

Table des matières

STATISTIQUE DESCRIPTIVE :	9
1. VARIABLE STATISTIQUE QUANTITATIVE DISCRETE	12
1.1. DÉFINITION DES PARAMÈTRES USUELS	12
1.1.1. mode	12
1.1.2. étendue	12
1.1.3. médiane Q_2 et quartiles Q_1, Q_2, Q_3	12
1.1.4. moyenne	13
1.1.5. variance, moment centré d'ordre 2, écart-type	13
1.1.6. moment centré d'ordre 3, coefficient d'asymétrie	13
1.1.7. moment centré d'ordre 4, coefficient d'aplatissement	13
1.2. FONCTION CUMULATIVE CROISSANTE (FONCTION DE RÉPARTITION)	14
2. VARIABLE STATISTIQUE QUANTITATIVE CONTINUE	15
2.1. DÉFINITION DES PARAMÈTRES USUELS	15
2.1.1. mode	15
2.1.2. étendue	15
2.1.3. médiane Q_2 et quartiles Q_1, Q_2, Q_3	15
2.1.4. Interpolation linéaire	16
2.1.5. moyenne	16
2.1.6. variance, moment centré d'ordre 2 & écart-type	16
2.1.7. moment centré d'ordre 3, coefficient d'asymétrie	16
2.1.8. moment centré d'ordre 4, coefficient d'aplatissement	17
2.2. FONCTION CUMULATIVE CROISSANTE (FONCTION DE RÉPARTITION)	17
2.3. COMPARAISON D'UNE VARIABLE X CONTINUE AVEC UNE VARIABLE DE LAPLACE-GAUSS LG(μ, σ) DE PARAMÈTRES $\bar{X} = \mu$ ET $\sigma_X = \sigma, \gamma_1 = 0, \beta_1 = 0, \gamma_2 = 3, \beta_2 = 0$	18
2.4. EXTRAIT DE LA TABLE DONNANT LA DISTRIBUTION DES FRÉQUENCES DE LA VARIABLE CENTRÉE RÉDUITE DE LAPLACE-GAUSS	19
3. QUELQUES REMARQUES IMPORTANTES	20
3.1. CHANGEMENT DE VARIABLE POUR UNE VARIABLE QUELCONQUE	20
3.2. DES PROPRIÉTÉS FONDAMENTALES	20
4. VARIABLE STATISTIQUE QUALITATIVE DISCRETE ORDONNEE	21
4.1. DÉFINITION DES PARAMÈTRES USUELS	21
4.1.1. mode	21
4.1.2. étendue	21
4.1.3. médiane Q_2 et quartiles Q_1, Q_2, Q_3	21
4.1.4. entropie	22
4.1.5. entropie relative	22
5. VARIABLE STATISTIQUE QUALITATIVE DISCRETE : VARIABLE NOMINALE	23
5.1. DÉFINITION DES PARAMÈTRES USUELS	23
5.1.1. mode	23
5.1.2. entropie	23
5.1.3. entropie relative	23
6. TRAITEMENTS GRAPHIQUES	24
7. REPRÉSENTATION BARYCENTRIQUE DANS LE TRIANGLE ÉQUILATÉRAL	27
7.1. PROPRIÉTÉS DU TRIANGLE ÉQUILATÉRAL	27
7.2. DÉFINITION DU BARYCENTRE :	28

7.3. USAGE DES PROPRIÉTÉS POUR RÉALISER UNE REPRÉSENTATION DES DONNÉES EN STATISTIQUE :	28
8. REPRÉSENTATION BARYCENTRIQUE DANS LE CARRÉ	29
8.1. THÉORÈME D'ASSOCIATIVITÉ :	29
8.2. PROCÉDURE DE CONSTRUCTION	29
9. LES MOYENNES D'UNE VARIABLE STATISTIQUE QUANTITATIVE DISCRETE	30
9.1. MOYENNE (ARITHMÉTIQUE)	30
9.2. MOYENNE GÉOMÉTRIQUE	30
9.3. MOYENNE HARMONIQUE	30
9.4. MOYENNE D'ORDRE 2 : MOYENNE QUADRATIQUE (MOMENT D'ORDRE 2)	30
9.5. MOYENNE D'ORDRE M	31
9.6. MOYENNE TRONQUÉE D'ORDRE 1	31
9.7. MOYENNE TRONQUÉE D'ORDRE Q	31
9.8. MOYENNE DE WINSOR D'ORDRE Q	31
9.9. REMARQUE IMPORTANTE SUR L'EXISTENCE D'UNE MOYENNE :	31
CALCULER :	32
10. ANALYSE COMBINATOIRE : CALCUL DE DÉNOMBREMENT.	32
EFFETS DES APPROXIMATIONS DANS LES CALCULS EN STATISTIQUE : DANGER!	
APPROXIMATIONS	34
A PROPOS DE L'ARTICLE D'OUEST FRANCE (FÉVRIER 1992)	36
ESTIMATION :	42
11. ESTIMER UN PARAMÈTRE	42
12. ESTIMATION D'UNE MOYENNE μ ET D'UNE VARIANCE σ^2	44
12.1. CONDITIONS D'UTILISATION:	44
12.2. ESTIMATEUR:	44
12.3. ESTIMATION PONCTUELLE DE LA MOYENNE:	44
12.4. ESTIMATION PONCTUELLE DE LA VARIANCE DE LA POPULATION:	44
12.5. ESTIMATION PAR INTERVALLE DE CONFIANCE BILATÉRAL SYMÉTRIQUE:	44
12.6. ESTIMATION PAR INTERVALLE DE CONFIANCE À DROITE:	45
12.7. ESTIMATION PAR INTERVALLE DE CONFIANCE À GAUCHE:	45
12.8. TAILLE DE L'ÉCHANTILLON POUR UNE PRÉCISION FIXÉE:	45
12.9. COMPLÉMENTS ET REMARQUES À PROPOS DE L'ESTIMATION D'UNE MOYENNE ET D'UNE VARIANCE.	45
13. ESTIMATION D'UNE PROPORTION π	46
13.1. CONDITIONS D'UTILISATION (CAS N°1):	46
13.2. ESTIMATEUR:	46
13.3. ESTIMATION PONCTUELLE:	46
13.4. ESTIMATION PAR INTERVALLE DE CONFIANCE BILATÉRAL SYMÉTRIQUE:	46
13.5. ESTIMATION PAR INTERVALLE DE CONFIANCE À DROITE:	47
13.6. ESTIMATION PAR INTERVALLE DE CONFIANCE À GAUCHE:	47
13.7. TAILLE DE L'ÉCHANTILLON POUR UNE PRÉCISION FIXÉE:	47
13.8. COMPLÉMENTS ET REMARQUES À PROPOS DE L'ESTIMATION D'UNE PROPORTION.	47
TESTER UNE HYPOTHESE	48

14. QUELQUES IDÉES GÉNÉRALES SUR LES TESTS STATISTIQUES.....	48
15. COMPARAISON DE DEUX PROPORTIONS π_1, π_2.....	51
15.1. CONDITIONS D'UTILISATION:	51
15.2. STATISTIQUE	51
15.3. TEST BILATÉRAL SYMÉTRIQUE: $H_0 (\pi_1 = \pi_2)$ CONTRE $H_1 (\pi_1 \neq \pi_2)$	51
15.4. TEST UNILATÉRAL À DROITE: $H_0 (\pi_1 \geq \pi_2)$ CONTRE $H_1 (\pi_1 < \pi_2)$	52
15.5. TEST UNILATÉRAL À GAUCHE: $H_0 (\pi_1 \leq \pi_2)$ CONTRE $H_1 (\pi_1 > \pi_2)$	52
15.6. COMPLÉMENTS ET REMARQUES À PROPOS DE L'ESTIMATION D'UNE PROPORTION.	52
16. COMPARAISON D'UNE PROPORTION π À UNE VALEUR π_0	53
16.1. CONDITIONS D'UTILISATION:	53
16.2. STATISTIQUE ET VARIABLE DE DÉCISION UTILISÉE.....	53
16.3. TEST BILATÉRAL SYMÉTRIQUE: $H_0 (\pi = \pi_0)$ CONTRE $H_1 (\pi \neq \pi_0)$	53
16.4. TEST UNILATÉRAL À DROITE: $H_0 (\pi \geq \pi_0)$ CONTRE $H_1 (\pi < \pi_0)$	54
16.5. TEST UNILATÉRAL À GAUCHE: $H_0 (\pi \leq \pi_0)$ CONTRE $H_1 (\pi > \pi_0)$	54
16.6. COMPLÉMENTS ET REMARQUES À PROPOS DE L'ESTIMATION D'UNE PROPORTION.	54
17. COMPARAISON D'UNE MOYENNE μ À UNE VALEUR DONNÉE μ_0	55
17.1. CONDITIONS D'UTILISATION:	55
17.2. STATISTIQUE ET VARIABLE DE DÉCISION UTILISÉE.....	55
17.3. TEST BILATÉRAL SYMÉTRIQUE: $H_0 (\mu = \mu_0)$ CONTRE $H_1 (\mu \neq \mu_0)$	55
17.4. TEST UNILATÉRAL À DROITE: $H_0 (\mu \geq \mu_0)$ CONTRE $H_1 (\mu < \mu_0)$	56
17.5. TEST UNILATÉRAL À GAUCHE: $H_0 (\mu \leq \mu_0)$ CONTRE $H_1 (\mu > \mu_0)$	56
17.6. COMPLÉMENTS ET REMARQUES À PROPOS DE L'ESTIMATION D'UNE PROPORTION.	56
18. COMPARAISON DE DEUX VARIANCES σ_1^2 ET σ_2^2	57
18.1. CONDITIONS D'UTILISATION:	57
18.2. STATISTIQUE ET VARIABLE DE DÉCISION UTILISÉE.....	57
18.3. TEST BILATÉRAL SYMÉTRIQUE: $H_0 (\sigma_1^2 = \sigma_2^2)$ CONTRE $H_1 (\sigma_1^2 \neq \sigma_2^2)$	58
18.4. COMPLÉMENTS ET REMARQUES À PROPOS DE L'ESTIMATION D'UNE PROPORTION.	58
19. COMPARAISON DE DEUX MOYENNES μ_1 ET μ_2 À PARTIR DE DEUX ÉCHANTILLONS INDÉPENDANTS.....	59
19.1. CONDITIONS D'UTILISATION:	59
19.2. STATISTIQUE ET VARIABLE DE DÉCISION UTILISÉE:	59
19.3. TEST BILATÉRAL SYMÉTRIQUE: $H_0 (\mu_1 - \mu_2 = \delta_0)$ CONTRE $H_1 (\mu_1 - \mu_2 \neq \delta_0)$	60
19.4. COMPLÉMENTS ET REMARQUES À PROPOS DE L'ESTIMATION D'UNE PROPORTION.	60
20. TEST DE COMPARAISON DES MOYENNES DE DEUX ÉCHANTILLONS APPARIÉS : LE TEST T DE STUDENT.....	61
20.1. CONDITIONS D'UTILISATION:	61
20.2. STATISTIQUE ET VARIABLE DE DÉCISION UTILISÉE:	61
20.3. TEST BILATÉRAL SYMÉTRIQUE: $H_0 (\mu_\Delta = 0)$ CONTRE $H_1 (\mu_\Delta \neq 0)$	61
21. TEST D'INDÉPENDANCE DE DEUX VARIABLES.....	63
21.1. CAS DES TABLEAUX 2 X 2	63
21.2. TEST DU KHI-DEUX D'INDÉPENDANCE	63
21.3. STATISTIQUES ET VARIABLES DE DÉCISION UTILISÉES.....	64

22. TEST EXACT DE FISCHER	65
23. ÉTUDE SIMULTANÉE DE DEUX VARIABLES QUALITATIVES	67
23.1. INTRODUCTION	67
23.2. TABLEAU CROISÉ.....	67
23.2.1. Transformations du tableau croisé.....	69
23.2.2. Représentations graphiques du tableau croisé.....	71
23.2.3. Raisonner statistiquement à partir du tableau croisé.....	71
23.3. LA NOTION FONDAMENTALE D'INDÉPENDANCE STATISTIQUE.....	71
23.3.1. Caractérisation de l'hypothèse d'indépendance.....	71
23.4. UNE MESURE D'ASSOCIATION : LE χ^2	73
23.4.1. Remarques et conditions d'utilisation :	74
23.5. LE TEST DU χ^2 D'INDÉPENDANCE DE DEUX VARIABLES QUALITATIVES	74
23.5.1. La démarche du test du χ^2	74
23.6. RETOUR SUR LE TABLEAU DE CONTINGENCE	76
23.6.1. Contribution d'une "case" à la valeur prise par $D^2 = \chi^2$:.....	77
23.6.2. Mesures d'association dérivées de la valeur prise par $D^2 = \chi^2$:.....	77
23.6.2.1. coefficient de contingence de Karl Pearson :.....	77
23.6.2.2. coefficient de Tschuprow :.....	77
23.6.2.3. coefficient de Cramer :.....	77
23.7. AUTRES UTILISATIONS DE LA MESURE D'ASSOCIATION χ^2	77
24. TEST D'HOMOGENÉITÉ	79
24.1. TEST D'HOMOGENÉITÉ DU KHI-DEUX	79
25. TEST DU χ^2 D'ADÉQUATION.....	80
25.1. PRÉSENTATION DU PREMIER EXEMPLE.....	80
25.2. LA DÉMARCHE DU TEST DU χ^2 D'ADÉQUATION.....	80
25.3. PRÉSENTATION DU SECOND EXEMPLE : TEST D'ADÉQUATION À UNE DISTRIBUTION DE LA LAPLACE-GAUSS	81
26. TEST DU χ^2 DE MAC NEMAR	84
27. ÉTUDE SIMULTANÉE DE DEUX VARIABLES QUANTITATIVES	86
27.1. LIAISON ENTRE DEUX VARIABLES QUANTITATIVES X ET Y : COEFFICIENT DE CORRELATION DE BRAVAIS-PEARSON	86
27.1.1. Caractère significatif du coefficient de corrélation de Bravais-Pearson :	87
STATISTIQUE DE RANG	89
28. TEST DE L'HYPOTHÈSE D'UN ÉCHANTILLON ALÉATOIRE. (PROBLÈME À UN ÉCHANTILLON).....	91
28.1. LE COEFFICIENT DE CORRÉLATION DES RANGS R_S DE SPEARMAN	92
28.2. LE COEFFICIENT DE CORRÉLATION DE RANG τ DE KENDALL.....	93
28.2.1. Méthodes de calcul:	93
28.3. COMPARAISON DES COEFFICIENTS DE SPEARMAN ET DE KENDALL.....	94
28.4. LE TEST DES "SIGNES"	95
29. TEST D'INDÉPENDANCE (PROBLÈME À DEUX ÉCHANTILLONS).....	98
29.1. LE COEFFICIENT DE CORRÉLATION DES RANGS R_S DE SPEARMAN.....	98
29.1.1. Cas d'ex æquo :.....	99
29.2. LE COEFFICIENT DE CORRÉLATION DE RANG τ DE KENDALL.....	100

29.2.1.	<i>Méthodes de calcul:</i>	100
29.2.2.	<i>Cas d'ex æquo :</i>	102
30.	TEST D'HOMOGENÉITÉ (PROBLÈME À DEUX ÉCHANTILLONS)	103
30.1.	LE TEST DE WILCOXON	103
30.1.1.	<i>Cas d'ex æquo :</i>	104
30.2.	LE TEST DE MANN ET WHITNEY	105
30.2.1.	<i>Conditions d'utilisation:</i>	105
30.2.2.	<i>Statistique et variable de décision</i>	105
30.2.3.	<i>Test bilatéral :</i>	105
30.2.4.	<i>Test unilatéral :</i>	106
30.3.	LE TEST DE LA MÉDIANE.....	106
30.3.1.	<i>cas où $N = 2p$:</i>	107
30.3.2.	<i>cas où $N = 2p+1$:</i>	107
30.3.3.	<i>Test bilatéral :</i>	107
30.3.4.	<i>Test unilatéral :</i>	108
30.3.5.	<i>Décision :</i>	108
31.	TEST D'INDÉPENDANCE (PROBLÈME À K ÉCHANTILLONS)	109
1.1.	TEST DE CONCORDANCE DE P RANGEMENTS DE N OBJETS DE M.G. KENDALL	109
31.1.1.	<i>Méthode de calcul:</i>	109
31.2.	TEST DE CONCORDANCE DE P RANGEMENTS DE N OBJETS DE FRIEDMAN	111
31.2.1.	<i>Méthode de calcul:</i>	111
32.	TEST D'HOMOGENÉITÉ (PROBLÈME À K ÉCHANTILLONS)	112
1.1.	COMPARAISON DE K ÉCHANTILLONS INDÉPENDANTS TEST H DE KRUSKAL-WALLIS.....	112
32.1.1.	<i>Conditions d'utilisation:</i>	112
32.1.2.	<i>Statistique et variable de décision</i>	112
32.1.3.	<i>Test unilatéral : H_0 (identité des K distributions) contre H_1 (deux distributions au moins sont différentes)</i>	113
32.1.4.	<i>En cas d'existence d'ex æquo</i>	113
33.	POUR ALLER ENCORE PLUS LOIN...	114
	ANALYSE DE VARIANCE	117
33.1.	ANALYSE DE VARIANCE À UN CRITÈRE	117
33.1.1.	<i>Conditions d'utilisation:</i>	117
33.1.2.	<i>Mise en place des notations pour le traitement statistique:</i>	117
33.1.3.	<i>Mise en place du modèle statistique:</i>	119
33.1.4.	<i>Estimation des effets du facteur contrôlé A:</i>	120
33.1.5.	<i>Tableau d'analyse de variance:</i>	121
33.1.6.	<i>Interprétation des résultats:</i>	121
33.1.7.	<i>Comparaison des variances des groupes A_i:</i>	122
	PROBABILITES	123
34.	DES OUTILS THÉORIQUES POUR UNE MODÉLISATION PROBABILISTE, POUR ESTIMER, POUR TESTER	123
34.1.	EXPERIENCE ALEATOIRE ET EVENEMENTS :	123
34.1.1.1.	Cas particuliers :	125
34.1.2.	<i>ESPACE PROBABILISÉ :</i>	125
34.1.3.	<i>Quelques propriétés usuelles:</i>	126
34.1.4.	<i>Produit d'espaces probabilisés :</i>	127
34.1.5.	<i>Lois de probabilité conditionnelle et indépendance :</i>	127
34.1.6.	<i>variable aléatoire réelle et loi de probabilité</i>	128
34.1.6.1.	Variable aléatoires discrètes.....	129
34.1.6.2.	Variable aléatoires (absolument) continues.....	129

34.1.6.3.	Fonction de répartition d'une variable aléatoire X	129
34.1.7.	<i>Indépendance de deux variables aléatoires</i>	129
34.1.8.	<i>Valeurs caractéristiques d'une variable aléatoire</i>	130
34.1.8.1.	Espérance mathématique de X	130
34.1.8.2.	Propriétés de l'espérance mathématique:.....	130
34.1.8.3.	Variance de X dont l'espérance est $E(X)=m$	130
34.1.8.4.	écart-type de X.....	130
34.1.8.5.	Propriétés de la variance:.....	130
34.1.8.6.	Lien entre l'espérance et l'écart-type : l'inégalité de Bienaymé-Tchebychev	131
34.1.8.7.	Variable centrée réduite	131
34.1.8.8.	Moments centrés d'ordre k de X.....	131
34.1.8.9.	Médiane(s) d'une variable aléatoire.....	131
34.1.8.10.	Quartiles d'une variable aléatoire	131
34.1.9.	<i>Égalité presque sûre de deux variables aléatoires</i>	132
34.2.	VARIABLES ALÉATOIRES ET LOIS DE PROBABILITÉ USUELLES.....	132
34.3.	VARIABLE ALÉATOIRE BINOMIALE ET APPROXIMATION DE SA DISTRIBUTION DE PROBABILITÉ PAR CELLE DE LA VARIABLE DE LAPLACE-GAUSS.....	136
34.4.	PROLONGEMENT DE LA VARIABLE BINOMIALE DISCRÈTE À VALEURS ENTIÈRES DANS N À UNE VARIABLE CONTINUE À VALEURS RÉELLES DANS R	138
34.4.1.	<i>Quelques relations exactes entre les distributions usuelles :</i>	141
34.4.1.1.	Loi de probabilité de la variable de Poisson et loi de probabilité de la variable de Fisher-Snédecor : 141	
1.1.1.2.	Loi de probabilité de la variable binomiale et loi de probabilité de la variable de Fisher-Snédecor : 141	
34.4.1.3.	Loi de probabilité de la variable T_n de Student et loi de probabilité de la variable de Fisher- Snédecor : 141	
34.4.2.	<i>Quelques approximations des distributions usuelles :</i>	142
34.4.2.1.	Convergence en loi de la variable binomiale $B(n,p)$ vers la variable de Laplace-Gauss.....	142
34.4.2.2.	Convergence en loi de la variable de Poisson $P(m)$ vers la variable de Laplace-Gauss	142
34.4.2.3.	Convergence en loi de la variable binomiale $B(n,p)$ vers la variable de Poisson $P(\lambda)$	142
34.4.2.4.	Convergence en loi de la variable hypergéométrique $H(N,n,p)$ vers la variable binomiale $B(n,p)$ 142	
34.4.3.	<i>Formules approchées de fonction de répartition de quelques variables continues usuelles :</i> 142	
34.4.3.1.	Variable de Pearson : variable du $\chi^2(n)$	142
34.4.3.2.	Variable de Fisher-Snedecor : variable $F(m,n)$	142
34.4.3.3.	Variable de "Student" : variable T_n	143
ECHANTILLONNAGE		145
35. ÉCHANTILLONNAGE D'UNE VARIABLE		145
35.1.	N-ÉCHANTILLON DE LA VARIABLE X.....	145
35.2.	RÉALISATION D'UN N-ÉCHANTILLON DE LA VARIABLE X	145
35.3.	STATISTIQUE ET LOI D'ÉCHANTILLONNAGE	145
36. QUELQUES STATISTIQUES USUELLES.....		145
36.1.	LA VARIABLE ALÉATOIRE «MOYENNE EMPIRIQUE» EST DÉFINIE PAR.....	146
36.2.	LA VARIABLE ALÉATOIRE «VARIANCE EMPIRIQUE» EST DÉFINIE PAR	146
36.3.	LA VARIABLE ALÉATOIRE «VARIANCE EMPIRIQUE MODIFIÉE» EST DÉFINIE PAR	146
36.4.	LA VARIABLE ALÉATOIRE «VARIANCE EMPIRIQUE MODIFIÉE» EST DÉFINIE PAR	146
36.5.	LA VARIABLE ALÉATOIRE «FRÉQUENCE EMPIRIQUE» EST DÉFINIE PAR.....	147
36.6.	LA VARIABLE ALÉATOIRE «FONCTION DE RÉPARTITION EMPIRIQUE».....	147
36.7.	LA VARIABLE ALÉATOIRE «COEFFICIENT DE CORRÉLATION EMPIRIQUE»	147
TABLES STATISTIQUES :		149
<u>VARIABLE ALEATOIRE DE PEARSON DE TYPE III</u>		152

VARIABLE ALEATOIRE DE PEARSON DE TYPE VI	155
COMPLEMENTS: CALCULER AVEC LES TABLES	156

STATISTIQUE DESCRIPTIVE :

Si le mathématicien se place toujours, même de façon implicite, sur un référentiel (ou univers) dont il étudie les sous-ensembles, les propriétés, les transformations, etc., le statisticien appelle cet ensemble de référence une **population**. Une **unité statistique** est un élément de la population dont les sous-ensembles sont appelés **échantillons**. Pour recueillir des données, le statisticien procède à des “*mesures*” sur les unités statistiques. Il observe sur ces unités, un même phénomène, une même propriété appelée **caractère statistique**. La *mesure* du caractère statistique s’effectue à l’aide d’une **variable statistique** qui peut être **quantitative** ou **qualitative**, **discrète** ou **continue**, **ordonnée** ou **non**. A chaque unité statistique, on fait correspondre le résultat observé du caractère statistique *mesuré* par la variable statistique. Cette application — au sens mathématique du terme — entre l’ensemble des unités statistiques et l’ensemble des **valeurs** prises par la variable statistique quantitative ou des **modalités** prises par la variable qualitative est appelée **série statistique**.

Le nombre d’unités statistiques de la population est l’**effectif (total)**. Ce nombre est le cardinal de l’ensemble. Il peut être connu ou inconnu, fini ou infini.

Dans le cas d’une étude sur une population finie ou un échantillon fini, la série statistique peut être représentée par un tableau dont les lignes représentent les unités statistiques, les individus, et la colonne représente la variable étudiée, chaque case comporte le résultat correspondant. Ce tableau permet de rapporter conjointement les résultats relatifs à plusieurs variables.

Tableau 1-1 des séries statistiques

unités statistiques	variable n°1	...	variable n°q
<i>code de l’unité</i>	<i>valeur</i> ou <i>intervalle</i> ou <i>modalité</i>	...	<i>valeur</i> ou <i>intervalle</i> ou <i>modalité</i>
...

L’observation des faits et la collecte de ces valeurs ou de ces modalités constituent un problème plus complexe qu’il n’y paraît.

Tableau 1-2 des précautions à prendre avant la collecte des données

De nombreuses précautions doivent être prises pour limiter les effets des ambiguïtés. Chacune des opérations est à considérer avec soin avant de réaliser la collecte des informations :

- ◆ définir le but de l'étude avec précision,
- ◆ caractériser les faits élémentaires à observer ou à expérimenter,
- ◆ caractériser la population ou l'échantillon auquel l'étude est référée,
- ◆ lister les précautions à prendre,
- ◆ circonscrire le champ des observations ou des expérimentations,
- ◆ élaborer un modèle statistique dans le cadre duquel le traitement des données s'effectuera,
- ◆ intégrer l'idée de l'imprévisible et celle de risque,

Le relevé des informations peut être

- exhaustif (recensement) ou partiel (sondage)
- continu, périodique ou occasionnel
- direct ou indirect .

Après la collecte des données a lieu le dépouillement. Cette activité comporte deux phases:

On **regroupe** d'abord les unités statistiques pour lesquelles la variable statistique prend la même valeur, ou presque dans le cas d'une variable continue, ou la même modalité dans le cas d'une variable qualitative. Ces groupements forment les **catégories ou les classes du caractère statistique**. Une catégorie est caractérisée soit par une valeur de la variable quantitative, soit par un intervalle dans le cas d'une variable quantitative continue, soit par une modalité de la variable qualitative. Ces catégories ou classes sont constituées d'unités considérées comme équivalentes au regard du caractère étudié. Ce sont des classes d'équivalence au sens mathématique du terme. L'ensemble des catégories ou des classes est l'ensemble-quotient de la population par la relation d'équivalence définie par la variable statistique. Les classes d'équivalence forment une partition de la population, c'est à dire qu'en particulier les catégories ou classes doivent être deux à deux disjointes. Cette remarque a son importance pour définir les **intervalles** dans le cas d'une variable continue.

On procède au **comptage** du nombre d'unités dans chaque catégorie. Ce nombre d'unités est l'**effectif de la catégorie** .

L'application qui à chaque catégorie associe son effectif est appelée **distribution des effectifs de la variable statistique**.

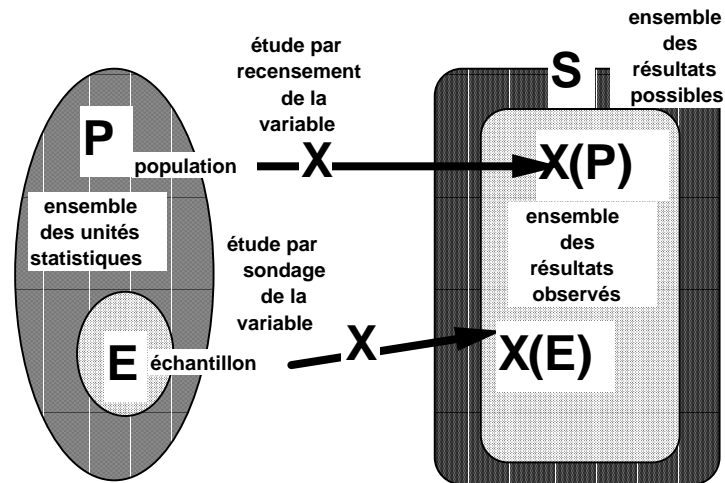
L'application qui à chaque catégorie associe sa fréquence est appelée **distribution des fréquences de la variable statistique**.

Une propriété que le sens commun peut difficilement appréhender concerne les variables statistiques quantitatives **continues**. Dans ce cas, la théorie indique que la fréquence relative à une valeur est nulle mais cela ne signifie pas l'impossibilité

d'apparition de cette valeur. Ceci se traduit par l'existence mathématique d'une fonction **densité de fréquences** qui n'est autre que la fonction dont la représentation graphique est l'**histogramme des fréquences**.

Le dépouillement est l'opération qui permet de passer de la **série statistique** à la **distribution des effectifs**, puis à la **distribution des fréquences**.

Pour schématiser notre propos, nous fournissons la représentation suivante :



1. VARIABLE STATISTIQUE QUANTITATIVE DISCRETE

X une variable statistique quantitative discrète définie sur la population statistique P d'effectif total N.

(i, o_i) avec $i = 1, \dots, N$ la série statistique observée, rangée dans l'ordre de recueil des observations.

(i, o'_i) avec $i = 1, \dots, N$ la série statistique observée, rangée selon l'ordre croissant des valeurs o_i .

(x_k, n_k) avec $k = 1, \dots, p$ la distribution des effectifs de la variable X.

(x_k, f_k) avec $k = 1, \dots, p$ la distribution des fréquences de la variable X.

Les valeurs x_k qui figurent dans le tableau statistique sont rangées dans l'ordre croissant.

valeurs de la variable	x_1	x_2	...	x_k	...	x_{p-1}	x_p	totaux
effectifs	n_1	n_2	...	n_k	...	n_{p-1}	n_p	N
fréquences	f_1	f_2	...	f_k	...	f_{p-1}	f_p	1

avec
$$N = \sum_{k=1}^{k=p} n_k \quad \text{et} \quad f_k = \frac{n_k}{N}$$

1.1. Définition des paramètres usuels.

1.1.1. mode

La valeur ou les valeurs de la variable d'effectif maximum ou de fréquence maximum.

1.1.2. étendue

L'intervalle dont la borne inférieure est la plus petite valeur prise par X et la borne supérieure est la plus grande valeur prise par X c'est à dire $[o'_1 ; o'_N]$. C'est aussi la mesure de l'amplitude de cet intervalle : $A = o'_N - o'_1$

1.1.3. médiane Q2 et quartiles Q1, Q2, Q3

La médiane coïncide avec le deuxième quartile.

Q1, Q2, Q3 sont trois valeurs de la variable X qui vérifient les propriétés suivantes :

$$\text{Prop}\{X \leq Q1\} \geq 0,25 \quad \text{et} \quad \text{Prop}\{X \geq Q1\} \geq 0,75$$

$$\text{Prop}\{X \leq Q2\} \geq 0,5 \quad \text{et} \quad \text{Prop}\{X \geq Q2\} \geq 0,5$$

$$\text{Prop}\{X \leq Q3\} \geq 0,75 \quad \text{et} \quad \text{Prop}\{X \geq Q3\} \geq 0,25$$

L'intervalle interquartile est l'intervalle $[Q1 ; Q3]$

Ces propriétés conduisent à repérer Q1, Q2, Q3 de la façon suivante :

N =	Q1	Q2	Q3
N = 4q	entre la valeur de rang q et celle de rang q+1	entre la valeur de rang 2q et celle de rang 2q+1	entre la valeur de rang 3q et celle de rang 3q+1
N = 4q + 1	entre la valeur de rang q et celle de rang q+1	la valeur de rang 2q+1	entre la valeur de rang 3q+1 et celle de rang 3q+2
N = 4q + 2	la valeur de rang q+1	entre la valeur de rang 2q+1 et celle de rang 2q+2	la valeur de rang 3q+2
N = 4q + 3	la valeur de rang q+1	la valeur de rang 2q+2	la valeur de rang 3q+3

1.1.4. moyenne

La moyenne est la valeur de la variable X obtenue par l'une des trois procédures de calcul suivante :

$$\bar{x} = m = \frac{1}{N} \sum_{i=1}^N o_i = \frac{1}{N} \sum_{k=1}^{k=p} n_k x_k = \sum_{k=1}^{k=p} f_k x_k$$

1.1.5. variance, moment centré d'ordre 2, écart-type

La variance ou moment centré d'ordre 2 est la valeur obtenue par l'une des deux procédures de calcul suivante :

$$V(x) = \frac{1}{N} \sum_{k=1}^{k=p} n_k (x_k - m)^2 = \left(\frac{1}{N} \sum_{k=1}^{k=p} n_k x_k^2 \right) - m^2$$

L'écart-type est la valeur obtenue par le calcul suivant :

$$\sigma_x = \sqrt{V(x)} \quad \text{ou} \quad \sigma_x^2 = V(x)$$

1.1.6. moment centré d'ordre 3, coefficient d'asymétrie

Le moment centré d'ordre 3 est la valeur obtenue par la procédure de calcul suivante :

$$M_3(x) = \mu_3 = \frac{1}{N} \sum_{k=1}^{k=p} n_k (x_k - m)^3$$

Le coefficient d'asymétrie de Pearson β_1 est la valeur obtenue par le calcul suivant :	Le coefficient d'asymétrie de Fisher γ_1 est la valeur obtenue par le calcul suivant :
$\beta_1 = \left(\frac{\mu_3}{\sigma^3} \right)^2$	$\gamma_1 = \frac{\mu_3}{\sigma^3}$

1.1.7. moment centré d'ordre 4, coefficient d'aplatissement

Le moment centré d'ordre 4 est la valeur obtenue par la procédure de calcul suivante :

$$M_4(x) = \mu_4 = \frac{1}{N} \sum_{k=1}^{k=p} n_k (x_k - m)^4$$

Le coefficient d'aplatissement de Pearson β_2 est la valeur obtenue par le calcul suivant :	Le coefficient d'aplatissement de Fisher γ_2 est la valeur obtenue par le calcul suivant :
$\beta_2 = \frac{\mu_4}{\sigma^4}$	$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$

1.2. fonction cumulative croissante (fonction de répartition)

Cette fonction est ainsi définie de \mathbb{R} dans $[0;1]$: $t \xrightarrow{F} F(t) = \text{prop}(\{X < t\})$. F est alors une fonction définie sur \mathbb{R} , croissante, positive et continue à gauche en tout point de \mathbb{R} avec :

$$\lim_{t \rightarrow +\infty} F(t) = 1 \quad \text{et} \quad \lim_{t \rightarrow -\infty} F(t) = 0$$

Par convention, nous désignons les catégories particulières suivantes :

$E_1 = \{X=t\}$ désigne l'ensemble des individus pour lesquels la valeur de la variable est égale au nombre réel t .

$E_2 = \{X < t\}$ désigne l'ensemble des individus pour lesquels la valeur de la variable est inférieure strictement au nombre réel t .

$E_3 = \{X \leq t\}$ désigne l'ensemble des individus pour lesquels la valeur de la variable est inférieure ou égale au nombre réel t .

De manière analogue on peut définir $E_4 = \{X \geq t\}$ et $E_5 = \{X > t\}$

On rappelle que $\text{card}(E_i)$ = effectif de cet ensemble E_i et $\text{prop}(E_i) = \frac{\text{card}(E_i)}{N}$

Dans le cas où X est une variable quantitative discrète, l'analyse des diverses possibilités donne :

variation de t dans \mathbb{R}	$\{X < t\}$	$\text{card}(\{X < t\})$	$F(t) =$	$F(t) =$
$t \leq x_1$	\emptyset	0	0	0
$x_1 < t \leq x_2$	$\{X=x_1\}$	n_1	$\frac{n_1}{N}$	f_1
$x_2 < t \leq x_3$	$\{X=x_1\} \cup \{X=x_2\}$	n_1+n_2	$\frac{n_1+n_2}{N}$	f_1+f_2
...
$x_k < t \leq x_{k+1}$	$\{X=x_1\} \cup \{X=x_2\} \cup \{X=x_3\} \cup \dots \cup \{X=x_k\}$	$n_1+n_2+n_3+\dots+n_k$	$\frac{n_1+n_2+n_3+\dots+n_k}{N}$	$f_1+f_2+f_3+\dots+f_k$
...
$x_p < t$	tous les individus	$N = \sum_{k=1}^p n_k$	$\frac{\sum_{k=1}^p n_k}{N} = 1$	$\sum_{k=1}^p f_k = 1$

La représentation graphique de la fonction F est la **courbe cumulative croissante**. Dans ce cas il s'agit d'une fonction en escalier, fonction constante par intervalle.

2. VARIABLE STATISTIQUE QUANTITATIVE CONTINUE

X une variable statistique quantitative continue définie sur la population statistique P d'effectif total N.

$(i, [a_i ; b_i[$ avec $i = 1, \dots, N$ la série statistique observée, rangée dans l'ordre de recueil des observations rapportées à des intervalles.

$([x_k, x_{k+1}[; n_k)$ avec $k = 1, \dots, p$ la distribution des effectifs de la variable X.

$([x_k, x_{k+1}[, f_k)$ avec $k = 1, \dots, p$ la distribution des fréquences de la variable X.

Les résultats sont ramenés à des intervalles sur chacun desquels on fait l'hypothèse forte que la distribution des effectifs (ou des fréquences) est une distribution uniforme.

Les valeurs x_k qui figurent dans le tableau statistique sont rangées dans l'ordre croissant.

Tableau 2-1 statistique d'une variable continue

valeurs de la variable	$[x_1, x_2[$	$[x_2, x_3[$...	$[x_k, x_{k+1}[$...	$[x_p, x_{p+1}[$	
centres des intervalles	$c_1 = \frac{x_1+x_2}{2}$	$c_2 = \frac{x_2+x_3}{2}$...	$c_k = \frac{x_k+x_{k+1}}{2}$...	$c_p = \frac{x_p+x_{p+1}}{2}$	totaux
effectifs	n_1	n_2	...	n_k	...	n_p	N
fréquences	f_1	f_2	...	f_k	...	f_p	1

avec $N = \sum_{k=1}^{k=p} n_k$ et $f_k = \frac{n_k}{N}$

2.1. Définition des paramètres usuels.

2.1.1. mode

La valeur (ou les valeurs) de la variable qui est l'abscisse du point d'ordonnée maximum de la courbe "polygone des fréquences". C'est la valeur qui correspond à la densité de fréquence maximum.

2.1.2. étendue

L'intervalle dont la borne inférieure est la plus petite valeur prise par X et la borne supérieure est la plus grande valeur prise par X c'est à dire $[x_1 ; x_{p+1}]$

C'est aussi la mesure de l'amplitude de cet intervalle : $A = x_{p+1} - x_1$

2.1.3. médiane Q2 et quartiles Q1, Q2, Q3

La médiane coïncide avec le deuxième quartile.

Q1, Q2, Q3 sont trois valeurs de la variable X qui vérifient les propriétés suivantes :

$$\text{Prop}(\{X \leq Q1\}) = 0,25 \text{ et } \text{Prop}(\{X \geq Q1\}) = 0,75$$

$$\text{Prop}(\{X \leq Q2\}) = 0,5 \text{ et } \text{Prop}(\{X \geq Q2\}) = 0,5$$

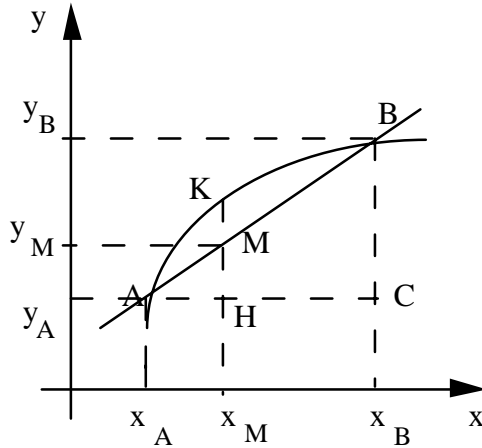
$$\text{Prop}(\{X \leq Q3\}) = 0,75 \text{ et } \text{Prop}(\{X \geq Q3\}) = 0,25$$

L'intervalle interquartile est l'intervalle [Q1 ; Q3]

Pour calculer les valeurs Q1, Q2, Q3 la procédure suivante est mise en œuvre :

- repérage des trois intervalles $[x_k; x_{k+1}[$ contenant respectivement Q1, Q2, Q3
- estimation par interpolation linéaire des valeurs Q1, Q2, Q3.

