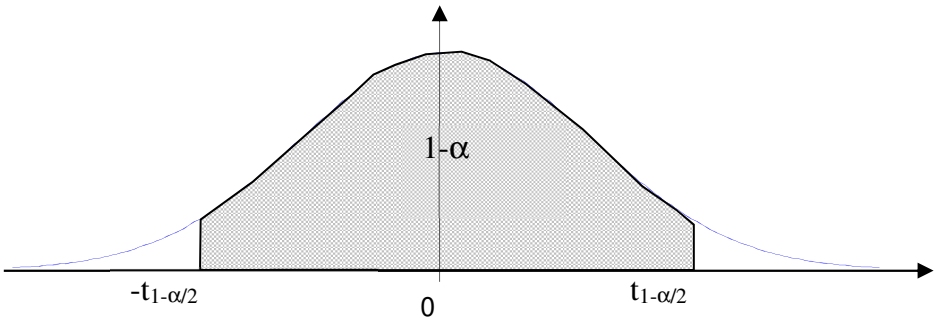
Filière IRC

1ème ANNEE

Cours de statistiques et probabilités



Bruno NEYRAN
Institut Universitaire de Technologie de Lyon
Université Claude Bernard LYON I

2018-2019

Plan

Bibliographie

STATISTIQUE APPLIQUEE A LA GESTION AVEC EXERCICES CORRIGES ET UTILISATION D'EXCEL. 7ème édition, Vincent Giard, ECONOMICA Collection GESTION, 576 pages, 33€.
On y trouve tout, ou presque (pas la fiabilité).

Mathématiques BTS Industriel. Spécialités des groupements B et C, Françoise Comparat, France Laplume, NATHAN Collection NATHAN-TECHNIQUE, 416 pages, 23,5€.
Les deux tiers concernent les statistiques et les probabilités présentées de manière claire (groupements B et C veut dire du secteur industriel). Je l'indique parce que l'éditeur me l'a envoyé gratuitement!

Mathématiques BTS/DUT industriels - Probabilités et statistique, Gérard CHAUVAT, Ediscience, 240 pages - 2005, 18 €.
A peu près tout ce qui est utile pour un DUT technique type GEII. Pour ceux qui ont perdu leur polycopié, mais n'existait pas quand j'ai fait le mien. Moins complet que ce polycopié. En plus l'éditeur me l'a envoyé gratuitement!

Mini Manuel de Probabilités et statistique - 2ème édition, Françoise Couty-Fredon, Jean Debord, Daniel Fredon, 256 pages – 2014, 17€.
Assez complet, avec des applications plutôt biologiques.

Probabilités Statistiques, assimiler et utiliser les statistique, cours et exercices corrigés, Luc PIBOULEAU, Technosup, 283 pages – 2006, 33€.
Cours bien résumé et des exercices détaillés.

Il se publie au moins 10 ouvrages de statistiques par an, beaucoup sont appliqués à la gestion et d'autres à la biologie et la médecine.

Sites internet:

http://www.agro-montpellier.fr/cnam-lr/statnet/
cours interactif et pédagogique avec exercices

http://www.fas.umontreal.ca/biol/legendre/BIO2041/index.html
http://biol10.biol.umontreal.ca/BIO2042/
Deux cours de statistique: de base pour Biostatistique I (*BIO2041*), de haut niveau pour Biostatistique II (*BIO 2042*) rédigés de manière très pédagogique, surtout les laîus des TP.

http://www.itl.nist.gov/div898/handbook/index.htm
Cours sur la fiabilité avec exemple

Statistiques
Introduction et définitions

La statistique a pour objet la description numérique d'ensemble nombreux. Elle vise au rassemblement, à la présentation et à l'analyse d'un grand nombre de données en vue de la prise de décision.
On distingue deux étapes dans le traitement des données.

- étape descriptive: on décrit l'ensemble des données grâce à des tableaux, graphiques et un nombre réduit de paramètres
- étape déductive: à partir de l'étude d'un sous ensemble des données, on cherche à estimer des caractéristiques de la population totale. On se servira des **probabilités** pour déterminer des intervalles de confiance et tester des hypothèses.

Quelques définitions

Population: désigne un ensemble fini ou infini d'éléments.

Individu: un élément de l'ensemble étudié

Echantillon: ensemble de n éléments prélevés dans une population, qu'on appelle alors population-mère.

Effectif: nombre d'individus dans une population (N) ou dans un échantillon (n).

Sondage: prélèvement d'un échantillon dans une population.

Caractère: caractéristique que l'on étudie. On peut lui attribuer plusieurs valeurs différentes ou **modalités**.

Variable: caractéristique mesurable qui s'exprime par un nombre, donc **quantitative**. Une variable dont la valeur est déterminée en fonction du résultat d'une expérience aléatoire est appelée variable **aléatoire**.

Evènement: résultats d'une expérience aléatoire

Probabilité: nombre réel compris entre 0 et 1 qui, associé à un évènement, mesure les chances de réalisation de cet évènement au cours d'une épreuve donnée. Si cette épreuve consiste en un prélèvement d'un élément dans un ensemble, la probabilité est égale à la fréquence de cet évènement au cours de l'épreuve.

Probabilités: La théorie des probabilités vise à **modéliser** le hasard, les phénomènes aléatoires, l'imprévisible. Les probabilités utilisent l'expérience passée et donc la statistique pour déterminer des lois sous-jacentes et ainsi prédire l'avenir possible.

Modèle statistique: désigne une loi de probabilité qui semble bien résumer des observations. On utilisera les propriétés théoriques de cette loi à des fins de prévisions pour guider l'action et prendre des décisions.

Statistique descriptive

1 Caractère qualitatif, quantitatif, discret, continu

Un caractère peut être **quantitatif** s'il s'exprime par un nombre. On l'appelle alors variable statistique (ex: "nombre d'enfants d'une famille"). On peut appliquer une notion d'ordre sur un paramètre quantitatif (3 enfants > 2 enfants).
Un caractère non quantitatif est **qualitatif** (ex: "commune de naissance", "couleur de cheveux"). On ne peut pas appliquer une notion d'ordre sur un paramètre qualitatif (Saint-Etienne <> Lyon, blond<>brun).
Une variable statistique est **discrète** si elle prend des valeurs isolées (si elle varie par saut) (ex: "nombre de résistances défectueuses"). Elle est **continue** si elle est susceptible de prendre toutes les valeurs d'un intervalle de $\Re$ (ex: valeur d'une résistance).

2 Présentation des données

2.1 Présentation des données par classe

On va voir cela sur un exemple.

Population = {étudiants du groupe 1 de $2^{ème}$ année}
Individu = 1 étudiant
Effectif : N = 23
Caractère : X = "nombre de frères et sœurs"; de type quantitatif discret

Tableau de distribution

Modalité x_i	effectifs n_i	fréquence $f_i = n_i / N$	effectifs cumulés $n_i^+ = \sum_{j=1}^{i} n_j$	Fréquences cumulées $f_i^+ = \sum_{j=1}^{i} f_j$
0	5	0,2173913	5	0,217391304
1	9	0,39130435	14	0,608695652
2	5	0,2173913	19	0,826086957
3	3	0,13043478	22	0,956521739
>3	1	0,04347826	23	1
Total	N=23	1		

fig 1-(a): Exemple de tableau de distribution

Les notations suivantes sont constamment reprises par la suite:

X : Caractère ou variable aléatoire
x_i : modalité (ou classe) i, $[x_{i\,min} ; x_{i\,max}]$ pour un caractère quantitatif continu
n_i : effectif de la classe i
f_i : fréquence de la classe i
N : effectif total

2.2 Présentation des données sous forme de graphe

2.2.1 Caractère qualitatif ou quantitatif discret

La représentation des effectifs ou fréquences simples est possible par les diagrammes rectangulaires (en bande) ou circulaires à secteurs (camemberts).

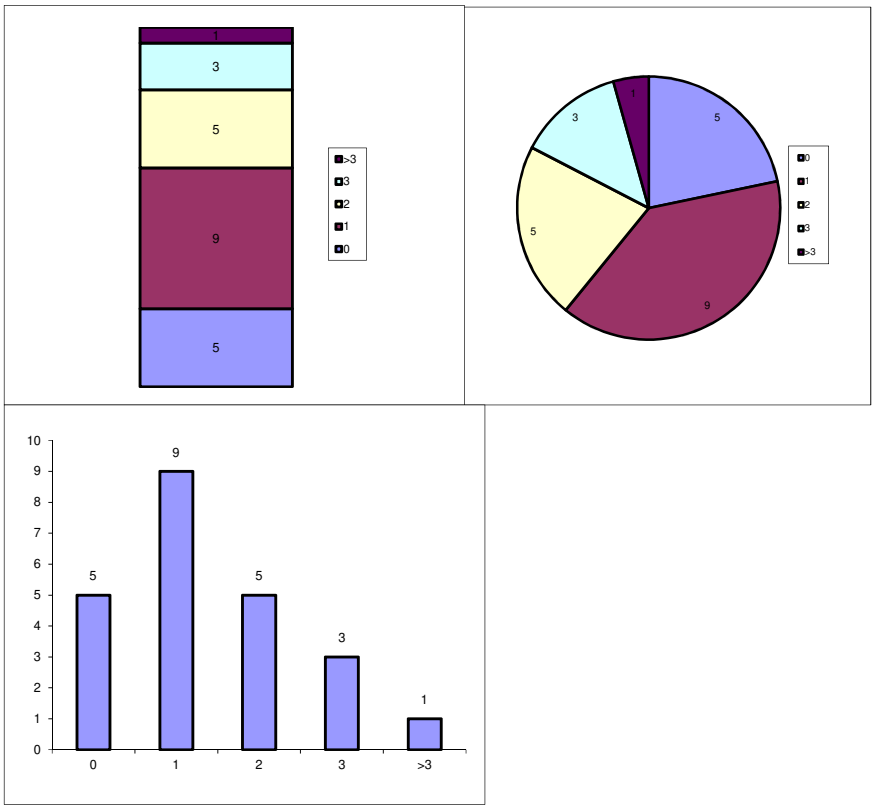


fig 1-(i): Diagramme en bande, en camembert, en tuyau d'orgue

Pour les caractères **quantitatifs** (notion d'ordre) discrets, on peut aussi tracer les effectifs ou fréquences cumulés.

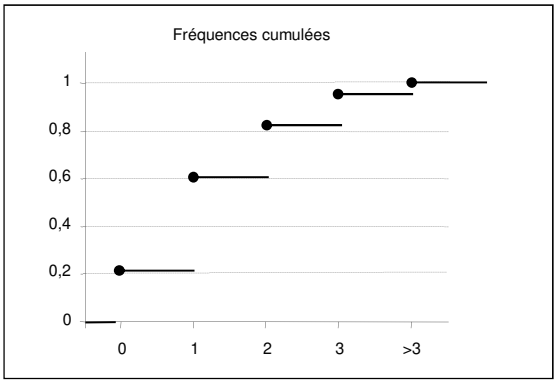


fig 1-(ii): Diagramme en escalier de la fréquence cumulée

2.2.2 Caractère quantitatif continu : histogramme

Dans ce cas, on regroupe les valeurs de la variable dans des classes correspondant chacune à un **intervalle** de valeurs de la variable.

3 Les différents paramètres résumant les données

Deux principaux paramètres résument un ensemble de données :

- un paramètre de position situant autour de quelle valeur centrale se trouvent les données,
- un paramètre de dispersion indiquant la répartition des données autour de la valeur centrale

Ces paramètres peuvent être calculés selon deux méthodes, une basée sur l'ordre à partir d'un classement des données, et une algébrique suivant une formule.

3.1 Paramètres basés sur l'ordre

3.1.1 Paramètre de position

3.1.1.1 Le mode

C'est la valeur d'une série, où modalité, correspondant à l'effectif le plus fort.

Ex: dans l'exemple précédent c'est 1 "frères et sœurs".
Pour un caractère quantitatif continu on parle de classe modale.

3.1.1.2 La médiane (notée me)

C'est la valeur qui partage la série **classée** en deux sous groupes de même effectif.

Caractère quantitatif discret

- l'effectif est un nombre impair: dans l'exemple précédent on a 23 étudiants. Après classement, l'étudiant qui est classé $12^{ième}$ sépare la classe en 2 sous-groupes de 11 élèves. De plus, il a 1 "frères et sœurs" qui correspond donc à la médiane
- l'effectif est un nombre pair: si dans l'exemple précédent on rajoute un étudiant qui a 2 "frères et sœurs". On a maintenant 24 étudiants. Les étudiants classés $12^{ième}$ et $13^{ième}$ séparent la classe en 2 sous-groupes de 11 élèves. La valeur moyenne de ces deux étudiants fournit la médiane: me = 1 "frères et sœurs" car ces deux étudiants ont 1 "frères et sœurs".

Caractère quantitatif continu

On la détermine par interpolation linéaire.

Dépense pour un article (€) x_i	Nombre ménages n_i	n_i^+
[0;100[	40	40
[100;200[	25	65
[200;300[	45	110
Total	N=110	

fig 1-1.: Caractère quantitatif continu regroupé en classe

La médiane est élément de la première classe dont l'effectif cumulé n_i^+ dépasse N/2=55 (ou dont la fréquence cumulée f_i^+ dépasse 0,5).
Puis on fait une interpolation linéaire entre les deux bornes de la classe médiane [100;200[.
me=100+(200-100)(55-40)/(65-40)
Ainsi 50% des ménages dépensent moins de 160€ et 50% des ménages dépensent plus de 160€.

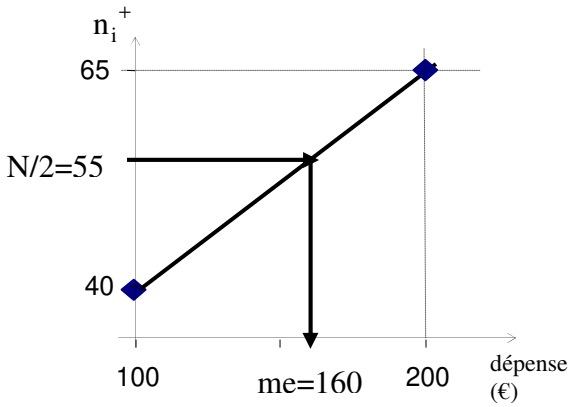


fig 1-2.: interpolation linéaire entre les deux bornes de la classe médiane

3.1.1.3 Les quantiles

Ce sont des valeurs qui partagent la série classée par ordre croissant en x partie de même effectif.
Si x=2 => médiane me
Si x=3 => 2 terciles T_1, T_2
Si x=4 => 3 quartiles Q_1, Q_2, Q_3. Dans la série de N valeurs classée (x_i), Q_1 est la donnée x_i dont l'indice i est le plus petit entier supérieur ou égal à N/4.
Si x=10 => 9 déciles D_1, ... , D_9
Pour leur calcul, on procède par interpolation linéaire comme cela a été décrit pour le calcul de la médiane.

3.1.2 Paramètre de dispersion

Notion de dispersion: soit deux séries de notes
8/8/10/11/13 ; $\bar{x}$ = me = 10
2/5/10/16/17 ; $\bar{x}$ = me = 10

Ces deux séries ont la même moyenne et la même médiane, mais elles ne sont pas comparables. La seconde a des valeurs plus aléatoires, dispersées.
Il existe plusieurs paramètres pour mesurer la dispersion.

3.1.3 L'étendue

C'est x_{max}-x_{min}. Dans le cas 1 elle vaut 13-8=5, dans le cas 2 elle vaut 17-2=15.

Problème : ce paramètre dépend uniquement des deux valeurs extrêmes qui peuvent être fausses ou aberrantes.

3.1.4 Intervalle interquartile

C'est l'écart absolu $Q_3 - Q_1$, (voir **Erreur ! Source du renvoi introuvable.** pour la définition de Q_1 et Q_3)
Pour comparer plusieurs séries, on utilise l'écart ou intervalle interquartile relatif $(Q_3 - Q_1)/me$

3.1.5 Résumé graphique : diagramme en boite (ou boite à moustaches) (box-plot)

On résume une série de valeurs par un graphique utilisant la médiane me, les deux quartiles Q_1, Q_3.et les valeurs minimum et maximum de la série.

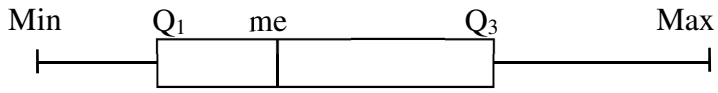


fig 1-(iii): Diagramme en boite (ou boite à moustaches) (box-plot)

Ce diagramme est souvent employé pour comparer la répartition de deux séries.

3.2 Paramètres algébriques (calculés)

3.2.1 Paramètre de position : la moyenne

On appelle $\bar{x}$ la moyenne arithmétique. $\bar{x} = \dfrac{\sum n_i x_i}{N} = \sum f_i x_i$

Cas discret : la formule s'applique directement.

Cas continu : si toutes les valeurs de la variable sont différentes on a constamment $n_i = 1$. Si les valeurs sont regroupées en classe, on prend pour x_i le centre de la classe et on obtient alors une valeur approchée de la moyenne.
Dans le cas du tableau de la figure **Erreur ! Source du renvoi introuvable.**, on obtient:
$\bar{x}$ = (50*40+150*25+250*45)/110 = 154,55

Propriété

Soit la variable Y, construite par une relation linéaire en fonction de la variable X : Y = a X + b
où a et b sont deux constantes réelles. Sa valeur moyenne $\bar{y}$ a pour valeur : $\bar{y} = a\bar{x} + b$

Preuve: $\bar{y} = \dfrac{\sum n_i y_i}{N} = \dfrac{\sum n_i (a x_i + b)}{N} = a\dfrac{\sum n_i x_i}{N} + b\dfrac{\sum n_i}{N} = a\bar{x} + b$

Définition: variable centrée

Soit une variable X, de moyenne $\bar{x}$, on appelle variable centrée X_c la variable définie par $X_c = X - \bar{x}$.

Une variable centrée est à valeur moyenne nulle.

3.2.2 Paramètres de dispersion : variance et écart-type

3.2.2.1 Variance expérimentale

Les écarts entre les valeurs et la moyenne rendent bien compte de la notion de dispersion. Si on somme tous ces écarts, il y a compensation entre les écarts positifs pour les valeurs qui sont supérieures à la moyenne et les écarts négatifs pour celles qui sont inférieures.

On pourrait utiliser la valeur absolue des écarts, mais la fonction "valeur absolue" pose des problèmes car elle présente une dérivée discontinue en 0. On préfère prendre le carré des écarts pour définir la variance expérimentale s_X^2 comme la valeur moyenne du carré des écarts:

$$s_X^2 = \frac{SS_X}{N} = \frac{\sum n_i (x_i - \overline{x})^2}{N}$$; SS_X désigne la somme des carrés des écarts (Sum Square)

Nous verrons plus tard dans le chapitre sur l'estimation la variance *estimée*.

Propriété

La variance peut aussi s'écrire : $s_X^2 = \frac{\sum n_i x_i^{\,2}}{N} - \overline{x}^2$. Cette formule est plus simple lors du calcul pratique de la variance.

Preuve: $s_X^2 = \frac{\sum n_i (x_i - \overline{x})^2}{N} = \frac{\sum n_i (x_i^{\,2} - 2x_i\overline{x} + \overline{x}^2)}{N} = \frac{\sum n_i x_i^{\,2}}{N} - 2\overline{x}\frac{\sum n_i x_i}{N} + \overline{x}^2 \frac{\sum n_i}{N}$

$$s_X^2 = \frac{\sum n_i x_i^{\,2}}{N} - 2\overline{x}\,\overline{x} + \overline{x}^2 = \frac{\sum n_i x_i^{\,2}}{N} - \overline{x}^2$$

Exemple: Dans le cas 1 précédent, la variance expérimentale vaut $(2(2)^2+1^2+3^2)/5=3,6$,
dans le cas 2 elle vaut $(8^2+5^2+6^2+7^2)/5=34,8$
ou encore en utilisant la propriété, dans le cas 1 $(2(8)^2+10^2+11^2+13^2)/5-10^2=103,6-10^2=3,6$
et dans le cas 2 $(2^2+5^2+10^2+16^2+17^2)/5-10^2=134,8-10^2=34,8$

Remarque : Si les valeurs d'une série continue sont regroupées en classe, on fait comme pour le cas du calcul de la moyenne, en prenant pour x_i le centre de la classe et on obtient alors une valeur approchée de la variance.
Dans le cas du tableau de la figure **Erreur ! Source du renvoi introuvable.**, on obtient:
$s_X^2 = (40(50-154,55)^2 + 25(150-154.55)^2 + 45(250-154.55)^2)/110 = 7706,6$

Remarque : la variance peut s'interpréter de manière analogue à la puissance moyenne de l'ondulation (valeur moyenne nulle) d'un signal périodique x(t) : $P_{moy\ ond} = \frac{1}{T}\int_{\theta}^{\theta+T} \left(x(t) - x_{moy}\right)^2 dt$

3.2.2.2 Ecart-type expérimentale

C'est la racine carrée de la variance expérimentale: $s_X = \sqrt{s_X^2}$

Exemple: Dans le cas 1 précédent $s_X = \sqrt{3,6} = 1,897$, dans le cas 2, $s_X = \sqrt{34,8} = 5,9$

Remarques : - l'écart type possède la **même unité** que la variable X.
- l'écart type et la variance sont toujours positifs.

Remarque : l'écart type peut s'interpréter de manière analogue à la valeur efficace de l'ondulation (valeur moyenne nulle) d'un signal périodique x(t) : $V_{eff\ ond} = \sqrt{P_{moy\ ond}} = \sqrt{\frac{1}{T}\int_{\theta}^{\theta+T} \left(x(t) - x_{moy}\right)^2 dt}$

Propriété

Soit la variable Y, construite par une relation linéaire en fonction de la variable X : Y = a X + b
où a et b sont deux constantes réelles.
La variance de la variable Y vaut: $s_Y^2 = a^2 s_X^2$
L'écart type de la variable Y vaut: $s_Y = |a| s_X$

La variance et l'écart type de la variable Y ne dépendent pas du terme additif b.

Preuve: $s_Y^2 = \frac{\sum n_i (y_i - \overline{y})^2}{N} = \frac{\sum n_i (ax_i + b - (a\overline{x} + b))^2}{N} = \frac{\sum n_i (ax_i - a\overline{x})^2}{N} = a^2 \frac{\sum n_i (x_i - \overline{x})^2}{N}$

3.2.3 Avec Excel

Si les valeurs de X se trouvent sur la ligne 1 entre les colonnes A et H, les commandes suivantes:

MOYENNE(A1:H1) donne la moyenne
MEDIANE(A1:H1) donne la médiane
VAR.P.N (A1:H1) donne la variance expérimentale (P pour Population)
ECARTYPE.PEARSON (A1:H1) donne l'écart type expérimentale (pour Population)

Attention: VAR.S fait une estimation de la variance en supposant que les données sont un échantillon de la population totale (voir chapitre sur l'estimation). De même ECARTYPE.STANDARD fait une estimation l'écart type en supposant que les données sont un échantillon de la population totale.

$$\text{VAR.S} = (\text{VAR.P.N}) . \frac{N}{N-1} ; \ \text{ECARTYPE.STANDARD} = (\text{ECARTYPE. PEARSON }) . \sqrt{\frac{N}{N-1}}$$

3.2.4 Les calculatrices ATTENTION

Dans leur documentation les calculatrices donnent :

- σ_x écart-type de la population (pour la population)
- s_x écart-type de l'échantillon (pour l'échantillon)

MAIS le sens donné ici au mot **population** veut dire qu'on calcule l'écart-type sur des données correspondant à TOUTE la **population (échantillon=population)** pour σ_x.
MAIS le sens donné ici au mot **échantillon** veut dire qu'on calcule l'écart-type de la population sur des données correspondant à un **échantillon de la** population pour s_x. Donc pour la calculatrice

- σ_x écart-type de l'ensemble des données considérées comme une **population**
- s_x écart-type de la population mère calculée à partir des données considérées comme un **échantillon**

Cette notation est **contraire à ce qui prit dans la majorité des documents** où

- s_x est l'écart-type de l'échantillon (notion statistique)
- σ_x écart-type de la population, **estimé** à partir de cet échantillon (voir estimateur dans le chapitre probabilité).

4 Tracé et autre résumés des données

4.1 Le coefficient de variation

On peut regarder chaque écart-type et constater qu'il est élevé ou pas, mais on ne pourrait comparer chaque écart-type qu'à un autre écart-type exprimé dans la même unité. Le coefficient de variation est égal à l'écart-type divisé par la moyenne, c'est-à-dire : $C_v = \frac{s_x}{\overline{x}}$. On peut ainsi comparer l'étendu relative de deux séries de mesure d'unité différentes.

4.2 Diagramme de Pareto

Le diagramme de Pareto est un histogramme particulier, un moyen simple pour classer les phénomènes par ordre d'importance. Il permet souvent d'illustrer que pour un phénomène, 20% des causes produisent 80% des effets.
Si on considère les causes des défauts de fabrication et leur fréquence

Défauts	**fréquence**
rayures	3639
poussières	666
taches	132
hors tolérance	389

fonctionnement	313
erreur jugement	4583
inversions	846
manquants	152
emballage	2168
autres	67

Un diagramme de Pareto consiste à tracer l'histogramme du nombre (ou fréquence) de défaut en fonction des défauts ORDONNES du plus fréquent au plus rare. On trace également la somme cumulée des fréquences sur ce même graphique.

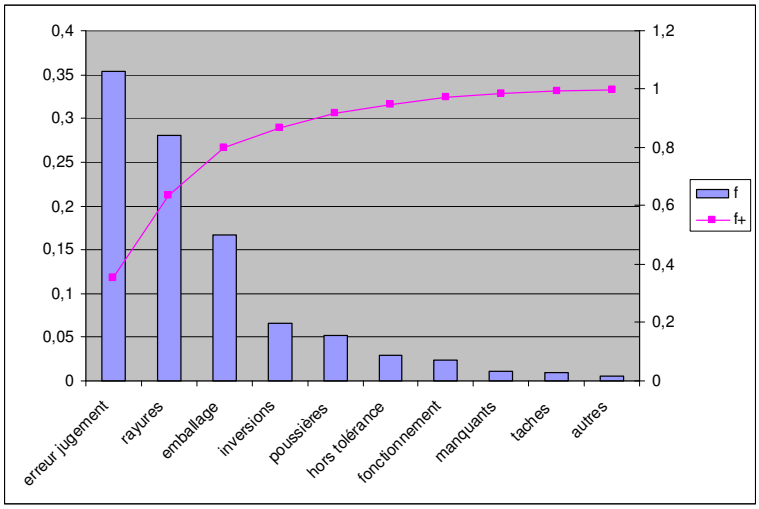


fig 1-(b): Diagramme de Pareto des causes de défaut

4.3 Tracé de Bland et Altman

Dans le but d'étudier l'accord entre deux techniques de mesures, en particulier si on veut comparer une nouvelle technique avec une technique de référence, on dispose d'une série de deux mesures appariées, c'est-à-dire deux mesures effectuées sur un même échantillon. Pour évaluer la concordance des mesures, on trace la différence entre les résultats des deux méthodes en fonction de la moyenne des deux méthodes.

Cette technique a été décrite par JM Bland et D.G. Altman (1986)

Ce tracé permet de révéler une relation entre les différences et les moyennes, pour la recherche de tout biais systématique et pour identifier les valeurs aberrantes possibles.

Probabilité

En probabilité, on dit qu'une expérience est aléatoire lorsque tous les résultats possibles sont connus à l'avance, mais que seul le hasard réalise un résultat plutôt qu'un autre.

1 Définition

1.1 Evènements

Lors d'une expérience aléatoire un résultat possible est appelé évènement élémentaire. L'ensemble Ω des résultats possibles forme un ensemble appelé univers.
Un évènement A est un sous ensemble de l'univers. Il est réalisé lorsque le résultat de l'expérience aléatoire est l'un des évènements élémentaires de A.
Deux évènement A et B sont incompatibles si $A \cap B = \varnothing$
L'évènement contraire ou complémentaire de A est noté $\overline{A}$: $A \cap \overline{A} = \varnothing$ et $A \cup \overline{A} = \Omega$

1.2 Probabilité

Soit un univers Ω associé à une expérience aléatoire, et A un évènement. La probabilité de A est un nombre réel , noté P(A), tel que:

$$0 \le P(A) \le 1 ; P(\Omega) = 1 ; P(\varnothing) = 0 ; P(\overline{A}) = 1 - P(A)$$

Propriété

Si A et B sont deux évènements : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Si A et B sont incompatibles: $P(A \cup B) = P(A) + P(B)$

1.3 Pourcentage et probabilité

On dit que les N évènements élémentaires d'un univers Ω sont équiprobables si la probabilité de chacun d'eux est 1/N.

Dans un univers équiprobable, la probabilité d'un évènement A est : $P(A) = \dfrac{Card(A)}{Card(\Omega)} = \dfrac{Card(A)}{N}$ où

Card(A) désigne le cardinal, le nombre d'élément de l'ensemble A.
La proportion ou pourcentage du nombre d'éléments de A peut s'interpréter comme une probabilité.

Exemple: on considère la valeur du lancement d'un dé à 6 faces. L'ensemble Ω des résultats possibles contient 6 éléments $\Omega = \{1;2;3;4;5;6\}$. Chaque élément est un évènement élémentaire. Chaque évènement élémentaire est équiprobable et sa probabilité est de 1/6. L'évènement "le résultat est pair" est le sous-ensemble $A = \{2;4;6\}$. L'évènement "le résultat est multiple de 3" est le sous-ensemble $B = \{3;6\}$. On a:

$$P(A) = \frac{Card(A)}{Card(\Omega)} = 3/6 = 1/2 ; P(B) = \frac{Card(B)}{Card(\Omega)} = 2/6 = 1/3 ; \ A \cap B = \{6\};$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/3 - 1/6 = 4/6 ;$$
on retrouve $P(A \cup B) = \dfrac{Card(A \cup B)}{Card(\Omega)} = \dfrac{Card(\{2;3;4;6\})}{Card(\Omega)} = 4/6$

2 Probabilité conditionnelle

2.1 Définition

Soit A et B deux évènements d'une expérience aléatoire d'univers Ω. La probabilité de réalisation de A sachant que B est réalisé, encore appelée probabilité conditionnelle, est notée P(A/B) ou $P_B(A)$. Elle a pour valeur:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Illustration:

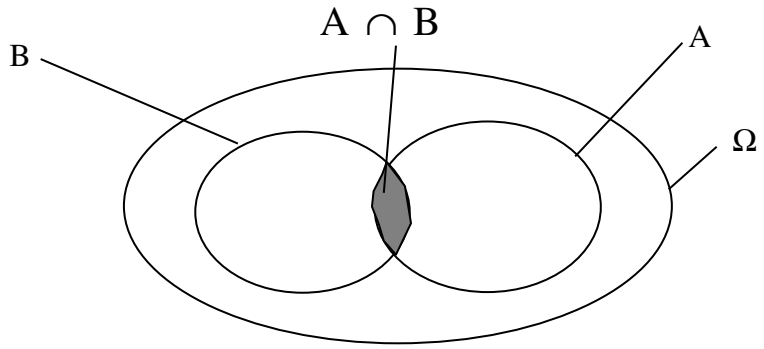


fig 3-1: probabilité conditionnelle

2.2 Théorème de Bayes

$$P(A/B) = P(B/A)\frac{P(A)}{P(B)}\text{; car } P(A \cap B) = P(A/B)xP(B) = P(B/A)xP(A)$$

2.3 Evènements indépendants

Deux évènements A et B sont indépendants, si et seulement si, on a:

$$P(B/A) = P(B) \text{ ou } P(A/B) = P(A)$$

Illustration:

Cela signifie que la proportion de B si A est déjà vérifiée est la même que la proportion de B dans l'univers entier

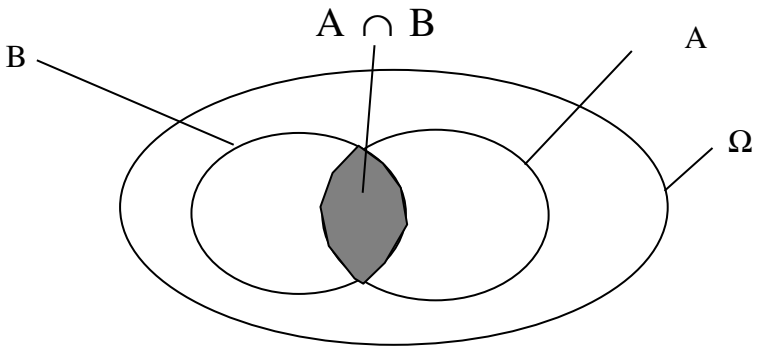


fig 3-2: Evènements indépendants

Sur le dessin, le rapport de surface B/Ω est le même que le rapport de surface $A \cap B/A$.
Alors le rapport de surface A/Ω est le même que le rapport de surface $A \cap B/B$.

Exemple: on reprend le cas du lancement d'un dé à 6 faces. La probabilité de réalisation de A (tirage pair) sachant que B (tirage multiple de 3) est réalisé vaut:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/3} = 1/2\text{ ; elle est égale à P(A)}$$

De même: $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{1/2} = 1/3$ est égale à P(B)

Les évènements A et B sont donc indépendants

Exemple: on teste deux cartes d'un appareil électronique. Si ces deux cartes ont été réalisées de manière totalement indépendante (chaîne de fabrication, approvisionnement en composant , ... différents), la probabilité de panne d'une carte est indépendante de la probabilité de panne de l'autre carte. Si maintenant on suppose que les cartes sont câblées en utilisant des composants communs (même composant provenant du même fournisseur), si une des cartes A tombent en

panne à cause de ces composants communs, il y a de fortes chances que l'autre carte B qui contient les mêmes composants tombe aussi en panne. Les deux évènements "carte A en panne" et "carte B en panne" ne sont plus indépendants. Ainsi $P(B/A)$ qui représente la probabilité que la carte B tombe en panne **sachant** que la carte A et en panne est supérieure à $P(B)$ la probabilité simple que la carte B tombe en panne.

2.4 Probabilité composée d'évènements indépendants

Si deux évènements A et B sont indépendants alors :

$$P(A \cap B) = P(A)\,P(B)$$

car $P(A/B) = \dfrac{P(A \cap B)}{P(B)} = P(A)$

Exemple: un appareil électronique contient 2 cartes. La première a une probabilité p_1=0,001 de panne, la seconde une probabilité p_1=0,003 de panne. Pour le calcul de la probabilité d'avoir un appareil qui ne tombe pas en panne il faut raisonner en terme d'appareil en "état de marche", c'est-à-dire avec 3 cartes en fonctionnement. La probabilité d'avoir une première carte en "état de marche". est $(1-p_1)$= 0,999. La probabilité d'avoir une seconde carte en "état de marche". est $(1-p_2)$= 0,997. Les possibilités de panne sur chacune des cartes sont supposées indépendantes, la probabilité d'avoir un appareil en "état de marche" est donc est $(1-p_1)(1-p_2)$= 0,9969x0,997=0,996. La probabilité d'avoir un appareil en panne est de 1-0,996=0,004.

2.5 Exemple

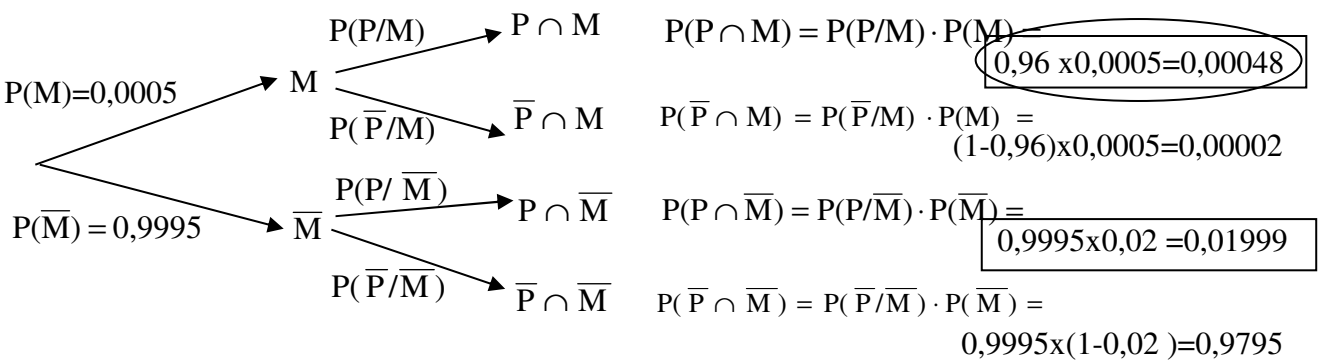
Pour dépister une maladie, on applique un test. Si le patient est effectivement atteint, le test donne un résultat positif dans 96% des cas. Mais il se peut aussi que le résultat du test soit positif alors que le patient est en bonne santé, et ceci se produit dans 2% des cas. En moyenne 0,05% des patients sont atteints de la maladie à dépister.

Soit P l'événement: "le test est positif", M l'événement: "le patient a la maladie" et $\overline{M}$ l'événement complémentaire de M: "le patient n'a pas la maladie".

b) Calculer la probabilité pour qu'un client soit atteint sachant que son test a été positif.

La résolution peut se faire soit par application du théorème de Bayes, soit en construisant un arbre où tous les cas élémentaires sont envisagés.

- on construit l'arbre suivant où sont portés les différentes probabilités



A chaque niveau (chaque colonne), la somme de tous les évènements possibles donne 1.

Tous les cas où le test est positif sont encadrés. Le cas où le test est positif et le patient malade est encerclé. La probabilité pour qu'un client soit atteint sachant que son test a été positif est donc:

$$\frac{0,00048}{0,00048+0,01999} = 0,02345, \text{ soit } 2,3\%$$

- Par la formule de Bayes, l'énoncé s'écrit:

  $P(P/M) = 0,96$; $P(P/\overline{M}) = 0,02$ et $P(M) = 0,0005$ et $P(\overline{M}) = 1-P(M)$

  $$P(M/P) = P(P/M)\frac{P(M)}{P(P)} \text{ or } P(P) = P(P/M)P(M) + P(P/\overline{M})P(\overline{M})$$

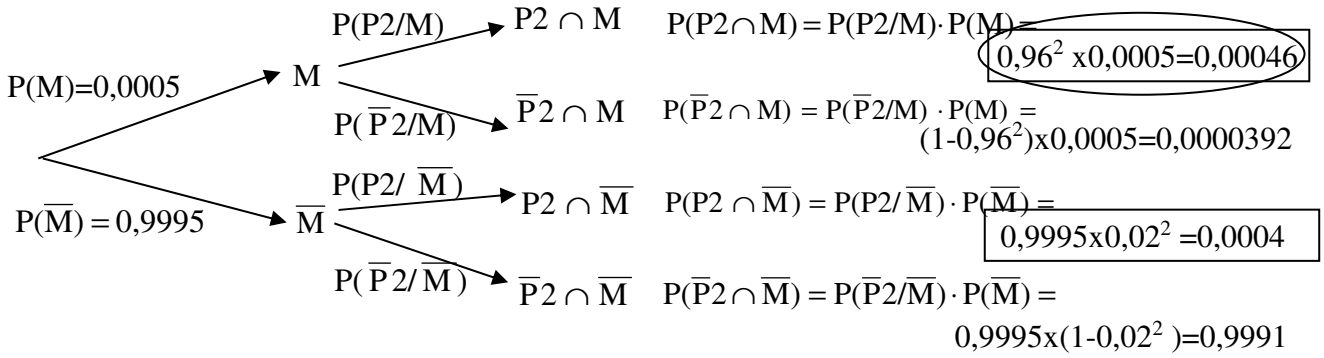
$P(M/P) = 0,96\times0,0005/[0,96\times0,0005+0,02\times0,9995] = 0,02345$

- On peut construire le tableau des probabilités

	P	$\overline{P}$	total
M	$P(P\cap M) = 0,96x0,0005$	$P(\overline{P}\cap M) = (1-0,96)x0,0005$	$P(M) = 0,0005$
$\overline{M}$	$P(P\cap \overline{M}) = 0,9995x0,02$	$P(\overline{P}\cap \overline{M}) = 0,9995x(1-0,02)$	$P(\overline{M}) = 0,9995$
total	$P(B) = 0,02047$	$P(\overline{B}) = 0,97953$	1

c) Sous les mêmes conditions calculer la probabilité pour qu'un client soit atteint sachant que deux tests (indépendants) ont été positifs.
Soit P2 l'événement:"deux tests sont positifs"
$P(P2|M) = P(P|M)^2$ et $P(P2|M) = P(P|M)^2$ car les tests sont indépendants

- on construit l'arbre suivant où sont portés les différentes probabilités



Tous les cas où le test est positif sont encadrés. Le cas où le test est positif et le patient malade est encerclé. La probabilité pour qu'un client soit atteint sachant que deux tests ont été positifs est donc: $\frac{0,00046}{0,00046+0,0004} = 0,5354$, soit 53,5%

- Par la formule de Bayes:

$$P(M/P2) = P(P2/M)\frac{P(M)}{P(P2)} \text{ or } P(P2) = P(P2/M)P(M) + P(P2/\overline{M})P(\overline{M})$$

$P(M/P2) = 0,96^2\times0,0005/[0.96^2\times0.0005+0.02^2\times0.9995] = 0,5354$

2.6 Formule des probabilités totales

Soit un système complet d'évènement (A_i forme une partition de l'ensemble Ω des évènements, les A_i sont disjoints et leur union égale Ω), soit B un évènement, alors :

$$P(B) = \sum_i P(B\cap A_i) = \sum_i P(A_i)P(B/A_i)$$

Cette formule permet de calculer la probabilité d'un événement B en le décomposant suivant un système complet d'événements.

Exemple: On dispose de 3 urnes U1, U2, U3, chacune contient 10 boules; parmi elles, U1 contient 1 blanche, U2 contient 2 blanches, et U3 contient 6 blanches. On tire au hasard une boule. Quelle est la probabilité d'obtenir une blanche?
On note B l'événement "on obtient une boule blanche" et A_i l'événement "on tire la boule dans l'urne Ui". $\{A_1,A_2,A_3\}$ forme un système complet d'événements, et :

$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3)$$

$$P(B) = \frac{1}{3}.\frac{1}{10} + \frac{1}{3}.\frac{2}{10} + \frac{1}{3}.\frac{6}{10} = \frac{3}{10}$$

2.7 Sensibilité, spécificité, valeurs prédictives d'un test

Soit un test (Positif/Négatif) recherchant une caractéristique (Présent/Absent).

- La sensibilité est la probabilité que le test soit positif si la caractéristique est présente.
  Se = P(Positif */* Présent *)*

Un test est d'autant plus sensible qu'il y a peu de faux négatifs.

