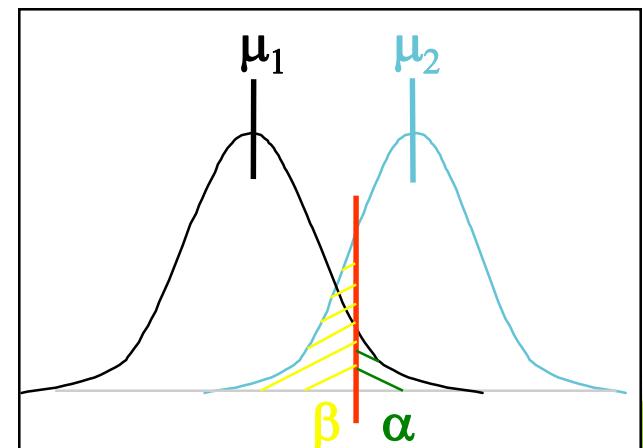
Il y a un risque de α % que nous nous trompons lorsque nous disons que le processus modifié est meilleur.

α -- est le risque de trouver une différence quand il n'y en n'a pas.
Utilisé comme critère de décision pour rejeter H_0 .

β -- est le risque de ne pas trouver de différence quand en réalité il y en a une. C'est aussi le paramètre de déterminer la taille nécessaire des échantillons.