2.1.4. Interpolation linéaire



Ici, on suppose x_M connu, et on calcule y_M

valeur approchée de y_K par la relation $\frac{x_M - x_A}{x_B - x_A} =$

$$\frac{y_M - y_A}{y_B - y_A}$$

ou encore par la relation

$$\boxed{\frac{y_M - y_A}{x_M - x_A} = \frac{y_B - y_A}{x_B - x_A}}$$

qui permet de calculer

x_M ou y_M

2.1.5. moyenne

La moyenne est la valeur de la variable X obtenue par l'une des deux procédures de calcul suivante :

$$\bar{x} = m = \mu = \frac{1}{N} \sum_{k=1}^{k=p} n_k c_k = \frac{\sum_{k=1}^{k=p} f_k c_k}{\sum_{k=1}^{k=p} f_k}$$

2.1.6. variance, moment centré d'ordre 2 & écart-type

La variance ou moment centré d'ordre 2 est la valeur obtenue par l'une des deux procédures de calcul suivante :

$$V(x) = \frac{1}{N} \sum_{k=1}^{k=p} n_k (c_k - m)^2 = \left(\frac{1}{N} \sum_{k=1}^{k=p} n_k c_k^2 \right) - m^2$$

L'écart-type est la valeur obtenue par le calcul suivant : $\sigma_x = \sqrt{V(x)}$ ou $\sigma_x^2 = V(x)$

2.1.7. moment centré d'ordre 3 , coefficient d'asymétrie

Le moment centré d'ordre 3 est la valeur obtenue par la procédure de calcul suivante :

$$M_3(x) = \mu_3 = \frac{1}{N} \sum_{k=1}^{k=p} n_k (c_k - m)^3$$

Le coefficient d'asymétrie de Pearson β_1 est la valeur obtenue par le calcul suivant :	Le coefficient d'asymétrie de Fisher γ_1 est la valeur obtenue par le calcul suivant :
$\beta_1 = \left(\frac{\mu_3}{\sigma^3}\right)^2$	$\gamma_1 = \frac{\mu_3}{\sigma^3}$

2.1.8. moment centré d'ordre 4 , coefficient d'aplatissement

Le moment centré d'ordre 4 est la valeur obtenue par la procédure de calcul suivante.

$$M_4(x) = \mu_4 = \frac{1}{N} \sum_{k=1}^{k=p} n_k (c_k - m)^4$$

Le coefficient d'aplatissement de Pearson β_2 est la valeur obtenue par le calcul suivant :	Le coefficient d'aplatissement de Fisher γ_2 est la valeur obtenue par le calcul suivant :
$\beta_2 = \frac{\mu_4}{\sigma^4}$	$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$

2.2. fonction cumulative croissante (fonction de répartition)

Cette fonction est ainsi définie de \mathbb{R} dans $[0;1]$: $t \xrightarrow{F} F(t) = \text{prop}(\{X < t\})$. F est alors une fonction définie sur \mathbb{R} , croissante, positive et continue à gauche en tout point de \mathbb{R} avec : $t \xrightarrow{F} \lim F(t) = 1 \rightarrow +\infty$ et $t \xrightarrow{F} \lim F(t) = 0 \rightarrow -\infty$

Dans le cas où X est une variable quantitative continue, l'analyse des diverses possibilités donne :

variation de t dans \mathbb{R}	$\{X < t\}$	$F(t) =$
$t \leq x_1$	\emptyset	0
$x_1 < t \leq x_2$	$\{x_1 < X < t\}$	$\left(\frac{f_1}{x_2 - x_1}\right) (t - x_1)$
$x_2 < t \leq x_3$	$\{x_1 < X \leq x_2\} \cup \{x_2 < X < t\}$	$f_1 + \left(\frac{f_2 - f_1}{x_3 - x_2}\right) (t - x_2)$
\dots	\dots	\dots
$x_k < t \leq x_{k+1}$	$\{x_1 < X \leq x_2\} \cup \{x_2 < X \leq x_3\} \cup \dots \cup \{x_{k-1} < X \leq x_k\} \cup \{x_k < X < t\}$	$f_1 + f_2 + \dots + \left(\frac{f_{k+1} - f_k}{x_{k+1} - x_k}\right) (t - x_k)$
\dots	\dots	\dots
$x_{p+1} < t$	\mathbb{P}	1

La représentation graphique de la fonction F est la **courbe cumulative croissante**.

Dans ce cas il s'agit d'une fonction affine par intervalle. Sa construction repose sur l'hypothèse de la distribution uniforme des fréquences sur chaque intervalle et l'interpolation linéaire.

2.3. Comparaison d'une variable X continue avec une variable de Laplace-Gauss $LG(\mu, \sigma)$ de paramètres $\bar{x} = \mu$ et $\sigma_X = \sigma$, $\gamma_1 = 0$, $\beta_1 = 0$, $\gamma_2 = 3$, $\beta_2 = 0$

L'histogramme est la représentation graphique de la fonction densité de fréquences

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

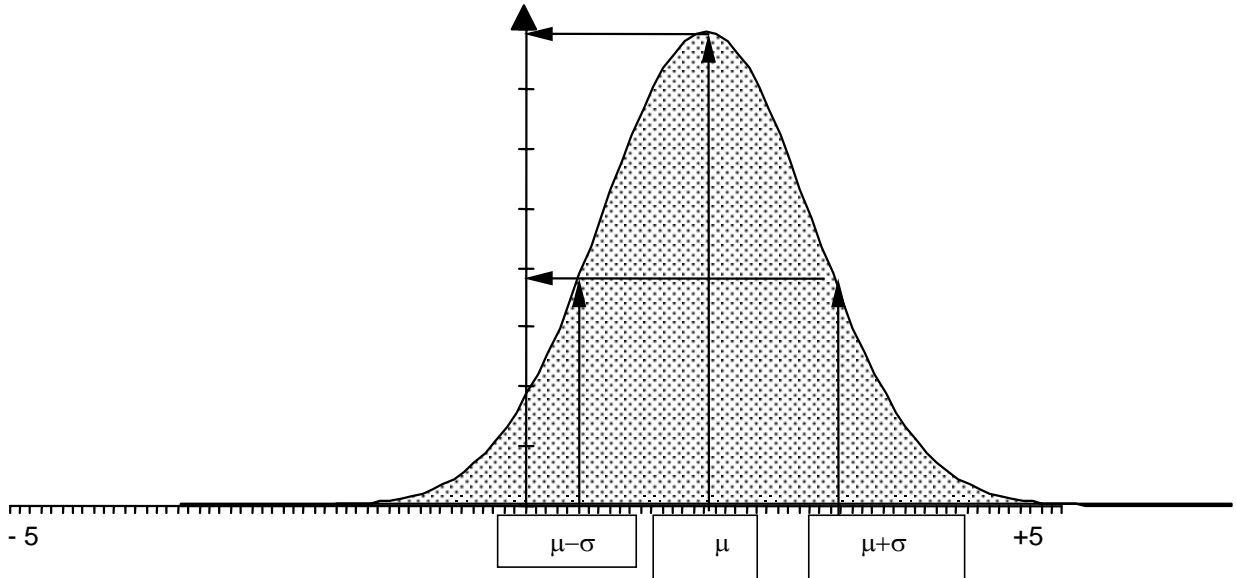


Figure 2.3-1 Courbe de Laplace-Gauss

Dans l'intervalle interquartile **A** = $[\mu - 0,675 \sigma ; \mu + 0,675 \sigma]$

il y a *environ* **50 %** des observations

Dans l'intervalle **B** = $[\mu - \sigma ; \mu + \sigma]$

il y a *environ* **68 %** des observations

Dans l'intervalle **C** = $[\mu - 1,96 \sigma ; \mu + 1,96 \sigma]$

il y a *environ* **95 %** des observations

Dans l'intervalle **D** = $[\mu - 2,58 \sigma ; \mu + 2,58 \sigma]$

il y a *environ* **99 %** des observations

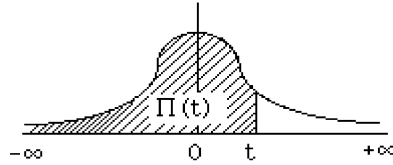
Dans l'intervalle **E** = $[\mu - 3 \sigma ; \mu + 3 \sigma]$

il y a *environ* **99,7 %** des observations

2.4. Extrait de la table donnant la distribution des fréquences de la variable centrée réduite de Laplace-Gauss

$$Z = \frac{X - \mu}{\sigma} = \text{LG}(0,1)$$

$$\Phi(t) = \text{Prop} \{ Z < t \}$$



t	0	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

Table pour les grandes valeurs de t

t	3,0	3,1	3,2	3,3	3,4	3,4	3,5	3,6	3,8	4,0
$\Pi(t)$	0,99 865	0,999 032	0,9993 13	0,9995 17	0,9996 63	0,9996 63	0,999 767	0,999 840	0,9999 27	0,999 968

Nota Bene : La table donne les valeurs de $\Pi(t)$ pour t positif.
Lorsque t est négatif, il suffit de prendre le complément à 1.

Exemple :

pour t = 1,37 on lit dans la table $\Phi(t=1,37) = 0,9147$

pour t = -1,37 on obtient $\Phi(t=-1,37) = 1-0,9147 = 0,0853$

3. Quelques remarques importantes

3.1. Changement de variable pour une variable quelconque

On désigne Y la nouvelle variable reliée à la variable X par : $Y = aX + b$

Les relations entre :

- la moyenne \bar{Y} de la variable Y et la moyenne \bar{X} de la variable X : $\bar{Y} = a\bar{X} + b$

- la variance de la variable Y et la variance de la variable X :

$$\sigma^2 Y = V(Y) = a^2 V(X) = a^2 \sigma^2 X$$

$$\sigma Y = \sqrt{V(Y)} = |a| \sqrt{V(X)} = |a| \sigma X$$

Un cas particulier très important : $Z = \frac{X - \bar{X}}{\sigma X}$, variable centrée réduite, pour

laquelle:

$$\bar{Z} = 0 \text{ et } \sigma_Z = 1.$$

Ce changement de variable peut être interprété comme un changement d'origine et un changement d'unité de mesure. Les valeurs de Z sont les valeurs transformées de X en choisissant la moyenne comme origine et l'écart-type comme unité de mesure.

3.2. Des propriétés fondamentales

La **moyenne** est la valeur de la variable qui rend minimum l'**écart moyen** à une valeur

c quelconque : $\overline{E_c} = \frac{1}{N} \sum_{k=1}^p n_k (x_k - c) = 0$ si et seulement si $c = m$.

La **moyenne** est la valeur de la variable qui rend minimum le **moment centré en une**

valeur c d'ordre 2 : $\frac{1}{N} \sum_{k=1}^p n_k (x_k - c)^2$ est minimum si et seulement si $c = m$. Ainsi parmi

toutes les positions possibles c par rapport auxquelles on peut mesurer une fluctuation à l'aide du moment centré d'ordre 2 par rapport à c, la variance est la valeur minimale de cette fluctuation.

La **médiane** est la (les) valeur(s) de la variable qui rend(ent) minimum l'**écart absolu**

moyen à une valeur c : $\overline{E_c} = \frac{1}{N} \sum_{k=1}^p n_k |x_k - c|$ est minimum si et seulement si $c = Q_2$ ou

si c appartient à l'intervalle médian.

4. VARIABLE STATISTIQUE QUALITATIVE DISCRETE ORDONNEE

X une variable statistique qualitative discrète ordonnée définie sur la population statistique P d'effectif total N.

(i, o_i) avec $i = 1, \dots, N$ la série statistique observée, rangée dans l'ordre de recueil des observations.

(i, o'_i) avec $i = 1, \dots, N$ la série statistique observée, rangée selon l'ordre défini sur l'espace des modalités o_i .

(x_k, n_k) avec $k = 1, \dots, p$ la distribution des effectifs de la variable X.

(x_k, f_k) avec $k = 1, \dots, p$ la distribution des fréquences de la variable X.

Les modalités x_k qui figurent dans le tableau statistique sont rangées dans l'ordre "croissant".

modalités de la variable	x_1	x_2	...	x_k	...	x_{p-1}	x_p	totaux
effectifs	n_1	n_2	...	n_k	...	n_{p-1}	n_p	N
fréquences	f_1	f_2	...	f_k	...	f_{p-1}	f_p	1

avec $N = \sum_{k=1}^{k=p} n_k$ et $f_k = \frac{n_k}{N}$

4.1. Définition des paramètres usuels.

4.1.1. mode

La modalité ou les modalités de la variable d'effectif maximum ou de fréquence maximum.

4.1.2. étendue

L'intervalle dont la borne inférieure est la modalité de rang 1 prise par X et la borne supérieure est la modalité de rang p prise par X c'est à dire $[x_1 ; x_p]$

L'amplitude de cet intervalle n'est pas mesurable au sens habituel.

4.1.3. médiane Q2 et quartiles Q1, Q2, Q3

La médiane coïncide avec le deuxième quartile.

Q1, Q2, Q3 sont trois modalités de la variable X qui vérifient les propriétés suivantes :

$$\text{Prop}(\{X \text{ avant } Q1\}) \geq 0,25 \text{ et } \text{Prop}(\{X \text{ après } Q1\}) \geq 0,75$$

$$\text{Prop}(\{X \text{ avant } Q2\}) \geq 0,5 \text{ et } \text{Prop}(\{X \text{ après } Q2\}) \geq 0,5$$

$$\text{Prop}(\{X \text{ avant } Q3\}) \geq 0,75 \text{ et } \text{Prop}(\{X \text{ après } Q3\}) \geq 0,25$$

L'intervalle interquartile est l'intervalle $[Q1 ; Q3]$

Ces propriétés conduisent à repérer Q1, Q2, Q3 de la façon suivante :

N =	Q1	Q2	Q3
-----	----	----	----

$N = 4q$	entre la modalité de rang q et celle de rang $q+1$	entre la modalité de rang $2q$ et celle de rang $2q+1$	entre la modalité de rang $3q$ et celle de rang $3q+1$
$N = 4q + 1$	entre la modalité de rang q et celle de rang $q+1$	la modalité de rang $2q+1$	entre la modalité de rang $3q+1$ et celle de rang $3q+2$
$N = 4q + 2$	la modalité de rang $q+1$	entre la modalité de rang $2q+1$ et celle de rang $2q+2$	la modalité de rang $3q+2$
$N = 4q + 3$	la modalité de rang $q+1$	la modalité de rang $2q+2$	la modalité de rang $3q+3$

4.1.4. entropie

L'entropie renvoie à l'idée de quantité d'information nécessaire à la réduction de l'incertitude de l'observateur, à l'idée d'une "mesure du désordre", de la dispersion qui oppose les deux cas extrêmes suivants :

modalités de la variable	x_1	x_2	...	x_k	...	x_{p-1}	x_p	totaux
effectifs	$\frac{N}{p}$	$\frac{N}{p}$...	$\frac{N}{p}$...	$\frac{N}{p}$	$\frac{N}{p}$	N
fréquences	$\frac{1}{p}$	$\frac{1}{p}$...	$\frac{1}{p}$...	$\frac{1}{p}$	$\frac{1}{p}$	1

modalités de la variable	x_1	x_2	...	x_k	...	x_{p-1}	x_p	totaux
effectifs	0	0	...	N	...	0	0	N
fréquences	0	0	...	1	...	0	0	1

L'indice utilisé est : $H = - \sum_{k=1}^{k=p} f_k \log_2(f_k)$ et $0 \leq H \leq \log_2(p)$

La fonction \log_2 est la fonction logarithme à base 2, c'est à dire $y = \log_2(x) = \frac{\ln(x)}{\ln(2)}$

équivalent à $x = 2^y$. Le calcul se fait avec une machine à partir du \ln , logarithme népérien. On peut constater que dans le premier cas $H = \log_2(p)$ est maximale et dans le second cas $H = 0$ est minimale.

4.1.5. entropie relative

L'entropie relative renvoie à l'idée d'un taux de désordre observé relativement au désordre maximum.

$$H_{rel} = \frac{H}{H_{max}} = \frac{- \sum_{k=1}^{k=p} f_k \log_2(f_k)}{\log_2(p)}$$

5. VARIABLE STATISTIQUE QUALITATIVE DISCRETE : VARIABLE NOMINALE

X une variable statistique qualitative discrète définie sur la population statistique P d'effectif total N.

(i, o_i) avec $i = 1, \dots, N$ la série statistique observée, rangée dans l'ordre de recueil des observations.

(x_k, n_k) avec $k = 1, \dots, p$ la distribution des effectifs de la variable X.

(x_k, f_k) avec $k = 1, \dots, p$ la distribution des fréquences de la variable X.

modalités de la variable	x_1	x_2	...	x_k	...	x_{p-1}	x_p	totaux
effectifs	n_1	n_2	...	n_k	...	n_{p-1}	n_p	N
fréquences	f_1	f_2	...	f_k	...	f_{p-1}	f_p	1

avec $N = \sum_{k=1}^{k=p} n_k$ et $f_k = \frac{n_k}{N}$

5.1. Définition des paramètres usuels.

5.1.1. mode

La modalité ou les modalités de la variable d'effectif maximum ou de fréquence maximum.

5.1.2. entropie

L'entropie renvoie à l'idée d'une "mesure du désordre":

L'indice utilisé est : $H = - \sum_{k=1}^{k=p} f_k \log_2(f_k)$ et $0 \leq H \leq \log_2(p)$

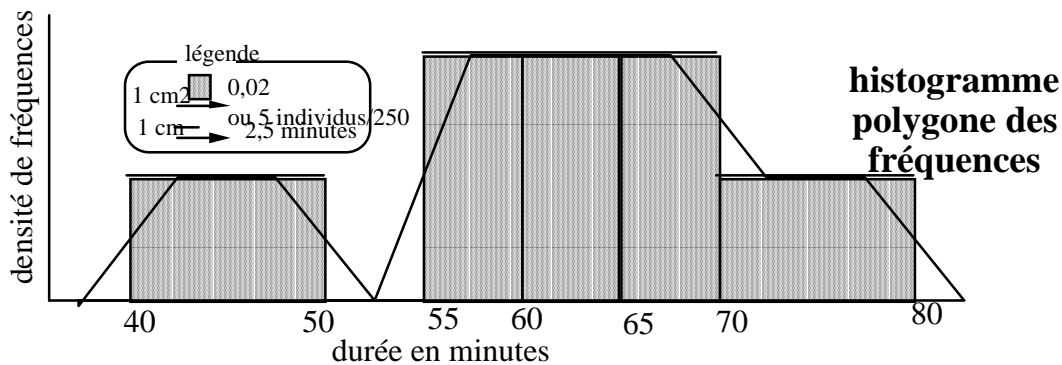
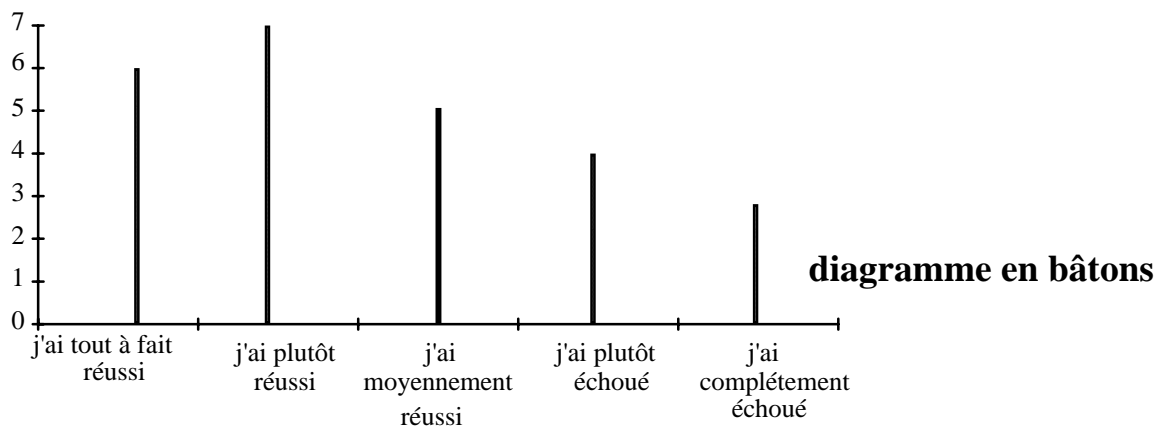
5.1.3. entropie relative

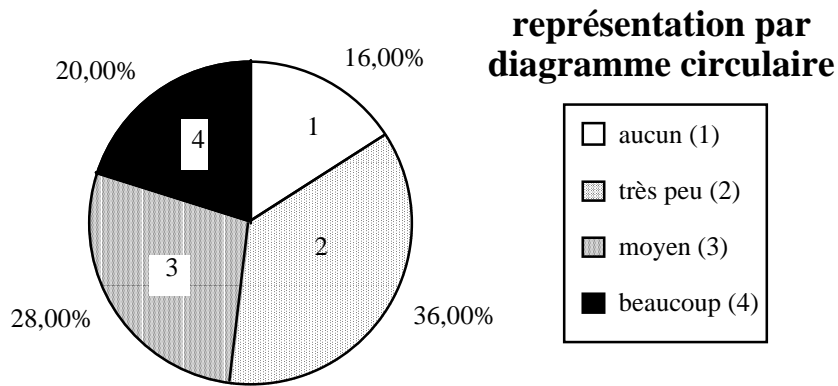
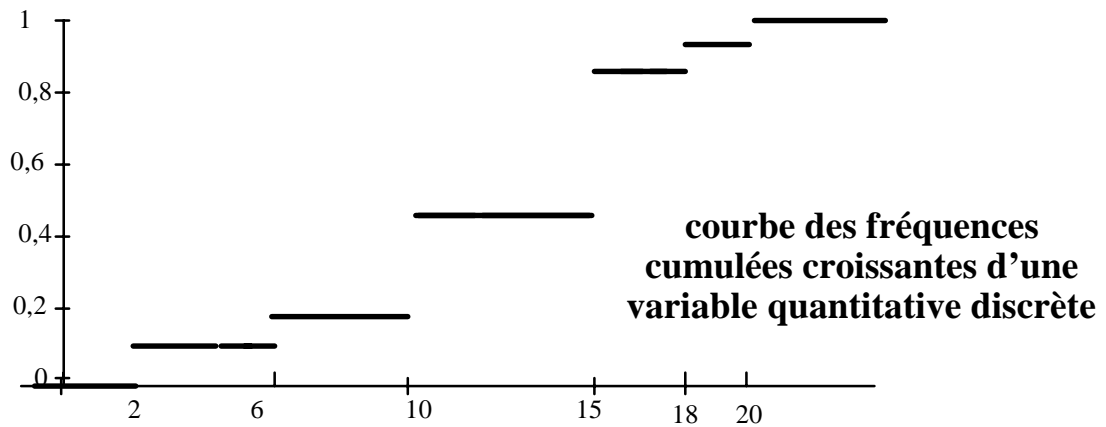
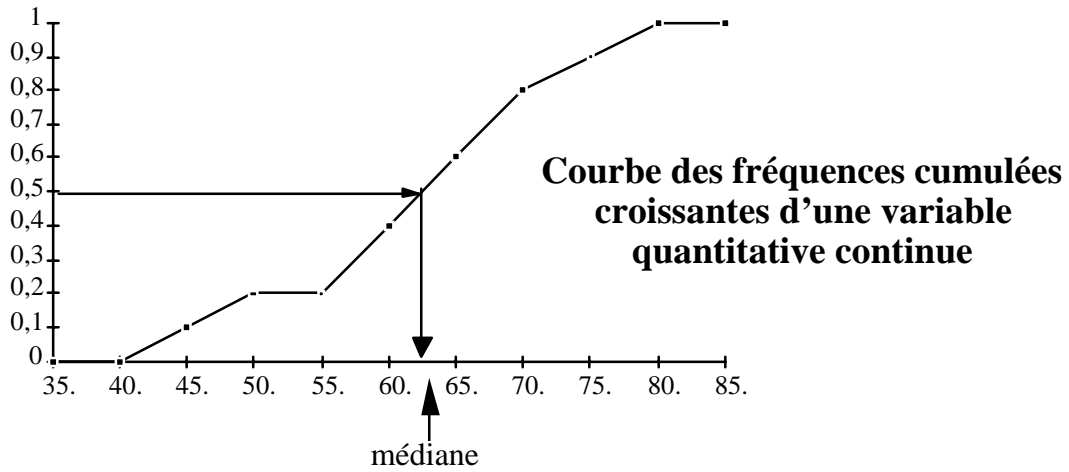
L'entropie relative renvoie à l'idée d'un taux de désordre observé relativement au désordre maximum.

$$H_{rel} = \frac{H}{H_{max}} = \frac{- \sum_{k=1}^{k=p} f_k \log_2(f_k)}{\log_2(p)}$$

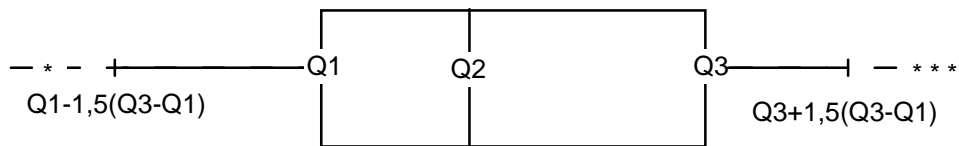
6. TRAITEMENTS GRAPHIQUES

variables	types habituels de graphiques
variable qualitative nominale discrète	- diagramme en bâtons des effectifs ou des fréquences - diagramme circulaire ou rectangulaire
variable qualitative ordonnée discrète	- diagramme en bâtons des effectifs ou des fréquences - diagramme circulaire ou rectangulaire
variable quantitative discrète	- diagramme en bâtons des effectifs ou des fréquences - diagramme circulaire ou rectangulaire - courbe cumulative croissante courbe des effectifs cumulés croissants ou des fréquences ou des fréquences cumulées croissantes. - diagramme en boîtes (Tukey)
variable quantitative continue	- diagramme en bâtons des effectifs ou des fréquences - diagramme circulaire ou rectangulaire - histogramme des effectifs ou des fréquences - polygone des effectifs ou des fréquences (lissage) - courbe cumulative croissante courbe des effectifs cumulés croissants ou des fréquences ou des fréquences cumulées croissantes. - diagramme en boîtes (Tukey)





Une représentation graphique de type **diagramme en boîtes** (J.W. Tukey):



Lorsque la variable étudiée est indexée par le temps et possède un caractère périodique (par exemple certaines variables chronologiques) la représentation en diagramme polaire est particulièrement bien adaptée en faisant apparaître la propriété de périodicité. En effet la variable "temps" est représentée par une mesure angulaire tandis que la variable est représentée par une mesure radiale.

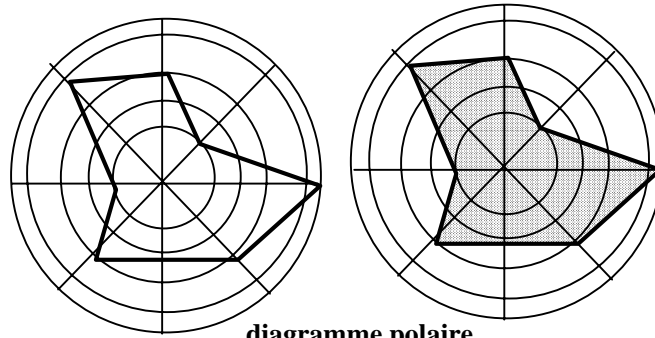


diagramme polaire

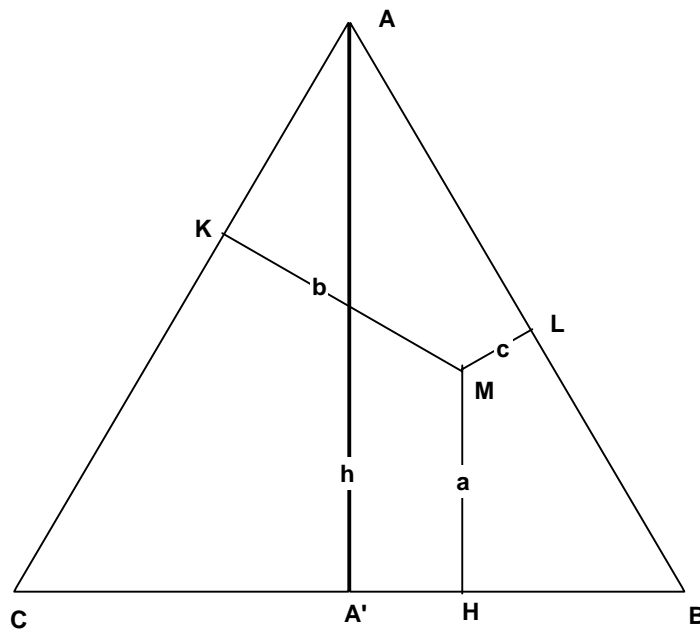
Cependant cette représentation peut être utilisée dans d'autres cas. En fait elle peut constituer un mode de représentation d'un vecteur statistique dans lequel chaque composante serait représenté par un axe de la "cible" et chaque valeur ou modalité par un niveau déterminé par un des cercles concentriques.

Il convient de préciser que la ligne polygonale joignant les points sur chaque axe n'est pas toujours interprétable quantitativement comme elle peut l'être pour certaines variables chronologiques. Elle est à comparer selon le cas, à la ligne polygonale tracée soit sur les diagramme en bâtons soit sur les histogrammes

7. Représentation barycentrique dans le triangle équilatéral

7.1. Propriétés du triangle équilatéral

Le triangle équilatéral est une figure géométrique familière que l'on caractérise en général comme étant un triangle dont les trois côtés sont isométriques ou les trois angles mesurent chacun 60° ou $\frac{\pi}{3}$ radian. La représentation graphique ci-dessous réalise une telle figure.



Parmi les multiples propriétés vérifiées cet objet géométrique, il en est une qui est particulièrement utile à la statistique exploratoire.

Il s'agit du théorème de Viviani (1622-1703) :

La somme des distances d'un point M quelconque situé sur le domaine délimité par le triangle, à chacun des côtés, est égale à la hauteur.

En écriture formelle, cela revient à $MH + MK + ML = AA' = BB' = CC' = h$

Cette propriété est démontrée en remarquant que l'aire du triangle ABC est décomposable en la somme des aires respectives des triangles AMB, BMC, CMA.

Ainsi $\text{Aire}(ABC) = \text{Aire}(AMB) + \text{Aire}(BMC) + \text{Aire}(CMA)$

Si nous notons la longueur du côté $m = AB = AC = BC$, alors $\frac{m h}{2} = \frac{m c}{2} + \frac{m a}{2} + \frac{m b}{2}$

Par un simple calcul algébrique on déduit alors que $h = a + b + c$.

Nous rappelons aussi le lien entre la hauteur et le côté : $h = \frac{a\sqrt{3}}{2}$

La seconde notion mathématique utile est celle de barycentre d'un système de trois points pondérés (A,a) , (B,b) , (C,c) , les nombres a , b , c sont des nombres réels. La notion de moyenne (arithmétique pondérée) en statistique en est un cas particulier.

7.2. Définition du barycentre :

Étant donné trois points pondérés (A,a) , (B,b) , (C,c) tels que $a+b+c \neq 0$

Alors il existe un unique point G , barycentre de trois points pondérés, caractérisé par la relation vectorielle : $a\overrightarrow{GA} + b\overrightarrow{GB} + c\overrightarrow{GC} = \vec{0}$

Or le point M intérieur vérifie cette propriété si on affecte les sommets A , B , C des poids respectifs a , b , c correspondants aux distances du point M aux côtés.

Le point de concours des trois médianes du triangle correspond à la situation où les trois points A , B , C sont affectés du même poids, c'est à dire $a = b = c = \frac{h}{3}$.

7.3. Usage des propriétés pour réaliser une représentation des données en statistique :

Supposons que la variable statistique X étudiée soit une variable à trois modalités M_1 , M_2 , M_3 . Supposons que l'étude conduite à une répétition de cette variable, caractérisée par la suite X_1, X_2, \dots, X_n . De là à chaque unité statistique U_i nous pouvons associer un triplet (f_{i1}, f_{i2}, f_{i3}) de nombres tel que f_{ik} = fréquence d'apparition de la modalité M_k dans la série des n observations pour l'individu U_i . Sous cette hypothèse $f_{i1} + f_{i2} + f_{i3} = 1$.

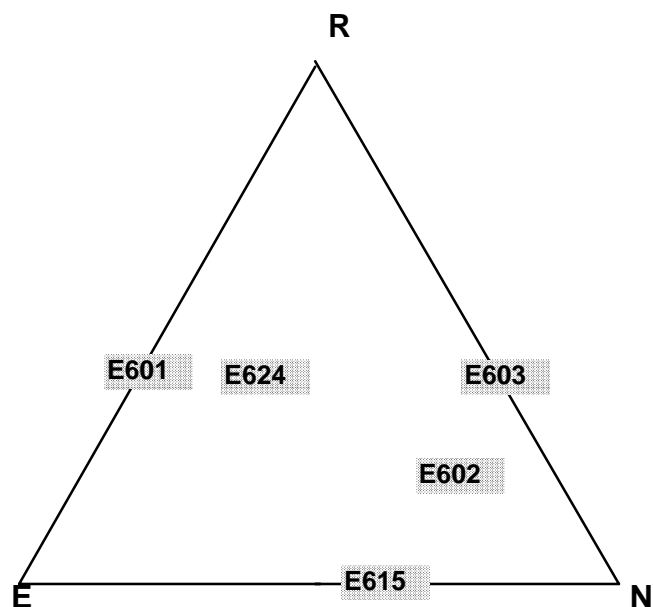
Ainsi l'unité statistique U_i peut être représentée graphiquement dans le triangle équilatéral $M_1M_2M_3$ en considérant que ce point est le barycentre des points pondérés $(M_1, f_{i1}), (M_2, f_{i2}), (M_3, f_{i3})$.

Une situation typique est celle d'un questionnaire de n questions pour chacune desquelles on attribue l'une des trois modalités R = réponse correcte, E = réponse erronée, N = non-réponse. A chaque questionnaire on peut alors associer le triplet (x,y,z) avec

$$x = \frac{\text{nombre de réponses correctes}}{n},$$

$$y = \frac{\text{nombre de réponses erronées}}{n},$$

$$z = \frac{\text{nombre de non-réponses}}{n}$$



8. Représentation barycentrique dans le carré

Dans l'étude de certaines variables qualitatives à quatre modalités, la représentation barycentrique dans le carré est tout à fait intéressante. On place le point G dans un carré ABCD de telle sorte que G soit le barycentre des points pondérés (A, a), (B, b), (C, c), (D, d) c'est à dire que : $a\overrightarrow{GA} + b\overrightarrow{GB} + c\overrightarrow{GC} + d\overrightarrow{GD} = \vec{0}$ avec $a+b+c+d \neq 0$

La construction du point G s'appuie sur le théorème suivant :

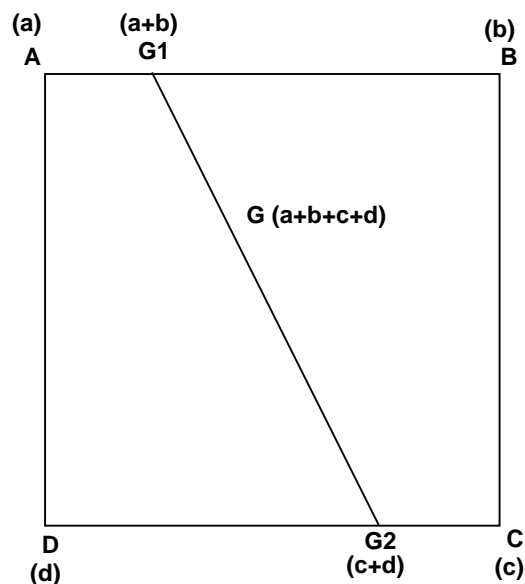
8.1. Théorème d'associativité :

On ne modifie pas le barycentre de quatre points, en remplaçant deux de ces points par leur barycentre affecté de la somme de leurs coefficients.

8.2. Procédure de construction

Ainsi la démarche suit la procédure :

- construire sur la droite AB le point G₁ en tant que barycentre des points pondérés (A, a), (B, b),
- construire sur la droite CD le point G₂ en tant que barycentre des points pondérés (C, c), (D, d)
- construire sur la droite G₁G₂ le point G en tant que barycentre des points pondérés (G₁, a+b), (G₂, c+d)



9. LES MOYENNES d'une VARIABLE STATISTIQUE QUANTITATIVE DISCRETE

X une variable statistique quantitative discrète définie sur la population statistique P d'effectif total N.

(i, o_i) avec $i = 1, \dots, N$ la série statistique observée, rangée dans l'ordre de recueil des observations.

(i, o'_i) avec $i = 1, \dots, N$ la série statistique observée, rangée selon l'ordre croissant des valeurs o_i .

(x_k, n_k) avec $k = 1, \dots, p$ la distribution des effectifs de la variable X.

(x_k, f_k) avec $k = 1, \dots, p$ la distribution des fréquences de la variable X.

Les valeurs x_k qui figurent dans le tableau statistique sont rangées dans l'ordre croissant.

valeurs de la variable	x_1	x_2	...	x_k	...	x_{p-1}	x_p	totaux
effectifs	n_1	n_2	...	n_k	...	n_{p-1}	n_p	N
fréquences	f_1	f_2	...	f_k	...	f_{p-1}	f_p	1

$$\text{avec } N = \sum_{k=1}^{k=p} n_k \text{ et } f_k = \frac{n_k}{N}$$

9.1. moyenne (arithmétique)

$$M = \frac{1}{N} \sum_{k=1}^{k=p} n_k x_k$$

9.2. moyenne géométrique

$$G = \sqrt[N]{\prod_{k=1}^{k=p} x_k^{n_k}}$$

9.3. moyenne harmonique

$$H = \frac{N}{\sum_{k=1}^{k=p} \frac{n_k}{x_k}}$$

9.4. moyenne d'ordre 2 : moyenne quadratique (moment d'ordre 2)

$$Q = \sqrt{\frac{1}{N} \sum_{k=1}^{k=p} n_k x_k^2}$$

9.5. moyenne d'ordre m

Soit m un nombre réel strictement positif

$$Q_m = \left(\frac{1}{N} \sum_{k=1}^{k=p} n_k x_k^m \right)^{\frac{1}{m}}$$

9.6. moyenne tronquée d'ordre 1

$$MT_1 = \frac{1}{N-2} \sum_{i=2}^{i=N-1} o'_i$$

9.7. moyenne tronquée d'ordre q

Soit q un nombre entier

$$MT_q = \frac{1}{N-2q} \sum_{i=q+1}^{i=N-q} o'_i$$

9.8. moyenne de Winsor d'ordre q

Soit q un nombre entier

$$MW_q = \frac{1}{N} \left(q o'_{q+1} + \sum_{i=q+1}^{i=N-q} o'_i + q o'_{N-q} \right)$$

9.9. Remarque importante sur l'existence d'une moyenne :

Il est important de prendre en compte les conditions d'existence de ces valeurs et les conditions algébriques qui rendent calculables les différentes *moyennes*. Par exemple la moyenne harmonique H suppose qu'aucun des termes x_k ne soit nul. La moyenne d'ordre m nécessite pour certaines valeurs de m que x_k soient positifs, par exemple pour

$$m = 0,5 \text{ car } x_k^{0,5} = x_k^{\frac{1}{2}} = \sqrt{x_k}$$

CALCULER :

En statistique, calculer est une activité dominante. Que cette activité de calcul concerne l'activité de comptage, de dénombrement ou bien qu'elle concerne les traitements conduits directement sur les valeurs des variables quantitatives, il convient d'une part de posséder des instruments techniques et conceptuels, d'autre part de s'astreindre à maintenir une attention critique à son égard.

10. Analyse combinatoire : calcul de dénombrement.

L'analyse combinatoire a pour objet l'effectuation des dénombrements c'est à dire la détermination à partir d'un ensemble fini A donné du nombre d'éléments, du cardinal d'un autre ensemble défini par des propriétés sur les éléments de A.