- La spécificité est la probabilité que le test soit négatif si la caractéristique est absente.
  Sp = P(Négatif */* Absent *)*

Un test est d'autant plus sensible qu'il y a peu de faux positifs.

- La valeur prédictive positive est la probabilité que la caractéristique soit présente si le test est positif
  VPP = P(Présent / Positif *)*

La valeur prédictive positive est plus grande s'il y a peu de faux positifs.

- La valeur prédictive négative est la probabilité que la caractéristique soit absente si le test est négatif
  VPN = P(Absent / Négatif *)*

La valeur prédictive négative est plus grande s'il y a peu de faux négatifs.

Résumé :

Caractéristique / test	Présent	Absent	Total
Positif	a vrai positif	b faux positif	a+b positifs
Négatif	c faux négatif	d vrai négatif	c+d négatifs
Total	a+c présent	b+d absent	N=a+b+c+d

$$Se = P(\text{ Positif / Présent }) = \frac{a}{a+c}$$

$$Sp = P(\text{ Négatif / Absent }) = \frac{d}{b+d}$$

$$VPP = P(\text{ Présent / Positif }) = \frac{a}{a+b}$$

$$VPN = P(\text{ Absent / Négatif }) = \frac{d}{c+d}$$

3 Analyse combinatoire, notion de dénombrement

Pour une grande partie du calcul des probabilités discrètes on cherche à calculer le nombre d'évènement réalisables sur une ensemble important. Dans ce paragraphe, E désigne un ensemble à n éléments que l'on suppose distincts.

3.1 Permutations

On appelle permutation des n éléments toute disposition ordonnée de ces n éléments.
Le nombre de permutations d'un ensemble E de n éléments est égal à : n!
Exemple: le nombre de permutations de l'ensemble {1;2;3} est 3!=6: (1;2;3), (1;3;2), (2;1;3), (2;3;1), (3;1;2), (3;2;1)

3.2 Arrangement sans répétition

On appelle arrangement sans répétition de p éléments pris parmi les n éléments d'un ensemble E, toute disposition ordonnée de p éléments de E.
Le nombre d'arrangements sans répétition de p éléments pris dans un ensemble à n éléments est égal à:

$$A_n^p = \frac{n!}{(n-p)!} = n(n-1)...(n-p+1)$$

Exemple: parmi l'ensemble (urne) {1;2;3} on tire 2 éléments l'un après l'autre (on s'intéresse à l'ordre) sans les remettre dans l'urne. Les arrangements sans répétition sont au nombre de

$$A_3^2 = \frac{3!}{(3-2)!} = \frac{3x2}{1} = 6$$

6 possibilités: (1;2), (2;1), (1;3), (3;1), (2;3), (3;2)

3.3 Combinaisons sans répétition

On appelle combinaisons sans répétition de p éléments pris parmi les n éléments d'un ensemble E, toute disposition non ordonnée de p éléments de E.
Le nombre d'arrangements sans répétition de p éléments pris dans un ensemble à n éléments est égal à:

$$C_n^p = \frac{n!}{p!(n-p)!} = \frac{n(n-1)...(n-p+1)}{p!}$$

Exemple: parmi l'ensemble (urne) {1;2;3} on tire simultanément 2 éléments (on ne s'intéresse pas à l'ordre) sans les remettre dans l'urne. Les combinaisons sans répétition de 2 éléments de l'ensemble {1;2;3} sont : (1;2), (2;3); (1;3); $C_3^2 = \frac{3!}{2!(3-2)!} = \frac{3x2}{2x1} = 3$

Propriété: $C_n^p = \frac{n!}{p!(n-p)!} = C_n^{n-p} = \frac{n!}{(n-p)!(n-n+p)!}$

Notation: dans les livres récents C_n^p est noté $\binom{n}{p}$

3.4 Arrangement avec répétition

On appelle arrangement avec répétition de p éléments pris parmi les n éléments d'un ensemble E, toute disposition ordonnée de p éléments, non nécessairement distincts, de E.
Le nombre d'arrangements avec répétition de p éléments pris dans un ensemble à n éléments est n^p.

Exemple: parmi l'ensemble (urne) {1;2;3} on tire 2 éléments l'un après l'autre (on s'intéresse à l'ordre) **avec remise dans l'urne**. Les arrangements avec répétition sont au nombre de 3^2.
9 possibilités: (1;1), (1;2), (1;3), (2;1), (2;2), (2;3), (3;1), (3;2), (3;3)

3.5 Combinaisons avec répétition

On appelle combinaison avec répétition de p éléments pris parmi les n éléments d'un ensemble E, toute disposition non ordonnée de p éléments, non nécessairement distincts, de E.
Le nombre de combinaison avec répétition de p éléments pris dans un ensemble à n éléments est égal à:

$$C_{n+p-1}^p = \frac{(n+p-1)!}{p!\,(n-1)!} = \frac{(n+p-1)...n}{p!}$$

Exemple: parmi l'ensemble (urne) {1;2;3} on tire 2 éléments (on ne s'intéresse pas à l'ordre) **avec remise dans l'urne**. Les combinaisons avec répétition de 2 éléments de l'ensemble {1;2;3} sont :

(1;1), (1;2), (1;3); (2;2); (2;3), (3;3); $C_{3+2-1}^2 = C_4^2 = \frac{4!}{2!(4-2)!} = \frac{4x3x2}{2x2} = 6$

3.6 Résumé du nombre de tirages

On considère une urne de n boules distinguables où on effectue p tirages successifs. Le nombre de tirages possibles est:

p tirages parmi n objets	Sans remise	Avec remise
Avec ordre, objets discernables	A_n^p	n^p
Sans ordre ou objets indiscernables (combinaison)	C_n^p	C_{n+p-1}^p

La notion de remise est claire. La notion d'ordre revient à différencier les combinaisons en tenant compte de l'ordre dans lequel elles ont été tirées. Sans notion d'ordre, cela revient à considérer des boules non distinguables

3.7 Placement de p objets dans n cases

Il y a quatre manières de placer p objets dans n cases.

p objets dans n cases	un objet par case	sans limitation du nombre d'objets par case
Objets discernables (ordre)	A_n^p	n^p
Objets non discernables (sans ordre)	C_n^p	C_{n+p-1}^p

Dans le cas d'objets discernables, si le nombre d'objets par case n'est pas limité chaque objet a le choix entre les n cases. Si on se limite à un objet par case, le deuxième objet a le choix entre n-1 cases ... .
Dans le cas d'objets indiscernables, il faut enlever les placements qu'on ne peut différencier.

4 Variable aléatoire réelle (v.a.r.)

4.1 Définition d'une variable aléatoire réelle X

Soit Ω l'ensemble des résultats possibles d'une expérience aléatoire. Soit P(A) la probabilité de chacun des évènements $A \in \Omega$. Soit A un ensemble d'évènements de Ω. (Ω, A, P) s'appelle un espace probabilisé.
Une variable aléatoire réelle (v.a.r.) sur l'espace probabilisé (Ω, A, P), est une fonction de Ω dans $\Re$.

Remarque : en pratique, une variable aléatoire est une grandeur mesurable ou observable par une expérience et dont le résultat ne peut être prévu, comportant donc une partie de "hasard". Chaque mesure est la réalisation de la v.a.r.

Remarque : à toute v.a.r. on associe une loi statistique, connue théoriquement. Elle va permettre non pas de prévoir le résultat d'une nouvelle réalisation mais les caractéristiques générales (position, dispersion) de la v.a.r. et ainsi de calculer un intervalle de confiance pour une nouvelle réalisation.

Exemple: On lance 2 fois un dé. Ω est l'ensemble des couples (i,j) avec i et j éléments de (1;2;3;4;5;6). On s'intéresse à la valeur X de la somme des deux tirages.

Ω possède 36 éléments.

Soit A l'ensemble des évènements "le total X fait 9" $A = \{(3;6),(4;5),(5;4),(6,3)\}$

On a $P(A) = P(X = 9) = \frac{card(A)}{card(\Omega)} = \frac{4}{36} = \frac{1}{9}$

Dans cet exemple on appelle X la v.a.r. définie comme la somme des deux tirages. Pour chaque couple $(i,j)\in\Omega$ on a $X\big((i,j)\big)=i+j$

On note : X= "somme des deux tirages indépendants".

Une v.a.r. est donc une fonction qui à chaque évènement élémentaire d'une expérience aléatoire associe un nombre réel.

Elle est dite **discrète** si elle ne prend que **certaines valeurs** d'un intervalle de $\Re$.

Elle est **continue** si elle est susceptible de prendre toutes les valeurs d'un intervalle de $\Re$.

4.1.1 Loi de probabilité d'une variable aléatoire discrète

La loi de probabilité f(k) d'une v.a.r. X est la fonction qui à chaque valeur réelle k, associe la probabilité de l'évènement (X=k). On la note généralement $f(k) = P(X=k)$.

On définit aussi la fonction de répartition F(k) qui à chaque nombre réel k, associe la probabilité de l'évènement $(X\le k)$: $F(k) = P(X\le k)$.

C'est la somme de toutes les probabilités pour les évènements tels que $X\le k$

Exemple: Dans l'exemple précédent la v.a.r. X est discrète et ne peut prendre que des valeurs entières comprises entre 2 et 12. (X=9) constitue un évènement pour la variable X, il correspond à l'ensemble A des évènements élémentaires de Ω.

La probabilité de cet évènement se note P(X=9). On avait P(X=9) =1/9. Pour toutes les valeurs de $k\in[2;12]$ on obtient le tracé de la loi de probabilité P(X=k) en fonction de k :

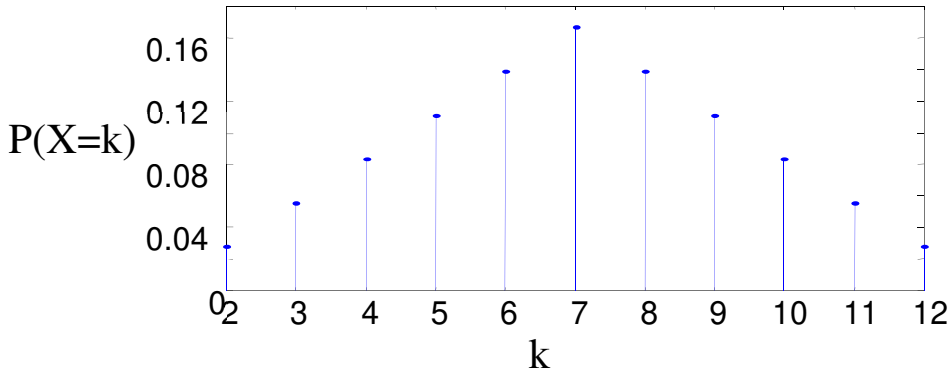


fig 3-3: Loi de probabilité de la variable aléatoire X

On peut aussi tracer la fonction de répartition P(X≤k) en fonction de k.

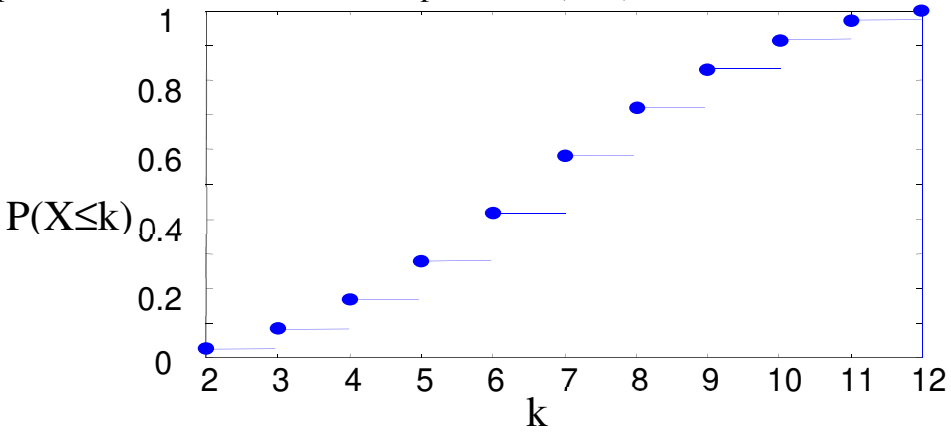


fig 3-4: Fonction de répartition de la variable aléatoire X

4.1.2 Densité de probabilité d'une variable aléatoire continue

Dans le cas d'une v.a.r. continue, la probabilité $P(X=k)$ est nulle car on ne peut avoir égalité parfaite entre deux réels prix au hasard. On ne peut alors définir facilement que la probabilité $P(X\le k)$. Cette probabilité va s'exprimer sous forme de surface en utilisant la fonction **densité de probabilité.**

Soit X une v.a.r. continue. Elle est définie par la fonction f appelée **densité de probabilité** de la v.a.r. X qui vérifie:

- pour tout réel x, f(x) est positif
- $\int_{-\infty}^{+\infty} f(x)\,dx = 1$; la surface comprise entre l'axe des abscisses et la courbe de f(x), est égale à 1.
- pour tout réel t: $P(X \le t) = \int_{-\infty}^{t} f(x)\,dx$

Propriété: Pour tout réel a et b (a<b), on a:

$$P(a < X \le b) = P(X \le b) - P(X \le a) = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_{a}^{b} f(x)\,dx$$

Cette probabilité est l'aire de la surface comprise entre la courbe de f(x), l'axe des abscisses et les droites verticales d'équations respective x=a et x=b.
En conséquence, pour tout réel a, on a $P(X = a) = 0$.

C'est pour cela que la courbe f(x) s'appelle **densité de probabilité**. Elle ne peut s'interpréter qu'en considérant la surface comprise sous la courbe et non pas les valeurs particulières de f(x). Pour un intervalle de largeur dt petit on peut considérer f(x) comme constant et écrire

$$P(t < X \le t + dt) = \int_{t}^{t+dt} f(x)\,dx = f(t)\,dt$$

On définit également la **fonction de répartition d'une variable aléatoire continue** X, la fonction F définie sur $\Re$ par:

$F(t) = P(X \le t)$

Propriété: $F(t) = P(X \le t) = \int_{-\infty}^{t} f(x)\,dx$ ou encore $f(x) = \frac{dF(x)}{dx}$.

La densité de probabilité est la dérivée de la fonction de répartition.

4.2 Propriétés des variables aléatoires

4.2.1 Espérance (ou moyenne) d'une v.a.r.

L'espérance de la v.a.r. X est égale à la moyenne des valeurs prises par X, pondérées par leur probabilité de réalisation.

Ce qui s'écrit pour une variable aléatoire **discrète** : $E(X) = \sum_{i=1}^{n} k_i\, p_i$

Exemple: on reprend l'exemple de la somme du lancé de 2 dés. On a n= 11 valeurs possibles pour la somme. On somme pour ces 11 valeurs le produit $k_i\, p_i$, qui donne par exemple, pour k=9 :

$k_i\, p_{i=}\, 9\frac{1}{9} = 1$.

On obtient: $E(X) = 7$.

Cela était prévisible en regardant la figure 3 qui donne la loi de probabilité de X

Ce qui s'écrit pour une variable aléatoire **continue** : $E(X) = \int_{-\infty}^{+\infty} x\, f(x)\,dx$

Exemple: on considère le tirage aléatoire d'un réel compris dans l'intervalle [0 ;1] avec une densité de probabilité uniforme.

Le calcul de l'espérance donne: $E(X) = \int_{-\infty}^{+\infty} x\, f(x)\,dx = \int_{0}^{1} x\,dx = 1/2$.

Propriété: soit la v.a.r. X et a, b deux constantes réelles, on a: $E(aX + b) = aE(X) + b$

4.2.2 Variable aléatoire centrée

Lorsque l'espérance d'une v.a.r. est nulle, on dit que cette variable est centrée.
La variable aléatoire X-E(X) est toujours centrée. On l'appelle v.a.r. X centrée.

4.2.3 Moments d'une variable aléatoire

Soit X une v.a.r., on appelle moment d'ordre m, l'espérance de la v.a.r. X^m : $E(X^m)$

Ce qui s'écrit pour une variable aléatoire **discrète** : $E(X^m) = \sum_{i=1}^{n} k_i{}^m \, p_i$

Ce qui s'écrit pour une variable aléatoire **continue** : $E(X^m) = \int_{-\infty}^{+\infty} x^m \, f(x)\,dx$

4.2.4 Moments d'ordre 2: variance, écart type

On appelle variance de la v.a.r. X, le moment **d'ordre 2 de la v.a.r. X centrée**. C'est l'espérance de la variable $(X - E(X))^2$: $V(X) = E\left((X - E(X))^2\right)$

On appelle écart type de la v.a.r. X le nombre $\sigma_X = \sqrt{V(X)}$

Ce qui s'écrit pour une variable aléatoire **discrète** : $V(X) = \sum_{i=1}^{n} (k_i - E(X))^2 \, p_i$

Ce qui s'écrit pour une variable aléatoire **continue** : $V(X) = \int_{-\infty}^{+\infty} (x - E(x))^2 \, f(x)\,dx$

Exemple: on reprend l'exemple de la somme du lancer de 2 dés. On a n = 11 valeurs possibles pour la somme. On somme pour ces 11 valeurs le produit $(k_i - 7)p_i$ qui donne par exemple, pour k=9 :

$(k_i - 7)p_i$: $(9-7)^2 \frac{1}{9} = 0{,}444$.

On obtient : $V(X) = 5{,}833$ et $\sigma_X = 2{,}415$

Exemple: on considère le tirage aléatoire d'un réel compris dans l'intervalle [0 ;1] avec une densité de probabilité uniforme.

Le calcul de la variance donne: $V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 \, f(x)\,dx = \int_{0}^{1} (x - 1/2)^2 \, dx = 1/12$.

Ou $V(X) = \int_{-\infty}^{+\infty} x^2 \, f(x)\,dx - E(X)^2 = \int_{0}^{1} x^2 \, dx - (1/2)^2 = 1/12$

Propriété: comme en statistique, on a: $V(X) = E(X^2) - E(X)^2$ où $E(X^2) = \int_{-\infty}^{+\infty} x^2 \, f(x)\,dx$

Propriété: soit la v.a.r. X et a,b deux constantes réelles, on a: $V(aX + b) = a^2 V(X)$

4.2.5 Variable aléatoire centrée réduite

Lorsque la variance d'une v.a.r. est égale à 1 on dit que cette variable est réduite.
Lorsque l'espérance d'une v.a.r. est nulle et que sa variance est égale à 1 on dit que cette variable est centrée réduite.

La variable aléatoire $\frac{X - E(X)}{\sigma_X}$ est **centrée réduite**.

4.2.6 Indépendance de deux variables aléatoires

Deux v.a.r. X et Y sont indépendantes si, et seulement si, pour toute valeur a prise par X et toute valeur b prise par Y, les évènements (X≤a) et (Y≤b) sont indépendants.
Cela signifie que la probabilité de (Y≤b) ne dépend pas du fait que X≤a, et inversement, que la probabilité de (X≤a) ne dépend pas du fait que Y≤b. Les deux évènements (X≤a) et (Y≤b) ne possèdent aucun lien de causalité.

4.2.7 Covariance de deux variables aléatoires, coefficient de corrélation

Comme dans le cas discret on définit la covariance de deux v.a.r., X et Y par le nombre:

$$COV(X,Y) = E\big((X - E(X))(Y - E(Y))\big)$$

C'est l'espérance du produit des v.a.r. centrées.
Le coefficient de corrélation vaut:

$$\rho = \frac{COV(X,Y)}{\sqrt{V(X)\,V(Y)}} = \frac{COV(X,Y)}{\sigma_X\,\sigma_Y}$$

Ce qui s'écrit pour une variable aléatoire **discrète**, X et Y deux v.a. prenant n et m valeurs différentes :

$$COV(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}\big(x_i - E(X)\big)\big(y_j - E(Y)\big)P(X = x_i \text{ et } Y = y_j)$$

Remarque : $COV(X,X) = E\big((X - E(X))(X - E(X))\big) = E\big((X - E(X))^2\big) = V(X)$

4.2.8 Somme de deux variables aléatoires

Soit deux v.a.r. X et Y quelconques:

$$E(X+Y) = E(X) + E(Y)$$

$$V(X+Y) = V(X) + V(Y) + 2COV(X,Y)$$

Si X et Y sont deux v.a.r. **indépendantes**:

$$V(X+Y) = V(X) + V(Y)$$

Exemple: on reprend l'exemple de la somme du lancé de 2 dés. Si on considère la variable X : "résultat du lancement de 1 dé". L'espérance et la variance de X sont : E(X) = 3,5 et V(X) = 2,917.
La v.a.r. Y : "somme du lancement de 2 dés" est la somme du résultat d'un premier lancement correspondant à la v.a.r. X_1 et d'un second lancement correspondant à la v.a.r. X_2. Les deux v.a.r. ont les mêmes paramètres. Les deux tirages sont **indépendants**. On a donc $Y = X_1 + X_2$.
On vérifie que :

$E(Y) = E(X_1) + E(X_2) = 3,5 + 3,5 = 7$
$V(Y) = V(X_1) + V(X_2) = 2,917 + 2,917 = 5,833$

Remarque: différence entre la variable Y=2X et la variable $Z=X_1+X_2$.
La variable 2X correspond à **un seul tirage aléatoire** mais dont le résultat est multiplié par 2.
La variable X_1+X_2 correspond à la somme de **deux tirages aléatoires.**
Si les tirages correspondent à la même expérience aléatoire réalisée deux fois de manière indépendante: $V(X_1 + X_2) = V(X_1) + V(X_2) = 2\,V(X)$ où V(X) est la variance de un tirage
alors que $V(2X) = 2^2 V(X) = 4\,V(X)$
Doubler la valeur d'un tirage revient à prendre un second tirage de même valeur, c'est à dire non indépendant du premier et dont la covariance $COV(X_1,X_1)=V(X_1)=V(X)$.
On obtient donc:

$$V(2X) = V(X_1 + X_1) = V(X_1) + V(X_1) + 2COV(X_1, X_1) = V(X) + V(X) + 2V(X) = 4V(X)$$

Exemple: au lieu de lancer 2 dés, on ne lance qu'un seul dé et on double le résultat du lancement.
lancement de 1 dé". L'espérance et la variance de X sont : E(X) = 3,5 et V(X) = 2,917.

La v.a.r. Z : "double du lancement de 1 dés" s'écrit Z = 2X. On vérifie que :

$E(Z) = E(2X) = 2 \times 3{,}5 = 7$

$V(Z) = V(2X) = 2^2\, V(X) = 4 \times 2{,}917 = 11{,}666$

On remarque que $V(X_1) + V(X_2) \neq V(2X)$

4.2.9 Somme de n variables aléatoires

Soit X_i, n v.a.r. quelconques:

$$E(X_1 + ... + X_n) = E(X_1) + ... + E(X_n)$$

Soit X_i, n v.a.r. indépendantes de même écart type σ

$$V(X_1 + ... + X_n) = V(X_1) + ... + V(X_n) = n\,\sigma^2$$

$$\sigma_{X_1+...+X_n} = \sigma\sqrt{n}$$

Remarque: soit X une v.a.r. d'espérance m est d'écart type σ

la variable aléatoire nX a pour espérance n m et pour variance $n^2\sigma^2$, donc pour écart type $n\sigma$

la variable aléatoire $\sum_{i=1}^{n} X_i$ a aussi pour espérance n m, mais pour variance $n\,\sigma^2$, donc pour écart type $\sigma\sqrt{n}$

4.3 Loi associée à une variable discrète

4.3.1 Loi de Bernoulli

Une v.a.r. X suit une loi de Bernoulli de paramètre p si elle ne prend que deux valeurs 0 ou 1 avec la probabilité P(X=1)=p avec $p \in [0;1]$ (et P(X=0)=1-p).

Cette loi modélise l'issue d'une expérience en ne s'intéressant qu'au "succès" ou à "l'échec" de l'expérience.

Propriété: E(X)=0.(1-p)+1.p=p

$E(X^2)=0^2.(1-p)+1^2\ p=p$; donc $V(X)=E(X^2)-(E(X))^2=p-p^2=p(1-p)$

4.3.2 Loi binomiale

Une v.a.r. X suit une loi binomiale de paramètre (n,p), si elle est la **somme** de n variables de Bernoulli **indépendantes** de paramètres p. Elle est à valeur dans $\{0;1;...;n\}$ avec pour tout k=0,...,n :

$$P(X = k) = C_n^k p^k (1-p)^{n-k}.$$

Cette loi modélise une succession de "succès" ou "échec", p étant la probabilité du "succès" de chaque expérience.

Notation: on dit que X suit une loi $\mathcal{B}(n;p)$ ou $X \sim \mathcal{B}(n;p)$. Une loi de Bernoulli peut s'écrire comme une loi $\mathcal{B}(1;p)$.

Exemple: La probabilité qu'une personne parle anglais est p=1/5. La v.a.r. Y indiquant si une personne parle anglais suit une loi de Bernoulli de paramètre p. On considère la v.a.r. X qui associe le nombre de personnes parlant anglais dans un groupe de n=3 personnes. La variable X suit une loi binomiale de paramètre (n,p) et:

$$P(X = 0) = C_3^0 (1/5)^0 (1-1/5)^{3-0} = 1x1x(4/5)^3 = 64/125$$

$$P(X = 1) = C_3^1 (1/5)^1 (1-1/5)^{3-1} = 3(1/5)^1 (4/5)^2 = 48/125$$

$$P(X = 2) = C_3^2 (1/5)^2 (1-1/5)^{3-2} = 3(1/5)^2 (4/5)^1 = 12/125$$

$$P(X = 3) = C_3^3 (1/5)^3 (1-1/5)^{3-3} = 1x(1/5)^3 x1 = 1/125$$

on vérifie que P(X=0) + P(X=1) + P(X=2) + P(X=3) = (64+48+12+1)/125=1

Exemple: avec le logiciel Excel ou OpenOffice_Calc, $P(X=2)=12/125=0{,}096$ est obtenu par la commande
LOI.BINOMIALE(2;n;p;0) qui donne la loi de probabilité
LOI.BINOMIALE(2;3;1/5;0) = 0,096
LOI.BINOMIALE(2;n;p;1) fournit la fonction de répartition

Propriété: si X suit une loi $\mathcal{B}$(n;p).
E(X)=n p; (somme de l'espérance des n variables de Bernoulli)
V(X)=n p(1-p); (somme de la variance des n variables de Bernoulli indépendantes)
Dans l'exemple précédent, sur un groupe de 3 personnes, il y en a en moyenne 3/5 qui parle anglais

Utilité pratique: Pour appliquer un modèle binomiale $\mathcal{B}$(n;p).en pratique, il faut remplir les conditions suivantes:

- répéter une même épreuve n fois de suite
- prendre comme variable le nombre de "succès" ou "d'échec"
- avoir une probabilité de "succès" p, identique pour toutes les épreuves, autrement dit que les épreuves sont indépendantes.

Cette dernière condition est réalisée lorsqu'on réalise une succession de tirage **avec remise**.
Dans la pratique, si les tirages s'effectuent sans remise on peut utiliser un modèle binomiale à condition que la taille de prélèvement soit petite devant la taille de la population totale (<1/10).

4.3.3 Loi géométrique

On considère une épreuve de Bernoulli dont la probabilité de succès est p. On renouvelle cette épreuve de manière indépendante jusqu'au premier succès. On appelle X la variable aléatoire donnant le rang du premier succès. Les valeurs de X sont les entiers naturels non nuls. La probabilité que X = k est alors donné par la loi géométrique.
Une v.a.r. X, à valeur dans $\mathbb{N}^*$(entier strictement positif), suit une loi géométrique de paramètre p, si elle s'écrit:

$P(X=k)=p(1-p)^{k-1}$; on a (k-1) échecs suivi de un succès

alors $P(X>k)=(1-p)^k$

Propriété: E(X)=1/p
$V(X)=(1-p)/p^2$

Cette loi modélise le temps d'attente (qui est aussi le rang) du premier succès dans des épreuves de Bernoulli répétées ou dans un tirage avec remise. On l'appelle parfois loi exponentielle discrète.

Exemple : on lance un dé à 6 faces, soit X le nombre d'essai pour obtenir un "6", il suit une loi géométrique de paramètre p=1/6.
La probabilité d'avoir un "6" avant le 3eme tirage est : $P(X\le 3)=1-P(X>3)=1-(1-p)^3$.

Exemple : un serveur reçoit une requête en moyenne toute les 7 s. Soit X le temps d'attente d'une nouvelle requête exprimé en seconde.
Hypothèse : le temps est mesuré en seconde, et toute seconde commencée est comptée comme entière.
P(X=1) est la probabilité d'avoir une requête pendant la première seconde, P(X=1)=1/7.
A chaque seconde on a une probabilité de 1/7 d'avoir une requête.
Le temps d'attente d'une requête suit une loi géométrique de moyenne 7=1/p, soit de paramètre p=1/7.
La probabilité de recevoir une requête après plus de 7 s est : $P(X>7)=(1-p)^7$.

4.3.4 Loi de Pascal (binomiale négative)

La loi de Pascal (binomiale négative) est la loi de probabilité de la variable aléatoire X donnant le nombre d'essais nécessaires pour obtenir n succès d'une variable de Bernoulli.
Par exemple, on effectue des tirages **avec remise** dans une urne contenant une proportion p de boules gagnantes et (1-p) de boules perdantes (avec q = 1 - p). On appelle X le nombre de tirages nécessaires pour obtenir r boules gagnantes. La probabilité que X = k est alors donné par la loi de Pascal.

Une v.a.r. X, à valeur dans {r, r+1, … }, suit une loi de Pascal de paramètre r,p, si elle s'écrit:

$$P(X = k) = C_{k-1}^{r-1} p^r (1-p)^{k-r}$$

Propriété: E(X)=r/p
V(X)=r(1-p)/p^2

Si r = 1, on retrouve la loi géométrique donnant le nombre de tirage jusqu'à obtention de la première boule gagnante

Exemple : on lance un dé à 6 faces, soit X le nombre d'essai pour obtenir deux "6", il suit une loi de Pascal de paramètre r=2, p=1/6.
La probabilité d'avoir deux "6" avec 3 lancés est : $P(X = 3) = C_{3-1}^{2-1} (1/6)^2 (5/6)^{3-2}$.

Exemple : un serveur reçoit une requête en moyenne toute les 7 s. Soit X le temps d'attente pour recevoir 3 nouvelles requêtes.
Hypothèse : le temps est mesuré en seconde, et toute seconde commencée est comptée comme entière.
Le temps d'attente de 3 requêtes suit une loi géométrique de Pascal de paramètre r=3, p=1/7.
La probabilité de recevoir 3 requêtes en 15 s est : $P(X = 15) = C_{15-1}^{3-1} (1/7)^3 (6/7)^{15-3}$.
La probabilité de recevoir 3 requêtes en moins de 15 s est : $P(X \le 15) = \sum_{i=3}^{15} C_{i-1}^{3-1} (1/7)^3 (6/7)^{i-3}$.

4.3.5 Loi hypergéométrique

On tire simultanément (**sans remise**) n boules dans une urne contenant N boules, pN boules gagnantes et (1-p)N boules perdantes (avec q = 1 - p). On appelle X la variable aléatoire donnant le nombre de boules gagnantes extraites. X prend des valeurs entières de 0 à n. X suit une loi hypergéométrique.

Une v.a.r. X, suit une loi hypergéométrique de paramètre n, p, N avec n entier n (≤N). Elle est à valeurs dans l'ensemble {0,1,2, ...,n}, et elle s'écrit:

$$P(X = k) = \frac{C_{pN}^{k} C_{(1-p)N}^{n-k}}{C_N^n}$$

Propriété: E(X)=np
V(X)=np(1-p)(N-n)/(N-1)

Cette loi modélise la situation d'une population de N individus dont pN présentent un caractère donné (ou celui d'une urne de N boules dont pN sont d'une couleur donnée). La loi hypergéométrique est la loi de la variable aléatoire qui compte les observations du caractère donné (ou des boules de la couleur donnée) dans un tirage sans remise de n individus; on cherche la probabilités d'observer k fois ce caractère.
Lorsque N est grand, on approxime la loi hypergéométrique par la loi binomiale $\mathcal{B}$(n;p).

4.3.6 Loi de Poisson

Une loi binomiale dont le nombre de tirages est très grand et la probabilité du "succès" (ou de "l'échec") est très faible peut être approximée par une loi plus simple, la loi de Poisson de même paramètres (espérance np et variance np(1-p)).

Une v.a.r. X, à valeur dans $\mathbb{N}$ (entier positif), suit une loi de Poisson de paramètre λ ($\lambda>0$), si sa loi de probabilité vérifie pour tout $k \in \mathbb{N}$:

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!}.$$

Cette loi modélise le nombre X de réalisations d'un évènement pendant des intervalles de temps (ou d'espace ...) ayant tous la même valeur.

Notation: on dit que X suit une loi $\mathcal{P}(\lambda)$ ou $X \sim \mathcal{P}(\lambda)$.

Exemple 1: On considère la désintégration nucléaire d'un gramme d'uranium radioactif (contenant donc un très grand nombre d'atomes). On observe en moyenne 45 désintégrations par seconde. Il se passe en moyenne 1/45 s entre chaque désintégration. Le nombre de désintégrations pendant un intervalle de temps donné dépend de la durée de cet intervalle mais pas de l'instant où il se situe. La variable X qui associe le nombre de désintégrations par seconde suit une loi de Poisson de paramètre 45.

$$P(X=1)=\frac{e^{-45}\,45^{1}}{1!}=1{,}2\ 10^{-18}\,;\ P(X=45)=\frac{e^{-45}\,45^{45}}{45!}=0{,}0594\,;$$

$$P(X=46)=\frac{e^{-45}\,45^{46}}{46!}=0{,}0581\,;\ P(X=100)=\frac{e^{-45}\,45^{100}}{100!}=6{,}4\ 10^{-13}$$

Exemple 2: Une machine produisant des résistances fournit 2 résistances défectueuses parmi les 10 000 qu'elle produit chaque minute. Ce taux moyen d'erreur est constant tout au long de la journée de production. La variable X qui associe le nombre de résistances défectueuses par minute suit une loi de Poisson de paramètre 2.

$$P(X=0)=\frac{e^{-2}\,2^{0}}{0!}=0{,}1353\,;\ P(X=1)=\frac{e^{-2}\,2^{1}}{1!}=0{,}2707\,;\ P(X=2)=\frac{e^{-2}\,2^{2}}{2!}=0{,}2707\,;$$

$$P(X=3)=\frac{e^{-2}\,2^{3}}{3!}=0{,}1804$$

Exemple 3: Les fautes d'impression d'un quotidien ont été analysées. On en compte, en moyenne, une tous les 800 mots. Ce taux ne dépend pas de la page considérée dans le journal. La variable aléatoire qui associe le nombre de fautes par page (2 500 mots) suit une loi de Poisson de paramètre 2500/800=3,125.

$$P(X=0)=\frac{e^{-3{,}125}\,3{,}125^{0}}{0!}=0{,}0439\,;\ P(X=1)=\frac{e^{-3{,}125}\,3{,}125^{1}}{1!}=0{,}1373\,;$$

$$P(X=2)=\frac{e^{-3{,}125}\,3{,}125^{2}}{2!}=0{,}2145\,;\ P(X=3)=\frac{e^{-3{,}125}\,3{,}125^{3}}{3!}=0{,}2235\,;$$

$$P(X=4)=\frac{e^{-3{,}125}\,3{,}125^{4}}{4!}=0{,}1746$$

Remarque : la loi suivie par X est en fait une loi binomiale $\mathcal{B}(2500;1/800)$, car on imprime 2500 mots avec chacun la probabilité 1/800 de comporter une erreur. Cette loi binomiale est approximée par une loi de Poisson de même valeur moyenne $\mathcal{P}(\lambda)$ avec $\lambda=np$, comme cela est indiqué un peu plus loin si $n>50$ et $p\leq 0{,}1$ et $np<17$ (dans notre cas $np=3{,}125$).

Exemple: avec le logiciel Excel ou OpenOffice_Calc, $P(X=2)$ est obtenu par la commande

LOI.POISSON(2;λ;0) qui donne la loi de probabilité
LOI.POISSON (2;3,125;0) = 0,21454
LOI.POISSON (2; λ;1) fournit la fonction de répartition

Allure: La figure suivante donne l'allure de la loi de probabilité de la loi de Poisson pour différentes valeurs du paramètre λ.

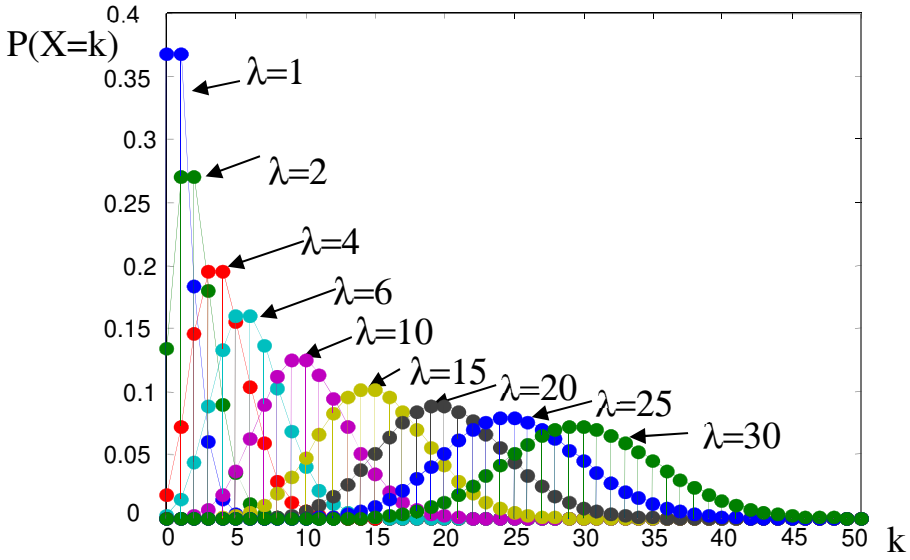


fig 3-5: loi de probabilité de la loi de Poisson pour différentes valeur du paramètre λ.

Propriété: si X suit une loi de Poisson $\mathcal{P}(\lambda)$.

$E(X)=\lambda$; $V(X)=\lambda$; $\sigma_X=\sqrt{\lambda}$

Utilité pratique: Pour appliquer une modèle de Poisson $\mathcal{P}(\lambda)$ en pratique, il faut remplir les conditions suivantes:

- la proportionnalité entre le nombre d'évènements et la longueur de la plage de mesure (temps, espace, ...) est stationnaire. Le présent ne dépend pas du passé, il n'y a pas de vieillissement.
- la probabilité d'avoir un évènement est identique pour tous les évènements, indépendamment des autres évènements, autrement dit les évènements sont indépendants.
- l'isolement des différents évènements fait apparaître la loi de Poisson comme la loi de petites probabilités.

Approximation d'une loi binomiale par une loi de Poisson: On montre que la loi de Poisson est une approximation de la loi binomiale dans le cas où la probabilité du "succès" (ou de "l'échec") est très faible et le nombre de tirages important. Le paramètre λ qui définit la loi de Poisson est aussi son espérance, cette espérance vaut np pour une loi binomiale. Si elle est constante et que n est très grand, cela signifie que la probabilité p est très faible. Voilà pourquoi la loi de Poisson s'appelle la **loi des évènements rares**. En pratique si **n>50 et p≤0,1 et np<17**, alors la loi binomiale $\mathcal{B}(n;p)$ est très proche de la loi de Poisson $\mathcal{P}(\lambda)$ **avec λ=np**. La loi de Poisson ne comporte qu'un seul paramètre et se trouve donc d'utilisation plus simple que la loi binomiale qui en comporte deux.

4.4 Loi associée à une variable continu

4.4.1 Loi normale

4.4.1.1 Définition de la loi normale (ou Gaussienne)

Une v.a.r. continue X d'espérance μ et de variance σ^2, suit une loi normale de paramètre μ et σ^2, si sa densité de probabilité f(x) est définie sur $\Re$ par.

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

Notation: on dit que X suit une loi $\mathcal{N}(\mu,\sigma^2)$ ou $X \sim \mathcal{N}(\mu,\sigma^2)$. La loi normale est aussi appelée loi de Gauss ou loi de Laplace-Gauss. Attention, suivant les auteurs, on note parfois $X \sim \mathcal{N}(\mu,\sigma)$.

Courbe: La figure suivante donne l'allure de la loi normale $\mathcal{N}(0,1)$, généralement appelée Z, qui est aussi appelée la loi gaussienne centrée réduite. Cette courbe est appelée courbe de Gauss.

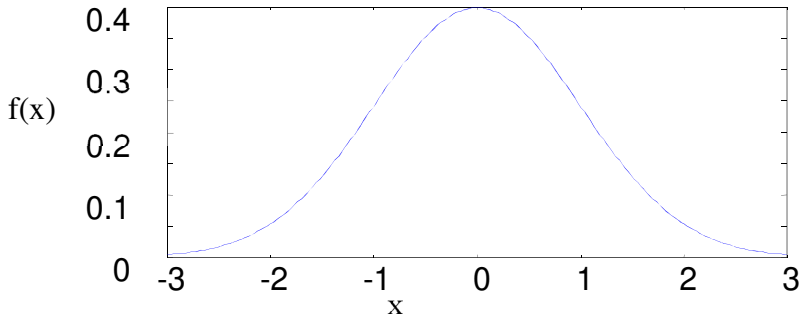


fig 3-6: densité de probabilité de la loi normale $Z \sim \mathcal{N}(0,1)$.ou courbe de Gauss.

Propriétés:

- Soit X une v.a.r. qui suit une loi normale $\mathcal{N}(\mu,\sigma^2)$, alors l'espérance de X vaut μ et la variance de X vaut σ^2.
- Soit X une v.a.r. qui suit une loi normale $\mathcal{N}(\mu,\sigma^2)$ et a,b deux constantes réelles, la v.a.r. définie par aX+b suit une loi normale $\mathcal{N}(a\mu+b,(a\sigma)^2)$.
- Soit X une v.a.r. qui suit une loi normale $\mathcal{N}(\mu,\sigma^2)$, alors la loi définie par $Z = \frac{X-\mu}{\sigma}$ suit la loi centrée réduite $\mathcal{N}(0,1)$. On note conventionnellement Z une variable qui une loi $\mathcal{N}(0,1)$.
- Soit X et Y deux v.a.r. **indépendantes** qui suivent une loi normale $\mathcal{N}(\mu_X, {\sigma_X}^2)$ et $\mathcal{N}(\mu_Y, {\sigma_Y}^2)$. La loi définie par X+Y suit une loi normale $\mathcal{N}(\mu_X + \mu_Y, {\sigma_X}^2 + {\sigma_Y}^2)$.

4.4.1.2 Fonction de répartition de la loi normale

La fonction gaussienne f(x) n'est **pas intégrable**. La probabilité $P(Z \le t) = \int_{-\infty}^{t} f(x)\,dx$ ne peut pas se calculer analytiquement. Seules les valeurs **numériques** de la fonction de répartition de la loi **normale centrée réduites** sont accessibles soit dans des tables, soit par une machine à calculer.

