Statistiques descriptives
Analyse de données
Algorithmes

(2) PROBABILITÉS ET STATISTIQUES

Jérôme Lacaille
Expert Émérite Safran

PROBABILITÉS ET STATISTIQUES

1. **Variables aléatoires**
2. **Densités, fonctions de répartitions**
3. **Statistiques élémentaires**
4. **Lois limites**
5. **Simulations**

→ **Tests**

VARIABLES ALÉATOIRES

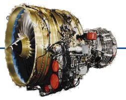
Variable aléatoire

- Une valeur observée x est le résultat d'une expérience.
- La fonction X produisant cette variable modélise l'expérience.
- On parle de variable aléatoire. Mathématiquement il s'agit de fonctions mesurables. Si l'on dispose d'un espace mesurable $(\Omega, \mathcal{A}, P)$
  - Ω est l'espace des observations
  - $\mathcal{A}$ est une tribu, essentiellement partie de l'ensemble des parties de Ω stable par réunion.
  - P une mesure de probabilité de $\mathcal{A} \rightarrow [0,1]$, (une mesure telle que $P(\Omega) = 1$).

X est mesurable si pour tout $B \in \Lambda, X^{-1}(B) \in \mathcal{A}$.
Avec $\forall B \in \mathcal{A}, P_X(B) = P\left(X^{-1}(B)\right) = P(X \in B)$.

Exemples

- Le résultat du lancer d'un dé.
- Le temps d'attente à une caisse de supermarché.
- La mesure renvoyée par un capteur.
- Etc.



RÉPARTITION

Il y a des variables

- Catégorielles, dont les valeurs ne sont pas nécessairement ordonnées, par exemple {A,B,C}.
- Discrètes, voire booléennes, qui prennent leurs valeurs dans un ensemble fini ou dénombrable de possibles (événements individuels) : $\{0,1\}, \{1,2, \dots n\}, \mathbb{N}$
- Continues : dont les mesures sont réelles ou complexes.

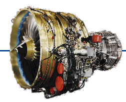
- Un vecteur de variables aléatoires et appelé généralement un vecteur aléatoire.
- La répartition des valeurs x produites par la variable X est appelée distribution.

Fonction de répartition

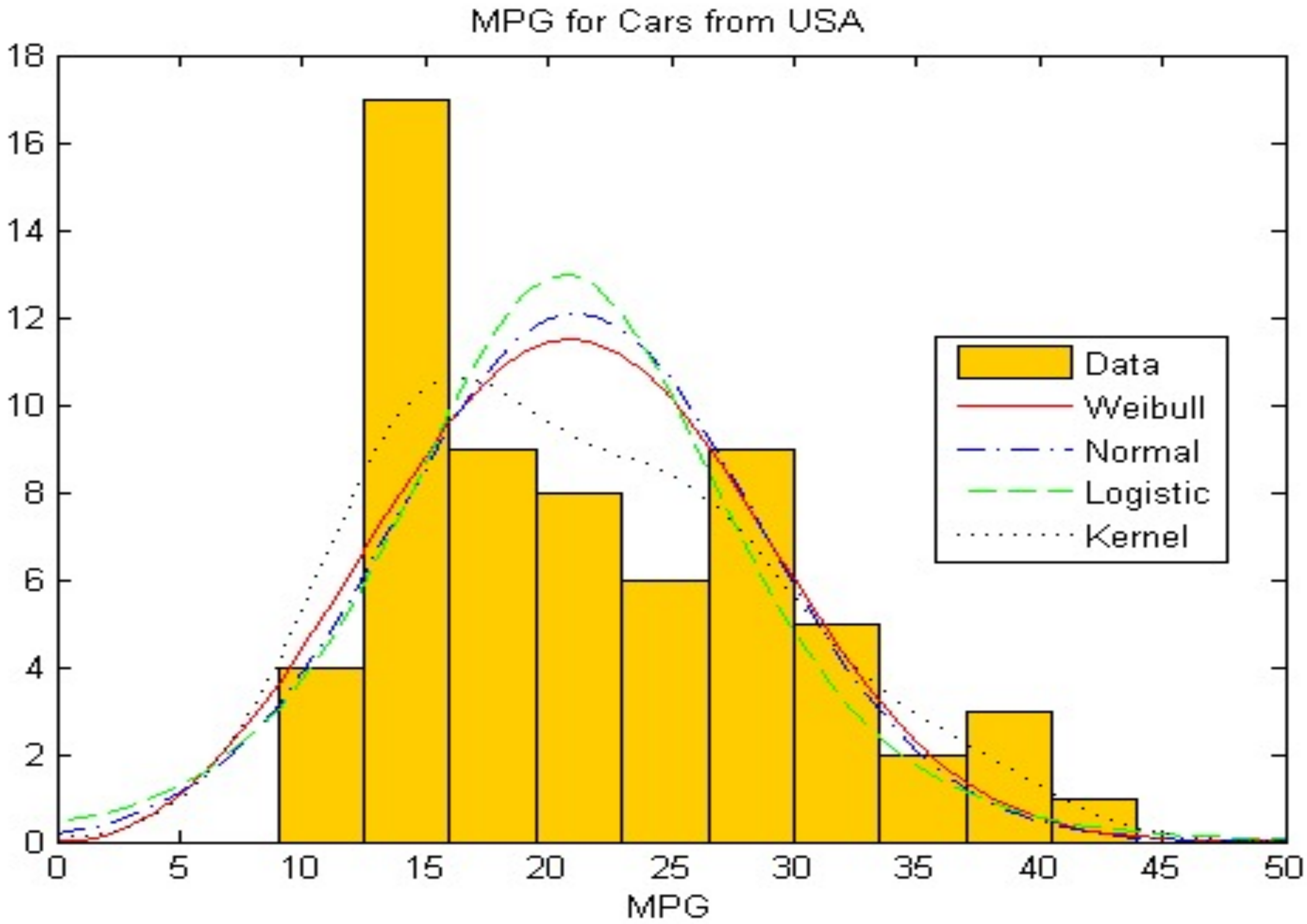
$$F_X(x) = P(X < x)$$

On peut afficher un histogramme de ses valeurs, ou sa densité f_X si la variable est absolument continue par rapport à la mesure de Lebesgue.

