Statistiques Appliquées à la Psychologie
UV PY213B
Présentation du cours 2000/2001

Organisation matérielle

Cours magistral : 13 heures
    mardi 11h30 - 12h30 - Amphi 3
Travaux dirigés : séances de deux heures
    lundi 13h45 à 15h45 - A222
    lundi 16h à 18h - A221
    lundi 16h à 18h - A212
    mercredi 13h45-15h45 - A326
    mercredi 16h à 18h - B213

Contrôle des connaissances : (contrôle continu)
    Examen de 3 heures en janvier

Bibliographie

- N. Guéguen. Manuel de Statistiques pour Psychologues, Dunod

- B. Cadet Méthodes statistiques en psychologie. P.U. de Caen

- G. Mialaret. Statistiques appliquées aux sciences humaines. PUF

- B. Beaufils. Statistiques appliquées à la psychologie. Tome 1. LEXIFAC Bréal

- M. Reuchlin. Précis de Statistiques. PUF Coll. Le Psychologue.

- Colin, Lavoie, Delisle. Initiation aux méthodes quantitatives en Sc. Humaines CDR. No 310-LAV V2098/A

Contenu

Statistiques descriptives : résumer, donner une vue synthétique d'un ensemble de données
Statistiques "mathématiques" : distributions théoriques
Corrélation linéaire, Test du χ^2, aperçu sur les autres tests.

Transparents du CM et polycopiés de TD 1998/99 sur internet (au format .pdf lisible par Acrobat Reader) :

http://geai.univ-brest.fr/enseignements/py213b.html

Pourquoi faut-il étudier les statistiques ?

Les statistiques sont-elles utiles au psychologue ?

Les statistiques, il y a des calculatrices et des logiciels pour faire cela ?

Introduction – Vocabulaire

Collecte des données

Sur qui ?

Population

Individu statistique, unité statistique, sujet

A propos de quoi ?

Attribut, caractère, variable statistique
Modalités d'une variable : exhaustives - exclusives
Champ ou domaine de variation

Nature d'une variable statistique

Variables nominales – échelle nominale

Variables ordinales – échelle ordinale

Variables numériques discrètes ou continues
Echelles d'intervalles, échelles de rapports

Recueil et présentation d'un ensemble de données

Individus : s_1, s_2, . . . , s_N
Variables étudiées : X, Y, . . .
Tableau protocole :

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
. . .	. . .	. . .
s_N	x_N	y_N

Recensement ou tri à plat : **tableau d'effectifs**

Modalités	Effectifs	Fréquences
a_1	n_1	f_1
a_2	n_2	f_2
. . .	. . .	
a_k	n_k	f_k
	N	1 (ou 100%)

Variable regroupée en classes :

Classes	Effectifs	Fréquences
$[a_1, a_2[$	n_1	f_1
$[a_2, a_3[$	n_2	f_2
. . .	. . .	
$[a_k, a_{k+1}[$	n_k	f_k
	N	1 (ou 100%)

Etude conjointe de deux variables : **tableau de contingence**

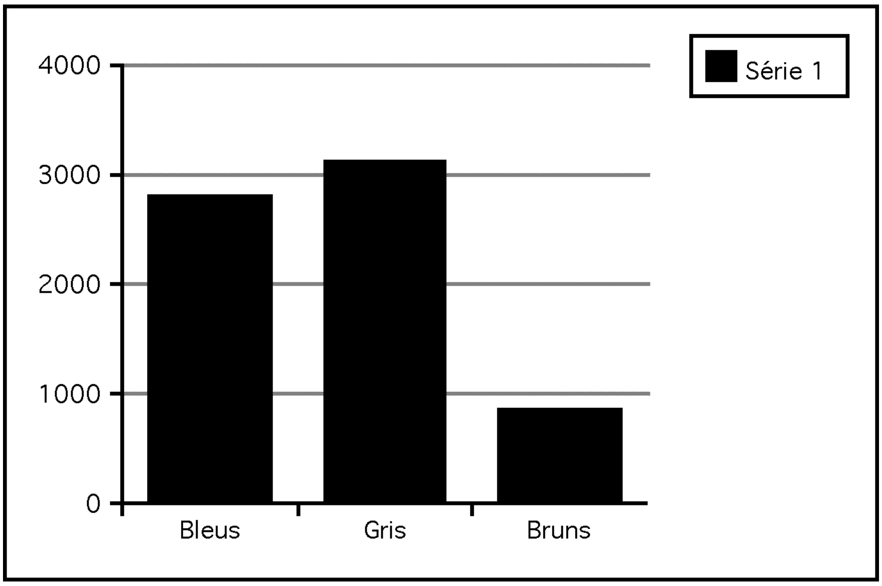
$X \setminus Y$	b_1	b_2	$\ldots$	b_j	$\ldots$	b_l	
a_1	n_{11}	n_{12}		n_{1j}	$\ldots$	n_{1l}	
a_2	n_{21}	n_{22}			$\ldots$		
$\ldots$							
a_i	n_{i1}	n_{i2}		n_{ij}	$\ldots$		
$\ldots$							
a_k	n_{k1}	n_{k2}			$\ldots$	n_{kl}	
	$N_{.1}$						N