Courbe: La figure suivante donne l'allure de la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.

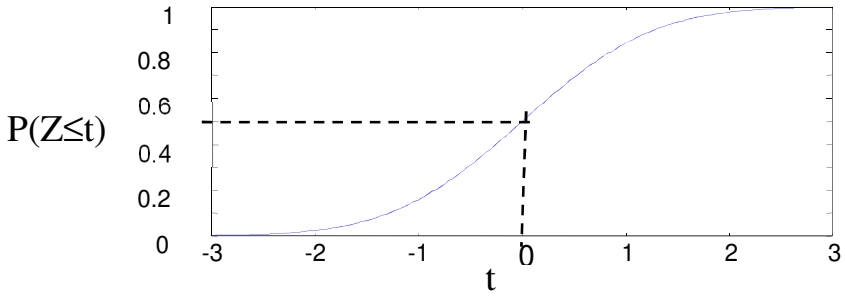


fig 3-7: fonction de répartition de la loi normale $\mathcal{N}(0,1)$.

Les tables de la loi Normale: On trouve conventionnellement dans les tables la fonction de répartition d'une variable T qui suit une loi normale centrée réduite $\mathcal{N}(0,1)$. La probabilité $P(Z \le t)$ est parfois appelée $\Pi(t)$. Les valeurs de $P(Z \le t)$ sont données en fonction de t pour $t \ge 0$. Elles varient donc dans l'intervalle [0,5;1].

La valeur $P(Z \le t)$ pour $t<0$ est obtenue en prenant le complément de la valeur du tableau pour la valeur positive (-t): $P(Z \le t) = P(Z > (-t))$

$P(Z \le t) = 1$ - (la valeur du tableau pour (-t)).

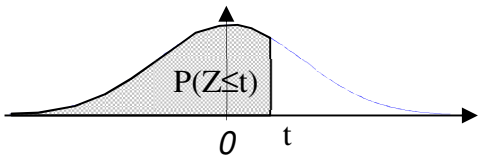


fig 3-8: valeurs données dans la table de la loi normale $\mathcal{N}(0,1)$.

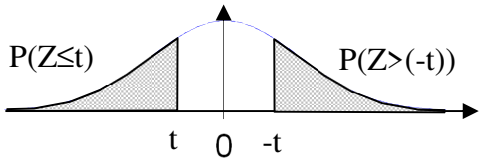


fig 3-9: illustration de l'utilisation de la table pour les valeurs t<0

Quand on désire la loi inverse, c'est à dire trouver la valeur t_p tel que $P(Z \le t_p)$ égale une valeur donnée, on utilise une seconde table donnant, t_p pour $p = P(Z \le t_p)$ connue. Sur cette table les valeurs de p portées à gauche et en haut sont comprises entre 0 et 0,5. Pour ces valeurs, il convient de prendre les valeurs du tableau $P(Z \le t)$ **mais avec un signe négatif**. Les valeurs de p portées à droite et en bas sont elles comprises entre 0,5 et 1. Les valeurs du tableau sont à prendre telles quelles, positives.

Exemple d'utilisation de la table inverse:

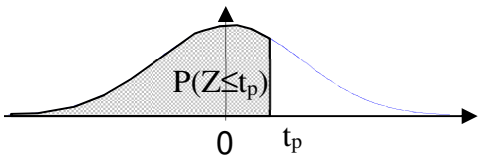
Détermination de t_p pour $p=P(Z \le t_p)$ connue

Je veux calculer la valeur t_p telle que $P(Z \le t_p) \le 0,122$

La valeur 0,122 est inférieure à 0,5, j'utilise les valeurs de p portées à gauche et en haut de la table. Au croisement de la ligne 0,12 et de la colonne 0,002 je lis la valeur t_p=1,165. Comme la valeur 0,122 est inférieur à 0,5, je sais que t_p est négatif et j'obtient : $P(Z \le -1,165) \le 0,122$.

Je veux calculer la valeur t_p telle que $P(Z \le t_p) \le 0,878$

La valeur 0,878 est supérieure à 0,5, j'utilise les valeurs de p portées à droite et en bas de la table. Au croisement de la ligne 0,87 et de la colonne 0,008 je lis la valeur t_p=1,165. Comme la valeur 0,878 est supérieure à 0,5, je sais que t_p est positif et j'obtient : $P(Z \le 1,165) \le 0,878$.



P<0,5		0,002		
0,12		***1,165***		0,87
		0,008		$p \ge 0,5$

fig 3-10: Exemple d'utilisation de la table inverse.

Exemple: Pour calculer la probabilité que X qui suit une loi normale $\mathcal{N}(1,4)$, ait une valeur comprise entre 0,5 et 2,5, on fait:

$P(0,5 < X \le 2,5) = P(X \le 2,5) - P(X \le 0,5))$

on se ramène à la table de la loi Z normale centrée réduite $\mathcal{N}(0,1)$:

$$P(X \le 2,5) = P(Z \le \frac{2,5-1}{2}) = P(Z \le 0,75) \text{ et } P(X \le 0,5) = P(Z \le \frac{0,5-1}{2}) = P(Z \le -0,25)$$

$P(Z \le 0,75) = 0,7734$ se lit directement dans la table

mais $P(Z \le -0,25)$ ne s'y trouve pas car la table ne donne que $P(Z \le t)$ pour $t \ge 0$. Il faut prendre le complément de la valeur du tableau pour la valeur positive (-t):

$P(Z \le -0,5) = P(Z > 0,5) = 1 - P(Z \le 0,5) = 1 - 0,59871 = 0,40129$

$P(0,5 < X \le 2,5) = 0,77337 - 0,40129 = 0,37208$

Avec Excel ou OpenOffice Calc: on dispose de la fonction LOI.NORMALE.N(x; μ;σ;0) qui fournit la densité de probabilité de la loi normale $\mathcal{N}(\mu,\sigma^2)$ pour la valeur x, et, ce qui est plus utile, la fonction LOI.NORMALE.N(t; μ;σ;1) qui fournit la **fonction de répartition** de la loi normale $\mathcal{N}(\mu,\sigma^2)$ pour la valeur t.

$P(X \le 2,5)$ =LOI.NORMALE.N(2,5;1;2;1); $P(X \le 0,5))$ = LOI.NORMALE.N(0,5;1;2;1)

La fonction LOI.NORMAL.STANDART.N donne directement la densité de probabilité de la loi normale $\mathcal{N}(0,1)$

$$P(X \le 2,5) = \text{LOI.NORMALE.STANDART.N}(\frac{2,5-1}{2};1)$$

Pour la loi inverse, la fonction =LOI.NORMALE.INVERSE.N

LOI.NORMALE.INVERSE.N(α;μ;σ) donne la valeur t_α telle que $P(X \le t_\alpha) = \alpha$ pour une loi $\mathcal{N}(\mu,\sigma^2)$. Il existe aussi la fonction LOI.NORMALE.STANDARD.INVERSE.N(α) qui donne la valeur t_α telle que $P(Z \le t_\alpha) = \alpha$ pour une loi $\mathcal{N}(0,1)$.

LOI.NORMALE.STANDARD.INVERSE.N(0,975) = 1,95996

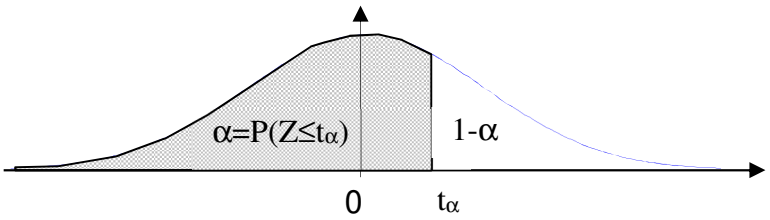


fig 3-11: Illustration de la table de la loi Normale $\mathcal{N}(0,1)$ inverse

Valeur **très** remarquable:

$$P(\mu-\sigma \le X \le \mu+\sigma) = P(-1 \le Z \le 1) = P(Z \le 1) - P(Z \le -1) = P(Z \le 1) - (1 - P(Z \le 1))$$

$$P(\mu-\sigma \le X \le \mu+\sigma) = 2P(Z \le 1) - 1 = 2 \text{ x } 0,8413 - 1 = 0,6826 \approx 0,7$$

La probabilité pour qu'une variable suivant une loi de Gauss se trouve dans l'intervalle μ±σ est d'environ 70%.

$$P(\mu-2\sigma \le X \le \mu+2\sigma) = 2P(Z \le 2) - 1 = 2 \text{ x } 0,9772 - 1 = 0,9544 \approx 0,95$$

De même, la probabilité pour qu'une variable suivant une loi de Gauss se trouve dans l'intervalle μ±2σ est supérieur à 95%.

Remarque pratique: que l'on calcule P(X<a) ou P(X>a) avec a négatif ou positif, **IL FAUT TOUJORS FAIRE UN DESSIN** pour représenter la surface que l'on veut calculer et voir à quoi elle correspond par rapport à ce que nous donne la table. Au DS vous n'aurez à votre disposition que la table inverse, on prendra alors dans la table la valeur la plus proche de t_p recherché pour trouver $P(T \le t_p)$

4.4.1.3 Approximation de la loi binomiale par la loi Normale

Si n est grand et si ni p et (1-p) ne sont très petits (p n'est ni proche de 0 et de 1), la loi normale fournit une bonne approximation de la loi binomiale.

Les conditions d'application de l'approximation de la loi binomiale par la loi Normale diffèrent suivant les auteurs:

- si n>5 et $\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \frac{1}{\sqrt{n}} < 0,3$
- n≥30 et (n p>10) et (n(1-p)>10)
- si np(1-p)>9

alors on peut remplacer la loi binomiale $\mathcal{B}(n,p)$ par la loi $\mathcal{N}(\mu,\sigma^2)$ avec :

$\mu = n\,p$

$$\sigma^2 = n\,p\,(1-p)$$

Il y a donc conservation de la moyenne et de l'écart type.
Si l'approximation de la loi binomiale par la loi Normale ne s'applique pas, on essayera l'approximation de la loi par une loi de Poisson (n>50 et p<0,1); sinon on garde la loi binomiale qui donne des calculs plus fastidieux.

Correction de continuité

L'approximation de la loi binomiale, qui est liée à une variable aléatoire discrète, par la loi Normale, qui est une loi continue, pose un problème traité ici sur un exemple.
Soit X une v.a.r. discrète qui suit une loi binomiale $\mathcal{B}(n,p)= \mathcal{B}(30;\ 0,6)$. La probabilité P(X=20) vaut

$$P(X=20) = C_{30}^{20}(0,6)^{20}(1-0,6)^{30-20} = 0,1152$$

Comme la loi Normale $\mathcal{N}\left(n\,p, n\,p\,(1-p)\right) = \mathcal{N}\left(30x0,6;30x0,6\,(1-0,6)\right) = \mathcal{N}(18;7,2)$ est une loi continue,

$P(X=20) = P(Z=\frac{20-18}{2,683}) = 0$ car la loi de ne peut s'interpréter qu'en considérant la surface comprise sous la courbe et non pas ses valeurs particulières.
Pour avoir une approximation de la loi binomiale en X=20 il faut prendre la surface sous la loi normale entre 19,5 et 20,5:

$$P(\frac{19,5-18}{2,683} < Z \le \frac{20,5-18}{2,683}) = P(Z \le \frac{20,5-18}{2,683}) - P(Z \le \frac{19,5-18}{2,683}) = P(Z \le 0.9317) - P(Z \le 0.5590) = 0,1123$$

4.4.1.4 Approximation de la loi de Poisson par la loi Normale

Si λ>18 alors on peut remplacer la loi de Poisson $\mathcal{P}(\lambda)$ par la loi $\mathcal{N}(\lambda,\lambda)$. Il y a conservation de la moyenne et de l'écart type.
Une correction de continuité identique à celle expliquée dans le cadre de l'approximation de la loi binomiale est à appliquer.

4.4.1.5 Théorème central limite

Soit X_i, n v.a.r. indépendantes qui suivent la même loi d'espérance m et d'écart type σ. Soit M_n la variable aléatoire définie par : $M_n = \frac{X_1 + ... + X_n}{n}$

Si n est suffisamment grand (**n>30)**, alors M_n suit approximativement la loi Normale $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Ce théorème permet d'utiliser la loi Normale dans le cadre de la théorie des **grands échantillons**.

4.4.2 Loi de Student

La loi de Student est utilisée à la place de la loi Normale $\mathcal{N}(0,1)$ lorsque dans la sommation de v.a.r. indépendantes (en nombre <30), l'écart type σ des variables est inconnu et fait l'objet d'une estimation.
On dit qu'une v.a.r. continue X suit une loi de Student à n degrés de liberté (ddl). Cette distribution est symétrique de moyenne égale à 0, son écart type diminue et tend vers 1 quand le ddl augmente.
Le nombre ddl correspond au nombre d'informations non redondantes utilisées.
La densité de probabilité f(x) d'une loi de Student à n degrés de liberté est définie sur $\Re$ par.

$$f(x) = \frac{\Gamma\left((n+1)/2\right)}{\sqrt{n\pi}\ \Gamma\left(n/2\right)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

avec la fonction $\Gamma(x)$ qui est la généralisation de la fonction factorielle : $\Gamma(x+1) = x\,\Gamma(x)\ \ \Gamma(1) = 1$;
et ainsi: $\Gamma(n+1) = n!$ pour n entier.

Notation: on dit qu'une v.a.r. X suit une loi $\mathcal{S}(n)$.

Par exemple, on va voir dans le chapitre sur l'estimation que si une v.a.r. a pour moyenne μ et un écart type inconnu mais que celui calculé sur un échantillon de n valeurs vaut σ_e, alors la v.a.r. X="moyenne d'un échantillon de n valeurs" suit une loi $\mathcal{S}(n)(\mu;\sigma(n))$, loi de Student à n degrès de liberté de moyenne m et de variance $\sigma^2(n)=\frac{\sigma_e^{\,2}}{n-1}$.

La variable $\frac{X-m}{\sigma(n)}$ suit une loi $\mathcal{S}(n)$, de moyenne 0 et d'écart type tendant sers 1 quand n tend vers l'infini.

Propriété: La loi normale $\mathcal{N}(0,1)$. peut être utilisée comme approximation de la loi de Student quand le ddl est grand (ddl>30).

Table: La loi $\mathcal{S}(n)(0;1)=\mathcal{S}(n)$ est tabulée pour différents degrés de liberté.

ATTENTION, la table utilisée en pratique, ainsi que les fonctions programmées dans les logiciels, est une table **bilatérale**. Elle donne la valeur t_α telle que :
$P(|S|>t_\alpha)=P(S>t_\alpha)+P(S<-t_\alpha)=\alpha$ pour une loi $\mathcal{S}(n)$
Dans Excel ou OpenOffice_Calc : LOI.STUDENT.INVERSE.BILATERALE (α;n)

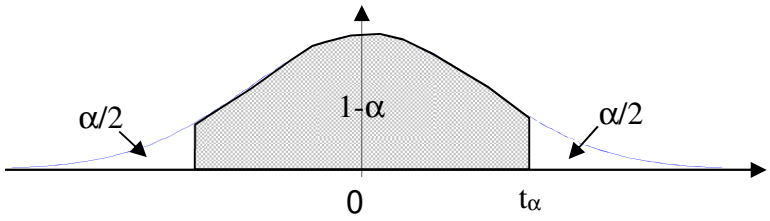


fig 3-12: Illustration de la fonction LOI.STUDENT.INVERSE

Exemple: la commande : LOI.STUDENT.INVERSE.BILATERALE (0,05;8)= 2,3060.
La valeur 2,306 est lue dans la table pour $\upsilon=8$ degrés de liberté et α=0,95
Pour α=0,95 et $\upsilon>30$, la valeur tend vers 1,960 qui est aussi lu dans la table $\mathcal{N}(0,1)$ pour t_p=0,975, car la table Z est unilatérale.
Pour connaitre la valeur des probabilités : LOI.STUDENT.BILATERALE (2,3060;8)= 0,05
Il existe aussi les fonctions : LOI.STUDENT.N et LOI.STUDENT.DROITE qui renvoient les probabilités unilatérale gauche et droite, et la fonctions LOI.STUDENT.INVERSE.N qui donne la valeur d'une variable de Student possédant une probabilité unilatérale gauche donnée.

4.4.3 Loi du Khi-deux

Si les v.a.r. X_1, ... , X_n suivent toutes une loi $\mathcal{N}(0,1)$ (centrées réduites), la somme des carrés de ces variables suit une loi du Khi-deux à n degrés de liberté (ddl). Le nombre de degrés de liberté n correspond au nombre de termes linéairement indépendants impliqués dans le calcul d'une somme de carrés basée sur n observations indépendantes.

Notation: on dit que X suit une loi $\mathcal{X}^2(n)$. Cette distribution n'est pas symétrique.
La densité de probabilité, mais surtout la fonction de répartition sont tabulées

Exemple: si on dispose de n réalisations indépendantes d'une même variable X suivant la loi $\mathcal{N}(\mu,\sigma)$, on peut écrire que la variable $\sum_{i=1}^{n}\left(\frac{X-\mu}{\sigma}\right)^2$ suit une loi $\mathcal{X}^2(n)$.

Exemple: avec le logiciel Excel ou OpenOffice_Calc, on dispose de la fonction

LOI.KHIDEUX.DROITE (t_α;n) qui donne la probabilité $\alpha : P(\chi^2 > t_\alpha) = \alpha$ pour une loi $\mathcal{X}^2$(n), ainsi la commande : LOI.KHIDEUX.DROITE (30,14;19)= 0,05

LOI.KHIDEUX.N (t_α;n) qui donne la probabilité $\alpha : P(\chi^2 < t_\alpha) = \alpha$ pour une loi $\mathcal{X}^2$(n)

La fonction LOI.KHIDEUX.INVERSE.DROITE (α;n) donne t_α telle que $P(\chi^2 > t_\alpha) = \alpha$ pour une loi $\mathcal{X}^2$(n). Cela correspond à 1- (valeur de la table du $\mathcal{X}^2$) :

LOI.KHIDEUX.INVERSE renvoie donne t_α telle que $P(\chi^2 < t_\alpha) = \alpha$.

4.4.4 Loi exponentielle

La loi exponentielle est définie pour x≥0. Une v.a.r. continue X suit une loi exponentielle de paramètre λ, si sa densité de probabilité f(x) est définie sur $\Re^+$ par.

$$f(x) = \lambda e^{-\lambda x}$$

La fonction de répartition F(t), probabilité $P(X \le t) = \int_{-\infty}^{t} f(x)\,dx$ a pour valeur:

$$F(t) = P(X \le t) = 1 - e^{-\lambda t}$$

Cette loi est utilisée pour les calculs de fiabilité. C'est la loi d'attente du premier évènement, ou de l'intervalle entre deux évènements consécutifs, d'un processus de poisson de paramètre λ. Cela correspond à la loi du temps entre deux pannes d'un appareil dont le taux de panne suit une loi de poisson et n'est donc pas sensible au vieillissement.

Propriété: $E(X) = \frac{1}{\lambda}$, $V(X) = \frac{1}{\lambda^2}$; $\sigma_X = \frac{1}{\lambda}$

Exemple : le temps d'attente entre 2 demandes de renseignement suit une loi exponentielle de paramètre λ= 1/(3mn) (temps d'attente moyen : E(X)=1/λ=3min), calculer la probabilité que deux demandes soient espacées de moins de 2mn

$P(X \le 2) = 1 - e^{-2/3} = 0{,}4866$

Avec Excel, la fonction : =LOI.EXPONENTIELLE.N(2;1/3;VRAI) donne 0,4866

LOI.EXPONENTIELLE.N(x;lamda; cumulative)

Ou =LOI.GAMMA.N(2;1;3;VRAI) qui donne aussi 0,4866

LOI.GAMMA.N(x;alpha;beta;cumulative) ;

4.4.5 Processus de Poisson

Le processus de Poisson d'intensité λ (réel strictement positif) est un processus de comptage d'occurrences qui vérifie les conditions suivantes :

- Les nombres d'occurrences dans des intervalles de temps disjoints sont indépendants
- La probabilité d'une occurrence dans un intervalle de temps est proportionnelle à la longueur de cet intervalle, le coefficient de proportionnalité étant λ
- La probabilité qu'il y ait plus d'une occurrence dans un petit intervalle de temps est négligeable

Ces deux dernières conditions forment la propriété dite des « événements rares ».

Si on trace le nombre N d'occurrence en fonction du temps t écoulé, on obtient une fonction en escalier à largeur de marches inégales mais dont la hauteur reste constante (1). Elle est de pente moyenne λ (nombre d'occurrence par unité de temps)

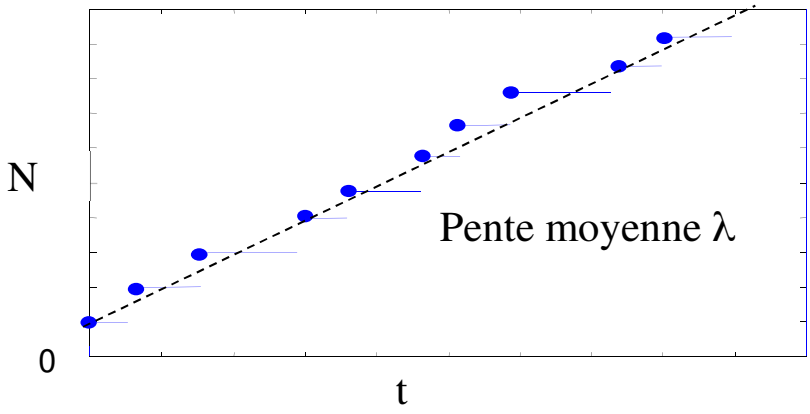


fig 3-13: Nombre d'occurrence en fonction du temps d'un processus de Poisson

- On peut appréhender un processus de Poisson par l'étude du nombre Nt "nombres d'occurrences dans des intervalles de temps t donné (strictement positif)". Cette variable suit une loi de Poisson de paramètre λt, c'est-à-dire que :

  $$P(Nt = k) = \frac{e^{-(\lambda t)}(\lambda t)^k}{k!}.$$

  Nt est une loi discrète.

Exemple : soit processus de Poisson d'intensité λ= 3 min^{-1} (3 occurrences par minutes). On considère un intervalle de temps de 10 min et la v.a. X="nombres d'occurrences dans un intervalle de temps de 10 min".

- Quelle est la loi suivie par X : X suit une loi de Poisson de paramètre (λt=3 min^{-1}.10min=30)
- Calculer P(X> 40). Pour le calcul de $P(X > 40)$ on approxime X par une loi $\mathcal{N}(30,30)$

- On peut aussi appréhender un processus de Poisson par l'étude des temps d'arrivée entre deux occurrences. On démontre aussi que le temps S s'écoulant entre deux incrémentations du processus de comptage (rappelons que la probabilité que le processus de comptage augmente d'un coup de deux unités ou plus est nulle d'après la définition) est une variable aléatoire suivant une loi exponentielle de paramètre λ, c'est-à-dire que :

  $$P(S \le t) = 1 - e^{-\lambda t}$$

  S est une loi continue.

Exemple : une occurrence peut représenter une arrivée dans une file d'attente, une panne, un accident, un appel téléphonique… . Ces processus décrivent des événements qui surviennent successivement mais qui sont espacés par des durées imprévisibles.

Exemple : soit processus de Poisson d'intensité λ= 3/min (3 occurrences par minutes). Soit la v.a. S="temps d'attente d'une nouvelle occurrence à partir de l'instant d'arrivée d'une occurrence précédente".

- Quelle est la loi suivie par S : S suit une exponentielle loi exponentielle de paramètre λ=3/min=3/60s=1/20s=0,05s^{-1} (temps d'attente moyen : E(X)=1/λ=(1/3)min=20s),
- Calculer P(S<30s) : $P(S \le t) = 1 - e^{-\lambda t} = 1 - e^{-\frac{30}{20}} = 0{,}777$

Exemple : le temps d'attente X entre 2 demandes de renseignement suit une loi exponentielle de paramètre λ= 1/(3min) (temps d'attente moyen : E(X)=1/λ=3min). Pour calculer la probabilité que deux demandes soient espacées de moins de 2min, on peut aussi dire que le nombre Y de

demandes en t=2 min suit une loi de Poisson de moyenne λt=2min/3min=2/3, et qu'on cherche : $P(Y \geq 1) = 1 - P(Y < 1) = 1 - P(Y = 0) = 1 - 0{,}5134 = 0{,}4866$

De plus, les temps s'écoulant entre deux incrémentations du processus de comptage sont des variables aléatoires indépendantes.

Mais la somme de v.a. exponentielles indépendantes n'est pas une v.a. exponentielle. Du coup, si l'on étudie la durée T_n entre PLUSIEURS (n) tops, donc une somme d'espacements indépendants $T_n = S_1 + ... + S_n$, on se tourne vers une loi gamma (et plus précisément d'Erlang puisque le nombre de périodes est un entier naturel). En effet, la loi exponentielle est un cas particulier de la loi d'Erlang qui ne vérifie pas la propriété d'additivité alors que loi d'Erlang a l'avantage de vérifier la propriété d'additivité.

Exemple : L'arrivée d'autobus forme un processus de Poisson d'intensité $\lambda = 4/h$. Chaque autobus s'arrête un temps fixe $\tau = 1mn$ à la station. Un passager qui arrive à l'instant θ monte dans le bus s'il est présent, attend pendant un temps $\tau' = 5mn$ puis si l'autobus n'est pas arrivé s'en va à pied. Déterminer la probabilité qu'il prenne l'autobus.

Le passager rentre à pied si aucun autobus n'est présent à l'arrêt entre les temps $[\theta - \tau, \theta + \tau']$, soit un intervalle de temps de largeur $\tau + \tau'$. La probabilité qu'aucun bus ne se présente pendant cette durée est $P(X \leq t) = 1 - e^{-\lambda(\tau+\tau')} = 1 - e^{-\frac{4}{60}(1+5)} = 1 - e^{-0,4} = 0{,}3297$

4.4.6 Loi gamma et loi d'Erlang

La distribution gamma est définie pour x≥0. Une v.a.r. continue X suit une loi gamma $\Gamma(a, 1/\lambda)$, avec a (ou α), paramètre de forme et λ paramètre d'intensité, si sa densité de probabilité f(x) est définie sur $\Re^+$ par.

$$f(x) = \frac{1}{\Gamma(a)} \lambda^a x^{a-1} e^{-\lambda x}$$

ou encore $f(x) = \frac{1}{\Gamma(a)} \frac{1}{b^a} x^{a-1} e^{-x/b}$ avec $b(ou\ \beta) = 1/\lambda$ le paramètre d'échelle

Pour les valeurs de k entier, la fonction $\Gamma(k) = (k-1)!$ et on obtient la loi d'Erlang

$$f(x) = \frac{\lambda^k x^{k-1} \exp(-\lambda x)}{(k-1)!} \text{ pour } x>0.$$

La distribution d'Erlang est la distribution de la somme de k variables aléatoires indépendantes et identiquement distribuées selon une loi exponentielle.
Si on fait k=1 on retrouve la loi exponentielle : $f(x) = \lambda \exp(-\lambda x)$

La fonction de répartition de la distribution d'Erlang est : $P(X < x) = 1 - \sum_{n=0}^{k-1} e^{-\lambda x} \frac{(\lambda x)^n}{n!}$

Propriété: $E(X) = \frac{a}{\lambda}$, $V(X) = \frac{a}{\lambda^2}$; $\sigma_X = \frac{\sqrt{a}}{\lambda}$

Additivité : la somme de deux v.a. suivant une loi d'Erlang $\Gamma(a_1, 1/\lambda)$ et $\Gamma(a_2, 1/\lambda)$, suit une loi d'Erlang $\Gamma(a_1 + a_2, 1/\lambda)$. On retrouve cette propriété dans la loi de Poisson dont la loi gamma normalisée en fait la version continue.

Exemple : le temps d'attente entre 2 demandes de renseignement suit une loi exponentielle de paramètre λ= 1/(3min), (b ou β= 3min), calculer la probabilité d'avoir plus de 2 demandes (a ou alpha=2)

espacées de moins de 6min. On considère la v.a. X="temps d'arrivée de 2 demandes" qui suit une loi d'Erlang ou Gamma $\Gamma(2,3)$.

$P(X \le 6) = 0{,}594$ avec la fonction Excel : =LOI.GAMMA.N(6;2;3;VRAI)

LOI.GAMMA.N(x;alpha;beta;cumulative) ;

On pourrait aussi dire que le nombre de demandes en 6 min Y suit une loi de Poisson de moyenne 2, et qu'on cherche

$P(Y \ge 2) = 1 - P(Y < 2) = 1 - (P(Y = 0) + P(Y = 1)) = 1 - 0{,}406 = 0{,}534$

Application au télé-trafic : la propriété d'additivité lui vaut d'être utilisée dans la gestion du télé-trafic, par exemple pour déterminer l'effectif optimal d'un centre d'appels téléphoniques ou le nombre de canaux à ouvrir dans une station GSM. Ces applications utilisent en fait des formules dérivées de la loi d'Erlang (Erlang-B avec perte des appels non aboutis et Erlang-C avec attente jusqu'au service).

Voir : http://hp.vector.co.jp/authors/VA002244/erlang.htm pour un calculateur gratuit

4.4.7 Loi de Weibull

La loi de Weibull est définie pour x≥0. Une v.a.r. continue X suit une loi de Weibull de paramètre de forme α et de paramètre d'échelle β, si sa densité de probabilité f(x) est définie sur $\Re^+$ par:

$$f(x) = \frac{\alpha x^{\alpha-1}}{\beta^{\alpha}} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$$

La fonction de répartition F(x), probabilité $P(X \le t) = \int_{-\infty}^{t} f(x)\,dx$ a pour valeur:

$$F(x) = P(X \le t) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$$

Cette loi est utilisée pour les calculs de fiabilité.

5 Un peu d'histoire

BAYES Thomas, anglais, 1702-1761

Théologien (protestant), il s'adonna aux mathématiques sous la houlette de De Moivre. Travaux sur le calcul des probabilités et en calcul différentiel. Il sera le premier, avant Laplace, à exposer le problème de la probabilité des causes : calcul de la probabilité d'un événement complexe dont on sait qu'un de ses composants ("causes") s'est produit. Il fut membre de la Royal Society.

FISHER A., anglais, 1890-1962

Ronald A. Fisher (1890–1962) est, comme Karl Pearson, l'un des principaux fondateurs de la théorie moderne de la statistique. Fisher étudia à Cambridge où il obtint en 1912 un diplôme en astronomie. C'est en étudiant la théorie de l'erreur dans les observations astronomiques que Fisher s'intéressa à la statistique. Fisher est l'inventeur de la branche de la statistique appelée l'analyse de la variance.

GAUSS Karl Friedrich, allemand, 1777-1855

Enfant prodige, illustre mathématicien et physicien (importants travaux et publications en électricité, optique et magnétisme, théorie du potentiel), astronome (succédant à Mayer, il fut directeur de l'observatoire de Göttingen). Il établit l'orbite de Cérès (découverte en 1801 par l'astronome italien Giuseppe Piazzi) en utilisant la méthode des moindres carrés. Ce grand savant sera surnommé par ses pairs Prince des mathématiciens. Le "gauss" est l'unité d'induction magnétique.

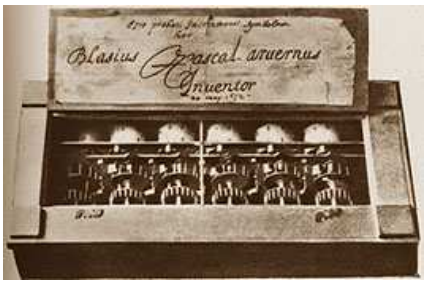
GOSSET William S. dit "STUDENT", anglais, 1876-1937

Le théorème dit "de Student" fut obtenu par William S. Gosset (1876–1937) aux alentours de 1910. Gosset, qui avait étudié les mathématiques et la chimie, travaillait comme statisticien pour la brasserie Guinness en Angleterre. A l'époque, on savait que si $X_1,...,X_n$ sont des variables aléatoires indépendantes et identiquement distribuées $\mathcal{N}(m,\sigma)$, alors la variable $\dfrac{\overline{X}-m}{\sigma/\sqrt{n}}$ suit une loi $\mathcal{N}(0,1)$. Toutefois, dans les applications statistiques on s'intéressait plutôt à la quantité $\dfrac{\overline{X}-m}{\hat{\sigma}/\sqrt{n}}$ où $\hat{\sigma}$ est l'écart type estimé sur un échantillon de taille n. On se contentait alors de supposer que cette quantité suivait, à peu près, la loi $\mathcal{N}(0,1)$. Suite à de nombreuses simulations, Gosset arriva à la conclusion que cette approximation était valide seulement lorsque n est suffisamment grand. Il décida donc d'essayer d'obtenir la distribution exacte de la quantité $\dfrac{\overline{X}-m}{\hat{\sigma}/\sqrt{n}}$.

Après avoir suivi un cours de statistique avec Karl Pearson, il obtint son célèbre résultat. Gosset publia tous ses travaux de statistique sous le pseudonyme de Student. Ainsi on appelle loi de Student la loi de probabilité qui aurait dû être appelée la loi de Gosset.

PASCAL Blaise, français, 1623-1662

Philosophe de renom, auteur des célèbres Pensées, mathématicien et physicien. Sa mère mourut alors qu'il n'avait que 3ans; il fut élevé et instruit par son père Étienne Pascal.
A 12 ans, Blaise découvrait et démontrait des théorèmes classiques de géométrie euclidienne. A 16 ans, il écrivait, en latin, un *Essay pour les coniques* inspiré des travaux de Desargues. A 19 ans, il mit au point et fit construire, en plusieurs exemplaires, une machine à calculer que l'on peut admirer à Clermont-Ferrand, sa ville natale, et qu'il présenta à la reine Christine de Suède par ces mots : "*cet ouvrage, Madame, est une machine pour faire les règles d'arithmétique sans plume et sans jetons*".
Sa principale contribution en physique porte sur l'hydrostatique et l'étude de la pression atmosphérique (Le *pascal* est une unité de pression correspondant à 1 newton par mètre carré) suite aux découvertes et travaux



de Torricelli. Pascal eut une santé fragile. A la mort de son père (1651), il se retire quelques temps du monde scientifique. Il entre au couvent de Port-Royal (1654) suite à une révélation mystique, tout en poursuivant son œuvre scientifique et philosophico-religieuse.

PEARSON Karl, anglais, 1857-1936

Le statisticien Karl Pearson fut l'un des premiers à démontrer le rôle important de la loi du khi-deux en statistique. Sa principale contribution fut sans doute le célèbre test d'ajustement du khi-deux. Karl Pearson est considéré à juste titre comme l'un des pères de la théorie moderne de la statistique. Son fils, Egon S. Pearson (1895–1980), fît également d'importants travaux en statistique. Dans les années 1930, il développa, avec Jersey Neyman (1894–1981), la théorie moderne des tests d'hypothèse.

POISSON Siméon Denis, français, 1781- 1840

Brillant polytechnicien, élève de Fourier et de Laplace, astronome et physicien. On le connaît bien sûr pour sa célèbre loi de probabilités portant son nom (Théorie du calcul des probabilités, 1838), mais ses travaux portent cependant principalement en électricité, magnétisme, mécanique et mouvements vibratoires (théorie de la chaleur, théorie des ondes) où, introduisant de nombreux concepts mathématiques liés aux équations de Laplace (théorie du potentiel électrostatique, équations aux dérivées partielles), il apparaît, à la suite de Daniel Bernoulli et Fourier comme le bâtisseur de la physique mathématique moderne (étude, au moyen de la seule analyse mathématique, du comportement d'un phénomène, en tant que conséquence des lois -attribuées par l'expérience- qui le régissent). Membre éminent de l'Académie des Sciences, il rejettera les travaux du jeune Galois en 1831.

Echantillonnage et estimation

1 Introduction

On va étudier les relations qui peuvent exister entre un échantillon et une population mère.

- Si on connaît la population et si on se pose des questions sur l'échantillon (comme l'échantillon est-il représentatif de la population ?) -> on fait de l'échantillonnage.
- Si on connaît l'échantillon et si l'on souhaite estimer un paramètre de la population totale -> on fait de l'estimation

On utilise la même théorie:

- dans le cas de l'échantillonnage on connaît les paramètres de la population mère
- dans le cas de l'estimation cherche les paramètres de la population mère.

On s'intéresse ici principalement à l'estimation. Dans cette partie on cherche à **estimer** la valeur d'un paramètre probabiliste d'une population mère (comme la moyenne par exemple) à partir d'un **échantillon** car pour des raisons techniques ou financières, il est impossible de mesurer le paramètre sur l'ensemble des individus. Dans ce cas on peut faire de **l'estimation ponctuelle** en donnant une valeur unique proche de la valeur inconnue du paramètre ou construire un intervalle qui a de fortes chances de contenir le paramètre inconnu. Cet intervalle étant appelé **intervalle de confiance**.

Exemple: on veut vérifier les qualités d'une production en prélevant un échantillon de produits et en extrapolant à l'ensemble de la production

Notation:

Caractéristique	Echantillon	Population totale
Taille	n	N
Moyenne	$\overline{x} = \frac{\sum x_i}{n}$	μ
Ecart-type	$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$	σ
Proportion	p_e	p

2 Propriétés d'un estimateur

On note avec un "chapeau" $\hat{\mu}$ (resp $\hat{\sigma}$) l'estimateur du paramètre μ (resp σ). On va essayer de définir ici ce qu'est un bon estimateur.

2.1 Estimateur consistant

Un estimateur est consistant s'il tend vers la vraie valeur du paramètre quand on fait tendre la taille de l'échantillon vers l'infini

2.2 Estimateur sans biais (JUSTE)

Un estimateur est sans biais si son espérance mathématique (la moyenne d'un nombre infini d'estimation) est égale à la vraie valeur du paramètre.

2.3 Estimateur à variance minimale (PRECIS)

Un estimateur est à variance minimale s'il possède la plus petite variance de tous les estimateurs.

2.4 Estimateur absolument efficace (optimalement efficace)

Un estimateur est absolument efficace s'il est consistant, sans biais et s'il possède la plus petite variance de tous les estimateurs consistant et sans biais.

3 Estimation ponctuelle

Soit X une v.a.r. défini sur une population mère d'espérance μ et de variance σ^2.

3.1 Moyenne, proportion

Une estimation convergente, sans biais et efficace de la moyenne μ inconnue de la population mère est estimée par la moyenne de l'échantillon:

$$\hat{\mu} = \overline{x}$$

De même une estimation convergente, sans biais et efficace de la probabilité p (ou la proportion, fréquence en %) de la population totale est prise égale à celle de l'échantillon p_e:

$$\hat{p} = p_e$$

3.2 Variance, écart type

Une estimation convergente, sans biais $\hat{V}$ de la variance inconnue de la population mère est donnée à partir de la variance de l'échantillon:

$$\hat{V} = \hat{\sigma}^2 = \frac{n}{n-1} s^2$$

Une estimation convergente, quasi sans biais $\hat{\sigma}$ de l'écart type inconnue de la population mère est donnée à partir de l'écart type de l'échantillon:

$$\hat{\sigma} = \sqrt{\frac{n}{n-1} s^2} = \sqrt{\frac{n}{n-1}}\ s$$

4 Estimation par intervalle de confiance : définition

L'estimation ponctuelle ne donne pas de renseignement sur la qualité de cette estimation. On va chercher un intervalle I (autour de la valeur de l'estimation ponctuelle) ayant un certain degrés de chances de contenir le paramètre inconnu θ.

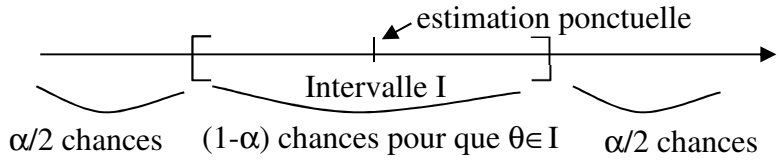


fig 4-1: notion de niveau de confiance d'un intervalle de confiance

On peut dire que $\theta \in I$ avec la probabilité ou niveau de confiance $(1-\alpha)$, et que le risque que I ne contienne pas θ est α.

5 Echantillonnage : loi de probabilité des mesures sur un échanillon

Soit une population de N unités statistiques sur laquelle on prélève un échantillon de taille n.
L'échantillonnage est dit **exhaustif** si le tirage des n individus constituant l'échantillon a lieu sans remise. Il est dit **non-exhaustif** si le tirage est réalisé avec remise.
En fait, le plus souvent, la taille d'un échantillon est faible par rapport à celle de la population et on appelle :

Taux de sondage : n/N.

Si le taux de sondage est inférieur à 10 %, l'échantillonnage est considéré avec remise.

5.1 Distribution d'échantillonnage de la moyenne

On étudie sur une population, une variable aléatoire X.

Dans cette population : X a comme moyenne $E(X) = \mu$ et comme variance $V(X) = \sigma^2$.

Si on considère **k** échantillons de taille **n**, on peut calculer, pour chacun d'eux, leur moyenne $\overline{x}_j$.

On appelle alors $\overline{X}$ ="moyenne d'un échantillon de taille **n**" la variable moyenne qui prend comme valeurs : $\{\overline{x}_1; \overline{x}_2; .....; \overline{x}_k\}$.

On peut écrire la variable $\overline{X}$ en fonction des variables X de chaque tirage :

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$ somme des n tirages ensuite divisée par n

Calculons les paramètres de la va $\overline{X}$

$$E(\overline{X}) = E\left(\frac{1}{n}(X_1 + X_2 + ... + X_n)\right) = \frac{1}{n}(E(X_1) + E(X_2) + ... + E(X_n))$$

$$E(\overline{X}) = \frac{1}{n}(n\,E(X)) = E(X) = \mu$$

et : $$V(\overline{X}) = V\left(\frac{1}{n}(X_1 + X_2 + ... + X_n)\right) = \frac{1}{n^2}(V(X_1) + V(X_2) + ... + V(X_n))$$

car les tirages X_i sont indépendant.

$$V(\overline{X}) = \frac{1}{n^2}(n\,V(X)) = \frac{V(X)}{n} = \frac{\sigma^2}{n}$$

Connaissant ces paramètres, on pourra alors construire un intervalle de confiance de la moyenne d'un échantillon de taille n.

Remarque : si le tirage est sans remise, on montre que : $V(\overline{X}) = \frac{\sigma^2}{n} \times \frac{N-n}{N-1}$, où $\frac{N-n}{N-1}$ est appelé facteur d'exhaustivité.