- *situation n° 1* : nombre de suites de k termes constituées en extrayant un élément de chacun des k ensembles A_1, A_2, \dots, A_k , de cardinal respectif n_1, n_2, \dots, n_k ,

$$\prod_{i=1}^k n_i = n_1 \times n_2 \times \dots \times n_k$$

- *situation n° 2* : nombre d'échantillons ordonnés de *taille* k obtenus par tirage avec remise à partir d'une population P d'effectif total N :

nombre de suites de k termes constituées en extrayant k éléments d'un ensemble P de cardinal égal à N, c'est à dire constituées en considérant que l'élément extrait est remis dans l'ensemble P après tirage,

$$N^k$$

- *situation n° 3* : nombre d'échantillons ordonnés de *taille* k obtenus par tirage sans remise à partir d'une population P d'effectif total N :

nombre d'*arrangements sans répétition* de N éléments pris k à k éléments dans l'ensemble P, c'est à dire de suites de k termes distincts.

$$A_N^k = \frac{N!}{(N-k)!}$$

- *situation n° 4* : nombre d'échantillons ordonnés de *taille* N obtenus par tirage sans remise à partir d'une population P d'effectif total N :

nombre de *permutations* de N éléments dans l'ensemble P, c'est à dire de suites de N termes distincts.

$$P_N = N!$$

- *situation n° 5* : nombre d'échantillons de *taille* n obtenus par tirage sans remise à partir d'une population P d'effectif total N :

nombre de *combinaisons* de n éléments dans l'ensemble P , c'est à dire de sous-ensembles de n termes distincts de l'ensemble P :

$$C_N^n = \frac{N!}{n! (N-n)!}$$

Quelques propriétés :

$$C_N^n = C_N^{N-n}$$

Formule à partir de laquelle peut être construit le triangle de Pascal :

$$C_{N+1}^n = C_N^n + C_N^{n-1}$$

Formule du binôme de Newton :

$$(a + b)^n = \sum_{i=0}^n (C_N^i a^i b^{n-i})$$

- *situation n° 6* : nombre de partitions de k sous-ensembles A_1, A_2, \dots, A_k , de cardinal respectif n_1, n_2, \dots, n_k , obtenues à partir d'une population P d'effectif total N :

la suite de sous-ensembles A_1, A_2, \dots, A_k , de cardinal n_1, n_2, \dots, n_k , est une partition de P si et seulement si les sous-ensembles sont deux à deux disjoints et si la réunion des k sous-ensembles donne l'ensemble P .

$$\frac{N!}{n_1! n_2! \dots n_k!}$$

- *situation n° 7* : nombre d'*arrangements avec répétitions* associés à r_1, r_2, \dots, r_N , éléments obtenus à partir d'une population

$P = \{ w_i, i = 1, \dots, N \}$ d'effectif total N c'est à dire les suites de $r_1 + r_2 + \dots + r_N$ termes comportant r_i fois l'individu w_i pour $i = 1$ à N :

$$\frac{(r_1 + r_2 + \dots + r_N)!}{r_1! r_2! \dots r_N!}$$

- *situation n° 8* : nombre de *combinaisons avec répétitions* associées à r_1, r_2, \dots, r_N , obtenus à partir d'une population $P = \{ w_i, i = 1, \dots, N \}$ d'effectif total N c'est à dire les suites de $r_1 + r_2 + \dots + r_N$ termes comportant r_i fois au rang i l'individu w_i pour $i = 1$ à N :

$$C_{(r_1 + r_2 + \dots + r_N) + N - 1}^{N-1} = \frac{(r_1 + r_2 + \dots + r_N + N - 1)!}{(N-1)! (r_1 + r_2 + \dots + r_N)!}$$

Effets des approximations dans les calculs en statistique : Danger! Approximations...

Les deux exemples suivants ont pour but de faire comprendre l'intérêt des précautions à prendre lorsque l'on remplace une valeur par une "valeur approchée tout à fait raisonnable".

Lorsque l'on utilise une machine à calculer, comme ceci est pratiquement une nécessité pour réaliser des calculs statistiques, il convient de travailler avec la précision maximale de l'instrument dans le déroulement des opérations. Le recours à une valeur approchée n'intervient alors que pour communiquer les informations et les conclusions de l'étude.

Dans le premier exemple, nous considérons une étude portant sur le prix unitaire d'un objet en fonction des 80 points de vente existants. La variable est considérée comme une variable quantitative discrète. On cherche à obtenir le prix moyen et la dispersion des valeurs à l'aide de l'écart-type.

valeurs de la variable	6,55	6,8	6,85	6,9	7	7,1	7,35	7,4	effectif total
effectifs	7	6	8	13	16	18	7	5	80

Les calculs donnent les résultats suivants:

somme des 80 valeurs	559,4		
prix unitaire moyen	6,9925	prix unitaire moyen arrondi "raisonnablement"	7
somme des carrés des écarts à la valeur moyenne exacte	3,8005	somme des carrés des écarts à la valeur moyenne approchée	3,805
variance calculée selon la définition	0,04750625	variance calculée selon la définition	0,048
somme des carrés des valeurs de la variable	3915,405		
variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	0,04750625	variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	-0,06

Nous repérons une aberration dans le résultat issu du calcul de la variance par la seconde méthode. En utilisant dans le cours des calculs, une valeur approchée tout à fait raisonnable **7** pour **6,9925**, on obtient une valeur négative **-0,06** ce qui est impossible puisque la variance est par définition.

Dans le second exemple, nous considérons une étude portant sur la taille d'individu. La variable est considérée comme une variable quantitative continue. On cherche à obtenir la taille moyenne et la dispersion des valeurs à l'aide de l'écart-type.

valeurs de la variable	[148;152[[152;156[[156;160[[160;164[[164;168[[168;172[[172;176[[176;180[
valeurs centrales	150	154	158	162	166	170	174	178	effectif total
effectifs	5	12	21	39	33	10	2	3	125

Les calculs donnent les résultats suivants:

somme des 125 valeurs	20294		
TAILLE moyenne	162,352	TAILLE moyenne arrondie "raisonnablement"	162,5
somme des carrés des écarts à la valeur moyenne exacte	4032,512	somme des carrés des écarts à la valeur moyenne approchée	4035,25
variance calculée selon la définition	32,260096	variance calculée selon la définition	32,282
somme des carrés des valeurs de la variable	3298804		
variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	32,260096	variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	-15,818

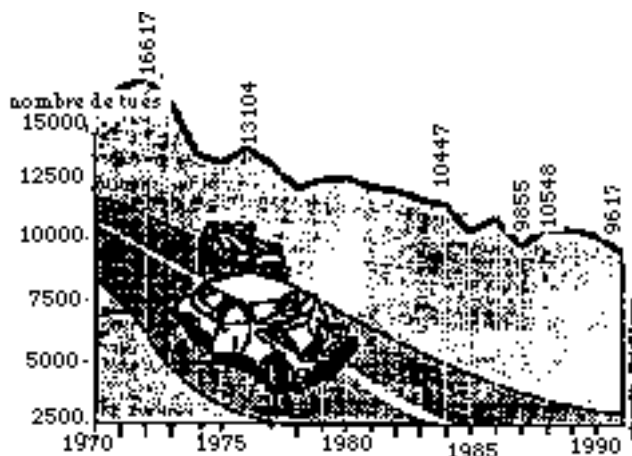
Nous repérons encore une aberration dans le résultat issu du calcul de la variance par la seconde méthode. En utilisant dans le cours des calculs, une valeur approchée tout à fait raisonnable **162,5** pour **162,352**, on obtient une valeur négative **-15,818**

A propos de l'article d'OUEST FRANCE (février 1992)

Le chiffre le plus bas depuis trente et un ans

9617 tués sur les routes en 1991

9617 morts sur les routes de France l'an dernier. Une hécatombe ! Et pourtant, le ministère des Transports a de quoi se féliciter de ce chiffre. Jamais depuis 1960 il n'y avait eu si peu de tués. Par rapport à 1990, plus de 600 vies humaines ont été épargnées.



16600 morts en 1972

149000 accidents de la route ont provoqué la mort de 9617 personnes en 1991. Un chiffre en baisse de 6,5% par rapport à l'année précédente. Ces accidents ont fait 206 000 blessés (- 8,8 %). Un Martien descendant sur notre planète serait sans doute effaré devant la satisfaction relative affichée par les pouvoirs publics devant ce désastre qui, bon an mal an raye de la carte une ville de la taille de La Ferté-Bernard

Si l'on oublie les bonheurs perdus, les vies brisées, les victimes à tout jamais handicapées, c'est pourtant vrai que ces statistiques sont encourageantes. Songez qu'en 1972, la courbe ascendante des accidents de la route avait culminé à 16600 morts et près de 400000 blessés. De cette année-là, date la véritable mobilisation contre le massacre du bitume. Trois mesures furent prises coup sur coup, la limitation de vitesse, le port de la ceinture de Sécurité hors-agglomération et le casque obligatoire pour les motards

Vitesse limitée à 50 km/h en ville : la principale explication au bon résultat 1991, selon le secrétariat d'Etat aux Transports.



Un mieux dans l'ouest aussi

Les chiffres pour l'Ouest épousent la courbe encourageante notée au plan national. Le bilan fait état de 1334 tués. 52 de moins qu'en 1990. Les accidents sont en baisse de 11,70%, les tués de 3,75% et les blessés de 12,2%. Les services de police et de gendarmerie ont constaté 4753 accidents en Bretagne (504 morts. et 6120 blessés). 3021 accidents en Basse-Normandie (274 morts et 4181 blessés) et 6129 accidents en Pays de Loire (556 tués et 8539 blessés)

Comment un article de presse régionale que nombre de lecteurs parcourent en diagonale s'avère requérir un traitement parallèle écrit explicitant le modèle mathématique sous-jacent pour parvenir à la compréhension et au contrôle de la validité de ce qui est énoncé.

Voici un article de journal destiné à tous les publics. Certes une lecture en diagonale demeure toujours possible mais :

- quelles informations le lecteur en tire-t-il ?
- comment contrôle-t-il sa compréhension du texte ?
- à quelles activités se livre-t-il pour procéder à ce contrôle ?

Tant de questions qu'en particulier un enseignant peut être amené à se poser non seulement pour lui-même, mais aussi en relation avec la part qu'il prend dans le développement de la capacité à lire, du savoir lire des individus. Il se trouve que ce texte fait aussi référence au domaine de la statistique. Plus en profondeur, le contrôle de la validité de l'information que le lecteur peut retirer de la lecture de cet article, va requérir des connaissances déclaratives et des connaissances procédurales rattachées au domaine des mathématiques. Pour contrôler cette compréhension, nous suggérons au lecteur de compléter les tableaux ci-dessous. Les informations contenues dans les cases sont des données numériques quantitatives.

année 1991	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Bretagne	a	a	a	b	c	c
Basse-Normandie	a	a	a	b	c	c
Pays de Loire	a	a	a	b	c	c
Ouest France	b	b	a	b	c	c
	a	a	a	b	c	c

	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Ouest 1990	g	g	g-e	b	c	c
Ouest 1991	d	d	a	d	c	c
variation 91/90	a	a	a	f		
France 1990	g	g	e	b	c	c
France 1991	a	a	a	b	c	c
variation 91/90	a	a	f	f		

(a) ces données sont accessibles directement dans le texte. Le recours au surlignage, sorte de pratique écrite de la lecture d'un texte, permet de mettre en relief cette information.

(b) ces données sont indirectement accessibles par l'intermédiaire des données explicites. Elles résultent de ces dernières par le recours à une procédure additive qui met en jeu l'addition ou l'opération réciproque, la soustraction.

Si nous procédons à un traitement des colonnes : le nombre de victimes est la somme du nombre de tués et du nombre de blessés.

Si nous procédons à un traitement des lignes : la zone géographique OUEST réunit les trois régions Bretagne, Basse-Normandie et Pays de Loire. Il en résulte en vertu de la propriété selon laquelle le cardinal - c'est à dire le nombre d'éléments, l'effectif - de la réunion de trois ensembles disjoints deux à deux est égal à la somme des cardinaux de chacun des ensembles. que les effectifs recherchés concernant l'OUEST s'obtiennent par addition des effectifs relatifs à chacune des trois régions.

(c) Ceci ne constitue plus du tout une information accessible même indirectement dans le texte. Il s'agit d'un traitement choisi par le lecteur pour contrôler la validité de l'information que l'auteur de l'article souhaite faire passer qui pourrait se résumer à cette phrase : **la situation semble s'améliorer !**

Pour ce faire, nous recourons à la notion de "quantité relative" qui s'appuie sur une procédure multiplicative qui met en jeu multiplication ou son opération réciproque, la division. Dans un langage formalisé, nous pouvons coder cette procédure dans le registre algébrique, de la façon suivante : $Q = 100 \frac{X}{Y}$. Nous obtenons de cette façon un indicateur quantitatif rendant comparables les données entre les diverses régions ou zones géographiques. Bien qu'il se rattache concrètement à un nombre entier, le nombre de tués comptabilisés sur 100 accidents, ce nombre peut être aussi considéré comme un nombre réel et être éventuellement fourni par une approximation décimale au dixième ou au centième près. Il convient de le considérer comme une construction mentale humaine visant rendre compte d'une certaine réalité et non comme un objet ayant une existence propre dans la nature. Même un platonicien convaincu ne serait sans doute pas choqué par un telle affirmation.

(d) Ces données sont lisibles dans le premier tableau.

(e) L'obtention de cette donnée s'appuie sur une procédure additive. La phrase du texte "Par rapport à 1990, plus de 600 vies...épargnées" indique la procédure ainsi formalisée : $X_{1990} - Y = X_{1991}$ et en nous plaçant dans l'hypothèse minimale où $Y = 600$ cela donne $X_{1990} - 600 = X_{1991}$. Le résultat recherché est alors $X_{1990} = X_{1991} + 600$

(f)(g) Il s'agit sans doute là de la procédure mathématique la plus complexe et subtile impliquée dans le traitement requis pour comprendre le texte.

Nous la formalisons ainsi :

$$Q = \frac{X_{1991} - X_{1990}}{X_{1990}}$$

Le recours à une représentation du registre algébrique offre des possibilités de traitement particulièrement pertinentes pour la résolution de notre problème.

Ainsi nous pouvons tour à tour exprimer chacune des trois variables de la façon suivante, ces transformations étant permises par des règles algébriques que nous ne développons pas ici, :

$$X_{1991} = (1+Q)X_{1990}$$

$$X_{1990} = \frac{X_{1991}}{1+Q}$$

L'application adéquate de ces formules permet d'obtenir les résultats recherchés .

année 1991	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Bretagne	4753	6120	504	6624	10,60	139,36
Basse-Normandie	3021	4181	274	4455	9,06	147,46
Pays de Loire	6129	8539	556	9095	9,07	148,39
Ouest	13903	18840	1334	20174	9,59	145,10
France	149000	206000	9617	215617	6,45	144,70

	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Ouest 1990	15745	21457	1386	22843	8,80	145,08
Ouest 1991	13903	18840	1334	20174	9,59	145,10
variation 91/90	-11,7%	-12,20%	-3,75%	-11,68%		
France 1990	159358	225877	10217	236094	6,41	148,15
France 1991	149000	206000	9617	215617	6,45	144,70
variation 91/90	-6,5%	-8,8%	-5,8%	-8,67%		

Il est encore possible d'exploiter un peu loin cet article. Mais cette fois cela implique la mise en œuvre de la lecture d'un graphique, seule information exprimée dans un registre sémiotique iconique exploitable. En effet l'image du panneau de signalisation n'est là que pour renforcer l'information relative à la mesure de limitation de vitesse. Ce graphique débarrassé de l'image des automobiles accidentées est une représentation graphique du

domaine mathématique. Il s'agit de la représentation graphique d'une variable chronologique, celle qui fournit le nombre de tués chaque année, c'est une fonction qui à la variable "temps" associe la variable "nombre de tués". L'axe temporel est l'axe des abscisses. L'axe des ordonnées fournit le nombre de tués.

Nous pouvons alors observer une tendance, le nombre de tués a tendance à être de plus en faible, ou de moins en moins fort, le nombre de tués a tendance à diminuer globalement.

Une autre façon de traiter cette information est de réaliser une conversion de ce graphique, exprimé dans un registre géométrico-iconique, en un tableau, registre iconique.

Ce que nous pouvons repérer immédiatement, ce sont les points pour lesquels la valeur de l'ordonnée est rappelée. Certes l'attention du lecteur est requise pour éviter de commettre une erreur sur la détermination de la valeur de l'abscisse, c'est à dire la date. Nous pouvons alors compléter le tableau suivant :

année	1972	1977	1985	1987	1988	1991
nombre de tués	16617	13104	10447	9855	10548	9617

Pour constituer un tableau rapportant les valeurs relativement aux 22 années de 1970 à 1991, il y a nécessité de recourir à une règle graduée et à une règle de proportionnalité pour obtenir les valeurs. Cette procédure paraît coûteuse pour le lecteur.

Le texte nous rapporte aussi une valeur approximative du nombre des victimes en 1972 : 400000. Ce nombre peut alors être comparé à celui du nombre des victimes en 1991 : 215617. Une procédure du type (f) permet d'obtenir la réponse :

$$Q = \frac{X_{1991} - X_{1972}}{X_{1972}} = \frac{215617 - 400000}{400000} = -0,4609575$$

d'où une chute de 46,09% du nombre des victimes entre 1972 et 1991.

Il s'est effectivement passé quelque chose !

Examinons maintenant la variation entre 1990 et 1991. Les valeurs absolues sont effectivement en baisse, ce qui est un acquis incontestable puisqu'il y a moins d'accidents, moins de victimes et moins de tués. Nous pouvons aussi regarder les effets des accidents. Le facteur Q dont le mode de calcul est fourni en (f)(g), nous apporte une information tout à fait intéressante :

année 1991	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents	années 1990 1991	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Bretagne	10,60	139,36	Ouest 1990	8,80	145,08
Basse-Normandie	9,06	147,46	Ouest 1991	9,59	145,10
Pays de Loire	9,07	148,39	France 1990	6,41	148,15
Ouest	9,59	145,10	France 1991	6,45	144,70
France	6,45	144,70			

Pour accroître les contrastes, réalisons une sorte de zoom en rapportant les nombres des victimes et des tués à 1000 accidents. Cette démarche est recevable puisque le nombre des accidents est de plusieurs milliers.

année 1991	nombre de tués pour 1000 accidents	nombre de victimes pour 1000 accidents	années 1990 1991	nombre de tués pour 1000 accidents	nombre de victimes pour 1000 accidents
Bretagne	106,0	1393,6	Ouest 1990	88,0	1450,8
Basse-Normandie	90,6	1474,6	Ouest 1991	95,9	1451,0
Pays de Loire	90,7	1483,9	France 1990	64,1	1481,5
Ouest	95,9	1451,0	France 1991	64,5	1447,0
France	64,5	1447,0			

Nous sommes alors frappé par le fait le taux des victimes par accident est le plus faible en Bretagne mais que ce gain est cruellement compensé par un perte sur le taux de mortalité par accident. 1000 accidents provoquent la mort de 16 personnes de plus en Bretagne qu'en Pays de Loire ou en Basse-Normandie.

Si maintenant nous nous intéressons à la variation entre 1990 et 1991, force est de constater que sur la France, 1000 accidents ont impliqué en moins 4 victimes en 1991, ce qui constitue un gain. Mais nous voyons que le taux de mortalité reste inchangé 64 tués pour 1000 accidents. Ainsi les accidents sont-ils toujours aussi mortels mais ils sont moins "blessants". En revanche, en OUEST si le nombre des victimes pour 1000 accidents restent le même :1451, le nombre de tués s'est accru de 7 personnes.

En OUEST les accidents tuent davantage en 1991 qu'en 1990 !

Cette conclusion n'était nullement explicite dans le texte. Il y a eu nécessité d'un traitement spécifique prenant appui sur des notions et des méthodes statistiques pour faire surgir cette information de celles que l'article nous apportaient. Une lecture en diagonale ne paraît pas conduire à une telle conclusion qui nuance quelque peu celle qui est donnée dans l'article : **Un mieux dans l'OUEST aussi !**

ESTIMATION :

11. Estimer un paramètre

Estimer, c'est attribuer une valeur ou une modalité à un paramètre inconnu tel qu'en particulier une moyenne, une variance, un écart-type, une proportion, une médiane, un mode, un effectif

Estimer une proportion, ou une moyenne, une variance, un écart-type d'une variable sur une **population** de taille N finie ou de taille infinie, c'est chercher à attribuer une valeur numérique approximative à l'un de ces paramètres inconnus à partir des données observées (x_1, x_2, \dots, x_n) , réalisation d'un **n-échantillon** (X_1, X_2, \dots, X_n) de la **variable X**. La réalisation des observations (x_1, x_2, \dots, x_n) est effectuée par un **tirage au hasard dans la population**.

Le résultat de cette recherche est une **estimation** de la moyenne, de la proportion ou de la variance. Mais l'opération s'appelle aussi une **estimation**.

L'outil permettant de réaliser cette estimation s'appelle un **estimateur**.

Un estimateur est une **fonction des valeurs observées** sur un échantillon, relativement au paramètre Θ (moyenne, proportion, variance) de la population (population-mère).

Un estimateur est considéré comme une **variable aléatoire** dont la distribution de probabilité et les propriétés permettent de préciser les informations relatives à l'estimation qui en découle.

Parmi les propriétés, nous citerons celle qui caractérise la notion de **biais d'un estimateur**.

Un **estimateur sans biais** est un estimateur dont l'espérance mathématique est égale à la valeur théorique du paramètre que l'on cherche à estimer.

L'estimation peut être **ponctuelle** ou **par intervalle** (par intervalle de confiance, fourchette d'estimation).

L'estimation ponctuelle est la valeur unique, considérée comme la meilleure sur l'espace des valeurs du paramètre, attribuée par l'estimateur choisi au paramètre inconnu.

L'estimation par intervalle de confiance revient à fournir un intervalle aléatoire $(t_1; t_2)$ dépendant de l'échantillon de telle sorte que la probabilité pour que l'intervalle $(t_1; t_2)$ contienne le paramètre soit connue et égale au niveau de confiance noté $1 - \alpha$.

L'idée est que si l'expérience était répétée un grand nombre de fois dans des conditions rigoureusement identiques, cet intervalle recouvrirait le paramètre θ dans $100(1-\alpha)\%$ des cas.

Cet intervalle peut être

- unilatéral à droite :

$$\text{Prob} \{ [t_1; +\infty[\text{ contient } \theta \} = \text{Prob} \{ t_1 \leq \theta \} = 1 - \alpha$$

- unilatéral à gauche :

$$\text{Prob} \{]-\infty; t_2] \text{ contient } \theta \} = \text{Prob} \{ \theta \leq t_2 \} = 1 - \alpha$$

- bilatéral :

$$\text{avec } \alpha_1 + \alpha_2 = \alpha \quad \text{Prob} \{ t_1 \leq \theta \} = 1 - \alpha_1 \text{ et } \text{Prob} \{ \theta \leq t_2 \} = 1 - \alpha_2$$

- bilatéral symétrique:

$$\text{Prob} \{ t_1 \leq \theta \} = 1 - \alpha/2 \text{ et } \text{Prob} \{ \theta \leq t_2 \} = 1 - \alpha/2$$

$$\text{Prob} \{ [t_1; t_2] \text{ contient } \theta \} = 1 - \alpha$$

12. Estimation d'une moyenne μ et d'une variance σ^2

12.1. Conditions d'utilisation:

- La variance σ^2 de la population est inconnue
 - L'échantillon est obtenu aléatoirement par n tirages avec ou sans remise dans une population de taille inconnue ou avec remise dans une population de taille N

- La taille n de l'échantillon est supérieure à 30 ou quelconque si la variable est distribuée selon une loi gaussienne sur la population

12.2. Estimateur:

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

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

La variable $Y = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}}$ est une variable de Student $T_{(n-1)}$ à n-1 ddl

12.3. Estimation ponctuelle de la moyenne:

Elle est obtenue à partir de la "moyenne calculée avec les valeurs observées sur l'échantillon", c'est à dire le résultat de \bar{X} sur l'échantillon: $m = \frac{1}{n} \sum_{i=1}^{i=n} x_i$

12.4. Estimation ponctuelle de la variance de la population:

Elle est obtenue à partir de la "variance calculée avec les valeurs observées sur l'échantillon", c'est à dire le résultat de S^2 sur l'échantillon

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{i=n} (x_i - m)^2 = \frac{n}{n-1} \sigma_{\text{échantillon}}^2$$

12.5. Estimation par intervalle de confiance bilatéral symétrique:

On choisit un niveau de confiance $1 - \alpha$. On détermine la valeur k à partir d'une table de la variable de Student T_{n-1} à n-1 degrés de liberté telle que $\text{Prob}\{T_{n-1} < k\} = 1 - \frac{\alpha}{2}$

$m - k \frac{s}{\sqrt{n}} < \mu < m + k \frac{s}{\sqrt{n}}$

12.6. Estimation par intervalle de confiance à droite:

On choisit un niveau de confiance $1 - \alpha$. On détermine la valeur k à partir d'une table de la variable de Student T_{n-1} à $n-1$ degrés de liberté telle que $\text{Prob}\{T_{n-1} < k\} = 1 - \alpha$

$$m - k \frac{s}{\sqrt{n}} < \mu$$

12.7. Estimation par intervalle de confiance à gauche:

$$\mu < m + k \frac{s}{\sqrt{n}}$$

12.8. Taille de l'échantillon pour une précision fixée:

Pour obtenir une estimation par intervalle bilatéral symétrique de m au niveau de confiance $1 - \alpha$ avec une précision absolue fixée à l'avance de $\pm \Delta$, où $\Delta > 0$ et $|t_1 - t_2| < 2\Delta$ si l'intervalle est $[t_1; t_2]$, il convient de prélever un échantillon de taille n avec la valeur k obtenue à partir d'une table de la variable de Student T_{n-1} à $n-1$ degrés de liberté telle que $\text{Prob}\{T_{n-1} < k\} = 1 - \frac{\alpha}{2}$

$$n \geq \frac{s^2}{\Delta^2} k^2$$

12.9. Compléments et remarques à propos de l'estimation d'une moyenne et d'une variance.

13. Estimation d'une proportion π

13.1. Conditions d'utilisation (cas n°1):

- L'échantillon est obtenu aléatoirement par n tirages
 - Si la taille N de la population est finie, les tirages doivent être avec remise mais cette condition peut être négligée si le taux de sondage est tel que $\frac{n}{N} < 0,1$

- Pour des conditions optimales, la taille n de l'échantillon devrait être supérieure à 100 et l'estimation ponctuelle devrait être comprise entre 0,1 et 0,9, sinon il conviendrait de consulter des documents de statistique précisant d'autres conditions

13.2. Estimateur:

La formule donnant "proportion des cas favorables calculée à partir des valeurs observées sur l'échantillon": $F_n = \frac{R_n}{n}$ où R_n est la variable qui associe à chaque échantillon le nombre d'observations ayant le caractère étudié parmi les n observations de l'échantillon. $X = nF_n = R_n = B(n, \pi)$ variable binomiale telle que l'on approche

$$Y = \frac{R_n - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{n\left(\frac{R_n}{n} - \pi\right)}{\sqrt{n}\sqrt{\pi(1-\pi)}} = \frac{\left(\frac{R_n}{n} - \pi\right)}{\frac{1}{\sqrt{n}}\sqrt{\pi(1-\pi)}} \text{ par une variable de Laplace-Gauss LG}(0;1)$$

Ainsi approximativement la variable $Y = \frac{F_n - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$ suit une loi de Laplace-Gauss

LG(0;1)

13.3. Estimation ponctuelle:

$$p = f_n$$

où f_n est la valeur de F_n calculée à partir de l'échantillon obtenu à l'issue de n tirages

La variance $\frac{\pi(1-\pi)}{n}$ est estimée ponctuellement soit par $\frac{f_n(1-f_n)}{n-1}$ soit par sa valeur maximale $\frac{1}{4(n-1)}$.

13.4. Estimation par intervalle de confiance bilatéral symétrique:

On choisit un niveau de confiance $1 - \alpha$. On détermine la valeur k à partir d'une table de la variable de Laplace-Gauss LG(0;1) telle que $\text{Prob}(LG(0;1) < k) = 1 - \frac{\alpha}{2}$

Les deux situations usuelles sont:

- Intervalle de confiance à 95% : $k=1,9600$
- Intervalle de confiance à 99% : $k=2,5758$

$$f_n - k\sqrt{\frac{f_n(1-f_n)}{n-1}} \leq \pi \leq f_n + k\sqrt{\frac{f_n(1-f_n)}{n-1}}$$

13.5. Estimation par intervalle de confiance à droite:

On choisit un niveau de confiance $1 - \alpha$. On détermine la valeur k à partir d'une table de la variable de Laplace-Gauss $LG(0;1)$ telle que $\text{Prob}(LG(0;1) < k) = 1 - \alpha$

Les deux situations usuelles sont:

- Intervalle de confiance à 95% : $k=1,6449$
- Intervalle de confiance à 99% : $k=2,3263$

$$f_n - k\sqrt{\frac{f_n(1-f_n)}{n-1}} \leq \pi \leq 1$$

13.6. Estimation par intervalle de confiance à gauche:

$$0 \leq \pi \leq f_n + k\sqrt{\frac{f_n(1-f_n)}{n-1}}$$

13.7. Taille de l'échantillon pour une précision fixée:

Pour obtenir une estimation par intervalle bilatéral symétrique de m au niveau de confiance $1 - \alpha$ avec une précision absolue fixée à l'avance de $\pm \Delta$, où $\Delta > 0$ et $|t_1 - t_2| < 2\Delta$ si l'intervalle est $[t_1; t_2]$, il convient de prélever un échantillon de taille n telle que: $\text{Prob}(LG(0;1) < k) = 1 - \alpha/2$

$$n \geq \frac{p_0(1-p_0)}{\Delta^2} k^2$$

où p_0 est une estimation ($0 < p_0 \leq 0,5$) *a priori* de la proportion maximum que l'on peut raisonnablement envisager dans la population étudiée.

13.8. Compléments et remarques à propos de l'estimation d'une proportion.

TESTER UNE HYPOTHESE

14. Quelques idées générales sur les tests statistiques

Un **test** est une opération sur laquelle on prend appui pour prendre une décision de choix entre deux **hypothèses** alternatives sur la base des informations issues d'un **échantillon**.

Un test statistique d'hypothèse consiste à **définir une règle de décision**.

Une hypothèse est dite «*hypothèse simple*» si elle peut se ramener par exemple à une confrontation simple avec une valeur isolée du type $\theta = \theta_0$.

Une hypothèse est dite «*hypothèse composite*» dans le cas contraire, c'est à dire par exemple si elle peut se ramener à une confrontation complexe avec plusieurs valeurs du type $\theta > \theta_0$, $\theta < \theta_0$, $\theta \neq \theta_0$ ou $\theta \in E$

Si H_0 (dite «*hypothèse nulle* ») et H_1 (dite «*hypothèse alternative* ») sont ces deux hypothèses alternatives dont une seule est vraie, nous avons les 4 cas de figure suivants:

réalité inconnue	H_0 vraie	H_1 vraie
décision prise		
H_0 vraie	correcte	erreur de seconde espèce
H_1 vraie	erreur de première espèce	correcte

Risques et Probabilités d'erreur

réalité inconnue	H_0 vraie	H_1 vraie
décision prise		
ne pas rejeter H_0 ou accepter H_0 comme vraie	Prob{accepter H_0 H_0 vraie} = $1-\alpha$	Prob{accepter H_0 H_0 fausse} = β
rejeter H_0 ou accepter H_1 comme vraie	Prob{rejeter H_0 H_0 vraie} = α	Prob{rejeter H_0 H_0 fausse} = $1-\beta$

Dans l'usage courant l'hypothèse nulle H_0 joue un rôle prééminent, ce qui conduit à contrôler le risque de première espèce α en posant généralement sa valeur égale à 0,01 ou le plus souvent 0,05 et parfois 0,001.

Cette hypothèse nulle H_0 est établie à partir de divers facteurs:

- hypothèse de prudence,

- hypothèse subjective à laquelle on tient particulièrement,
 - hypothèse facile à formuler,
 - hypothèse solidement établie et non contredite jusqu'alors par l'expérience,
- dans le but de ne pas la rejeter.

Toutefois il faut avoir conscience que l'acceptation (c'est à dire le non rejet) de l'hypothèse H_0 ne signifie pas qu'elle est réellement vraie mais seulement que les informations recueillies sur l'échantillon ne la contredisent pas, et donc que rien ne nous indique que le choix de l'hypothèse alternative H_1 lui est raisonnablement préférable.

Le risque de seconde espèce β est obtenu par le calcul. Cependant cela n'est possible que dans les cas où les lois de probabilités sous l'hypothèse H_1 sont connues.

Notons que α et β sont des valeurs qui varient en sens contraire entre 0 et 1.

La probabilité $1-\beta$ d'opter avec raison pour H_1 s'appelle la **puissance du test**.

Lorsqu'on a fixé α , il faut alors choisir une variable de décision:

- qui apporte le plus possible d'informations relative au problème posé,
- dont la loi de probabilité est différente selon que H_0 ou H_1 est vraie,
- dont la loi est bien connue au moins sous la condition «*Ho est vraie*».

On définit alors la **région critique K**, c'est à dire l'ensemble des valeurs de la variable de décision qui conduisent à rejeter l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1 .

La région critique est déterminée par : $\text{Prob}\{\text{rejeter } H_0 \mid H_0 \text{ vraie}\} = \alpha$

On définit aussi la **région d'acceptation A**, c'est à dire l'ensemble des valeurs de la variable de décision qui conduisent à ne pas rejeter l'hypothèse nulle H_0 au détriment de l'hypothèse alternative H_1 .

La région critique K et la région d'acceptation A sont deux **ensembles complémentaires**.

Construire un test revient donc à déterminer *a priori* une région critique.

Démarche à suivre pour tester une hypothèse:

- identifier et formuler les deux hypothèses H_0 et H_1 ,
- déterminer la variable de décision,
- caractériser l'allure de la région critique K en fonction de l'hypothèse H_1 ,
- déterminer par le calcul la région critique K en fonction du risque α de première espèce,
- calculer, quand cela est possible, la puissance du test $1-\beta$,
- calculer la valeur expérimentale de la variable de décision à partir des valeurs obtenues sur l'échantillon,
- formuler la conclusion en terme de rejet ou non rejet de l'hypothèse nulle H_0 .

Les tests sont classés selon leur objet. Citons pour exemple ceux que l'on est amené à utiliser dans des recherches en Sciences de l'éducation :

Tests paramétriques, c'est à dire visant à tester une hypothèse relative à un ou plusieurs paramètres (moyenne, variance, proportion, etc.) d'une variable aléatoire spécifiée ou non:

Tests non-paramétriques dont la construction est établie sur la base d'une *fonction des observations* issues d'un échantillon aléatoire, de loi de probabilité indépendante de la connaissance de la distribution de la loi de probabilité sur la population.

Tests d'homogénéité :

- tests de comparaison d'une moyenne à une valeur fixée,
- tests de comparaison d'une variance à une valeur fixée,
- tests de comparaison d'une proportion à une valeur fixée,
- tests de comparaison d'une différence de deux moyennes à une valeur fixée,
- tests de comparaison de deux moyennes ,
- tests de comparaison de deux variances ,
- tests de comparaison de deux proportions ,

Tests d'adéquation (ou d'ajustement)d'une loi observée avec d'une loi théorique.

Tests d'indépendance de deux caractères.

15. Comparaison de deux proportions π_1 , π_2

15.1. Conditions d'utilisation:

Deux échantillons sont obtenus aléatoirement par n_1 et n_2 tirages de façon indépendante dans chacune des deux populations au sein desquelles on observerait respectivement les proportions inconnues π_1 et π_2

Les tirages doivent être avec remise mais cette condition peut être négligée si le taux de sondage est tel que $\frac{n}{N} < 0,1$

Les tailles n_1 et n_2 sont supérieures à 100

15.2. Statistique

La variable donnant "proportion des cas favorables calculée à partir des valeurs observées sur l'échantillon": $F_n = \frac{R_n}{n}$ où R_n désigne la variable donnant le nombre d'observations ayant le caractère étudié parmi les n observations de l'échantillon.

On considère la variable de décision suivante $D = F_{n1} - F_{n2} = \frac{R_{n1}}{n_1} - \frac{R_{n2}}{n_2}$

Sous l'hypothèse H_0 , on suppose que D est approximativement une variable de Laplace-

Gauss de paramètres $m_D=0$ et $\sigma_D = \sqrt{f_0(1-f_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ et $f_0 = \frac{r_1+r_2}{n_1+n_2}$ où r_1 et r_2 sont supérieurs à 10

On note la valeur expérimentale de la statistique D : $d = f_{n1} - f_{n2} = \frac{r_1}{n_1} - \frac{r_2}{n_2}$

15.3. Test bilatéral symétrique: $H_0 (\pi_1 = \pi_2)$ contre $H_1 (\pi_1 \neq \pi_2)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur k à partir d'une table de la variable de Laplace-Gauss $LG(0;1)$. telle que $\text{Prob}(LG(0;1) < k) = 1 - \frac{\alpha}{2}$

Les deux situations usuelles sont: $\alpha=5\%$: $k=1,9600$ $\alpha=1\%$: $k=2,5758$

si $-k \sigma_D < d < k \sigma_D$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

15.4. Test unilatéral à droite: $H_0 (\pi_1 \geq \pi_2)$ contre $H_1 (\pi_1 < \pi_2)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur k à partir d'une table de la variable de Laplace-Gauss $LG(0;1)$, telle que $\text{Prob}(LG(0;1) < k) = 1 - \alpha$

Les deux situations usuelles sont: $\alpha=5\%$: $k=1,6449$ $\alpha=1\%$: $k=2,3263$

si $-k \sigma_D < d$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

15.5. Test unilatéral à gauche: $H_0 (\pi_1 \leq \pi_2)$ contre $H_1 (\pi_1 > \pi_2)$

si $d < k \sigma_D$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

15.6. Compléments et remarques à propos de l'estimation d'une proportion.

16. Comparaison d'une proportion π à une valeur π_0

16.1. Conditions d'utilisation:

Un échantillon est obtenu aléatoirement par n tirages dans une population au sein de laquelle on observerait une proportion inconnue π

Les tirages doivent être avec remise mais cette condition peut être négligée si le taux de sondage est tel que $\frac{n}{N} < 0,1$

16.2. Statistique et variable de décision utilisée

La variable donnant la "proportion des cas favorables calculée à partir des valeurs observées sur l'échantillon": $F_n = \frac{R}{n}$ où R désigne la variable donnant le nombre d'observations ayant le caractère étudié parmi les n observations de l'échantillon. Pour un échantillon donné F_n prend la valeur expérimentale f_n

$$D = F_n - \pi_0 = \frac{R}{n} - \pi_0$$

On admet que D est approximativement une variable de Laplace-Gauss de paramètres $\mu_D = 0$ et $\sigma_D = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$.

On note la valeur expérimentale de D : $d = f_n - \pi_0 = \frac{r}{n} - \pi_0$

16.3. Test bilatéral symétrique: $H_0 (\pi = \pi_0)$ contre $H_1 (\pi \neq \pi_0)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur k à partir d'une table de la variable de Laplace-Gauss $LG(0;1)$. telle que $\text{Prob}(LG(0;1) < k) = 1 - \frac{\alpha}{2}$

Les deux situations usuelles sont: $\alpha=5\%$: $k=1,9600$ $\alpha=1\%$: $k=2,5758$

Les deux situations usuelles sont: $\alpha=5\%$: $k=1,9600$ $\alpha=1\%$: $k=2,5758$

si $-k \sigma_D < d < k \sigma_D$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

16.4. Test unilatéral à droite: $H_0 (\pi \geq \pi_0)$ contre $H_1 (\pi < \pi_0)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur k à partir d'une table de la variable de Laplace-Gauss $LG(0;1)$, telle que $\text{Prob}(LG(0;1) < k) = 1 - \alpha$

Les deux situations usuelles sont: $\alpha=5\%$: $k=1,6449$ $\alpha=1\%$: $k=2,3263$

si $-k \sigma_D < d$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

16.5. Test unilatéral à gauche: $H_0 (\pi \leq \pi_0)$ contre $H_1 (\pi > \pi_0)$

si $d < k \sigma_D$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

16.6. Compléments et remarques à propos de l'estimation d'une proportion.

17. Comparaison d'une moyenne μ à une valeur donnée μ_0

17.1. Conditions d'utilisation:

La variance σ^2 de la population est inconnue ainsi que sa moyenne μ

L'échantillon est obtenu aléatoirement par n tirages avec remise si la population est finie de taille N ou sans remise si la population est de taille N inconnue ou infinie, ou si le taux de sondage est inférieure à 10%

La taille n de l'échantillon est supérieure à 5 ou quelconque si la variable est distribuée selon une loi gaussienne sur la population

17.2. Statistique et variable de décision utilisée

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

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

Les valeurs expérimentales sont obtenues sur l'échantillon recueilli en calculant

$$m = \frac{1}{n} \sum_{i=1}^{i=n} x_i \quad \text{et} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{i=n} (x_i - m)^2 = \frac{n}{n-1} \sigma_{\text{échantillon}}^2$$

La variable $D = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$ est une variable de Student $T_{(n-1)}$ à $n-1$ ddl

D est considérée comme une variable aléatoire de Student T_{n-1} à $n-1$ degrés de liberté

17.3. Test bilatéral symétrique: $H_0 (\mu = \mu_0)$ contre $H_1 (\mu \neq \mu_0)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur k à partir d'une table de la variable de Student T_{n-1} à $n-1$ degrés de liberté telle que $\text{Prob}\{T_{n-1} < k\} = 1 - \frac{\alpha}{2}$

Si $\mu_0 - k \frac{s}{\sqrt{n}} < m < \mu_0 + k \frac{s}{\sqrt{n}}$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

17.4. Test unilatéral à droite: $H_0 (\mu \geq \mu_0)$ contre $H_1 (\mu < \mu_0)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur k à partir d'une table de la variable de Student T_{n-1} à ddl = $n-1$ telle que $\text{Prob}\{T_{n-1} < k\} = 1 - \alpha$

Si $\bar{m} - k \frac{s}{\sqrt{n}} < \mu_0$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

17.5. Test unilatéral à gauche: $H_0 (\mu \leq \mu_0)$ contre $H_1 (\mu > \mu_0)$

Si $\bar{m} < \mu_0 + k \frac{s}{\sqrt{n}}$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

17.6. Compléments et remarques à propos de l'estimation d'une proportion.

18. Comparaison de deux variances σ_1^2 et σ_2^2

18.1. Conditions d'utilisation:

Deux échantillons sont obtenus aléatoirement par n_1 et n_2 tirages de façon indépendante dans chacune des deux populations au sein desquelles on observerait respectivement les variances inconnues σ_1^2 et σ_2^2

Dans chacune de deux populations, la variable suit une loi de Laplace-Gauss sinon on doit avoir n_1 et n_2 supérieurs à 30.