$$f_X(x) = \frac{dF_X(x)}{dx}$$



DENSITÉS DE PROBABILITÉS



EXEMPLES DE DISTRIBUTIONS DE PROBABILITÉS

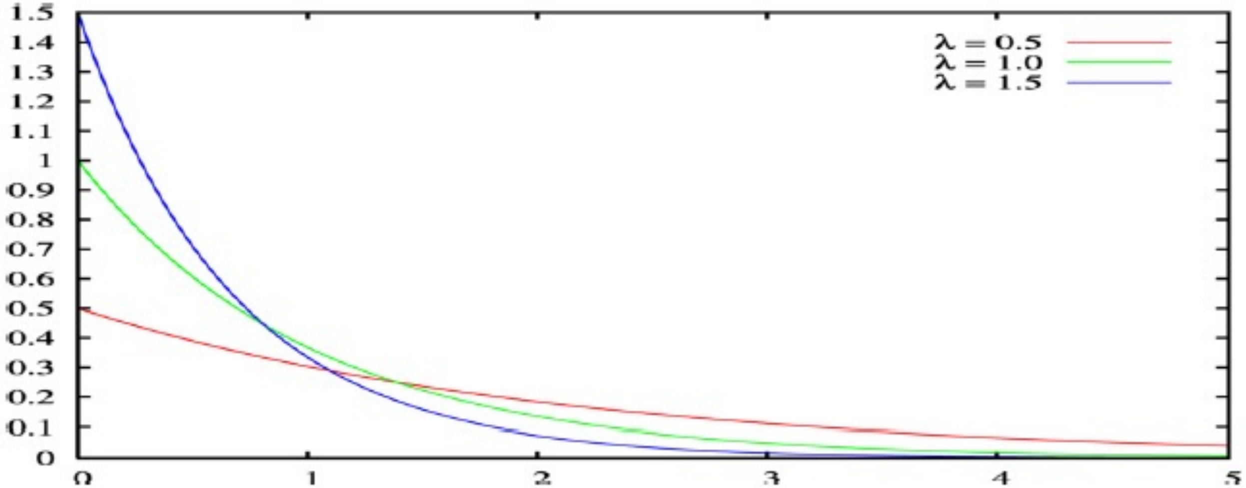
Discrètes

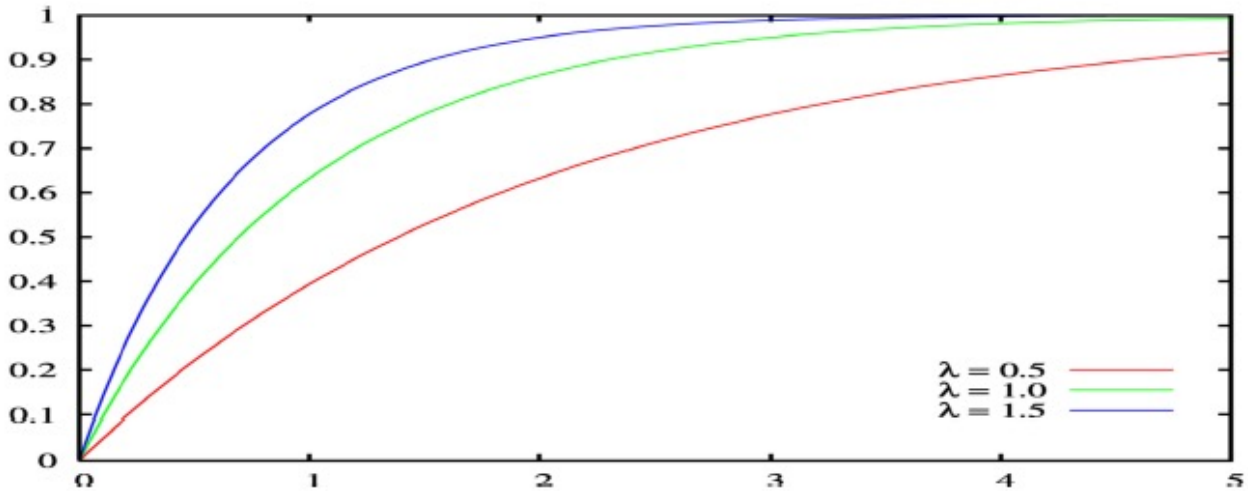
Loi de Bernoulli $B(p)$	$p \in [0,1]$	$X \in \{0,1\}$	$P(X=1)=p$
Loi binomiale $B(n,p)$	$p \in [0,1]$ $n \in \mathbb{N}$	$X \in \{0..n\}$	$P(X=k)=C_n^k p^k (1-p)^{n-k}$
Loi de Poisson $P(\lambda)$	$\lambda \in \mathbb{R}^{+*}$	$X \in \mathbb{N}$	$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$

Continues

Uniforme $U[a,b]$	$[a,b] \subset \mathbb{R}$	$X \in [a,b]$	$f_X(x)=\frac{1}{b-a}$
Normale $N(\mu,\sigma^2)$ gaussienne	$\mu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}^+$	$X \in \mathbb{R}$	$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Exponentielle $Exp(\lambda)$	$\lambda \in \mathbb{R}^+$	$X \in \mathbb{R}^+$	$f_X(x)=\lambda e^{-\lambda x}$

DENSITÉS ET FONCTIONS DE RÉPARTITIONS DE LA LOI EXPONENTIELLE POUR DIFFÉRENTES VALEURS DES PARAMÈTRES.



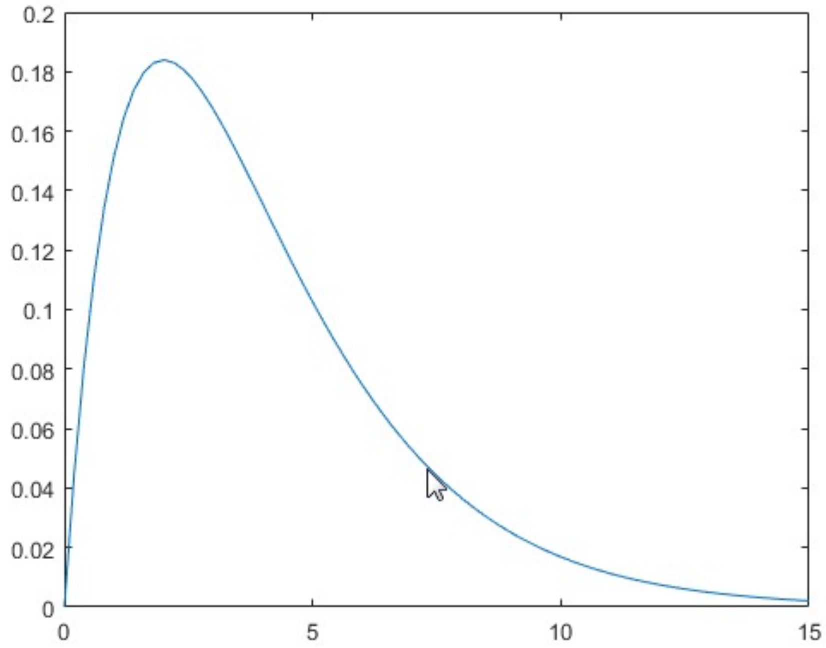


Loi du $\chi^2(n)$

Si $X \sim N(0,1)$	alors $X^2 \sim \chi^2(1)$

$Y = \sum_{i=1}^{n} X_i^2$	avec chaque X_i indépendant de loi N(0,1)
	alors Y $\sim \chi^2(n)$

Densité de la loi du $\chi^2(4)$.



Loi de Student $T(k)$

Si $Z \sim N(0,1)$ et $S \sim \chi^2(k)$	alors $T = \frac{Z}{S\sqrt{k}}$ suit une loi de Student T(k).

THÉORÈME DE COCHRAN

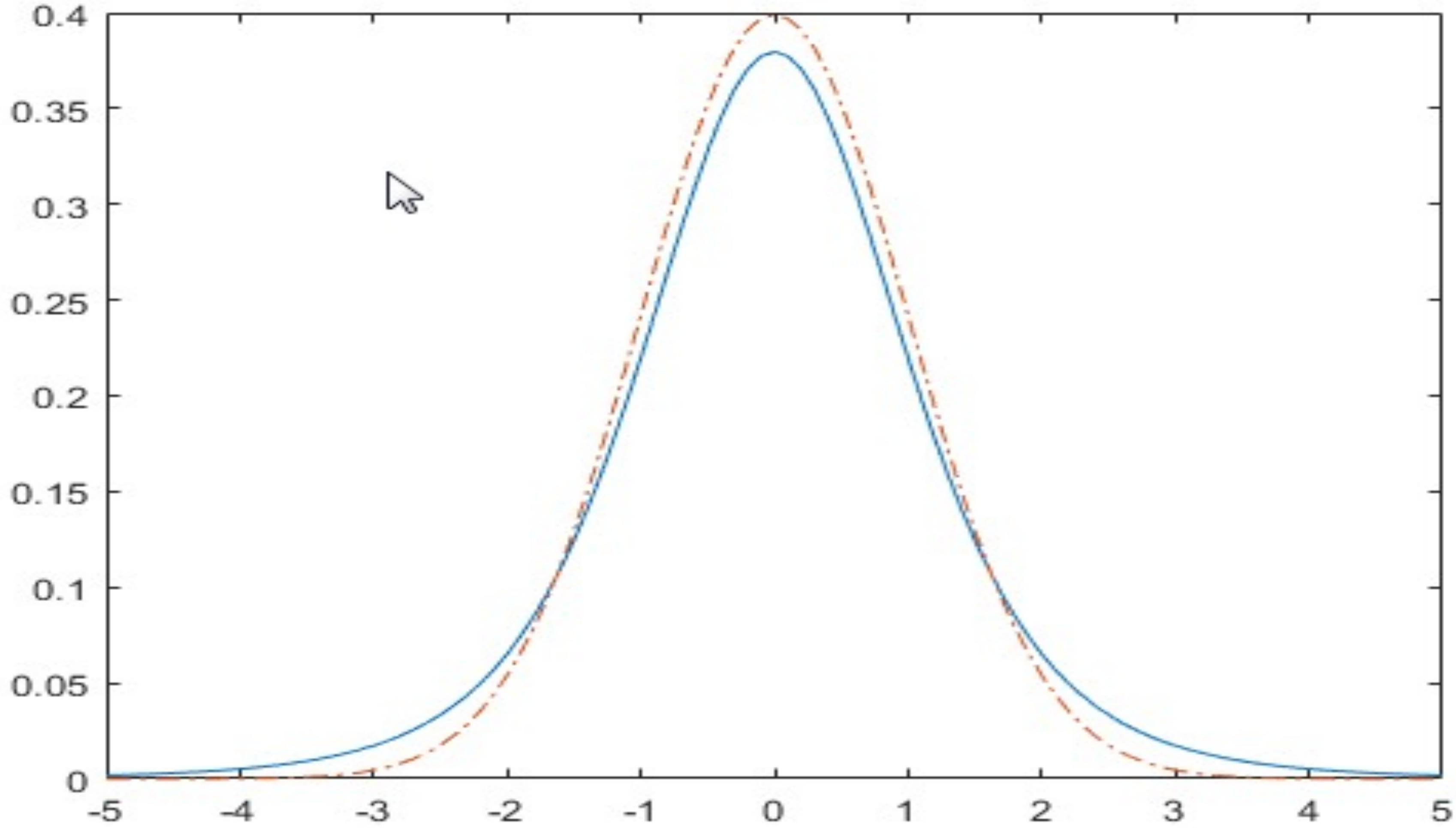
Soient $x_1, \ldots\ x_n$ un échantillon de n observations gaussiennes $N(\mu, \sigma^2)$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	est un estimateur de la moyenne μ et suit une loi $\mathrm{N}(\mu, \frac{\sigma^2}{n})$.
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	est un estimateur sans biais de la variance σ^2 et suit une loi $\chi^2(n-1)$.