Risque α - aussi appelé:
Erreur de type I
Risque du producteur

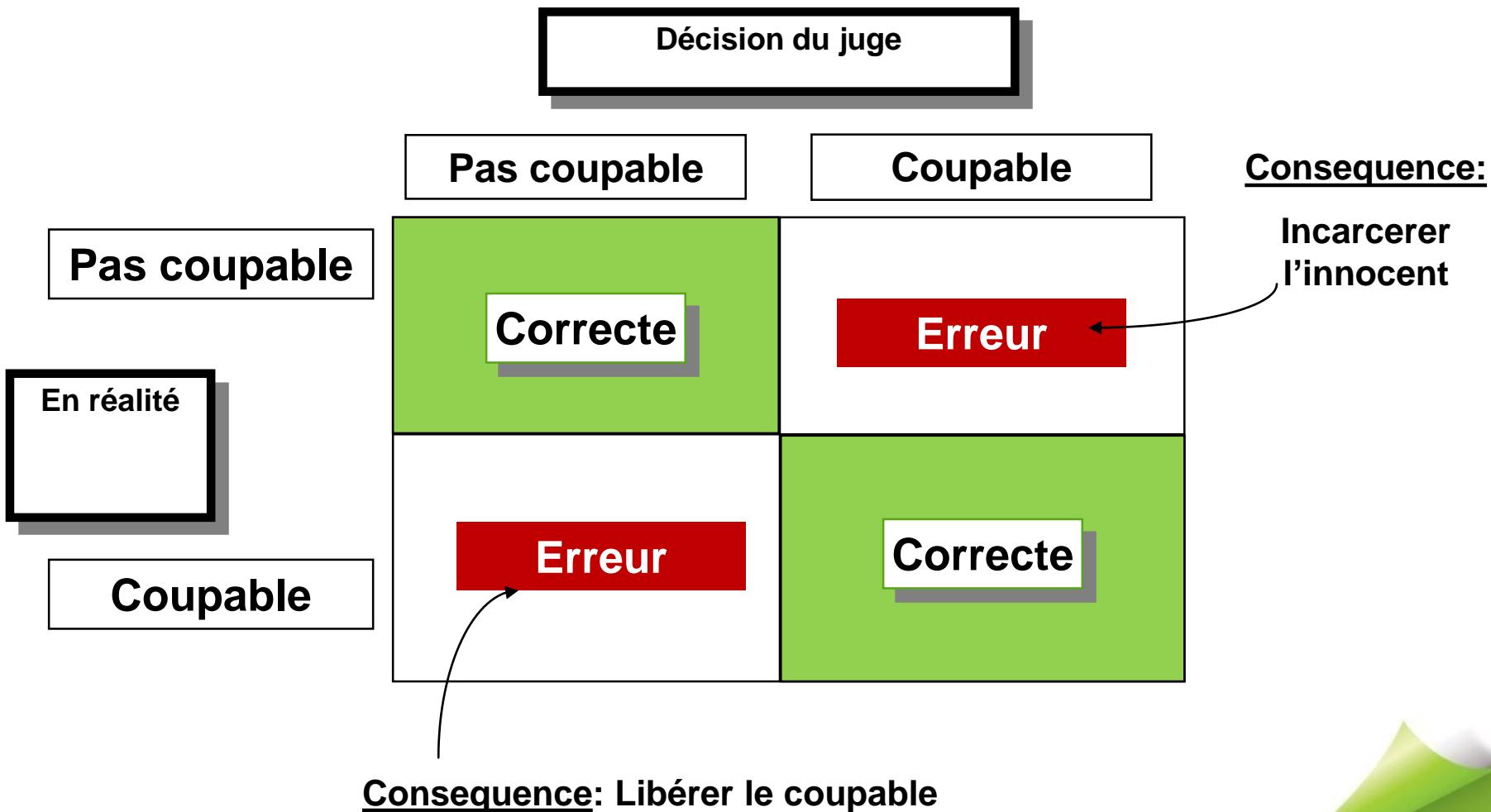




Concernant l' Hypothèses nulle

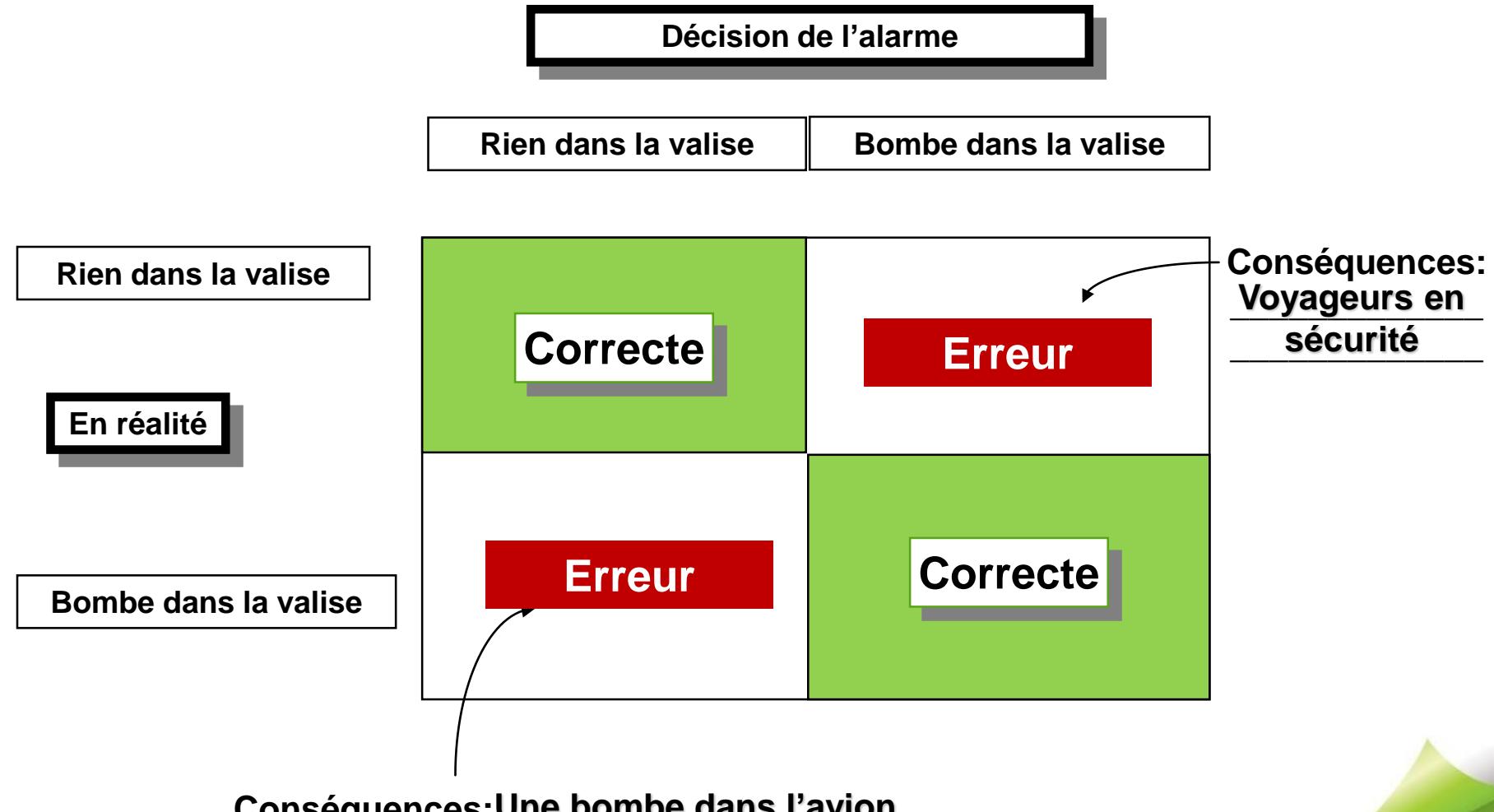
- L' Hypothèse nulle (H_0) est supposée être vraie
 - Comme présumer que l'accusé est “non coupable”
 - ↖ **Rappel:** Le système judiciaire: l'accusé est innocent jusqu'à prouver qu'il est coupable
 - ↖ On assume pas qu'il y a un effet seulement si la probabilité de “non effet” est si faible de le croire
 - ↖ H_0 : Pas d' Effet
 - ↖ H_a : Effet
 - Vous êtes l'avocat de l'accusé tu doit présenter des évidences raisonnables au delà des doutes

L'Hypothèse nulle (H_0) est "pas coupable"