18.2. Statistique et variable de décision utilisée

On pose :

La variable donnant la "moyenne calculée à partir des valeurs observées sur les échantillons":

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{i=n_1} X_{1i} \quad \text{et} \quad \bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{i=n_2} X_{2i}$$

Les valeurs expérimentales obtenues sur les deux échantillons recueillis sont alors

$$m_1 = \frac{1}{n_1} \sum_{i=1}^{i=n_1} x_{1i} \quad \text{et} \quad m_2 = \frac{1}{n_2} \sum_{i=1}^{i=n_2} x_{2i}$$

La variable décision suivante

$$D = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\frac{\sum_{i=1}^{i=n_1} Z_{1i}^2}{(n_1-1)}}{\frac{\sum_{i=1}^{i=n_2} Z_{2i}^2}{(n_2-1)}} = \frac{\frac{\chi^2_{(n_1-1)}}{n_1-1}}{\frac{\chi^2_{(n_2-1)}}{n_2-1}} = F_{(n_1-1; n_2-1)}$$

en considérant

$$Z_{1i}^2 = \frac{(X_{1i} - \bar{X}_1)^2}{\sigma_1^2} \quad \text{et} \quad Z_{2i}^2 = \frac{(X_{2i} - \bar{X}_2)^2}{\sigma_2^2}$$

Sous l'hypothèse H_0 énonçant $\sigma_1^2 = \sigma_2^2$ la variable de décision D devient

$$D = \frac{S_1^2}{S_2^2} = \frac{\sum_{i=1}^{i=n_1} (X_{1i} - \bar{X}_1)^2}{n_1 - 1} \cdot \frac{n_2 - 1}{\sum_{i=1}^{i=n_2} (X_{2i} - \bar{X}_2)^2}$$

La valeur expérimentale est

$$d = \frac{s_1^2}{s_2^2} = \frac{\sum_{i=1}^{i=n_1} (x_{1i} - m_1)^2}{n_1 - 1} \cdot \frac{n_2 - 1}{\sum_{i=1}^{i=n_2} (x_{2i} - m_2)^2}$$

Par commodité on suppose que la numérotation des échantillons permet de mettre la plus forte des variances estimées au numérateur. Ainsi les valeurs recueillies par D sont toujours supérieures à 1

D est considérée comme une variable de Fisher-Snedécor $F(n_1-1; n_2-1)$

18.3. Test bilatéral symétrique: $H_0 (\sigma_1^2 = \sigma_2^2)$ contre $H_1 (\sigma_1^2 \neq \sigma_2^2)$

On choisit un niveau de risque de 1ère espèce α . On détermine les valeurs k à partir d'une table d'e Fisher-Snedecor $ddl=n_1-1$ et $ddl= n_2-1$

si $d \leq k$ alors on accepte H_0 (on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α .

18.4. Compléments et remarques à propos de l'estimation d'une proportion.

19. Comparaison de deux moyennes μ_1 et μ_2 à partir de deux échantillons indépendants

19.1. Conditions d'utilisation:

Deux échantillons sont obtenus aléatoirement par n_1 et n_2 tirages de façon indépendante dans chacune des deux populations au sein desquelles on observerait respectivement **les variances inconnues σ_1^2 et σ_2^2** mais avec $\sigma_1^2 = \sigma_2^2 = \sigma^2$ de moyennes inconnues μ_1 et μ_2 ;

Dans chacune de deux populations, la variable suit une loi de Laplace-Gauss sinon on doit avoir n_1 et n_2 supérieurs à 30.

19.2. Statistique et variable de décision utilisée:

On pose la variable de décision

$$D = (\bar{X}_1 - \bar{X}_2) = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1} - \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}$$

la valeur expérimentale est

$$d = (m_1 - m_2) = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1} - \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

D est considérée comme une variable aléatoire de Laplace-Gauss de paramètres

$$\mu_D = 0 \text{ et } \sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Pour estimer σ_D on pose $(S_D)^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ où les valeurs

expérimentales de S_1^2 et S_2^2 sont

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - m_1)^2}{n_1 - 1} = \frac{n_1}{n_1 - 1} \sigma_{\text{écha } n_1}^2 \text{ et } s_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - m_2)^2}{n_2 - 1} = \frac{n_2}{n_2 - 1} \sigma_{\text{écha } n_2}^2$$

Alors la variable $Y = \frac{D}{\sqrt{\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ est considérée comme

une variable aléatoire de Student à $n_1 + n_2 - 2$ degrés de liberté

19.3. Test bilatéral symétrique: $H_0 (\mu_1 - \mu_2 = \delta_0)$ contre $H_1 (\mu_1 - \mu_2 \neq \delta_0)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur k à partir d'une table de la variable de Student $T_{(n_1 + n_2 - 2)}$ à $n_1 + n_2 - 2$ degrés de liberté telle que $\text{Prob}\{T_{(n_1 + n_2 - 2)} < k\} = 1 - \frac{\alpha}{2}$

si $-k s_D \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < d < +k s_D \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ alors on accepte H_0 (on rejette H_1) en

prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (on accepte H_1) en prenant un risque de première espèce de niveau α

19.4. Compléments et remarques à propos de l'estimation d'une proportion.

20. Test de comparaison des moyennes de deux échantillons appariés : le test T de Student

20.1. Conditions d'utilisation:

On soumet successivement un échantillon de n individus à deux « mesures » d'une variable X . Le but est de tester l'hypothèse de l'identité des deux distributions de la variable par comparaison des moyennes. La variable est une variable quantitative

20.2. Statistique et variable de décision utilisée:

Pour soutenir notre décision concernant les deux énoncés hypothétiques contradictoires, nous utilisons les outils suivants :

X_1 la variable X correspondant à la première série de mesures

X_2 la variable X correspondant à la deuxième série de mesures

X_1 et X_2 sont supposées être séparément des variables de Laplace-Gauss respectivement de moyennes μ_1 et μ_2 et d'écart-types σ_1 et σ_2 de telle sorte que la variable « différence des mesures 1 et 2 » $\Delta = X_1 - X_2$ soit aussi une variable de Laplace-Gauss $LG(\mu_\Delta ; \sigma_\Delta)$ avec $\mu_\Delta = \mu_1 - \mu_2$.

Le test repose sur l'étude de la statistique, variable de décision $D = \frac{\bar{\Delta} - \mu_\Delta}{\frac{S_\Delta}{\sqrt{n}}}$

dans laquelle $\bar{\Delta}$ est la moyenne empirique et S_Δ , l'écart-type empirique, construits sur la variable Δ . La variable D est une variable de Student $T_{(n-1)}$ à $n-1$ ddl

La mise en application du test consiste à calculer la valeur d empirique prise par D sur

un échantillon : $d = \frac{\bar{\delta} - \mu_\Delta}{\frac{s_\Delta}{\sqrt{n}}}$ et sous l'hypothèse $H_0 \mu_\Delta = 0$ donc $d = \frac{\bar{\delta}}{\frac{s_\Delta}{\sqrt{n}}}$. Ce calcul est

conduit de la même manière que celui que nous avons réalisé dans le travail sur l'estimation. $\bar{\delta}$ n'est autre qu'une estimation ponctuelle de la moyenne de la variable $\Delta = X_1 - X_2$ « différence des mesures 1 et 2 » et s_Δ une estimation ponctuelle de l'écart-type de cette même variable Δ .

20.3. Test bilatéral symétrique: $H_0 (\mu_\Delta = 0)$ contre $H_1 (\mu_\Delta \neq 0)$

On choisit un niveau de risque de 1ère espèce α . On détermine la valeur κ à partir d'une table de la variable de Student T_{n-1} à $n-1$ degrés de liberté telle que $\text{Prob}\{T_{n-1} <$

$\kappa = 1 - \frac{\alpha}{2}$ Notons de plus, que, pour $n > 80$, nous revenons à l'usage de la table de Laplace-Gauss LG (0 ; 1) qui fournit alors des valeurs approchées tout à fait acceptables.

Si $-\kappa < \delta < \kappa$ alors on accepte H_0 (et on rejette H_1) en prenant un risque de seconde espèce de niveau β sinon on rejette H_0 (et on accepte H_1) en prenant un risque de première espèce de niveau α . Le signe de la valeur δ permet d'interpréter le sens de la variation lorsque l'on a rejeté l'hypothèse H_0

21. Test d'indépendance de deux variables

21.1. cas des tableaux 2 x 2

Le problème abordé dans cette fiche est celui de l'interprétation de données classées dans un tableau 2 x 2. Ce problème peut être découpé en trois types:

Type1	col1	col2		Type2	col1	col2		Type3	col1	col2	
ligne 1	a	c	m connu	ligne 1	a	c	m connu	ligne 1	a	c	m
ligne2	b	d	n connu	ligne2	b	d	n connu	ligne2	b	d	n
	r connu	s connu	N connu		r	s	N connu		r	s	N connu

Un article de E.S. Pearson aborde ce problème statistique "The choice of statistical tests illustrated on the interpretation of data classed in a 2x2 table" publié dans *BIOMETRIKA* n°34 (pp 139-169)

21.2. Test du Khi-deux d'indépendance

	variable Y		
variable X	y1	y2	
x1	a	c	a+c
x2	b	d	b+d
	a+b	c+d	a+b+c+d = n

Il s'agit de définir une règle de décision relative à la validité de l'hypothèse d'indépendance des deux variables X et Y. Ces deux variables ne prennent que deux valeurs ou deux modalités.

Le but est de tester l'hypothèse nulle H_0 selon laquelle «la variable X est indépendante de la variable Y» contre l'hypothèse alternative selon laquelle « la variable X et la variable Y sont statistiquement dépendantes »

La notion d'indépendance d'événements ou de variables se fonde en statistique et en probabilité sur la relation mathématique suivante:

« Deux caractères A et B sont dits indépendants statistiquement si la proportion des individus possédant simultanément les deux caractères A et B est égale au produit de la proportion des individus possédant le caractère A et de celle des individus possédant le caractère B »

En appliquant ce principe, nous obtenons la distribution d'effectifs théoriques:

	variable X		
variable X	y1	y2	
x1	$a' = \frac{(a+b)(a+c)}{n}$	$c' = \frac{(c+d)(a+c)}{n}$	$a'+c' = a+c$
x2	$b' = \frac{(a+b)(b+d)}{n}$	$d' = \frac{(b+d)(c+d)}{n}$	$b'+d' = b+d$
	$a'+b' = a+b$	$c'+d' = c+d$	$a'+b'+c'+d' = n$

21.3. Statistiques et variables de décision utilisées

Il s'agit de mesurer l'écart entre la distribution des effectifs théoriques (T) et la distribution des effectifs observés (O). Le critère, variable de décision, $D^2 = \sum_{i=1}^{i=s} \frac{(O_i - T_i)^2}{T_i}$

où O_i = effectif observé et T_i = effectif théorique sous la contrainte d'indépendance des deux variables, est le critère utilisé ici.

Si aucun des effectifs théoriques n'est inférieur à 5, la variable D^2 suit approximativement une loi du Khi-Deux à 1 degré de liberté.

Si un des effectifs théoriques est inférieur à 5, on recourt à la variable hypergéométrique du test exact de Fisher.

Grâce à une table de la loi du Khi-Deux (variable de Pearson), nous pouvons connaître la valeur théorique critique k telle que la probabilité de l'évènement $\{\chi^2 > k\}$ soit approximativement égale à $\alpha = 1\%$ ou $\alpha = 5\%$. Ainsi par lecture pour un Khi-Deux à $ddl = 1$, nous obtenons les valeurs : $k = 6,63$ pour 1% et $k = 3,84$ pour 5% .

Si l'écart expérimental d^2 calculé pour D^2 à partir de l'observation reste inférieur à k alors nous ne pouvons rejeter H_0 en revanche si cet écart expérimental d^2 dépasse k alors nous rejetons l'hypothèse H_0 avec 1% ou 5% de risque de commettre une erreur de première espèce c'est à dire de rejeter à tort l'hypothèse nulle. Le non-rejet de l'hypothèse d'indépendance des deux variables revient à rejeter l'hypothèse de dépendance avec un risque de seconde espèce de niveau β , en général inconnu.

22. Test exact de Fischer

Test réalisé à partir d'un "tableau à quatre cases" contenant des "petits nombres" Ce que nous développons ici, est abordé selon la perspective d'un problème de type I. Dans une classe de 15 élèves de CE2, on se propose de comparer l'efficacité de deux algorithmes de la multiplication. Après expérience, on obtient les résultats suivants à une épreuve de contrôle finale, évaluée selon deux modalités "réussite" ou "échec" :

	réussites	échecs	
algorithme A	2	6	8
algorithme B	4	3	7
	6	9	15

A première vue, nous pouvons constater que 4 sur 7 (soit 57,1%) réussissent avec l'algorithme B contre 2 sur 8 (soit 25%) avec l'algorithme A . L'efficacité paraît être ainsi mise en évidence. Examinons cette situation d'un point de vue probabiliste, et testons l'hypothèse selon laquelle "l'algorithme B n'est pas plus efficace que l'algorithme A". Sous cette hypothèse, le résultat à l'épreuve est indépendant de l'algorithme choisi. Cela signifie que sur les 15 élèves, 6 réussiront de toute façon. En constituant le groupe de 7 élèves utilisateurs de l'algorithme B, nous avons tout simplement extrait au hasard 4 élèves qui réussiront parmi les 6.

Quelle est alors la probabilité α d'obtenir au moins 4 élèves qui réussiront avec l'algorithme B ?

T1	R	E		T2	R	E		T3	R	E	
A	2	6	8	A	1	7	8	A	0	8	8
B	4	3	7	B	5	2	7	B	6	1	7
	6	9	15		6	9	15		6	9	15

Il revient à calculer la probabilité d'obtenir l'un des trois tableaux ci-dessus. La probabilité d'obtenir un tableau est calculée à partir de la loi d'une variable hypergéométrique dont les paramètres sont déterminés par les valeurs marginales du tableau. Le cas général peut être représenté ainsi

T	col1	col2	
ligne 1	a	b	m
ligne 2	c	d	n
	r	s	N

Il est d'abord remarquable que les valeurs m, n, r, s étant connues et telles que $N=m+n+r+s$, le tableau est complètement déterminé dès qu'une case est connue. Il s'agit d'un problème à 1 degré de liberté. Considérons la variable X donnant la valeur a de la

case (1;1). $X = H(N; m; \frac{r}{N})$. Avec $P(X = a) = \frac{m! n! r! s!}{a! b! c! d! N!}$ et par ailleurs on a $E(X) = \frac{rm}{N}$ et

$V(X) = \frac{mnrs}{N^2(N-1)}$. Dans notre exemple, le calcul porte sur la valeur de la case (2;1).,

d'où $X = H(15; 7; \frac{6}{15})$

$$P(T1) = P(X= 4) = \frac{8! 7! 6! 9!}{2! 6! 4! 3! 15!} = \frac{28}{143}$$

$$P(T2) = P(X= 5) = \frac{8! 7! 6! 9!}{1! 7! 5! 2! 15!} = \frac{24}{715}$$

$$P(T3) = P(X= 6) = \frac{8! 7! 6! 9!}{0! 8! 6! 1! 15!} = \frac{1}{715}$$

$$P(T1) + P(T2) + P(T3) = \frac{3}{13} \approx 0,23$$

C'est dire qu'un tel événement a plus de 20% de chance de se produire par hasard. Aussi nous ne pouvons attribuer les résultats à une efficacité supérieure de l'algorithme B sur l'algorithme A. En d'autres termes, nous ne pouvons rejeter l'hypothèse nulle, H_0 selon laquelle "l'algorithme B n'est pas plus efficace que l'algorithme A" (dans le sens : *aussi efficace...*) . A moins que l'on accepte de se tromper environ une fois sur 4.

23. Étude simultanée de deux variables qualitatives

23.1. Introduction

Dans une enquête par questionnaire, on conduit d'abord une analyse dite « univariée » des caractères qualitatifs et quantitatifs fondée sur la description (tris à plat, analyses des tendances et de la dispersion) en procédant question par question. Ensuite on essaie d'exploiter les résultats pour répondre à des interrogations mettant en œuvre des procédures d'estimation ou des tests statistiques d'hypothèses mais en prenant en considération qu'une seule question à la fois. Mais il est aussi très intéressant d'aborder des analyses qui vont tenir compte de l'étude conjointe de deux variables. Dans cette perspective nous allons aborder les questions suivantes :

Comment peut-on analyser simultanément les réponses à deux questions qualitatives ? Peut-on ou non établir une dépendance entre les réponses fournies à deux questions qualitatives fermées ? Comment tester les hypothèses de dépendance/indépendance ? Ceci est l'objet de la partie traitant du **test d'indépendance par la méthode du Khi-deux (χ^2)**

Peut-on établir ou non l'homogénéité des réponses fournies à une question qualitatives fermées selon les catégories établies par les réponses à une autre question qualitative fermée ? Ceci est l'objet de la partie traitant du **test d'homogénéité par la méthode du Khi-deux (χ^2)** visant établir l'égalité des profils lignes ou des profils colonnes d'un tableau croisé.

La distribution des fréquences observée sur l'échantillon pour une question à réponses qualitatives diffère-t-elle significativement d'une distribution connue pour l'ensemble de la population ? Ceci est l'objet de la partie traitant du **test d'adéquation par la méthode du Khi-deux (χ^2)**

Deux séries successives de données de type présence-absence ou échec-réussite relevées sur le même échantillon font-elles apparaître des différences statistiquement significatives (Test de Mac Nemar) ? Ceci est l'objet de la partie traitant du **test de Mac Nemar**.

Pour répondre à la première question, nous sommes amenés à procéder à un « tri croisé » et à construire ce que nous appelons un « tableau croisé » qui constitue un outil très efficace.

23.2. Tableau croisé

Considérons les N individus interrogés qui ont répondu à tout un ensemble de questions. Supposons que l'on souhaite étudier simultanément 2 caractères (ou

variables) nominaux ou qualitatifs¹. Rappelons que le tri à plat donne la répartition des individus interrogés selon toutes les modalités de chaque caractère. C'est une perte d'information qui empêche d'analyser la répartition conjointe. Essayons de modéliser cette situation.

Soit V_A et V_B les deux variables à étudier ayant respectivement l et c modalités. Ainsi, (A_1, A_2, \dots, A_l) sont les modalités de V_A et (B_1, B_2, \dots, B_c) sont les modalités de V_B . Le tri croisé est l'opération qui consiste à dénombrer les individus relatifs à tous les croisements : $(A_1, B_1), (A_1, B_2), (A_1, B_3) \dots (A_l, B_c)$. Le nombre de croisements est $l \times c$.

Par exemple, si l'objectif est d'analyser l'effet d'un suivi personnalisé sur la qualité de production d'un mémoire professionnel de fin d'étude, il convient de croiser les deux variables :

- V_A désignant le suivi personnalisé dont les modalités sont : $(A_1, A_2) =$ (suivi personnalisé, suivi non personnalisé)
- V_B désignant la qualité de la production dont les modalités sont : $(B_1, B_2, B_3) =$ (mémoire validé avec mention, mémoire validé sans mention, mémoire non validé)

Supposons que le « tableau croisé » obtenu pour 250 étudiants soit le suivant :

Tableau 23-1 observé de la répartition des qualités de mémoires et des suivis

Variable V_A suivi personnalisé :	Tableau de contingence	Variable V_B « qualité de la production »		
		B_1 mémoire validé avec mention	B_2 mémoire validé sans mention	B_3 mémoire non validé
A_1 suivi personnalisé		80	40	27
A_2 suivi non personnalisé		40	30	33

Ce tableau croisé est également appelé « tableau de contingence », on le complète généralement par ces marges-lignes et marges-colonnes ou totaux marginaux de la manière suivante :

Variable V_A suivi personnalisé :	Tableau de contingence	Variable V_B « qualité de la production »			Totaux
		B_1 mémoire validé avec mention	B_2 mémoire validé sans mention	B_3 mémoire non validé	
A_1 suivi personnalisé		80	40	27	147
A_2 suivi non personnalisé		40	30	33	103
Totaux		120	70	60	250

Généralisons maintenant la notation à une situation quelconque en recourant à l'usage de double indice. Pour repérer la ligne et la colonne. :

¹ On peut également étudier 2 caractères numériques découpés en classes ou bien encore un numérique découpé en classes et un qualitatif

$V_A :$	$V_B :$	B_1	B_2	B_j	B_c	Total
A_1		N_{11}	N_{12}		N_{1j}		N_{1c}	$N_{1.}$
A_2		N_{21}	N_{22}		N_{2j}		N_{2c}	$N_{2.}$
...								
A_i		N_{i1}	N_{i2}		N_{ij}		N_{ic}	$N_{i.}$
.....								
A_L		N_{L1}	N_{L2}		N_{Lj}		N_{Lc}	$N_{L.}$
Total		$N_{.1}$	$N_{.2}$		$N_{.j}$		$N_{.c}$	N

Pour lire le tableau, il faut donc savoir que :

- N_{ij} désigne l'effectif de la case (i, j)
- $N_{.j}$ désigne l'effectif de la colonne j
- $N_{i.}$ désigne l'effectif de la ligne i

23.2.1. Transformations du tableau croisé

Afin d'analyser les éléments remarquables de ce tableau, on peut être amené à le transformer. Selon le type d'information recherché, on peut soit le remplacer par un tableau de pourcentage, en divisant tous les nombres par l'effectif total N , ou bien, le plus souvent, calculer des pourcentages lignes ou des pourcentages colonnes.

Dans le cas des pourcentages lignes (profils lignes), à la modalité A_i , on associe la suite des pourcentages, selon la variable V_B des $N_{i.}$ individus qui possèdent la modalité A_i . Sur la ligne "total", on calcule également des pourcentages correspondant aux pourcentages moyens, ce sont les pourcentages de la répartition des modalités du caractère B

Dans le cas des pourcentages colonnes (profils colonnes), à la modalité B_j , on associe la suite des pourcentages, selon la variable V_A des $N_{.j}$ individus qui possèdent la modalité B_j . Sur la colonne total, on calcule également des pourcentages correspondant aux pourcentages moyens, ce sont les pourcentages de la répartition des modalités de la variable V_A

Ainsi, dans l'exemple donné précédemment, on peut construire les profils lignes puis les profils colonnes.

Tableau 23-2 des profils lignes

Variable V_A suivi personnalisé :	Tableau de contingence	Variable V_B « qualité de la production »			
		B_1 mémoire validé avec mention	B_2 mémoire validé sans mention	B_3 mémoire non validé	Totaux
A_1 suivi personnalisé		54,4%	27,2%	18,4%	100%
A_2 suivi non personnalisé		38,8%	29,2%	32%	100%
Profil moyen		48%	28%	24%	100%

Parmi, les étudiants ayant bénéficié d'un suivi personnalisé, on compte : 54,4% de mémoire validé avec mention, 27,2% de mémoire validé sans mention et 18,4% de mémoire non validé

Parmi, les étudiants n'ayant pas bénéficié d'un suivi personnalisé, on compte : 38,8% de mémoire validé avec mention, 29,2% de mémoire validé sans mention et 32% de mémoire non validé

Par ailleurs, la répartition des qualités de mémoire, indépendamment de la connaissance du suivi personnalisé ou non est la suivante : 48% de mémoire validé avec mention, 28% de mémoire validé sans mention et 24% de mémoire non validé

En mettant en œuvre une procédure analogue, construire le tableau des profils colonnes.

L'analyse de ces profils permet des comparaisons entre groupes d'individus. On peut suivre la même démarche pour les profils colonnes.

Tableau 23-3 des profils colonnes

Variable V_A suivi personnalisé :	Tableau de contingence	Variable V_B « qualité de la production »			
		B_1 mémoire validé avec mention	B_2 mémoire validé sans mention	B_3 mémoire non validé	Profil moyen
A_1 suivi personnalisé		66,6%	57,1%	45%	58,2%
A_2 suivi non personnalisé		33,4%	42,9%	55%	41,8%
Totaux		100%	100%	100%	100%

Parmi les étudiants ayant validé leur mémoire avec mention, on en compte 66,6% ayant eu un suivi personnalisé et 33,4% dans la situation inverse.

Parmi les étudiants ayant validé leur mémoire sans mention, on en compte 57,1% ayant eu un suivi personnalisé et 42,9% dans la situation inverse.

Parmi les étudiants n'ayant pas validé leur mémoire, on en compte 45% ayant eu un suivi personnalisé et 55% dans la situation inverse.

Enfin, en moyenne, on compte 58,2% d'étudiants ayant eu un suivi personnalisé et 41,8% dans la situation inverse

23.2.2. Représentations graphiques du tableau croisé.

On peut également représenter graphiquement ces profils lignes et ces profils colonnes. Dans le cas des profils lignes, on pourra représenter la répartition des qualités de production des étudiants ayant eu un suivi personnalisé, puis la répartition de ceux n'ayant pas eu de suivi personnalisé. On pourra comparer les deux diagrammes obtenus en fonction de la répartition moyenne. Dans le cas des profils colonnes, on construira trois diagrammes de répartition du type de suivi, le premier concernera les étudiants ayant validé avec mention, le second concernera les étudiants ayant validé sans mention, et le troisième concernera les étudiants n'ayant pas validé.

23.2.3. Raisonner statistiquement à partir du tableau croisé.

L'analyse du tableau croisé conduite précédemment est essentielle mais ne permet pas d'extrapoler les résultats obtenus à l'ensemble des individus de la population. Pour cela plaçons-nous dans la situation d'un sondage aléatoire dans une population. On admet également que la taille de l'échantillon est petite devant la taille de la population afin de négliger le problème des individus interrogés plusieurs fois (tirage sans remise équivalent alors au tirage avec remise quand le taux de sondage est inférieur à 10%).

23.3. La notion fondamentale d'indépendance statistique.

Désignons maintenant les proportions calculées sur la population totale :

- π_{ij} la proportion d'individus admettant simultanément les modalités A i et B j
- $\pi_{i.}$ la proportion de ceux qui ont la modalité A i
- $\pi_{.j}$ la proportion de ceux ayant la modalité B j

Si on ne connaît pas les réponses relatives à la population totale, on peut formuler des hypothèses concernant ces valeurs, en particulier se poser la question de l'indépendance des deux variables V_A et V_B

Par définition :

deux variables statistiques V_A et V_B sont **statistiquement indépendantes** si pour tous les couples (i ,j) on a $\pi_{ij} = \pi_{i.} \times \pi_{.j}$

23.3.1. Caractérisation de l'hypothèse d'indépendance

Dans un sondage portant sur N personnes, si les deux variables V_A et V_B sont indépendantes, on peut donc déterminer les effectifs correspondant à cette hypothèse d'indépendance, à partir des effectifs marginaux, appelés effectifs théoriques.

On peut vérifier que l'effectif théorique de la case (i, j) est égal à : $N \cdot \pi_{i.} \times \pi_{.j} / N$

On peut également vérifier que dans un tableau d'effectifs théoriques, tous les profils lignes sont égaux entre eux ainsi que tous les profils colonnes

Constatons cette dernière propriété en calculant les effectifs théoriques associés au tableau précédent.

Calculons tout d'abord les effectifs théoriques de chaque case grâce à la formule précédente

Par exemple pour la case (2,3), c'est à dire pour la 2^{ème} ligne et la troisième colonne on a : $24,72 = 103 \times 60/250$ et 24,72 est l'effectif théorique auquel il faudrait s'attendre dans le cas de l'indépendance des deux variables V_A et V_B

Tableau 23-4 des effectifs théoriques de la répartition des qualités de mémoires et des suivis

V_A suivi personnalisé :	Tableau de contingence Effectifs théoriques	Variable V_B « qualité de la production »			
		B_1 mémoire validé avec mention	B_2 mémoire validé sans mention	B_3 mémoire non validé	Totaux
A_1 suivi personnalisé		70,56	41,16	35,28	147
A_2 suivi non personnalisé		49,44	28,84	24,72	103
Totaux		120	70	60	250

On peut remarquer qu'il suffit de calculer seulement l'effectif théorique de 2 cases, car les autres chiffres se déduisent du calcul des marges (effectifs totaux du tableau), marges identiques à celles du tableau observé.

Construire le tableau des profils colonnes et celui des profils lignes sous l'hypothèse d'indépendance.

En reprenant le tableau d'effectifs théoriques, on trouve :

Tableau 23-5 des profils colonnes sous l'hypothèse d'indépendance

V_A suivi personnalisé :	Tableau de contingence	Variable V_B « qualité de la production »			
		B_1 mémoire validé avec mention	B_2 mémoire validé sans mention	B_3 mémoire non validé	Profil moyen
A_1 suivi personnalisé		58,2%	58,2%	58,2%	58,2%
A_2 suivi non personnalisé		41,8	41,8	41,8	41,8%
Totaux		100%	100%	100%	100%

Tableau 23-6 des profils lignes sous l'hypothèse d'indépendance

V _A suivi personnalisé :	Tableau de contingence	Variable V _B « qualité de la production »			
		B ₁ mémoire validé avec mention	B ₂ mémoire validé sans mention	B ₃ mémoire non validé	Totaux
A ₁ suivi personnalisé		48%	28%	24%	100%
A ₂ suivi non personnalisé		48%	28%	24%	100%
Profil moyen		48%	28%	24%	100%

23.4. Une mesure d'association : le χ^2

Afin de comparer les effectifs observés lors du sondage aux effectifs théoriques de l'hypothèse d'indépendance, on utilise un indice baptisé le χ^2 (à prononcer Khi-deux 22^{ème} lettre de l'alphabet grec) construit à partir des écarts entre le tableau croisé des effectifs observés que nous rebaptiserons O et le tableau croisé des effectifs théoriques que nous appellerons T

Pour chaque case (i, j) on calcule l'écart : $O_{ij} - T_{ij}$, on élève cet écart au carré, puis on le divise par l'effectif théorique T_{ij} . On calcule ensuite le χ^2 en sommant sur toutes les cases du tableau. La formule qui décrit l'opération précédente, s'écrit :

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

Cette expression établie par le statisticien Pearson, exprime l'importance de l'écart entre une distribution observée et une distribution théorique. Il faut associer à cette valeur de χ^2 un nombre de degrés de liberté qui dépend de la taille du tableau. Ce nombre noté par son acronyme ddl est calculé par : $ddl = (L-1) \times (C-1)$.

Ainsi dans notre exemple $L=2$; $C=3$, donc $ddl = (2-1) \times (3-1) = 2$. On peut remarquer que le chiffre 2 correspond exactement au nombre de cases pour lesquelles il a fallu calculer les effectifs théoriques, les autres se déduisant en référence aux marges du tableau.

Vérifier que $\chi^2 = 7,8614$ dans l'exemple que nous avons proposé sur le croisement des deux variables V_A et V_B

Des tables statistiques fournissent des valeurs critiques théoriques du χ^2 associées au nombre de degrés de liberté n'ayant que 10, 5, ou 1 chances sur 100 d'être dépassées par suite de fluctuations d'échantillonnage

23.4.1. Remarques et conditions d'utilisation :

- L'indice du χ^2 est toujours positif ou nul. S'il est nul, les deux variables V_A et V_B sont strictement indépendantes
- Les valeurs du χ^2 sont d'autant plus grandes que les écarts entre effectifs observés et effectifs théoriques sont grands
- L'usage du χ^2 pour tester l'hypothèse d'indépendance avec un risque contrôlé n'est pertinent que si tous les effectifs théoriques sont supérieurs ou égaux à 5. Dans le cas contraire, il faut procéder à des regroupements de modalités selon le sens des données. Cela tient au fait que les tables du Khi-deux fournissent une approximation de la véritable distribution de fréquence de la mesure d'association χ^2 et que lorsque les effectifs sont inférieurs à 5, cette approximation n'est plus pertinente. Si après regroupement maximum qui conduit au tableau 2x2, un au moins des effectifs théoriques demeure inférieur à 5, il faut alors avoir recours au test exact de Fisher.
- Si on multiplie tous les effectifs du tableau observé par un nombre k , le χ^2 calculé est alors multiplié par k . D'autres mesures d'association peuvent être définies qui neutralisent les effets de cette propriété qui rend incomparables deux tableaux de contingence construits à partir de deux échantillons de tailles différentes

23.5. Le test du χ^2 d'indépendance de deux variables qualitatives

23.5.1. La démarche du test du χ^2

Nous y repérons quatre grandes étapes comme dans la plupart des tests statistiques :

- Étape 1 : formulation d'hypothèses

Comme pour tout test statistique, deux hypothèses contradictoires sont à prendre en compte : l'hypothèse dite « nulle » notée H_0 et une hypothèse alternative dite « expérimentale » notée H_1

Par exemple, voici trois formulations équivalentes en ce qui concerne la propriété qui nous intéresse ici

H_0 : Il n'y a pas de différence entre la répartition des effectifs observés et la répartition des effectifs théoriques

H_1 : Il y a une différence entre la répartition des effectifs observés et la répartition des effectifs théoriques

Ou bien :

H_0 : Il n'y a pas de liaison entre les variables V_A et V_B

H_1 : Il y a une liaison entre les variables V_A et V_B

Ou encore :

H_0 : les variables V_A et V_B sont indépendantes

H_1 : les variables V_A et V_B sont dépendantes

Nous rappelons que, lorsque les hypothèses sont formulées, elles sont ensuite soumises à l'épreuve des faits pour décider laquelle est la plus vraisemblable. C'est l'hypothèse nulle qui est testée de manière privilégiée. La décision statistique consiste donc à conserver l'hypothèse nulle H_0 ou à la rejeter comme nous l'avons présenté dans la partie introductive aux tests statistiques. Dans la pratique on se donne α de l'ordre de 0,05 (5%) ou 0,01 (1%), et l'on calcule β le cas échéant. Le calcul de β sort du cadre de ce cours.

- Étape 2 : calcul du χ^2 et du nombre de degrés de liberté ddl

Le calcul du χ^2 est fait à partir des effectifs observés et des effectifs théoriques, comme il a été indiqué précédemment. C'est la statistique du test. On calcule également le nombre de degrés de liberté associé. Ici, dans l'exemple, rappelons que $\chi^2=7,8614$ et $ddl = 2$; On vérifie également que les effectifs théoriques sont bien supérieurs ou égaux à 5

- Étape 3 : lecture du χ^2 théorique pour un risque α donné

Dans la table statistique du χ^2 , sur la ligne $ddl=2$, on peut lire que pour un risque de première espèce α donné de 0,05 (5%), le χ^2 théorique ou lu, noté χ^2_t , est égal à 5,99. C'est à dire qu'il n'y a que 5 chances sur 100 pour que la statistique du χ^2 dépasse cette valeur de 5,99. Pour un risque de 0,01 (1%), on lit une valeur de 9,21

- Étape 4 : décision statistique

Si $\chi^2 > \chi^2_t$ on rejette H_0 avec un risque α

Si la valeur du χ^2 calculé est supérieure au χ^2 théorique, alors on rejette l'hypothèse nulle H_0 pour le risque donné. La différence entre les effectifs observés et les effectifs théoriques est trop grande, elle ne peut résulter du hasard ou des seules fluctuations d'échantillonnage. Au risque α près, les deux variables V_A et V_B sont statistiquement dépendantes ou liées, l'hypothèse H_1 est retenue. L'analyse des profils lignes ou colonnes ou bien l'analyse des effectifs théoriques et observés peuvent alors nous permettre d'interpréter le sens de ce lien qu'il faut se garder de considérer d'emblée comme un lien causal..

Si $\chi^2 < \chi^2_t$ on conserve H_0 avec un risque β inconnu

Dans le cas contraire, si la valeur du χ^2 calculé est inférieure ou égale au χ^2 théorique alors la différence entre les effectifs observés et les effectifs théoriques n'est pas significative, elle résulte des seules fluctuations d'échantillonnage. Il est possible, pour un risque β , dit de 2^{ème} espèce, inconnu, de retenir l'hypothèse H_0 .

Dans l'exemple que nous avons proposé sur le croisement des deux variables V_A et V_B quelle hypothèse conserveriez-vous avec quel risque et à quel niveau ?

Pour un risque α de 5%, $7,8614 > 5,99$ donc on peut rejeter l'hypothèse H_0 et admettre une dépendance entre les 2 caractères. L'analyse des effectifs théoriques et observés montre qu'un suivi personnalisé conduit plus souvent que la moyenne à une validation avec mention.

Pour un risque α de 1%, $7,8614 < 9,21$ donc on ne peut pas rejeter l'hypothèse H_0 , il convient d'accepter H_0 avec un risque inconnu β . On dit encore que les différences observées ne sont pas significatives

Remarquons, pour terminer, que le test du χ^2 n'a de sens que si l'on étudie un recueil de données à partir d'un échantillon. Dans le cas d'un recensement, il servira seulement comme mesure d'association entre variables. Il pourra servir, par exemple, à la recherche des caractères les plus liés à un caractère donné

23.6. Retour sur le tableau de contingence

Quand il s'agit de deux variables X et Y qui possèdent plus de deux modalités, nous avons alors à traiter un tableau à q colonnes et p lignes donc à pq cases.

	y ₁	...	y _j	...	y _q	effectifs totaux relatifs à X
x ₁	n ₁₁	...	n _{1j}	...	n _{1q}	n _{1.}
...
x _i	n _{i1}	...	n _{ij}	...	n _{iq}	n _{i.}
...
x _p	n _{p1}	...	n _{pj}	...	n _{pq}	n _{p.}
effectifs totaux relatifs à Y	n _{.1}	...	n _{.j}	...	n _{.q}	n = n _{..}

On calcule la valeur empirique d^2 de D^2 par la formule

$$d^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - \frac{n_i \cdot n_j}{n})^2}{\frac{n_i \cdot n_j}{n}}$$

avec n_{ij} = effectif observé relatif à la réalisation

simultanée de la modalité i de la première variable X et de la modalité j de la seconde variable Y.

23.6.1. Contribution d'une "case" à la valeur prise par $D^2 = \chi^2$:

On peut calculer les **contributions absolues** de chaque "case" :

$$CTA_{ij} = \frac{(n_{ij} - \frac{n_i \cdot n_j}{n})^2}{\frac{n_i \cdot n_j}{n}}$$

Puis les **contributions relatives** de chaque "case" :

$$CTR_{ij} = \frac{1}{d^2} \frac{(n_{ij} - \frac{n_i \cdot n_j}{n})^2}{\frac{n_i \cdot n_j}{n}}$$

Cette information peut être présentée sous forme d'un tableau analogue au tableau de contingence. Cela fait apparaître les croisements qui contribuent le plus à écarter le tableau observé du tableau sous l'hypothèse d'indépendance.

23.6.2. Mesures d'association dérivées de la valeur prise par $D^2 = \chi^2$:

On peut tout d'abord énoncer une propriété remarquable :

$$\frac{d^2}{n} \leq \inf \{p-1, q-1\}$$

où p est le nombre de lignes et q est le nombre de colonnes du tableau de contingence.