$\bar{x}$ et s^2 sont indépendants et

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim T(n-1)$$

LOIS GAUSSIENNE ET DE STUDENT



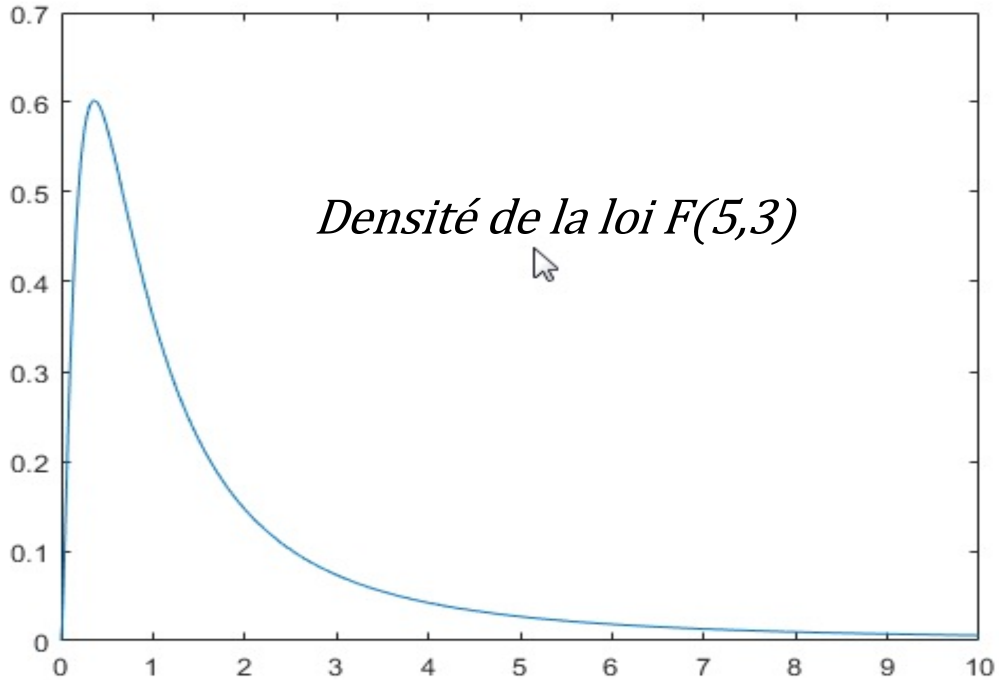
En bleu, la loi T(5) comparée à la loi normale centrée réduite N(0,1) en rouge.
Les queues de distribution sont plus épaisses.

LOI DE FISHER-SNEDECOR

Si X_1 et X_2 sont deux variables aléatoires indépendantes de lois respectives $\chi^2(k_1)$ et $\chi^2(k_2)$ alors

$$\frac{X_1/k_1}{X_2/k_2} \sim F(k_1, k_2)$$

Cette loi sert à comparer des variances.



MOYENNE, VARIANCE, QUANTILES

Si une variable est intégrable, sa moyenne est notée

$$m = E(X) = \int x f_X(x) dx$$

Sa variance est l'écart au carré autour de la moyenne

$$\sigma^2 = E(X - m)^2 = EX^2 - m^2 = \int x^{\mathbf{2}} f_X(x) dx - \boldsymbol{m^2}$$

L'écart-type σ est la racine-carrée de la variance.

Pour toute valeur $\alpha \in [0,1]$ le quantile q_α est la valeur obtenue par l'inverse de la fonction de répartition

$$\alpha = P(X < q_\alpha)$$

ECHANTILLON

Un échantillon est un ensemble d'observations issues d'une loi donnée (généralement inconnue).

- Le tableau suivant montre un échantillon d'un vecteur aléatoire issu de mesures faites en vol par un avion. Chaque ligne est un vol dont on a la date et l'heure et chaque colonne est la mesure d'une caractéristique.

ACARS (10 x 6)

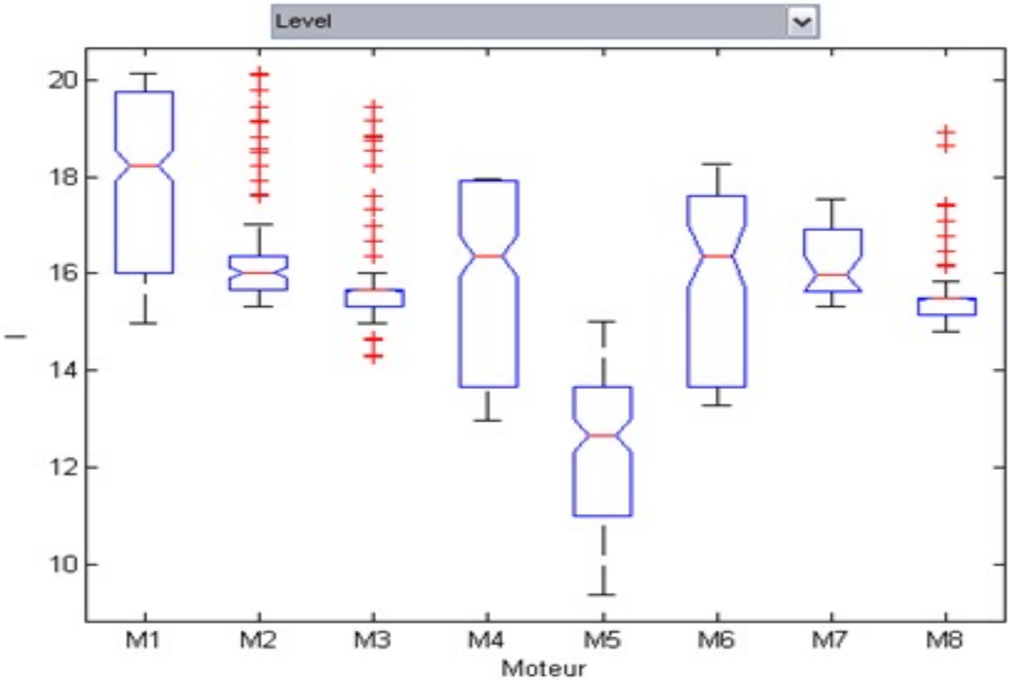
ID Date	EGT deg_C	FF lb/h	N1 %_rpm	N2 %_rpm	PS3 psi	T3 deg_C
01-Feb-2008 01:34:32	631	2098	87	90.8	94.8	403
01-Feb-2008 05:44:57	681	2436	90	93.6	106.5	443
01-Feb-2008 15:33:59	693	2396	90.5	93.7	103.3	446
01-Feb-2008 18:20:44	648	2211	86.7	90.6	99.3	409
01-Feb-2008 19:34:55	671	2302	88.6	92.8	102	432
02-Feb-2008 06:25:30	607	2382	86.1	89.7	110.5	389
02-Feb-2008 10:22:20	662	2143	88.7	92.3	94.8	425
02-Feb-2008 12:06:17	667	2349	88.4	93.1	104.5	438
02-Feb-2008 16:38:21	662	2165	88.6	92.2	95.8	424
02-Feb-2008 18:21:50	575	2376	82	88.8	115.5	380

- Les échantillons d'observations sont généralement placés dans des tables.

BOITES À MOUSTACHES (BOXPLOTS)

On peut afficher les caractéristiques statistiques de chaque variable d'un échantillon. Par exemple un boxplot (ou boite à moustaches) est un affichage représentant la médiane ($q_{0.5}$) et les deux quantiles à 25% et 75% ainsi que

- Des moustaches généralement placées à $q_{.75} + 1.5(q_{.75} - q_{.25})$ et symétriquement ;
- Des outliers (au-delà des moustaches) sont affichés (ici en rouge) ;
- Des angles donnent l'intervalle de confiance autour de la médiane à $q_{.5} \pm 1.57(q_{.75} - q_{.25})/\sqrt{n}$.



LOIS LIMITES, LOI FORTE

Si on prend deux échantillons issus de la même loi et qu'on calcule un estimateur, la moyenne par exemple, on ne trouve pas systématiquement le même résultat. En fait l'estimateur de la moyenne est lui-même une variable aléatoire.

- On dit que cet estimateur est sans biais si sa moyenne est égale à la valeur théorique estimée.

Loi forte des grands nombres

Soit (X_n) une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de même loi qu'une variable X.

On suppose $E(X) = m < +\infty$ alors

$$\bar{X}_N = \frac{1}{N}(X_1 + X_2 + \cdots + X_N) \overset{ps}{\rightarrow} m$$

quand $N \rightarrow +\infty$.

THÉORÈME CENTRAL LIMITE (TCL)

Si $E(X) < +\infty$ **et** $E(X^2) = \sigma^2 < +\infty$ **alors**

$$\sqrt{N}\ \frac{\bar{X}_N - m}{\sigma} \xrightarrow{L} N(0,1)$$

quand $N \to +\infty$**.**

Deux conséquences

- La loi normale apparait naturellement quand on observe des moyennes d'échantillons de variables.
- La convergence vers l'estimateur de la moyenne a une vitesse en $1/\sqrt{N}$.