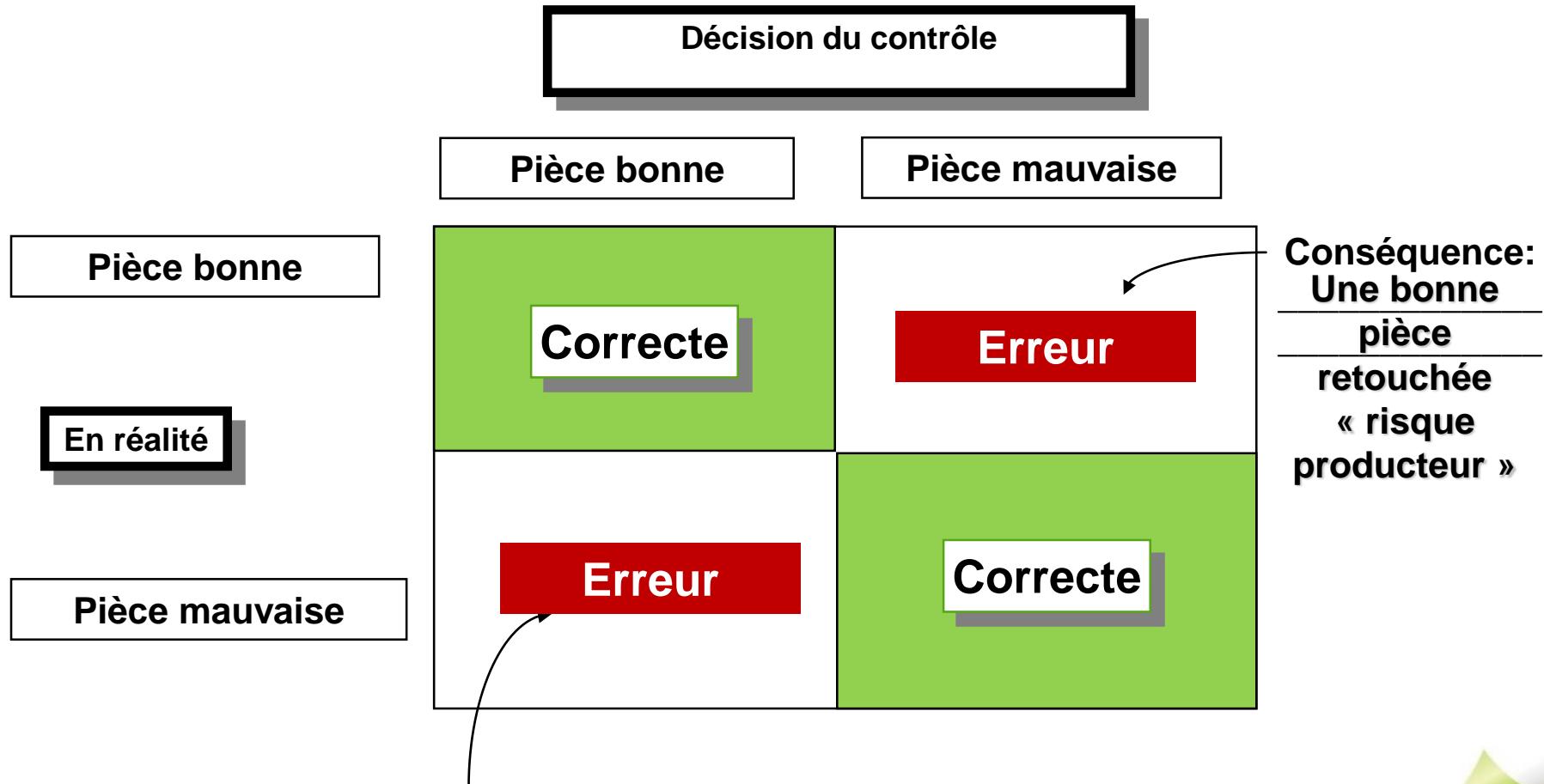


Exemple: Sécurité Airport

L'Hypothèse nulle (H_0) est "il n'y a rien dans la valise"



L'Hypothèse nulle (H_0) est “La pièce est bonne”



- **Notre décision de rejeter ou de confirmer l'hypothèse est basée sur les data**
- **Notre décision peut être erronée sous deux formes:**
 - En rejetant l'hypothèse nulle par erreur: Erreur **Type I**
 - Rejeter l'hypothèse quand il est vrai
 - Cette erreur est connue comme “risque du producteur”
 - Accepter l'hypothèse nulle par erreur : Erreur **Type II**
 - Echouer à rejeter l'hypothèse nulle quand il est faut
 - Cette erreur est connu sous “risque consommateur”
- **On doit spécifier avant l'investigation l'ampleur du risque qu'on est prêt à prendre en commettant ces erreurs**

- La décision est correcte
- Accepter l'hypothèse Nulle
 - C'est l'intervalle de confiance
 - La valeur typique de $(1 - \alpha) = 0.95$
- Rejeter l'hypothèse nulle
 - Elle est appelé “puissance du test”
 - La valeur typique de $(1 - \beta) = 0.80$

**La valeur du P est très Importante :Retenir cette parole:
*If P is Low , H_0 Must Go! , If P is high H_0 is the guy!***

A quel degré P doit être faible? Ça dépend des conséquences de l'erreur type I

les termes de significativité, d'hypothèse nulle, et l'utilisation de la valeur-p. L'hypothèse nulle ne peut jamais être acceptée, mais peut seulement être rejetée par le test statistique. Dans cette approche, la valeur-p est considérée comme une mesure d'à quel point les données plaident contre l'hypothèse nulle. Les seuils suivants sont généralement pris pour référence :

La valeur du P est très importante : Retenir cette parole:
If P is Low , H_0 Must Go! , If P is high H_0 is the guy!