Pourquoi le facteur d'exhaustivité?:

Le nombre de tirages (arrangements) de n éléments parmi N vaut A_N^n sans remise et N^n avec remise. Dans le cas des tirages sans remises, on ne peut tirer deux fois la même valeur.
La variance de la moyenne est plus faible lorsque le tirage se fait sans remise. La valeur de la variance est celle du tirage avec remise multipliée par le nombre $\frac{(N-n)}{(N-1)}$ qui est inférieur à 1 (la démonstration sort du cadre de ce cours).
La variance de la moyenne est multiplié par le facteur d'exhaustivité $\frac{N-n}{N-1}$.
Lorsque la population est très grande (N beaucoup plus grand que n), le facteur d'exhaustivité est négligeable (presque égal à 1).

5.2 Distribution d'échantillonnage des fréquences

On s'intéresse à la proportion p d'individus, supposée connue, possédant une caractéristique donnée dans la population mère. La probabilité qu'un individu tiré au hasard possède la caractéristique et donc p.
Si on appelle X="le nombre d'individus possédant cette caractéristique dans un échantillon de taille n".
La v.a.r. X suit une loi binomiale $\mathcal{B}(n;p)$.
On a $E(X) = np$ et $V(X) = np(1-p)$
Quand les conditions d'approximation de la loi Binomiale par la loi Normale s'applique, c'est à dire si np(1-p)>9, alors X suit une loi $\mathcal{N}(np; n\,p(1-p))$.
La variable aléatoire qui nous intéresse est la **fréquence** de la caractéristique observée dans un échantillon de taille n. Elle correspond à la variable X/n. On la note $\mathbf{p_n}$ et elle prend comme valeurs : $\{p_1; p_2; .....; p_k\}$ pour les k échantillons.

Les paramètres de la variables $\mathbf{p_n}$ sont : $E(p_n) = \frac{E(X)}{n} = p$ et $V(p_n) = \frac{V(X)}{n^2} = \frac{p(1-p)}{n}$

La variable $\mathbf{p_n}$ suit une loi $\mathcal{N}(p; \frac{p(1-p)}{n})$

Connaissant ces paramètres, on pourra alors construire un intervalle de confiance de la fréquence d'un échantillon de taille n.

Remarque : si le tirage est sans remise, on montre que : $V(p_n) = \frac{p(-p)}{n} \times \frac{N-n}{N-1}$, où $\frac{N-n}{N-1}$ est appelé facteur d'exhaustivité.

6 Estimation par intervalle de confiance

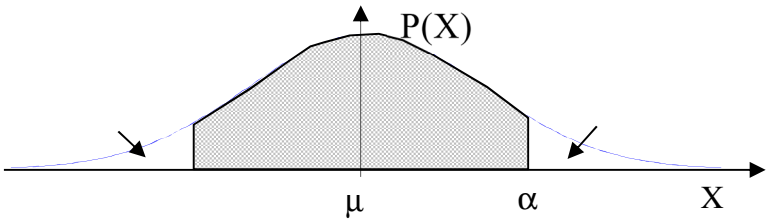
6.1 Estimation de la moyenne

6.1.1 Inégalité de Bienaymé-Tchebychev

Soit X une v.a.r. défini sur une population mère de moyenne μ et de variance σ^2. La probabilité que l'écart entre une réalisation de X est la moyenne μ soit supérieure à α est inférieure à $(\sigma/\alpha)^2$.

$$\forall \alpha > 0 ; P(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$$

La surface indiquée par les deux flèches est inférieure à $\frac{\sigma^2}{\alpha^2}$



Soit une va X défini sur une population mère, de moyenne μ et de variance σ^2.
En appliquant l'inégalité de Bienaymé-Tchebychev à la variable $\overline{X}$ ="moyenne d'un échantillon de taille n", qui a comme moyenne μ et comme variance $\frac{\sigma^2}{n}$ .on obtient

$$\forall \alpha > 0 ; P(|\overline{X} - \mu| \geq \alpha) \leq \frac{\sigma^2}{n\,\alpha^2}$$

Cette inégalité donne une idée de l'écart entre la moyenne de l'échantillon $\overline{X}$ est l'espérance de la variable X. L'écart tend vers 0 quand n tend vers l'infini.
L'intervalle de confiance qu'on peut ainsi construire n'est pas très efficace.

Exemple: Soit la variable X, d'espérance μ inconnue et de variance $\sigma^2 = 100$.
A partit d'un échantillon de 25 personnes où on a $\overline{x} = 120$, *calculer un IC de* μ, *avec un niveau de confiance de 95 % ?*

$$P\left(|\overline{X} - \mu| \geq \alpha\right) \leq \frac{\sigma^2}{n\,\alpha^2} = 0{,}05 \;\Rightarrow\; \alpha = \frac{\sigma}{\sqrt{0{,}05\ n}} = \frac{10}{\sqrt{0{,}05 \times 25}} = 8{,}94 \;\Rightarrow\; I = [\overline{x} - \alpha ; \overline{x} + \alpha]$$

Donc $I = [111 ; 129]$

6.1.2 Construction d'un intervalle de confiance (cas loi normale)

On se place par la suite dans le cas où, soit la variable X, soit la variable $\overline{X}$ ="moyenne d'un échantillon de taille n" suit une loi Normale afin d'obtenir un IC plus précis.

Cas d'une loi Normale: Si X suit une loi normale de paramètres σ^2 connus, ou si on ne connaît pas la loi de X mais que n est strictement supérieur à 30, alors la loi suivie par la variable

$\overline{X}$ ="moyenne de tous les échantillons de taille n" est normale $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Dans le cas ou seulement la variable X suit une loi normale avec σ^2 inconnus, l'intervalle de confiance s'écrit:

$$I = \left[\overline{x} - t \times \frac{\sigma}{\sqrt{n}} ; \overline{x} + t \times \frac{\sigma}{\sqrt{n}}\right]$$

avec - n: taille de l'échantillon
- $\overline{x}$: moyenne de l'échantillon
- t : obtenu dans la table d'une loi connue (normale ou Student) et dépendant du niveau de confiance α
- **mais** σ^2 variance de la **population totale**

Donc deux situations peuvent se présenter

- l'écart type de la population totale est connu
- l'écart type de la population totale est inconnu

6.1.2.1 L'écart type de la population totale est connu

On l'utilise donc directement et t est obtenu dans la table de la loi $\mathcal{N}(0,1)$ pour la valeur de probabilité (1-α/2), car la loi $\mathcal{N}(0,1)$ est généralement **unilatérale P(Z<t)**.

Exemple: On étudie la variable X qui représente le quotient intellectuel d'un individu dans une population de grande taille. On sait que X suit une loi normale $\mathcal{N}(\mu;10^2)$.

A partit d'un échantillon de 25 personnes où on a $\bar{x}=120$, *calculer un IC de μ, avec un niveau de confiance de 95 % ?*

- la taille de l'échantillon est inférieure à 30, mais la loi de X est normale $\mathcal{N}(\mu,\sigma^2)$. On connaît donc la loi de la variable $\overline{X}$ moyenne d'un échantillon de 25 personnes, $\overline{X}$ suit une loi $\mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)$. Il est donc possible de construire un intervalle de confiance.
- La taille de la population est très grande ⇒ le tirage est avec remise.
- L'estimation ponctuelle est $\hat{\mu}=\bar{x}=120$.
- $\frac{\sigma^2}{n}=\frac{10^2}{25}=2^2$
- $t_{1-\frac{\alpha}{2}}=t_{0,975}=1,96$
- $\left[\bar{x}-t_{1-\frac{\alpha}{2}}\times\frac{\sigma}{\sqrt{n}};\bar{x}+t_{1-\frac{\alpha}{2}}\times\frac{\sigma}{\sqrt{n}}\right]$ = [120 – 1,96×2 ; 120 + 1,96×2] = [116 ; 124]
  On a une probabilité de 0,95 que l'espérance E(X)=μ soit comprise entre 116 et 124.
- Avec l'inégalité de Bienaymé-Tchebychev on avait : $I=[111;129]$ ce qui est moins précis

6.1.2.2 L'écart type de la population totale est inconnu

Il faut donc l'estimer à partir de celui de l'échantillon car $\hat{\sigma}^2=\frac{n}{n-1}s^2$.

Dans ces conditions I devient: $I=\left[\bar{x}-t\frac{\hat{\sigma}}{\sqrt{n}};\bar{x}+t\frac{\hat{\sigma}}{\sqrt{n}}\right]$

MAIS ici, t est obtenu dans la table de **Student** à (n-1) degrés de liberté pour un risque α, confiance 1-α. Les tables de Student sont généralement bilatérale **P(|T|<t).**

En théorie, on montre que la variable $\frac{\overline{X}-\mu}{\hat{\sigma}/\sqrt{n}}$ suit une loi de Student à (n-1) degrés de libertés.

Remarque: si n>30, la loi de Student est remplacée par la loi normale.

Exemple: On étudie la variable X="note d'un étudiant au DS de math" qui suit une loi normale $\mathcal{N}(\mu;\sigma^2)$.

A partit d'un échantillon de 9 personnes où on a $\bar{x}=11$ et $s^2=4,2^2$, *calculer un IC de μ, avec un niveau de confiance de 95 % ?*

L'intervalle de confiance vaut:

$$I = \left[\bar{x} - t\frac{\hat{\sigma}}{\sqrt{n}} ; \bar{x} + t\frac{\hat{\sigma}}{\sqrt{n}} \right]$$

avec $\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{9}{9-1} 4{,}2^2$ => $\frac{\hat{\sigma}}{\sqrt{n}} = \frac{s\sqrt{n}}{\sqrt{n-1}} \frac{1}{\sqrt{n}} = \frac{s}{\sqrt{n-1}} = \frac{4{,}2}{\sqrt{9-1}}$

t est obtenu dans la table de Student à (9-1)=8 degrés de liberté pour une probabilité pour un risque 0,05, confiance 0,95 : t = 2,306

$I = [7{,}576; 14{,}424]$

Avec Excel, la commande LOI.STUDENT.INVERSE.BILATERALE (0,05;8) = 2,306, donne directement la valeur d'une variable aléatoire suivant une loi T de Student à 8 ddl, pour une probabilité **bilatérale** donnée 0,05.

6.1.2.3 Illustration : dispersion des mesures

- *L' écart type des mesures*

La variance permet d'évaluer la dispersion des mesures autour de la valeur moyenne. La variance est estimée par : $\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum n_i (x_i - \bar{x})^2$

Dans ce cas, la signification de l'écart type est la suivante: si une mesure **supplémentaire** de la grandeur est réalisée, la probabilité pour que **cette mesure** se situe entre $\bar{x} - \hat{\sigma}$ et $\bar{x} + \hat{\sigma}$ vaut $P(-\sigma \le X \le \sigma) = 2P(T \le 1) - 1 = 2 \times 0{,}8413 - 1 = 0.6826 \approx 0{,}7$

- *L' écart type de la **moyenne** des mesures*

La variance de la **moyenne** d'un échantillon de taille n: $\frac{\hat{\sigma}^2}{n}$

Sa signification est la suivante : si une nouvelle **série de n mesures** est réalisée, la probabilité pour que la **nouvelle valeur moyenne** obtenue se situe entre $\bar{x} - \frac{\hat{\sigma}^2}{n}$ et $\bar{x} + \frac{\hat{\sigma}^2}{n}$ vaut 0.7

Exemple: on effectue 100 mesures du temps de fermeture d'un transistor (ns).

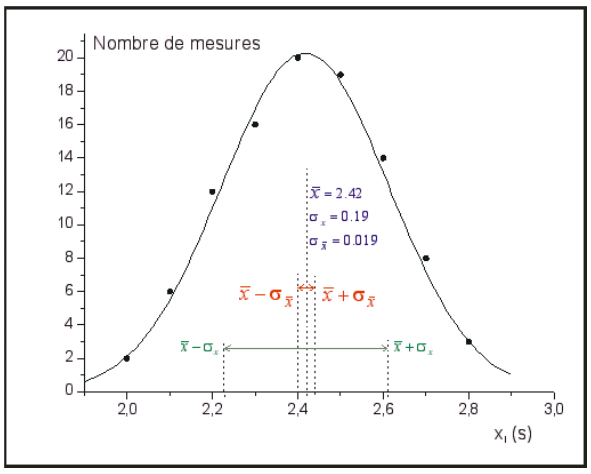


fig 4-2: illustration de la dispersion de une mesure et de la moyenne de100 mesures

On obtient comme résultat pour la valeur moyenne et la **variance**: $\bar{x} = 2{,}42\text{ns}$ et $\hat{\sigma}^2 = (0{,}19\text{ns})^2$.

On notera : $x = 2{,}42 \pm 0{,}2\text{ns}$.

La variance de la moyenne quant à elle vaut :

$\frac{\hat{\sigma}^2}{n} = 0.19^2 / 10^2 = 0.019^2$

Puisque cette valeur représente la confiance que l'on peut accorder à la valeur moyenne, c'est elle qui permet de décider quels chiffres de la valeur moyenne sont significatifs.

On notera : $\bar{x} = 2{,}42 \pm 0{,}02\text{ns}$. La valeur de la variance de la moyenne montre bien que le fait d'avoir réalisé une moyenne sur un grand nombre de mesures a fait "gagner" une décimale par rapport à une mesure unique.

6.2 Estimation d'une proportion (ou %, ou fréquence)

Quand les conditions d'approximation de la loi Binomiale par la loi Normale s'applique, c'est à dire si np(1-p)>9 , on a vu que la proportion suit donc une loi $\mathcal{N}(p;\frac{p(1-p)}{n})$.

A partir d'une proportion p_e obtenue sur un échantillon de taille n, l'intervalle de confiance s'écrit:

$$I=\left[p_e-t\sqrt{\frac{p_e(1-p_e)}{n}};p_e+t\sqrt{\frac{p_e(1-p_e)}{n}}\right]$$

où t est lu dans la table de la loi $\mathcal{N}(0,1)$ pour 1-α/2.

Les conditions d'applications supplémentaires sont les suivantes:

- n supérieur à 30
- un nombre de la population mère très grand par rapport au nombre n de l'échantillon
- un nombre d'individus X supérieur ou égal à 5

Exemple: lors d'un sondage politique sur un échantillon de 800 personnes, 224 ont déclaré vouloir voter pour le candidat A. Dans quelle intervalle est connue la proportion p=224/800=28%.?

Dans notre cas (n=800)>5 et

$$\left|\sqrt{\frac{p}{1-p}}-\sqrt{\frac{1-p}{p}}\right|\frac{1}{\sqrt{n}}=\left|\sqrt{\frac{0,28}{1-0,28}}-\sqrt{\frac{1-0,28}{0,28}}\right|\frac{1}{\sqrt{800}}=0,0346<0,3$$

ou np(1-p) = 800 x 0,28 x (1-0,28) = 161 >9

La proportion p=28% est connue avec l'intervalle de confiance:

$$I=\left[28-t\sqrt{\frac{0,28(1-0,28)}{800}};28+t\sqrt{\frac{0,28(1-0,28)}{800}}\right]$$

où t est lu dans la table de la loi $\mathcal{N}(0,1)$ pour 1-α/2.

Si on se fixe un niveau de confiance de 95%, on a $t_{1-\alpha/2}= t_{0,975} = 1,96$

et un intervalle de confiance de largeur $\pm 1,96\sqrt{\frac{0,28(1-0,28)}{800}}$ soit $\pm 3,11\%$.

Ce qui donne $I=[0,2488\%;0,3111\%]$

6.3 Remarques

- Lorsque l'on estime un paramètre par intervalle de confiance, deux qualités s'opposent: **la précision et la sécurité**. Si on exige beaucoup de sécurité (α très petit), on obtient un intervalle de confiance de large amplitude et inversement. La seule façon d'obtenir à la fois une bonne précision et une bonne sécurité est de prendre n très grand ... , mais ce n'est pas économique!

Exemple: lors d'une enquête politique 16% des personnes se sont déclarées favorable au candidat B. Sur quelle taille d'échantillon doit-on faire le sondage pour avoir un intervalle de confiance de 90% et de 99% de largeur ±0,5%

Dans la table $\mathcal{N}(0,1)$ pour 1-α/2 on trouve respectivement t =1,645 et t =2,576 pour un niveau de confiance de 90% et de 99%.

On veut $t\sqrt{\frac{0,16(1-0,16)}{n}}=0,5\%=0,005$

Pour un niveau de confiance de 90% et t=1,645 on obtient :

$$n=\left(\frac{t}{0,005}\right)^2 0,16(1-0,16)=14548$$

Pour niveau de confiance de 99% et t=2,576 on obtient :

$$n = \left(\frac{t}{0{,}005}\right)^2 0{,}16(1-0{,}16) = 35674$$

Réduire la largeur de l'intervalle de confiance augmente considérablement le coût du sondage!

- Les formules données dans ce chapitre sont valables pour un échantillon prélevé avec remise ou supposé dans une population très grande par rapport à la taille de l'échantillon (de sorte que le prélèvement de ces unités ne modifie pas trop la proportion). Cela est vérifié en pratique si n/N≤0,1.

6.4 Estimation d'une variance

Soit X une v.a.r. défini sur une population mère **suivant une loi normale** $\mathcal{N}(\mu, \sigma^2)$ de moyenne μ et de variance σ^2. On distingue deux cas :

- Si la moyenne μ est connue, la variable Y="rapport $\frac{n\,v}{\sigma^2}$ de n fois la variance $v = \frac{1}{n}\sum(x_i - \mu)^2$ d'un échantillon de taille n sur la variance de la population mère" suit une loi du $\chi^2(n)$
- Si la moyenne μ est inconnue, la variable Y="rapport $\frac{n\,s^2}{\sigma^2}$ de n fois la variance d'un échantillon de taille n sur la variance de la population mère" suit une loi du $\chi^2(n-1)$

L'allure de la loi du $\chi^2(n)$ est donnée par la figure suivante :

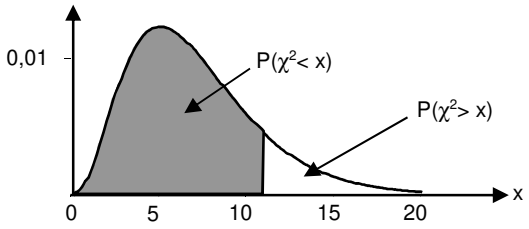


fig 4-3: densité de probabilité de la loi du $\chi^2(7)$ à 7 degrés de liberté.

Exemple: avec le logiciel Excel ou OpenOffice_Calc

LOI.KHIDEUX.DROITE (x ;ddl) donne $P(\chi^2 > x)$ pour un degrés de liberté de ddl

LOI.KHIDEUX.INVERSE.DROITE (p ;ddl) donne x tel que $P(\chi^2 > x) = p$ pour un degrés de liberté de ddl

En pratique cela donne pour la construction de l'intervalle de confiance I :

- Si μ est connu : $I = \left[\frac{n\,v}{\chi^2_{1-\frac{\alpha}{2}}(n)}; \frac{n\,v}{\chi^2_{\frac{\alpha}{2}}(n)}\right]$ avec $v = \frac{1}{n}\sum(x_i - \mu)^2$, $\chi^2_{1-\frac{\alpha}{2}}(n)$ est la valeur x telle que $P(\chi^2(n) < x) = 1 - \frac{\alpha}{2}$ et $\chi^2_{\frac{\alpha}{2}}(n)$ est la valeur x telle que $P(\chi^2(n) < x) = \frac{\alpha}{2}$ pour la loi χ^2 à n degrés de liberté.
- Si μ est inconnu : $I = \left[\frac{n\,s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}; \frac{n\,s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$ avec $s^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$ variance de l'échantillon et $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ est la valeur x telle que $P(\chi^2(n-1) < x) = 1 - \frac{\alpha}{2}$ et $\chi^2_{\frac{\alpha}{2}}(n-1)$ est la valeur x telle que $P(\chi^2(n-1) < x) = \frac{\alpha}{2}$ pour la loi χ^2 à (n-1) degrés de liberté.

Exemple: Après avoir fait passer un test noté sur 100, on choisit un échantillon de 20 personnes. On suppose que les notes suivent une loi normale $\mathcal{N}(m,\sigma)$. La variance de l'échantillon est mesurée à 182, calculer un intervalle de confiance I pour σ^2 au niveau de confiance de 95%.

La loi est normale et la moyenne m est inconnue, risque α=5% : $I = \left[\frac{n s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} ; \frac{n s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right]$ avec

n=20 ; s^2=182 ; $\chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0,975}(19) = 32,9$ valeur de x telle que $P(\chi^2(19) < x) = 0,975$;

$\chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0,025}(19) = 8,91$ de x telle que $P(\chi^2(19) < x) = 0,025$

$I = \left[\frac{20x182}{32,9} ; \frac{20x182}{8,91} \right] == [110,6;\ 408,5]$; estimation ponctuelle : $\frac{20}{19}182 = 191,57$

Test d'hypothèse

1 Généralités définitions

1.1 Hypothèses soumises au test

Les statistiques développent des techniques et des méthodes qui permettent d'analyser les données issues de l'observation, afin de cerner les caractéristiques de la population concernée et d'identifier un modèle capable d'interpréter ces données.
Dans ce cadre, on est amené à faire des hypothèses, c'est-à-dire à émettre des assertions concernant ces caractéristiques ou ce modèle.
Le plus souvent, la situation se résume en une alternative constituée de deux hypothèses H_0 et H_1, qui s'excluent mutuellement et qui sont appelées respectivement l'**hypothèse nulle**, ou fondamentale, et l'**hypothèse alternative**, ou contraire.
Souvent, les hypothèses H_0 et H_1 jouent des rôles symétriques (test bilatérale), mais ce n'est pas toujours le cas (test unilatérale).
On choisit pour hypothèse nulle H_0 une hypothèse de type **égalité**, car elle **permet de faire des calculs sur la validité du test**.

1.2 Le test

1.2.1 Définition

Les hypothèses à confronter, H_0 et H_1, étant identifiées, leur validité est soumise à l'épreuve à l'aide d'un test d'hypothèses.
Un **test d'hypothèses** est une règle de décision qui permet, sur la **base des données observées** et avec des **risques d'erreur déterminés**, d'accepter ou de refuser une hypothèse statistique.

1.2.2 Erreur, risque, niveau, puissance

La règle de décision d'un test étant basée sur l'observation d'un échantillon et non sur la base d'une information exhaustive, on n'est jamais sûr de l'exactitude de la conclusion : il y a donc toujours un risque d'erreur.
L'erreur de première espèce consiste à rejeter H_0 à tort : le **risque d'erreur de première espèce** est noté α, c'est le risque d'erreur que l'on prend en rejetant H_0 alors qu'elle est vraie. C'est le risque de voir une fausse différence. On l'appelle aussi le **niveau du test**. C'est la probabilité qu'il n'y a pas de différence, compte tenu des observations.
L'erreur de deuxième espèce consiste à accepter H_0 à tort : le **risque d'erreur de deuxième espèce** est noté β, c'est le risque d'erreur que l'on prend en acceptant H_0 à qu'elle est fausse. C'est le risque de ne pas voir une différence existante. $\eta = 1 - \beta$ est appelé **la puissance du test, c'est la capacité de voir une vraie différence avec ces observations**.
On s'efforce de construire des tests qui limitent les risques à des niveaux jugés acceptables.
En règle générale, on impose un seuil α à ne pas dépasser (par exemple 5 %, par défaut) et, compte tenu de cette contrainte, on cherche à construire les tests ayant la plus grande puissance possible.

1.2.3 Fonction discriminante

Dans la pratique, on définit une variable aléatoire T, que l'on appelle **variable de décision**, ou **fonction discriminante**, et dont la loi est connue, au moins **sous l'hypothèse H_0**.
La loi de T dépend de la taille n de l'échantillon.

1.2.4 Probabilité critique (**pvalue**)

Si l'on note t_{obs} la réalisation de la fonction discriminante T, obtenu sur un échantillon de taille n, on appelle **probabilité critique** ou **pvalue** de l'hypothèse H_0 :

- $p = P(|T| > |t_{obs}|)$ si T a tendance à s'éloigner de 0 lorsque H_0 n'est pas vraie (test **bilatéral**).

On peut aussi considérer un test unilatéral à droite (ou à gauche)

- $p = P(T > t_{obs})$ si T a tendance à prendre de grandes valeurs lorsque H_0 n'est pas vraie (test unilatéral à gauche)

- $p = P(T < t_{obs})$ si *T* a tendance à prendre de petites valeurs lorsque H_0 n'est pas vraie (test unilatéral à droite)

La probabilité critique fournit une mesure de **crédibilité de l'hypothèse** H_0 :

- une valeur très faible de la probabilité critique signifie que H_0 n'est pas valable (probabilité p de validité de H_0),
- une valeur trop élevée permet de dire que la différence entre les observations et l'hypothèse est sans doute due au hasard (probabilité p).

Le reste de ce chapitre est illustré dans le cas de test bilatéral. Pour une même valeur du risque total α, il est nécessaire d'adapter les formules quand on calcule t_{obs} dans le cas unilatéral. Dans le cas bilatéral on a $\frac{\alpha}{2}$ de chaque côté, dans le cas unilatéral on a α sur le côté considéré, la valeur de t_{obs} sera donc plus faible dans le cas unilatéral.

1.2.5 Décision avec la pvalue

La décision de refuser ou ne pas refuser l'hypothèse H_0 se fait en comparant la probabilité critique (pvalue) et le risque de première espèce α:

- pvalue < α => on refuse H_0 en prenant un risque pvalue faible de se tromper
- pvalue > α => on ne peut refuser H_0 car on prendrait alors un risque pvalue trop grand de se tromper

1.3 Illustration des risques de 1er et 2eme espece

Le tableau suivant résume ces deux risques:

		Réalité	
		H_0 vraie	H_0 fausse
Conclusion test	Rejet de H_0	α erreur de type I	(1-β) Puissance
	Non rejet de H_0	(1-α) Confiance	β erreur de type II

α= Proba (rejet de H_0 alors que H_0 est vraie)
α est le risque de conclure qu'il y a une différence alors qu'il n'y en a pas.
Par exemple dans le cas où on test l'efficacité d'un produit, α représente le risque de conclure que le produit est efficace alors qu'il ne l'est pas.
Par exemple dans le cas où on test la toxicité d'un produit, α représente le risque de retirer par erreur un produit non toxique.
(1-α) représente la confiance du test, la probabilité qu'il n'y a pas de différence, compte tenu des observations.

β= Proba (non rejet de H_0 alors que H_0 est fausse)
β est le risque de conclure à une fausse égalité.
Par exemple dans le cas où on test l'efficacité d'un produit, β représente le risque de ne pas détecter un produit qui est efficace.
Par exemple dans le cas où on test la toxicité d'un produit, β représente le risque de ne pas détecter une toxicité réelle.
(1-β) représente la puissance du test, la capacité de détection d'une différence réelle.

1.4 Illustration de la décision avec t_{obs} (valeur de la fonction discriminante)

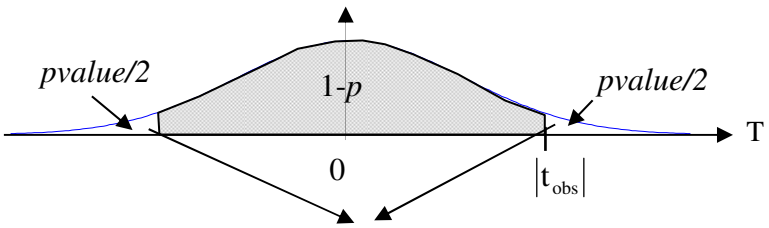
- On définit les hypothèses H_0 et H_1.
- On définit la variable aléatoire de décision *T*, dont la loi est connue, au moins sous l'hypothèse H_0.

- On calcule t_{obs} la valeur de la fonction discriminante T, obtenu sur un échantillon de taille n.

A partir de la valeur de t_{obs}, on décide le rejet éventuel de H_0 soit en utilisant la pvalue, soit directement en comparant t_{obs} avec $t_{théo}$.

1.4.1 Décision avec la pvalue

- compte tenu de la loi de probabilité de T, on calcul $P(|T| > |t_{obs}|)$, la probabilité critique (pvalue) que le fonction discriminante soit supérieur à $|t_{obs}|$ mesuré sur l'échantillon.



pvalue = probabilité que, avec l'hypothèse H_0, que la fonction discriminante soit supérieur à $|t_{obs}|$ mesuré sur l'échantillon

fig 5-1: probabilité que la fonction discriminante soit supérieur à $|t_{obs}|$ observé => rejet à tort de l'hypothèse

- si cette probabilité *pvalue* est faible ($<\alpha$), on dit que les fluctuations aléatoires ne peuvent expliquer un écart observé aussi important et on refuse l'hypothèse H_0
- on dit qu'on a refusé l'hypothèse H_0 avec un risque *pvalue* (**risque de 1ière espèce**) de se tromper (risque de rejeter H_0 alors que celle-ci est vraie)

1.4.2 Décision en comparant t_{obs} avec $t_{théo}$

En pratique, si on travaille avec des tables statistiques, et non avec un logiciel, on ne peut calculer la pvalue car la table de Student directe n'est pas disponible. On est alors amener à comparer $|t_{obs}|$ avec $t_{théo}$, la valeur de la variable décision qui correspond au risque α de 1ere espèce: $P(|T| > |t_{théo}|) = \alpha$.

- On refuse H_0 si $|t_{obs}| > |t_{théo}|$
- On ne peut refuser H_0 si $|t_{obs}| < |t_{théo}|$

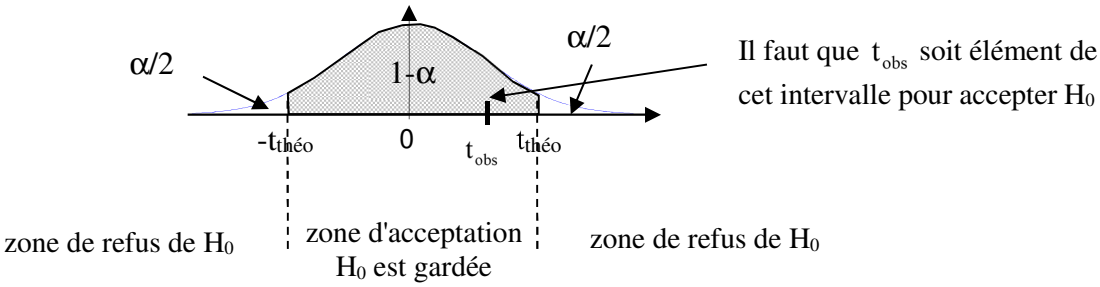


fig 5-2: explication graphique des critères d'acceptation d'un test bilatéral au risque α

L'ensemble des valeurs de t_{obs} où on refuse H_0 est appelé la **zone critique**, ou **zone de refus** de H_0. Le complémentaire est appelé **zone d'acceptation** de H_0. Ces régions dépendent du risque α de première espèce du test.

1.5 Notation

Pour une variable aléatoire définie sur une population, on note :

μ : Espérance de la variable aléatoire

σ^2 : variance de la variable aléatoire, peut parfois être supposé connu

On considère un échantillon de n mesures de la variable aléatoire :

$\bar{x} = \frac{\sum x_i}{n}$: moyenne de l'échantillon

$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^{\,2}}{n} - \bar{x}^2$: variance de l'échantillon

$\hat{\sigma}^2 = \frac{n}{n-1} s^2$: estimation de la variance de la population totale à partir de l'échantillon de taille n

2 Comparaison d'une moyenne d'échantillon à une valeur donnée (test de conformité)

2.1 X suit une loi Normale et variance de la population σ^2 connue, ou n>30

Question: Est-ce bien d'une population de moyenne égale à μ_0 qu'est issu l'échantillon observé de taille n et de moyenne $\bar{x}$.

Condition: la loi de X est une loi Normale $\mathcal{N}(\mu, \sigma^2)$

Hypothèse nulle: H_0: $\mu = \mu_0$

Hypothèse alternative: H_1: $\mu \neq \mu_0$; on parle alors de test bilatéral

si on avait $\begin{cases} H_0 : "\mu = \mu_0" \\ \text{contre } H_1 : "\mu < \mu_0" \end{cases}$ ou $\begin{cases} H_0 : "\mu = \mu_0" \\ \text{contre } H_1 : "\mu > \mu_0" \end{cases}$ on parle de test bilatéral

La variable aléatoire $\bar{X}$ ="moyenne d'un échantillon de taille n" suit une loi $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

La variable aléatoire écart E="différence entre la moyenne d'un échantillon de taille n et la valeur μ_0" suit une loi $\mathcal{N}\left(\mu - \mu_0, \frac{\sigma^2}{n}\right)$.

Si on a l'hypothèse H_0 que $\mu=\mu_0$, la fonction discriminante $T = \frac{E}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit une loi $\mathcal{N}(0,1)$

Remarque : quel que soit la loi de X, si n>30, la variable aléatoire $\bar{X}$ suit encore une loi normale.

La fonction discriminante $T = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit une loi $\mathcal{N}(0,1)$

Mise en œuvre du test : A partir des observations on calcule $t_{obs} = \frac{\bar{x} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$, on trouve $t_{théo}$ dans la table $\mathcal{N}(0,1)$ au risque α, comme la table $\mathcal{N}(0,1)$ est unilatérale on lit $t_{théo}$ pour $p = 1 - \frac{\alpha}{2}$:

- On refuse H_0 si $|t_{obs}| > |t_{théo}|$
- On ne peut refuser H_0 si $|t_{obs}| < |t_{théo}|$

- On calcule la pvalue, la probabilité p : $p = P(|T| > |t_{obs}|)$ sur la table $\mathcal{N}(0,1)$. Comme la table $\mathcal{N}(0,1)$ est unilatérale on lit $(1-p/2)$ pour $|t_{obs}|$, ou $p/2$ pour $-|t_{obs}|$
  - On refuse H_0 si $p < \alpha$
  - On ne peut refuser H_0 si $p > \alpha$
- On écrit le résultat : il (n') existe (pas) une (de) différence significative (p= ……)

Exemple: soit X qui suit une loi $\mathcal{N}(800;(49)^2)$. On étudie la moyenne d'un échantillon de n=21 éléments. On trouve une valeur moyenne de 790. Cet échantillon est-il représentatif de la population totale avec un risque de première espèce α=3% (risque de rejeter cette hypothèse H_0 alors que celle-ci est vraie)

H_0 : l'échantillon est représentatif d'une population centrée sur 800: $\mu = 800$

H_1 : $\mu \neq 800$

$$t_{obs} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{790 - 800}{\frac{49}{\sqrt{21}}} = -0{,}935$$

avec la table ou Excel : *p/2* = LOI.NORMALE.STANDARD (-0,975)= 0,175

car la fonction LOI.NORMALE.STANDARD travail pour t réel et donne la surface de moins l'infini à t.

donc : p = 0,35 = 35%, Si on refuse H_0 on a un risque de 35% que H_0 soit quand même vraie (ce qui est très supérieur au risque maximum α=3%)

A partir du risque maximum: $t_{théo} = t_{1-\alpha/2} = t_{0,985} = 2{,}1701$

$|t_{obs}| < t_{théo}$: **on garde H_0**; l'échantillon est représentatif d'une population centrée sur 800.

2.2 Variance de la population σ^2 inconnue mais estimée

Condition: la loi de X est une loi Normale ou n>30, avec σ^2 inconnu, sinon on ne peut rien faire.

Hypothèse nulle: H_0: $\mu = \mu_0$

Hypothèse alternative: H_1: $\mu \neq \mu_0$; on parle alors de test bilatéral

si on avait $\begin{cases} H_0 : "\mu = \mu_0" \\ \text{contre } H_1 : "\mu < \mu_0" \end{cases}$ ou $\begin{cases} H_0 : "\mu = \mu_0" \\ \text{contre } H_1 : "\mu > \mu_0" \end{cases}$ on parle de test bilatéral

La variable aléatoire $\bar{X}$ ="moyenne d'un échantillon de taille n" qui nous intéresse, suit une loi de Student à n-1 degrés de liberté $\mathcal{S}(n-1)\left(\mu, \frac{\hat{\sigma}^2}{n}\right)$ avec.

La variable aléatoire écart E="différence entre la moyenne d'un échantillon de taille n et la valeur a" suit une loi suit une loi de Student à n-1 degrés de liberté $\mathcal{S}(n-1)\left(\mu - a, \frac{\hat{\sigma}^2}{n}\right)$.

Si on a l'hypothèse H_0 que μ=a, la variable E suit une loi de Student à n-1 degrés de liberté $\mathcal{S}(n-1)\left(0, \frac{\hat{\sigma}^2}{n}\right)$

Si n>30, on remplace la loi de Student par la loi Normale $\mathcal{N}\left(0, \frac{\hat{\sigma}^2}{n}\right)$

On calcule l'écart observée E_{obs} de la variable E, a pour valeur : $E_{obs} = \bar{x} - a$.

On calcul la probabilité p que l'écart E soit supérieur à $|E_{obs}|$ mesuré sur l'échantillon.

Pour utiliser les tables, on se ramène à une loi normalisée $\mathcal{S}(n-1)(0,1)$ en calculant non pas E_{obs} mais:

$$t_{obs} = \frac{E_{obs}}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{\overline{x} - a}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

Mise en œuvre du test : A partir des observations on calcule $t_{obs} = \frac{\overline{x} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$, on trouve $t_{théo}$ dans la table $\mathcal{S}(n-1)$ au risque α, comme la table $\mathcal{S}$ est bilatérale on lit $t_{théo}$ directement pour $p = \alpha$:

- On refuse H_0 si $|t_{obs}| > |t_{théo}|$
- On ne peut refuser H_0 si $|t_{obs}| < |t_{théo}|$

- On calcule la pvalue, la probabilité *p* : $p = P(|T| > |t_{obs}|)$ avec un calculateur (Excel, …)
  - On refuse H_0 si $p < \alpha$
  - On ne peut refuser H_0 si $p > \alpha$
- On écrit le résultat : il (n') existe (pas) une (de) différence significative (p= ……)

Exemple 1: Pour qu'un type de détecteur de fumée soit efficace il faut qu'il se déclenche **en moyenne** dans les deux secondes qui suivent le début de l'incendie (ni trop vite, ni trop tard). Afin de vérifier un nouvel appareil, un service de contrôle en teste 12. Le nouvel appareil est-il conforme à la norme au risque de 3%?
On considère la v.a.r. X="temps de déclenchement d'un détecteur de fumée"
On ne connaît pas la loi de X, n=12<30 => on ne peut rien faire

Exemple 2: même question mais on contrôle 36 appareils avec les résultats suivants:

temps de déclenchement (s)	1,2	1,4	1,7	1,8	1,9	2,0	2,2
nombre de détecteurs	4	3	2	7	12	3	5

Le nouvel appareil est-il conforme à la norme avec un risque de première espèce α=3% (risque de rejeter cette hypothèse H_0 alors que celle-ci est vraie)?
H_0 : le nouvel appareil est conforme à la norme: $\mu = 2$
H_1 : $\mu \neq 2$

$\overline{X}$ suit une loi $\mathcal{S}(36-1)\left(\mu, \frac{\hat{\sigma}^2}{n}\right)$, comme n>30, on peut dire que $\overline{X}$ suit une loi Normale $\mathcal{N}\left(\mu, \frac{\hat{\sigma}^2}{n}\right)$.

avec $\frac{\hat{\sigma}^2}{n} = \frac{0{,}2818^2}{36} = 0{,}00221$

La variable écart E suit une loi de Student $\mathcal{S}(35)\left(\mu - a, \frac{\hat{\sigma}^2}{n}\right)$, ou par approximation une loi Normale $\mathcal{N}\left(\mu - a, \frac{\hat{\sigma}^2}{n}\right)$.

Avec l'hypothèse H_0 que μ=a, la variable E suit une loi de Student $\mathcal{S}(35)\left(0, \frac{\hat{\sigma}^2}{n}\right)$, ou par approximation une loi Normale $\mathcal{N}\left(0, \frac{\hat{\sigma}^2}{n}\right)$.

On calcule l'écart observée E_{obs} de la variable E, il a pour valeur : $E_{obs} = \overline{x} - a = 1{,}8 - 2 = -0{,}2$.

Pour utiliser les tables, on se ramène à une loi normalisée en calculant:

$$t_{obs} = \frac{\overline{x} - a}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{1,8-2}{0,04863} = -4,11$$

avec Excel : *p/2* = LOI.STUDENT.DROITE (ABS(t_{obs});35)= 1,13 10^{-4}

car la fonction Excel travail pour t≥0, avec ici 35 degrés de liberté et un risque unilatéral. Elle donne la surface de t à l'infini, donc le risque. Pour un risque bilatéral on a:

p = LOI.STUDENT.BILATERALE (ABS(t_{obs});35)= 2,27 10^{-4}

L'approximation par la loi normale donne:

p/2 = LOI.NORMALE.STANDARD.N (-4,11 ;1)= 1,98 10^{-5}

ou *p/2* = LOI.NORMALE.STANDARD.N (-ABS(t_{obs}) ;1)= 1,98 10^{-5}

car la fonction LOI.NORMALE.STANDARD.N travail pour t réel et donne la surface de moins l'infini à t.

donc : *p* = 0,023%, On refuse H_0 avec un risque de 0,023% que H_0 soit quand même vraie (ce qui est très inférieur au risque maximum α=5%)

Si on considère $t_{théo}$ avec un risque α=3%: $t_{théo} = t_{1-\alpha/2} = t_{0,985} = 2,1701$ car on prend la loi Normale (n>30).

Si n<30 **et** si X suit une loi normale dont on ne connaît pas l'écart type, on prend la loi de Student $\mathcal{S}(n)\left(\mu, \hat{\sigma}^2\right)$.

$\left|t_{obs}\right| > t_{théo}$: **on rejette H_0**; l'échantillon ne représente pas une population centrée sur 2, mais on a un risque inférieur à 3% (même très inférieur) de se tromper en refusant H_0.

Exemple 3: même question et même mesure, mais on désire savoir si le temps de déclenchement moyen est **inférieur** à deux secondes. On veut donc savoir la probabilité d'avoir un temps moyen supérieur à 2s, compte tenu des tests effectués.

C'est un problème d'estimation de moyenne. On a vu dans le chapitre sur l'estimation que si n est strictement supérieur à 30, alors la loi suivit par la "moyenne de tous les échantillons de taille n" est normale $\mathcal{N}\left(\mu, \frac{\hat{\sigma}^2}{n}\right)$. La moyenne $\overline{X}$ suit une loi $\mathcal{N}\left(\mu; \frac{\hat{\sigma}^2}{n}\right)$, soit

$\frac{\hat{\sigma}^2}{n} = \frac{0,2818^2}{36} = 0,00221$. La probabilité pour que cette moyenne soit effectivement supérieure à 2, compte tenu de nos mesures est:

$$P(X > 2) = 1 - P(X \le 2) = 1 - P(T \le \frac{2-1,8}{0,04863}) = 1,13\ 10^{-4}$$

Exemple 4: la durée de vie (en heures) des ampoules électriques produites par une usine est une v.a.r. X de moyenne de 1120h et d'écart type 120h. On désire vérifier l'affirmation du fabricant sur la durée de vie moyenne au seuil de risque de 5%, en testant un échantillon de 36 ampoules. Quelle doit être la valeur de la durée de vie moyenne si on veut rejeter l'affirmation du constructeur au risque de 5%.