SIMULATION, RELAXATION, OPTIMISATION

Beaucoup d'algorithmes utilisent désormais des tirages aléatoires.

- L'optimisation stochastique utilise souvent un la convergence d'un processus aléatoire (chaine de Markov) vers une mesure limite.
- La relaxation stochastique (**recuit simulé**) est un procédé de descente de gradient optimal qui permet sous de bonnes conditions d'atteindre un minimum global d'énergie.
- Les filtres particulaires utilisent des pluralités de trajectoires pour estimer la loi d'un processus.
- Pour le calcul de la **robustesse** d'une méthode par bootstrap, K-folds, out-of-the-bag, leave-one-out, ou d'autres méthodes.
- La **cryptographie** utilise des séquences de nombres aléatoires pour générer des clés.

Générateur de nombres pseudo-aléatoires

- Les outils informatiques modernes disposent *désormais* toujours de générateurs de nombres pseudo-aléatoires **uniformes**.

Générateurs de nombres aléatoires

- Il existe des sociétés spécialisées dans la génération de nombres aléatoires, notamment en exploitant les propriétés de l'optique quantique.

SAFRAN Snecma

GÉNÉRATION DE NOMBRES ALÉATOIRES

Quantis Random Number Generator

TRUE RANDOM NUMBER GENERATOR EXPLOITING THE RANDOMNESS OF QUANTUM PHYSICS

- En 1990, en collaboration avec l'IEF (Institut d'électronique fondamentale, *Patrick Garda*) nous avons fait appel à un laboratoire d'optique pour générer un vrai mouvement brownien (processus aléatoire gaussien) en envoyant un faisceau laser à travers une substance contenant des particules en mouvement aléatoire. Le faisceau était ensuite diffusé par une lentille pour être projeté sur des capteurs photoélectriques connectés à un réseau de neurones de Hinton.

SIMULATION D'UNE PROBABILITÉ UNIFORME

Supposons de l'on dispose d'un générateur de nombres pseudo-aléatoires sous la forme d'une fonction rand() qui renvoie un tirage uniforme sur [0,1].

Pour générer une expérience X de loi binomiale $B(p)$ avec $0 < p < 1$, c'est-à-dire

- $P(X = 1) = p$
- $P(X = 0) = 1 - p$

On execute

```
r = rand() ;
if r <= p
        then x=1
        else x=0
end
```

Ou plus simplement `x = (rnd() <= p)` sous Matlab par exemple.

SIMULATION D'UNE NOMBRE GAUSSIEN

Pour tirer un nombre aléatoire gaussien, si l'on ne dispose pas d'une fonction (comme `randn()` sous Matlab par exemple) on utilise le théorème de Box-Muller :

- Soient U_1 et U_2 deux variables aléatoires indépendantes et identiquement distribuées de lois uniformes $U(]0,1[)$ alors les variables aléatoires X_1
- et X_2 définies par :

$$X_1 = \sqrt{-2\ \ln \mathrm{U}_1}\cos(2\pi U_2) \quad \textbf{et} \quad X_2 = \sqrt{-2\ \ln \mathrm{U}_1}\sin(2\pi U_2)$$

Sont indépendantes, gaussiennes N(0,1).

Une solution, moins couteuse en temps de calcul, peut aussi d'utiliser le TCL en tirant une somme de 12 valeurs uniformes sur $\left[-\frac{1}{2},\frac{1}{2}\right]$.

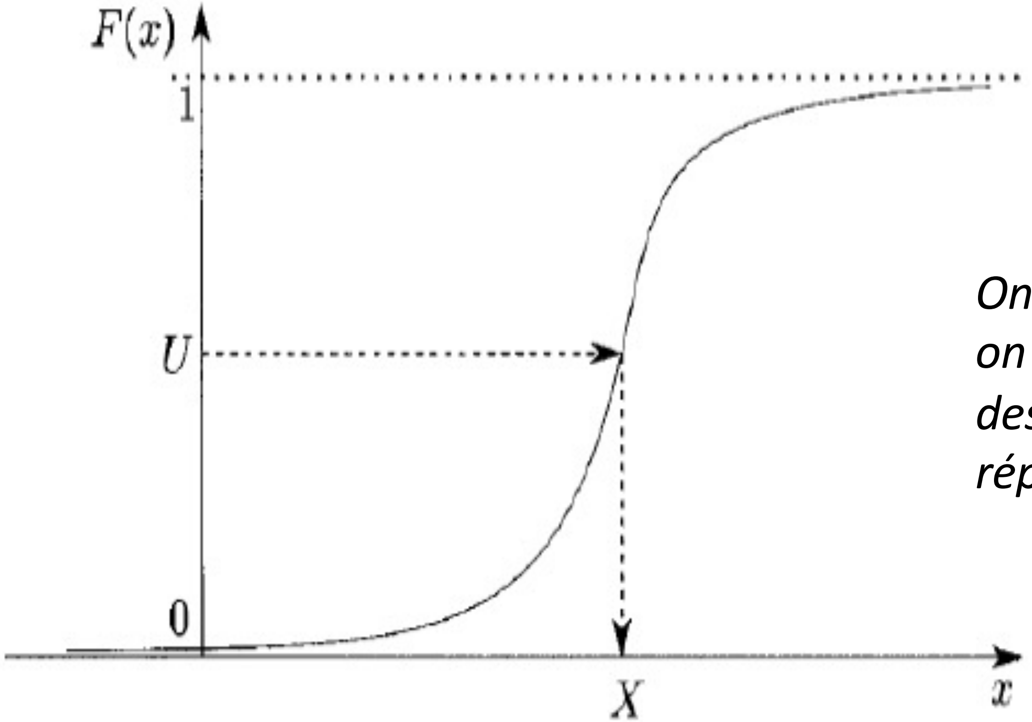
- En effet si $X_i \sim U([-\frac{1}{2},\frac{1}{2}])$ alors $E(X) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x\, dx = 0$ et $E(X^2) = 2\int_0^{\frac{1}{2}} x^2 dx = \frac{1}{12}$
- Donc $S_{12} = X_1 + X_2 + \cdots + X_{12} \to N(0,1)$

TRANSFORMATION INVERSE

Si on dispose de la fonction de répartition F_X de la loi à échantillonner alors en posant

$$F_X^{-1}(u) = \inf\{x : F_X(x) > u\}$$

Si $U \sim U([0,1])$ **alors** $F_X^{-1}(U) \sim X$



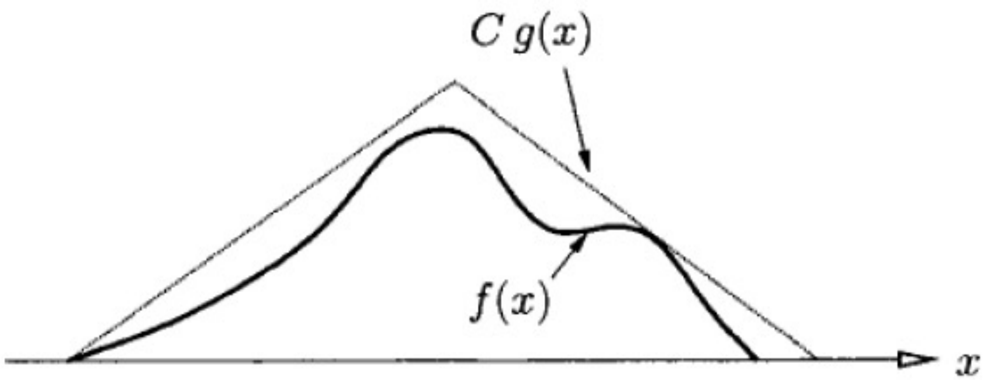
On tire U uniformément sur [0,1] et on cherche l'image X de U sur l'axe des abscisses par la fonction de répartition.

MÉTHODE D'ACCEPTATION ET DE REJET

C'est une méthode souvent utilisée en "*importance sampling*" quand on dispose d'une approximation ou une majoration de la distribution attendue (que l'on cherche en général à estimer).

On suppose que la densité de probabilité $g(x)$ majore la densité cherchée $f(x)$:

$$\exists C > 0, \text{ telle que } f(x) \leq Cg(x)$$



Alors :

1. On génère X de densité $g(x)$.
2. On génère $U \sim U([0,1])$ indépendante de X.
3. Si $U \leq f(X)/Cg(X)$ alors on prend $Z = X$, sinon on recommence (1).

SAFRAN Snecma

TESTS

1. **Vraisemblance, statistiques, estimateurs**
2. **Risque quadratique**
3. **Information de Fisher, borne de Cramer-Rao**
4. **Statistique exhaustive**
5. **Intervalle de confiance**
6. **Tests**
7. **Erreurs**

→ **Applications**

STATISTIQUE

Soient une série d'observations X_i, formant un échantillon de données. On appelle statistique toute fonction mesurable des observations

$$Z = f(X_1, \ldots, X_n)$$

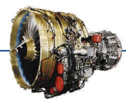
- La moyenne, variance, les moments sont des statistiques usuelles.