A quel degré P doit être faible? Ça dépend des conséquences de l'erreur type I

$p < 0.01$: très forte présomption contre l'hypothèse nulle
 $0.01 < p < 0.05$: forte présomption contre l'hypothèse nulle
 $0.05 < p < 0.1$: faible présomption contre l'hypothèse nulle
 $p > 0.1$: pas de présomption contre l'hypothèse nulle

- Les données sont collectées
- Un test statistique est calculé sur la base d'un taux **signal-parasite (S/P)** pour ces données comme Z- ou T-Score), et valeur- P
- Si H_0 est vraie (aucune différence entre _____ & _____), alors
 - le taux S/P est très faible et force le test à produire une “valeur-p” élevée
- Si H_a est vraie (véritable différence entre _____ & _____), alors
 - le taux S/P sera élevé et forcera la “valeur-p” à être faible
- La “valeur-p” est la probabilité de l’hypothèse nulle se produisant par hasard.
- La valeur-p est basée sur une distribution de référence supposée ou réelle (distribution normale, distribution-t, chi-au carré ou distribution-F).

Enoncez une “Hypothèse nulle” (H_0)



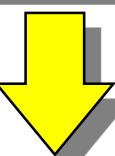
Enoncez l’ “Hypothèse alternative” (H_a)



Etablissez vos critères (α)



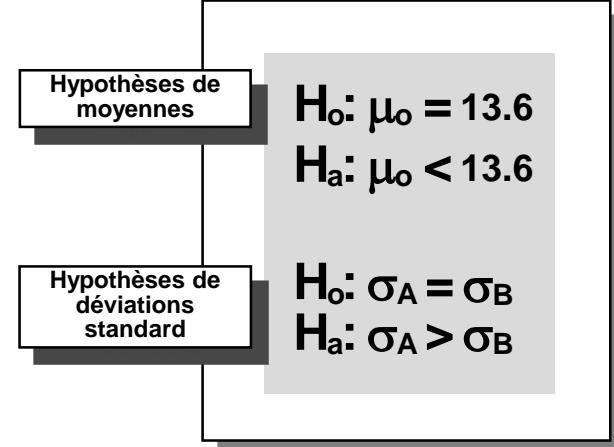
Rassembliez des preuves (échantillon de réalité)



DECIDEZ:

Que suggèrent les preuves ?

Rejeter H_0 ? ou ne pas rejeter H_0 ?



Exemple

- Votre centrale dispose de plusieurs générateurs. Tous se valent à peu près au niveau des performances. Le responsable de la maintenance a décidé de dépenser 100000 \$ pour modifier l'un d'entre eux afin d'améliorer son rendement. Avant de dépenser davantage d'argent, de temps et de ressources à modifier les autres, il veut savoir s'il a amélioré le rendement de façon substantielle. Après avoir prélevé un échantillon du rendement de deux générateurs (l'un modifié et l'autre pas) comment déterminer s'il existe une « réelle différence » entre les deux rendements ?
- Regardons les résultats. Le générateur B est celui qui a été modifié.

Générateur A	Générateur B
89.7	84.7
81.4	86.1
84.5	83.2
84.8	91.9
87.3	86.3
79.7	79.3
85.1	82.6
81.7	89.1
83.7	83.7
84.5	88.5

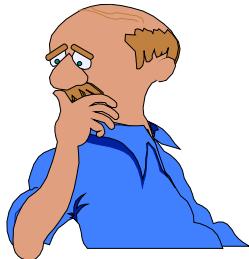
“Y-a-t-il vraiment une différence entre le générateur A et le générateur B?”

Exemple

- **Question pratique:** les modifications du générateur B vont-elles améliorer le rendement par rapport au processus actuel, représenté par le générateur A

Descriptive Statistics				
Variable	Machine	N	Mean	StDev
Yield	A	10	84.24	2.90
	B	10	85.54	3.65

- **Question statistique:** la moyenne du générateur B (85,54) diffère-t-elle assez de celle du générateur A (84,24) pour être considérée comme significative ? Ou les moyennes sont-elles assez rapprochées pour être le résultat du hasard et des variations au jour le jour? Concept statistique:



- Hypothèse statistique: il n'y a pas de différence entre les générateurs
- Ceci s'appelle une hypothèse nulle (Ho)