La v.a.r. $\overline{x}$ qui associe à chaque échantillon de 36 ampoules sa durée de vie moyenne suit une loi $\mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right) = \mathcal{N}\left(1120; \frac{120^2}{36}\right) = \mathcal{N}(1120;4000)$

H_0 : l'affirmation du fabricant est vraie: $\mu = 1120$

H_1 : l'affirmation du fabricant est fausse $\mu < 1120$ (une durée de vie supérieure n'est pas gênante, au contraire!)

On a affaire à un test unilatéral où la zone critique ne se situe que d'un seul coté de la moyenne.

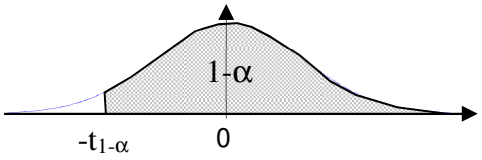


fig 5-3: explication graphique du test unilatéral d'une moyenne à une valeur m

Avec l'hypothèse H_0 la variable écart E suit une loi Normale $\mathcal{N}(0;20^2)$.

L'écart observée E_{obs} de la variable E a pour valeur : $E_{obs} = \overline{x} - 1120$.

Pour utiliser les tables, on se ramène à une loi normalisée en calculant:

$$t_{obs} = \frac{\overline{x} - a}{\sigma(n)} = \frac{\overline{x} - 1120}{20}$$

Pour avoir un risque α=5% il faut $-t_{1-\alpha} = -1{,}645$.

On refuse l'hypothèse si le risque est inférieur à 5% donc si $t_{obs} < -t_{1-\alpha}$. Ce qui donne pour la moyenne $\overline{x}$ un seuil de $\overline{x} < 20 \times (-1{,}645) + 1120 = 1087$.

Si la moyenne de l'échantillon de 36 ampoules observé est inférieure à 1087, on refuse H_0. On a 5% de chance de rejeter H_0 alors que H_0 est vraie.

3 Comparaison d'une fréquence à une valeur donnée

Question: A partir de la mesure p_{mes} sur un échantillon de n essais, peut-on déduire que la fréquence d'apparition d'un évènement est bien la valeur $p=p_1$.

Condition: les conditions d'approximation de la loi Binomiale par la loi Normale s'applique (n est grand et p n'est ni proche de 0 et de 1).

Hypothèse nulle: H_0: $p = p_1$

Hypothèse alternative: H_1: $p \neq p_1$; on parle alors de test bilatéral

si on avait $\begin{cases} H_0 : "p = p_1" \\ \text{contre } H_1 : "p < p_1" \end{cases}$ ou $\begin{cases} H_0 : "p = p_1" \\ \text{contre } H_1 : "p > p_1" \end{cases}$ on parle de test bilatéral

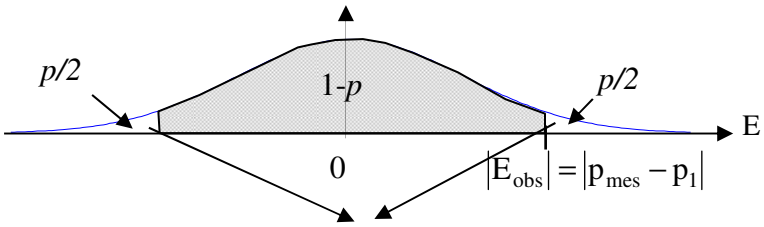
La proportion mesurée sur n essais suit une loi binomiale $\mathcal{B}(n;p)$ qui est approchée par $\mathcal{N}(p;\frac{p(1-p)}{n})$.

La variable aléatoire écart E="différence entre la proportion d'un échantillon de taille n et la valeur p_1" suit une loi $\mathcal{N}(p-p_1;\frac{p(1-p)}{n})$.

Avec l'hypothèse H_0: ($p = p_1$), cette écart suit une loi $\mathcal{N}(p_1-p_1;\frac{p_1(1-p_1)}{n}) = \mathcal{N}(0;\frac{p_1(1-p_1)}{n})$

On calcule l'écart observée E_{obs} de la variable E: $E_{obs} = p_{mes} - p_1$.

On calcul la probabilité *p* que l'écart E soit supérieur à $|E_{obs}|$ mesuré sur l'échantillon.



p = probabilité que, avec l'hypothèse H_0 ($p=p_1$), l'écart E soit supérieur à $|E_{obs}|$ mesuré sur l'échantillon

fig 5-4: illustration de la position de l'écart observé pour que H_0 soit acceptée

Pour utiliser les tables, on se ramène à une loi normalisée $\mathcal{N}(0;1)$ en calculant non pas E_{obs} mais:

$$t_{obs} = \frac{p_{mes} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}$$

Ensuite:

- Soit on se reporte sur la table (ou plutôt la fonction Excel) pour trouver la probabilité p correspondante. Si p est inférieur à α alors on refuse H_0 en prenant un risque p faible de se tromper. Si p est supérieur à α alors on ne peut refuser H_0 car le risque de se tromper serait supérieur à α. On accepte alors H_0.
- soit on se donne un risque maximum α et la table nous donne la valeur maximale $t_{théo}$ que doit prendre t_{cal} pour que le risque p pris en refusant H_0 soit inférieur à α . Dans le cas d'un test bilatéral $t_{théo} = t_{1-\alpha/2}$. Dans le cas d'un test unilatéral $t_{théo} = t_{1-\alpha}$.

Si $|t_{obs}| < t_{théo}$: on garde H_0;

Si $|t_{obs}| > t_{théo}$: on rejette H_0 avec α% de chance de se tromper (voir figure **Erreur ! Source du renvoi introuvable.**)

Exemple: un joueur tire n=800 fois une carte, au hasard, dans un jeu de 32 cartes. Un roi apparaît 134 fois. Avec seuil de risque de 5%, le jeu est-il truqué et le joueur un tricheur?

Soit F la v.a.r. qui associe la fréquence d'apparition d'un roi sur 800 tirages. F suit une loi binomiale $\mathcal{B}(n;p)$ avec p=4/32=0,125.

Dans notre cas (n=800)>5 et

$$\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \frac{1}{\sqrt{n}} = \left|\sqrt{\frac{0,125}{1-0,125}} - \sqrt{\frac{1-0,125}{0,125}}\right| \frac{1}{\sqrt{800}} = 0,08 < 0,3 ,$$

La loi binomiale peut être approchée par une loi $\mathcal{N}(p;\frac{p(1-p)}{n})$.

H_0 : l'échantillon est représentatif d'une proportion de: $p = 0,125$

H_1 : $p \neq 0,125$

La proportion mesurée vaut: p_{mes}=134/800 = 0,1675

$$t_{obs} = \frac{p_{mes} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{134/800 - 0,125}{\sqrt{\frac{0,125(1-0,125)}{800}}} = 3,635$$

avec la table ou Excel : *p/2* = LOI.NORMALE.STANDARD.N(-3,635 ;1)= 0,00014

ou *p/2* = LOI.NORMALE.STANDARD.N(-ABS(t_{cal}) ;1)= 1,4 10^{-4}

car la fonction LOI.NORMALE.STANDARD.N travail pour t réel et donne la surface de moins l'infini à t.

donc : p = 0,028%, On refuse H_0 avec un risque de 0,028% que H_0 soit quand même vraie (ce qui est très inférieur au risque maximum α=5%)
A partir du risque maximum: $t_{théo} = t_{1-\alpha/2} = t_{0,975} = 1{,}96$

$|t_{obs}| > t_{théo}$: **on rejette H_0**; l'échantillon n'est pas représentatif d'une proportion de 0,125. Le jeu est truqué ou le joueur est un tricheur.

4 Comparaison de deux moyennes

4.1 Echantillons indépendants

4.1.1 Populations normales de variances connues

Soit deux populations Normales:
- population 1 de moyenne μ_1, d'écart type σ_1 dont on connaît un échantillon de taille n_1 de moyenne $\bar{x}_1$
- population 2 de moyenne μ_2, d'écart type σ_2 dont on connaît un échantillon de taille n_2 de moyenne $\bar{x}_2$

Hypothèse nulle H_0: les deux échantillons représentent une même population: $\mu_1 = \mu_2$
Hypothèse alternative H_1: $\mu_1 \neq \mu_2$;

On considère la variable aléatoire E="différence entre les deux moyennes des échantillons de taille n_1 et n_2". Comme chaque moyenne suit une loi Normale $\mathcal{N}\left(\mu_i; \frac{\sigma^2_i}{n_i}\right)$, E suit une loi Normale $\mathcal{N}\left(\mu_1 - \mu_2; \frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}\right)$ (somme des moyennes et sommes des variances).

Si on a l'hypothèse H_0 $\mu_1 = \mu_2$, la variable E suit une loi $\mathcal{N}\left(0; \frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}\right)$

L'écart observée E_{obs} de la variable E, a pour valeur : $E_{obs} = \bar{x}_1 - \bar{x}_2$. On calcul la probabilité p que l'écart E soit supérieur à $|E_{obs}|$ mesuré sur l'échantillon. Si cette probabilité p est faible on dit que les fluctuations aléatoires ne peuvent expliquer un écart observé aussi important et on refuse l'hypothèse H_0 ($\mu_1 = \mu_2$). On conclut que $\mu_1 \neq \mu_2$, mais avec une probabilité p, faible, de se tromper.
Pour utiliser les tables, on se ramène à une loi normalisée en calculant non pas E_{obs} mais:

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}}$$

Ensuite:

- Soit on se reporte sur la table (ou plutôt la fonction Excel) pour trouver la probabilité p correspondante. Si p est inférieur à α alors on refuse H_0 en prenant un risque p faible de se tromper. Si p est supérieur à α alors on ne peut refuser H_0 car le risque de se tromper serait supérieur à α. On accepte alors H_0.
- soit on se donne un risque maximum α et la table nous donne la valeur maximale $t_{théo}$ que doit prendre t_{obs} pour que le risque p pris en refusant H_0 soit inférieur à α. Dans le cas d'un test bilatéral $t_{théo} = t_{1-\alpha/2}$. Dans le cas d'un test unilatéral $t_{théo} = t_{1-\alpha}$.

Si $|t_{obs}| < t_{théo}$: on garde H_0; on prend $|t_{cal}|$ car la loi normale est symétrique, on ne peut pas mettre en évidence de différence entre les deux échantillons
Si $|t_{obs}| > t_{théo}$: on rejette H_0 avec α% chance de se tromper

Exemple: on compare deux échantillons issus de deux populations: $\sigma_1=\sigma_2= 2$, $n_1=100$, $\bar{x}_1=12$, $n_2=50$, $\bar{x}_2=14$. Sont-ils issus de populations de même valeur moyenne, avec un risque $\alpha=5\%$ d'erreur?

H_0: $\mu_1 = \mu_2$ contre H_1: $\mu_1 \neq \mu_2$;

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{12-14}{\sqrt{\frac{2^2}{100} + \frac{2^2}{50}}} = -5{,}77$$

avec la table ou Excel : *p/2* = LOI.NORMALE.STANDARD.N(-5,77 ;1)= $4\ 10^{-9}$
ou *p/2* = LOI.NORMALE.STANDARD.N(-ABS(t_{cal}) ;1)= $4\ 10^{-9}$
car la fonction LOI.NORMALE.STANDARD.N travail pour t réel et donne la surface de moins l'infini à t.
donc : $p = 8\ 10^{-7}\%$, On refuse H_0 avec un risque de $8\ 10^{-7}\%$ que H_0 soit quand même vraie (ce qui est très inférieur au risque maximum $\alpha=5\%$)
A partir du risque maximum: $t_{théo} = t_{1-\alpha/2} = t_{0,975} = 1{,}96$

$|t_{obs}| > t_{théo}$: **on rejette H_0**; les deux échantillons ne sont pas issus de population avec la même valeur moyenne

4.1.2 Populations de lois et de variances inconnues: grands échantillons (n>30)

Soit deux populations:
- population 1 de moyenne μ_1, dont on connaît un échantillon de taille n_1 de moyenne $\bar{x}_1$, d'écart type s_1
- population 2 de moyenne μ_2, dont on connaît un échantillon de taille n_2 de moyenne $\bar{x}_2$, d'écart type s_2

Hypothèse nulle H_0: les deux échantillons représentent une même population -> $\mu_1 = \mu_2$
Hypothèse alternative H_1: : $\mu_1 \neq \mu_2$;

On considère la variable aléatoire E="différence entre les deux moyennes des échantillons de taille n_1 et n_2". Dans le cas des grands échantillons, chaque moyenne suit une loi Normale $\mathcal{N}\left(\mu_i; \frac{\hat{\sigma}^2_i}{n_i}\right)$ avec $\hat{\sigma}^2_i = \frac{n_i}{n_i - 1} s^2_i$, E suit une loi Normale $\mathcal{N}\left(\mu_1 - \mu_2; \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right) = \mathcal{N}\left(\mu_1 - \mu_2; \frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right)$.

Si on a l'hypothèse H_0 $\mu_1 = \mu_2$, la variable E suit une loi $\mathcal{N}\left(0; \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)$

L'écart observée E_{obs} de la variable E, a pour valeur : $E_{obs} = \bar{x}_1 - \bar{x}_2$. On calcul la probabilité *p* que l'écart E soit supérieur à $|E_{obs}|$ mesuré sur l'échantillon. Si cette probabilité *p* est faible on dit que les fluctuations aléatoires ne peuvent expliquer un écart observé aussi important et on refuse l'hypothèse H_0 ($\mu_1 = \mu_2$). On conclut que $\mu_1 \neq \mu_2$, mais avec une probabilité *p*, faible, de se tromper.
Pour utiliser les tables, on se ramène à une loi normalisée en calculant non pas E_{obs} mais:

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

Ensuite:

- Soit on se reporte sur la table (ou plutôt la fonction Excel) pour trouver la probabilité *p* correspondante. Si *p* est inférieur à α alors on refuse H_0 en prenant un risque *p* faible de se tromper. Si *p* est supérieur à α alors on ne peut refuser H_0 car le risque de se tromper serait supérieur à α. On accepte alors H_0.

- soit on se donne un risque maximum α et la table nous donne la valeur maximale $t_{théo}$ que doit prendre t_{cal} pour que le risque *p* pris en refusant H_0 soit inférieur à α. Dans le cas d'un test bilatéral $t_{théo} = t_{1-\alpha/2}$. Dans le cas d'un test unilatéral $t_{théo} = t_{1-\alpha}$.

Si $|t_{obs}| < t_{théo}$: on garde H_0; on prend $|t_{obs}|$ car la loi normale est symétrique, on ne peut pas mettre en évidence de différence entre les deux échantillons

Si $|t_{obs}| > t_{théo}$: on rejette H_0 avec α% chance de se tromper

Exemple: on compare deux échantillons issus de deux populations: n_1=100, $\bar{x}_1$=14, s_1^2=10, n_2=50, $\bar{x}_2$=12, s_2^2=4. Sont-ils issus de populations de même valeur moyenne, avec un risque α=5% d'erreur?

H_0: $\mu_1 = \mu_2$ contre H_1: $\mu_1 \neq \mu_2$;

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} = \frac{14-12}{\sqrt{\frac{10}{99} + \frac{4}{49}}} = 4{,}48$$

avec la table ou Excel : *p/2* = LOI.NORMALE.STANDARD.N(-4,48 ;1)= 3,7 10^{-6}
ou *p/2* = LOI.NORMALE.STANDARD.N(-ABS(t_{cal}) ;1)= 3,7 10^{-6}
car la fonction LOI.NORMALE.STANDARD.N travail pour t réel et donne la surface de moins l'infini à t.
donc : *p* = 7,4 10^{-4}%, On refuse H_0 avec un risque de 7,4 10^{-4}% que H_0 soit quand même vraie (ce qui est très inférieur au risque maximum α=5%)
A partir du risque maximum: $t_{théo} = t_{1-\alpha/2} = t_{0,975} = 1{,}96$

$|t_{obs}| > t_{théo}$: **on rejette H_0**; les deux échantillons ne sont pas issus de population avec la même valeur moyenne

4.1.3 Populations normales et variances inconnues: petits échantillons (n≤30)

Test préliminaire d'égalité des variances

Pour de petit échantillons, l'estimation des variances est imprécise, on est amené à utiliser les deux séries de mesures pour estimer la variance commune aux deux échantillons $\hat{\sigma}_c$.

C'est pourquoi avant de réaliser un test de comparaison de deux moyennes pour n≤30, il faut **toujours effectuer un test d'égalité des variances de ces 2 échantillons**. Ce test est décrit au paragraphe 7.1, le résultat en est rappelé ici :

La variable F_{obs} : $F_{obs} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} > 1$ suit une loi de Fischer-Snedecor $F_{n1-1,n2-1}$. Le test est validé si F_{obs} est inférieure à la valeur $F_{théo}$ lue dans la table de Fisher à (n1-1), n2-1) degrés de libertés.

Mise en œuvre du test de comparaison des moyennes

Soit deux populations Normales:
- population 1 de moyenne m_1, dont on connaît un échantillon 1 de taille n_1 de moyenne $\bar{x}_1$, d'écart type s_1
- population 2 de moyenne m_2, dont on connaît un échantillon 2 de taille n_2 de moyenne $\bar{x}_2$, d'écart type s_2

On suppose que les variances des deux populations sont **identiques** mais inconnues. On peut calculer une estimation $\hat{\sigma}_c$ de la variance commune σ_c :

$$\hat{\sigma}_c^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2} \quad \text{ou} \quad \hat{\sigma}_c^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

Hypothèse nulle H_0: les deux échantillons représentent une même population -> $\mu_1 = \mu_2$
Hypothèse alternative H_1: : $\mu_1 \neq \mu_2$;

On considère la variable aléatoire E="différence entre les deux moyennes des échantillons de taille n_1 et n_2". E suit une loi de Student à n_1+n_2-2 degrés de liberté $\mathcal{S}(n_1+n_2-2)\left(\mu_1-\mu_2, \frac{\hat{\sigma}_c^2}{n_1}+\frac{\hat{\sigma}_c^2}{n_2}\right)$, de moyenne nulle et de variance la somme des variances communes $\hat{\sigma}^2{}_c$:

$$\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}=\frac{\hat{\sigma}_c^2}{n_1}+\frac{\hat{\sigma}_c^2}{n_2}=\hat{\sigma}_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

L'écart observée E_{obs} de la variable E, a pour valeur : $E_{obs}=\overline{x}_1-\overline{x}_2$. On calcul la probabilité *p* que l'écart E soit supérieur à $|E_{obs}|$ mesuré sur l'échantillon. Si cette probabilité *p* est faible on dit que les fluctuations aléatoires ne peuvent expliquer un écart observé aussi important et on refuse l'hypothèse H_0 ($\mu_1=\mu_2$). On conclut que $\mu_1 \neq \mu_2$, mais avec une probabilité *p*, faible, de se tromper.
Pour utiliser les tables, on se ramène à une loi normalisée en calculant non pas E_{obs} mais:

$$t_{obs}=\frac{\overline{x}_1-\overline{x}_2}{\sqrt{\hat{\sigma}_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=\frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{(n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Ensuite:

- Soit on se reporte sur la table de Student à n_1+n_2-2 degrés de liberté (ou plutôt la fonction Excel) pour trouver la probabilité *p* correspondante. Si *p* est inférieur à α alors on refuse H_0 en prenant un risque *p* faible de se tromper. Si *p* est supérieur à α alors on ne peut refuser H_0 car le risque de se tromper serait supérieur à α. On accepte alors H_0.
- soit on se donne un risque maximum α et la table de Student à n_1+n_2-2 degrés de liberté nous donne la valeur maximale $t_{théo}$ que doit prendre t_{cal} pour que le risque *p* pris en refusant H_0 soit inférieur à α. Dans le cas d'un test bilatéral $t_{théo}=t_{1-\alpha/2}$. Dans le cas d'un test unilatéral $t_{théo}=t_{1-\alpha}$.

Si $n_1+n_2-2 > 30$, on remplace la loi de Student par la loi Normale

Si $|t_{obs}| < t_{théo}$: on garde H_0; on prend $|t_{cal}|$ car la loi normale est symétrique, on ne peut pas mettre en évidence de différence entre les deux échantillons
Si $|t_{obs}| > t_{théo}$: on rejette H_0 avec α% chance de se tromper

Exemple: deux populations suivent une loi Normale de même variance. On compare deux échantillons: $n_1=17$, $\overline{x}_1=15$, $s_1^2=16$, $n_2=26$, $\overline{x}_2=13$, $s_2^2=25$. Sont-ils issus de populations de même valeur moyenne, avec un risque α=5% d'erreur?

Test préliminaire d'égalité des variances

F_{obs} : $F_{obs} = \frac{\hat{\sigma}_1^{\,2}}{\hat{\sigma}_2^{\,2}} = \frac{\frac{26}{25}25}{\frac{17}{16}16} = \frac{26}{17} = 1{,}53$; $F_{theo} = F(25,16) = 2{,}227$

=INVERSE.LOI.F.DROITE (0,05;25;16) avec Excel

F_{obs} est inférieure à $F_{théo}$ => les variances ne sont pas significativement différentes

Mise en œuvre du test de comparaison des moyennes

H_0: $m_1 = m_2$ contre H_1: $m_1 \neq m_2$;

la variable aléatoire écart E suit une loi de Student à $n_1+n_2-2=17+26-2=41>30$ degrés de liberté. On peut utiliser la loi Normale

$$t_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{15-13}{\sqrt{\frac{17x16+26x25}{17+26-2}}\sqrt{\frac{1}{17}+\frac{1}{26}}} = 1{,}35$$

avec Excel : *p* = LOI.STUDENT.BILATERALE(ABS(t_{obs});41)= 0,18442

pour un risque bilatéral.

avec la loi normale: *p/2* = LOI.NORMALE.STANDARD.N(-1,35 ;1)= 0,0885

ou *p/2* = LOI.NORMALE.STANDARD.N(-ABS(t_{cal}) ;1)= 0,0885

car la fonction LOI.NORMALE.STANDARD.N travail pour t réel et donne la surface de moins l'infini à t.

donc : *p* = 18,4%, Si on refuse H_0 on a un risque de 18,4% que H_0 soit quand même vraie (ce qui est supérieur au risque maximum α=5%). **On ne peut refuser H_0**.

A partir du risque maximum: $t_{théo} = t_{1-\alpha/2} = t_{0,975} = 1{,}96$ car $n_1+n_2-2 > 30$, on prend la loi normale.

$|t_{obs}| < t_{théo}$: on ne peut refuser H_0; les deux échantillons sont issus de population de valeur moyenne non significativement différente.

Comparaison des moyennes si les variances sont différentes

Hypothèse nulle H_0: les deux échantillons représentent une même population -> $\mu_1 = \mu_2$

Hypothèse alternative H_1: : $\mu_1 \neq \mu_2$;

On se ramène à une loi normalisée en calculant t_{obs}:

$$t_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}}$$ qui suit une loi de Student

MAIS le degrés υ de liberté de t_{obs} est donnée par la formule de "l'équation de Welch-Satterthwaite"

$$\upsilon = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\hat{\sigma}_1^4}{n_1^2(n_1-1)}+\frac{\hat{\sigma}_2^4}{n_2^2(n_2-1)}}$$

qui est plus faible que n_1+n_2-2, ce qui donne une valeur de $t_{théo}$ plus grande et diminue la puissance du test.

Si $|t_{obs}| < t_{théo}$: on garde H_0; on ne peut pas mettre en évidence de différence entre les deux échantillons

Si $|t_{obs}| > t_{théo}$: on rejette H_0 avec α% chance de se tromper

A l'usage, on constate que tenir compte de l'inégalité des variances n'est vraiment déterminant que pour les effectifs déséquilibrés, avec n_1 très différent de n_2. Certains auteurs précisent même que l'on devrait toujours utiliser la variante pour variances inégales dès que n_1 et n_2 sont très différents.

Mise en œuvre du test de comparaison des moyennes à variances différentes
Si on reprend l'exemple précédent, mais en n'utilisant pas le calcul de la variance commune.
H_0: $m_1 = m_2$ contre H_1: $m_1 \neq m_2$;

$$t_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} = \frac{15 - 13}{\sqrt{\frac{16}{17-1} + \frac{25}{26-1}}} = 1{,}414$$

qui suit une loi de Student à υ de liberté :

$$\upsilon = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\hat{\sigma}_2^4}{n_2^2(n_2-1)}} = \frac{\left(\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right)^2}{\frac{s_1^4}{(n_1-1)^3} + \frac{s_2^4}{(n_2-1)^3}} = \frac{\left(\frac{16}{17-1} + \frac{25}{26-1}\right)^2}{\frac{16^2}{(17-1)^3} + \frac{25^2}{(26-1)^3}} = 39$$

Le degrés de liberté de 39 n'est pas très inférieur $n_1+n_2-2=17+26-2=41$ du calcul avec la variance commune, car dans notre exemple les variances et les effectifs ne sont pas très différents.
On trouve une pvalue de *p* = LOI.STUDENT.BILATERALE(1,414;39) = 0,165 qui est peu différent de 0,183 trouvé avec la variance commune.

4.2 Echantillons appariés

On parle d'échantillons appariés lorsque chaque valeur x_i de E_1 est associée à une valeur x'_i de E_2 (appariés = associés par paires)

Exemple: E_1 est un groupe de malades avant un traitement et E_2 est le même groupe de malades après traitement

Deux échantillons appariés sont de même taille, on va donc calculer les n différences $d_i = x_i - x_i'$. L'échantillon $\{d_1;...;d_n\}$ a pour moyenne $\overline{d}$ et comme écart type σ_d

Notation: $\overline{d}$: moyenne des différences dans l'échantillon
σ_d : écart type connu de la variable différence
$\hat{\sigma}_d$: écart type estimé de la variable différence
s_d : écart type des différences dans la population expérimentales

On va tester si la variable des différences a une moyenne nulle, et on se ramène ainsi au problème de la comparaison d'une moyenne à une valeur donnée, zéro.

Hypothèse nulle : H_0: $\delta=0$
Hypothèse alternative : H_1: $\delta \neq 0$;

La variable aléatoire $\overline{d}$, moyenne des différences dans l'échantillon suit une loi de Student à n-1 degrés de liberté $\mathcal{S}(n-1)\left(0, \frac{\sigma^2_d}{n}\right)$, de moyenne nulle et de variance $\frac{\sigma^2_d}{n}$:

$$\frac{\sigma^2_d}{n} = \frac{\hat{\sigma}^2_d}{n} = \frac{s^2_d}{n-1}$$

Pour utiliser les tables, on se ramène à une loi normalisée en calculant non pas $\bar{d}$ mais:

$$t_{obs} = \frac{\bar{d}}{\frac{\sigma_d}{\sqrt{n}}}$$

Ensuite:

- Soit on se reporte sur la table de Student à n-1 degrés de liberté (ou plutôt la fonction Excel) pour trouver la probabilité *p* correspondante. Si *p* est inférieur à α alors on refuse H_0 en prenant un risque *p* faible de se tromper. Si *p* est supérieur à α alors on ne peut refuser H_0 car le risque de se tromper serait supérieur à α. On accepte alors H_0.
- soit on se donne un risque maximum α et la table de Student à n-1 degrés de liberté nous donne la valeur maximale $t_{théo}$ que doit prendre t_{cal} pour que le risque *p* pris en refusant H_0 soit inférieur à α. Dans le cas d'un test bilatéral $t_{théo} = t_{1-\alpha/2}$. Dans le cas d'un test unilatéral $t_{théo} = t_{1-\alpha}$.

Si n-1 >30, on remplace la loi de Student par la loi Normale

Si $|t_{obs}| < t_{théo}$: on garde H_0; on ne peut pas mettre en évidence de différence entre les deux échantillons

Si $|t_{obs}| > t_{théo}$: on rejette H_0 avec α% chance de se tromper

Exemple: chez un groupe de 10 malades on expérimente les effets d'un traitement destiné à diminuer la pression artérielle. On observe les résultats suivants (valeurs de la tension artérielle systolique en cm Hg)

N° malade	1	2	3	4	5	6	7	8	9	10
avant traitement	15	18	17	20	21	18	17	15	19	16
après traitement	12	16	17	18	17	15	18	14	16	18
Différence	3	2	0	2	4	3	-1	1	3	-2

Le traitement a-t-il une action significative, au risque de 5%?
On précise que les distributions sont gaussiennes (en effet n<30).

Après avoir calculé les différences sur la dernière ligne du tableau, on calcule la moyenne des différences et son écart type:

$\bar{d} = 1{,}5$; s_d=1,8574; $\hat{\sigma}^2_d = \frac{n}{n-1} s^2_d = \frac{10}{10-1} 1{,}8574^2 = 1{,}9579^2$;

Hypothèse nulle : H_0: δ=0

Hypothèse alternative : H_1: $\delta \neq 0$

La variable aléatoire $\bar{d}$, moyenne des différences dans l'échantillon suit une loi de Student à 10-1=9 degrés de liberté $\mathcal{S}(9)\left(0, \frac{\hat{\sigma}^2_d}{n}\right)$ avec $\frac{\hat{\sigma}^2_d}{n} = \frac{s^2_d}{n-1} = \frac{1{,}8574^2}{9}$

$$t_{obs} = \frac{\bar{d}}{\frac{\hat{\sigma}_d}{\sqrt{n}}} = \frac{1{,}5}{\frac{1{,}8574}{\sqrt{9}}} = 2{,}42$$

avec Excel : *p* = LOI.STUDENT.BILATERALE (ABS(t_{cal});9)= 0,0386
pour un risque bilatéral. On ne peut utiliser la loi normale.

donc : p = 3,86%, Si on refuse H_0 on a un risque de 3,86% que H_0 soit quand même vraie (ce qui est inférieur au risque maximum α=5%). **On refuse H_0**.
A partir du risque maximum: $t_{théo}(9) = 2{,}26$

$|t_{obs}| > t_{théo}$: on rejette H_0; le traitement a une action significative

5 Comparaison de deux fréquences

Soit deux populations et deux fréquences d'apparition d'un évènement sur ces populations:
- population 1 avec la fréquence de succès p_1 dont on connaît un échantillon de taille n_1 possédant une proportion de succès $\hat{p}_1$
- population 2 avec la fréquence de succès p_2 dont on connaît un échantillon de taille n_2 possédant une proportion de succès $\hat{p}_2$

Hypothèse nulle H_0: les deux échantillons représentent une même proportion: $p_1 = p_2$
Hypothèse alternative H_1: $p_1 \neq p_2$;

On considère la variable aléatoire E="différence entre les deux proportions des échantillons de taille n_1 et n_2". Chaque proportion p_i sur un échantillon de n_i valeurs suit une loi binomiale $\mathcal{B}(n_i;p_i)$ qui est approchée par $\mathcal{N}(p_i;\frac{p_i(1-p_i)}{n_i})$. La v.a.r. E suit une loi Normale $\mathcal{N}\left(p_1-p_2;\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}\right)$ ou avec l'hypothèse $p_1 = p_2$: $\mathcal{N}\left(0;\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}\right)$

comme on ne connaît pas p_1 et p_2 , on peut l'approcher par $\mathcal{N}\left(0;\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)$

L'écart observée E_{obs} de la variable E, a pour valeur : $E_{obs} = \hat{p}_1 - \hat{p}_2$. On calcul la probabilité p que l'écart E soit supérieur à $|E_{obs}|$ mesuré sur l'échantillon. Si cette probabilité p est faible on dit que les fluctuations aléatoires ne peuvent expliquer un écart observé aussi important et on refuse l'hypothèse H_0 ($p_1 = p_2$). On conclut que $p_1 \neq p_2$, mais avec une probabilité p, faible, de se tromper.
Pour utiliser les tables, on se ramène à une loi normalisée en calculant non pas E_{obs} mais:

$$t_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Ensuite:

- Soit on se reporte sur la table de la loi normale (ou plutôt la fonction Excel) pour trouver la probabilité p correspondante. Si p est inférieur à α alors on refuse H_0 en prenant un risque p faible de se tromper. Si p est supérieur à α alors on ne peut refuser H_0 car le risque de se tromper serait supérieur à α. On accepte alors H_0.
- soit on se donne un risque maximum α et la table de la loi normale nous donne la valeur maximale $t_{théo}$ que doit prendre t_{cal} pour que le risque p pris en refusant H_0 soit inférieur à α . Dans le cas d'un test bilatéral $t_{théo} = t_{1-\alpha/2}$. Dans le cas d'un test unilatéral $t_{théo} = t_{1-\alpha}$.

Si $|t_{obs}| < t_{théo}$: on garde H_0;

Si $|t_{obs}| > t_{théo}$: on rejette H_0 avec α% chance de se tromper

Exemple: dans un ensemble de 40 joueurs, 23 arrivent au bout du jeu vidéo; dans un autre ensemble de 60 joueurs, 30 ont leur BAC. Ces proportions sont-elles égales avec un seuil de risque de 5%?

$$\hat{p}_1 = \frac{23}{40} = 0{,}575 \; ; \hat{p}_2 = \frac{30}{60} = 0{,}5$$

$$t_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{0{,}575 - 0{,}5}{\sqrt{\frac{0{,}575(1-0{,}575_1)}{40} + \frac{0{,}5(1-0{,}5)}{60}}} = 0{,}74$$

avec la table ou Excel : *p/2* = LOI.NORMALE.STANDARD.N (-0,74 ;1)= 0,22965
ou *p/2* = LOI.NORMALE.STANDARD.N(-ABS(t_{cal}) ;1)= 0,22965
car la fonction LOI.NORMALE.STANDARD.N travail pour t réel et donne la surface de moins l'infini à t.
donc : *p* = 45,93%, Si on refuse H_0 on a un risque de 46% que H_0 soit quand même vraie (ce qui est très supérieur au risque maximum α=5%)
A partir du risque maximum: $t_{théo} = t_{1-\alpha/2} = t_{0,975} = 1{,}96$

$|t_{obs}| < t_{théo}$: **on accepte H_0**; on ne peut pas dire que ces deux proportions sont différentes.
La différence de fréquences observées entre ces deux proportions, sur ces échantillons, n'est pas significative.

6 Comparaison d'une variance d'échantillon à une valeur donnée

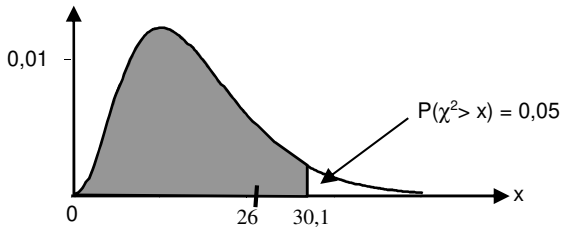
Soit X une v.a.r. défini sur une population mère **suivant une loi normale** $\mathcal{N}(m,\sigma)$ de moyenne m et d'écart type σ. Comme pour l'estimation, on considère deux cas

- Si la moyenne m est connue, la variable Y="rapport $\frac{n\,v}{\sigma^2}$ de n fois la variance $v = \frac{1}{n}\sum (x_i - m)^2$ d'un échantillon de taille n sur la variance de la population mère" suit une loi du χ^2(n)
- Si la moyenne m est inconnue, la variable Y="rapport $\frac{n\,V_e}{\sigma^2} = \frac{n\,s^2}{\sigma^2}$ de n fois la variance d'un échantillon de taille n sur la variance de la population mère" suit une loi du χ^2(n-1)

Remarque : dans le second cas le nombre de degrés de liberté est égal au nombre d'observations moins 1 car on a perdu un degré au moment d'estimer la moyenne.

On considère l'hypothèse H0 : $s^2 = \sigma^2$ fixé. Dans chaque cas, on compare la valeur expérimentale de la variable Y avec la valeur obtenue avec l'hypothèse H0 de la valeur de la variance. Si la valeur expérimentale est située dans une zone où le risque de refus est inférieur à la valeur α donnée, on peut rejeter H0, sinon on ne peut rejeter H0 et on accepte H1.

Exemple: On étudie la dispersion des dépenses hebdomadaires des étudiantes de l'IUT B. Pour cela, on sélectionne un échantillon aléatoire de 20 étudiants et on obtient les résultats suivants:
Moyenne : 153.75 ; Variance : $s^2 = 812.83$
On test H0: $s^2 = \sigma^2 = 625$ et H1: $s^2 > \sigma^2 = 625$
Comme la moyenne est inconnue, la variable $\frac{nV_e}{\sigma^2} = \frac{n\,s^2}{\sigma^2}$ suit une loi du χ^2(n-1).
La valeur expérimentale de cette variable est $20\frac{813}{625} = 26$. La valeur limite pour un risque de 5% est donnée par la valeur de x telle que $P(\chi^2(19) > x) = 0{,}05$, o trouve x = 30,14



Comme la valeur critique est à l'intérieur de la région d'acceptation, on ne rejette pas l'hypothèse nulle. Ce qui signifie que l'hypothèse de variance de 625 de la moyenne des dépenses hebdomadaires ne peut être rejetée.

7 Comparaison de deux variances

7.1 Comparaison de deux variances d'échantillons (fisher-Snedecor)

Dans le chapitre sur l'estimation, on a vu que l'estimation de la variance suit une loi du χ^2. La variable construite comme la différence de la variance de deux échantillons ne prend pas seulement des valeurs positives, elle ne correspond donc pas à une loi du χ^2.

Pour tester l'égalité de deux variances, on s'intéresse **au rapport des deux variances** et en vérifiant que ce rapport ne dépasse pas une certaine valeur théorique donnée dans la table de la loi de Fisher.

Soit X_1 et X_2 deux v.a.r. défini sur une population mère **suivant deux lois normales** $\mathcal{N}(m_1, \sigma^2_1)$ de moyenne m_1 et d'écart type σ_1 et $\mathcal{N}(m_2, \sigma^2_2)$ de moyenne m_2 et d'écart type σ_2

Loi de Fisher (ou Fisher-Snedecor) : La variable Z définie par $Z = \dfrac{Y_1 / \nu_1}{Y_2 / \nu_2}$ ="rapport d'une variable suivant une loi $\chi^2(\nu_1)$ et celle suivant une loi $\chi^2(\nu_2)$ divisées par leur degrés de liberté respectifs " suit une loi de Fischer $F_{\nu 1,\nu 2}$. On parle aussi de Loi de Fisher comme celle du rapport de deux χ^2 normalisés. Par convention on place au numérateur la variable possédant la plus grande variance afin de travailler avec une table de Fisher pour des valeurs supérieures à 1.

Mis en œuvre du test

On étudie deux populations de variances σ_1^2 et σ_2^2 inconnues.

On dispose de deux échantillons aléatoires, tirés de façon indépendante dans chacune des populations. La distribution de la variable, dans chacune des populations, suit une loi Normale sinon l'effectif de chaque échantillon doit au moins être égal à 30.

Les hypothèses

$H_0 : \sigma_1^2 / \sigma_2^2 = 1$

$H_1 : \sigma_1^2 / \sigma_2^2 \neq 1$ (bilatéral) ou $\sigma_1^2 / \sigma_2^2 > 1$ (unilatéral).

- Variable aléatoire étudiée : $\hat{\sigma}_1^2 / \hat{\sigma}_2^2$ (on porte par principe toujours au numérateur la plus forte des variances estimées).

  Loi de la v.a : Loi F (ν_1 ; ν_2) avec $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$
- Critère statistique calculé : $F_{obs} = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} > 1$
- Critère statistique théorique : $F_{1-\alpha/2}$, car comme on a posé **par principe** F_{obs} comme le rapport de la plus grande sur la plus petite variance, le test devient unilatéral.
- Conclusion : Si $F_{obs} < F_{théo}$ on conserve H_0.

Exemple: on veut comparer les dispersions des dépenses hebdomadaires des étudiants des départements GEii et TechCo. Pour cela on suppose que les dépenses suivent des lois Normales. On

sélectionne deux échantillons aléatoires de 20 et 30 étudiants respectivement et obtient les réponses suivantes:

	N	moyenne	$\hat{\sigma}^2 = s^2$
TechCo (1)	30	145.27	985.93
GEii (2)	20	153.75	812.83

L'hypothèse nulle (l'hypothèse à tester) et l'alternative sont les suivantes:

H0: σ^2 TechCo = σ^2 GEii, soit $\sigma_1^2 / \sigma_2^2 = 1$

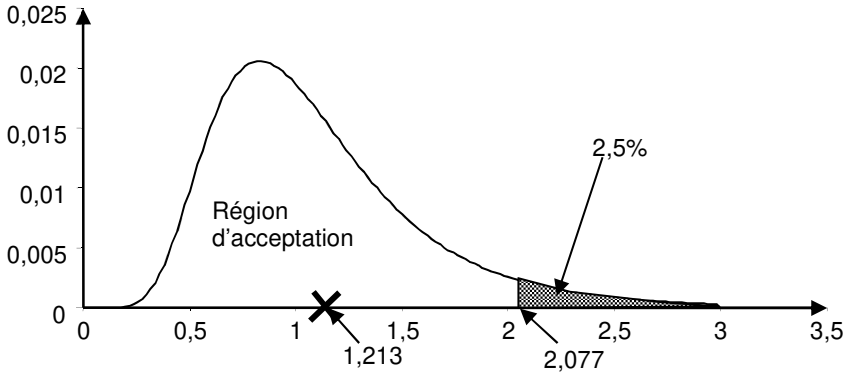
H1: σ^2 TechCo > σ^2 GEii, soit $\sigma_1^2 / \sigma_2^2 > 1$

où σ^2 représente la vraie variance des dépenses hebdomadaires des étudiants.

Le critère statistique calculé : $F_{obs} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$

où $\hat{\sigma}^2$ ou s^2 sont les variances estimées à partir des échantillons.

- La valeur calculée est : $F_{obs} = 1{,}213$
- On sélectionne un seuil de signification, par exemple, 5%
- Le nombre de degrés de liberté du numérateur et du dénominateur sont 29 et 19 respectivement
- La valeur critique est donc : 2.402 (avec Excel INVERSE.LOI.F(0,025 ;29;19))
- la valeur observée (1.213) est inférieure à la valeur critique (2.402), on ne rejette pas l'hypothèse nulle. Ce qui signifie que les dispersions moyennes à la moyenne des dépenses hebdomadaires ne sont pas différentes.



7.2 Comparaison de plusieurs variances : test de Bartlett

Ce test est utilisé lorsque l'on doit vérifier l'homogénéité des variances intra groupes de plusieurs séries statistiques.

On étudie k populations de variances σ_1^2, σ_2^2,…σ_i^2,.....σ_k^2 inconnues.

On dispose de k échantillons aléatoires, tirés de façon indépendante dans chacune des populations.

La distribution de la variable, dans chacune des populations, suit une loi Normale et aucune des variances empiriques n'est nulle ni trop petite.