"Faire des statistiques" c'est donc résumer l'information contenue dans l'échantillon.

On suppose que notre échantillon suit une loi de probabilité dépendant d'un paramètre θ.

$$X \sim P_\theta$$

$$f_X(x) = f_\theta(x)$$



VRAISEMBLANCE ET ESTIMATEUR

Pour une valeur θ donnée, la vraisemblance de notre échantillon sous P_θ est notée

$$L_\theta(X_1, \dots, X_n) = f_\theta(X_1) f_\theta(X_2) \dots f_\theta(X_n)$$

- C'est bien la probabilité d'observer l'échantillon, ou sa densité dans le cas continu si les observations sont indépendantes.
- Ce n'est pas une statistique car elle dépend du paramètre θ, la vraisemblance observée est obtenue en remplaçant les variables aléatoires X_i par leurs observations x_i.
- Plus la vraisemblance est grande, plus le paramètre θ est proche de la loi réellement observée.

Un estimateur du paramètre θ est une statistique $T = f(X_1, \dots, X_n)$

- Un estimateur va estimer le paramètre (mais n'est pas nécessairement proche de celui-ci).

Estimateur du maximum de vraisemblance

$$\hat{\theta} = \text{Arg Max } L_\theta(X_1, \dots, X_N)$$

est appelé estimateur du maximum de vraisemblance.

RISQUE QUADRATIQUE

On mesure la qualité d'un estimateur par le risque $\boldsymbol{\lambda}(T,\theta) = E_\theta(T-\theta)^2$

Notons que si l'on fait intervenir la moyenne de l'estimateur

$$\begin{aligned}
\boldsymbol{\lambda}(T,\theta) &= E_\theta(T - E_\theta T + E_\theta T - \theta)^2 \\
&= E_\theta[(T - E_\theta T)^2 + 2(T - E_\theta T)(E_\theta T - \theta) + (E_\theta T - \theta)^2] \\
&= E_\theta(T - E_\theta T)^2 + (E_\theta T - \theta)^2
\end{aligned}$$

- Le premier terme $E_\theta(T - E_\theta T)^2$ est la variance de l'estimateur.
- Le second terme $(E_\theta T - \theta)^2$ est le carré de son biais.

- L'estimateur est dit "sans biais" si son biais est nul !
- Il est asymptotiquement sans biais si le biais converge vers 0 quand la taille de l'échantillon augmente.
- Il est convergeant s'il est asymptotiquement sans biais (ou sans biais) et que sa variance converge vers 0.
- Il est efficace si il est sans biais et que sa variance égale la borne inférieure de Cramer-Rao.

Pour deux estimateurs T et T' sans biais, T est dit meilleur que T' ssi $var(T) \leq var(T')$.

EXEMPLES DE STATISTIQUES

La moyenne empirique

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

est un estimateur sans biais et convergeant de la moyenne.

La variance empirique

$$S'^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \bar{X})^2$$

est un estimateur biaisé ($E(S'^2) = \frac{N-1}{N}\sigma^2$), asymptotiquement sans biais et convergeant de la variance.

C'est l'estimateur (d'auto-corrélation) qu'il faut utiliser cependant pour construire la matrice de Hankel à utiliser pour l'estimation du rang d'un processus autorégressif.

ESTIMATEUR SANS BIAIS DE LA VARIANCE

Si par contre la moyenne μ est connue, la variance empirique dans laquelle la moyenne empirique est remplacée par sa valeur réelle est sans biais.

$$S'^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

- Si la moyenne est estimée par $\bar{X}$ l'estimateur

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

  est sans biais et convergeant.

La fréquence d'apparition d'un événement est la moyenne empirique d'une loi de Bernoulli $B(p)$

$$F \sim \frac{1}{N} B(\mathbf{N}, p)$$

est un estimateur sans biais ($E(F) = p$) et convergeant ($\mathrm{var}(F) = \frac{p(1-p)}{N}$) de la fréquence théorique p.

INFORMATION DE FISHER
BORNE DE CRAMER-RAO

L'information de Fisher est définie par

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log L_{\theta}(X)\right)^2\right]$$

Le théorème de Cramer-Rao montre que pour tout estimateur sans biais T de θ, sa variance est minorée par l'inverse de l'information de Fisher.

$$var(T) \geq I(\theta)^{-1}$$

Notons que l'information de Fisher est additive pour les variables indépendantes, donc l'information de Fisher de n événements indépendants n'est autre que n fois l'information de Fisher d'un unique événement.

$$I_n(\theta) = nI_1(\theta)$$

SCORE ET INFORMATION DE FISHER

Le score de l'échantillon est défini par

$$s_n(\theta) = \frac{\partial}{\partial\theta}\log L_\theta(X_1, \dots X_n)$$

- Pour une série d'observations données, le score est une fonction du paramètre θ qui s'annule au maximum de vraisemblance $\hat{\theta}$.
- L'information de Fisher est la variance du score.

$$I_n(\theta) = E(s_n(\theta)^2)$$

- L'information de Fisher mesure l'information apportée par un échantillon sur le paramètre. Si $I_n(\theta)$ est faible, c'est que l'échantillon n'est pas très informatif.
- Le score mesure la sensibilité de la vraisemblance en $\boldsymbol{\theta}$. C'est la dérivée de la vraisemblance donc un score faible montre que la vraisemblance est peu sensible aux variations du paramètre.
- L'information de Fisher peut aussi s'expliquer comme une courbure de la géométrie autour de θ :

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log L_\theta(X)\right] = -E\left[\frac{\partial}{\partial\theta}s_n(\theta)\right]$$

Intégration par partie de $\int s(\theta)s(\theta)dP$ (page suivante)

  - *En dimension plus grande que 1, on crée une matrice de Fisher qui représente bien la courbure de géométrie de l'information autour de la valeur du paramètre.*

DÉMONSTRATION

L'information de Fisher est la covariance du score

  - $I(\theta) = E[ss']$ où $s(\theta) = \frac{\partial}{\partial\theta}\log L_\theta(X_1, \ldots X_n)$=$\frac{\partial}{\partial\theta}l(\theta)$
- Si $t(X,\theta)$ est une variable aléatoire de la même dimension que $\boldsymbol{\theta}$ est sous des conditions de régularité suffisantes, alors
  - $E[st'] = \frac{\partial}{\partial\theta}E(t') - E(\frac{\partial t\prime}{\partial\theta})$
- En effet
  - $E(t') = \int t'LdX$ donc en dérivant sous le signe somme (régularité)
  - $\frac{\partial}{\partial\theta}E(t') = \int t'\frac{\partial L}{\partial\theta}dX + \int\frac{\partial t\prime}{\partial\theta}LdX$ or $\frac{\partial L}{\partial\theta} = L\frac{\partial \log L}{\partial\theta}$
  - $\frac{\partial}{\partial\theta}E(t') = \int\frac{\partial \log L}{\partial\theta}t'LdX + \int\frac{\partial t\prime}{\partial\theta}LdX$
  - $\frac{\partial}{\partial\theta}E(t') = \mathrm{E}(\mathrm{st}') + \mathrm{E}\left(\frac{\partial t'}{\partial\theta}\right)$

Corolaires

- En prenant $t = cte$
  - $E(s) = 0$
- En prenant $t = s$
  - $I(\theta) = var(s) = E(ss') = -E\left(\frac{\partial s'}{\partial\theta}\right) = -E(\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'})$

EXEMPLES D'INFORMATIONS DE FISHER

Lois discrètes

Bernoulli	$B(p)$	$I = \frac{1}{p(1-p)}$
Binomiale	$B(n,p)$	$I = \frac{n}{p(1-p)}$
Poisson	$P(\lambda)$	$I = \frac{1}{\lambda}$

Lois continues

Gaussienne	$N(\mu,\sigma^2)$	$I = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$
Exponentielle	$Exp(\lambda)$	$I = \frac{1}{\lambda^2}$

STATISTIQUE EXHAUSTIVE

Pour une statistique $T(X_1, \dots, X_n)$, il est possible de calculer son information de Fisher. Elle est inférieure à celle de la loi estimée. En cas d'égalité on dit que la statistique T est exhaustive.

$$I_T(\theta) \leq I_n(\theta)$$

- Si la vraisemblance peut se décomposer en deux termes de la forme

$$L_\theta(X_1, \dots, X_n) = g(X_1, \dots, X_n)\mathrm{h}(\theta, T(X_1, \dots, X_n))$$

- Alors la statistique T est exhaustive.

$$I_n(\theta) = -E\left[\cancel{\frac{\partial^2}{\partial\theta^2}\log g(X)}\right] - E\left[\frac{\partial^2}{\partial\theta^2}\log \mathrm{h}(\theta, T)\right] = I_T(\theta)$$

Ne dépend pas de θ

Par exemple $\bar{X}_n$ est un estimateur exhaustif de la moyenne μ d'une loi normale.

$$L_n(\mu, \sigma^2) = \frac{1}{(2\pi\sigma)^{\frac{n}{2}}} e^{-\frac{(\bar{X}_n - \mu)^2}{2\sigma^2}}$$

- L'estimateur du maximum de vraisemblance $\hat{\theta}$, s'il existe est :
- Sans biais, convergeant, asymptotiquement gaussien (TCL) et converge en loi vers $N\left(\theta, I_n(\theta)^{-1}\right)$.