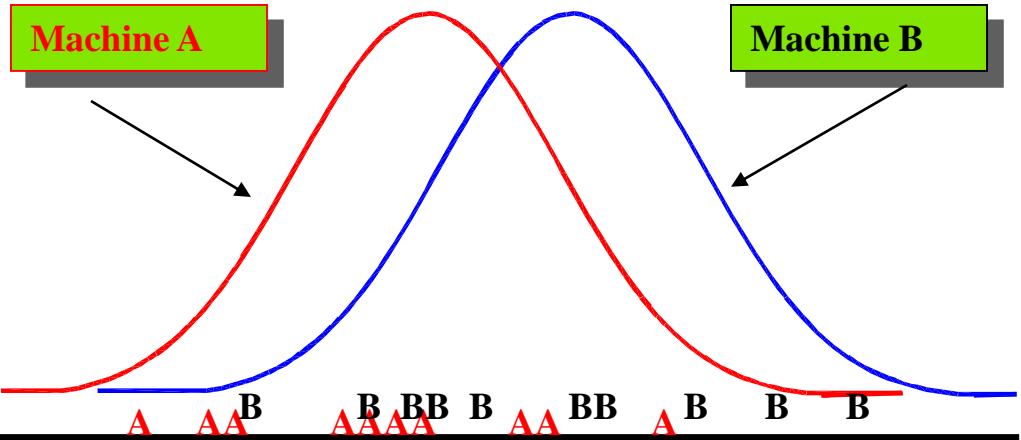
- Hypothèse réelle: le générateur modifié améliore son rendement
- Ceci s'appelle une hypothèse alternative (Ha)

$$\begin{array}{l} \text{Ho: } \mu_a = \mu_b \\ \text{Ha: } \mu_a \neq \mu_b \end{array}$$

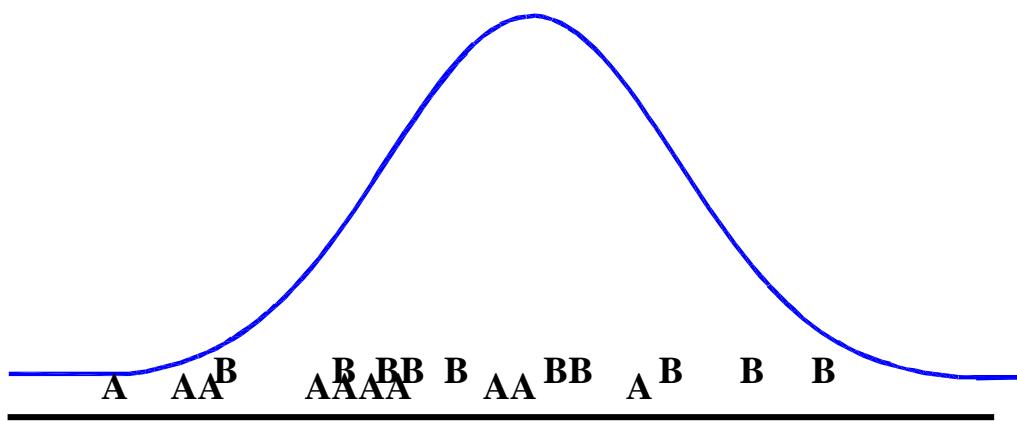
Nous devons montrer qu'il est tellement peu probable que les valeurs observées proviennent du même processus, que Ho doit être erronée.

Statistical Concept:

- En réalité, les rendements des générateurs représentent-ils deux populations différentes ?



- Ou les rendements des générateurs viennent-ils d'une seule population ?



La valeur critique

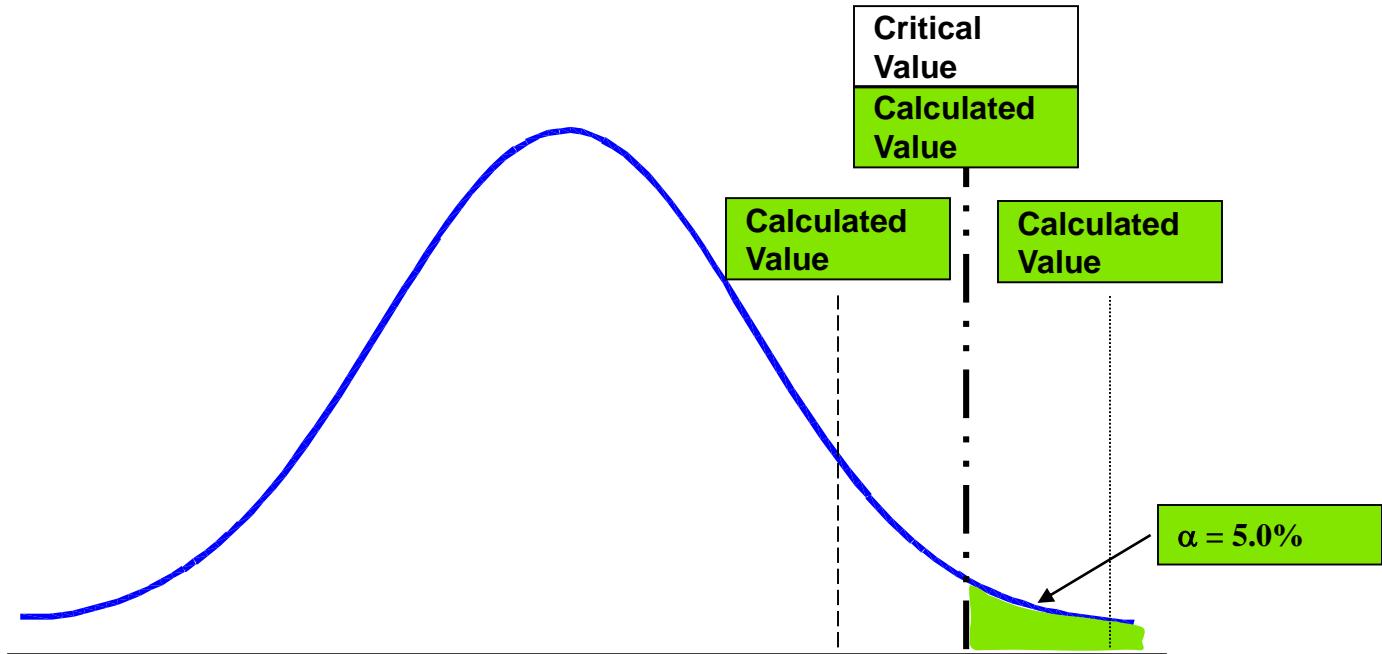
Cette valeur dépend de la forme de l'hypothèse alternative, en particulier savoir si le test est bilatéral, unilatéral à gauche, ou unilatéral à droite. Pour un test donné, la valeur critique peut-être vue comme la valeur limite a partir de laquelle on pourra rejeter avec un seuil de significativité donné.