♦ Les hypothèses

H_0 : les variances sont homogènes.

H_1 : au moins une variance est supérieure aux autres.

♦ Critère statistique calculé : $X^2 \text{ calc} = \nu \text{ Ln } \hat{\sigma}^2 - \sum_{i=1}^{i=k} (\nu_i \text{Ln} \hat{\sigma}_i^2)$

$\nu_i = n_i - 1$ $\quad \nu = \sum_{i=1}^{i=k} \nu_i$ $\quad \hat{\sigma^2}_i = \frac{SCE_{xij}(\overline{x_i})}{n_i - 1}$ $\quad \hat{\sigma^2} = \frac{1}{\nu} \sum_{i=1}^{i=k} \nu_i \hat{\sigma_i^2}$

Avec $\hat{\sigma^2}_i$ l'estimation de la variance σ_i^2, $SCE_{xij}(\overline{x_i})$: Somme des Carré des Ecart des x_{ij} par rapport à $\overline{x_i}$

♦ Critère statistique théorique : $X^2_{1-\alpha}$ (k-1)

♦ Conclusion : X^2 calc < X^2 théo on ne peut pas mettre en évidence qu'au moins une variance est supérieure aux autres, on garde l'hypothèse de l'homogénéité des variances.

Ce test est utilisé dans le chapitre sur l'analyse de la variance ANOVA.

8 Estimation et test pour la régression linéaire

8.1 Estimation par intervalle de confiance des paramètres de la droite

Pour N points de mesure (x_i, y_i), l'estimation de la variance du résidu est donnée par : $\hat{\sigma}_E^2 = \frac{1}{N-2} \sum_{i=1}^{N} e_i^2$.

L'estimation de la variance du coefficient a de la droite est donnée par :

$$\hat{\sigma}_a^2 = \frac{\hat{\sigma}_e^2}{SS_X} = \frac{\sum_{i=1}^{N} e_i^2}{(N-2)\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

Le coefficient $\frac{a - \hat{a}}{\hat{\sigma}_a}$ suit une loi de Student à N-2 degrés de libertés ($\hat{a} = \frac{s_{XY}}{s_X^2}$).

L'estimation de la variance du coefficient b de la droite est donnée par :

$$\hat{\sigma}_b^2 = \hat{\sigma}_e^2 \left[\frac{1}{N} + \frac{\overline{x}^2}{SS_X} \right] = \frac{\sum_{i=1}^{N} e_i^2}{(N-2)} \left[\frac{1}{N} + \frac{\overline{x}^2}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right] = \frac{\left(\sum_{i=1}^{N} e_i^2 \right) \left(\sum_{i=1}^{N} x_i^2 \right)}{(N-2) N \sum_{i=1}^{N} (x_i - \overline{x})^2}$$

Le coefficient $\frac{b - \hat{b}}{\hat{\sigma}_b}$ suit une loi de Student à N-2 degrés de libertés ($\hat{b} = \overline{y} - \hat{a}\,\overline{x}$).

8.2 Estimation par intervalle de confiance des points de la droite de régression

Pour une valeur x_0 fixée, les points de la droite de régression ont pour coordonnées (x_0, $(x_0, \hat{y}_0) = (x_0, a\,x_0 + b)$. L'estimation $\hat{y}_0$ suit une loi de Student à N-2 degrès de libertés de variance :

$$\hat{\sigma}_y^2 = \hat{\sigma}_e^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \overline{x})^2}{SS_X} \right]$$

8.3 Intervalle de confiance de l'espérance de nouveaux points de mesures issus d'une même population

On veut estimer la valeur Y **moyenne (espérance)** pour une sous-population possédant la valeur x_0 fixée. Cette valeur **moyenne** est estimé exactement de la même façon qu'au paragraphe précédent en utilisant l'équation de la droite de régression $\hat{y}_0 = a\,x_0 + b$. L'intervalle de confiance de cette espérance est plus réduit que celle obtenue pour une seule mesure, comme dans le paragraphe précédent.

On a encore $\hat{y}_0$ qui suit une loi de Student à N-2 degrès de libertés de variance :

$$\hat{\sigma}_y^2 = \hat{\sigma}_e^2 \left[\frac{1}{N} + \frac{(x_0 - \overline{x})^2}{SS_X} \right]$$

8.4 Test sur la validité de la régression

8.4.1 Test de non nullité de la pente

Pour tester la significativité de la régression, on teste si la pente est significativement différente de zéro, ce qui se produit si la covariance est nulle entre X et Y. Pour cela, on lit dans la table de Student à (N-2) degrés de libertés, la valeur $t_{\alpha/2}$ au risque α bilatéral d'avoir une valeur supérieure à $t_{\alpha/2}$. Si la valeur expérimentale $\frac{a}{\hat{\sigma}_a}$ est supérieure à $t_{\alpha/2}$ alors on peut dire que a est significativement différent de zéro, avec un risque α de se tromper.

Excel donne directement p, la probabilité de se tromper en disant que le coefficient est différent de zéro, compte tenu de la variance de ce coefficient.

8.4.2 Test de Fischer du rapport des variances

On teste si le rapport de la variance expliquée sur la variance résidu $F = \frac{s_M^2}{s_E^2}$, est significativement supérieure à 1. Pour cela on utilise *l'estimation* de chacun des termes :

$$F = \frac{s_M^2}{s_E^2} = \frac{SS_M}{\frac{SS_E}{N-2}} = (N-2)\frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} e_i^{\,2}}$$

F suit une loi de Fischer-Snedecor à 1 et (N-2) degrès de libertés, loi F(1 ;N-2).
Si la valeur expérimentale de F est supérieur à la valeur de la table au risque α, $F_\alpha(1 ;N-2)$, on rejette l'hypothèse que le rapport F est égal à 1. ;

Remarque : on peut aussi écrire F en fonction de r^2 : $F = \frac{r^2}{(1-r^2)/(N-2)}$. Certain logiciel donne directement la probabilité p de se tromper en disant que F est significativement supérieur à 1.

Exemple : On trouve $r^2 = 0.82$ avec une régression sur 12 points : F = 45.5. La table donne $F_{0,05}(1 ;10) = 4,96$. Le rapport des variances est significativement supérieur à 1, au risque 5% de se tromper. En pratique, on considère que F doit être au moins quatre fois supérieur au F théorique. C'est le cas ici : 45.5 > 4 x 4,96.

8.4.3 Comparaison du deux droites de régression expérimentales

Soit Y et Y' deux variables aléatoires liés à la même variable X connues respectivement sur n et n' points. On test ici l'égalité entre les pentes des régressions respectives a et a'.
On suppose que les variances des résidus $\hat{\sigma}_E^{\,2}$ et $\hat{\sigma}'_E{}^2$ sont égales, ce qui est l'objet d'un test préalable (voir pour cela le paragraphe 7.1 Comparaison de deux variances d'échantillons utilisant la loi de Snédécor).
Si cela est vérifié, on estime la variance commune par :

$$\hat{\sigma}_{EC}^{\,2} = \frac{(n-2)\hat{\sigma}_E^{\,2} + (n'-2)\hat{\sigma}'_E{}^2}{n+n'-4}$$

Alors sous l'hypothèse H0 d'égalité des pentes a=a', la variable T :

$$T = \frac{a-a'}{\hat{\sigma}_{EC}^{\,2}\sqrt{\frac{1}{ns_X^2} + \frac{1}{n's_{X'}^2}}}$$; où s_X^2 est la variance expérimentale des x_i. $s_X^2 = \frac{\sum(x_i - \bar{x})^2}{n} =$

suit une loi de Student à n+n'-4 degrès de libertés.

9 Estimation et test du coefficient de corrélation

9.1 Estimation par intervalle de confiance du coefficient de corrélation

Soit ρ le coefficient de corrélation entre les variables X et Y. Soit r le coefficient de corrélation expérimentale entre les mesure x_i et y_i : $r = \frac{s_{XY}}{\sqrt{s_X^2}\sqrt{s_Y^2}} = \frac{s_{XY}}{s_X\, s_Y}$.

Soit Z la variable aléatoire qui prend la valeur z : $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \arctan h(r)$.

Pour n assez grand (>20) Z suit une loi Normale $\mathcal{N}\left(\zeta, \frac{1}{n-3}\right)$, où ζ est le nombre défini par :

$\zeta = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \arctan h(\rho)$.

On peut ainsi construire un intervalle de confiance $[z_1; z_2]$ de ζ puis en utilisant la fonction tanh() un intervalle $[r_1; r_2] = [\tanh(z_1); \tanh(z_2)]$ de ρ.

9.2 Comparaison du coefficient de corrélation à une valeur théorique

9.2.1 Comparaison à zéro

Dans ce cas, si H0 est vérifiée (ρ=0) , la variable aléatoire T qui prend pour valeur :

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

suit une loi de Student à n-2 degrés de libertés. On peut ainsi construire un intervalle de confiance de t et rejeter H0 si la valeur expérimentale de t et hors de cet intervalle.

9.2.2 Comparaison à un valeur non nulle

Hypothèse H0 : $\rho = \rho_0$.

On calcule $\zeta_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = \arctan h(\rho_0)$

Soit Z la variable aléatoire qui prend la valeur z : $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \arctan h(r)$. Sous H0, elle suit une loi Normale $\mathcal{N}\left(\zeta_0, \frac{1}{n-3}\right)$. On peut donc tester la valeur $(z - \zeta_0)\sqrt{n-3}$ par rapport à la loi $\mathcal{N}(0,1)$.

9.3 Comparaison de deux coefficients de corrélations expérimentaux

Soit ρ et ρ' deux coefficients de corrélation inconnus et supposés égaux sous H0.
Soit r et r' les deux coefficients de corrélation expérimentaux.

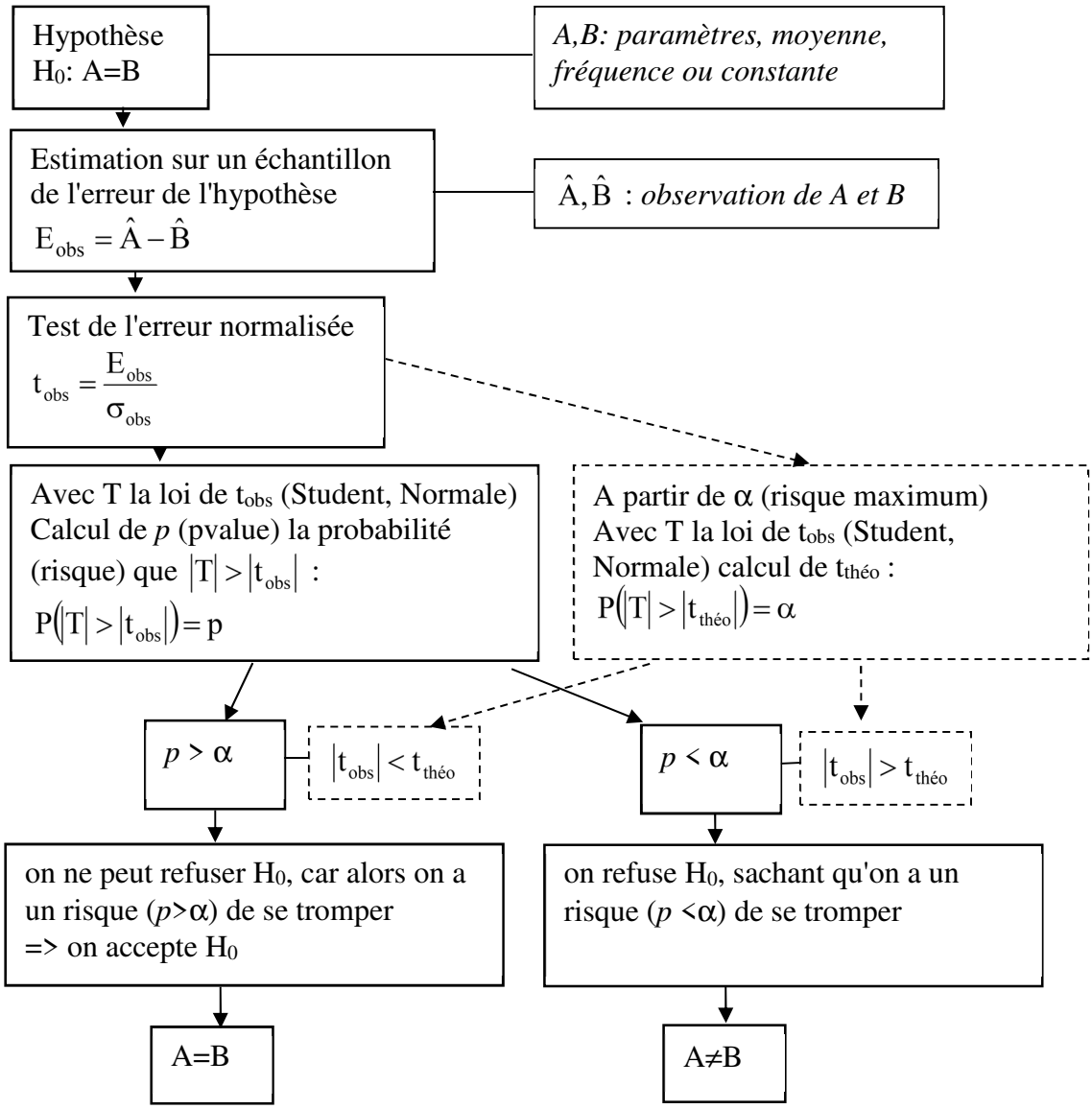
Soit $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \arctan h(r)$ et $z' = \frac{1}{2}\ln\left(\frac{1+r'}{1-r'}\right) = \arctan h(r')$, Z et Z' leur variable aléatoire correspondante.

Sous H0 et si n et n' sont assez grand (>20), la variable U qui prend le valeur u :

$$u = \frac{z - z'}{\sqrt{\frac{1}{n-3} + \frac{1}{n'-3}}}$$

suit une loi $\mathcal{N}(0,1)$. Le test de H0 en résulte directement.

Résumé de la démarche d'un test



Remarque sur la terminologie anglo-saxonne : T-test et P value

T-test

The **t-test** is the most commonly used method to evaluate the differences in means between two groups. The groups can be independent (e.g., blood pressure of patients who were given a drug vs. a control group who received a placebo) or dependent (e.g., blood pressure of patients "before" vs. "after" they received a drug, see below). Theoretically, the t-test can be used even if the sample sizes are very small (e.g., as small as 10; some researchers claim that even smaller n's are possible), as long as the variables are approximately normally distributed and the variation of scores in the two groups are not too different.

P value

The p value is the probability of having observed our data (or more extreme data) when the null hypothesis is true. In other words, if the null hypothesis is true, the p value gives the probability of observing our data (or more extreme) by chance, so it can be thought of as a measure of strength in the belief of the null hypothesis. To illustrate: we sample a classroom of 30 children to test the null hypothesis that the population of boys and girls are on average of equal heights. A p value of 0.01 suggests that the probability of collecting the observed heights of the 30 children (or with a greater height difference between the boys and girls) is 0.01 when the overall population of boys and girls are truly of equal height.
p-value : probabilité que sous l'hypothèse H_0 on observe de telles mesures. Si la p-value est grande, on accepte H_0, si la p-value est faible on refuse H_0.

Site permettant de réaliser les test avec importation possible de données au format Excel
http://www.u707.jussieu.fr/biostatgv/index.html

Comparaison de la répartition d'une population : test du χ^2

1 Généralités définitions

Les chapitres précédents portaient sur des caractères quantitatifs, nous allons considérer ici un caractère qualitatif à k classes.

Pour les différentes classes, on observe des effectifs o_1, ... , o_k. On fait une hypothèse sur la répartition supposée de la population suivant les classes. Elle conduit à des effectifs théoriques calculés t_1, ... , t_k de la population de chaque classe, différents des effectifs observés. On étudie l'écart entre les effectifs calculés et observés pour en déduire, avec un certain risque α, si l'hypothèse sur la répartition est valable.

Exemple: on désire savoir si le tabac a une influence sur l'apparition d'un cancer. Pour cela, on considère deux populations, celle des fumeurs et celle des non-fumeurs et on étudie la proportion d'apparition du cancer dans chacune des deux populations. La répartition cancer/pas de cancer de la population des non-fumeurs est la répartition de référence et on étudie si la répartition de la population des fumeurs est comparable à celle des non-fumeurs

Attention:
- le test du χ^2 s'applique aux **effectifs** des diverses catégories et non aux pourcentage.
- le test du χ^2 ne peut être utilisé que si **tous les effectifs calculés** sont suffisamment **grands**: il faut qu'ils atteignent ou dépassent **5**.

Nombre de degrés de liberté: l'effectif de la dernière classe est égale à l'effectif total N moins la somme des effectifs des autres classes. La répartition est parfaitement définie avec les effectifs de (k-1) classes. (k-1) est le **nombre de degrés de liberté** (ddl) du test.

2 Loi et table du χ^2

On construit la variable d'écart entre la répartition calculée et la répartition observée.

Classe	1	2	...	k
Proportion théorique	p_1	p_2	...	p_k
Effectif théorique calculé	t_1	t_2	...	t_k
Effectif observé	o_1	o_2	...	o_k
Ecart	$o_1 - t_1$	$o_2 - t_2$	...	$o_k - t_k$
Carré écart normalisé	$\frac{(o_1 - t_1)^2}{t_1}$	$\frac{(o_2 - t_2)^2}{t_2}$	...	$\frac{(o_k - t_k)^2}{t_k}$

Pour une population totale N, l'effectif calculé t_1 vaut p_1 N.

On ne saurait utiliser comme indice la somme (ou la moyenne) des écarts à cause de la compensation des écarts positifs et négatifs. La somme de leur valeur absolue ne se prête pas commodément à des calculs de probabilité. La somme des carrés des écarts évite les inconvénients ci-dessus, cependant il donne le même poids à tous les écarts, qu'ils se rapportent à de petits ou grands effectifs. Ces considérations on conduit à adopter l'indice suivant, dit du χ^2.

$$\chi^2 = \frac{(o_1 - t_1)^2}{t_1} + \frac{(o_2 - t_2)^2}{t_2} + ... + \frac{(o_k - t_k)^2}{t_k} = \sum_{i=1}^{k} \frac{(o_i - t_i)^2}{t_i}$$

Le χ^2 suit une loi $\mathcal{X}^2$(k-1) dite du Khi-deux à (k-1) degrés de liberté (ddl).

La fonction de répartition de la loi $\mathcal{X}^2$(k-1) est tabulée.

3 Comparaison d'une répartition à celle d'une loi théorique: test du χ^2

Pour comparer la répartition observée à la répartition théorique d'un caractère qualitatif à k classes, on forme $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - t_i)^2}{t_i}$ et on cherche le risque correspondant α (risque de se tromper si on refuse la conformité des deux répartitions) dans la table du $\mathcal{X}^2$(k-1) dite du Khi-deux à (k-1) degrés de liberté.

- si $\alpha > 5\%$ la différence n'est pas significative
- si $\alpha \leq 5\%$ la différence est significative et le risque correspondant à la valeur du χ^2 mesure le degrés de signification

NB: la méthode n'est valable que si tous les effectifs calculés égalent ou dépassent 5.

Exemple: on reprend l'exemple du cancer. On sait que sur une population de non-fumeur 15% des personnes développent un cancer. Sur une population de 35 fumeurs, 23 ont développé un cancer. L'apparition des cancers dans la population fumeur est-elle conforme à celle de la population non-fumeur?

Classe	cancer	pas de cancer
Proportion théorique (non-fumeur)	15%	85%
Effectif théorique calculé (non-fumeur)	5,25	29,75
Effectif observé (fumeur)	23	12
Ecart	17,75	-17,75
Carré écart normalisé	60,01	10,59

χ^2=70,6; on a deux classes, on regarde donc la table pour la ligne ddl=υ=*1*
70,6 correspond à une probabilité supérieure à 0,9995, soit un risque inférieure à 0,05%. On peut rejeter l'hypothèse que la répartition observée soit identique à la répartition théorique avec un risque d'erreur inférieur à 0,05%.

Avec Excel ou OpenOffice_Calc, la fonction LOI.KHIDEUX (t_α;n) donne la probabilité α :
$P(\chi^2 > t_\alpha) = \alpha$ pour une loi $\mathcal{X}^2$(n) : LOI.KHIDEUX (70,6;1)= 4 ;4 10^{-17} , proche de zéro.

Exemple: ce test est souvent utilisé pour vérifier si une répartition suit une loi Gaussienne $\mathcal{N}(m,\sigma)$ ou autre. Si la moyenne et l'écart type de la distribution sont calculés à partir des données, il convient de prendre (k-3) degrés de liberté si on travaille avec k classes.

4 Test d'indépendance de deux caractères qualitatif par le test du χ^2

On considère maintenant deux caractères qualitatifs X_1 et X_2, qui donne deux répartitions observées sur une même population. On ne connaît la loi théorique d'aucun des deux caractères, on veut juste savoir si ces deux répartitions indiquent que les deux caractères sont liés entre eux ou indépendants.

Le but est ici de prouver l'indépendance des deux caractères, pour cela on fait l'hypothèse d'indépendance et on calcule le risque α de dire que cette hypothèse est fausse, c'est à dire que les deux caractères sont indépendants, alors qu'ils ne sont pas indépendants.

On est amené à construire le tableau de contingence. Pour un caractère X_1 à k_1 classes et un caractère X_2 à k_2 classes, ce tableau comporte k_1 lignes et k_2 colonnes. On porte à l'intersection de la ligne i et de la colonne j, l'effectif de la population $o_{i,j}$ qui présente les caractères i et j.
Dans une ligne supplémentaire et une colonne supplémentaire, on fait la somme des effectifs de chaque ligne l_i et de chaque colonne c_i. Ainsi, à l'intersection de cette ligne et colonne supplémentaire, on retrouve l'effectif total de la population.

X_1 \ X_2	1	2	...	k_2	Total
1	$o_{1,1}$	$o_{1,2}$	...	$o_{1,k2}$	l_1
2	$o_{2,1}$	$o_{2,1}$	...	$o_{2,k2}$	l_2
...	...	...	...	...	...
k_1	$o_{k1,1}$	$o_{k1,2}$	...	$o_{k1,k2}$	l_{k1}
Total	c_1	c_2	...	c_{k2}	N

Nombre de degrés de liberté: pour chaque ligne et chaque colonne, l'effectif de la dernière classe est égale à l'effectif total de la ligne/colonne moins la somme des effectifs des autres classes. La répartition est parfaitement définie avec les effectifs de $(k_1\text{-}1)(k_2\text{-}1)$ classes. $(k_1\text{-}1)(k_2\text{-}1)$ est le **nombre de degrés de liberté** (ddl) du test d'indépendance.

On construit ensuite le tableau de contingence théorique $t_{i,j}$ correspondant à l'hypothèse d'indépendance des deux caractères. On sait que pour deux évènements A et B indépendants $P(A \cap B) = P(A)\,P(B)$. Donc, $P((X_1 = i) \cap (X_2 = j) = P(X_1 = i)\,P(X_2 = j) = \frac{c_i}{N}\frac{l_j}{N}$. Ce qui donne pour la population de la case i,j: $l_j c_i/N$.
En chaque case on fait le produit du total de sa ligne par le total de sa colonne divisé par N, l'effectif total de la population.

X_1 \ X_2	1	2	...	k_2	Total
1	$t_{1,1}=l_1c_1/N$	$t_{1,2}=l_1c_2/N$	...	$t_{1,1}=l_1c_{k2}/N$	l_1
2	$t_{2,1}=l_2c_1/N$	$t_{2,2}=l_2c_2/N$	...	$t_{2,1}=l_2c_{k2}/N$	l_2
...	...	...	...	...	...
k_1	$t_{k1,1}=l_{k2}c_1/N$	$t_{k1,2}=l_{k2}c_2/N$	...	$t_{k1,k2}=l_{k1}c_{k2}/N$	l_{k1}
Total	c_1	c_2	...	c_{k2}	N

Pour comparer la répartition observée à la répartition théorique d'indépendance, on forme $\chi^2 = \sum_{i,j=1}^{k_1,k_2} \frac{(o_{i,j} - t_{i,j})^2}{t_{i,j}}$ et on cherche le risque correspondant α (risque de se tromper si on refuse l'indépendance) dans la table du $\mathcal{X}^2((k_1\text{-}1)(k_2\text{-}1))$ dite du Khi-deux à $(k_1\text{-}1)(k_2\text{-}1)$ degrés de liberté.

- si $\alpha > 5\%$ la différence n'est pas significative
- si $\alpha \leq 5\%$ la différence est significative et le risque correspondant à la valeur du χ^2 mesure le degrés de signification

NB: la méthode n'est valable que si tous les effectifs calculés égalent ou dépassent 5.

Coefficient de contingence : on définit le coefficient de contingence C par : $C = \sqrt{\frac{\chi^2}{N + \chi^2}}$. C vaut 0 en cas d'indépendance, mais sa valeur dépend du nombre N d'individus. Il n'est possible de comparer deux coefficients de contingence que s'ils proviennent de tables de contingence de même taille. Son intérêt est donc plus limité que celui du khi 2.

Exemple: on reprend l'exemple du cancer. Un échantillon de 350 personnes se répartie de la manière suivante

Classe	cancer	pas de cancer	Total
Non-fumeur	47/*63*	268/*252*	315
Fumeur	23/*7*	12/*28*	35
Total	70	280	350

On porte en *italique* sur le même tableau les effectifs correspondants à l'hypothèse d'indépendance entre "être fumeur" et "avoir un cancer". On calcule le χ^2:
$\chi^2=(47-63)^2/63+(23-7)^2/7+(268-252)^2/252+(12-28)^2/28=50,79$
On regarde la table pour la ligne ddl=υ=(2-1)(2-1)=1
50,79 correspond à une probabilité supérieure à 0,9995, soit un risque inférieure à 0,05%. On peut rejeter l'hypothèse d'indépendance entre "être fumeur" et "avoir un cancer" avec un risque d'erreur inférieur à 0,05%.
Le résultat diffère légèrement de celui trouvé au paragraphe 3, car ici la proportion de cancer sur les non-fumeurs est obtenue sur un échantillon de 315 personnes.

Remarque: si on veut comparer des données qui sont sous forme de proportion en %, on dit que la population totale est de 100 (ou 1000) suivant la précision des données.

5 Autre test d'adéquation à une loi théorique : test de Kolmogorov et Smirnov

Le test de Kolmogorov-Smirnov est un test d'hypothèse utilisé pour déterminer si un échantillon suit bien une loi donnée connue par sa fonction de répartition continue, ou bien si deux échantillons suivent la même loi.

Le test de Kolmogorov-Smirnov compare la fonction de répartition théorique F_0 avec la fonction de répartition de l'échantillon empirique. On calcule la distance maximale entre les fonctions théoriques et empiriques. Si cette distance dépasse une certaine valeur, on dira que l'échantillon est mauvais.

Le test de Kolmogorov-Smirnov s'étend à la **comparaison de deux fonctions de répartition** empiriques.

Mise en œuvre :

Soit $\{k_i\}$ les n valeurs prises par une loi discrète X (ou les bornes d'intervalles d'une loi continue), la fonction de répartition expérimentale de X est définie par : $F_e(k_i) = P(X \le k_i)$. En pratique $P(X \le k_i)$ représente le nombre de mesure de valeur inférieure à k_i divisé par le nombre total n de mesure. Soit $F_0(k_i)$ les valeurs de la loi théorique à laquelle on veut comparer la loi expérimentale $F_0(k_i) = P(X_0 < k_i)$ où X_0 est une variable suivant la loi théorique considéré.
On mesure la distance entre les deux fonctions de répartition par :

$$D_{KS}(F_0, F_e) = \max_i \left\{ \left|F_o(k_i) - F_e(k_i)\right|, \left|F_0(k_i) - F_e(k_{i-1})\right| \right\}$$

Sous l'hypothèse H_0 d'égalité des lois F_e et F_0 la distance D_{KS} vérifie la propriété :

$$P\left[D_{KS}(F_A, F_B)_n \ge t\right] = 2 \sum_{k=0}^{Int\left[n(1-t)\right]} C_n^k \, t \left(t + \frac{k}{n}\right)^{k-1} \left(1 - t - \frac{k}{n}\right)^{n-k}$$

Où $Int\left[n(1-t)\right]$ est la partie entière de $n(1-t)$. Pour des valeurs de n>100 on a l'approximation :

$$P\left[D_{KS}(F_A, F_B)_n \ge \frac{t}{\sqrt{n}} \right] \approx 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 t^2)$$

Ces formules sont facilement programmables en Basic, C, Scilab, … . Voir en Annexe le code Scilab.
Le test est nécessairement unilatéral à droite (rejet des valeurs trop grandes).

Utilisation pour tester la normalité de mesures:

Soit l'ensemble de mesures : {482 ;489 ;490 ;491 ;491 ;493 ;497 ;499 ;499 ;509}

Ces mesures suivent-elles un loi normale $\mathcal{N}(494\ ;7{,}3636)$?

Valeurs triées	Effectifs	Effectifs cumulés	z	$F_e(k_i)$	$F_0(k_i)$	$\lvert F_0(k_i) - F_e(k_{i-1})\rvert$	$\lvert F_0(k_i) - F_e(k_i)\rvert$
k_i	n_i	n_i^+	$\frac{k_i - \overline{x}}{\sigma}$	$P(X \le k_i) = f_i^+$	$P(T < z_i)$	Ecart à gauche	Ecart à droite
482	1	1	-1,63	0,1000	0,0516	0,0516	0,0484
489	1	2	-0,68	0,2000	0,2486	0,1486	0,0486
490	1	3	-0,54	0,3000	0,2935	0,0935	0,0065
491	2	5	-0,41	0,5000	0,3419	0,0419	**0,1581**
493	1	6	-0,14	0,6000	0,4460	0,0540	0,1540
497	1	7	0,41	0,7000	0,6581	0,0581	0,0419
499	2	9	0,68	0,9000	0,7514	0,0514	0,1486
509	1	10	2,04	1,0000	0,9792	0,0792	0,0208

On a 8 points de la fonction de répartition, l'écart maximum est de 0,1581.

$$P[D_{KS}(F_0, F_e)_8 \ge 0{,}1581] = 1 > 0{,}05$$

Avec 8 points de la fonction de répartition, on est sur de trouver un écart max supérieur à 0,1581. L'écart est donc trop faible pour dire que les mesures ne suivent pas une loi Normale.

On accepte H0, les mesures suivent un loi normale $\mathcal{N}(494\ ;7{,}3636)$?

Il faudrait une écart $D_{KS}(F_0, F_e)_8 > 0{,}4543$ pour rejeter H0 avec un risque de 5%.

Utilisation pour comparer deux séries de mesures :

On veut savoir si les groupes de mesures A={ 4,6;4,7;4,9;5,1;5,2;5,5;5,8;6,1;6,5;6,5;7,2} de n_1=11 éléments et B={5,1;5,3;5,4;5,6;6,2;6,3;6,8;7,7;8;8,1} de n_1=10 éléments sont significativement différentes

On classe toutes les mesures et on somme les effectifs.

k_i	Prov	Cumul A $F_A(k_i)$	Cumul B $F_B(k_i)$	$\lvert F_A(k_i) - F_B(k_{i-1})\rvert$	$\lvert F_A(k_i) - F_B(k_i)\rvert$
4,6	A	0,0909	0	0,0909	0,0909
4,7	A	0,1818	0	0,1818	0,1818
4,9	A	0,2727	0	0,2727	0,2727
5,1	A/B	0,3636	0,1	0,3636	0,2636
5,2	A	0,4545	0,1	0,3545	0,3545
5,3	B	0,4545	0,2	0,3545	0,2545
5,4	B	0,4545	0,3	0,2545	0,1545
5,5	A	0,5455	0,3	0,2455	0,2455
5,6	B	0,5455	0,4	0,2455	0,1455
5,8	A	0,6364	0,4	0,2364	0,2364
6,1	A	0,7273	0,4	0,3273	0,3273
6,2	B	0,7273	0,5	0,3273	0,2273
6,3	B	0,7273	0,6	0,2273	0,1273
6,5	A/A	0,9091	0,6	0,3091	0,3091
6,8	B	0,9091	0,7	0,3091	0,2091
7,2	A	1,0000	0,7	0,3000	0,3000
7,7	B	1,0000	0,8	0,3000	0,2000
8	B	1,0000	0,9	0,2000	0,1000
8,1	B	1,0000	1	0,1000	0,0000

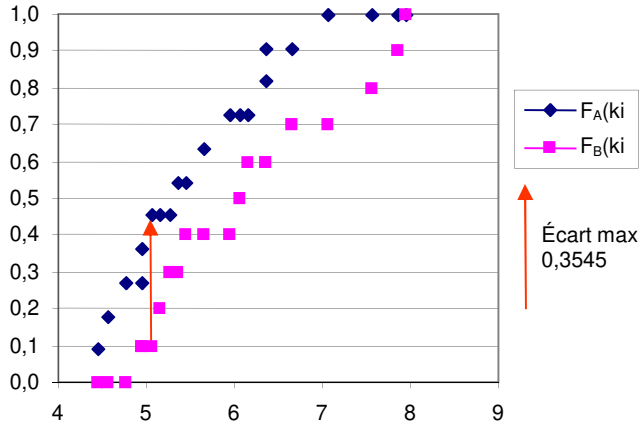


fig 5-5: Fonctions de répartition des données A et B, mesure de l'écart maximum

On a 19 points des fonctions de répartition, l'écart maximum est de 0,3545.

$P\left[D_{KS}(F_A, F_B)_{19} \geq 0{,}3545\right] = 0{,}0122 < 0{,}02$

L'écart est donc trop grand pour dire que les mesures sont issues d'une même population.
On refuse H0, les mesures sont significativement différentes avec un risque < 2%
Il faudrait un écart $D_{KS}(F_A, F_B)_{19} < 0{,}301$ pour ne pas rejeter H0 avec un risque de 5%.

Remarque : les valeurs où sont estimées les fonctions de distribution correspondent ici aux valeurs expérimentales k_i et ne sont donc, en général, pas régulièrement réparties. La valeur de l'écart entre les deux fonctions est donc biaisée. Certains auteurs préconisent de calculer les deux fonctions de distributions en choisissant de manière régulière les valeurs où on les calcule.

Théorie des files d'attente

Les théories des files d'attente sont des outils probabilistes permettant de prendre en compte et de modéliser les goulots d'étranglement au niveau de la logistique, des centrales téléphoniques, des requêtes sur les serveurs, des caisses de grands magasins ou encore dans les toilettes des grands stades sportifs (...) en fonction des hypothèses et contraintes de départ.

Que ce soit pour le client ou pour une entreprise l'attente est une activité sans valeur ajoutée. Elle correspond aussi bien à l'attente d'un client ou utilisateur que celle des employés ou des équipements inoccupés. L'analyse et la minimisation des files d'attente relève donc d'une importance stratégique. Ces théories se révèlent notamment utiles pour justifier des investissements, des embauches ou des achats d'équipements.

La problématique type peut s'exprimer ainsi:

- Quel est le nombre optimal de services (stations/terminaux) pour traiter la demande tout en évitant une file d'attente trop importante et le rejet de certaines requêtes?
- Quel est le temps d'attente moyen d'une requête?
- Quel est le nombre moyen de requêtes en attente dans la file?

1 Présentation du problème

1.1 Notation de Kendall

Une notation a été développée par Kendall pour représenter les files d'attentes. La forme réduite de cette notation est:

$$A/B/C/D$$

où *A* représente le processus d'arrivée des clients dans le système, *B* représente la distribution des services des clients du système, *C* le nombre de serveurs du système et D la longueur maximale de la file d'attente.

Par exemple, la notation $M/M/1/\infty$ signifie que les clients arrivent au système selon une loi de Poisson (modélisée par une chaîne de Markov), que le temps de traitement est du type exponentiel (modélisée par une chaîne de Markov aussi) et le système constitué d'un seul serveur selon le principe du premier arrivé premier servi dans une file d'attente à population infinie et régime permanent.

1.2 Définitions

On considère un faisceau de service (circuits) de capacité N, composé de N services élémentaires identiques, soumis à un flux de requêtes et un temps moyen de service constant.

Flux d'arrivée (λ) de requêtes dans la file d'attente, appelé également "taux moyen d'arrivée ", ou encore "fréquence moyenne d'arrivées". L'inverse λ^{-1} donne le temps moyen entre requêtes.

Flux de départ (μ) correspondant au taux de traitement des requêtes. L'inverse μ^{-1} donne le temps moyen d'attente **pendant** le service $\mathbf{t_m}$ (donc une fois arrivé en fin de file d'attente) appelé aussi "temps moyen de service".

Charge de référence du service :

C'est le nombre **moyen** de services élémentaires utilisés. Il est égal au rapport $\lambda/\mu = \lambda\, t_m$ et doit être strictement inférieur à N le nombre de serveurs pour éviter l'engorgement (plus de sortie que d'arrivée). Pour un faisceau à N liens, la charge maximum vaut N, elle correspond à une occupation permanente des canaux. En pratique, comme les appels ont lieu de manière aléatoire, il arrive que le faisceau soit congestionné, c'est-à-dire que les N liens soient occupés.

$$A = \frac{\lambda}{\mu} = \lambda t_m$$

La charge est une grandeur sans unité qui s'exprime en *Erlang*.

Mesure de la charge, le Erlang. En téléphonie, un Erlang correspond à l'occupation maximale d'un équipement permettant l'acheminement d'une seule communication téléphonique, ce qui peut être réalisé par exemple dans le cas d'une communication d'une heure, ou de dix communication de six minutes exactement consécutives, sans que l'équipement ne reste inactif).
Ainsi :

- 1 Erlang correspond à une communication téléphonique permanente sur la durée d'observation.
- 0,3 Erlang correspond à l'utilisation de l'équipement pendant 30 % du temps considéré, par exemple, pour une heure d'observation, à une communication de 18 minutes.
- 10 Erlangs correspondent à l'occupation d'un ensemble de dix équipements pendant une heure d'observation, ou bien à l'occupation d'un ensemble de vingt équipements pendant 50 % de la durée d'observation.

Exemple: l'analyse des appels d'une société révèle un taux d'appel de 40 appels par heure (λ=40appels/60mn) et une durée moyenne par appel de 5 minutes (t_m=5mn, μ=1/5mn). La charge vaut donc : $A = \frac{40}{60}5 = 3{,}3$ Erlang

Ce qui signifie qu'il y a en moyenne 3,3 lignes occupées si la distribution est uniforme.

Dans ce chapitre on considère les modèles M/M/1 et M/M/N

1.3 Hypothèse du modèle

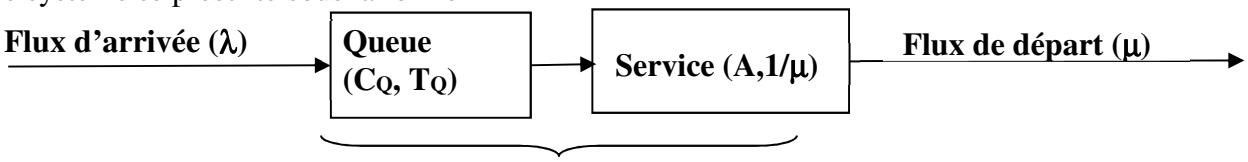
On considère un faisceau de service (circuits) de capacité N.
Les flux de de départ et de service doivent vérifier les hypothèses correspondant à un processus de Poisson d'intensité respective λ et μ et donc les conditions suivantes :

- Les appels arrivent suivant un processus de poisson de taux λ. Le taux d'arrivée de appels est donc indépendant des appels passés et la source est infini (pas de limite sur le nombre d'appels).
- La durée de service suit une loi exponentielle de paramètre μ. La durée moyenne de service est donc $t_m=1/\mu$.
- L'accessibilité au service est totale
- Les appels rejetés au service par manque de ressources, blocage sont perdus et ne sont pas corrélés aux appels entrant (pas de répétition d'appels). Cela correspond à ce qu'on appelle le modèle de Erlang B. On verra plus loin le modèle avec attente jusqu'au service qui correspond au modèle de Erlang C.

En conséquence de l'hypothèse de Poisson, le temps s'écoulant entre deux arrivée est suit une loi exponentielle et la probabilité d'avoir une arrivée pendant le temps *dt* est (λt).
De même la probabilité qu'un service se termine durant le temps *dt* est (μt).
En absence de blocage (nombre de service illimité), le taux d'occupation moyenne des lignes, appelé aussi charge offerte du faisceau est A : $A = \frac{\lambda}{\mu}$

2 Modélisation des arrivées et départs d'un système M/M/1/∞

Le système se présente sous la forme



C désigne la v.a. donnant le nombre de clients présent dans le système, C_Q le nombre de clients présent dans la queue, 1 client est dans le service.
T désigne la v.a. donnant le temps passé par un client dans le système, T_Q le temps passé par un client dans la queue. Le temps moyen passé dans le service est $t_m=1/\mu$.
On étudie le régime stationnaire du système.

La charge $A=\frac{\lambda}{\mu}=\lambda t_m$ doit être inférieure à 1.

On démontre que :

Probabilité système vide	Notation	1-A
Probabilité d'attente		A
Nombre moyen de clients dans le système	L	$E(C)=\frac{A}{1-A}$
Nombre moyen de clients en attente	L_q	$E(C_Q)=L-A=\frac{A^2}{1-A}$
Nombre moyen de clients en service		A
Temps moyen de séjour dans le système	W	$E(T)=\frac{E(C)}{\lambda}=\frac{1}{\mu-\lambda}=\frac{1}{\mu}\left(\frac{1}{1-A}\right)$
Temps moyen d'attente dans la queue	W_q	$E(T_Q)=W-\frac{1}{\mu}=\frac{1}{\mu}\left(\frac{A}{1-A}\right)$
Condition d'atteinte de l'équilibre		$A=\frac{\lambda}{\mu}<1$
Probabilité d'avoir *k* clients dans le système		$p_k=A^k(1-A)$

La relation $W=E(T)=\frac{E(C)}{\lambda}=\frac{L}{\lambda}$ s'appelle "relation de Little" (ce dernier ayant démontré rigoureusement que la relation est valable pour n'importe quel type de file d'attente). De plus $W_q=W-\frac{1}{\mu}$ et $L_q=L-A$.