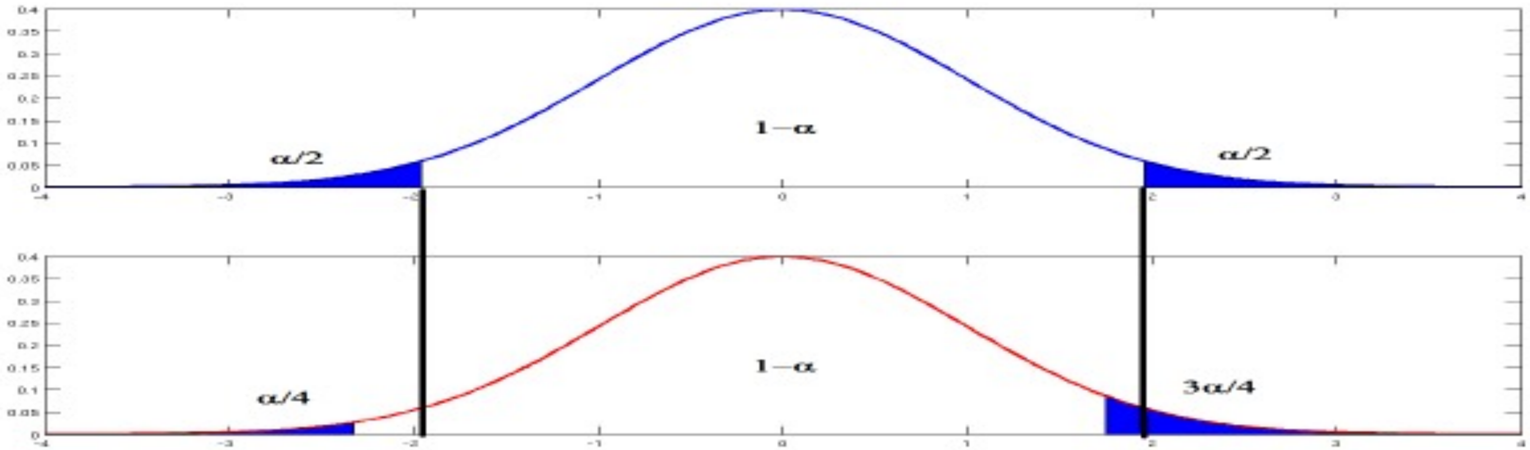
INTERVALLE DE CONFIANCE

Un intervalle de confiance est une méthode d'estimation de paramètre $\theta \in \mathbb{R}$ par encadrement.

Soit $\alpha \in [0,1]$, un intervalle de confiance de niveau α est donné par un couple (A_α, B_α) tel que

$$P_\theta(\theta \in [A_\alpha, B_\alpha]) \geq 1 - \alpha$$

où A_α et B_α sont deux statistiques telles que $A_\alpha \leq B_\alpha$.



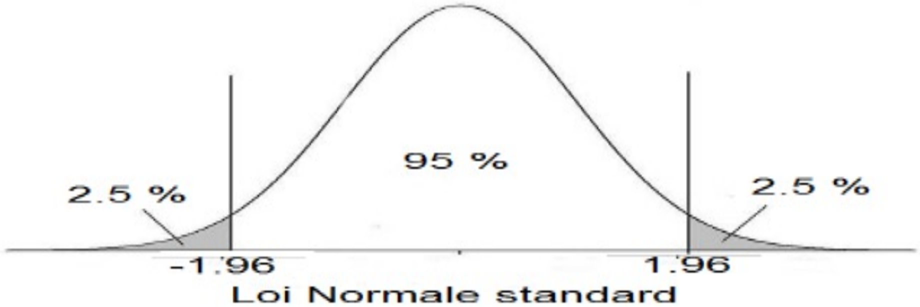
Un intervalle de confiance n'est pas nécessairement unique. Ci-dessus deux intervalles de confiance à 95% pour une gaussienne.

SAFRAN Snecma

IC MOYENNE GAUSSIENNE VARIANCE CONNUE

$X \sim N(\mu, \sigma^2)$ **donc**

$$\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$



soit $u_{1-\frac{\alpha}{2}}$ le quantile de la loi normale réduite (qui est symétrique), alors

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Un intervalle de confiance de niveau α calculé à partir d'un échantillon est donc

$$IC_n(1-\alpha) = \left[\bar{x}_n - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

- *Notons que* $u_{97.5\%} = 1.96$.
- *Le quantile à* 3σ *correspond à peu près à* 10^{-3}.
- *Le quantile à* 6σ *correspond à* 10^{-9}.

  Ce qui explique l'usage répandu qu'il en est fait.

IC MOYENNE GAUSSIENNE VARIANCE INCONNUE

$X \sim N(\mu, \sigma^2)$ **donc**

$$\frac{(\bar{X}_n - \mu)}{S/\sqrt{n}} \sim T(n-1)$$

avec

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

$$P\left(-t_{n-1,1-\frac{\alpha}{2}} < \frac{\bar{X}_n - \mu}{S/\sqrt{n}} < t_{n-1,1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Un intervalle de confiance de niveau α calculé à partir d'un échantillon est donc

$$IC_n(1-\alpha) = \left[\bar{x}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + t_{n-1,1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

La loi de Student a une variance plus grande que la loi normale centrée réduite mais tend vers la loi normale quand n tend vers l'infini.

TESTS D'HYPOTHÈSES

La formulation d'un test fait intervenir deux hypothèses

- Une hypothèse de base H_0 (hypothèse nulle) ;
- Une hypothèse alternative H_1.

Souvent à chaque hypothèse est associée un paramètre ou un ensemble (souvent un intervalle) de paramètres.

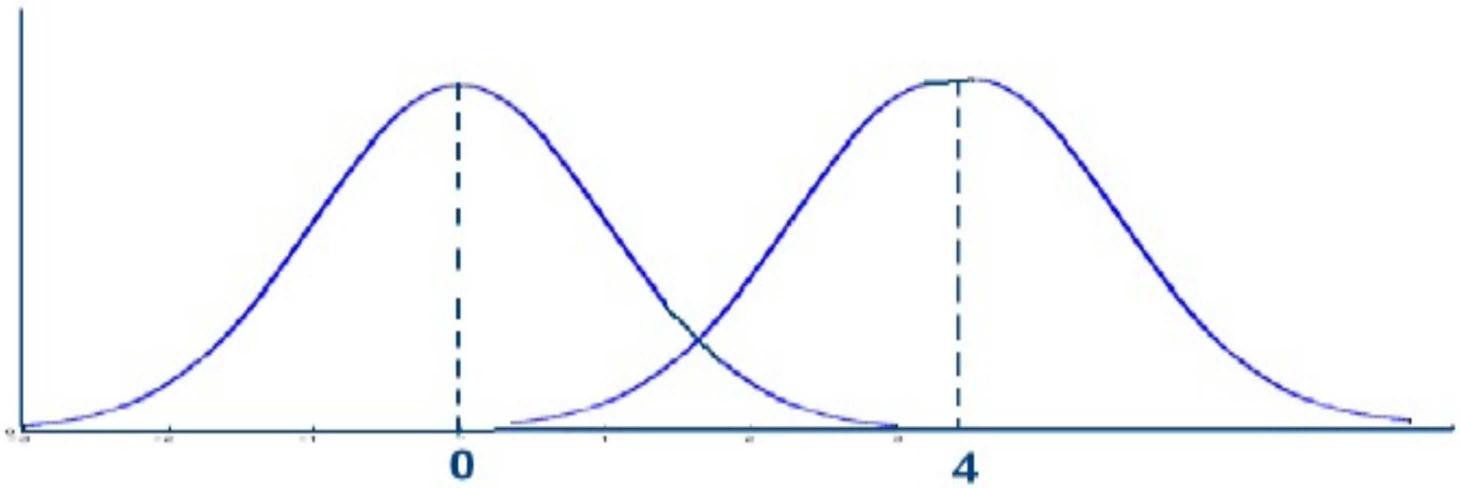
$$H_0 \Leftrightarrow \theta \in \Theta_0$$
$$H_1 \Leftrightarrow \theta \in \Theta_1$$

- Le test de H_0 contre H_1 est une règle de décision qui est construite à partir d'une région de rejet $R \subset \Omega$. Si l'échantillon $(X_1, \dots, X_n) \in R$ alors on rejette H_0 pour H_1.

- On construit la région de rejet à l'aide d'une statistique exprimant clairement la différence entre les deux sous-ensembles de paramètres.

SAFRAN
Snecma

EXEMPLE DE TEST UNILATÉRAL



Pour tester $\mu = 0$ *contre* $\mu = 4$ *d'une distribution gaussienne de variance 1, il suffit de prendre une statistique exhaustive de* μ : $\bar{X}_n$.