p-value vs. critical value

Calc < Critical
P-value > alpha

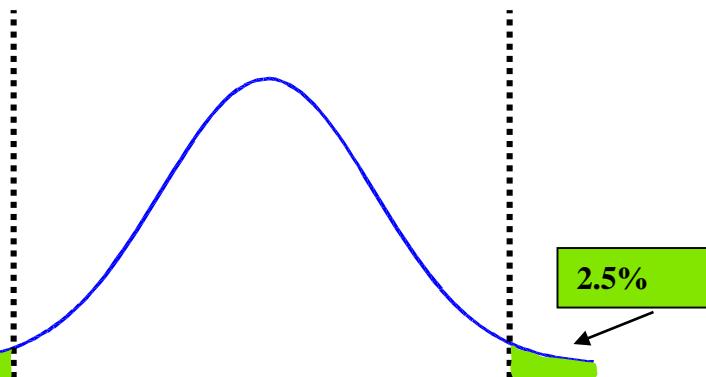
Calc = Critical
P-value = alpha

Calc > Critical
P-value < alpha

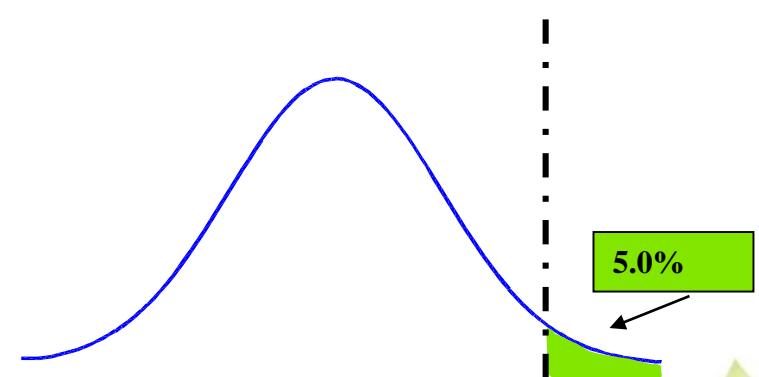


- Le test d'égalité ; dévise le risque en deux “cotés”
- Le test de supérieur/inférieur à une certaine valeur; met tous le risque en un seul “coté”