Exemple: Supposons un service de dépannage traitant une requête à la fois. Supposons que λ=8 requêtes/heure (nombre de requêtes arrivant en moyenne par heure) et que la résolution d'une requête dure en moyenne 6min, μ=10 résolutions/heure (nombre de résolutions en moyenne par heure). Nous avons alors:

$$A=\frac{8}{10}=0{,}8$$

ce qui correspond à la charge ou taux d'occupation du service. Donc il y a 20% de probabilité pour que le service soit vide et 80% de probabilité pour qu'il y ait une attente.

$$L=E(C)=\frac{A}{1-A}=4$$

Ce qui correspond donc au nombre moyen de requêtes dans le système (service + en attente).

$$L_Q = E(C_Q) = \frac{A^2}{1-A} = L - A = 3{,}2$$

Ce qui correspond donc au nombre moyen de requêtes en attente.

$$W = E(T) = \frac{E(C)}{\lambda} = \frac{1}{\mu - \lambda} = \frac{1}{\mu}\left(\frac{1}{1-A}\right) = 0{,}5\text{h}$$

Ce qui correspond à un temps moyen de résolution de 30 minutes.

$$W_Q = E(T_Q) = \frac{1}{\mu}\left(\frac{A}{1-A}\right) = 0{,}4\text{h}$$

Ce qui correspond à une attente moyenne de 24 minutes dans la file d'attente.
Et la probabilité qu'il y a ait 5 requêtes dans le système (exécution + attente):

$$p_k = A^k(1-A) = A^5(1-A) = 6{,}5\%$$

3 Modélisation des arrivées et départs d'un système M/D/1/∞

La distribution des services des clients du système est **constante** de durée t_m.

On note $\mu = \frac{1}{t_m}$, l'inverse du temps de service. La charge $A = \frac{\lambda}{\mu} = \lambda t_m$ doit être inférieure à 1.

Le temps moyen d'attente est : $W_Q = \frac{1}{2\mu}\left(\frac{A}{1-A}\right)$

Le temps moyen d'attente est ici plus faible, **la moitié moins**, que pour un système M/M/1/∞ de même durée moyenne de service (1/μ).

Le temps moyen dans le service est : $W = E(T) = W_Q + \frac{1}{\mu} = \frac{1}{2\mu}\frac{2-A}{1-A}$

Le nombre moyen de clients en attente: $L_Q = \frac{A^2}{2(1-A)}$,est **moitié moins** que pour un système M/M/1/∞.

Nombre moyen de clients dans le système : $L = L_Q + A$

Exemple: Supposons que l'on dispose d'une machine à commande numérique traitant des pièces une à la fois. Supposons que λ=8pièces/heure (nombre de pièces arrivant en moyenne par heure) et que la durée du traitement est constante et dure 6 min, μ=10pièces/heure (nombre de pièces usinés par heure). Nous avons alors:

$$A = \frac{8}{10} = 0{,}8$$

ce qui correspond au trafic ou taux d'occupation de la machine. Donc il y a 20% de probabilité pour que le système soit vide et 80% de probabilité pour qu'il y ait une attente.

Le temps moyen d'attente est : $E(T_Q) = W_Q = \frac{1}{2\mu}\left(\frac{A}{1-A}\right) = 0{,}2$ soit 12 min

Le temps moyen d'attente est la moitié que pour un système M/M/1/∞ de même durée moyenne de service (1/μ) : 24 min dans l'exemple précédent.

Le temps moyen dans le service est : $W = W_Q + \frac{1}{\mu} = 0{,}2 + 0{,}1 = 0{,}3$ soit 18 min

La relation de Little donne : $E(C) = \lambda E(T) = 2{,}4$ qui est plus faible que pour un système M/M/1/∞ de même durée moyenne de service (1/μ) : 4 dans l'exemple précédent.

On a $L_Q = \frac{A^2}{2(1-A)} = 1{,}6$ qui est moitié plus faible que 3,2 dans l'exemple précédent.

On vérifie $L = L_Q + A$

4 Modèle d'un système M/M/N/0 avec perte : modèle d'ERLANG B

Nous allons nous intéresser ici à un système disposant de N canaux de communication (chaque canal censé supporter un débit de un appel avec réponse immédiate). Si les N canaux sont occupés, les appels qui arrivent sont considérés comme perdus (pas de file d'attente). Nous parlons alors de blocage du système. Il s'agit donc d'une file d'attente limitée de type M/M/N/0 selon la notation de Kendall, appelée également "système à perte".

On appelle taux d'utilisation (ou agent occupancy), le rapport $\rho = \frac{A}{N}$, le nombre moyen de canaux occupés.

4.1 Etude du taux d'occupation

À tout moment, des tentatives d'appel et des arrêts peuvent avoir lieu sur la ligne; la charge fluctue donc entre 0 et N appels en cours. Pour un intervalle de temps très court *dt* , la probabilité d'avoir exactement une tentative d'appel (n =1) vaut $e^{-(\lambda dt)}(\lambda dt) \approx \lambda dt$

De même la probabilité d'avoir un appel qui se termine vaut μdt .

En supposant que k lignes sur un total de N lignes soient occupées à l'instant t, on calcule respectivement trois probabilités sur un intervalle de temps dt :

- P_1 : la probabilité d'une tentative d'appel, $P_1 \approx \lambda dt$
- P_2 : la probabilité d'un arrêt, $P_2 = C_k^1 (\mu dt)^k (1-\mu dt)^{k-1} \approx k\mu dt$, car n'importe quelle ligne parmi les k lignes peut se libérer pendant l'intervalle de temps dt .
- P_3 : la probabilité d'un statu quo en matière d'occupation de lignes, c'est ce qui se produit quand il n'y a pas de tentative d'appel et qu'aucune des k lignes occupées ne se libère :
  $P_2 = (1-\lambda dt)(1-k\mu dt) \approx 1-\lambda dt - k\mu dt$

De par la nature poissonienne, on suppose qu'il n'y a pas simultanément une tentative d'appel et un arrêt.

On définit p(k;t+dt) la probabilité qu'il y ait k lignes occupées à l'instant t + dt. On peut écrire p(k;t+dt) en utilisant les probabilités P_1, P_2, P_3 :

$$p(k,t+dt) = P_1\, p(k-1,t) + P_2\, p(k+1,t) + P_3\, p(k,t)$$

Le premier terme correspond au fait qu'il y avait k-1 lignes occupées à l'instant t mais qu'une tentative d'appel ait eu lieu pendant dt. Le deuxième terme correspond au fait qu'il y avait k+1 lignes occupées à l'instant t et qu'une ligne se soit libérée pendant dt. Le dernier terme correspond au fait qu'il y avait déjà k lignes occupées à l'instant t mais qu'aucune tentative d'appel ni d'arrêt n'ait eu lieu.

$$p(k,t+dt) \approx \lambda dt\, p(k-1,t) + (k+1)\mu dt\, p(k+1,t) + (1-\lambda dt - k\mu dt)p(k,t)$$

Il existe cependant deux cas particuliers correspondant à k=0 et k=N :

$$p(0,t+dt) \approx (1-\lambda dt)\, p(0,t) + \mu dt\, p(1,t) \text{ (pas de possibilité de relâchement ligne)}$$

$$p(N,t+dt) \approx \lambda dt\, p(N-1,t) + (1-\lambda dt - N\mu dt)p(N,t) \text{ (pas de nouvelle tentative d'appel).}$$

En régime stationnaire, on fait l'hypothèse que les probabilités ne sont pas fonction du temps :

$$p(k,t+dt) = p(k,t) = P_k$$

L'équation de transition s'écrit

$$P_k = \lambda dt\, P_{k-1} + (k+1)\mu dt\, P_{k+1} + (1-\lambda dt - k\mu dt)P_k$$

$$0 = (\lambda\, P_{k-1} + (k+1)\mu P_{k+1} + (-\lambda - k\mu)P_k)dt$$

$$(\lambda + k\mu)P_k = \lambda\, P_{k-1} + (k+1)\mu P_{k+1}$$

Et pour les cas particulier :

$$k=0 : \lambda P_0 = \mu P_1$$

$$k=N : (\lambda - N\mu)P_N = \lambda\, P_{N-1}$$

De plus, les probabilités P_k doivent respecter la condition suivante :

$$\sum_{k=0}^{N} P_k = 1$$

On peut montrer que l'expression de P_k vérifiant toutes ces conditions est donnée par l'expression :

$$P_k = \frac{\frac{(\lambda/\mu)^k}{k!}}{\sum_{i=0}^{N}\frac{(\lambda/\mu)^i}{i!}} \ ;$$

Or pour une variable X suivant une loi de Poisson d'espérance A on a : $P(X=k) = \frac{e^{-A}A^k}{k!}$

On peut écrire $P_k = \frac{\frac{(A)^k}{k!}}{\sum_{i=0}^{N}\frac{(A)^i}{i!}} = \frac{\frac{e^{-A}(A)^k}{k!}}{\sum_{i=0}^{N}e^{-A}\frac{(A)^i}{i!}} = \frac{P(X=k)}{P(X\le N)}$ avec $A = \frac{\lambda}{\mu}$

Cette formule représente ainsi la probabilité d'avoir k lignes occupées. Elle est valable $k \in [0, N]$

4.2 Congestion

L'état qui résulte d'une occupation de toutes les lignes est appelé congestion. La probabilité de cet événement de blocage si un appel est rejeté en raison d'une occupation des N lignes est (k = N) :

$$B = P_N = \frac{\frac{(\lambda/\mu)^N}{N!}}{\sum_{i=0}^{N}\frac{(\lambda/\mu)^i}{i!}} = \frac{\frac{(A)^N}{N!}}{\sum_{i=0}^{N}\frac{(A)^i}{i!}} = \frac{P(X=N)}{P(X\le N)} \text{ avec } A = \frac{\lambda}{\mu}$$

Cette expression de la probabilité de blocage est la formule dite d'ERLANG B.

4.3 Nombre moyen de lignes occupées

L'espérance des probabilités P_k fournit le nombre moyen de lignes occupées, c'est-à-dire la charge du trafic écoulé. Cette espérance vaut

$$E(k) = \frac{\lambda}{\mu}(1-B) = A(1-B)$$

Comme A est la charge de référence du faisceau (sans blocage), A(1 - B) est la charge réelle (avec blocage).

Dans le cas du modèle d'ERLANG B, les appels en cas de blocage **sont perdus (les clients aussi !) car il n'y a pas de file d'attente**. On cherche généralement à avoir un taux de blocage très faible de l'ordre de quelques pourcents.

Exemple 1 : Calculer la probabilité de saturation d'une hotline (dont la durée de service suit une loi exponentielle et la distribution des arrivées suit une loi de Poisson) sachant que le trafic A de la ligne est estimé à 2 Erlang (λ=1 appel/mn, t_m=2mn, μ=1/2mn) pour une seule ligne téléphonique (N=1).

$$B = P_1 = \frac{\frac{(2)^1}{1!}}{\frac{2^0}{0!}+\frac{2^1}{1!}} = 67\%$$

Le taux d'occupation moyen est : $E(k) = A(1-B) = 2(1-0.666) = 0{,}667$

Exemple 2 : Dans une entreprise, on a dénombré aux heures de pointes 200 appels d'une durée moyenne de 6 minutes à l'heure (temps de service moyen). Quelle est la probabilité de saturation avec 20 opérateurs (sachant que la durée de service suit une loi exponentielle et la distribution des arrivées une loi de Poisson).

La plus grosse difficulté ici est de calculer le trafic! Il y a donc 200 appels par heures avec 10 appels traités seulement par heure (puisque 6 minutes par appel dans une heure de 60 minutes

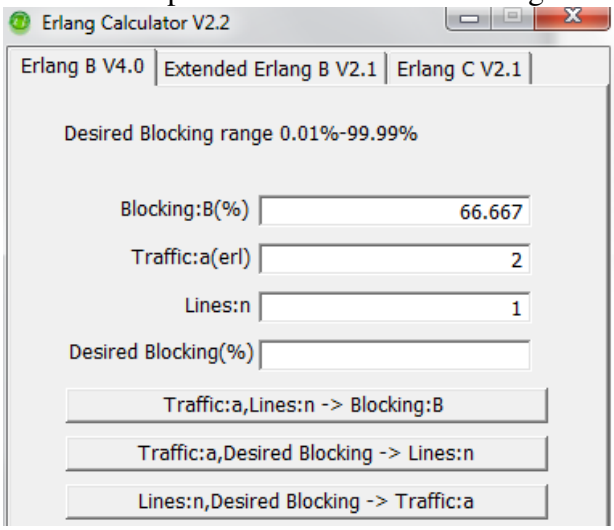
fait 10 appels). Le trafic A est donc de 200/10 soit 20 Erlang. En appliquant alors la relation précédente, nous avons.

$$B = P_{20} = \frac{\frac{(20)^{20}}{20!}}{\sum_{i=0}^{20}\frac{(20)^{i}}{i!}} = 15{,}8\%$$

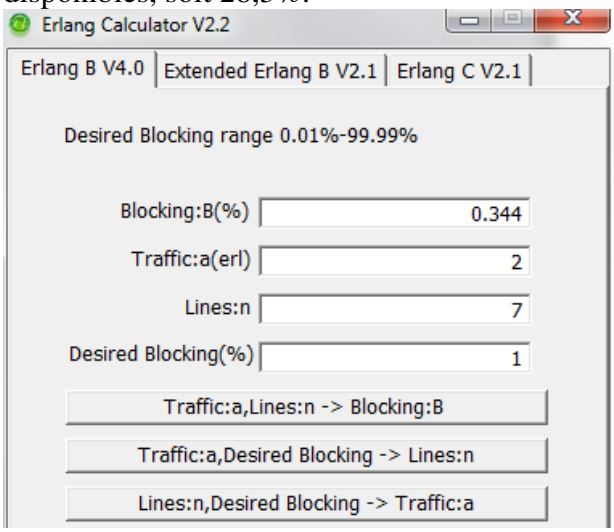
Dans l'industrie on admet un taux de saturation de 1%. Avec un logiciel, on trouve que N doit alors être égal à 30.

Utilisation de logiciel: avec un programme trouvé en ligne

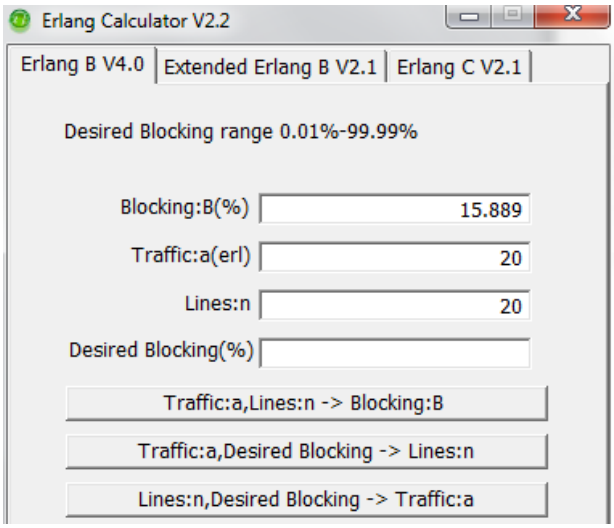
http://www.erlang.com/calculator/exeb/ en ligne (Erlang B et C)
ou téléchargeable :
http://home.earthlink.net/~malcolmhamer/Erlang-B.xls pour Excel (Erlang B seul)
http://www.softpedia.com/get/Science-CAD/Erlang-Calculator.shtml ou
http://hp.vector.co.jp/authors/VA002244/erlang.htm Erlang Calculatot V2.2, executable (Erlang B et C)
Avec l'exemple 1 : on a un taux de blocage de 66.67%



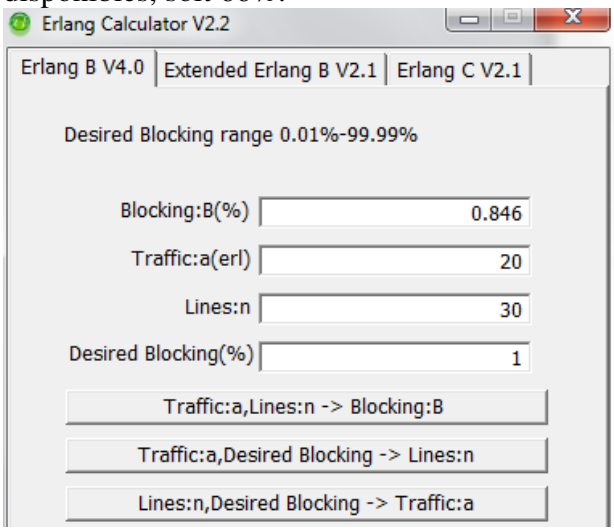
Si on désire un taux de blocage inférieur à 1% il faut 7 lignes et le taux de blocage est 0,34%. Le taux d'occupation moyen est : $E(k) = A(1-B) = 2(1-0{,}0034) = 1{,}99$ sur les 7 lignes disponibles, soit 28,3%.



Pour l'exemple 2 : on a un taux de blocage de 15,89%



Si on désire un taux de blocage inférieur à 1% il faut 30 lignes et le taux de blocage est 0,85%. et le taux d'occupation moyen est : $E(k) = A(1-B) = 20(1-0{,}0085) = 19{,}8$ sur les 30 lignes disponibles, soit 66%.

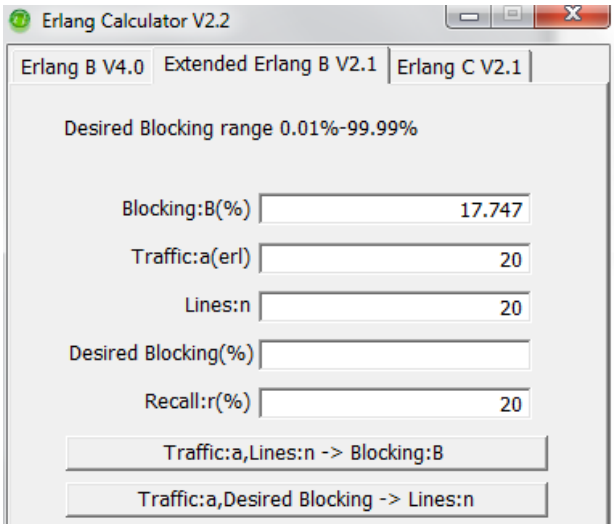


4.4 Loi d'Erlang B étendue

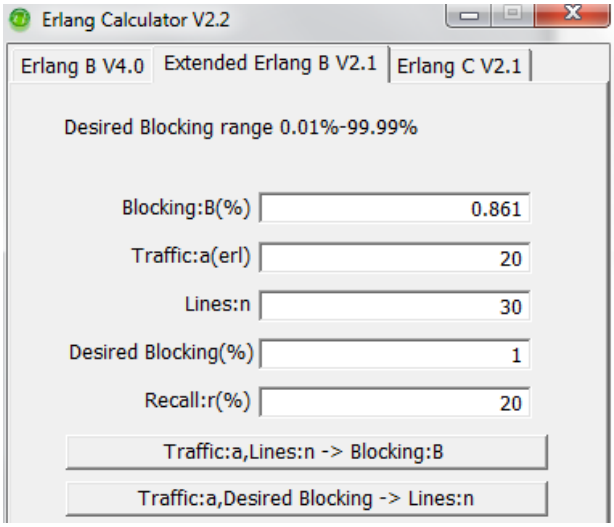
On considère toujours un système M/M/N/0 mais avec un taux de rappel de r (proportion qui en cas de blocage vont rappeler et se rajouter à la charge de référence). La résolution est itérative. Les étapes sont les suivantes : on part d'une charge de référence A_0

1. on calcule la probabilité de blocage B comme avec la formule d'Erlang B
2. on calcule le nombre probable de rappel : R=B r
3. on calcule le nouveau trafic : $A_{i+1} = A_i + R$
5. on retourne à l'étape 1 (calcul de B) jusqu'à stabilisation de B.

Exemple 2 : si on prend un taux de rappel de 20% on a un taux de blocage de 17,7% :



Si on désire un taux de blocage inférieur à 1% il faut 30 lignes et le taux de blocage est 0,86%, quasi inchangé de celui obtenu avec un taux de rappel de 0 car le taux de blocage est très faible !



4.5 Charge d'un système avec attente jusqu'au service

L'effet d'une tentative d'appel ayant échoué mais reconduite jusqu'à obtention d'une ligne peut être modélisé assez facilement. Appelons A' la charge réelle qui tient compte de la reconduite des tentatives échouées. On a :

$$A' = A + AB + (AB)B + (AB^2)B + ... = \frac{A}{1-B}$$

La charge réelle A' est supérieure à la charge de référence A du faisceau.
Il est important de noter qu'en principe le réseau doit être dimensionné pour cette charge A' et non pour la charge A. Cela revient à prendre un taux de rappel de 100%.

5 Modèle d'un système M/M/N/∞ avec mise en attente: modèle d'ERLANG C

Considérons maintenant un système pour lequel les appels peuvent être mis en attente avant d'être servis lorsque les N serveurs sont bloqués. Il s'agit donc d'une file d'attente à capacité illimitée de type M/M/N/∞ selon la notation de Kendall.
On peut montrer que la probabilité P_k qu'il y ait k lignes occupées est donnée par :

$$\text{pour } 0 \le k \le N-1: \qquad P_k = \frac{A^k}{k!} P_0 = \frac{\frac{A^k}{k!}}{\sum_{i=0}^{N-1} \frac{A^i}{i!} + \frac{A^N}{N!}\frac{N}{N-A}}$$

pour $k \geq N$: $$P_k = P_N\left(\frac{A}{N}\right)^{k-N} = \frac{A^N}{N!}P_0\left(\frac{A}{N}\right)^{k-N} = \frac{\frac{A^k}{N!N^{k-N}}}{\sum_{i=0}^{N-1}\frac{A^i}{i!}+\frac{A^N}{N!}\frac{N}{N-A}}$$

La probabilité cumulée de mise en file d'attente se note C(N, A), elle est égale à :

$$C(N,A) = \sum_{k=N}^{\infty} p_k = \frac{\frac{A^N}{N!}\frac{N}{N-A}}{\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}}$$

En reprenant une variable X suivant une loi de Poisson d'espérance A : $P(X=k) = \frac{e^{-A}A^k}{k!}$

On a : $C(N,A) = \sum_{k=N}^{\infty} p_k = \frac{\frac{A^N}{N!}\frac{N}{N-A}}{\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}} = \frac{P(X=N)}{P(X=N)+\frac{N-A}{N}P(X \leq N-1)}$

Avec le taux d'utilisation $\rho = \frac{A}{N}$, on a : $C(N,A) = \frac{P(X=N)}{P(X=N)+(1-\rho)P(X \leq (N-1))}$

Elle donne la probabilité de mise en attente (de saturation/blocage, tous les serveurs sont bloqués) dans un système disposant de *N* canaux pour un trafic *A* (exprimé en "Erlang") et dans lequel les communications peuvent être mises en attente (à l'opposé du modèle Erlang-B) dans un file à capacité infinie selon le principe du premier/arrivé premier servi (FIFO). Cette dernière relation est appelée "formule d'Erlang C". Traditionnellement on note :

$C(N,A) = P(W)$ le "*W*" venant de l'anglais "Wait" (attendre).

Pour avoir le nombre moyen de personnes dans le service on fait $E(C) = \sum_{k=0}^{\infty} k\,P_k$

Pour avoir la longueur moyenne de la file d'attente on fait :

$$E(C_Q) = \sum_{k=N}^{\infty}(k-1)P_k = \frac{A^N}{N!}P_0\sum_{k=N}^{\infty}(k-1)\left(\frac{A}{N}\right)^{k-N}$$

soit $E(C_q) = C(N,A)\frac{A}{N-A} = C(N,A)\frac{\rho}{1-\rho}$

On applique la relation de Little pour avoir la durée moyenne de service $E(T) = \frac{E(C)}{\lambda}$ et la durée moyenne d'attente : $E(T_q) = E(T) - \frac{1}{\mu}$. On a $E(T_q) = \frac{C(N,A)\cdot t_m}{N\cdot(1-\rho)}$.

En anglais la durée d'attente moyenne est appelée "Average Speed of Answer" ou ASA.

On peut aussi calculer la probabilité qu'une demande soit prise en moins d'un temps fixé (target waiting time.

$$W(t) = Prob(waiting\ time < t) = 1 - C(N,A) \cdot \exp\left(-(N-A) \cdot \frac{t}{t_m}\right)$$

Pour avoir le nombre moyen de services utilisés : $\sum_{k=1}^{N} k\,P_k + \sum_{k=N+1}^{\infty} N\,P_k = A$, la charge de référence !

Le modèle proposé ci-dessus est bien évidemment critiquable car en réalité la capacité de la file d'attente est finie et certains clients abandonnent lorsque l'attente est trop longue.

Calcul du nombre de serveurs

Si le trafic A est connu, on veut dimensionner le nombre de serveurs N nécessaire pour obtenir un probabilité d'attente fixée. On travail de manière itérative en calculant C(N,A) en partant d'un nombre de serveur égal au trafic, puis en incrémentant N jusqu'à obtenir la valeur de C(N,A) souhaitée.

Exemple: Si nous prenons un taux d'arrivé de 1 appel par minute et une durée moyenne de service de 5 minutes, nous avons :

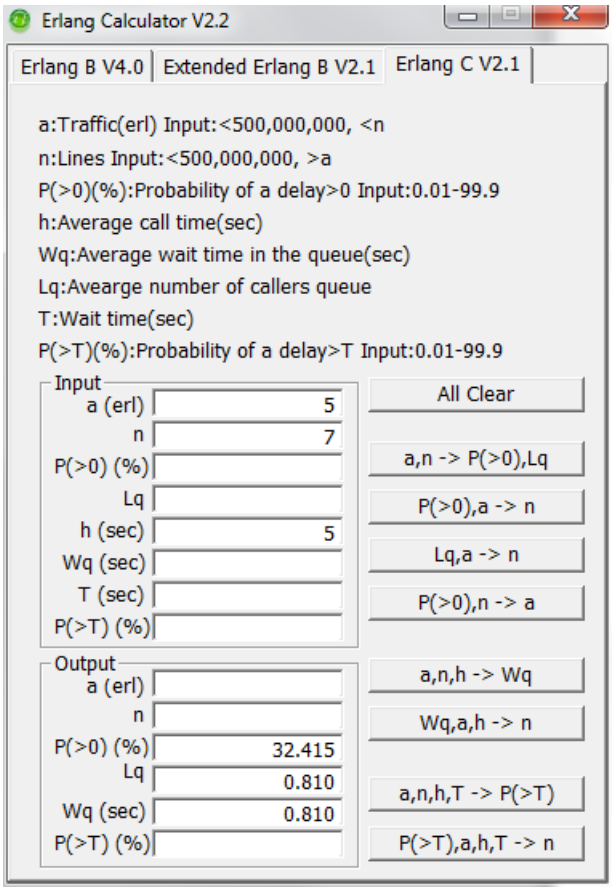
$$A = \frac{\lambda}{\mu} = 5$$

Si nous prenons un nombre N de 7 serveurs, nous avons:

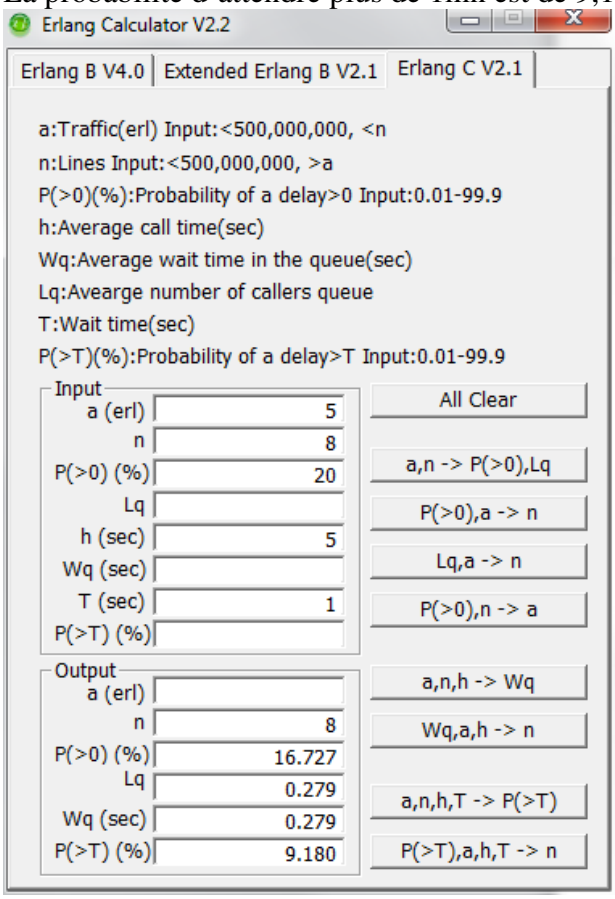
$$C(N,A) = \frac{\frac{5^7}{7!}\frac{7}{7-5}}{\sum_{i=0}^{6}\frac{5^i}{i!} + \frac{5^7}{7!}\frac{7}{7-5}} = 0{,}324$$

On a donc une probabilité cumulée de 32,4% d'être mis en attente. Ce qui un peu beaucoup.. (une règle empirique consiste à chercher le nombre N de serveurs afin que cette dernière valeur descende en-dessous des 20%).

Utilisation de logiciel: Avec A=5 et N=7, on a un taux de blocage (probabilité d'attente) de 32,4% et un nombre moyen de clients en attente Lq=0,81. Le temps moyen d'attente dans la queue est Wq=0,81mn (car débit A=5 et durée de service h=5mn)



Si on veut une probabilité d'attente inférieur à 20% on obtient n=8, un taux de blocage de 16,7%<20% et un nombre moyen de clients en attente Lq=0,279. Le temps moyen d'attente dans la queue est Wq=0,279mn.
La probabilité d'attendre plus de 1mn est de 9,18%.



Bibliographie : de nombreux cours se trouvent sur le web, entre autre, ces cours m'ont beaucoup inspiré

http://gestion.coursgratuits.net/technique-de-gestion/theories-des-files-d-attentes.php
http://www2.lifl.fr/~petitot/cours/trafic/poly.pdf

6 Exercices

6.1 Exercice

Un dépanneur opère avec un caissier unique. Il arrive en moyenne 24 clients à l'heure et 30 clients peuvent être servis à l'heure. On a constaté que chaque minute de réduction d'attente de clients sauve 75 euros par semaine (perte de vente évitée) et un employé supplémentaire coute 150 euros par semaine. Etudier les trois solutions envisageables :ajouter l'employé à la même caisse sachant que ceci augmente le taux de service à 40 clients par heure, ajouter une deuxième caisse avec file séparée, ou ajouter une deuxième caisse avec une seule file d'attente pour les 2 caisses.

Avant : on a un système M/M/1/∞, λ=24, μ=30 donc A=24/30=4/5=0,8 et $L = E(C) = \frac{A}{1-A} = 4$ puis

$W = E(T) = \frac{L}{\lambda} = \frac{4}{24/60} = 10\text{mn}$; $W_q = E(T_q) = W - \frac{1}{\mu} = 8\text{mn}$ et $L_q = E(C_q) = L - A = 3,2$

En moyenne 3,2 personnes qui attendent 8mn.

Solution 1 : on a un système M/M/1/∞, λ=24, μ=40 donc A=24/40=0,6 et $L = E(C) = \frac{A}{1-A} = 1,5$ puis

$W = E(T) = \frac{L}{\lambda} = \frac{1,5}{24/60} = 3,75\text{mn}$; $W_q = E(T_q) = W - \frac{1}{\mu} = 2,25\text{mn}$ et $L_q = E(C_q) = L - A = 0,9$

On gagne donc $(8-2,25)x75 = 431,25$, moins 150 euros du cout de l'employé

Solution 2 : on a deux systèmes M/M/1/∞, pour chaque file λ=12, μ=30 donc A=12/30=0,4 et

$L = E(C) = \frac{A}{1-A} = 2/3$ puis $W = E(T) = \frac{L}{\lambda} = \frac{2x60}{3x12} = 3,3\text{mn}$; $W_q = E(T_q) = W - \frac{1}{\mu} = 1,33\text{mn}$ et

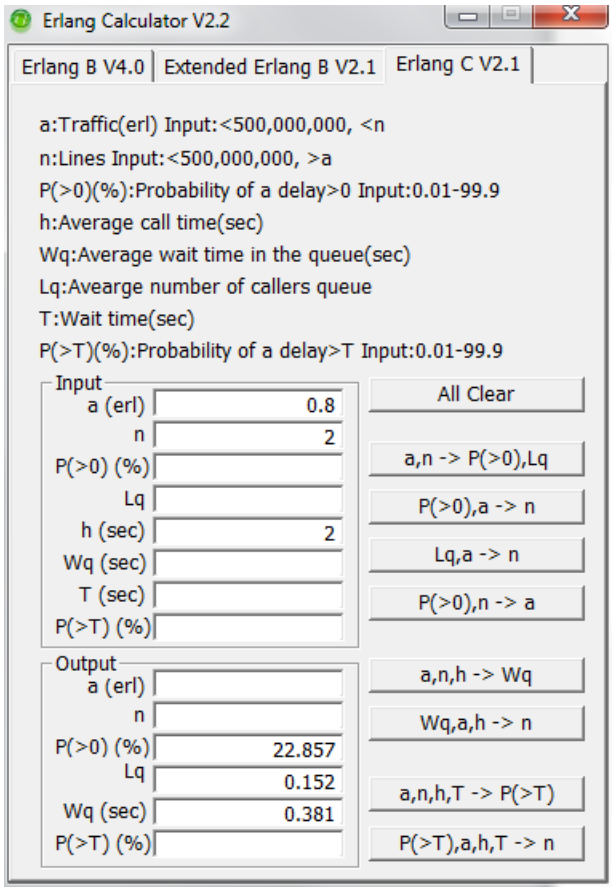
$L_q = E(C_q) = L - A = 0,27$

On gagne donc $(8-1,33)x75 = 500$, moins 150 euros du cout de l'employé

Solution 3 : on a un système M/M/2/∞, file λ=24 h^{-1}=24/60 min^{-1}, μ=30 h^{-1}= 30/60 min^{-1} ; durée de service (average call time)=1/μ=2 donc A=24/30=0,8.

Probabilité d'attente de 22,86% ; $P_0 = \frac{1}{\sum_{k=0}^{2-1} \frac{A^k}{k!} + \frac{A^2}{2!}\frac{2}{2-A}} = \frac{1}{1+A+\frac{A^2}{2-A}} = 0,43$

$L_q = E(C_q) = \sum_{k=N}^{\infty}(k-1)P_k = \frac{A^N}{N!}P_0\sum_{k=N}^{\infty}(k-1)\left(\frac{A}{N}\right)^{k-N} = 0,152$



$W_q = 0{,}38$. On gagne donc $(8-0{,}38)x75 = 571$, moins 150 euros du cout de l'employé

La solution 3 est la meilleure.

Fonction de corrélation

1 Lien avec l'analyse de données

On rappelle que la mesure de la ressemblance entre deux séries de données x_i et y_i, correspondant à deux variables X et Y, se fait à l'aide de la covariance.
Elle se définit comme la valeur moyenne du produit de ces deux variables centrées.

$$COV(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Afin d'obtenir un coefficient sans dimension compris entre -1 et 1, on considère le coefficient de corrélation: $r = \frac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$ encore appelé "Pearson's r corrélation coefficient"

On va étudier ici la mesure de ressemblance entre deux signaux pour **différents décalage temporel entre ces signaux**.

2 La fonction d'autocorrélation d'un signal

2.1 Définition

On définit la **fonction d'autocorrélation** d'un signal à énergie finie comme :

$$C_{xx}(\tau) = \int_{-\infty}^{+\infty} x(t)^* x(t+\tau) dt$$

Pour un signal à puissance moyenne finie, la fonction d'autocorrélation devient :

$$C_{xx}(\tau) = \lim_{T \to +\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)^* x(t+\tau) dt$$

et pour un signal périodique :

$$C_{xx}(\tau) = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)^* x(t+\tau) dt$$

- La fonction d'autocorrélation traduit la similitude d'un signal au niveau de la forme en fonction du décalage temporel τ.
- C'est une mesure de la ressemblance du signal avec lui même au cours du temps.
- Si le signal est périodique mais noyé dans du bruit, sa fonction d'autocorrélation sera aussi périodique et mais avec un bruit très inférieur. Elle permettra de détecter cette périodicité.
- Intuitivement, la corrélation est maximale si on ne décale pas temporellement le signal

2.2 Propriétés

- La fonction d'autocorrélation d'un signal réel est paire :
  $C_{xx}(\tau) = C_{xx}(-\tau)$
- La fonction d'autocorrélation d'un signal complexe est à symétrie hermitienne :
  $C_{xx}(\tau) = C_{xx}^*(-\tau)$
- La fonction d'autocorrélation est maximale en 0, et cette valeur représente l'énergie du signal :
  $|C_{xx}(\tau)| \leq C_{xx}(0)$

  $$C_{xx}(0) = \int_{-\infty}^{+\infty} x(t)^* x(t) dt = E > 0$$

- D'où la possibilité de normaliser les fonctions d'autocorrélation pour les comparer entre elles et d'obtenir la **fonction d'autocorrélation normalisée** :

$$\rho_{xx}(\tau) = \frac{C_{xx}(\tau)}{C_{xx}(0)}$$

- La fonction d'autocorrélation d'un signal périodique est périodique, elle est donc maximale pour toutes les valeurs multiple de la période T:

$$C_{xx}(0) = C_{xx}(kT) = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)^* x(t)dt = \text{Puissance moyenne} \geq 0$$

3 La fonction d'intercorrélation de 2 signaux

On définit la **fonction d'intercorrélation** de 2 signaux à énergie finie :

$$C_{xy}(\tau) = \int_{-\infty}^{+\infty} x(t)^* y(t+\tau)dt$$

Pour 2 signaux à puissance moyenne finie, la fonction d'intercorrélation devient :

$$C_{xy}(\tau) = \lim_{T\to+\infty} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)^* y(t+\tau)dt$$

- Exemple d'application

  Une impulsion x(t) est émise avec une célérité c, elle se réfléchit sur une cible avant de revenir à son point de départ.

  Pour en déduire la distance d de la cible, on cherche à mesurer le retard t_0 entre impulsion émise x(t) et le signal de retour y(t). Cette mesure s'avère difficile si les signaux n'ont pas de caractéristiques facilement repérables:

  On va donc calculer la fonction d'intercorrélation $C_{xy}(\tau)$ entre le signal émis et le signal reçu. Le maximum de cette fonction correspond à la similitude maximale entre les 2 signaux et donc au retard t_0, d'où: $2d = c\,t_0$

  On peut considérer également que l'intercorrélation consiste à déplacer le signal émis jusqu'à ce que l'on ait un maximum de ressemblance, ce qui correspondra au retard t_0.

- Les formules de la convolution et de la corrélation sont très proches:

$$C_{xx}(\tau) = \int_{-\infty}^{+\infty} x(t)^* x(t+\tau)dt \;;\; x(t) \otimes y(t) = \int_{-\infty}^{\infty} x(t')\, y(t-t')\, dt'$$

$$C_{xx}(\tau) = x^*(\tau) \otimes x(-\tau)$$

4 Relation fonction de corrélation et transformée de Fourier.

Le théorème de Wiener-Khinchine énonce que la densité spectrale de puissance est la transformée de Fourier de la fonction d'autocorrélation :

$$S_{xx}(\nu) = TF[C_{xx}(t)]$$

Définition: densité interspectrale de puissance :

$$S_{xy}(\nu) = TF[C_{xy}(t)]$$

Propriété:

$S_{xy}(\nu) = X^*(\nu).Y(\nu)$ où $X(\nu)$ et $Y(\nu)$ sont les transformée de Fourier de x(t) et y(t).

$S_{xx}(\nu) = X^*(\nu).X(\nu)$ => $S_{xx}(\nu) = |X(\nu)|^2$

Inversement, la fonction d'autocorrélation est la transformée de Fourier inverse de la densité spectrale :

$$C_{xx}(t) = TF^{-1}\left[S_{xx}(\nu)\right] = TF^{-1}\left[\left|X(\nu)\right|^2\right]$$

Ce qui est beaucoup plus rapide à calculer.

5 Fonction de corrélation et étude des système linéaire

Soit le système linéaire:

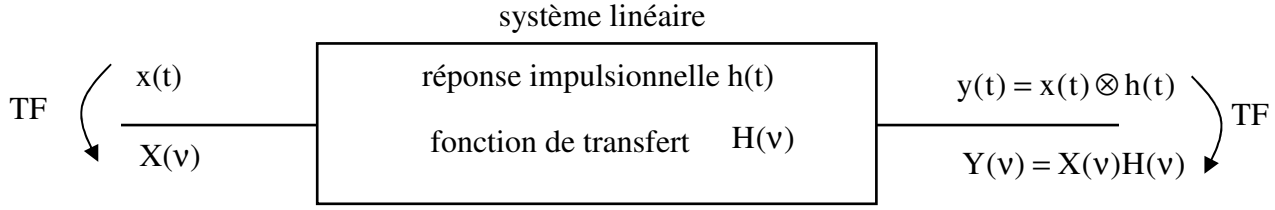
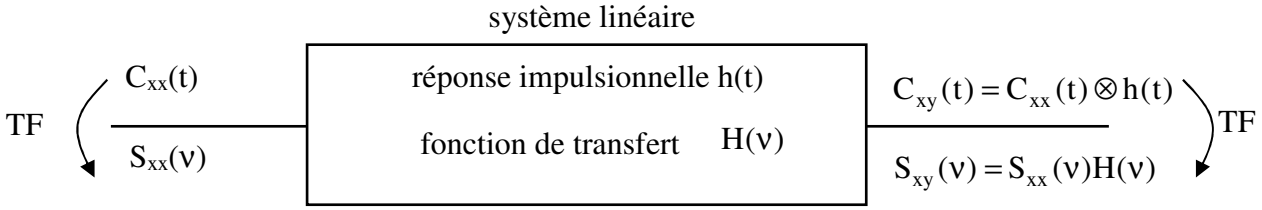
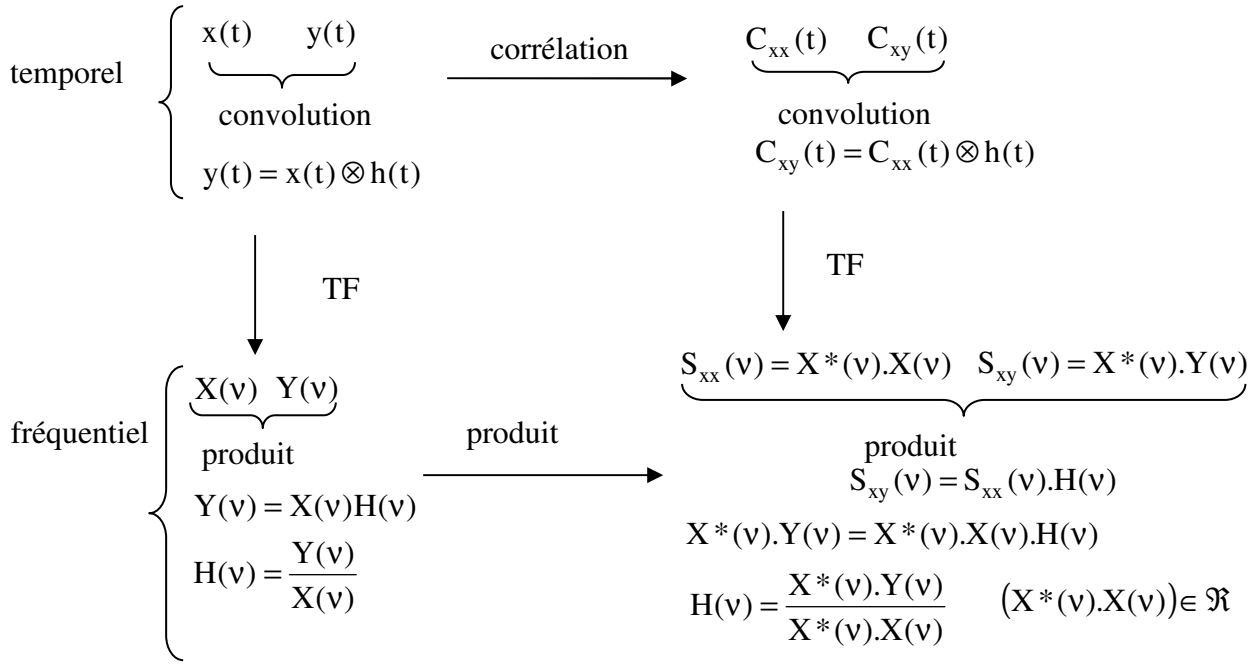


fig 5-(1): relation entrée-sortie d'un système linéaire

Les même relations entrée-sortie peuvent être écrite entre la fonction d'autocorrélation de l'entrée et la fonction d'intercorrélation entrée-sortie.