La région de rejet est une zone suspecte, ou peu probable, de valeur de l'estimateur. Dans le cas de l'exemple précédent, on cherche un seuil assez éloigné de zéro, vers la droite, mais pas trop car sinon on risque de créer de faux négatifs.

NIVEAU D'ERREUR

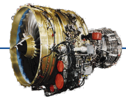
On choisit une probabilité $\alpha \in [0,1]$ correspondant au risque de rejeter H_0 à tort. La région de rejet R est définie par

$$\sup_{\theta \in \Theta_0} P_\theta(T \in R) = \quad P_{H_0}(T \in \mathrm{R}) \leq \alpha$$

On définit la <u>puissance</u> du test $\beta \in [0,1]$ comme étant la probabilité d'accepter H_1 si elle est vraie.

$$P_{H_1}(T \in R) = \beta$$

- α est la probabilité de rejeter H_0 à tord, c'est l'erreur de première espèce ;
- $1 - \beta$ est la probabilité de garder H_0 alors qu'elle est fausse, c'est l'erreur de seconde espèce.



TEST DE NEYMAN-PEARSON

Il s'agit d'un cas particulier où les deux hypothèses sont définies par deux valeurs d'un paramètre θ_0 et θ_1.

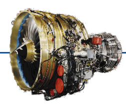
Pour tout niveau α il existe un test plus puissant que tous les tests de niveau α dont la région de rejet est donnée par :

$$R = \left\{ \frac{L_{\theta_0}(X_1, \dots, X_n)}{L_{\theta_1}(X_1, \dots, X_n)} < k_\alpha \right\}$$

avec k_α définit par $P_{\theta_0}(R) = \alpha$.

Ce test peut se calculer pour des statistiques exhaustives usuelles simples, quand les vraisemblances s'expriment facilement à partir des statistiques et pour des hypothèses unilatérales.

Il n'est pas exploitable dans le cas général.



TEST DU MAXIMUM DE VRAISEMBLANCE

On peut néanmoins utiliser la vraisemblance pour construire un test dans un cas entièrement général.

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L_\theta(X_1, \dots, X_n)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L_\theta(X1, \dots, X_n)}$$

Cette variable prend ses valeurs entre 0 et 1, elle est grande quand H_0 est vraie et petite quand H_0 est fausse.

Sous H_0,

$$-2 \log \Lambda \sim \chi^2(m)$$

où m est la différence entre le nombre de paramètres du cas général et le nombre de paramètres utilisés sous H_0.

SAFRAN
Snecma

APPLICATION

1. Description des données de deux moteurs
2. Observation des données
3. Intervalle de confiance pour la moyenne des marges
4. Test d'égalité des marges moyennes
5. Test d'égalité des variances

MOTEUR TURBOFAN
TEMPÉRATURE DE SORTIE DES GAZ

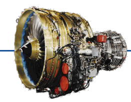
Nous disposons de deux moteurs d'avion à réaction sur lesquels on extrait des données des mesures de la température des gaz de sortie au décollage.

- *La température des gaz de sortie (EGT : Exhaust Gas Température) est un indicateur important de l'usure du moteur. Plus cette température est élevée, plus de l'énergie est perdue en agitation thermique au lieu d'être convertie en poussée. C'est en particulier un signe de l'usure des jeux entre les aubes des compresseurs et des turbine et le carter du turbofan.*

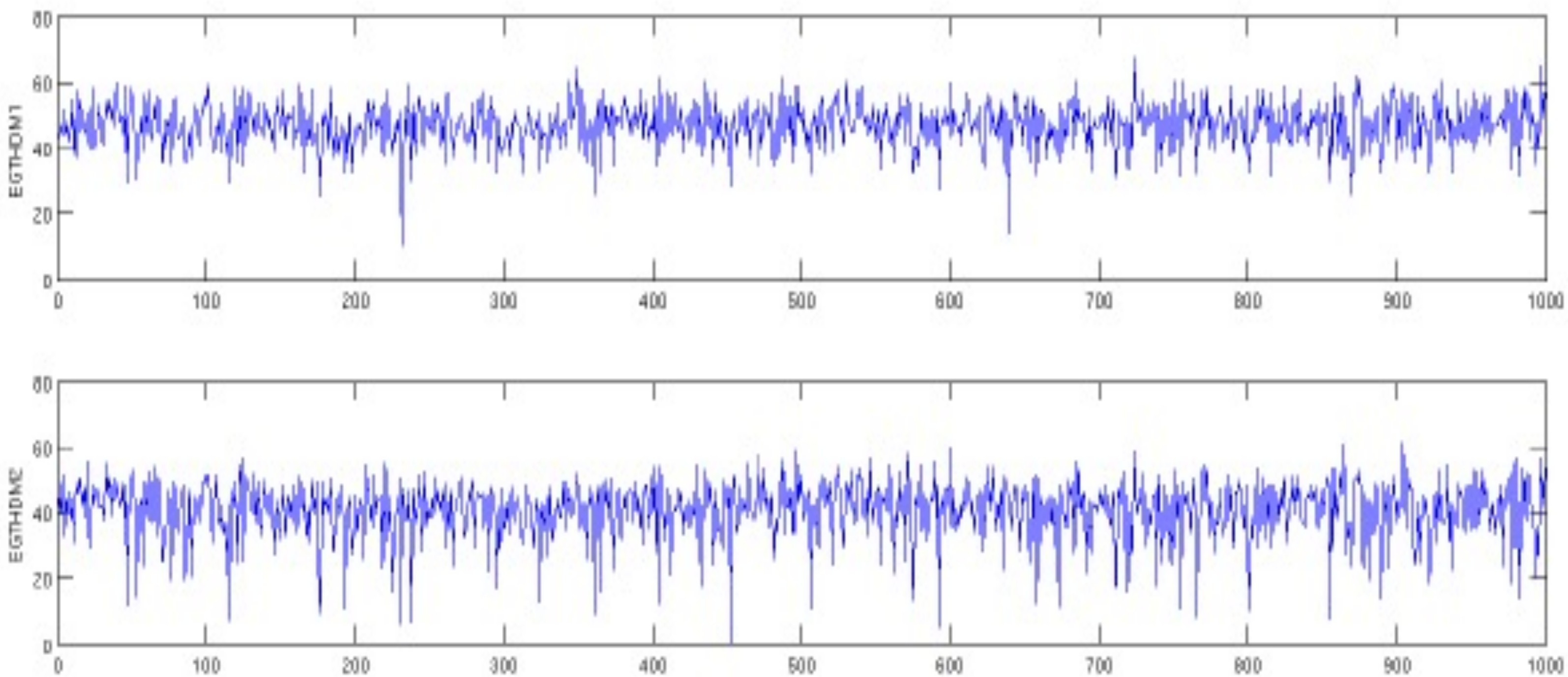
Données

- On dispose de 1000 observations pour chaque moteur.

- Pour que les données soient comparables on a pris soin de les normaliser en les rapportant à des conditions extérieures (météo, poids de l'avion, position de l'aéroport, etc.) équivalentes et on affiche la différence entre une valeur théorique maximale acceptable et la mesure ainsi renormalisée.
  - On construit ainsi une marge de tolérance : EGTHDM.