$$H_a : \mu_A \neq \mu_B$$



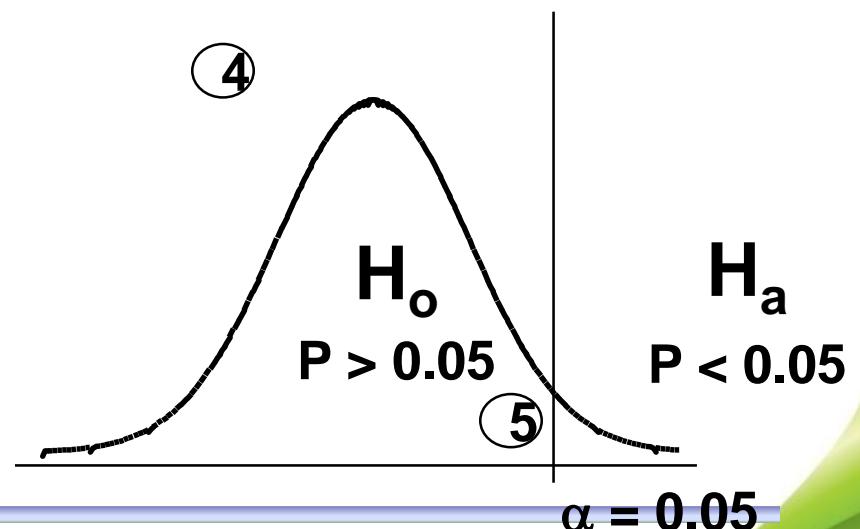
$$H_a : \mu_A > \mu_B$$



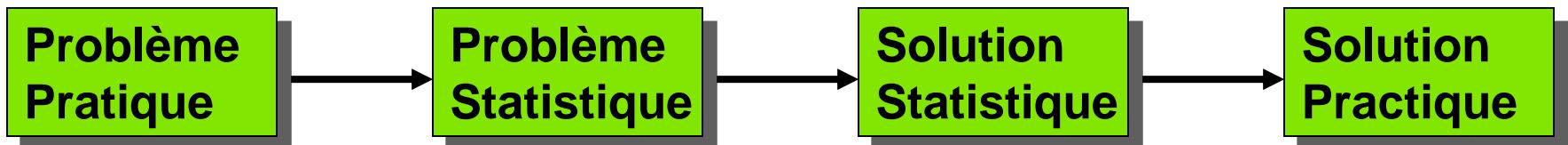
L'exemple qui suit concerne un test-t sur 2 échantillons, toutefois cette méthode est valable pour toutes les vérifications d'hypothèse:

1. Identifiez l'hypothèse nulle
2. Identifiez l'hypothèse alternative
3. Etablissez votre risque alpha
4. Dessinez un schéma (comme celui ci-dessous) pour représenter la vérification.
5. Collectez les données effectuez la vérification, déterminez la valeur-p et décidez.

- 1 $H_0: \mu_A = \mu_B$
- 2 $H_a: \mu_A > \mu_B$
- 3 $\alpha = 0.05$



1. Caractériser le problème et définir les objectifs
2. Elaborer les hypothèses
 - Enoncer l'hypothèse nulle (H_0)
 - Enoncer l'hypothèse alternative (H_a).
3. Décider du test statistique approprié (distribution de probabilité supposée, Z, t, χ^2 , F)
4. Indiquer le niveau Alpha (en général 5%)
5. Définir la taille de l'échantillon
6. Développer le plan d'échantillonnage
7. Effectuer la vérification et collecter les données
8. Calculer la statistique du test (z, t, ou F) à partir des données.
9. Déterminer la probabilité que se produise par hasard cette probabilité de test calculée = valeur-P .
 - Si p-value < α , rejeter H_0
 - Si p-value > α , accepter H_0
10. Reproduire les résultats et transposer la conclusion statistique en une solution pratique



Conclusion

- **Un hypothèse est accepté ou rejeté selon le niveau acceptable du risque (alpha).**
- **La collecte des données est destinée à avoir des échantillons représentatives sous des conditions spécifiques d'intérêts.**
- **Les statistiques et les distributions déterminent la valeur critique, la valeur calculée et la valeur P pour les analyses.**
- **On se basons sur les critères ci dessus on peut conclure :**
 - Les Tables de Référence – déterminent la valeur critique et le compare avec les statistiques calculés
 - Minitab ou Excel – détermine le risque (p-value) d'avoir se trompé si on rejette l'hypothèse nul.

For all tests:

$p > 0.05$ Fail to Reject H_0 (null)

$p < 0.05$ Reject H_0

Hypothesis Testing

Continuous Data (one factor only)

Non Normal

Normality Test

$H_0: s_1 = s_2 = s_3 = \dots$
 $H_a: \text{at least one is different}$
Minitab:
 Stat - Anova - Test for Equal Variances

For only two s's this is similar to an F-Test: $F = (S1)^2 / (S2)^2$
 If $F_{\text{calc}} > F_{\text{table}}$, then reject null.
 (Use Chi-Squared for one sample)

$H_0: M_1 = M \text{ target}$
 $H_a: M_1 \neq M \text{ target}$
Minitab:

Stat - Nonparametric - 1 Sample-Sign (OR)
 Stat - Nonparametric - 1 Sample-Wilcoxon
 (This is also used for paired comparisons:
 $H_0: M_1 - M_2 = 0$)

$M_1 = \text{Median of sample 1}$
 $M \text{ target} = \text{Target Median}$

1 Sample

2 or More Samples

$H_0: M_1 = M_2 = M_3 = \dots$
 $H_a: \text{at least one is different}$
Minitab:

Stat - Nonparametric - Mann-Whitney (OR)
 Stat - Nonparametric - Kruskal-Wallis (OR)
 Stat - Nonparametric - Mood's Median (OR)
 Stat - Nonparametric - Friedmans
 $M_1 = \text{Median of sample 1, etc...}$

Attribute Data (2 factors only)

$H_0: \text{Data is Normal}$
 $H_a: \text{Data is NOT Normal}$
Minitab:
 Stat - Basic Stat - Normality Test
 Use Anderson-Darling

Contingency Table

$H_0: \text{Two factors are independent}$
 $H_a: \text{Two factors are dependent}$
Minitab:
 Stat - Tables - Chi-square Test

Two or More Samples

Levene's Test

$H_0: \sigma_1 = \sigma \text{ target}$
 $H_a: \sigma_1 \neq \sigma \text{ target}$
Minitab:
 Stat - Basic Stats - Display Descriptive Statistics
 Graphs: Graphical Summary
 If σ target falls with σ CI, then fail to reject H_0 .

Normal

One Sample

Two or More Samples

Chi-Squared

$H_0: \sigma_1 = \sigma_2 = \sigma_3 = \dots$
 $H_a: \text{at least one is different}$
Minitab:
 Stat - Anova - Test for Equal Variance
 (For only two s's this is the same as an F-Test: $F = (S1)^2 / (S2)^2$
 If $F_{\text{calc}} > F_{\text{table}}$, then reject null.)