Il existe toute une série de coefficient proposés pour obtenir une mesure comprise entre 0 (dans le cas de l'indépendance) et 1 (dans le cas d'une liaison fonctionnelle).

23.6.2.1. coefficient de contingence de Karl Pearson :

$$C = \sqrt{\frac{d^2}{n+d^2}}$$

23.6.2.2. coefficient de Tschuprow :

$$T = \sqrt{\frac{d^2}{n\sqrt{(p-1)(q-1)}}$$

23.6.2.3. coefficient de Cramer :

$$V = \sqrt{\frac{d^2}{n \inf\{p-1, q-1\}}}$$

23.7. Autres utilisations de la mesure d'association χ^2

Il y a d'autres formes d'utilisation de la mesure du χ^2 en fonction des questions que l'on se pose. Tout d'abord, s'il s'agit de comparer des proportions relatives aux modalités d'une variable qualitative au sein de plusieurs populations, il suffit de se ramener au cas précédent en écrivant un tableau de contingence où les lignes correspondent aux divers échantillons et les colonnes représentent les modalités de la variable qualitative. **C'est le**

test d'homogénéité du χ^2 . La procédure mathématique et la démarche sont rigoureusement les mêmes que le test du χ^2 d'indépendance

S'il s'agit de comparer la distribution de fréquence d'une variable qualitative ou d'une variable quantitative, observée sur un échantillon, à une distribution théorique connue sur la population. **C'est le test du χ^2 d'adéquation.**

24. Test d'homogénéité

Un test d'homogénéité consiste à expliciter une règle de décision relative à une hypothèse portant sur une **comparaison**. Cette comparaison peut par exemple être réalisée sur des moyennes, des variances, des proportions.

24.1. Test d'homogénéité du Khi-deux

Nous nous trouvons dans cette situation lorsque l'on cherche à comparer des proportions relatives aux modalités d'une variable qualitative au sein de deux ou plusieurs populations. Les données recueillies peuvent alors être mises dans un tableau de contingence où les lignes correspondent aux diverses populations et les colonnes aux modalités de la variable qualitative.

L'hypothèse d'égalité des proportions d'apparition d'une modalité au sein de chaque population s'avère conduire à une propriété analogue à celle utilisée sous l'hypothèse d'indépendance des modalités de deux variables qualitatives. Il advient que la procédure utilisée pour ce test est la même que celle développées pour le test d'indépendance de deux variables (Test d'indépendance du Khi-deux)

variable qualitative X	X ₁	X ₂	X ₃	X ₄	effectifs totaux
Population A	a ₁	a ₂	a ₃	a ₄	a
Population B	b ₁	b ₂	b ₃	b ₄	b

l'hypothèse Ho d'homogénéité correspond à l'énoncé de la propriété suivante :

$$\frac{a_k}{a} = \frac{b_k}{b} \quad \text{pour } k = 1, 2, 3, 4$$

25. Test du χ^2 d'adéquation

S'il s'agit de comparer la distribution de fréquence d'une variable qualitative ou d'une variable quantitative, observée sur un échantillon, à une distribution théorique connue sur la population. C'est le **test du χ^2 d'adéquation** qui sera mis en pratique. Nous exposerons ce test à l'aide de deux exemples.

25.1. Présentation du premier exemple

Ayant relevé la répartition des diplômes dans un échantillon de 250 personnes représentatif (au sens du sondage aléatoire) de la population française, on cherche à la comparer à une répartition connue de ces mêmes diplômes dans la population entière. On se demande donc, si l'échantillon est représentatif (au sens du modèle réduit) relativement aux diplômes .

<i>Diplômes</i>	<i>Répartition dans la Population</i>	<i>Effectifs de l'échantillon</i>
Aucun ou Certificat d'études	34%	72
BEPC	8%	24
CAP-BEP	25%	65
Bac ou équivalent	11%	25
Diplôme supérieur au Bac	22%	64
	Total : 100%	Total : 250

25.2. La démarche du test du χ^2 d'adéquation

Nous suivons les quatre grandes étapes :

- Étape 1 : formulation d'hypothèses

Les deux hypothèses à prendre en compte sont :

H_0 : Il n'y a pas de différence entre la répartition des effectifs observés dans l'échantillon et la répartition des effectifs théoriques de la population.

H_1 : Il y a une différence entre la répartition des effectifs observés et la répartition des effectifs théoriques.

- Étape 2 : calcul du χ^2 et du nombre de degrés de liberté (ddl)

Le calcul du χ^2 est fait à partir des effectifs observés et des effectifs théoriques. On considère que le total des effectifs théoriques et le total des effectifs observés sont égaux. Ce χ^2 est la statistique du test. On calcule également le nombre de degrés de liberté associé. Ici, il est égal au nombre de modalités diminué de 1. On vérifie également que les effectifs théoriques sont bien supérieurs ou égaux à 5. Dans cet exemple, les effectifs théoriques sont calculés à partir des fréquences connues sur la population appliquées à un échantillon de 250 personnes. Ainsi l'effectif théorique de 85 personnes correspond au 34% de 250 etc..:

Diplômes	Répartition dans la Population	Effectifs théoriques attendus sur un échantillon de taille 250	Effectifs observés sur l'échantillon
Aucun ou Certificat d'études	34%	85	72
BEPC	8%	20	24
CAP-BEP	25%	62,5	65
Bac ou équivalent	11%	27,5	25
Diplôme supérieur au Bac	22%	55	64
	Total : 100%	Total 250	Total : 250

$$\chi^2 = \frac{(72 - 85)^2}{85} + \frac{(24 - 20)^2}{20} + \frac{(65 - 62,5)^2}{62,5} + \frac{(25 - 27,5)^2}{27,5} + \frac{(64 - 55)^2}{55} = 4,58 \text{ avec}$$

$$ddl=5 - 1= 4$$

- Étape 3 : lecture du χ^2 théorique pour un risque α donné

Dans la table statistique du χ^2 , sur la ligne $ddl=4$, on peut lire que pour un risque α donné de 0,05 (5%), le χ^2 théorique ou lu, noté χ^2_t , est égal à 9,49. C'est à dire qu'il n'y a que 5 chances sur 100 pour que la statistique du χ^2 dépasse cette valeur de 9,49.

- Étape 4 : décision statistique

La valeur du χ^2 calculé est inférieure au χ^2 théorique, alors la différence entre les effectifs observés et les effectifs théoriques n'est pas significative, elle résulte plutôt des seules fluctuations d'échantillonnage. Il est possible, avec un risque β inconnu de retenir l'hypothèse H_0 d'absence de différence entre la répartition des diplômes observés dans l'échantillon et des diplômes théoriques de la population.

25.3. Présentation du second exemple : test d'adéquation à une distribution de la Laplace-Gauss

Ce test est parfois appelé test de normalité en raison du fait que la variable de Laplace-Gauss est souvent nommée variable normale. Nous avons laissé de côté cette dénomination ambiguë.

Comparer la variable V3 « hauteur du saut. » à une variable de Laplace-Gauss en utilisant la forme du polygone des fréquences et les effectifs théoriques sous l'hypothèse V3 est une variable de Laplace-Gauss. Tester les hypothèses correspondantes (H_0 et H_1) au seuil de 5% avec le test du Khi-Deux d'adéquation.

Nous traitons la variable V3 comme une variable continue.

intervalles	[95;105[[105;110[[110;115[[115;120[[120;125[[125;135]	Totaux
effectifs	2	4	8	6	3	2	25
fréquences	0,08	0,16	0,32	0,24	0,12	0,08	1

Le détail des calculs permettant d'aboutir à ces résultats est fourni après

<i>paramètres</i>	<i>valeurs</i>	<i>valeurs approchées à utiliser</i>
étendue	40	40 cm
moyenne	114,5	114,5 cm
variance	55	55 cm ²
écart-type	7,41	7,4 cm
Q2 = médiane = second quartile	114,0625	114 cm

Il s'agit de comparer la variable V3 "hauteur du saut.." à une variable de Laplace-Gauss en utilisant la forme du polygone des fréquences et les effectifs théoriques sous l'hypothèse V3 est une variable de Laplace-Gauss. Le polygone des fréquences (lissage de la densité de fréquences) suggère une forme proche de la courbe de Laplace-Gauss. La hauteur médiane:114 cm, la hauteur modale : 112,5 cm et la hauteur moyenne 114,5 cm confirme par leur proximité cette remarque.

Nous allons procéder à un test d'hypothèse à l'aide du test du Khi-deux d'adéquation.

- Étape 1 : formulation d'hypothèses

L'hypothèse H_0 est alors formulée ainsi :

La variable V3 « hauteur du saut » est une variable de Laplace-Gauss $LG(\mu, \sigma)$ de paramètres estimés $m = 114,5$ et $s = 7,5$ "

L'hypothèse alternative H_1 peut être énoncée comme suit :

La variable V3 « hauteur du saut » n'est pas une variable de Laplace-Gauss

- Étape 2 : calcul du χ^2 et du nombre de degrés de liberté (ddl)

La variable de décision est la variable $D = \chi^2 = \sum_{i=1}^{i=s} \frac{(O_i - T_i)^2}{T_i}$ où O_i = effectif observé

et T_i = effectif théorique sous la contrainte de la loi de référence choisie pour chacune des s classes.

Si aucun des effectifs théoriques n'est inférieur à 5, la variable D suit approximativement une loi du Khi-Deux à $s-1-e$ degrés de liberté (où e désigne le nombre de paramètres estimés parmi les r paramètres ici $e=r-2$ ce sont μ et σ).

Si un des effectifs est inférieur à 5, on procède à des regroupements de classes pour se mettre dans les conditions d'application.

On calcule les effectifs théoriques sous l'hypothèse H_0 puis la valeur de la variable D critère de décision

Les intervalles de la variable centrée réduite $Z = \frac{V3-114,5}{7,5}$	effectifs théoriques	effectifs théoriques après regroupement	effectifs observés après regroupement	effectifs observés	$\frac{(O_i - T_i)^2}{T_i}$
$]-\infty ; -2,6[$	0,12			0	
$[-2,6 ; -1,27[$	2,43	6,86	6	2	0,1078
$[-1,27 ; -0,6[$	4,31			4	
$[-0,6 ; 0,07[$	6,34	6,34	8	8	0,4346
$[0,07 ; 0,73[$	5,98	5,98	6	6	0,00006
$[0,73 ; 1,4[$	3,8			3	
$[1,4 ; 2,73[$	1,94	5,82	5	2	0,1155
$[2,73 ; +\infty[$	0,08			0	
Totaux	25	25	25	25	0,65796

Ainsi la valeur expérimentale de la variable de décision est $d = \chi^2 = 0,65796$

Le calcul du degré de liberté donne $ddl = 4 - 2 - 1 = 1$

- Étape 3 : lecture du χ^2 théorique pour un risque α donné

Choisir la valeur α du risque de première espèce. Rechercher dans une table du Khi-Deux à $s-1-e$ degrés de liberté, la valeur critique κ telle que $\text{Prob}\{\chi^2_{(s-1-e)} > \kappa\} = \alpha$.

Au seuil de risque de 0,05 nous pouvons lire la valeur critique $k = 3,84$

- Étape 4 : décision statistique

Ayant calculé d la valeur empirique de D avec les s valeurs O_i observées sur l'échantillon extrait.

Si $d < \kappa$, c'est à dire si $d \in A = [0, \kappa]$, région d'acceptation, alors on ne rejette pas l'hypothèse nulle H_0 , on accepte alors H_0 en prenant un risque β d'erreur de seconde espèce

Si $d > \kappa$, c'est à dire si $d \in K = [\kappa, +\infty[$, région critique, alors on rejette l'hypothèse nulle H_0 et on accepte H_1 .

La valeur expérimentale $d^2 \approx 0,657$ obtenue sur l'échantillon est inférieure à $k = 3,84$. Par conséquent, nous ne sommes pas dans les conditions de rejet de l'hypothèse H_0 . Il nous est possible d'affirmer qu'il est raisonnable de tenir la variable $V3$ pour une variable de Laplace-Gauss $LG(114,5 ; 7,5)$. Toutefois cette affirmation nous place dans la situation d'une prise de risque de seconde espèce de valeur β inconnue.

26. Test du χ^2 de Mac Nemar

Si l'on cherche à comparer deux séries successives de données de type présence-absence ou échec-réussite relevées sur le même échantillon, on pourra utiliser le test du χ^2 de Mac Nemar

Si l'on veut comparer la difficulté de deux épreuves scolaires passées par le même groupe d'individus (séries dites appariées), on peut résumer le problème par l'étude d'un tableau croisé simple à 2 lignes et 2 colonnes, à condition de ne s'intéresser qu'à l'échec ou à la réussite.

Mac Nemar a montré que le test du χ^2 d'indépendance n'est pas approprié car il ne faut prendre en compte que les cas de discordances entre les épreuves, c'est à dire le cas de réussite à l'une et d'échec à l'autre et son complémentaire. Cela revient à savoir si ces deux nombres sont ou non égaux.

Prenons un exemple :

50 élèves passent deux épreuves A et B. Le tableau suivant résume l'ensemble des résultats :

Épreuve A:	Épreuve B :	Réussite	Échec	Total
Réussite		16	12	28
Échec		10	12	22
Total		26	24	50

Les fréquences de réussite aux deux épreuves sont-elles différentes significativement ?

Nous présentons dans le tableau ci-dessous les notations de ce problème :

Épreuve 2:	Épreuve 1 :	Réussite	Échec	Total
Réussite		N1	N2	N1 + N2
Échec		N3	N4	N3 + N4
Total		N1 + N3	N2 + N4	N1 + N2 + N3 + N4

Il s'agit de comparer N2 et N3, ou bien de comparer $N2 / (N2 + N3)$ à $1/2$ dans l'hypothèse d'une équivalence entre les épreuves. On compare une fréquence observée à une fréquence théorique de $1/2$.

Mac Nemar a montré qu'il suffisait de calculer la valeur du χ^2 suivante :

$$\chi^2_{\text{(Mac Nemar)}} = \frac{(N2 - N3)^2}{N2 + N3}$$

Cette valeur ne peut se calculer que pour un tableau de 4 cases. De plus le test requiert également que $N2 + N3$ soit au moins égale à 10

Les 4 étapes de la démarche du test sont identiques aux précédentes :

- Étape 1 : formulation d'hypothèses

H_0 : égalité des changements d'état entre les deux épreuves

H_1 : Non-égalité des changements d'état entre les deux épreuves

- Étape 2 : calcul du $\chi^2_{(Mac\ Nemar)}$ et du nombre de degrés de liberté (ddl)

Le calcul du $\chi^2_{(Mac\ Nemar)}$ donne : $\chi^2_{(Mac\ Nemar)} = (12-10)^2 / (12 + 10) = 4 / 22 = 0,18$

Quant au nombre de degrés de liberté (ddl) il est donc toujours de 1 dans le cas de cette statistique..

- Étape 3 : lecture du χ^2 théorique pour un risque α donné

La lecture du χ^2_t dans la table du χ^2 pour un risque α donné, sur la ligne ddl=1 permet de conclure. Ainsi, si l'on prend un risque α de 5%, la valeur théorique au-delà de laquelle l'hypothèse H_0 sera rejetée est 3,84. Pour un risque α de 1%, elle vaut 6,63.

- Étape 4 : décision statistique

Comme $0,18 < 3,84$ Nous ne rejetons pas l'hypothèse nulle . Les deux épreuves sont équivalentes pour un risque de 2^{ème} espèce β inconnu.

27. Étude simultanée de deux variables quantitatives

L'étude de la corrélation a pour objet la recherche d'une liaison entre deux variables quantitatives à partir de l'information fournie par le coefficient de corrélation ρ

27.1. Liaison entre deux variables quantitatives X et Y : Coefficient de corrélation de Bravais-Pearson

Tableau de contingence (pxq) issu du traitement des observations conjointes de X et de Y. Mise en place des paramètres utiles au calcul.

y \ x	y ₁	y ₂	y _j	y _q
x ₁						
x ₂						
...						
...						
x _i				n _{ij}		n _{i.}
...						
x _p						

$n_{i.} = \sum_{j=1}^q n_{ij}$ (effectif marginal de la valeur x_i)

$n_{.j} = \sum_{i=1}^p n_{ij}$ (effectif marginal de la valeur y_j)

$$N = \sum_{i=1}^p n_{i.} = \sum_{j=1}^q n_{.j} = \sum_{i=1}^p \sum_{j=1}^q n_{ij} \text{ effectif total.}$$

variable X	variable Y
estimation de l'espérance $\bar{X} = \frac{1}{N} \sum_{i=1}^{i=p} n_{i.} x_i = \sum_{i=1}^{i=p} f_{i.} x_i$	estimation de l'espérance $\bar{Y} = \frac{1}{N} \sum_{j=1}^{j=q} n_{.j} y_j = \sum_{j=1}^{j=q} f_{.j} y_j$
estimation de la variance $V(X) = \frac{1}{N-1} \sum_{i=1}^{i=p} n_{i.} (x_i - \bar{X})^2 = \frac{N}{N-1} \left[\left(\frac{1}{N} \sum_{i=1}^{i=p} n_{i.} x_i^2 \right) - \bar{X}^2 \right]$	estimation de la variance $V(Y) = \frac{1}{N-1} \sum_{j=1}^{j=q} n_{.j} (y_j - \bar{Y})^2 = \frac{N}{N-1} \left[\left(\frac{1}{N} \sum_{j=1}^{j=q} n_{.j} y_j^2 \right) - \bar{Y}^2 \right]$

covariance de X et Y	coefficient de corrélation
$\text{COV}(X,Y) = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} (x_i - \bar{X})(y_j - \bar{Y})$	$R_{BP} = \frac{\text{Cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$

27.1.1. Caractère significatif du coefficient de corrélation de Bravais-Pearson :

Avant toute étude plus approfondie, il convient de réaliser une représentation graphique du nuage statistique des N points de coordonnées (x_i, y_j) pour $i = 1$ à p et $j = 1$ à q . La forme de ce nuage orientera l'analyse.

Supposons que les N observations proviennent d'une population dans laquelle les deux variables X et Y sont indépendantes. Dans ce cas la valeur réelle du coefficient de corrélation est $\rho = 0$.

On peut alors s'interroger sur la distribution de probabilité de la variable R_{BP} correspondant à cet échantillonnage.

Il est établi que si $\rho = 0$ et si le couple (X,Y) est un couple de variables de Laplace - Gauss, la distribution de probabilité est obtenue de la façon suivante :

la variable $\frac{R_{BP} \sqrt{N-2}}{\sqrt{1-R_{BP}^2}}$ suit une loi de Student de ddl = N-2

Par ailleurs l'espérance de R_{BP} vaut $E(R_{BP}) = 0$ et la variance $V(R_{BP}) = \frac{1}{n-1}$.

Dans le cas général où r est quelconque dans $[-1 ; +1]$, on peut utiliser la transformée de Fisher :

$Z = \frac{1}{2} \ln \left(\frac{1+R_{BP}}{1-R_{BP}} \right)$ tend (en loi) vers LG $\left(\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right); \frac{1}{\sqrt{N-3}} \right)$ quand N tend vers l'infini

Cette transformation permet traiter le cas général, y compris lorsque le couple (X,Y) n'est pas une variable de Laplace-Gauss de dimension 2, dès que N est grand ($N > 30$). Cependant le fait de ne pas rejeter l'hypothèse selon laquelle le coefficient de corrélation est nul, n'entraîne pas nécessairement l'indépendance des deux variables. Il n'y a qu'une présomption d'indépendance. La nullité de ρ est une condition nécessaire mais pas suffisante pour l'indépendance. En d'autres mots l'absence de corrélation n'implique pas l'indépendance.

STATISTIQUE de RANG

En Sciences humaines, nous pouvons déterminer une classe importante de phénomènes dont l'étude peut mettre en jeu le rang d'une observation (valeur ou modalité) dans l'échantillon. Le rang d'une observation est le nombre d'observations qui lui sont "*inférieures ou égales*" ou qui sont *placées avant*, au sens de la relation d'ordre définie sur l'espace des observations possibles.

Nous pouvons formaliser cette définition par la relation suivante :

soit l'échantillon de la variable $X : (X_1, X_2, \dots, X_k, \dots, X_n)$

le rang de X_i est $R_i = \sum_{k=1}^{k=n} \mathbf{1}_{R^+}(X_i - X_k)$ avec $\mathbf{1}_{R^+}(X_i - X_k) = 1$ si X_k est "*inférieure ou égale*" à X_i et $\mathbf{1}_{R^+}(X_i - X_k) = 0$ sinon².

La statistique de rang R est le vecteur statistique des rangs $R = (R_1, R_2, \dots, R_k, \dots, R_n)$. L'ensemble des réalisations r possibles de ce vecteur statistique est l'ensemble p_n des $n!$ permutations des nombres entiers de 1 à n . La statistique R est uniformément distribuée sur p_n :

$$\text{Prob}(\{R=r\}) = \frac{1}{n!}$$

De là on peut déduire quelques résultats utiles :

- la probabilité de l'événement «le rang R_i est k » est

$$\text{Prob}(\{R_i = k\}) = \sum_{r \in p_n} \text{Prob}(\{R=r\} \text{ et } \{R_i = k\}) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

- la probabilité de l'événement «le rang R_p est k et le rang R_q est l » est

$$\text{Prob}(\{R_p = k\} \text{ et } \{R_q = l\}) = \sum_{r \in p_n} \text{Prob}(\{R=r\} \text{ et } \{R_p = k\} \text{ et } \{R_q = l\}) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$$

- l'espérance mathématique de la variable R_p est :

$$E(R_p) = \frac{1}{n} \sum_{k=1}^{k=n} k = \frac{n+1}{2}$$

- la variance de la variable R_p est :

$$V(R_p) = E(R_p^2) - E(R_p)^2 = \frac{(n^2-1)}{12}$$

- la covariance entre la variable R_p et la variable R_q est :

$$\text{cov}(R_p, R_q) = E(R_p R_q) - E(R_p)E(R_q) = -\frac{n+1}{12} \quad \text{avec } p \neq q$$

Les variables S_n utilisées pour traiter les informations fournies par les rangs sont les R-estimateurs dont l'expression peut être formulée ainsi :

$$S_n = \sum_{k=1}^{k=n} c_k f_n(R_k)$$

où n est le nombre total d'observations issues de l'échantillon de la variable X : $(X_1, X_2, \dots, X_k, \dots, X_n)$, R_k est le rang de X_k dans l'échantillon ordonné, f_n est une fonction des nombres entiers caractérisant les rangs, dont le résultats est appelés parfois "scores" et c_k sont des constantes, appelées constantes de régression.

² Nous utilisons la notation de la fonction indicatrice sur l'ensemble des nombres réels non négatifs même pour des variables qualitatives ordonnées par la relation d'ordre "*avant ou à la même place*".

28. Test de l'hypothèse d'un échantillon aléatoire. (Problème à un échantillon)

L'hypothèse nulle H_0 traduit l'idée selon laquelle : "l'ordre dans lequel on effectue les observations n'a pas d'importance"

Une des hypothèses alternatives H_1 traduit une idée contraire selon laquelle : "l'ordre dans lequel on effectue les observations a une importance"

Pour résoudre ce problème, nous proposons trois procédures fondées chacune sur une statistique différente :

- le coefficient de corrélation de rang R_s de Spearman

C. Spearman, " The proof and measurement of association between two things " *American Journal of Psychology*, n°15, janvier 1904, pp 72-101.

- le coefficient de corrélation de rang τ de Kendall

M.G. Kendall, " A new measure of rank correlation " , *Biometrika* n°30 -1938-pp 81-93

M.G. Kendall, " The treatment of ties in ranking problems" *Biometrika* n°33 -1943/1946-pp 239-251

G.P. Sillitto, " The distribution of Kendall's τ coefficient of rank correlation in rankings containig ties ",

Biometrika n°34 -1947 -pp 36-40 + tables

M.G. Kendall, " The variance of τ when both rankings contain ties "

Biometrika n°34 -1947 -pp 297 -298

Wassily Höfdding, " On the distribution of the rank correlation coefficient τ when the variates are not independent "

Biometrika n°34 -1947 -pp 183 -196

- le nombre de différences positives (nombre de signes +)

28.1. Le coefficient de corrélation des rangs R_S de Spearman

A chaque observation, nous associons son rang k dans l'ordre du recueil des observations et son rang R_k dans l'ordre de l'échantillon.

Rangs des observations	1	2	...	k	...	n
Rangs dans l'échantillon ordonné	R_1	R_2	...	R_k	...	R_n

La statistique est

$$S_n = R_S = 1 - \frac{\sum_{k=1}^{k=n} [R_k - k]^2}{n(n^2 - 1)}$$

l'espérance de R_S est $E(R_S) = 0$

la variance de R_S est $V(R_S) = \frac{1}{n-1}$

Pour tester la significativité de la valeur obtenue, on prend en référence la situation théorique d'indépendance des deux rangements dans la population, c'est à dire que les $n!$ classements sont équiprobables. Dans ce cas de l'indépendance, la valeur est $\rho_S = 0$. Cependant il convient de rappeler que la réciproque est fautive.

Dans le cas d'une tendance monotone croissante parfaite, les classements sont identiques

$$R_k = k$$

$\rho_S = 1$: les deux classements sont identiques

Dans le cas d'une tendance monotone décroissante parfaite, les classements sont inversés $R_k = n+1-k$

$\rho_S = -1$: les deux classements sont inverses

Il s'agit alors de prendre une décision sur la base d'une valeur r_S , réalisation de la variable R_S sur un échantillon. A l'aide de la table du coefficient de corrélation de Spearman ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'indépendance des deux rangements. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1 tendance monotone croissante dépendance positive	H1 tendance monotone dépendance quelconque	H1 tendance monotone décroissante dépendance négative
$W = \{ r_S, r_S > c \}$	$W = \{ r_S, r_S > c \}$	$W = \{ r_S, r_S < -c \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

Asymptotiquement la variable $Z_S = R_S \sqrt{n-1}$ suit la loi de la variable de Laplace-Gauss LG(0,1). Cette approximation est jugée acceptable pour $n > 30$.

Asymptotiquement la variable $T_S = R_S \sqrt{\frac{n-2}{1-R_S^2}}$ suit la loi de la variable de Student à $n-2$ ddl. Cette approximation est jugée acceptable pour $n > 10$.

28.2. Le coefficient de corrélation de rang τ de Kendall

Ce test est fondé sur les procédures suivantes qui déterminent la variable τ de

Kendall: $\tau = \frac{4A}{n(n-1)} - 1 = \frac{2S}{n(n-1)}$ où $A = \sum_{p=1}^{p=n-1} \sum_{q=p+1}^{q=n} \mathbf{1}_{R_+^*(X_q - X_p)}$ est la variable qui

donne le nombre de paires respectant l'ordre des observations.

S = le nombre de paires respectant l'ordre - le nombre de paires inversant l'ordre

$$S = \sum_{p=1}^{p=n-1} \sum_{q=p+1}^{q=n} \mathbf{1}_{R_+^*(X_q - X_p)} - \sum_{p=1}^{p=n-1} \sum_{q=p+1}^{q=n} \mathbf{1}_{R_+^*(X_p - X_q)}$$

- l'espérance de la variable τ est $E(\tau) = 0$

- la variance de la variable τ est $V(\tau) = \frac{2(2n+5)}{9n(n-1)}$

28.2.1. Méthodes de calcul:

méthode n°1

- dénombrer parmi les paires $\{p, q\}$ telles que $p < q$, le nombre a_i de paires telles que $x_p < x_q$,

- calculer $A = \sum_{i=1}^{i=n} a_i$

$$\text{le coefficient de Kendall } \tau = \frac{4A}{n(n-1)} - 1$$

méthode n°2

- considérer les $\frac{n(n-1)}{2}$ paires d'observations,

- à chaque paire $\{p; q\}$:

on associe 1 si les deux classements sont en accord $p < q$ et $x_p < x_q$

on associe -1 si les deux classements sont en désaccord $p < q$ mais $x_q < x_p$

- calculer la somme $S = \sum_{k=1}^{n(n-1)} \delta_k$ avec $\delta_k = 1$ ou -1

$$\text{le coefficient de Kendall } \tau = \frac{2S}{n(n-1)}$$

Pour tester la significativité de la valeur obtenue, on prend en référence la situation théorique d'indépendance des deux rangements dans la population, c'est à dire que les $n!$ classements sont équiprobables. Dans ce cas la valeur exacte est $\tau = 0$.

Dans le cas d'une tendance monotone croissante exacte, les classements sont identiques $R_k = k$ $\tau = 1$: les deux classements sont identiques

Dans le cas d'une tendance monotone décroissante exacte, les classements sont inversés $R_k = n+1-k$ $\tau = -1$: les deux classements sont inverses

Il s'agit alors de prendre une décision sur la base d'une valeur τ , réalisation de la variable τ sur un échantillon. A l'aide de la table du coefficient de corrélation de Kendall ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'indépendance des deux rangements. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1	H1	H1
tendance monotone croissante dépendance positive	tendance monotone dépendance quelconque	tendance monotone décroissante dépendance négative
$W = \{ \tau, \tau > c \}$	$W = \{ \tau, \tau > c \}$	$W = \{ \tau, \tau < -c \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

Asymptotiquement la variable $Z_K = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}$ suit la loi de la variable de Laplace-

Gauss $LG(0,1)$. Cette approximation est jugée acceptable pour $n > 7$

28.3. Comparaison des coefficients de Spearman et de Kendall

Sous l'hypothèse H_0 , le coefficient de corrélation entre les deux variables R_s et τ

vaut : $\text{Corr}(R_s, \tau) = \frac{2(n+1)}{\sqrt{2n(2n+5)}}$ qui tend vers 1 quand n tend vers l'infini. Par ailleurs

$V(\tau - \frac{2}{3} R_s) = E[(\tau - \frac{2}{3} R_s)^2] = \frac{2}{9n(n-1)}$ qui indique que ces deux coefficients de corrélation de rang sont approximativement dans un rapport $\frac{2}{3}$ ou $\frac{3}{2}$ puisque la variance de la variable $\tau - \frac{2}{3} R_s$ tend vers 0 quand n tend vers l'infini.

28.4. Le test des "signes"

A chaque observation, nous associons son rang k dans l'ordre du recueil des observations. Sous l'hypothèse H_0 , la $k^{\text{ème}}$ observation a autant de chance d'être rangée avant la $(k+1)^{\text{ème}}$ qu'après., au sens de la relation d'ordre permettant de ranger les résultats de la variable ordinale étudiée.

Considérons l'échantillon $X_1, X_2, \dots, X_k, \dots, X_n$. de la variable X et la variable de Bernoulli Z_k définie pour $k=1$ à $n-1$ par les règles suivantes : $Z_k = 0$ si X_k est avant X_{k+1} , $Z_k=1$ sinon. Sous H_0 , $\text{Prob}\{Z_k=0\} = \text{Prob}\{Z_k=1\} = \frac{1}{2}$. et $Z_k = B(1; \frac{1}{2})$

S'il existe une tendance monotone croissante exacte, pour tout $k=1$ à $n-1$, $Z_k=0$ et s'il existe une tendance monotone décroissante exacte, pour tout $k=1$ à $n-1$, $Z_k=1$.

En considérant le tableau des rangs, nous pouvons écrire :

A chaque observation, nous associons son rang k dans l'ordre du recueil des observations et son rang R_k dans l'ordre de l'échantillon.

Rangs des observations	1	2	...	k	...	n
Rangs dans l'échantillon ordonné	R_1	R_2	...	R_k	...	R_n

Et la statistique mise en œuvre est :

$$S_n = \sum_{k=1}^{k=n-1} Z_k = \sum_{k=1}^{k=n-1} \mathbf{1}_{R_+}(R_k - R_{k+1})$$

L'espérance de cette variable est : $E(S_n) = \sum_{k=1}^{k=n-1} E(Z_k)$

$$E(S_n) = \frac{n-1}{2}$$

La variance de cette variable est :

$$V(S_n) = \sum_{k=1}^{k=n-1} V(Z_k) + \sum_{p, q} \text{COV}(Z_p, Z_q) = \frac{n-1}{4} + \sum_{p, q} \text{COV}(Z_p, Z_q)$$

$$V(S_n) = \frac{n-1}{4} + 2 \sum_{p < q} \text{COV}(Z_p, Z_q)$$

$$\text{avec } \text{COV}(Z_p, Z_q) = E(Z_p Z_q) - E(Z_p) E(Z_q) = \text{Prob}\{Z_p Z_q = 1\} - \frac{1}{4}$$

Calculons $\text{Prob}\{Z_p Z_q = 1\}$:

$$\text{Prob}\{Z_p Z_q = 1\} = \text{Prob}(\{Z_p = 1\} \text{ et } \{Z_q = 1\})$$

Si $p+1 < q$ alors $\text{Prob}\{Z_p Z_q = 1\} =$

$$\text{Prob}(\{R_p > R_{p+1}\} \text{ et } \{R_q > R_{q+1}\}) = \text{Prob}\{R_p > R_{p+1}\} \text{Prob}\{R_q > R_{q+1}\} = \frac{1}{4}$$

Si $p+1 = q$ alors $\text{Prob}\{Z_p Z_q = 1\} =$

$$\text{Prob}(\{R_p > R_{p+1} = R_q > R_{q+1}\}) = \frac{1}{3!} = \frac{1}{6}$$

$$\text{Donc } \text{COV}(Z_p, Z_q) = \left(\frac{1}{6} - \frac{1}{4}\right) \delta_p^{q+1} \quad (\delta_p^{q+1} \text{ symbole de Kronecker})$$

D'où

$$V(S_n) = \frac{n-1}{4} + 2 \sum_{p < q} \text{COV}(Z_p, Z_q) = \frac{n-1}{4} + 2 \sum_{p=1}^{p=n-2} \text{COV}(Z_p, Z_{p+1})$$

$$V(S_n) = \frac{n-1}{4} + 2(n-2)\left(\frac{-1}{12}\right) = \frac{n+1}{12}$$

$$\boxed{V(S_n) = \frac{n+1}{12}}$$

Il s'agit alors de prendre une décision sur la base d'une valeur s_n , réalisation de la variable S_n sur un échantillon.

Soit à l'aide d'une table donnant la distribution de probabilité de $S_n - E(S_n)$ établie pour des valeurs de $n < 13$ par Moore et Wallis à partir de la formule de récurrence suivante de P.A. Mac Mahon :

$$P_n(p) = (p+1)P_{n-1}(p) + (n-p)P_{n-1}(p-1)$$

où $P_n(p)$ = nombre de permutations de (R_1, R_2, \dots, R_n) comportant p inversions.

La valeur critique au seuil α s'obtient alors par la résolution de l'équation :

$$\text{Prob}\{S_n < c_1\} = \alpha \text{ ou } \text{Prob}\{S_n > c_2\} = \alpha \text{ ou } \text{Prob}(\{S_n < c_1\} \text{ ou } \{S_n > c_2\}) = \alpha$$

Soit à l'aide d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, puisque

$$\frac{S_n - E(S_n)}{\sigma(S_n)} = \frac{S_n - \frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}} \text{ tend en loi LG}(0,1) \text{ quand } n \text{ tend vers l'infini}$$

Dans ce cas, il convient de tenir du passage d'une variable discrète à une variable continue en réalisant une correction (*correction de continuité*) selon la démarche suivant:

$$\text{Prob} \{S_n = s\} = \text{Prob} \left\{ s - \frac{1}{2} < S_n < s + \frac{1}{2} \right\}$$

La valeur critique au seuil α s'obtient alors par la résolution de l'équation :

$$\text{Prob} \left\{ \frac{s - \frac{n-1}{2} - \frac{1}{2}}{\sqrt{\frac{n+1}{12}}} < \frac{S_n - \frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}} < \frac{s - \frac{n-1}{2} + \frac{1}{2}}{\sqrt{\frac{n+1}{12}}} \right\} = 1 - \alpha$$

$$\text{Prob} \left\{ \frac{s - \frac{n-1}{2} - \frac{1}{2}}{\sqrt{\frac{n+1}{12}}} < \text{LG}(0;1) < \frac{s - \frac{n-1}{2} + \frac{1}{2}}{\sqrt{\frac{n+1}{12}}} \right\} = 1 - \alpha$$

On peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'indépendance des deux rangements. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1	H1	H1
tendance monotone croissante dépendance positive	tendance monotone dépendance quelconque	tendance monotone décroissante dépendance négative
$W = \{ S, S < c_1 \}$	$W = \{ S, S < c_1 \text{ ou } S > c_2 \}$	$W = \{ S, S > c_2 \}$

On rappelle que la valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$ ou $\text{Prob}(W^c|H_0) = 1 - \alpha$.

29. Test d'indépendance (Problème à deux échantillons)

Cette fois nous disposons d'un échantillon du couple de variables (X,Y):

$((X_1, Y_1), (X_2, Y_2), \dots, (X_k, Y_k), \dots, (X_n, Y_n))$ auquel on associe les couples de rangs
 $((R_1, Q_1), (R_2, Q_2), \dots, (R_k, Q_k), \dots, (R_n, Q_n))$

L'hypothèse nulle H_0 traduit l'idée selon laquelle :

"les deux rangements ont été effectués indépendamment l'un de l'autre"

Une des hypothèses alternatives H_1 traduit une idée contraire selon laquelle :

"il existe une dépendance entre les deux rangements"

Pour résoudre ce problème, nous proposons deux procédures fondées chacune sur une statistique différente :

- le coefficient de corrélation de rang R_S de Spearman
- le coefficient de corrélation de rang τ de Kendall

29.1. Le coefficient de corrélation des rangs R_S de Spearman

A chaque couple d'observations, nous associons son rang k dans l'ordre du recueil des observations et le couple de rangs (R_k, Q_k) dans l'ordre de l'échantillon.

Rangs des observations	1	2	...	k	...	n
Rangs dans l'échantillon ordonné de X	R_1	R_2	...	R_k	...	R_n
Rangs dans l'échantillon ordonné Y	Q_1	Q_2	...	Q_k	...	Q_n

La statistique est

$$S_n = R_S = 1 - \frac{\sum_{k=1}^{k=n} [R_k - Q_k]^2}{n(n^2 - 1)}$$

l'espérance de R_S est $E(R_S) = 0$

la variance de R_S est $V(R_S) = \frac{1}{n-1}$

Pour tester la significativité de la valeur obtenue, on prend en référence la situation théorique d'indépendance des deux rangements dans la population, c'est à dire que les $n!$ classements sont équiprobables. Dans ce cas de l'indépendance, la valeur est $\rho_S = 0$. Cependant il convient de rappeler que la réciproque est fautive.

Dans le cas d'une tendance monotone croissante parfaite, les classements sont identiques, $R_k = k$, $\rho_S = 1$: les deux classements sont identiques

Dans le cas d'une tendance monotone décroissante parfaite, les classements sont inversés, $R_k = n+1-k$, $\rho_s = -1$: les deux classements sont inverses

Il s'agit alors de prendre une décision sur la base d'une valeur r_s , réalisation de la variable R_s sur un échantillon. A l'aide de la table du coefficient de corrélation de Spearman ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'indépendance des deux rangements. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1	H1	H1
tendance monotone croissante dépendance positive	tendance monotone dépendance quelconque	tendance monotone décroissante dépendance négative
$W = \{ r_s, r_s > c \}$	$W = \{ r_s, r_s > c \}$	$W = \{ r_s, r_s < -c \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

Asymptotiquement la variable $Z_s = R_s \sqrt{n-1}$ suit la loi de la variable de Laplace-Gauss $LG(0,1)$. Cette approximation est jugée acceptable pour $n > 30$.

Asymptotiquement la variable $T_s = R_s \sqrt{\frac{n-2}{1-R_s^2}}$ suit la loi de la variable de Student à $n-2$ ddl. Cette approximation est jugée acceptable pour $n > 10$.

29.1.1. Cas d'ex æquo :

La démarche consiste pour conserver la somme des rangs égale à $\frac{n(n+1)}{2}$ à attribuer la moyenne des rangs que les observations auraient occupés si elles avaient été distinctes.