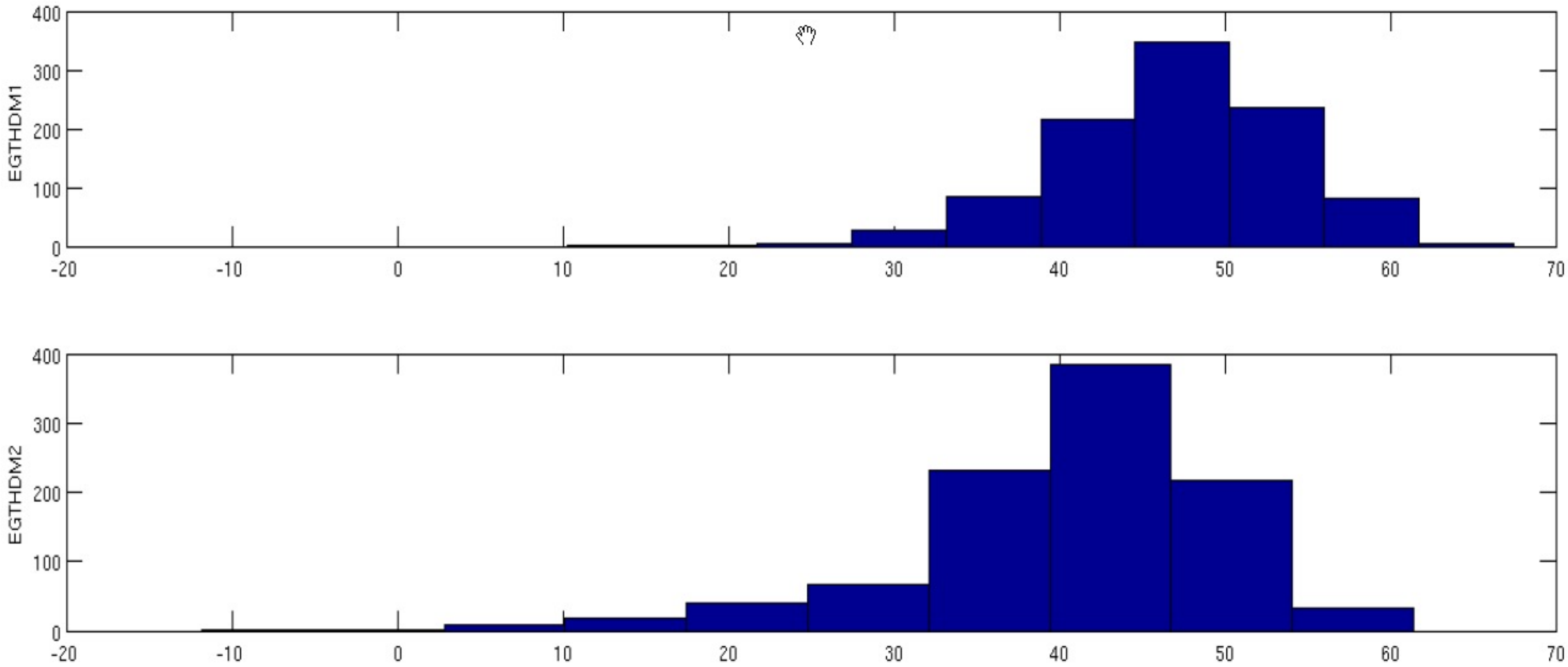
SAFRAN Snecma



1000 mesures (vols) des marges de température des gaz au décollage pour chaque moteur.

N=1000 vols	Moyenne empirique $\bar{X}_N$	Variance estimée S_N^2
Moteur 1	47.0073 °C	45.9138 °C²
Moteur 2	40.7666 °C	83.1309 °C²

RÉPARTITION DES DONNÉES



Distribution des observations.

INTERVALLE DE CONFIANCE POUR LA MOYENNE

Si on suppose la variance du moteur 1 exacte : $\sigma = \sqrt{45.9138}$

On rappelle qu'un intervalle de confiance de niveau α est donné par

$$IC_{\mu,n}(1-\alpha) = \left[\bar{x}_n - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \text{et } u_{97.5\%} = 1.96$$

Par exemple, sous Matlab pour estimer le quantile on utilise la formule

```
u = norminv(0.975,0,1) ;
```

L'intervalle de confiance à 95% pour la moyenne du moteur 1 est donc

$$IC_{\mu_1,1000}(95\%) = [46.5873\,, 47.4273]$$

pour

$\bar{x}_{1000} = 47.0073$

Le calcul équivalent, sans supposer la variance connue, en utilisant la loi de Student $T(999)$ **donne**

$$IC'_{\mu_1,1000}(95\%) = [46.5867\,, 47.4279]$$

La loi de Student est très proche de la loi normale dès que n est grand ($n > 20$).

INTERVALLE DE CONFIANCE POUR LA VARIANCE

Comme précédemment, on commence par supposer la moyenne connue :
$\mu = $ **47.0073 ° C**

L'estimateur sans biais de la variance est donc

$$S_n'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \frac{\sigma^2}{n}\chi^2(n)$$

Notons $k_{n,\frac{\alpha}{2}}$ et $k_{n,1-\frac{\alpha}{2}}$ les quantiles de la loi $\chi^2(n)$ au niveaux $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$.

Sous Matlab

$$k_{999,2.5\%} = \text{chi2inv}(0.025{,}999);$$

L'intervalle de confiance de la variance est

$$IC_{\sigma^2,n}(1-\alpha) = \left[\frac{nS_n'^2}{k_{n,1-\frac{\alpha}{2}}}, \frac{nS_n'^2}{k_{n,\frac{\alpha}{2}}}\right]$$

RÉSULTATS DES CALCULS DES BORNES DE L'INTERVALLE DE CONFIANCE

Finalement, pour le moteur n° 1 on obtient

$$IC_{\sigma_1^2} = [42.1421\,, 50.2198]$$

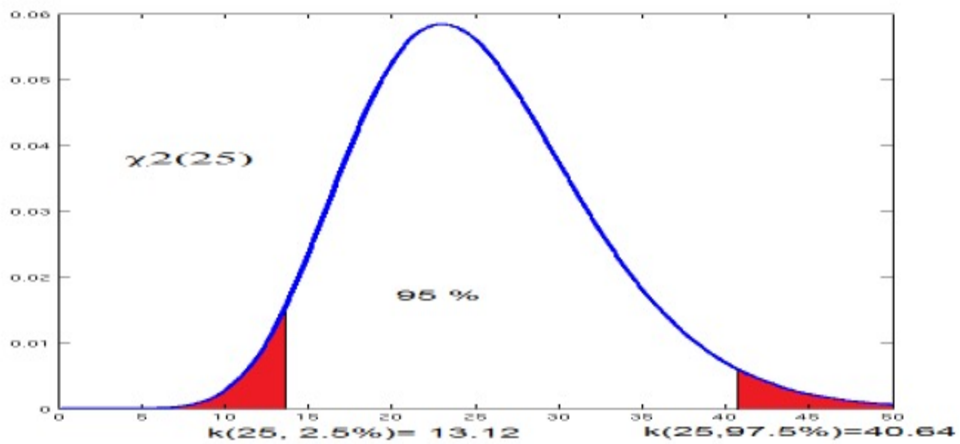
pour $s'^2_{1000} = 45.9138$

Quand la moyenne est inconnue, on doit l'estimer par sa valeur empirique S^2 et on perd un degré de liberté.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sim \frac{\sigma^2}{n-1}\,\chi^2(n-1)$$

On utilise les quantiles pour une loi $\chi^2(999)$ et on obtient.

$$IC'_{\sigma_1^2} = [42.1391\,, 50.2221]$$



Quantiles à 2.5% et 97.5% d'une loi $\chi^2(25)$.

TEST D'ÉGALITÉ DES MARGES MOYENNES DES DEUX MOTEURS

On cherche à tester si les moyennes des deux moteurs sont identiques.

Sinon comme le moteur 2 à une marge plus petite, on sait qu'il faudra en faire la maintenance plus tôt.

On veut donc tester

$$H_0:\ \mu_1 = \mu_2 \text{ contre } H_1:\ \mu_1 \neq \mu_2.$$

TEST DES MOYENNES
SI LA VARIANCE EST CONNUE

Comme n est très grand, la statistique de test est naturellement

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S'^2_1}{n_1} + \frac{S'^2_2}{n_2}}}$$

- Ici $n_1 = n_2 = 1000$, on sait aussi que comme n est grand l'utilisation la variance estimée est équivalente à considérer qu'elle est connue.
- On considère donc que $Z \sim N(0,1)$ sous H_0 au lieu de $T(1998)$.

La région de rejet à 95% est donc définie par $|Z| > 1.96$.

- Le calcul donne $z = 11.79$ donc l'hypothèse H_0 est rejetée et on peut conclure que avec une certitude de 95% on a $\mu_1 \neq \mu_2$.

En se contentant d'un test unilatéral avec H'_1: $\mu_1 > \mu_2$ on peut avoir un test plus précis.

- En effet, on a alors une région de rejet de type Z>1.64 à 95%. La conclusion, bien sûr est la même.

Par contre sachant qu'on avait pris un intervalle avec des quantiles symétriques et que 1.96 représente en fait un quantile supérieur à 97.5% on peut conclure que $\mu_1 > \mu_2$ avec une probabilité de 97.5%.

ET SI L'ÉCHANTILLON AVAIT ÉTÉ PLUS PETIT

Dans ce cas on aurait été obligé de calculer un estimateur sans biais de la variance et la statistique utilisée aurait été :

$$Z = (\bar{X}_1 - \bar{X}_2) / \sqrt{\frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right)}{n_1 + n_2 - 2}}$$

On teste d'abord que les deux variances sont égales, dans ce cas sous H_0, cette statistique suit une loi $T(n_1 + n_2 - 2)$.

SAFRAN
Snecma

TEST D'ÉGALITÉ DES VARIANCES

Aurait-on pu faire le test précédent facilement ?

Pour cela il faut s'assurer que les variances des deux moteurs sont égales.

- On veut donc tester $H_0\colon\ \sigma_1^2 = \sigma_2^2$ contre $H_1\colon\ \sigma_1^2 \neq \sigma_2^2$.
- La statistique naturelle est

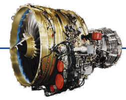
$$Z = \frac{S_1^2}{S_2^2}$$

Sous H_0 cette statistique suit une loi de Fisher $F(n_1 - 1, n_2 - 1)$.

- Pour $n_1 = n_2 = 1000$, les quantiles sont $f_{97.5\%} = 1.01$ et $f_{2.5\%} = 0.99$.
  Or

$$z = \frac{45.9138}{83.1309} = 0.57 < 0.99$$

donc on rejette H_0.



SAFRAN
Snecma

A SUIVRE

MODÈLES LINÉAIRES