6 Résume



Formulaire de probabilités

Régression linéaire: $a = \frac{s_{XY}}{s_X^2}$; $b = \bar{y} - a\,\bar{x}$; $r = \frac{s_{XY}}{s_X\, s_Y}$; avec $s_{XY} = \frac{\sum n_i (x_i - \bar{x})(y_i - \bar{y})}{N}$

Soient A et B deux évènements:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	$P(\overline{A}) = 1 - P(A)$
A et B incompatibles si $P(A \cup B) = P(A) + P(B)$	Probabilité conditionnelle: $P(A/B) = \frac{P(A \cap B)}{P(B)}$
Evènements indépendants: $P(B/A) = P(B)$ ou $P(A/B) = P(A)$	A et B indépendants: $P(A \cap B) = P(A)\,P(B)$
Théorème de Bayes: $P(A/B) = P(B/A)\frac{P(A)}{P(B)}$	

p tirages parmi n objets	Sans remise	Avec remise
Avec ordre	A_n^p	n^p
Sans ordre (combinaison)	C_n^p	C_{n+p-1}^p

$$C_n^p = \frac{n!}{p!(n-p)!} = \frac{n(n-1)...(n-p+1)}{p!}$$

p objets sur n cases	un objet par case	sans limitation du nombre d'objets par case
Objets discernables (ordre)	A_n^p	n^p
Objets non discernables (sans ordre)	C_n^p	C_{n+p-1}^p

Loi discrète : Espérance $E(X) = \sum_{i=1}^{n} k_i\, P(X = k_i)$; Variance $V(X) = E\left((X - E(X))^2\right) = E\left(X^2\right) - E(X)^2$

Loi binomiale $\mathcal{B}(n;p)$: $P(X = k) = C_n^k p^k (1-p)^{n-k}$ pour tout k de 0 à n; E(X)=n p; V(X)=n p(1-p);

Loi de Poisson $\mathcal{P}(\lambda)$: $P(X = k) = \frac{e^{-\lambda}\,\lambda^k}{k!}$; E(X)=λ ; V(X)= λ;

Approximation: loi binomiale => loi de Poisson: si n>50 et p≤0,1 et np <17, on remplace la loi binomiale $\mathcal{B}(n;p)$ par la loi de Poisson $\mathcal{P}(np)$.

Loi Normale $\mathcal{N}(\mu,\sigma^2)$: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$; E(X)=μ ; V(X)= σ^2;

Approximation: loi binomiale => loi Normale sous une des conditions :

- si n>5 et $\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \frac{1}{\sqrt{n}} < 0{,}3$
- n≥30 et (n p>10) et (n(1-p)>10)
- np(1-p)>9

on remplace la loi binomiale $\mathcal{B}(n,p)$ par la loi $\mathcal{N}(np,\ n\,p\,(1-p))$

Approximation: loi de Poisson => loi Normale: si λ>18 alors on remplace la loi de Poisson $\mathcal{P}(\lambda)$ par la loi $\mathcal{N}(\lambda,\lambda)$.

Résumé de l'estimation

Notation:

Caractéristique	Echantillon	Population totale
Taille	n	N
Moyenne	$\bar{x}$	μ
Variance	s^2	σ^2
Ecart-type	s	σ
Proportion	p_e	p

I ESTIMATION PONCTUELLE

$$\hat{\mu} = \bar{x} = \frac{\sum x_i}{n}$$

$$\hat{p} = p_e$$

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2 \quad \text{avec} \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^{\,2}}{n} - \bar{x}^2$$

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}}\, s$$

II INTERVALLES DE CONFIANCE

A Moyenne d'une loi X

***Hypothèses* : X suit une loi Normale ou la taille de l'échantillon est supérieure à 30.**

$$I = \left[\bar{x} - t\frac{\sigma}{\sqrt{n}} ; \bar{x} + t\frac{\sigma}{\sqrt{n}} \right]$$

Obtention de t :

- **si la variance de la population mère est connue**, t est obtenu dans la table de la loi **Normale** $\mathcal{N}(0,1)$ pour 1- $\frac{\alpha}{2}$
- **si la variance est inconnue mais estimée par** $\hat{\sigma}$, t est obtenu à partir de la loi de **Student** *bilatérale* à (n-1) degrés de liberté pour p = (1- α) et l'intervalle I devient :

$$I = \left[\bar{x} - t\frac{\hat{\sigma}}{\sqrt{n}} ; \bar{x} + t\frac{\hat{\sigma}}{\sqrt{n}} \right]$$

remarque : si n est supérieur à 30, on remplace la loi de Student par la loi Normale.

B Fréquence

***Hypothèses* : les conditions d'approximation de la loi Binomiale par la loi Normale s'appliquent.**

$$I = \left[p_e - t\sqrt{\frac{p_e(1-p_e)}{n}} ; p + t\sqrt{\frac{p_e(1-p_e)}{n}} \right]$$

t est déterminé à l'aide de la table $\mathcal{N}(0,1)$ pour la valeur $1-\frac{\alpha}{2}$.

B Variance

***Hypothèses* : X suit une loi Normale $\mathcal{N}(\mu, \sigma^2)$ ou la taille de l'échantillon est supérieure à 30.**

- Si μ est connu : $I = \left[\frac{n\,v}{\chi^2_{1-\frac{\alpha}{2}}(n)} ; \frac{n\,v}{\chi^2_{\frac{\alpha}{2}}(n)} \right]$ avec $v = \frac{1}{n}\sum (x_i - \mu)^2$, $\chi^2_{1-\frac{\alpha}{2}}(n)$ est la valeur x telle que $P\left(\chi^2(n) < x\right) = 1 - \frac{\alpha}{2}$ et $\chi^2_{\frac{\alpha}{2}}(n)$ est la valeur x telle que $P\left(\chi^2(n) < x\right) = \frac{\alpha}{2}$ pour la loi χ^2 à n degrés de liberté.
- Si μ est inconnu : $I = \left[\frac{n\,s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} ; \frac{n\,s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right]$ avec $s^2 = \frac{1}{n}\sum (x_i - \overline{x})^2$ variance de l'échantillon et $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ est la valeur x telle que $P\left(\chi^2(n-1) < x\right) = 1 - \frac{\alpha}{2}$ et $\chi^2_{\frac{\alpha}{2}}(n-1)$ est la valeur x telle que $P\left(\chi^2(n-1) < x\right) = \frac{\alpha}{2}$ pour la loi χ^2 à (n-1) degrés de liberté.

Résumé sur les tests d'hypothèse

Comparaison d'une moyenne d'échantillon à une valeur donnée

Loi Normale :

- Variance de la population σ^2 connue : $t_{obs} = \dfrac{\bar{x} - a}{\dfrac{\sigma}{\sqrt{n}}}$ suit une loi $\mathcal{N}(0,1)$
- Variance de la population σ^2 inconnue : $t_{obs} = \dfrac{\bar{x} - a}{\dfrac{\hat{\sigma}}{\sqrt{n}}}$ suit une loi $\mathcal{S}(n-1)$

Loi quelconque (n>30) : $t_{obs} = \dfrac{\bar{x} - a}{\dfrac{\hat{\sigma}}{\sqrt{n}}}$ suit une loi $\mathcal{N}(0,1)$

Comparaison d'une fréquence à une valeur donnée : $t_{obs} = \dfrac{p_{mes} - p_1}{\sqrt{\dfrac{p_1(1-p_1)}{n}}}$ suit une loi $\mathcal{N}(0,1)$

Comparaison de deux moyennes

Echantillons indépendants (2 échantillons (n_1, n_2))

- Populations normales de variances connues ou grands échantillons (n>30) :

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}} \text{ suit une loi } \mathcal{N}(0,1)$$

- Populations normales et variances inconnues: petits échantillons (n≤30) :

Test préliminaire d'égalité des variances: $F_{obs} = \dfrac{\hat{\sigma}_1^{\,2}}{\hat{\sigma}_2^{\,2}} > 1$ suit une loi de Fischer $F_{n1-1,n2-1}$.

On estime la variance commune : $\hat{\sigma}_c^2 = \dfrac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}_c^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \text{ suit une loi } \mathcal{S}(n_1+n_2-2)$$

Echantillons appariés : $t_{obs} = \dfrac{\bar{d}}{\dfrac{\hat{\sigma}_d}{\sqrt{n}}}$ suit une loi $\mathcal{S}(n-1)$ avec $\hat{\sigma}_d$ écart type estimé des écarts.

Comparaison de deux fréquences : $t_{obs} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$ suit une loi $\mathcal{N}(0,1)$

Dans chaque cas les valeurs de t_{obs} sont à comparer avec $t_{théo}$ lu suivant le cas

- dans la table de la loi **Normale** $\mathcal{N}(0,1)$ pour 1- $\dfrac{\alpha}{2}$
- dans la table de la loi de **Student** *bilatérale* pour p = (1- α)

Comparaison d'une variance d'échantillon à une valeur donnée (loi normale $\mathcal{N}(\mu, \sigma^2)$) :

- Moyenne connue : $\frac{n\,v}{\sigma^2}$ avec $v = \frac{1}{n}\sum (x_i - \mu)^2$ suit une loi du $\chi^2(n)$
- moyenne inconnue, $\frac{n\,s^2}{\sigma^2}$ suit une loi du $\chi^2(n-1)$

Comparaison de deux variances d'échantillons (n_1, n_2) : $F_{obs} = \frac{\hat{\sigma}_1^{\,2}}{\hat{\sigma}_2^{\,2}}$ suit une loi de Fischer $F_{n1-1,n2-1}$.

Test du χ^2 :

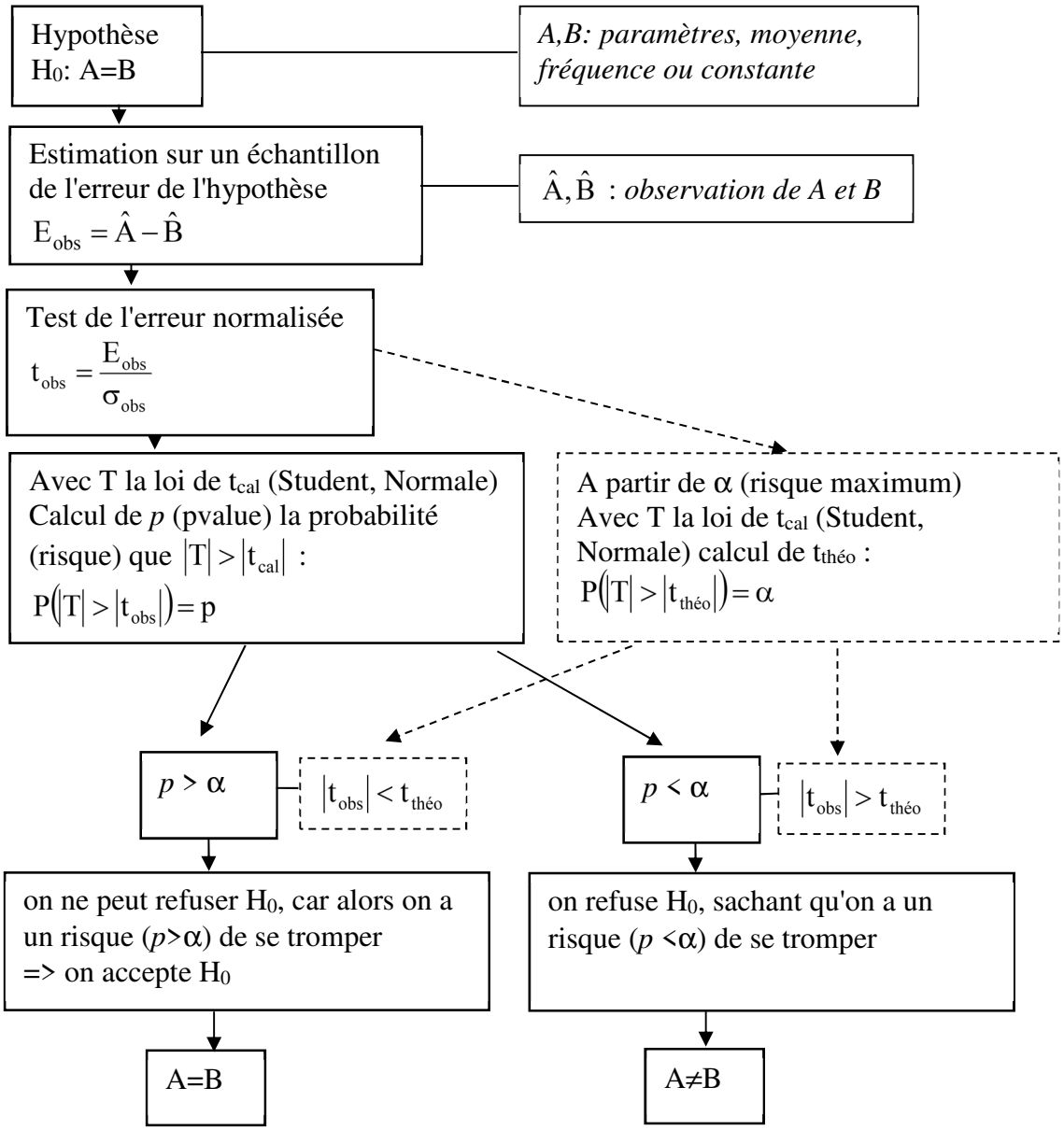
(1) Comparaison de la répartition de 2 populations, k classes

(2) Test indépendance X_1 à k_1 classes et X_2 à k_2 classes

population observée o_i, population théorique t_i : $\chi^2 = \sum_i \frac{(o_i - t_i)^2}{t_i}$

χ^2 suit une loi du X^2 à (k-1) ddl (1) , ou à $(k_1-1)(k_2-1)$ ddl (2) Effectif min des classes : 5

Résumé de la démarche d'un test



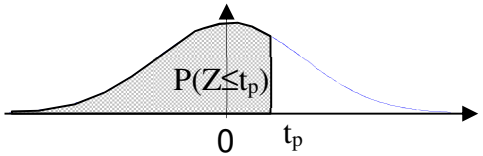
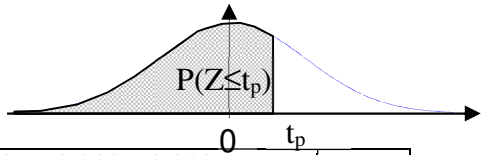


Table 1 : **Loi Normale centrée réduite $\mathcal{N}(0 ; 1)$: $P(Z<t_p)$, détermination de $p=P(Z\leq t_p)$ pour t_p connue**

tp	*0,00*	*0,01*	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,07*	*0,08*	*0,09*
0,0	**0,5000**	**0,5040**	**0,5080**	**0,5120**	**0,5160**	**0,5199**	**0,5239**	**0,5279**	**0,5319**	**0,5359**
0,1	**0,5398**	**0,5438**	**0,5478**	**0,5517**	**0,5557**	**0,5596**	**0,5636**	**0,5675**	**0,5714**	**0,5753**
0,2	**0,5793**	**0,5832**	**0,5871**	**0,5910**	**0,5948**	**0,5987**	**0,6026**	**0,6064**	**0,6103**	**0,6141**
0,3	**0,6179**	**0,6217**	**0,6255**	**0,6293**	**0,6331**	**0,6368**	**0,6406**	**0,6443**	**0,6480**	**0,6517**
0,4	**0,6554**	**0,6591**	**0,6628**	**0,6664**	**0,6700**	**0,6736**	**0,6772**	**0,6808**	**0,6844**	**0,6879**
0,5	**0,6915**	**0,6950**	**0,6985**	**0,7019**	**0,7054**	**0,7088**	**0,7123**	**0,7157**	**0,7190**	**0,7224**
0,6	**0,7257**	**0,7291**	**0,7324**	**0,7357**	**0,7389**	**0,7422**	**0,7454**	**0,7486**	**0,7517**	**0,7549**
0,7	**0,7580**	**0,7611**	**0,7642**	**0,7673**	**0,7704**	**0,7734**	**0,7764**	**0,7794**	**0,7823**	**0,7852**
0,8	**0,7881**	**0,7910**	**0,7939**	**0,7967**	**0,7995**	**0,8023**	**0,8051**	**0,8078**	**0,8106**	**0,8133**
0,9	**0,8159**	**0,8186**	**0,8212**	**0,8238**	**0,8264**	**0,8289**	**0,8315**	**0,8340**	**0,8365**	**0,8389**
1,0	**0,8413**	**0,8438**	**0,8461**	**0,8485**	**0,8508**	**0,8531**	**0,8554**	**0,8577**	**0,8599**	**0,8621**
1,1	**0,8643**	**0,8665**	**0,8686**	**0,8708**	**0,8729**	**0,8749**	**0,8770**	**0,8790**	**0,8810**	**0,8830**
1,2	**0,8849**	**0,8869**	**0,8888**	**0,8907**	**0,8925**	**0,8944**	**0,8962**	**0,8980**	**0,8997**	**0,9015**
1,3	**0,9032**	**0,9049**	**0,9066**	**0,9082**	**0,9099**	**0,9115**	**0,9131**	**0,9147**	**0,9162**	**0,9177**
1,4	**0,9192**	**0,9207**	**0,9222**	**0,9236**	**0,9251**	**0,9265**	**0,9279**	**0,9292**	**0,9306**	**0,9319**
1,5	**0,9332**	**0,9345**	**0,9357**	**0,9370**	**0,9382**	**0,9394**	**0,9406**	**0,9418**	**0,9429**	**0,9441**
1,6	**0,9452**	**0,9463**	**0,9474**	**0,9484**	**0,9495**	**0,9505**	**0,9515**	**0,9525**	**0,9535**	**0,9545**
1,7	**0,9554**	**0,9564**	**0,9573**	**0,9582**	**0,9591**	**0,9599**	**0,9608**	**0,9616**	**0,9625**	**0,9633**
1,8	**0,9641**	**0,9649**	**0,9656**	**0,9664**	**0,9671**	**0,9678**	**0,9686**	**0,9693**	**0,9699**	**0,9706**
1,9	**0,9713**	**0,9719**	**0,9726**	**0,9732**	**0,9738**	**0,9744**	**0,9750**	**0,9756**	**0,9761**	**0,9767**
2,0	**0,9772**	**0,9778**	**0,9783**	**0,9788**	**0,9793**	**0,9798**	**0,9803**	**0,9808**	**0,9812**	**0,9817**
2,1	**0,9821**	**0,9826**	**0,9830**	**0,9834**	**0,9838**	**0,9842**	**0,9846**	**0,9850**	**0,9854**	**0,9857**
2,2	**0,9861**	**0,9864**	**0,9868**	**0,9871**	**0,9875**	**0,9878**	**0,9881**	**0,9884**	**0,9887**	**0,9890**
2,3	**0,9893**	**0,9896**	**0,9898**	**0,9901**	**0,9904**	**0,9906**	**0,9909**	**0,9911**	**0,9913**	**0,9916**
2,4	**0,9918**	**0,9920**	**0,9922**	**0,9925**	**0,9927**	**0,9929**	**0,9931**	**0,9932**	**0,9934**	**0,9936**
2,5	**0,9938**	**0,9940**	**0,9941**	**0,9943**	**0,9945**	**0,9946**	**0,9948**	**0,9949**	**0,9951**	**0,9952**
2,6	**0,9953**	**0,9955**	**0,9956**	**0,9957**	**0,9959**	**0,9960**	**0,9961**	**0,9962**	**0,9963**	**0,9964**
2,7	**0,9965**	**0,9966**	**0,9967**	**0,9968**	**0,9969**	**0,9970**	**0,9971**	**0,9972**	**0,9973**	**0,9974**
2,8	**0,9974**	**0,9975**	**0,9976**	**0,9977**	**0,9977**	**0,9978**	**0,9979**	**0,9979**	**0,9980**	**0,9981**
2,9	**0,9981**	**0,9982**	**0,9982**	**0,9983**	**0,9984**	**0,9984**	**0,9985**	**0,9985**	**0,9986**	**0,9986**

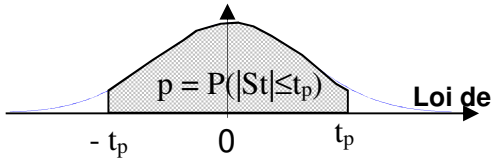
tp	*3,0*	*3,1*	*3,2*	*3,3*	*3,4*	*3,5*	*3,6*	*3,7*	*3,8*	*3,9*
P(T<tp)	**0,998650**	**0,999032**	**0,999313**	**0,999517**	**0,999663**	**0,999767**	**0,999841**	**0,999892**	**0,999928**	**0,999952**

Table 2 : **Loi Normale centrée réduite $\mathcal{N}(0 ; 1)$**
Détermination de t_p pour $p=P(Z\leq t_p)$ connue



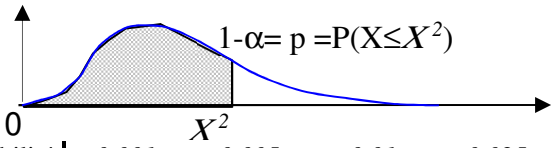
P<0,5	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009		
0,00		3,090	2,878	2,748	2,652	2,576	2,512	2,457	2,409	2,366	2,326	0,99
0,01	2,326	2,290	2,257	2,226	2,197	2,170	2,144	2,120	2,097	2,075	2,054	0,98
0,02	2,054	2,034	2,014	1,995	1,977	1,960	1,943	1,927	1,911	1,896	1,881	0,97
0,03	1,881	1,866	1,852	1,838	1,825	1,812	1,799	1,787	1,774	1,762	1,751	0,96
0,04	1,751	1,739	1,728	1,717	1,706	1,695	1,685	1,675	1,665	1,655	1,645	0,95
0,05	1,645	1,635	1,626	1,616	1,607	1,598	1,589	1,580	1,572	1,563	1,555	0,94
0,06	1,555	1,546	1,538	1,530	1,522	1,514	1,506	1,499	1,491	1,483	1,476	0,93
0,07	1,476	1,468	1,461	1,454	1,447	1,440	1,433	1,426	1,419	1,412	1,405	0,92
0,08	1,405	1,398	1,392	1,385	1,379	1,372	1,366	1,359	1,353	1,347	1,341	0,91
0,09	1,341	1,335	1,329	1,323	1,317	1,311	1,305	1,299	1,293	1,287	1,282	0,90
0,10	1,282	1,276	1,270	1,265	1,259	1,254	1,248	1,243	1,237	1,232	1,227	0,89
0,11	1,227	1,221	1,216	1,211	1,206	1,200	1,195	1,190	1,185	1,180	1,175	0,88
0,12	1,175	1,170	1,165	1,160	1,155	1,150	1,146	1,141	1,136	1,131	1,126	0,87
0,13	1,126	1,122	1,117	1,112	1,108	1,103	1,098	1,094	1,089	1,085	1,080	0,86
0,14	1,080	1,076	1,071	1,067	1,063	1,058	1,054	1,049	1,045	1,041	1,036	0,85
0,15	1,036	1,032	1,028	1,024	1,019	1,015	1,011	1,007	1,003	0,999	0,994	0,84
0,16	0,994	0,990	0,986	0,982	0,978	0,974	0,970	0,966	0,962	0,958	0,954	0,83
0,17	0,954	0,950	0,946	0,942	0,938	0,935	0,931	0,927	0,923	0,919	0,915	0,82
0,18	0,915	0,912	0,908	0,904	0,900	0,896	0,893	0,889	0,885	0,882	0,878	0,81
0,19	0,878	0,874	0,871	0,867	0,863	0,860	0,856	0,852	0,849	0,845	0,842	0,80
0,20	0,842	0,838	0,834	0,831	0,827	0,824	0,820	0,817	0,813	0,810	0,806	0,79
0,21	0,806	0,803	0,800	0,796	0,793	0,789	0,786	0,782	0,779	0,776	0,772	0,78
0,22	0,772	0,769	0,765	0,762	0,759	0,755	0,752	0,749	0,745	0,742	0,739	0,77
0,23	0,739	0,736	0,732	0,729	0,726	0,722	0,719	0,716	0,713	0,710	0,706	0,76
0,24	0,706	0,703	0,700	0,697	0,693	0,690	0,687	0,684	0,681	0,678	0,674	0,75
0,25	0,674	0,671	0,668	0,665	0,662	0,659	0,656	0,653	0,650	0,646	0,643	0,74
0,26	0,643	0,640	0,637	0,634	0,631	0,628	0,625	0,622	0,619	0,616	0,613	0,73
0,27	0,613	0,610	0,607	0,604	0,601	0,598	0,595	0,592	0,589	0,586	0,583	0,72
0,28	0,583	0,580	0,577	0,574	0,571	0,568	0,565	0,562	0,559	0,556	0,553	0,71
0,29	0,553	0,550	0,548	0,545	0,542	0,539	0,536	0,533	0,530	0,527	0,524	0,70
0,30	0,524	0,522	0,519	0,516	0,513	0,510	0,507	0,504	0,502	0,499	0,496	0,69
0,31	0,496	0,493	0,490	0,487	0,485	0,482	0,479	0,476	0,473	0,470	0,468	0,68
0,32	0,468	0,465	0,462	0,459	0,457	0,454	0,451	0,448	0,445	0,443	0,440	0,67
0,33	0,440	0,437	0,434	0,432	0,429	0,426	0,423	0,421	0,418	0,415	0,412	0,66
0,34	0,412	0,410	0,407	0,404	0,402	0,399	0,396	0,393	0,391	0,388	0,385	0,65
0,35	0,385	0,383	0,380	0,377	0,375	0,372	0,369	0,366	0,364	0,361	0,358	0,64
0,36	0,358	0,356	0,353	0,350	0,348	0,345	0,342	0,340	0,337	0,335	0,332	0,63
0,37	0,332	0,329	0,327	0,324	0,321	0,319	0,316	0,313	0,311	0,308	0,305	0,62
0,38	0,305	0,303	0,300	0,298	0,295	0,292	0,290	0,287	0,285	0,282	0,279	0,61
0,39	0,279	0,277	0,274	0,272	0,269	0,266	0,264	0,261	0,259	0,256	0,253	0,60
0,40	0,253	0,251	0,248	0,246	0,243	0,240	0,238	0,235	0,233	0,230	0,228	0,59
0,41	0,228	0,225	0,222	0,220	0,217	0,215	0,212	0,210	0,207	0,204	0,202	0,58
0,42	0,202	0,199	0,197	0,194	0,192	0,189	0,187	0,184	0,181	0,179	0,176	0,57
0,43	0,176	0,174	0,171	0,169	0,166	0,164	0,161	0,159	0,156	0,154	0,151	0,56
0,44	0,151	0,148	0,146	0,143	0,141	0,138	0,136	0,133	0,131	0,128	0,126	0,55
0,45	0,126	0,123	0,121	0,118	0,116	0,113	0,111	0,108	0,105	0,103	0,100	0,54
0,46	0,100	0,098	0,095	0,093	0,090	0,088	0,085	0,083	0,080	0,078	0,075	0,53
0,47	0,075	0,073	0,070	0,068	0,065	0,063	0,060	0,058	0,055	0,053	0,050	0,52
0,48	0,050	0,048	0,045	0,043	0,040	0,038	0,035	0,033	0,030	0,028	0,025	0,51
0,49	0,025	0,023	0,020	0,018	0,015	0,013	0,010	0,008	0,005	0,003	0,000	0,50
		0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	p≥0,5

p	0,9991	0,9992	0,9993	0,9994	0,9995	0,9996	0,9997	0,9998	0,9999
t_p	3,1214	3,1559	3,1947	3,2389	3,2905	3,3528	3,4316	3,5401	3,7190



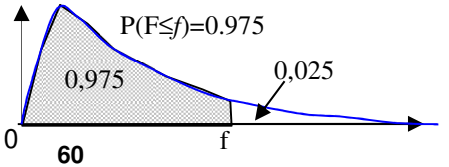
Loi de Student bilatérale à υ degrés de liberté, détermination de t_p pour p=P(|St|≤ t_p) connue

Risque bilatéral α	*80%*	*60%*	*40%*	*20%*	*10%*	*5%*	*2%*	*1%*	*0,5%*	*0,1%*
Probabilité p=1-α	*0,2*	*0,4*	*0,6*	*0,8*	*0,9*	*0,95*	*0,98*	*0,99*	*0,995*	*0,999*
υ=*1*	**0,325**	**0,727**	**1,376**	**3,078**	**6,314**	**12,706**	**31,821**	**63,656**	**318,289**	**636,578**
υ=*2*	**0,289**	**0,617**	**1,061**	**1,886**	**2,920**	**4,303**	**6,965**	**9,925**	**22,328**	**31,600**
υ=*3*	**0,277**	**0,584**	**0,978**	**1,638**	**2,353**	**3,182**	**4,541**	**5,841**	**10,214**	**12,924**
υ=*4*	**0,271**	**0,569**	**0,941**	**1,533**	**2,132**	**2,776**	**3,747**	**4,604**	**7,173**	**8,610**
υ=*5*	**0,267**	**0,559**	**0,920**	**1,476**	**2,015**	**2,571**	**3,365**	**4,032**	**5,894**	**6,869**
υ=*6*	**0,265**	**0,553**	**0,906**	**1,440**	**1,943**	**2,447**	**3,143**	**3,707**	**5,208**	**5,959**
υ=*7*	**0,263**	**0,549**	**0,896**	**1,415**	**1,895**	**2,365**	**2,998**	**3,499**	**4,785**	**5,408**
υ=*8*	**0,262**	**0,546**	**0,889**	**1,397**	**1,860**	**2,306**	**2,896**	**3,355**	**4,501**	**5,041**
υ=*9*	**0,261**	**0,543**	**0,883**	**1,383**	**1,833**	**2,262**	**2,821**	**3,250**	**4,297**	**4,781**
υ=*10*	**0,260**	**0,542**	**0,879**	**1,372**	**1,812**	**2,228**	**2,764**	**3,169**	**4,144**	**4,587**
υ=*11*	**0,260**	**0,540**	**0,876**	**1,363**	**1,796**	**2,201**	**2,718**	**3,106**	**4,025**	**4,437**
υ=*12*	**0,259**	**0,539**	**0,873**	**1,356**	**1,782**	**2,179**	**2,681**	**3,055**	**3,930**	**4,318**
υ=*13*	**0,259**	**0,538**	**0,870**	**1,350**	**1,771**	**2,160**	**2,650**	**3,012**	**3,852**	**4,221**
υ=*14*	**0,258**	**0,537**	**0,868**	**1,345**	**1,761**	**2,145**	**2,624**	**2,977**	**3,787**	**4,140**
υ=*15*	**0,258**	**0,536**	**0,866**	**1,341**	**1,753**	**2,131**	**2,602**	**2,947**	**3,733**	**4,073**
υ=*16*	**0,258**	**0,535**	**0,865**	**1,337**	**1,746**	**2,120**	**2,583**	**2,921**	**3,686**	**4,015**
υ=*17*	**0,257**	**0,534**	**0,863**	**1,333**	**1,740**	**2,110**	**2,567**	**2,898**	**3,646**	**3,965**
υ=*18*	**0,257**	**0,534**	**0,862**	**1,330**	**1,734**	**2,101**	**2,552**	**2,878**	**3,610**	**3,922**
υ=*19*	**0,257**	**0,533**	**0,861**	**1,328**	**1,729**	**2,093**	**2,539**	**2,861**	**3,579**	**3,883**
υ=*20*	**0,257**	**0,533**	**0,860**	**1,325**	**1,725**	**2,086**	**2,528**	**2,845**	**3,552**	**3,850**
υ=*21*	**0,257**	**0,532**	**0,859**	**1,323**	**1,721**	**2,080**	**2,518**	**2,831**	**3,527**	**3,819**
υ=*22*	**0,256**	**0,532**	**0,858**	**1,321**	**1,717**	**2,074**	**2,508**	**2,819**	**3,505**	**3,792**
υ=*23*	**0,256**	**0,532**	**0,858**	**1,319**	**1,714**	**2,069**	**2,500**	**2,807**	**3,485**	**3,768**
υ=*24*	**0,256**	**0,531**	**0,857**	**1,318**	**1,711**	**2,064**	**2,492**	**2,797**	**3,467**	**3,745**
υ=*25*	**0,256**	**0,531**	**0,856**	**1,316**	**1,708**	**2,060**	**2,485**	**2,787**	**3,450**	**3,725**
υ=*26*	**0,256**	**0,531**	**0,856**	**1,315**	**1,706**	**2,056**	**2,479**	**2,779**	**3,435**	**3,707**
υ=*27*	**0,256**	**0,531**	**0,855**	**1,314**	**1,703**	**2,052**	**2,473**	**2,771**	**3,421**	**3,689**
υ=*28*	**0,256**	**0,530**	**0,855**	**1,313**	**1,701**	**2,048**	**2,467**	**2,763**	**3,408**	**3,674**
υ=*29*	**0,256**	**0,530**	**0,854**	**1,311**	**1,699**	**2,045**	**2,462**	**2,756**	**3,396**	**3,660**
υ=*30*	**0,256**	**0,530**	**0,854**	**1,310**	**1,697**	**2,042**	**2,457**	**2,750**	**3,385**	**3,646**
υ=*∞*	**0,253**	**0,524**	**0,842**	**1,282**	**1,645**	**1,960**	**2,326**	**2,576**	**3,090**	**3,291**



Loi du χ^2 à υ degrés de liberté, détermination de X^2 pour p= P(X≤X^2) connue

probabilité	*0,001*	*0,005*	*0,01*	*0,025*	*0,05*	*0,1*	*0,5*	*0,9*	*0,95*	*0,975*	*0,99*	*0,995*	*0,999*	*0,9995*
υ=*1*				**0,001**	**0,004**	**0,02**	**0,45**	**2,71**	**3,84**	**5,02**	**6,63**	**7,88**	**10,83**	**12,12**
υ=*2*	**0,00**	**0,01**	**0,02**	**0,05**	**0,10**	**0,21**	**1,39**	**4,61**	**5,99**	**7,38**	**9,21**	**10,60**	**13,82**	**15,20**
υ=*3*	**0,02**	**0,07**	**0,11**	**0,22**	**0,35**	**0,58**	**2,37**	**6,25**	**7,81**	**9,35**	**11,34**	**12,84**	**16,27**	**17,73**
υ=*4*	**0,09**	**0,21**	**0,30**	**0,48**	**0,71**	**1,06**	**3,36**	**7,78**	**9,49**	**11,14**	**13,28**	**14,86**	**18,47**	**20,00**
υ=*5*	**0,21**	**0,41**	**0,55**	**0,83**	**1,15**	**1,61**	**4,35**	**9,24**	**11,07**	**12,83**	**15,09**	**16,75**	**20,51**	**22,11**
υ=*6*	**0,38**	**0,68**	**0,87**	**1,24**	**1,64**	**2,20**	**5,35**	**10,64**	**12,59**	**14,45**	**16,81**	**18,55**	**22,46**	**24,10**
υ=*7*	**0,60**	**0,99**	**1,24**	**1,69**	**2,17**	**2,83**	**6,35**	**12,02**	**14,07**	**16,01**	**18,48**	**20,28**	**24,32**	**26,02**
υ=*8*	**0,86**	**1,34**	**1,65**	**2,18**	**2,73**	**3,49**	**7,34**	**13,36**	**15,51**	**17,53**	**20,09**	**21,95**	**26,12**	**27,87**
υ=*9*	**1,15**	**1,73**	**2,09**	**2,70**	**3,33**	**4,17**	**8,34**	**14,68**	**16,92**	**19,02**	**21,67**	**23,59**	**27,88**	**29,67**
υ=*10*	**1,48**	**2,16**	**2,56**	**3,25**	**3,94**	**4,87**	**9,34**	**15,99**	**18,31**	**20,48**	**23,21**	**25,19**	**29,59**	**31,42**
υ=*11*	**1,83**	**2,60**	**3,05**	**3,82**	**4,57**	**5,58**	**10,34**	**17,28**	**19,68**	**21,92**	**24,73**	**26,76**	**31,26**	**33,14**
υ=*12*	**2,21**	**3,07**	**3,57**	**4,40**	**5,23**	**6,30**	**11,34**	**18,55**	**21,03**	**23,34**	**26,22**	**28,30**	**32,91**	**34,82**
υ=*13*	**2,62**	**3,57**	**4,11**	**5,01**	**5,89**	**7,04**	**12,34**	**19,81**	**22,36**	**24,74**	**27,69**	**29,82**	**34,53**	**36,48**
υ=*14*	**3,04**	**4,07**	**4,66**	**5,63**	**6,57**	**7,79**	**13,34**	**21,06**	**23,68**	**26,12**	**29,14**	**31,32**	**36,12**	**38,11**
υ=*15*	**3,48**	**4,60**	**5,23**	**6,26**	**7,26**	**8,55**	**14,34**	**22,31**	**25,00**	**27,49**	**30,58**	**32,80**	**37,70**	**39,72**
υ=*16*	**3,94**	**5,14**	**5,81**	**6,91**	**7,96**	**9,31**	**15,34**	**23,54**	**26,30**	**28,85**	**32,00**	**34,27**	**39,25**	**41,31**
υ=*17*	**4,42**	**5,70**	**6,41**	**7,56**	**8,67**	**10,09**	**16,34**	**24,77**	**27,59**	**30,19**	**33,41**	**35,72**	**40,79**	**42,88**
υ=*18*	**4,90**	**6,26**	**7,01**	**8,23**	**9,39**	**10,86**	**17,34**	**25,99**	**28,87**	**31,53**	**34,81**	**37,16**	**42,31**	**44,43**
υ=*19*	**5,41**	**6,84**	**7,63**	**8,91**	**10,12**	**11,65**	**18,34**	**27,20**	**30,14**	**32,85**	**36,19**	**38,58**	**43,82**	**45,97**
υ=*20*	**5,92**	**7,43**	**8,26**	**9,59**	**10,85**	**12,44**	**19,34**	**28,41**	**31,41**	**34,17**	**37,57**	**40,00**	**45,31**	**47,50**
υ=*21*	**6,45**	**8,03**	**8,90**	**10,28**	**11,59**	**13,24**	**20,34**	**29,62**	**32,67**	**35,48**	**38,93**	**41,40**	**46,80**	**49,01**
υ=*22*	**6,98**	**8,64**	**9,54**	**10,98**	**12,34**	**14,04**	**21,34**	**30,81**	**33,92**	**36,78**	**40,29**	**42,80**	**48,27**	**50,51**
υ=*23*	**7,53**	**9,26**	**10,20**	**11,69**	**13,09**	**14,85**	**22,34**	**32,01**	**35,17**	**38,08**	**41,64**	**44,18**	**49,73**	**52,00**
υ=*24*	**8,08**	**9,89**	**10,86**	**12,40**	**13,85**	**15,66**	**23,34**	**33,20**	**36,42**	**39,36**	**42,98**	**45,56**	**51,18**	**53,48**
υ=*25*	**8,65**	**10,52**	**11,52**	**13,12**	**14,61**	**16,47**	**24,34**	**34,38**	**37,65**	**40,65**	**44,31**	**46,93**	**52,62**	**54,95**
υ=*26*	**9,22**	**11,16**	**12,20**	**13,84**	**15,38**	**17,29**	**25,34**	**35,56**	**38,89**	**41,92**	**45,64**	**48,29**	**54,05**	**56,41**
υ=*27*	**9,80**	**11,81**	**12,88**	**14,57**	**16,15**	**18,11**	**26,34**	**36,74**	**40,11**	**43,19**	**46,96**	**49,65**	**55,48**	**57,86**
υ=*28*	**10,39**	**12,46**	**13,56**	**15,31**	**16,93**	**18,94**	**27,34**	**37,92**	**41,34**	**44,46**	**48,28**	**50,99**	**56,89**	**59,30**
υ=*29*	**10,99**	**13,12**	**14,26**	**16,05**	**17,71**	**19,77**	**28,34**	**39,09**	**42,56**	**45,72**	**49,59**	**52,34**	**58,30**	**60,73**
υ=*30*	**11,59**	**13,79**	**14,95**	**16,79**	**18,49**	**20,60**	**29,34**	**40,26**	**43,77**	**46,98**	**50,89**	**53,67**	**59,70**	**62,16**



Loi F de Fischer-Snedecor F(ddl1,ddl2, 0,025) risque α=2,5%

Ddl1 →

Ddl2 ↓

Ddl2 \ Ddl1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	976,7	984,9	993,1	998,1	1001	1005	1009
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,50	8,46	8,41	8,36
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,27	6,23	6,18	6,12
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,11	5,07	5,01	4,96
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,40	4,36	4,31	4,25
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,94	3,89	3,84	3,78
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,60	3,56	3,51	3,45
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,35	3,31	3,26	3,20
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,16	3,12	3,06	3,00
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,01	2,96	2,91	2,85
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,88	2,84	2,78	2,72
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,78	2,73	2,67	2,61
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,69	2,64	2,59	2,52
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,61	2,57	2,51	2,45
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,55	2,50	2,44	2,38
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,49	2,44	2,38	2,32
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,44	2,39	2,33	2,27
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,40	2,35	2,29	2,22
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,36	2,31	2,25	2,18
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,32	2,27	2,21	2,14
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,29	2,24	2,18	2,11
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,26	2,21	2,15	2,08
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,23	2,18	2,12	2,05
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,21	2,16	2,09	2,03
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,18	2,13	2,07	2,00
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,16	2,11	2,05	1,98
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,14	2,09	2,03	1,96
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,12	2,07	2,01	1,94
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	1,99	1,94	1,88	1,80
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32	2,22	2,11	1,99	1,92	1,87	1,80	1,72
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,87	1,82	1,74	1,67