Ainsi supposons qu'il y ait un groupe de q observations *ex æquo* aux rangs :

$p+1, p+2, \dots, p+q$, on leur attribue le même rang $p + \frac{q+1}{2}$

De là une expression modifiée du coefficient R_s de Spearman est à utiliser :

$$R_s^* = \frac{R^{*2} + Q^{*2} - \sum_{k=1}^{k=n} (R_k - Q_k)^2}{2R^*Q^*}$$

$$\text{avec } R^* = \frac{n(n^2-1)}{12} - \sum_{q \in E_X} \frac{q(q^2-1)}{12} \text{ et } Q^* = \frac{n(n^2-1)}{12} - \sum_{q \in E_Y} \frac{q(q^2-1)}{12}$$

E_X est l'ensemble des groupes d'*ex æquo* dans l'échantillon de X

E_Y est l'ensemble des groupes d'*ex æquo* dans l'échantillon de Y

29.2. Le coefficient de corrélation de rang τ de Kendall

A chaque couple d'observations, nous associons son rang k dans l'ordre du recueil des observations et le couple de rangs (R_k, Q_k) dans l'ordre de l'échantillon.

Rangs des observations	1	2	...	k	...	n
Rangs dans l'échantillon ordonné de X	R_1	R_2	...	R_k	...	R_n
Rangs dans l'échantillon ordonné Y	Q_1	Q_2	...	Q_k	...	Q_n

Puis nous réorganisons les résultats en ordonnant les couples selon le rang croissant dans l'échantillon de X.

Rangs dans l'échantillon ré-ordonné de X	1	2	...	k	...	n
Rangs dans l'échantillon ordonné Y	T_1	T_2	...	T_k	...	T_n

Ce test est fondé sur les procédures suivantes qui déterminent la variable τ de

Kendall: $\tau = \frac{4A}{n(n-1)} - 1 = \frac{2S}{n(n-1)}$ où $A = \sum_{p=1}^{p=n-1} \sum_{q=p+1}^{q=n} \mathbf{1}_{R+^*}(T_q - T_p)$ est la variable qui donne

le nombre de paires respectant l'ordre des observations.

S = le nombre de paires respectant l'ordre - le nombre de paires inversant l'ordre

$$S = \sum_{p=1}^{p=n-1} \sum_{q=p+1}^{q=n} \mathbf{1}_{R+^*}(T_q - T_p) - \sum_{p=1}^{p=n-1} \sum_{q=p+1}^{q=n} \mathbf{1}_{R+^*}(T_p - T_q)$$

- l'espérance de la variable τ est $E(\tau) = 0$

- la variance de la variable τ est $V(\tau) = \frac{2(2n+5)}{9n(n-1)}$

29.2.1. Méthodes de calcul:

méthode n°1

- dénombrer parmi les paires $\{p,q\}$ telles que $p < q$, le nombre a_i de paires telles que $T_p < T_q$

- calculer $A = \sum_{i=1}^{i=n} a_i$

le coefficient de Kendall $\tau = \frac{4A}{n(n-1)} - 1$

méthode n°2

- considérer les $\frac{n(n-1)}{2}$ paires d'observations,

- à chaque paire $\{p;q\}$:

on associe 1 si les deux classements sont en accord $p < q$ et $T_p < T_q$

on associe -1 si les deux classements sont en désaccord $p < q$ mais $T_q < T_p$

- calculer la somme $S = \sum_{k=1}^{\frac{n(n-1)}{2}} \delta_k$ avec $\delta_k = 1$ ou -1

le coefficient de Kendall $\tau = \frac{2S}{n(n-1)}$
--

Pour tester la significativité de la valeur obtenue, on prend en référence la situation théorique d'indépendance des deux rangements dans la population, c'est à dire que les $n!$ classements sont équiprobables. Dans ce cas la valeur exacte est $\tau = 0$.

Dans le cas d'une tendance monotone croissante exacte, les classements sont identiques, $R_k = k$, $\tau = 1$: les deux classements sont identiques

Dans le cas d'une tendance monotone décroissante exacte, les classements sont inversés, $R_k = n+1-k$, $\tau = -1$: les deux classements sont inverses

Il s'agit alors de prendre une décision sur la base d'une valeur τ , réalisation de la variable τ sur un échantillon. A l'aide de la table du coefficient de corrélation de Kendall ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'indépendance des deux rangements. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1 tendance monotone croissante dépendance positive	H1 tendance monotone dépendance quelconque	H1 tendance monotone décroissante dépendance négative
$W = \{ \tau, \tau > c \}$	$W = \{ \tau \mid \tau > c \}$	$W = \{ \tau, \tau < -c \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

Asymptotiquement la variable $Z_K = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}$ suit la loi de la variable de Laplace-Gauss

Gauss LG(0,1). Cette approximation est jugée acceptable pour $n > 7$

29.2.2. Cas d'*ex æquo* :

La démarche consiste pour conserver la somme des rangs égale à $\frac{n(n+1)}{2}$ à attribuer la moyenne des rangs que les observations auraient occupés si elles avaient été distinctes. Ainsi supposons qu'il y ait un groupe de m observations *ex æquo* aux rangs : $l+1, l+2, \dots, l+m$, on leur attribue le même rang $l + \frac{m+1}{2}$ à chaque paire $\{l+i, l+j\}$ avec $i=1$ à q et $j=1$ à m et $i \neq j$ on associe 0.

Cela revient à prendre la règle suivante :

- à chaque paire $\{p; q\}$:

on associe 1 si les deux classements sont en accord, c'est à dire $p < q$ et $T_p < T_q$

on associe -1 si les deux classements sont en désaccord, c'est à dire $p < q$ mais $T_q < T_p$

on associe 0 si les deux classements sont *ex æquo*, c'est à dire $p < q$ mais $T_p = T_q$

- calculer la somme $S = \sum_{k=1}^{\frac{n(n-1)}{2}} \delta_k$ avec $\delta_k = 1$ ou -1

La valeur maximum prise par S est alors : $\frac{n(n-1)}{2} - \frac{1}{2} \sum_{m \in E_X} m(m-1)$

$$\tau^* = \frac{S}{\sqrt{\frac{n(n-1)}{2} - \frac{1}{2} \sum_{m \in E_X} m(m-1)} \sqrt{\frac{n(n-1)}{2} - \frac{1}{2} \sum_{m \in E_Y} m(m-1)}}$$

E_X est l'ensemble des groupes d'*ex æquo* dans l'échantillon de X

E_Y est l'ensemble des groupes d'*ex æquo* dans l'échantillon de Y

Pour l'approximation de la loi de τ^* par la loi de la variable de Laplace-Gauss LG(0,1), il convient de tenir compte de la diminution de la variance de S .

30. Test d'homogénéité (Problème à deux échantillons)

Cette fois nous disposons d'un m-échantillon ($X_1, X_2, \dots, X_k, \dots, X_m$) de la variable X et d'un n-échantillon ($Y_1, Y_2, \dots, Y_l, \dots, Y_n$) de la variable Y. La statistique utilisée est alors de la forme :

$$S_N = \sum_{k=1}^{k=m} f_N(R_k)$$

où $N = m + n$ est le nombre total d'observations issues de l'échantillon global, ($X_1, X_2, \dots, X_k, \dots, X_m, Y_1, Y_2, \dots, Y_l, \dots, Y_n$), R_k est le rang de X_k dans l'échantillon global ordonné, f_N est une fonction des nombres entiers de 1 à N caractérisant les rangs et c_k sont telles que : $c_k = 1$ pour $k = 1$ à m et $c_k = 0$ pour $k = m + 1$ à $m + n$.

L'hypothèse nulle H_0 traduit l'idée selon laquelle :

"les deux variables X et Y sont régies par la même distribution de probabilité" ou encore *"X et Y admettent deux fonctions de répartition égales $F = G$ "*

Une des hypothèses alternatives H_1 traduit une idée contraire selon laquelle :

"la distribution de probabilité de X est différente de celle de Y" ou encore *" $F \neq G$ "* ou *" $F > G$ "* ou *" $F < G$ "*.

30.1. Le test de Wilcoxon

Rangs des observations X dans l'échantillon global ordonné de taille $N = m + n$	R_1	R_2	...	R_k	...	R_m	
Rangs des observations Y dans l'échantillon global ordonné de taille $N = m + n$	Q_1	Q_2	...	Q_k	Q_n

Ce test est fondé sur la statistique suivante :

$$S_N = W_N = \sum_{k=1}^{k=m} R_k$$

L'espérance est $E(W_N) = m \frac{N+1}{2} = \frac{m(m+n+1)}{2}$

La variance est $V(W_N) = \frac{mn(m+n+1)}{12}$

Sous l'hypothèse nulle H_0 , la loi de W_N est symétrique autour de $E(W_N)$.

Les valeurs extrêmes correspondent au cas où :

- si toutes les valeurs de X sont supérieures à celles de Y, le vecteur (R_1, R_2, \dots, R_m)

est une permutation de la réalisation $(n+1, n+2, \dots, n+m)$ et $W_N \max = nm + \frac{m(m+1)}{2}$

- si toutes les valeurs de X sont inférieures à celles de Y, le vecteur (R_1, R_2, \dots, R_m) est une permutation de la réalisation $(1, 2, \dots, m)$ et $W_N \min = \frac{m(m+1)}{2}$

Il s'agit alors de prendre une décision sur la base d'une valeur w_N , réalisation de la variable W_N sur un échantillon. A l'aide d'une table ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'identité des deux distributions. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1 F > G	H1 F = G	H1 F < G
$W = \{ w_N, w_N \leq c_1 \}$	$W = \{ w_N, w_N \leq c_1 \text{ ou } w_N \geq c_2 \}$	$W = \{ w_N, w_N \geq c_2 \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

A l'aide d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss $\frac{W_N - E(W_N)}{\sigma(W_N)}$ tend en loi vers $LG(0,1)$

Dans ce cas, il convient de tenir du passage d'une variable discrète à une variable continue en réalisant une correction (*correction de continuité*) selon la démarche suivante

$$: \text{Prob} \{W_N = s\} = \text{Prob} \left\{ s - \frac{1}{2} < W_N < s + \frac{1}{2} \right\}$$

30.1.1. Cas d'ex æquo :

La démarche consiste pour conserver la somme des rangs égale à $\frac{n(n+1)}{2}$ à attribuer la moyenne des rangs que les observations auraient occupés si elles avaient été distinctes.

Ainsi supposons qu'il y ait un groupe de q observations *ex æquo* aux rangs :

$$k+1, k+2, \dots, k+q, \text{ on leur attribue le même rang } k + \frac{q+1}{2}$$

on remplace alors la variance $V(W_N)$ par

$$V^*(W_N) = \frac{mn(n+m+1)}{12} - \frac{1}{n+m} \sum_{q \in E_X} \frac{q(q^2-1)}{12}$$

E_X est l'ensemble des groupes d'*ex æquo* dans l'échantillon de X

30.2. Le test de Mann et Whitney

Rangs des observations X dans l'échantillon global ordonné de taille $N = m + n$	R_1	R_2	...	R_k	...	R_m
Rangs des observations Y dans l'échantillon global ordonné de taille $N = m + n$	Q_1	Q_2	...	Q_k	...	Q_n

Ce test est fondé sur la statistique suivante dérivée de celle de Wilcoxon :

$$S_N = U_N = W_N - \frac{m(m+1)}{2} = \sum_{k=1}^{k=m} R_k - \frac{m(m+1)}{2}$$

on peut en déduire que :

- U_N varie entre 0 et mn
- l'espérance est $E(U_N) = \frac{mn}{2}$
- la variance est $V(U_N) = V(W_N) = \frac{mn(m+n+1)}{12}$

30.2.1. Conditions d'utilisation:

Etant donnés les deux échantillons indépendants $(X_1, X_2, X_3, \dots, X_m)$ et $(Y_1, Y_2, Y_3, \dots, Y_n)$ issus de deux populations P_1 et P_2 .

On mélange ces deux échantillons et on réordonne les valeurs.

On dénombre les couples (X_p, Y_q) tels que X_p a un rang plus grand que Y_q

U_N est la variable qui à chaque situation associe ce nombre. Elle varie entre 0 et nm selon les deux cas extrêmes:

$$X_1, X_2, X_3, \dots, X_m, Y_1, Y_2, Y_3, \dots, Y_n \text{ et } Y_1, Y_2, Y_3, \dots, Y_n, X_1, X_2, X_3, \dots, X_m,$$

Sous l'hypothèse de l'identité des distributions des deux variables X et Y, la loi exacte de U_N peut être calculée pour de faibles valeurs de n et de m. Toutefois dès que $n > 8$ et $m > 8$ on peut l'approcher par une loi de Laplace-Gauss.

$$\frac{U_N - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \approx LG(0;1)$$

30.2.2. Statistique et variable de décision

Pour plus de faciliter, on utilise de façon intermédiaire la variable W_X , somme des rangs de la variable X, puis on calcule $U_N = W_X - \frac{m(m+1)}{2}$

30.2.3. Test bilatéral :

H_0 (identité des deux distributions $F = G$) contre H_1 (Les deux distributions sont différentes $F \neq G$)

30.2.4. Test unilatéral :

Ho (identité des deux distributions $F = G$) contre H1 (Les deux distributions sont différentes soit $F < G$, soit $F > G$)

Il s'agit alors de prendre une décision sur la base d'une valeur u_N , réalisation de la variable U_N sur un échantillon. A l'aide d'une table ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle Ho postulant l'identité des deux distributions. La région critique W, région de rejet de Ho au profit de H1, est définie selon H1, l'une des trois hypothèses alternatives usuelles.

H1 F > G	H1 F ≠ G	H1 F < G
$W = \{ u_N, u_N \leq c_1 \}$	$W = \{ u_N, u_N \leq c_1 \text{ ou } u_N \geq c_2 \}$	$W = \{ u_N, u_N \geq c_2 \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

A l'aide d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss $\frac{U_N - E(U_N)}{\sigma(U_N)}$ tend vers LG(0,1)

Dans ce cas, il convient de tenir du passage d'une variable discrète à une variable continue en réalisant une correction (*correction de continuité*) selon la démarche suivante : $\text{Prob} \{U_N = s\} = \text{Prob} \{s - \frac{1}{2} < U_N < s + \frac{1}{2}\}$ La valeur critique au seuil α s'obtient alors par la résolution d'équation

30.3. Le test de la médiane

Cette approche a été introduite par J. Westenberg dans "Significance test for median and interquartile range in samples from continuous populations of any form" *Proceedings, Koninklijke Nederlands Akademie van Wetenschappen*, n°51, 1948, pp 252-261.

Elle est développée par Alexander McFarlane Mood dans *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950.

Rangs des observations X dans l'échantillon global ordonné de taille $N = m + n$	R ₁	R ₂	...	R _k	...	R _m
Rangs des observations Y dans l'échantillon global ordonné de taille $N = m + n$	Q ₁	Q ₂	...	Q _k	...	Q _n

La statistique utilisée est

$$S_N = M_N = \sum_{k=1}^{i=N} c_i f_N(R_i) = \sum_{i=1}^{i=m} \mathbf{1}_{R^+}(R_i - \frac{N+1}{2})$$

Cette variable associée à chaque échantillon global ordonné de taille $N = m+n$, le nombre d'observations concernant la variable X situées strictement au-delà de la médiane. Il s'agit des observations relatives à la variable X dont le rang est strictement plus grand que $\frac{N+1}{2}$. La loi exacte de M_N est connue : c'est une loi hypergéométrique. En effet l'événement $\{M_N = k\}$ est réalisé par toutes les permutations, réalisations du vecteur statistique des rangs $R = (R_1, R_2, \dots, R_i, \dots, R_m)$ qui contiennent k rangs strictement supérieurs à $\frac{N+1}{2}$ et $m-k$ rangs inférieurs ou égaux à $\frac{N+1}{2}$. Il convient pour aborder cette loi de distinguer deux cas en fonction de la parité de N .

30.3.1. cas où $N = 2p$:

$$\text{Prob} \{M_N = k\} = \frac{\binom{k}{p} \binom{m-k}{p}}{\binom{m}{N}} \text{ avec } \max\{0, m-p\} \leq k \leq \min\{p, m\}$$

car il s'agit de choisir k rangs parmi les rangs $p+1, p+2, \dots, p+p=2p$ et $m-k$ rangs parmi les rangs $1, 2, \dots, p$. Sous l'hypothèse H_0 , toutes les permutations sont considérées comme équiprobables. L'espérance de la variable est : $E(M_N) = \frac{m}{2}$ La variance de la variable est : $V(M_N) = \frac{m(N-m)}{4(N-1)}$

30.3.2. cas où $N = 2p+1$:

$$\text{Prob} \{M_N = k\} = \frac{\binom{k}{p+1} \binom{m-k}{p}}{\binom{m}{N}} \text{ avec } \max\{0, m-p+1\} \leq k \leq \min\{p, m\} \text{ car il s'agit de choisir } k$$

rangs parmi les rangs $p+2, p+2, \dots, p+p+1=2p+1$ et $m-k$ rangs parmi les rangs $1, 2, \dots, p+1$. Sous l'hypothèse H_0 , toutes les permutations sont considérées comme équiprobables. L'espérance de la variable est : $E(M_N) = \frac{m(N-1)}{2N}$ La variance de la variable est : $V(M_N) = \frac{m(N-m)(N^2-1)}{4N^2(N-1)}$

30.3.3. Test bilatéral :

H_0 (identité des deux distributions $F = G$) contre H_1 (Les deux distributions sont différentes $F \neq G$)

30.3.4. Test unilatéral :

H_0 (identité des deux distributions $F = G$) contre H_1 (Les deux distributions sont différentes soit $F < G$, soit $F > G$)

30.3.5. Décision :

Il s'agit alors de prendre une décision sur la base d'une valeur m_N , réalisation de la variable M_N sur un échantillon. A l'aide d'une table ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'identité des deux distributions. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H_1 $F > G$	H_1 $F \neq G$	H_1 $F < G$
$W = \{ m_N, m_N \leq c_1 \}$	$W = \{ m_N, m_N \leq c_1 \text{ ou } m_N \geq c_2 \}$	$W = \{ m_N, m_N \geq c_2 \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W|H_0) = \alpha$

A l'aide d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss

$\frac{M_N - E(M_N)}{\sigma(M_N)}$ tend en loi vers $LG(0;1)$ quand m et n tendent vers l'infini.

Il convient aussi de tenir compte du passage d'une variable M_N à valeur entière à une variable absolument continue.

$$\text{Prob} \{ M_N = c \} = \text{Prob} \left\{ c - \frac{1}{2} < M_N < c + \frac{1}{2} \right\} =$$

$$\text{Prob} \left\{ \frac{c - \frac{1}{2} - E(M_N)}{\sqrt{V(M_N)}} < \frac{M_N - E(M_N)}{\sigma(M_N)} < \frac{c + \frac{1}{2} - E(M_N)}{\sqrt{V(M_N)}} \right\} \approx \text{Prob} \left\{ \frac{c - \frac{1}{2} - E(M_N)}{\sqrt{V(M_N)}} < LG(0;1) < \frac{c + \frac{1}{2} - E(M_N)}{\sqrt{V(M_N)}} \right\}$$

31. Test d'indépendance (Problème à k échantillons)

31.1. Test de concordance de p rangements de n objets de M.G. Kendall

Objets Critères	1	2	...	i	...	n	totaux
1	R ₁₁	R ₂₁		R _{i1}		R _{n1}	R _{.1}
2	R ₁₂	R ₂₂		R _{i2}		R _{n2}	R _{.2}
...							
j	R _{1j}	R _{2j}		R _{ij}		R _{nj}	R _{.j}
...							
p	R _{1p}	R _{2p}		R _{ip}		R _{np}	R _{.p}
Totaux	R _{1.}	R _{2.}		R _{i.}		R _{n.}	r _{..}

31.1.1. Méthode de calcul:

Chaque ligne est une permutation des nombres entiers de 1 à n dont la somme est constante et vaut $\frac{n(n+1)}{2}$. Ainsi $r_{..} = p \frac{n(n+1)}{2}$. Dans l'hypothèse d'une concordance parfaite, les totaux des colonnes seraient égaux respectivement à p, 2p, 3p, ..., np à une permutation près. On utilise alors la statistique $S_K = \sum_{i=1}^{i=n} (R_{i.} - \frac{r_{..}}{n})^2$ dont la valeur

maximale est $S_{\max} = \frac{p^2(n^3-n)}{12}$

Le coefficient de concordance de Kendall est:

$$W = \frac{S_K}{S_{\max}} = \frac{S_K}{\frac{1}{12}p^2(n^3-n)}$$

On peut aussi $W = \frac{1}{p} + \frac{2}{p^2} \sum_{i=1}^{i=p-1} \sum_{j=i+1}^j=p R_{ij}$ D'où

- l'espérance de W : $E(W) = \frac{1}{p}$

- la variance de W : $V(W) = \frac{2(p-1)}{p^3(n-1)}$

- le moment centré d'ordre 3 de W : $\mu_3(W) = \frac{8(p-1)(p-2)}{p^5(n-1)^2}$

- le moment centré d'ordre 4 de W : $\mu_4(W) = \frac{12(p-1)^2}{p^6(n-1)^2} + \frac{48(p-1)(p-2)(p-3)}{p^7(n-1)^3} - \frac{48(p-1)}{p^7(n+1)(n-1)^2}$

w = 0 correspond au cas où chaque colonne a même total. De faibles valeurs de W suggèrent l'indépendance des classements.

Pour tester l'hypothèse nulle H_0 d'indépendance des p rangements, on procède selon les procédures suivantes:

- Pour $n \leq 7$ on utilise une table,

- Pour $n \leq 7$ et $2 < p \leq 20$ la variable $\frac{(p-1)W}{1-W}$ est distribuée comme la variable de Fisher-Snédecor $F(n_1 = n-1-\frac{2}{p} ; n_2 = (p-1)(n-1-\frac{2}{p}))$

- Pour $n > 7$ on utilise la variable $p(n-1)W$ qui est une variable de Pearson χ^2_{n-1}

Dans le cas où l'on est conduit à rejeter l'hypothèse d'indépendance des classements, on utilise souvent la règle de classement suivante:

les objets sont classés dans l'ordre défini par les totaux des colonnes.

Lorsqu'il y a des *ex æquo* on remplace le rang de ceux-ci par la moyenne

p	n	3	4	5	6
3		1	0,750	0,600	0,500
4		0,822	0,619	0,500	0,421
5		0,716	0,553	0,449	0,377
6		0,660	0,512	0,418	0,351
7		0,626	0,484	0,395	0,332
8		0,595	0,461	0,378	0,319
9		0,576	0,447	0,365	0,307
10		0,560	0,434	0,354	0,299
11		0,548	0,425	0,346	0,287
12		0,535	0,415	0,336	0,287
13		0,527	0,409	0,332	0,280
14		0,520	0,402	0,327	0,275
15		0,514	0,395	0,322	0,272
20		0,49	0,37	0,30	0,25
40		0,43	0,33	0,26	0,22
60		0,41	0,31	0,25	0,21
100		0,38	0,29	0,24	0,20
		0,33	0,25	0,20	0,17

arithmétique des rangs qu'ils auraient eu sans *ex æquo*.

$$W = \frac{S_K}{S_{\max}} = \frac{12S_K}{p^2(n^3-n)-p \sum_{j=1}^{j=p} (t_j^3 - t_j)}$$

avec t_j = nombre d'*ex æquo* au $j^{\text{ème}}$ classement

Cette table fournit les valeurs critiques k telles que $P(W \geq k) = \alpha = 0,05$

31.2. Test de concordance de p rangements de n objets de Friedman

31.2.1. Méthode de calcul:

Chaque ligne est une permutation des nombres entiers de 1 à n dont la somme est constante et vaut $\frac{n(n+1)}{2}$. Chaque colonne est constituée des rangs attribués par les

observations pour chaque objet. Ainsi $R_{i.} = \sum_{j=1}^{i=p} R_{ij}$ et $R_{.j} = \frac{1}{p} \sum_{i=1}^{j=p} R_{ij}$ fournit le rang

moyen de l'objet i . La moyenne générale des rangs $R = \frac{1}{np} \frac{n(n+1)}{2} = \frac{n+1}{2}$ Dans l'hypothèse d'une indépendance des classements, le rang moyen $R_{i.}$ de chaque colonne devrait peu s'écarter de la valeur R . On utilise alors la statistique de Friedman

$$S_F = \frac{p}{\frac{1}{12}n(n+1)} \sum_{i=1}^p (R_{i.} - \frac{n+1}{2})^2 = \frac{12}{pn(n+1)} \sum_{i=1}^p (R_{i.} - \frac{p(n+1)}{2})^2$$

La distribution exacte de la variable S_F sous l'hypothèse H_0 peut s'obtenir en considérant les $(n!)^{p-1}$ configurations possibles des rangs. En pratique on utilise une distribution approchée qui est celle de la variable de Pearson (Khi-Deux) à ddl = $n-1$. On fixe un seuil de rejet α pour déterminer la valeur critique k par la relation $\text{Prob}\{S_F > k\} = \alpha$ ou $\text{Prob}\{S_F \leq k\} = 1 - \alpha$

On calcule la valeur expérimentale s_F . Si $s_F > k$, on rejette H_0 en prenant un risque de première espèce de niveau α sinon on ne la rejette pas et H_0 est conservée avec un risque de seconde espèce de niveau β

On peut alors remarquer que $S_F = \frac{1}{\frac{1}{12}pn(n+1)} S_K = p(n-1) W$

32. Test d'homogénéité (Problème à k échantillons)

32.1. Comparaison de k échantillons indépendants Test H de Kruskal-Wallis

Rangs des observations X_1 dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{11}	R_{12}	...	R_{1k}	...	R_{1n_1}
Rangs des observations X_2 dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{21}	R_{22}	...	R_{2k}	...	R_{2n_1}
Rangs des observations X_p dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{p1}	R_{p2}	...	R_{pk}	...	R_{pn_p}
Rangs des observations X_k dans l'échantillon global ordonné de taille $N = \sum_{p=1}^{p=k} n_p$	R_{k1}	R_{k2}	...	R_{kk}	...	R_{kn_k}

32.1.1. Conditions d'utilisation:

Étant donnés les k échantillons indépendants respectivement de taille n_1, n_2, \dots, n_k issus de k populations $P_1, P_2 \dots P_k$. La variable étudiée est une variable ordinale

32.1.2. Statistique et variable de décision

On mélange ces k échantillons et on réordonne les $N = \sum_{p=1}^{p=k} n_p$ valeurs.

On prend en compte le rang de chaque observation dans le classement global

Sous l'hypothèse H_0 de l'identité des k distributions de la variable ordinale, les rangs sont distribués au hasard dans chaque échantillon. Considérons:

- la variable S_p = somme des rangs des n_p observations de l'échantillon n^o p

- la variable "rang moyen" est alors: $\frac{S_p}{n_p}$

- l'espérance de cette variable "rang moyen" sous H_0 :

$$E\left(\frac{S_p}{n_p}\right) = \frac{1}{n_p} E(S_p) = \frac{1}{n_p} \left(n_p \frac{N+1}{2} \right) = \frac{N+1}{2}$$

On mesure l'écart entre les résultats attendus sous l'hypothèse H_0 et les observations par la variable H :

$$H = \frac{12}{N(N+1)} \sum_{p=1}^{p=k} n_p \left(\frac{S_p}{n_p} - \frac{N+1}{2} \right)^2$$

La distribution exacte de H s'obtient par le dénombrement des m configurations équiprobables

$$m = \frac{n!}{n_1! n_2! \dots n_k!} \quad \text{où} \quad n = \sum_{p=1}^{p=k} n_p$$

L'espérance de la variable H : $E(H) = k-1$

$$\text{La variance de la variable H : } V(H) = 2(k-1) - \frac{2[3k^2 - 6k + N(2k^2 - 6k + 1)]}{5N(N+1)} - \frac{6}{5} \sum_{p=1}^{p=k} \frac{1}{n_p}$$

Cependant la distribution de H peut être approchée par la distribution de la variable de Pearson à k-1 ddl. Pour calculer la valeur expérimentale h, on peut aussi utiliser une

$$\text{expression équivalente de la variable H : } H = \frac{12}{n(n+1)} \left\{ \sum_{p=1}^{p=k} \frac{S_p^2}{n_p} \right\} - 3(n+1)$$

32.1.3. Test unilatéral : *Ho (identité des K distributions) contre H1 (deux distributions au moins sont différentes)*

On choisit un niveau de risque de 1ère espèce α . A ce seuil, on détermine la valeur critique c :

- soit à l'aide d'une table du H de Kruskal et Wallis
- soit à l'aide de la table de la variable de Pearson à ddl = k-1

Dans ce cas, on détermine la valeur c telle que $\text{Prob} \{ \chi^2(k-1) < c \} = 1 - \alpha$

On détermine la valeur expérimentale h de H par l'une des deux possibilités décrites ci-dessus: **si** $h > c$ alors **on rejette Ho** en prenant un risque de première espèce de niveau α **sinon on ne rejette pas Ho**, ce qui revient à accepter l'hypothèse H1 en prenant un risque de seconde espèce β .

32.1.4. En cas d'existence d'*ex æquo*

En cas d'*ex æquo* chaque observation reçoit un rang égal à la moyenne des rangs qu'elles occupent. Pour chaque groupe de q observations *ex æquo*, on pose $Q = (q-1)q(q+1)$. On calcule la somme $\sum_{i=1}^{i=r} Q_i$ des valeurs Q obtenues pour les r groupes d'*ex*

æquo Et on utilise la variable $H^* = \frac{H}{\sum_{i=1}^{i=r} Q_i} - \frac{1}{1 - \frac{i=1}{n(n-1)}}$

33. Pour aller encore plus loin...

Pour un approfondissement historique et épistémologique, on peut consulter l'article suivant :

William H. Kruskal , W. Allen Wallis , *Université de Chicago*,

" Use of ranks in one-criterion variance analysis " *Journal of the American Statistical Association* (J.A.S.A.) n°260, Volume n°47, décembre 1952, pp 583-621.

Cet article pose une problématique relative à la comparaison de plusieurs n_i -échantillons ($i=1$ à k), le but étant de savoir si ceux-ci sont issus ou non d'une même population.

" A common problem in practical statistics is to decide whether several samples should be regarded as coming from the same population. Almost invariably the samples differ, and the question is whether the differences signify differences among the populations, or are merely the chance variations to be expected among random samples from the same population. When this problem arises one may often assume that the populations are of approximately the same form, in the sense that if they differ it is by a shift or translation. " (page 584)

Il développe ensuite les avantages qu'apporte l'usage du rang.

" - The calculations are simplified (...)

- Only very general assumptions are made about the kind of distributions from which the observations come.(...)

- Data available only in ordinal form may often be used.

- When the assumptions of the usual test procedure are too far from reality, not only is there a problem of distribution theory if the usual test is used, but it is possible that the usual test may not have as good a chance as a rank test of detecting the kinds of difference of real interest." (page 585)

Un autre point intéressant de cet article est l'exploration de plusieurs tests déjà établis à cette époque et le lien avec l'approche que les auteurs proposent.

- Tests de permutation et rangs (*Permutation Tests and Ranks*)

- χ_r^2 de Milton Friedman (*Friedman's χ_r^2*)

- Test des deux échantillons de Frank Wilcoxon (*Wilcoxon's Two-Sample Test*)

Ce test fut abordé par plusieurs statisticiens :

- Milton Wilcoxon (1945, 1947)

- Léon Festinger (1946)

- H. B. Mann et D. R. Whitney (1947)

- J. B. S.Haldane et Cedric A. B. Smith (1948)

- Colin White (1952)

- Test des trois échantillons de D.R. Whitney (*Whitney's Three-Sample Test*)

- Test des k échantillons de T. J. Terpstra (*Terpstra's C-sample Test*)

- Test des k échantillons de Frederick Mosteller (*Mosteller's C-sample Test*)

- Test fondé sur les rangs normalisés de Ronald, Aylmer Fisher et Frank Yates (*Fisher and Yates' Normalized Ranks*)
- Test des séquences homogènes (*runs*) introduits par A. Wald et J. Wolfowitz
- Test de la médiane introduit par J. Westenberg
- Tests fondés sur la statistique d'ordre (*Order statistics*) introduits par Alexander Mc Farlane Mood et Brown , puis Frank J. Massey

ANALYSE DE VARIANCE

L'objet de l'analyse de variance est la construction de notions, de techniques de tests et d'estimations visant à apprécier l'effet d'une ou plusieurs variables qualitatives sur une variable quantitative. donner un sens.

Les « facteurs de variabilité » désignent les variables qualitatives susceptibles d'influer sur la variable quantitative et les « niveaux des facteurs » désignent les modalités des variables qualitatives. Dans le cas où il y a plusieurs facteurs, on désigne par « traitement » une combinaison des niveaux.

33.1. Analyse de variance à un critère

33.1.1. Conditions d'utilisation:

On s'intéresse à une variable quantitative X à expliquer par une variable A explicative (considérée comme une variable qualitative)

La situation étudiée conduit à ne tenir compte que d'une seule variable (explicative), c'est à dire d'un seul facteur de variabilité, à p modalités et à la mettre sous contrôle.

Chaque modalité du facteur contrôlé détermine un groupe d'individus dans la population.

Les observations (réalisations expérimentales de la variable X) sont réalisées sur la base de tirages aléatoires et indépendants d'individus dans chacun des groupes.

Ces observations (ces mesures) sont supposées sans erreur, ce qui en pratique revient à considérer que l'erreur de mesure est d'un ordre de grandeur négligeable devant la variabilité du facteur contrôlé expérimentalement.

33.1.2. Mise en place des notations pour le traitement statistique:

Le **plan d'expérience** conduit à recueillir des informations rassemblées dans le tableau ci-dessous:

modalités du facteur contrôlé	répétitions des observations	nombre de répétitions
A_1	$x_{11}, x_{12}, \dots, x_{1r}, \dots, x_{1n_1}$	n_1
A_2	$x_{21}, x_{22}, \dots, x_{2r}, \dots, x_{2n_2}$	n_2
...
A_i	$x_{i1}, x_{i2}, \dots, x_{ir}, \dots, x_{in_i}$	n_i
...
A_p	$x_{p1}, x_{p2}, \dots, x_{pr}, \dots, x_{pn_p}$	n_p

caractéristiques statistiques	algorithmes	notations mathématiques
somme des mesures sur le groupe A_j	$\mathbf{x}_{j.} = x_{j1} + x_{j2} + \dots + x_{jr} + \dots + x_{jn_j}$	$r = n_j$ $\mathbf{x}_{j.} = \sum_{r=1}^{n_j} x_{jr}$
somme des mesures sur les p groupes A_i	$\mathbf{x}_{..} = x_{1.} + x_{2.} + \dots + x_{i.} + \dots + x_{p.}$	$i = p$ $\mathbf{x}_{..} = \sum_{i=1}^p \mathbf{x}_{i.}$
effectif total des mesures	$N = n_1 + n_2 + \dots + n_i + \dots + n_p$	$i = p$ $N = \sum_{i=1}^p n_i$
moyenne des mesures sur le groupe A_j (moyenne du groupe)	$\bar{X}_{j.} = \frac{x_{j1} + x_{j2} + \dots + x_{jr} + \dots + x_{jn_j}}{n_j}$	$\bar{X}_{j.} = \frac{\mathbf{x}_{j.}}{n_j}$
moyenne des mesures sur les p groupes A_i (moyenne globale)	$\bar{X}_{..} = \frac{x_{1.} + x_{2.} + \dots + x_{i.} + \dots + x_{p.}}{n_1 + n_2 + \dots + n_i + \dots + n_p}$	$\bar{X}_{..} = \frac{\mathbf{x}_{..}}{N}$
somme des carrés des écarts à la moyenne globale, des mesures moyennées pondérées par les effectifs de chaque groupe (fluctuation intergroupe)	$SCE_A = \sum_{i=1}^p n_i (\bar{x}_{i.} - \bar{x}_{..})^2$	
somme des carrés des écarts à la moyenne du groupe, des mesures à l'intérieur de chaque groupe (fluctuation intragroupe)	$SCE_{A_i} = \sum_{r=1}^{n_i} (x_{ir} - \bar{x}_{i.})^2$	
somme des SCE_{A_i} fluctuation intragroupe totale	$SCE_R = \sum_{i=1}^p \sum_{r=1}^{n_i} (x_{ir} - \bar{x}_{i.})^2$	$i = p$ $SCE_R = \sum_{i=1}^p SCE_{A_i}$
somme des carrés des écarts à la moyenne, de toutes les mesures (fluctuation totale)	$SCE_T = \sum_{i=1}^p \sum_{r=1}^{n_i} (x_{ir} - \bar{x}_{..})^2$	
variance due au facteur A	$S_A^2 = \frac{SCE_A}{N}$	
variance résiduelle	$S_R^2 = \frac{SCE_R}{N}$	
variance totale	$S_T^2 = \frac{SCE_T}{N} = S_A^2 + S_R^2$	

33.1.3. Mise en place du modèle statistique:

Nous supposons que:

- le facteur A n'influe que sur les moyennes μ_i des variables X_i et non sur les variances σ^2 .

- le facteur A agit de façon additive

- les variables X_i sont des variables de Laplace-Gauss de paramètres μ_i et σ : $X_i = \text{LG}(\mu_i, \sigma)$ pour $i = 1$ à p

- la variance σ^2 est indépendante des variantes du facteur contrôlé, elle est appelée "erreur expérimentale"

Ainsi nous sommes conduits à tester l'égalité des p moyennes correspondant aux p groupes. Ceci revient à tester l'hypothèse H_0 ($\mu_1 = \mu_2 = \dots = \mu_p = \mu$) contre H_1 (il y a au moins deux moyennes différentes : $\mu_s = \mu_t$)

Pour traiter ce problème, on prend p échantillons correspondant chacun à un niveau du facteur, soit donc les n_i -échantillons ($X_{i1}, X_{i2}, \dots, X_{in_i}$) pour $i = 1$ à p .

On pose :

$$X_{ir} = \mu + \alpha_i + \varepsilon_{ir}$$

μ représente une constante, c'est à dire la valeur moyenne commune

α_i représente la variable exprimant l'effet du niveau i du facteur A

ε_{ir} représente la variable exprimant le résidu dont la loi de probabilité est celle de la variable Laplace-Gauss $\text{LG}(0; \sigma)$

L'hypothèse nulle peut alors être formalisée par ($\alpha_i = 0$ pour tout $i = 1$ à p) contre l'hypothèse alternative (il existe un $\alpha_i \neq 0$)

Nous pouvons considérer $\bar{X} = \frac{1}{N} \sum_{i=1}^p \sum_{r=1}^{n_i} X_{ir}$ et $\bar{X}_i = \frac{1}{n_i} \sum_{r=1}^{n_i} X_{ir}$

Puis nous remarquons que $X_{ir} - \bar{X} = X_{ir} - \bar{X}_i + \bar{X}_i - \bar{X}$.

$$[X_{ir} - \bar{X}]^2 = [X_{ir} - \bar{X}_i]^2 + [\bar{X}_i - \bar{X}]^2 + 2[X_{ir} - \bar{X}_i][\bar{X}_i - \bar{X}]$$

$$\text{d'où } \sum_{r=1}^{n_i} [X_{ir} - \bar{X}]^2 =$$

$$\sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}_i]^2 + \sum_{r=1}^{r=n_i} [\bar{X}_i - \bar{X}]^2 + 2 \sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}_i] [\bar{X}_i - \bar{X}] =$$

$$\sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}_i]^2 + n_i [\bar{X}_i - \bar{X}]^2 + 2 [\bar{X}_i - \bar{X}] \sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}_i] =$$

$$\sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}_i]^2 + n_i [\bar{X}_i - \bar{X}]^2$$

Puis

$$\sum_{i=1}^{i=p} \sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}]^2 = \sum_{i=1}^{i=p} \sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}_i]^2 + \sum_{i=1}^{i=p} n_i [\bar{X}_i - \bar{X}]^2$$

et donc que la variance totale

$$\frac{1}{N} \sum_{i=1}^{i=p} \sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}]^2 = \frac{1}{N} \sum_{i=1}^{i=p} \sum_{r=1}^{r=n_i} [X_{ir} - \bar{X}_i]^2 + \frac{1}{N} \sum_{i=1}^{i=p} n_i [\bar{X}_i - \bar{X}]^2 .$$

Cette formule d' «analyse de variance» peut être écrite :

$$S_T^2 = S_A^2 + S_R^2$$

avec

S_T^2 = variance totale

S_A^2 = variance due au facteur , variance interclasse

S_R^2 = variance résiduelle , variance intraclasse

33.1.4. Estimation des effets du facteur contrôlé A:

Les observations expérimentales permettent d'obtenir les estimations sans biais :

μ est estimée par m qui est une réalisation de l'estimateur $\bar{X}_{..}$ sur l'échantillon

La valeur prise par la variable α_j peut être estimée par $x_{j.} - \bar{X}_{..}$, qui est une réalisation de $\bar{X}_j - \bar{X}_{..}$,

La valeur prise par la variable ε_{ir} peut être estimée par $x_{ir} - \bar{X}_{i.}$, qui est une réalisation de $X_{ir} - \bar{X}_{i.}$

33.1.5. Tableau d'analyse de variance:

source de la variation	somme des carrés des écarts	degrés de liberté	Carrés moyens
entre les variantes du facteur A	SCE _A	p-1	CM _A = $\frac{SCE_A}{p-1}$
erreur aléatoire	SCE _R	N-p	CM _R = $\frac{SCE_R}{N-p}$
Globalement	SCE _T	N-1	

33.1.6. Interprétation des résultats:

- $CM_R = \frac{N}{N-p} S_R^2$ est un estimateur sans biais qui fournit une estimation de l'erreur σ^2 quel que soit l'effet du facteur A est une

On démontre que la variable aléatoire $\frac{N}{\sigma^2} S_R^2$ est une variable de Pearson χ^2 à ddl=N-1

- sous l'hypothèse Ho, tous les α_i sont nuls, alors $CM_A = \frac{N}{p-1} S_A^2$ est un estimateur sans biais, indépendant de CM_R , qui fournit une estimation, de l'erreur σ^2

On démontre que la variable aléatoire $\frac{N}{\sigma^2} S_A^2 = \sum_{i=1}^p \frac{n_i}{\sigma^2} S_{Ai}^2$ est une variable de Pearson χ^2 à ddl =N-p.

$$\text{Donc } F_A = \frac{\frac{N}{(p-1)\sigma^2} S_A^2}{\frac{N}{(N-p)\sigma^2} S_R^2} \text{ est une variable de Fisher-Snedecor à ddl}=(p-1, N-p)$$

On choisit un niveau de risque de 1ère espèce α .

hypothèse testée	Critère	Variable aléatoire de Fisher-Snédecor
le facteur A n'a aucun effet $\alpha_i = 0$ pour tout i	$F_A = \frac{CM_A}{CM_R} = \frac{\frac{S_R^2}{p-1}}{\frac{S_A^2}{N-p}}$ <p>f_A est une réalisation de F_A issue de l'expérience</p>	<p>F(p-1;N-p)</p> <p>Prob{F(p-1;N-p)>k} = α</p>

On détermine la valeur k à partir d'une table de variable F($\nu_1; \nu_2$) si $f_A > k$ alors on rejette Ho au seuil α choisi sinon on ne rejette pas Ho, ce qui revient à rejeter H1 avec un risque de niveau β de seconde espèce.