Bartlett's Test

Two Samples

Two or More Samples

One Way Anova

$H_0: \mu_1 = \mu \text{ target}$
 $H_a: \mu_1 \neq \mu \text{ target}$
Minitab:
 Stat - Basic Stats - 1 Sample-T
 Stat - Basic Stats - Display Descriptive Statistics
 Graphs: Graphical Summary
 If $X_{\bar{b}} \text{ target}$ falls with μ CI, then fail to reject H_0 .

2 Sample T Test (Variances Equal)

$H_0: \mu_1 = \mu_2$
 $H_a: \mu_1 \neq \mu_2$
Minitab:

Stat - Basic Stats - 2-Sample T
 (Compares Means using pooled Std Dev)
 Assume equal variances

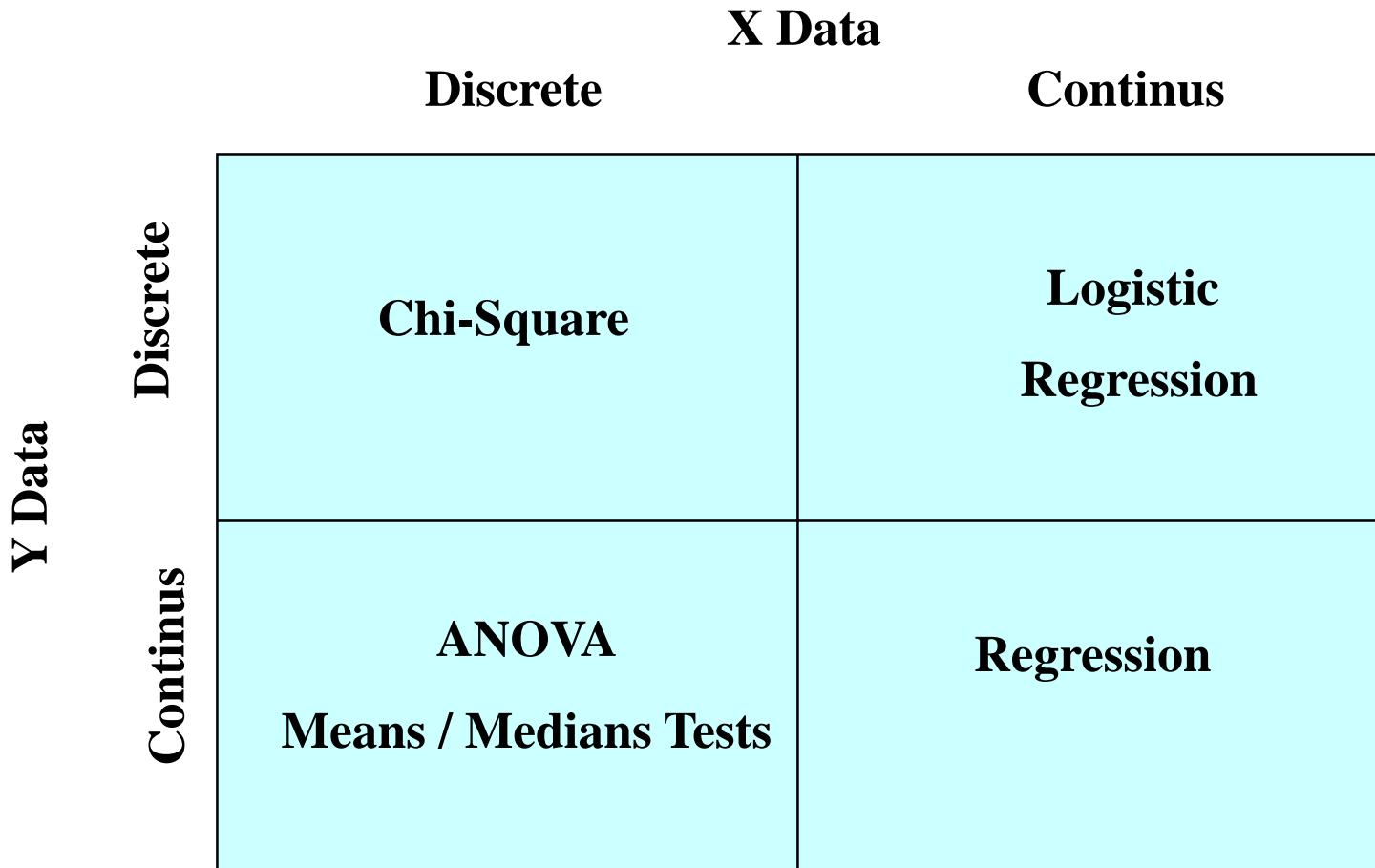
Fethi Derbaki 2021

$H_0: \mu_1 = \mu_2 = \mu_3 = \dots$
 $H_a: \text{at least one is different}$
Minitab:

Stat - Anova- One-way (or one-way unstacked)
 Assumes Equal Variances
 (Bartlett's test must fail to reject that variances are =.)



Analyze Roadmap





That's all Folks!