33.1.7. Comparaison des variances des groupes A_i :

On peut chercher à tester l'hypothèse d'égalité des p variances des groupes. Ainsi il s'agit alors de tester H_0 ($\sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2 = \dots = \sigma_p^2 = \sigma^2$) contre l'hypothèse H_1 (il existe au moins deux variantes différentes : $\sigma_s^2 \neq \sigma_t^2$).

Le **test de Barlett** est un test répondant à cette problématique. Pour cela on suppose les conditions suivantes réalisées :

- on considère p n_i -échantillons issus respectivement des groupes A_i ,
- les variables X_{ij} sont des variables de Laplace-Gauss $LG(\mu_i, \sigma_i)$,
- Aucune des valeurs des variances empiriques n'est nulle, ni trop petite du fait des arrondis,

La statistique utilisée est la suivante

$$B = (N-p) \ln \left(\frac{\sum_{i=1}^{i=k} (n_i-1) S_i^2}{N-p} \right) - \sum_{i=1}^{i=k} (n_i-1) \ln(S_i^2)$$

avec $S_i^2 = \frac{1}{n_i-1} \sum_{r=1}^{r=n_i} (X_{ir} - \bar{X}_i)^2$ et $\ln(\cdot)$ est la fonction logarithme népérien.

Sous l'hypothèse H_0 , B suit la loi d'une variable de Pearson χ^2 à $ddl=k-1$.

La région critique est déterminée par $W = \{ B > c \}$ avec $\text{Prob}\{ \chi^2 > c \} = \alpha$.

Si la réalisation b de B appartient à W , alors on rejette H_0 avec un risque de 1ère espèce de niveau α , sinon on conserve H_0 avec un risque de seconde espèce de niveau β inconnu.

PROBABILITES

34. Des outils théoriques pour une modélisation probabiliste, pour estimer, pour tester

Nombre d'études statistiques nécessitent :

- l'identification de la loi ou des lois de probabilité des variables qui génèrent les données issues du ou des phénomènes observés.

- la connaissance des caractéristiques de ces variables telles que l'espérance mathématique, l'écart-type, la variance, les coefficients de Fisher, la médiane,... en fonction des paramètres intervenant dans les fonctions de répartition ou dans les densités de probabilité.

- l'interprétation de ces caractéristiques fondamentales,

- le recours à des approximations de ces lois.

Ainsi la pratique de la modélisation statistique passe par une maîtrise des notions usuelles des calculs de probabilité. Ce chapitre apporte le vocabulaire et les propriétés de base afin de fournir des outils pour une modélisation probabiliste utiles à l'intelligibilité des situations étudiées.

L'approche théorique mathématique des probabilités permet d'échapper au moins partiellement à la question sur la nature de la probabilité. Néanmoins une réflexion sur les concepts de "hasard", de "probabilité" reste très importante, tout particulièrement lors du recours aux modèles probabilistes pour construire des réponses à des problématiques posées par la réalité concrète. Cependant nous nous limiterons ici à un exposé sur la formalisation d'une expérience aléatoire et sur l'axiomatique de Kolmogorov³.

34.1. EXPERIENCE ALEATOIRE et EVENEMENTS :

Une **expérience aléatoire** est une expérience dont on ne peut prévoir à l'avance avec certitude son résultat et quand bien même elle pourrait être répétée dans des conditions identiques, on pourrait obtenir des résultats différents à chaque fois. On représente cette expérience par l'ensemble de tous les résultats possibles. Ω = **univers des résultats ω possibles**. Cet univers peut être constitué certes d'un nombre **fini** ou **infini dénombrable** d'éléments mais aussi d'un nombre **infini non-dénombrable**.

³ Andreï Nikolaïevitch Kolmogorov (1903-1988) est un des fondateurs de l'école mathématique soviétique. Il a en particulier cherché à fonder rigoureusement une théorie des *processus stochastiques*, i.e. des phénomènes aléatoires dans l'évolution desquels le hasard intervient. L'axiomatique du calcul des probabilités fut rédigée entre 1929 et 1933.

Un **événement** est une assertion relative au résultat de l'expérience. Un événement **est réalisé** ou **n'est pas réalisé** selon qu'après l'expérience cette **assertion est vraie** ou **fausse**. A un événement on peut faire correspondre tous les résultats qui en assureraient la réalisation. Un événement est donc un **sous-ensemble** de l'univers des résultats possibles.

Un **événement élémentaire** est un événement qui ne peut être réalisé que par **un seul résultat**.

Si A et B sont deux événements, nous pouvons alors aussi envisager divers événements de la façon suivante:

- l'**événement «A ou B»** (réunion : $A \approx B$).
- l'**événement «A et B»** (intersection: $A \leftrightarrow B$).
- l'**événement contraire «nonA»** (complémentaire A).
- l'événement contraire de «A ou B» qui est l'événement «nonA et nonB».
- l'événement contraire de «A et B» qui est l'événement «nonA ou nonB».
- l'**événement certain Ω** qui est réalisé par tous les résultats.
- l'**événement impossible \emptyset** qui n'est réalisé par aucun résultat.
- l'**événement «A et non B»** (différence: $A \leftrightarrow B = A - B$)
- l'**événement ««A et non B» ou «nonA et B»** (différence symétrique: $A \Delta B$)

Par ailleurs deux événements sont deux **événements incompatibles** si la réalisation de l'un exclut celle de l'autre, c'est à dire l'événement « A et B» est l'événement impossible. Réciproquement, une question fondamentale se pose:

- Est-ce que tout sous-ensemble de l'univers Ω des résultats possibles peut être un événement ? et même est-il utile qu'il en soit ainsi ?

Pour avoir la qualité d'un événement, un sous-ensemble doit lui-même être élément d'une *classe*⁴ C de parties de l'univers, dotée d'une certaine **structure**, c'est à dire possédant quelques propriétés que nous allons énoncer à la manière d'un *règlement intérieur*. Cette structure est désignée sous le nom suivant : **tribu** ou **σ -algèbre de Boole**. Voici les propriétés exigibles de cette classe A dont un cas particulier est l'ensemble trivial de toutes les parties possibles de Ω , que l'on note usuellement $P(\Omega)$.

⁴ la notion de *classe* a été introduite par Von Neumann et Bernays dans les années 30 pour généraliser celle d'*ensemble*. Ainsi d'après le théorème de Cantor, l'objet mathématique dont tout ensemble est élément, est une *classe* et non un *ensemble*.

- l'événement certain Ω est un élément de la classe C ,
- Si l'événement A est un élément de la classe C , alors l'événement contraire «nonA» est aussi un élément de la classe C
- Si nous considérons une suite dénombrable d'événements A_i de cette classe C , alors l'événement B défini par la réunion de tous les événements A_i est encore un élément de la classe C

De ces trois axiomes, il est possible alors de déduire quelques propriétés que nous admettrons toutefois sans démonstration.

- l'événement impossible \emptyset est aussi un élément de la classe C
- Si nous considérons une suite dénombrable d'événements A_i de cette classe C , alors l'événement C défini par l'intersection de tous les événements A_i est encore un élément de la classe C
- l'événement «A et non B» = $A - B$ est un élément de la classe C
- l'événement ««A et non B» ou «nonA et B»» = $A \Delta B$ est un élément de la classe C

La suite $A_1, A_2, A_3, \dots, A_n$ d'événements forme un **système complet d'événements** si les événements sont deux à deux incompatibles et si l'événement « A_1 ou A_2 ou A_3 ou...ou A_n » est l'événement certain.

On appelle alors espace probablisable le couple $(\Omega ; C)$

34.1.1.1.Cas particuliers :

- si Ω est un ensemble fini ou infini dénombrable alors on choisit usuellement $C = P(\Omega)$

- si Ω est un ensemble infini non-dénombrable tel que un intervalle de \mathbb{R} ou $\mathbb{I}\mathbb{R}$ tout entier alors on choisit usuellement $C = \mathcal{B}$, qui est la tribu des boréliens⁵, c'est à dire la σ -algèbre de Boole engendrée par la classe des intervalles $[a ; + \infty[$ de \mathbb{R} .

34.1.2. ESPACE PROBABILISÉ :

L'idée consiste alors à associer à chaque événement un nombre compris entre 0 et 1 susceptible de représenter "son degré de réalisation". L'approche axiomatique est une façon de suspendre le débat philosophique qui est inévitablement attaché à cette problématique centrée sur le hasard. Pour ce faire, on appelle :

⁵ de nom de Emile.Borel (1871-1956), mathématicien et homme politique français, ces travaux fondamentaux ont porté sur les ensembles, les fonctions analytiques, la théorie de la mesure et la théorie des probabilités.

loi de probabilité ou probabilité sur $(\Omega ; \mathcal{C})$,

toute application P de \mathcal{C} dans $[0 ; 1]$ telle que :

- $P(\Omega) = 1$

- pour toute suite dénombrable $A_1, A_2, A_3, \dots, A_n$ d'événements incompatibles, on a

$$P(\text{«}A_1 \text{ ou } A_2 \text{ ou } A_3 \text{ ou...ou } A_n \text{»}) = P(A_1) + P(A_2) + P(A_3) + \dots + P(A_n) = \sum_{i=1}^n P(A_i)$$

- pour toute suite dénombrable $A_1, A_2, A_3, \dots, A_n, \dots$, d'événements incompatibles, on a

$$P(\text{«}A_1 \text{ ou } A_2 \text{ ou } A_3 \text{ ou...ou } A_n \text{ ou } \dots \text{»}) = \sum_{i=1}^{\infty} P(A_i)$$

De cette définition, il est possible alors de déduire quelques propriétés que nous admettrons encore sans démonstration.

34.1.3. Quelques propriétés usuelles:

(prob1) $P(\emptyset) = 0$

(prob2) $P(\text{«non}A\text{»}) = 1 - P(A)$

(prob3) Si A est inclus dans B , alors $P(A) \leq P(B)$

(prob4) $P(\text{«}A \text{ ou } B\text{»}) = P(A) + P(B) - P(\text{«}A \text{ et } B\text{»})$

(prob5) Soit un **système complet d'événements** $A_1, A_2, A_3, \dots, A_n$ alors pour tout

événement B est tel que $P(B) = \sum_{i=1}^n P(\text{«}A_i \text{ et } B\text{»})$

(prob6) Soit A et B deux événements tels que A est inclus dans B alors

$$P(B-A) = P(B) - P(A) \geq 0$$

(prob7) Si $P(A)=0$ alors A **n'est pas nécessairement** l'événement **impossible**, A est un **événement presque impossible**

(prob8) Si $P(A)=1$ alors A **n'est pas nécessairement** l'événement **certain**, A est un **événement presque certain**

On appelle alors **espace probabilisé, le triplet (Ω, \mathcal{C}, P)**

34.1.4. Produit d'espaces probabilisés :

Considérons deux espaces probabilisés dénombrables

$(\Omega_1, P(\Omega_1), P_1)$ et $(\Omega_2, P(\Omega_2), P_2)$, alors l'application P définie sur $P(\Omega_1 \times \Omega_2)$ telle

que pour tout couple d'événements $A_1 \in P(\Omega_1)$ et $A_2 \in P(\Omega_2)$ on ait :

$$p(A_1 \times A_2) = p_1(A_1) \cdot p_2(A_2)$$

est une **probabilité sur l'espace probabilisable** $(\Omega_1 \times \Omega_2; P(\Omega_1 \times \Omega_2))$.

Cette définition se généralise au produit cartésien d'un nombre fini d'univers. Soient $\Omega_1, \Omega_2, \dots, \Omega_n$ des univers dénombrables munis respectivement de probabilités p_1, p_2, \dots, p_n . Il existe une probabilité P , appelée **probabilité-produit**, et une seule sur l'ensemble-produit $\Omega = \Omega_1 \times \Omega_2 \times \dots \times \Omega_n$ telle que, pour toute suite (A_1, A_2, \dots, A_n) d'événements :

$$P(A_1 \times A_2 \times \dots \times A_n) = p_1(A_1) \cdot p_2(A_2) \cdot \dots \cdot p_n(A_n)$$

Cette notion permet de modéliser la répétition d'expériences aléatoires.

34.1.5. Lois de probabilité conditionnelle et indépendance :

Cette situation correspond à celle où l'on s'intéresse à la réalisation d'un événement A tout en prenant en compte la réalisation d'un autre événement B , tel qu'évidemment ces deux événements ne sont pas incompatibles. Pour ce faire, on appelle:

loi de probabilité conditionnelle ou **probabilité conditionnelle** de A sachant B , l'application $P(\dots \& B)$ de \mathbf{A} dans $[0 ; 1]$ telle que : $P(A \& B) = \frac{P(\text{«A et B»})}{P(B)}$

De cette définition, il est possible d'extraire une définition de l'**indépendance** de deux événements en partant du principe que la connaissance de l'événement B ne change pas les «chances» de réalisation de l'événement A .

L'événement A est indépendant de l'événement B si $P(A \& B) = P(A)$

Cela nous conduit à énoncer que :

Théorème (prob11) Deux événements A et B sont indépendants si et seulement si $P(\text{«A et B»}) = P(A) P(B)$

Définition de l'indépendance mutuelle de n événements : Les événements $A_1, A_2, A_3, \dots, A_n$ sont **mutuellement indépendants** si pour toute partie K de l'ensemble des indices

allant de 1 à n , on a $P(\bigcap A_k) = \prod P(A_k)$

Théorèmes de Bayes (sur la «probabilité des causes»)

- ◆ (Bayes1) $P(B\&A) = \frac{P(\text{«A et B»})}{P(A)} = \frac{P(A\&B) P(B)}{P(A)}$
- ◆ (Bayes2) $P(B\&A) = \frac{P(A\&B) P(B)}{P(A\&B)P(B) + P(A\&\text{«nonB»})P(\text{«nonB»})}$
- ◆ (Bayes3) $P(B_k\&A) = \frac{P(A\&B_k) P(B_k)}{\sum_{i=1}^n P(A\&B_i) P(B_i)}$ où B_1, B_2, \dots, B_n est un système complet

d'événements

34.1.6. variable aléatoire réelle et loi de probabilité

Le concept de variable aléatoire réelle formalise la notion de résultats quantitatifs issus d'une expérience aléatoire. Soit donc un **espace probabilisé**, $(\Omega ; \mathcal{C} ; P)$, **une variable aléatoire réelle X** est une application qui à tout résultat de l'univers Ω associe une valeur réelle x de \mathbb{R} de telle sorte que l'ensemble $\{ X = x \}$ des résultats ω de l'expérience aléatoire qui prennent cette valeur x soit toujours un **événement** de \mathcal{A} .

$$X : \Omega \longrightarrow \mathbb{R}$$

$$\omega \longrightarrow X(\omega) = x$$

Étudier une variable aléatoire revient d'abord à déterminer $X(\Omega)$, l'univers des résultats numériques possibles, puis l'espace probabilisable $(X(\Omega) ; \mathcal{B})$ enfin l'espace probabilisé $(X(\Omega) ; \mathcal{B} ; P_X)$ où P_X est la loi de probabilité de la variable aléatoire définie pour tout événement A de \mathcal{B} par $P_X(A) = P(\{ \omega ; \omega \in \Omega ; X(\omega) \in A \}) = P(\{ X^{-1}(A) \})$.

En particulier on convient des notations suivantes

$A = \{x\}$ on note $A = \{X=x\}$	$A =]a; b]$ on note $A = \{ a < X \leq b \}$
$A =]a; b[$ on note $A = \{ a < X < b \}$	$A = [a; +\infty[$ on note $A = \{ a \in X \}$
$A = [a; b[$ on note $A = \{ a \leq X < b \}$	$A =]-\infty ; b]$ on note $A = \{ X \leq b \}$
$A = [a; b]$ on note $A = \{ a \leq X \leq b \}$	

Si $X(\Omega)$ est fini ou infini dénombrable alors on choisit $\mathcal{B} = \mathcal{P}(X(\Omega))$ c'est à dire l'ensemble de toutes les parties de l'ensemble $X(\Omega)$

Si $X(\Omega)$ est infini non-dénombrable alors on choisit $\mathcal{B} = \mathcal{b}$, c'est à dire l'ensemble des parties construites à partir d'unions ou d'intersections d'intervalles de l'ensemble \mathbb{R} des nombres réels.

34.1.6.1. Variable aléatoires discrètes

Variable qui ne peut prendre qu'un nombre fini ou dénombrable de valeurs numériques

34.1.6.2. Variable aléatoires (absolument) continues

Variable qui peut prendre toutes les valeurs numériques d'un intervalle borné ou illimité ou même de \mathbb{R} tout entier

34.1.6.3. Fonction de répartition d'une variable aléatoire X

C'est l'application F de \mathbb{R} dans $[0 ; 1]$ définie par $F(x) = P(\{X < x\}) = P(-\infty ; x]$

Ses propriétés sont énoncées dans le théorème suivant :

La fonction de répartition F du variable aléatoire X est telle que

- F est non-décroissante
- F est continue à gauche
- $\lim_{x \rightarrow +\infty} F(x) = 1$ quand x tend vers l'infini positif
- $\lim_{x \rightarrow -\infty} F(x) = 0$ quand x tend vers l'infini négatif

et réciproquement : Toute fonction possédant les quatre propriétés précédentes peut être considérée comme la fonction de répartition d'une variable aléatoire.

Ainsi la fonction de répartition F caractérise la variable aléatoire dans le sens où elle permet de calculer la probabilité de tous les événements définis par la variable aléatoire X . En particulier nous pouvons remarquer que $P(\{a \leq X < b\}) = F(b) - F(a)$. Si la variable aléatoire est (absolument) continue alors P_X est une loi de probabilité absolument continue et il existe une fonction f , **densité de probabilité**, définie par

$$F'(x) = f(x) \quad \text{et} \quad F(x) = P(\{X < x\}) = \int_{-\infty}^x f(t) dt$$

$$\text{ainsi on peut obtenir} \quad P(\{a \leq X < b\}) = \int_a^b f(t) dt$$

La représentation graphique de la fonction **densité de probabilité de la variable aléatoire X** est l'**histogramme** de la distribution des probabilités des résultats de la variable X . On peut aussi remarquer que $P(\{X=x\}) = 0$ pour tout x et même que tout événement D réalisable par un nombre dénombrable de résultats est tel que $P(D)=0$. Ceci est même à considérer comme une caractéristique d'une variable aléatoire absolument continue.

34.1.7. Indépendance de deux variables aléatoires

Soient X et Y deux variables aléatoires réelles définies sur l'espace probabilisé $(\Omega; \mathcal{C}; P)$. Le couple (X, Y) est une application de Ω dans \mathbb{R}^2 muni de la tribu borélienne,

c'est à dire la σ -algèbre engendrée par la classe des pavés (les rectangles) ouverts bornés de \mathbb{R}^2 . X et Y sont indépendantes si pour tout couple A et B de boréliens,

$$P(\{X^{-1}(A)\} \text{ et } \{Y^{-1}(B)\}) = P(\{X^{-1}(A)\}) P(\{Y^{-1}(B)\})$$

X et Y sont indépendantes si et seulement si la fonction de répartition du couple (X,Y) définie par $H(x,y) = P(\{X < x\} \text{ et } \{Y > y\}) = F(x)G(y)$ où F est la fonction de répartition marginale de X et G est la fonction de répartition marginale de Y. Si X et Y admettent des densités f et g alors le couple (X,Y) admet pour densité $h(x,y) = f(x)g(y)$

34.1.8. Valeurs caractéristiques d'une variable aléatoire

Soit X une variable aléatoire réelle sur l'espace probabilisé $(\Omega ; \mathcal{C} ; P)$

34.1.8.1. Espérance mathématique de X

- X variable aléatoire discrète finie :
$$E(X) = \sum_{i=1}^n P(\{X=x_i\})x_i$$
- X variable aléatoire discrète infinie :
$$E(X) = \sum_{i=1} P(\{X=x_i\})x_i \text{ (si elle existe)}$$
- X variable aléatoire continue :
$$E(X) = \int xf(x)dx \text{ (si elle existe)}$$

34.1.8.2. Propriétés de l'espérance mathématique:

- Si X est une variable aléatoire qui ne prend que la valeur c alors $E(X) = c$
- $E(aX + b) = aE(X) + b$
- Si nous considérons une suite de variables X_1, X_2, \dots, X_n définies sur l'espace probabilisé $(\Omega ; \mathcal{C} ; P)$ alors la nouvelle variable aléatoire $S_n = X_1 + X_2 + \dots + X_n$, somme des variables aléatoires est telle que $E(S_n) = E(X_1) + E(X_2) + \dots + E(X_n)$

- Si nous considérons les deux variables X et Y **indépendantes** définies sur l'espace probabilisé $(\Omega ; \mathcal{C} ; P)$ alors la nouvelle variable produit XY est telle que $E(XY) = E(X)E(Y)$.

Attention, la réciproque est fautive!

34.1.8.3. Variance de X dont l'espérance est $E(X) = m$

$$\sigma^2 = V(X) = E((X-m)^2) \text{ (si elle existe)}$$

34.1.8.4. écart-type de X

$$\sigma = \sqrt{V(X)}$$

34.1.8.5. Propriétés de la variance:

- Si X est une variable aléatoire qui ne prend que la valeur c alors $V(X) = 0$
- $V(aX + b) = a^2V(X)$
- $V(X) = E(X^2) - (E(X))^2$

- Si nous considérons deux variables X, Y définies sur l'espace probabilisé $(\Omega; \mathcal{C}; P)$ alors la nouvelle variable aléatoire $S = X + Y$, somme des variables aléatoires est telle que $V(X+Y) = V(X) + V(Y) + 2 \text{coV}(X, Y)$ où la covariance de X et Y est définie par

$$\text{coV}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

- Si nous considérons les deux variables X et Y **indépendantes** définies sur l'espace probabilisé $(\Omega; \mathcal{A}; P)$ alors la nouvelle variable somme $X + Y$ est telle que

$$V(X+Y) = V(X) + V(Y) \text{ . Attention, la réciproque est fausse!}$$

- $V(X) = 0$ si et seulement si $X = c$ presque sûrement

34.1.8.6. Lien entre l'espérance et l'écart-type : l'inégalité de Bienaymé⁶-Tchebychev⁷

Pour tout nombre k réel positif $P(\{|X - E(X)| > k\sigma\}) \leq \frac{1}{k^2}$

34.1.8.7. Variable centrée réduite

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - E(X)}{\sigma} \text{ et } E(Z) = 0, V(Z) = 1, \sigma = 1$$

34.1.8.8. Moments centrés d'ordre k de X

$$\mu_k = \mu_k(X) = E[(X - E(X))^k]$$

Coefficients de Fisher	Coefficients de Pearson
$\gamma_1 = \frac{\mu_3}{\sigma^3} = E(Z^3) = \mu_3(Z)$	$\beta_1 = \left(\frac{\mu_3}{\sigma^3}\right)^2 = [E(Z^3)]^2 = [\mu_3(Z)]^2$
$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = E(Z^4) - 3 = \mu_4(Z) - 3$	$\beta_2 = \frac{\mu_4}{\sigma^4} = E(Z^4) = \mu_4(Z)$

34.1.8.9. Médiane(s) d'une variable aléatoire

Soit X une variable aléatoire réelles définie sur l'espace probabilisé $(\Omega; \mathcal{A}; P)$

On appelle **médiane** tout nombre m_d vérifiant $P(\{X < m_d\}) \leq 0,5$ et $P(\{X > m_d\}) \leq 0,5$

34.1.8.10. Quartiles d'une variable aléatoire

Soit X une variable aléatoire réelles définie sur l'espace probabilisé $(\Omega; \mathcal{A}; P)$

On appelle **premier quartile** tout nombre Q_1 vérifiant

$$P(\{X < Q_1\}) \leq 0,25 \text{ et } P(\{X > Q_1\}) \leq 0,75$$

On appelle **second quartile** (= médiane) tout nombre Q_2 vérifiant

$$P(\{X < Q_2\}) \leq 0,5 \text{ et } P(\{X > Q_2\}) \leq 0,5$$

On appelle **troisième quartile** tout nombre Q_3 vérifiant

⁶ Jules Bienaymé, (1796-1878), statisticien et administrateur français

⁷ Pafnoutiy Lvovitch Tchebichev, (1821-1894), mathématicien russe dont les travaux ont porté sur les nombres premiers, les formes quadratiques, les fonctions orthogonales, l'approximation des fonctions continues par des polynômes et sur le calcul des probabilités.

$$P(\{X < Q_3\}) \leq 0,75 \text{ et } P(\{X > Q_3\}) \leq 0,25$$

34.1.9. Égalité presque sûre de deux variables aléatoires

Soient X et Y deux variables aléatoires réelles définies sur l'espace probabilisé $(\Omega; \mathcal{C}; P)$. $X = Y$ presque sûrement si $P(\{\omega, \omega \in \Omega, X(\omega) \neq Y(\omega)\}) = 0$, c'est à dire que l'ensemble des résultats de Ω dont la valeur $X(\omega)$ est différente de la valeur $Y(\omega)$ est un ensemble négligeable⁸,

34.2. Variables aléatoires et lois de probabilité usuelles

Soit X une variable aléatoire réelle sur l'espace probabilisé $(\Omega; \mathcal{A}; P)$

nom de la variable	uniforme discrète	de Bernoulli $B(1;p)$
$X(\Omega)$	$\{0,1,2,\dots,n\}$	$\{0,1\}$
loi de probabilité	$P(\{X=k\}) = \frac{1}{n}$	$P(\{X=0\}) = 1-p$ $P(\{X=1\}) = p$
espérance $E(X)$	$\frac{n+1}{2}$	p
variance $V(X)$	$\frac{n^2-1}{12}$	$p(1-p)$
coefficient d'asymétrie de Fisher γ_1	0	
coefficient d'aplatissement de Fisher γ_2	$1,8 - \frac{2,4}{n^2-1} - 3 = -1,2 \frac{n^2+1}{n^2-1}$	

⁸ comme le point est négligeable pour la longueur d'un segment.

<i>nom de la variable</i>	binomiale B(n;p)	hypergéométrique H(N, n, p)
X(Ω)	{0,1,2,...,n}	{0,1,2,...,n}
loi de probabilité	$P(\{X=k\}) = \frac{n!}{k!(n-k)!} p^k(1-p)^{n-k}$	$P(\{X=k\}) = \frac{C_{Np}^k C_{N-Np}^{n-k}}{C_N^n}$
espérance E(X)	np	np
variance V(X)	np(1-p)	$\frac{N-n}{N-1} np(1-p)$
coefficient d'asymétrie de Fisher γ_1	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-2p}{\sqrt{np(1-p)}} \frac{N-2n}{N-2} \sqrt{\frac{N-1}{N-n}}$
coefficient d'aplatissement de Fisher γ_2	$\frac{1-6p(1-p)}{np(1-p)}$	$-3 + 3 \frac{(N-1)(N+6)}{(N-2)(N-3)} + \frac{(N+1)N(N-1)}{[(N-2)(N-3)(N-n)]} \left[\frac{1}{np(1-p)} \right] \left[1 - 6 \frac{N}{N+1} (p(1-p) + \frac{n(n-N)}{N^2}) \right]$

<i>nom de la variable</i>	de Poisson P(λ)	(vectorielle à m composantes) multinomiale M(n,p₁,p₂,...,p_m)
X(Ω)	N = {0,1, 2, ...}	{0,1,2,...,n} ^m
loi de probabilité	$P(\{X=k\}) = \exp(-\lambda) \frac{\lambda^k}{k!}$	$P(\{X=(k_1,k_2,\dots,k_m)\}) = \frac{n!}{k_1! \dots k_m!} (p_1)^{k_1} \dots (p_m)^{k_m}$
espérance E(X)	λ	
variance V(X)	λ	
coefficient d'asymétrie de Fisher γ_1	$\frac{1}{\sqrt{\lambda}}$	
coefficient d'aplatissement de Fisher γ_2	$\frac{1}{\lambda}$	

<i>nom de la variable</i>	uniforme continue	de Laplace-Gauss LG(0,1)
X(Ω)	[a; b]	R
loi de probabilité	$f(x) = \frac{1}{b-a} \mathbb{1}_{[a; b]}(x)$ $F(x) = \frac{x-a}{b-a}$ sur [a;b] $F(x) = 0$ sur]- ;a[$F(x) = 1$ sur]b; + [$f(x) = \frac{1}{\sqrt{2}} \exp(-\frac{x^2}{2})$ $F(x) = \int_{-\infty}^x f(t)dt$
espérance E(X)	$\frac{a+b}{2}$	0
variance V(X)	$\frac{(a-b)^2}{12}$	1
coefficient d'asymétrie de Fisher γ_1	0	0
coefficient d'aplatissement de Fisher γ_2	$\beta_2 = \frac{\mu_4}{\sigma^4} = 1,8$ $\gamma_2 = 1,8 - 3 = -1,2$	$\beta_2 = \frac{\mu_4}{\sigma^4} = 3$ $\gamma_2 = 3 - 3 = 0$

<i>nom de la variable</i>	de Laplace-Gauss LG(m,σ)	de Laplace-Gauss de dimension n LG(M,Σ)
X(Ω)	R	R ⁿ
loi de probabilité	$f(x) = \frac{1}{\sigma\sqrt{2}} \exp[-\frac{1}{2}(\frac{x-m}{\sigma})^2]$ $F(x) = \int_{-\infty}^x f(t)dt$	
espérance E(X)	m	
variance V(X)	σ	
coefficient d'asymétrie de Fisher γ_1	0	
coefficient d'aplatissement de Fisher γ_2	$\beta_2 = \frac{\mu_4}{\sigma^4} = 3$ $\gamma_2 = 3 - 3 = 0$	

nom de la variable	de "Student" ⁹ à ddl ¹⁰ n T_n	de Pearson ¹¹ à ddl n χ^2_n (Khi-deux)
$X(\Omega)$	R	$[0; +\infty[$
loi de probabilité	$f(x) = \frac{1}{\sqrt{n} B(\frac{1}{2}, \frac{n}{2})} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$ $F(x) = \int_{-\infty}^x f(t) dt$	$f(x) = \frac{1}{2^{n/2} \Gamma(\frac{n}{2})} \exp(-\frac{x}{2}) x^{\frac{n-2}{2}}$ $F(x) = \int_{-\infty}^x f(t) dt$
espérance $E(X)$	0	n
variance $V(X)$	$\frac{n}{n-2}$ si $n > 2$	$2n$
coefficient d'asymétrie de Fisher γ_1	0	$\sqrt{\frac{8}{n}}$
coefficient d'aplatissement de Fisher γ_2	$\frac{6}{n-4}$	$\frac{12}{n}$

nom de la variable	de Fisher ¹² -Snédécour $F(m;n)$	
$X(\Omega)$	$[0; +\infty[$	
loi de probabilité	$f(x) = \frac{\left(\frac{m}{n}\right)^{n/2}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} x^{(m/2)-1}$ $\left[1 + \frac{m}{n} x\right]^{-(m+n)/2}$ $F(x) = \int_{-\infty}^x f(t) dt$	
espérance $E(X)$	$\frac{n}{n-1}$	
variance $V(X)$	$2 \frac{n^2}{m} \frac{m+n-2}{(n-2)^2(n-4)}$	
coefficient d'asymétrie de Fisher γ_1	$\sqrt{\frac{8(n-4)}{m(m+n-2)}} \frac{2m+n-2}{n-6}$	
coefficient d'aplatissement de Fisher γ_2	en posant $p = m+n-2$ $\frac{12(n-2)^2(n-4) + (5n-22)mp}{m(n-6)(n-8)p}$	

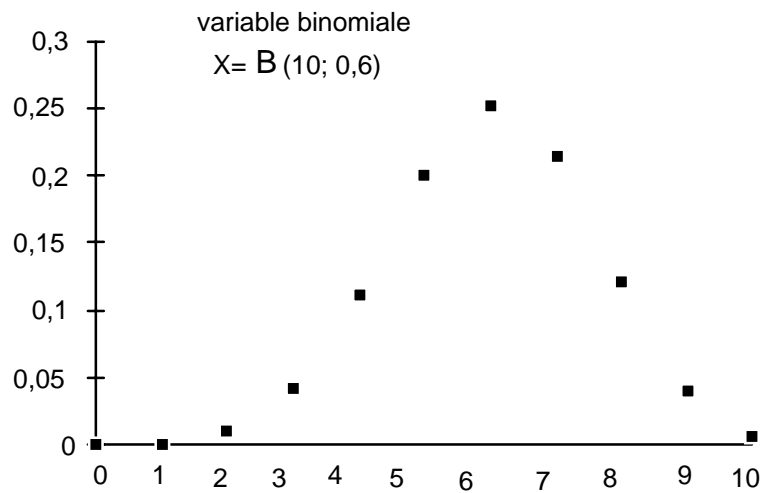
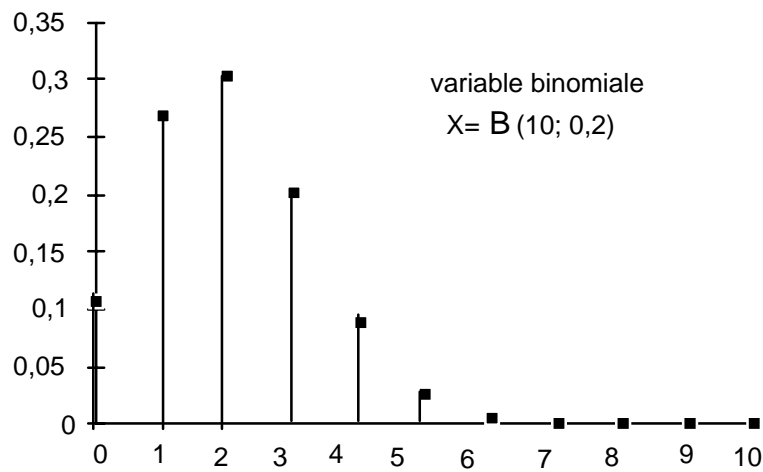
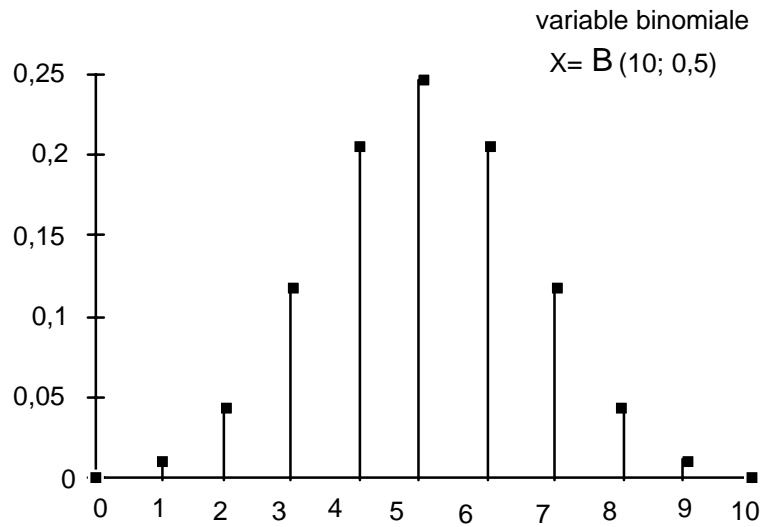
⁹ pseudonyme employé par William Sealy Gosset (Canterbury 1876; Londres 1937) statisticien dont les travaux ont contribué fortement au développement de la statistique. Il conduisit toute sa vie ses travaux au sein de la célèbre brasserie Guinness à Dublin, puis à Londres. Ce sont en particulier des problèmes liés à la fabrication de la bière qui le conduisirent à élaborer des outils et des techniques statistiques dont nous faisons usage.

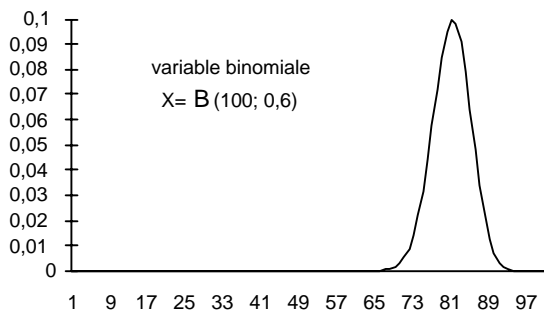
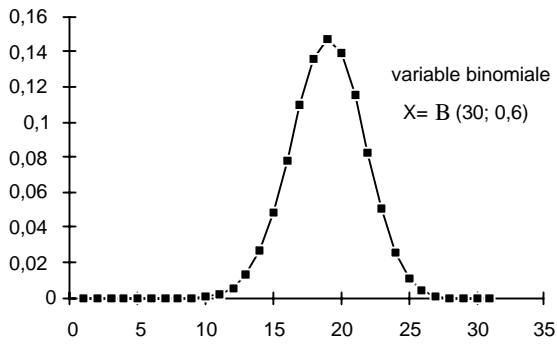
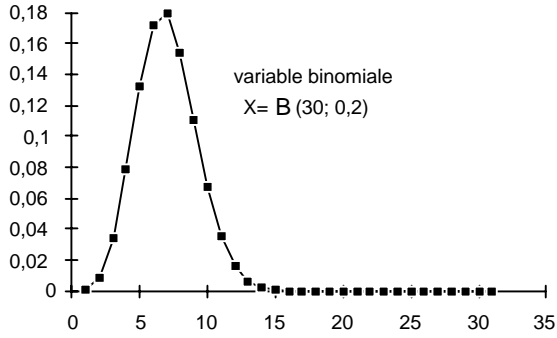
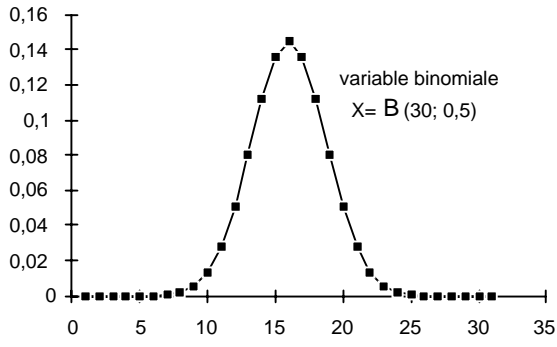
¹⁰ ddl = degré de liberté

¹¹ Karl Pearson (1857-1936) statisticien britannique dont contribua considérablement à l'évolution de la statistique. Son fils Egon Sharpe Pearson (1895-1980) fut aussi un éminent statisticien.

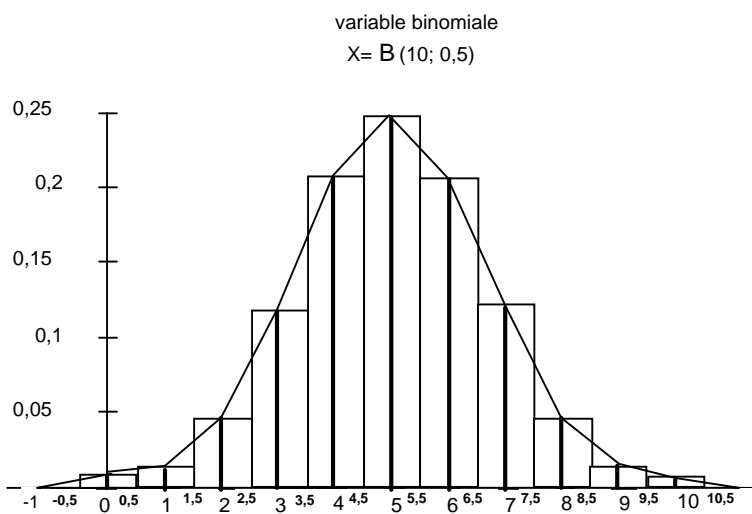
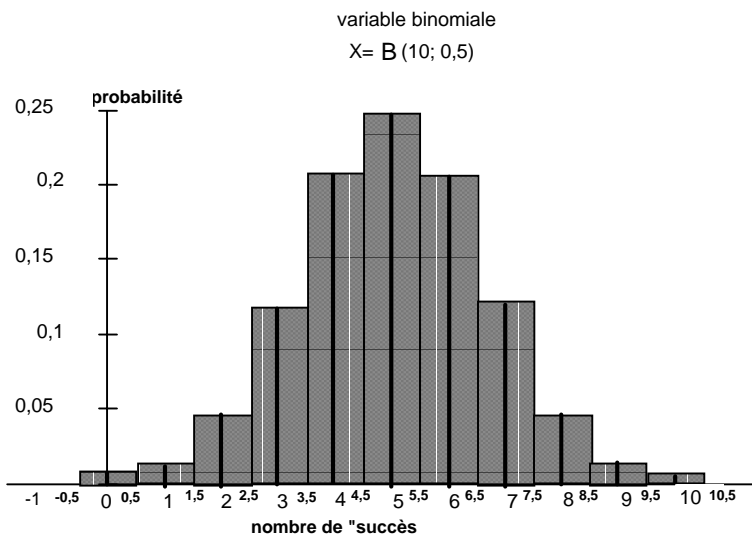
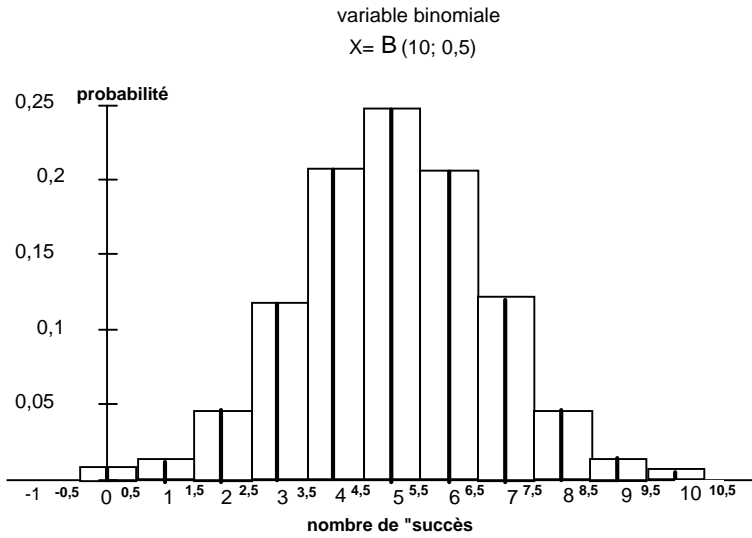
¹² Ronald Aylmer Fisher (1890, 1962), statisticien britannique, son œuvre abondante contribua à faire de la statistique une science moderne.

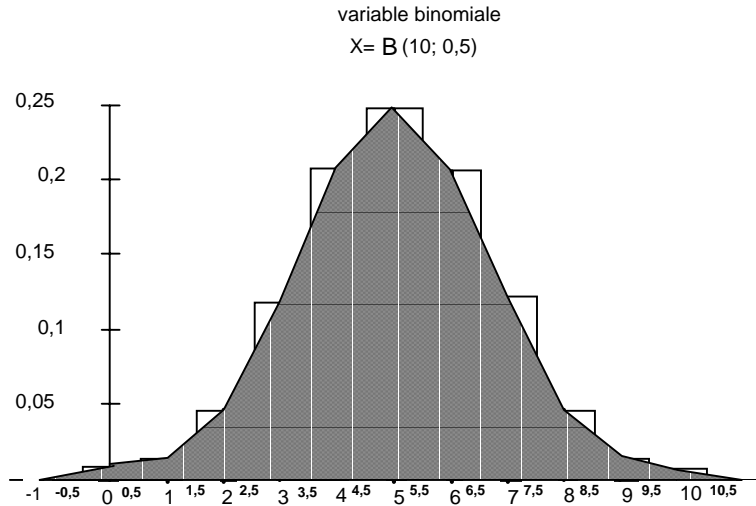
34.3. Variable aléatoire binomiale et approximation de sa distribution de probabilité par celle de la variable de Laplace-Gauss.





34.4. Prolongement de la variable binomiale discrète à valeurs entières dans N à une variable continue à valeurs réelles dans R .





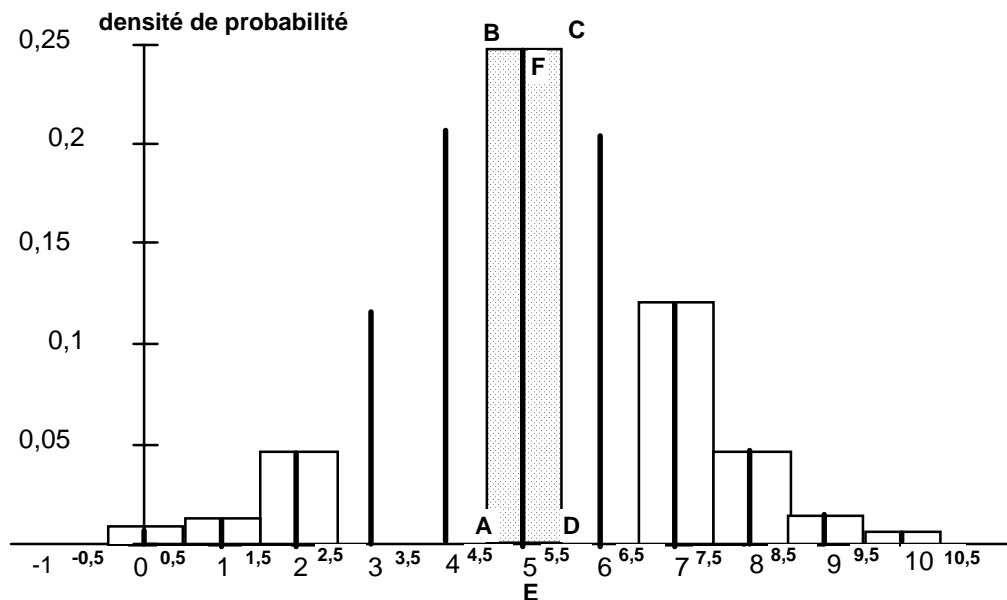
Les deux graphiques ci-dessus peuvent représenter deux points de vue :

- le point de vue "variable discrète" et son diagramme en bâtons
- le point de vue "variable continue" et son histogramme.

La démarche consiste à utiliser la propriété suivante valable d'ailleurs pour toutes variables discrètes à valeurs entières : l'ensemble $\{ X = k \} = \{ k - \frac{1}{2} \leq X < k + \frac{1}{2} \} = [k - \frac{1}{2} ; k + \frac{1}{2} [$

$$\text{Prob} (\{ X = k \}) = \text{Prob} (\{ k - \frac{1}{2} \leq X < k + \frac{1}{2} \}) = \frac{\text{aire du rectangle ABCD}}{\text{aire totale}}$$

Ceci conduit alors à un changement de point de vue. Les "bâtons" ne sont plus des segments mais des rectangles dégénérés, des rectangles de largeur nulle. Le diagramme en bâtons devient un histogramme.



La représentation graphique EF qui était initialement un *segment* dont la hauteur représentait la probabilité d'apparition de la valeur 5 , c'est à dire celle d'obtenir 5 succès en répétant 10 fois l'expérience à deux issues, devient un rectangle dont la largeur est nulle et la hauteur représente la densité de probabilité de cet événement.

Intuitivement, si l'on considère que la *courbe polygonale* suggère une approximation de la courbe de densité d'une variable de Laplace-Gauss $Y = LG(m, \sigma)$, cela revient alors à effectuer les calculs selon la procédure suivante :

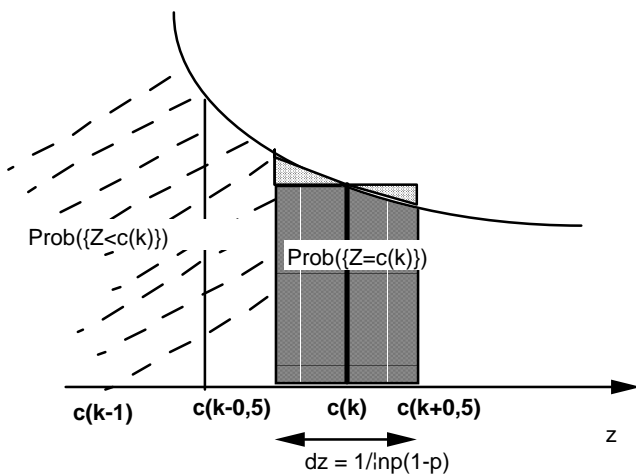
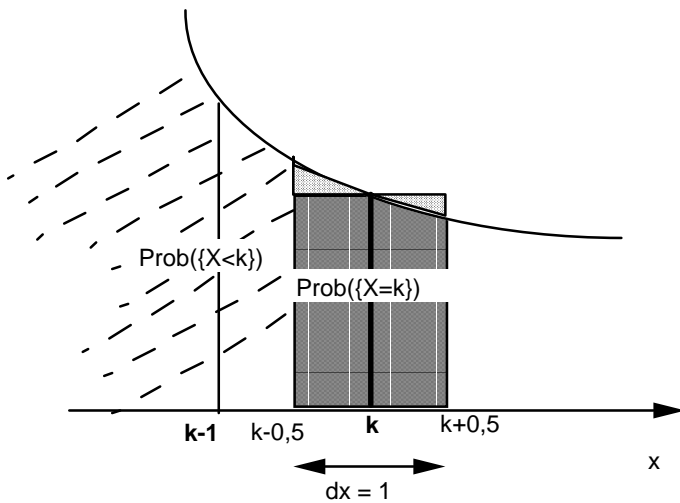
- il est tout d'abord logique de contraindre les deux variables X et Y a avoir leurs espérances et leurs variances respectivement égales :

$$E(X) = np = E(Y) = m$$

$$V(X) = np(1-p) = E(Y) = \sigma^2$$

$$\sigma_X = \sqrt{np(1-p)} = \sigma$$

- ensuite la loi de probabilité de la variable centrée réduite $Z = \frac{X-np}{\sqrt{np(1-p)}}$ est alors approchée par celle de $Z = LG(0;1)$. De là nous pouvons obtenir une première approximation :



$$Z = \frac{X-np}{\sqrt{np(1-p)}}$$

$$c(k) = \frac{k-np}{\sqrt{np(1-p)}}$$

$$c(k-0,5) = \frac{k-np-0,5}{\sqrt{np(1-p)}}$$

$$c(k+0,5) = \frac{k-np+0,5}{\sqrt{np(1-p)}}$$

$$\text{Prob} (\{ X = k \}) = \text{Prob} (\{ \frac{X-np}{\sqrt{np(1-p)}} = \frac{k-np}{\sqrt{np(1-p)}} \}) =$$

$$\text{Prob} (\{ k - \frac{1}{2} \leq X < k + \frac{1}{2} \}) =$$

$$\text{Prob} (\{ \frac{k-np-\frac{1}{2}}{\sqrt{np(1-p)}} \leq \frac{X-np}{\sqrt{np(1-p)}} < \frac{k-np+\frac{1}{2}}{\sqrt{np(1-p)}} \})$$

$$\left(\frac{k-np+\frac{1}{2}}{\sqrt{np(1-p)}} - \frac{k-np-\frac{1}{2}}{\sqrt{np(1-p)}} \right) f\left(\frac{k-np}{\sqrt{np(1-p)}} \right) = \frac{1}{\sqrt{np(1-p)}} f\left(\frac{k-np}{\sqrt{np(1-p)}} \right)$$

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

La seconde approximation concerne la fonction de répartition :

$$\text{Prob}(\{X < k\}) = \text{Prob}(\{0 \leq X < k\}) =$$

$$\text{Prob}(\{0 \leq X \leq k-1\}) = \sum_{i=0}^{i=k-1} \text{Prob}(\{X=i\}) = \text{Prob}\left(\left\{0 - \frac{1}{2} \leq X < k-1 + \frac{1}{2}\right\}\right)$$

$$\text{Prob}\left(\left\{-\frac{1}{2} \leq X < k - \frac{1}{2}\right\}\right) = \text{Prob}\left(\left\{\frac{-np-\frac{1}{2}}{\sqrt{np(1-p)}} \leq \frac{X-np}{\sqrt{np(1-p)}} < \frac{k-np-\frac{1}{2}}{\sqrt{np(1-p)}}\right\}\right)$$

$$\text{Prob}(\{X < k\}) \approx F\left(\frac{k-np-\frac{1}{2}}{\sqrt{np(1-p)}}\right) - F\left(\frac{-np-\frac{1}{2}}{\sqrt{np(1-p)}}\right) \text{ où } F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt$$

Cette démarche est justifiée par des propriétés démontrées dans la théorie des probabilités. Elle est applicable à d'autres variables discrètes entières que la variable binomiale $B(n,p)$. Ainsi en statistique de rang, plusieurs variables entières sont utilisées pour estimer un paramètre ou tester une hypothèse, citons pour exemplifier notre remarque : la statistique W_X de Wilcoxon, la statistique S de Spearman ou celle de Kendall.

L'opération consistant à introduire le facteur $\pm \frac{1}{2}$ dans les calculs est souvent dénommée " *correction de continuité*".

34.4.1. Quelques relations exactes entre les distributions usuelles :

34.4.1.1. *Loi de probabilité de la variable de Poisson et loi de probabilité de la variable de Fisher-Snédecor :*

$$\text{Soit } X = P(\lambda) \text{ et } Y = \chi^2(2(k+1)) \text{ et } \text{Prob}(\{X \leq k\}) = \text{Prob}(\{Y > 2\lambda\})$$

34.4.1.2. *Loi de probabilité de la variable binomiale et loi de probabilité de la variable de Fisher-Snédecor :*

$$\text{Soit } X = B(n,p) \text{ et } Y = F(2(k+1), 2(n-k))$$

$$\text{Prob}(\{X \leq k\}) = \text{Prob}\left(\left\{Y > \frac{n-k}{k+1} \frac{p}{1-p}\right\}\right)$$

34.4.1.3. *Loi de probabilité de la variable T_n de Student et loi de probabilité de la variable de Fisher-Snédecor :*

$$T_n^2 = F(1,n)$$

34.4.2. Quelques approximations des distributions usuelles :

34.4.2.1. Convergence en loi de la variable binomiale $B(n,p)$ vers la variable de Laplace-Gauss

La suite (X_n) étant une suite de variables binomiales $B(n,p)$ alors la variable

$$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}} \text{ tend vers } Z = \text{LG}(0;1)$$

34.4.2.2. Convergence en loi de la variable de Poisson $P(m)$ vers la variable de Laplace-Gauss

La famille (X_m) étant une famille de variables de Poisson $P(m)$ alors la variable

$$Z_n = \frac{X_n - m}{\sqrt{m}} \text{ tend vers } Z = \text{LG}(0;1)$$

34.4.2.3. Convergence en loi de la variable binomiale $B(n,p)$ vers la variable de Poisson $P(\lambda)$

La suite (X_n) étant une suite de variables binomiales $B(n,p)$ telles que $n \rightarrow \infty$ et $p \rightarrow 0$ de manière à ce que le produit np tende vers λ . Alors la suite (X_n) converge en loi vers une variable de Poisson $P(\lambda)$.

34.4.2.4. Convergence en loi de la variable hypergéométrique $H(N,n,p)$ vers la variable binomiale $B(n,p)$

Si $N \rightarrow \infty$ alors $H(N,n,p) \rightarrow B(n,p)$

34.4.3. Formules approchées de fonction de répartition de quelques variables continues usuelles :

34.4.3.1. Variable de Pearson : variable du $\chi^2(n)$

La formule de Wilson-Hilferty donne deux décimales exactes dès que $n > 2$

$$F(x) = \text{Prob}(\{\chi^2(n) < x\}) \approx \text{Prob}(\{\text{LG}(0;1) < (\frac{9n}{2})^{\frac{1}{2}} \left((\frac{x}{n})^{\frac{1}{3}} + \frac{2}{9n} - 1 \right) \})$$

34.4.3.2. Variable de Fisher-Snedecor : variable $F(m,n)$

La formule de Paulson donne deux décimales exactes dès que $n > 3$

$$\text{posons } A(x) = x^{\frac{1}{3}} \left(1 - \frac{2}{9n} \right) + \frac{2}{9m} - 1 \text{ et } B(x) = \frac{2}{9m} + x^{\frac{2}{3}} \frac{2}{9n}$$

$$F(x) = \text{Prob}(\{F(m,n) < x\}) \approx \text{Prob}(\{\text{LG}(0;1) < \frac{A(x)}{\sqrt{B(x)}}\})$$

34.4.3.3. Variable de "Student" : variable T_n

La formule se déduit de celle de l'approximation concernant la variable de Fisher-Snedecor :

$$A(x) = x^{\frac{2}{3}} \left(1 - \frac{2}{9n} \right) - \frac{7}{9} \quad \text{et} \quad B(x) = \frac{2}{9} + x^{\frac{4}{3}} \frac{2}{9n}$$

$$\text{Prob}(\{ |T_n| > x \}) \approx \text{Prob}(\{ \text{LG}(0;1) > \frac{A(x)}{\sqrt{B(x)}} \})$$

ECHANTILLONNAGE

35. Échantillonnage d'une variable

Soit une variable X définie sur l'espace probabilisé $(\Omega ; \mathcal{C} ; P)$

35.1. *n*-échantillon de la variable X

On désigne par ***n*-échantillon de la variable X** , la suite de variables X_1, X_2, \dots, X_n définies sur l'espace probabilisé $(\Omega ; \mathcal{C} ; P)$, indépendantes et identiquement distribuées, de même loi que la variable X .

35.2. réalisation d'un *n*-échantillon de la variable X

On désigne par **réalisation**, la suite des nombres x_1, x_2, \dots, x_n qui sont les résultats de l'expérience correspondant respectivement à la valeur obtenue par les variables X_1, X_2, \dots, X_n . Ainsi le nombre x_i est le résultat obtenu au tirage $n^{\circ}i$ ou encore le résultat de la variable X_i . Cette réalisation peut être obtenue par tirage avec ou sans remise, en tenant compte ou non de l'ordre d'obtention des résultats. Une **théorie de l'échantillonnage** est une théorie dont l'objet est l'étude des propriétés des *n*-échantillons, des caractéristiques qui peuvent résumer certaines propriétés, en liaison avec la distribution des probabilités de la variable X , nommée variable parente, l'étude des comportements des propriétés en fonction de la valeur n , nommée taille de l'échantillon

35.3. statistique et loi d'échantillonnage

On désigne aussi par **statistique**, une variable aléatoire T_n construite à partir du *n*-échantillon de la variable X . Cette variable est caractérisée par une loi de probabilité nommée **loi ou distribution d'échantillonnage**. La connaissance de cette loi de probabilité dépend de celle de la **variable-parente X** . Cette connaissance peut être obtenue soit directement soit par approximation selon les cas et les informations dont on dispose.

36. quelques statistiques usuelles

Soit un ***n*-échantillon X_1, X_2, \dots, X_n de la variable X** définie sur l'espace probabilisé $(\Omega ; \mathcal{C} ; P)$ dont les caractéristiques usuelles sont:

espérance $E(X) = m$ variance $V(X) = \sigma^2$	moment centré d'ordre k $M_k(X) = \mu_k$ coefficient d'asymétrie $\gamma_1(X) = \gamma_1$
--	--

écart-type $\sqrt{V(X)} = \sigma$	coefficient d'aplatissement $\gamma_2(X) = \gamma_2$
-----------------------------------	--

36.1. La variable aléatoire «moyenne empirique» est définie par

$$T_n = X_n = \frac{1}{n} \sum_{i=1}^n X_i$$

espérance $E(T_n)$	variance $V(T_n)$	écart-type $\sqrt{V(T_n)}$	$\gamma_1(T_n)$	$\gamma_2(T_n)$	tirage
m	$\frac{\sigma^2}{n}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{\gamma_1}{\sqrt{n}}$	$\frac{\gamma_2-3}{\sqrt{n}}$	remise
m	$\frac{N-n}{N-1} \frac{\sigma^2}{n}$	$\sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}$			exhaustif

36.2. La variable aléatoire «variance empirique» est définie par

$$T_n = S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - X_n)^2$$

espérance $E(T_n)$	variance $V(T_n)$	écart-type $\sqrt{V(T_n)}$	$\gamma_1(T_n)$	$\gamma_2(T_n)$	tirage
$\frac{n-1}{n} \sigma^2$	$\frac{n-1}{n^3} [(n-1)\mu_4 - (n-3)\sigma^4]$				remise
$\frac{n-1}{n} \frac{N}{N-1} \sigma^2$					exhaustif

36.3. La variable aléatoire «variance empirique modifiée» est définie par

$$T_n = S_n'^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - X_n)^2$$

espérance $E(T_n)$	variance $V(T_n)$	écart-type $\sqrt{V(T_n)}$	$\gamma_1(T_n)$	$\gamma_2(T_n)$	tirage
σ^2					remise
$\frac{N}{N-1} \sigma^2$					exhaustif

36.4. La variable aléatoire «variance empirique modifiée» est définie par

$$T_n = S_n''^2 = \frac{n(N-1)}{N(n-1)} \sum_{i=1}^n (X_i - X_n)^2$$

espérance $E(T_n)$	variance $V(T_n)$	écart-type $\sqrt{V(T_n)}$	$\gamma_1(T_n)$	$\gamma_2(T_n)$	tirage
σ^2					exhaustif

36.5. La variable aléatoire «fréquence empirique» est définie par

$$T_n = F_n = \frac{1}{n} \sum_{i=1}^n X_i$$

où la suite (X_n) est un n-échantillon de la variable $X = B(1;p)$, variable de Bernoulli.

espérance $E(T_n)$	variance $V(T_n)$	écart-type $\sqrt{V(T_n)}$	$\gamma_1(T_n)$	$\gamma_2(T_n)$	tirage
p	$\frac{p(1-p)}{n}$	$\sqrt{\frac{p(1-p)}{n}}$			remise
p	$\frac{N-n}{N-1} \frac{p(1-p)}{n}$	$\sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}}$			exhaustif

36.6. La variable aléatoire «fonction de répartition empirique»

Elle est définie pour chaque valeur x réelle par $T_n = F_n^*(x)$ = la proportion des réalisations du n-échantillon X_1, X_2, \dots, X_n de la variable X , qui sont **inférieures strictement à** x .

$$T_n = F_n^*(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{]-\infty, x[}(x_i)$$

La variable $\mathbf{1}_{]-\infty, x[}$ est une variable de Bernoulli $B(1; F(x))$

La variable $nF_n^*(x)$ est donc une variable binomiale $B(n; F(x))$

Si les n réalisations sont ordonnées par valeurs croissantes de x_1 à x_n , alors nous avons:

$$F_n^*(x) = 0 \quad \text{pour tout } x < x_1$$

$$F_n^*(x) = \frac{i-1}{n} \quad \text{pour tout } x_{i-1} < x < x_i$$

$$F_n^*(x) = 1 \quad \text{pour tout } x \geq x_n$$

Soit $F(x) = P(\{X < x\})$ la fonction de répartition de la variable X .

On peut remarquer que pour chaque x , la variable $F_n^*(x)$ est une variable binomiale $B(n; F(x))$

espérance $E(T_n)$	variance $V(T_n)$	écart-type $\sqrt{V(T_n)}$	$\gamma_1(T_n)$	$\gamma_2(T_n)$	tirage
$F(x)$	$\frac{F(x)(1-F(x))}{n}$	$\sqrt{\frac{F(x)(1-F(x))}{n}}$			remise
$F(x)$	$\frac{N-n}{N-1} \frac{F(x)(1-F(x))}{n}$	$\sqrt{\frac{N-n}{N-1}} \sqrt{\frac{F(x)(1-F(x))}{n}}$			exhaustif

36.7. La variable aléatoire «coefficient de corrélation empirique»

Soit un n-échantillon $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, du couple de variables de Laplace

Gauss (X,Y) indépendantes défini sur l'espace probabilisé $(\Omega; A; P)$ à valeurs dans \mathbb{R}^2

Soit la variable aléatoire «**coefficient de corrélation empirique**» définie par

$$T_n = R_{BP} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - X)(Y_i - Y)}{\sqrt{S^{(x)}_n} \sqrt{S^{(y)}_n}}$$

$$T_n = \frac{R_{BP} \sqrt{n-2}}{\sqrt{1-R_{BP}^2}}$$
 est une variable de "Student" à n-2 ddl

espérance $E(T_n)$	variance $V(T_n)$	écart-type $\sqrt{V(T_n)}$	$\gamma_1(T_n)$	$\gamma_2(T_n)$	tirage
0	$\frac{1}{n-1}$				remise
					exhaustif

TABLES STATISTIQUES :

1. **TABLE DE NOMBRES AU HASARD**

2. VARIABLE ALEATOIRE de PEARSON de type VII

variable T_n de STUDENT:

3. **VARIABLE ALEATOIRE de PEARSON de type III**

Variable χ^2 (Khi deux) :

La table a été obtenue par Jean-Claude Régnier à partir de la fonction KHI DEUX.INVERSE du logiciel Microsoft Excel 5. Elle fournit pour 11 valeurs particulières de probabilité α , une valeur approchée de la valeur k de la variable telle que $\text{Prob}(\chi^2 > k) = \alpha$.

4. Complément relatif à la variable aléatoire de Laplace-Gauss $Z = LG(0;1)$

Ce document rapporte l'extrait d'une table que nous avons construite pour donner les valeurs numériques de la fonction densité de probabilité et de la fonction de répartition de la variable Z pour des valeurs supérieures à 0.

La fonction densité est définie par la relation

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

La fonction de répartition est définie par :

$$F(x) = \text{Prob}(\{Z < x\}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt$$

Le calcul de F(x) est réalisé à partir d'une fonction approchée .

Nous avons choisi l'expression fournie dans l'ouvrage "Théorie des probabilités en vue des applications statistiques" de Tassi et Legait (1990) page 126 et page 329.

Pour x tel que $0 < x < 4$

$$F(x) \approx 1 - [au + bu^2 + cu^3] f(x)$$

avec $a = 0,4361836$

$b = -0,1201676$

$c = 0,9372980$

$x > 0$

$$u = \frac{1}{1 + 0,33267x}$$

qui donne une approximation de l'ordre de 10^{-5} .

Pour x tel que $4 < x$

$$F(x) \approx 1 - \left(1 - \frac{1}{x^2} + \frac{3}{x^4} - \frac{15}{x^6} + \frac{105}{x^8}\right) \frac{f(x)}{x}$$

qui donne une approximation de l'ordre de 10^{-7} .

Ainsi il convient de ne tenir pour significatives que les 5 premières décimales ou les 7 premières au-delà de $x = 4$.

Le calcul est obtenu par l'intermédiaire d'un tableur : "Excel" sur Mac-SE.

Pour obtenir les valeurs de F(x) sur $] -\infty ; 0[$ il suffit d'utiliser la symétrie de la distribution :

$$F(x) = 1 - F(-x) \text{ en effet } F(x) = \text{Prob}(\{Z < x\}) = \text{Prob}(\{Z > -x\}) = 1 - \text{Prob}(\{Z \leq -x\}) = 1 - F(-x)$$

5. VARIABLE ALEATOIRE de POISSON:

POUR DES VALEURS DU PARAMETRE λ COMPRISES ENTRE 0,1 et 10

Fonction de distribution :

$$P(\{X=k\}) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (\text{probabilité d'apparition d'une valeur égale à } k)$$

Pour la fonction de répartition $F(x) = \sum_{k=0}^x 1_{]-\infty; x[}(k)$ probabilité d'une valeur inférieure ou

égale à k , il suffit d'additionner les probabilités k premiers nombres.

k	$\lambda = 0,1$	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0	0,904837	0,818731	0,740818	0,670320	0,606531	0,548812	0,496585	0,449329	0,406570	0,367879
1	0,090484	0,163746	0,222245	0,268128	0,303265	0,329287	0,347610	0,359463	0,365913	0,367879
2	0,004524	0,016375	0,033337	0,053626	0,075816	0,098786	0,121663	0,143785	0,164661	0,183940
3	0,000151	0,001092	0,003334	0,007150	0,012636	0,019757	0,028388	0,038343	0,049398	0,061313
4	0,000004	0,000055	0,000250	0,000715	0,001580	0,002964	0,004968	0,007669	0,011115	0,015328
5		0,000002	0,000015	0,000057	0,000158	0,000356	0,000696	0,001227	0,002001	0,003066
6			0,000001	0,000004	0,000013	0,000036	0,000081	0,000164	0,000300	0,000511
7					0,000001	0,000003	0,000008	0,000019	0,000039	0,000073
8							0,000001	0,000002	0,000004	0,000009
9										0,000001

k	$\lambda = 1,5$	2	3	4	5	6	7	8	9	10
0	0,223130	0,1353	0,0497	0,0183	0,0067	0,0024	0,0009	0,0003	0,0001	0,0000
1	0,334695	0,2706	0,1493	0,0732	0,0336	0,0148	0,0063	0,0026	0,0011	0,0004
2	0,251021	0,2706	0,2240	0,1465	0,0842	0,0446	0,0223	0,0107	0,0049	0,0022
3	0,125511	0,1804	0,2240	0,1953	0,1403	0,0892	0,0521	0,0286	0,0149	0,0075
4	0,047067	0,0902	0,1680	0,1953	0,1754	0,1338	0,0912	0,0572	0,0337	0,0189
5	0,014120	0,0360	0,1008	0,1562	0,1754	0,1606	0,1277	0,0916	0,0607	0,0378
6	0,003530	0,0120	0,0504	0,1041	0,1462	0,1606	0,1490	0,1221	0,0910	0,0630
7	0,000756	0,0034	0,0216	0,0595	0,1044	0,1376	0,1490	0,1395	0,1171	0,0900
8	0,000142	0,0008	0,0081	0,0297	0,0652	0,1032	0,1303	0,1395	0,1317	0,1125
9	0,000024	0,0001	0,0027	0,0132	0,0362	0,0688	0,1014	0,1240	0,1317	0,1251
10	0,000004	0,0000	0,0008	0,0052	0,0181	0,0413	0,0709	0,0992	0,1185	0,1251
11		0,0000	0,0002	0,0019	0,0082	0,0225	0,0451	0,0721	0,0970	0,1137
12		0,0000	0,0000	0,0006	0,0034	0,0112	0,0263	0,0481	0,0727	0,0947
13			0,0000	0,0001	0,0013	0,0051	0,0141	0,0296	0,0503	0,0729
14			0,0000	0,0000	0,0004	0,0022	0,0070	0,0169	0,0323	0,0520
15				0,0000	0,0001	0,0008	0,0033	0,0090	0,0194	0,0347
16				0,0000	0,0000	0,0003	0,0014	0,0045	0,0109	0,0216
17					0,0000	0,0001	0,0005	0,0021	0,0057	0,0127
18					0,0000	0,0000	0,0002	0,0009	0,0028	0,0070
19					0,0000	0,0000	0,0000	0,0003	0,0013	0,0037
20						0,0000	0,0000	0,0001	0,0006	0,0018
21						0,0000	0,0000	0,0000	0,0002	0,0008
22							0,0000	0,0000	0,0001	0,0004
23							0,0000	0,0000	0,0000	0,0001
24								0,0000	0,0000	0,0000
25								0,0000	0,0000	0,0000

6. **VARIABLE ALEATOIRE de PEARSON de type VI**

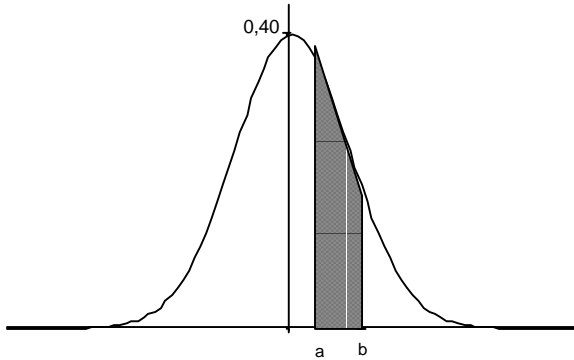
Variable $F(m,n)$ de Fisher-Snedecor à deux degrés de liberté m et n

7. COMPLEMENTS: calculer avec les tables

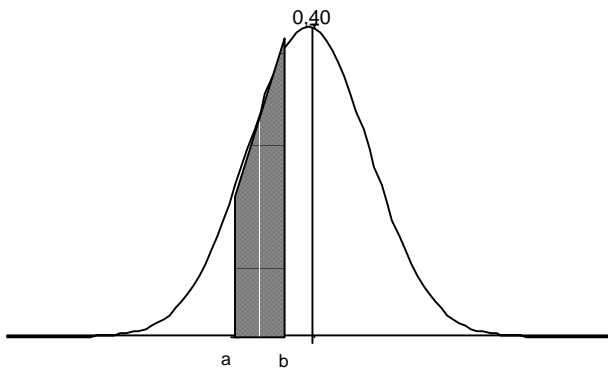
Quelques configurations usuelles :

Usage de la table donnant la distribution de probabilité de la variable centrée réduite de Laplace-Gauss LG(0;1) pour des valeurs $x \geq 0$ à partir de la fonction de répartition

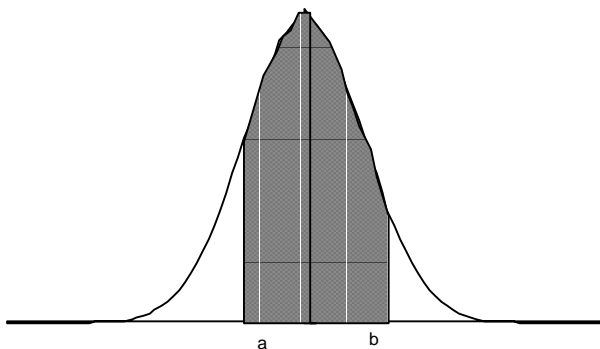
$$F(x) = \text{Prob}(\{Z = \text{LG}(0;1) < x\})$$



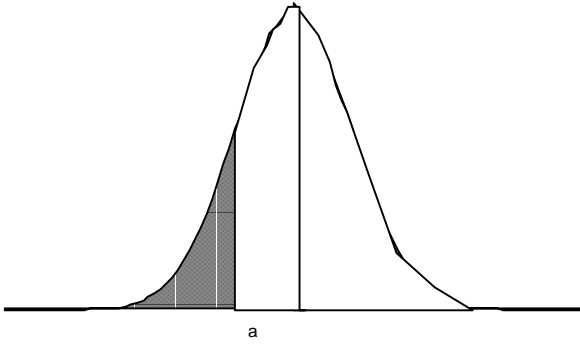
$$\begin{aligned} \text{Prob}(\{a \leq Z < b\}) &= \\ \text{Prob}(\{Z < b\}) - \text{Prob}(\{Z < a\}) &= \\ F(b) - F(a) \end{aligned}$$



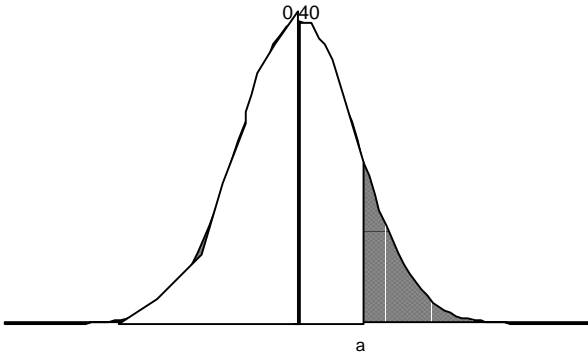
$$\begin{aligned} \text{Prob}(\{a < Z < b\}) &= \\ \text{Prob}(\{-b < Z \leq -a\}) &= \\ \text{Prob}(\{Z < -a\}) - \text{Prob}(\{Z < -b\}) &= \\ F(-a) - F(-b) \end{aligned}$$



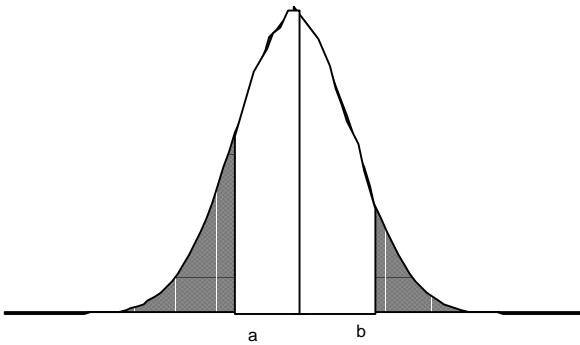
$$\begin{aligned} \text{Prob}(\{a \leq Z < b\}) &= \\ \text{Prob}(\{a \leq Z < 0\}) + \text{Prob}(\{0 \leq Z < b\}) &\text{ or} \\ \text{Prob}(\{0 \leq Z < b\}) &= F(b) - \frac{1}{2} \\ \text{Prob}(\{0 < Z \leq -a\}) &= F(-a) - \frac{1}{2} \\ \text{donc} \\ \text{Prob}(\{a \leq Z < b\}) &= F(-a) + F(b) - 1 \end{aligned}$$



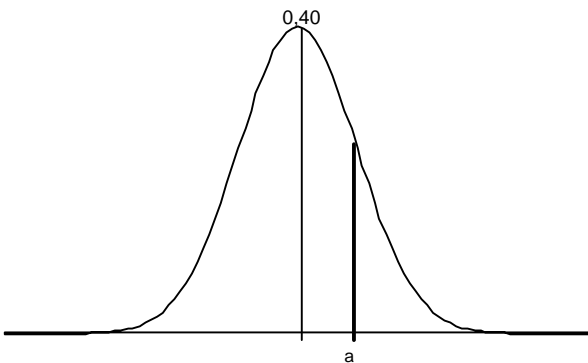
$\text{Prob}(\{Z < a\}) =$
 $\text{Prob}(\{Z \leq -a\}) = 1 - \text{Prob}(\{Z < -a\})$
 donc
 $\text{Prob}(\{Z < a\}) = 1 - F(-a)$



$\text{Prob}(\{Z \leq a\}) = 1 - \text{Prob}(\{Z < a\})$
 donc
 $\text{Prob}(\{Z \leq a\}) = 1 - F(a)$



$\text{Prob}(\{Z < a\} \text{ ou } \{Z > b\}) =$
 $\text{Prob}(\{Z < a\}) + \text{Prob}(\{Z > b\}) =$
 or
 $\text{Prob}(\{Z < a\}) = 1 - \text{Prob}(\{Z < -a\})$
 $\text{Prob}(\{Z > b\}) = 1 - \text{Prob}(\{Z \leq b\})$
 donc
 $\text{Prob}(\{Z < a\} \text{ ou } \{Z > b\}) =$
 $2 - F(-a) - F(b)$



$\text{Prob}(\{Z = a\}) = 0$

C'est sur la base d'un raisonnement analogue sur les histogrammes que l'on peut calculer les probabilités de ces événements sous les hypothèses des lois des variables de de Pearson, de Fisher-Snedecor, de "Student".

On peut aussi remarquer que l'équation $\text{Prob}(\{a \leq X < b\}) = \alpha$ comporte **trois inconnues** à savoir a , b les bornes de l'intervalle et α la valeur de la probabilité de l'événement $[a ; b]$. Le problème est donc résoluble si on fixe α et une information sur a ou b , ou encore si on fixe a et b .