Représentations graphiques

Variables nominales

Mod.	Eff.	Freq.	%
Bleus	2811	0,41	41%
Gris	3132	0,46	46%
Bruns	857	0,13	13%
Total	6800	1	100%

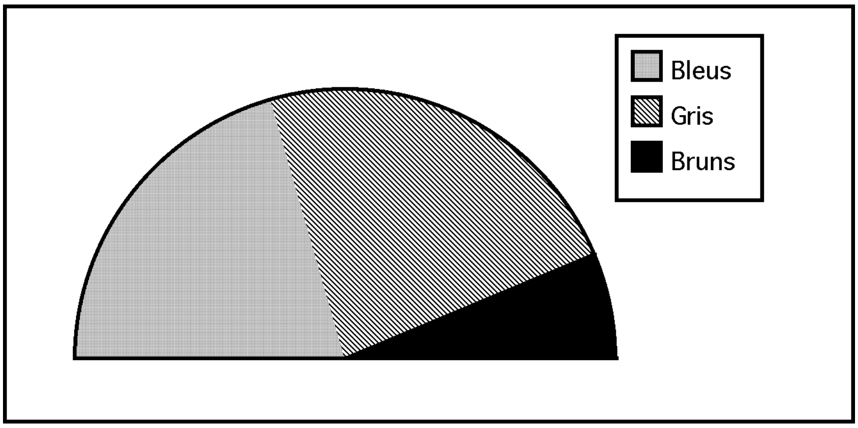
Diagrammes à bandes



Diagrammes circulaires ou semi-circulaires

Méthode de construction : construire un tableau de proportionnalité

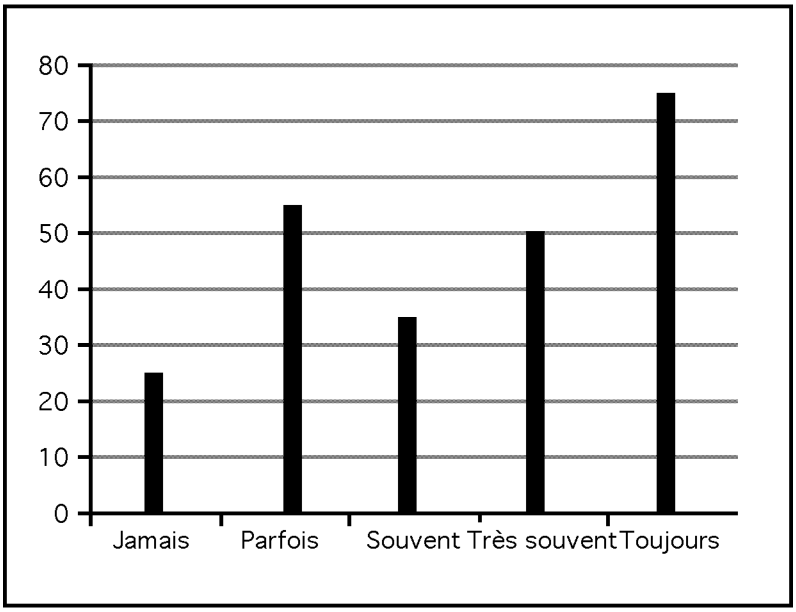
Modalités	Bleu	Gris	Bruns
Fréq.	0,41	0,46	0,13
Angle	74	83	23



Variable ordinale : diagramme en bâtons

On a demandé à 240 sujets s'ils fermaient à clef la porte de leur appartement.

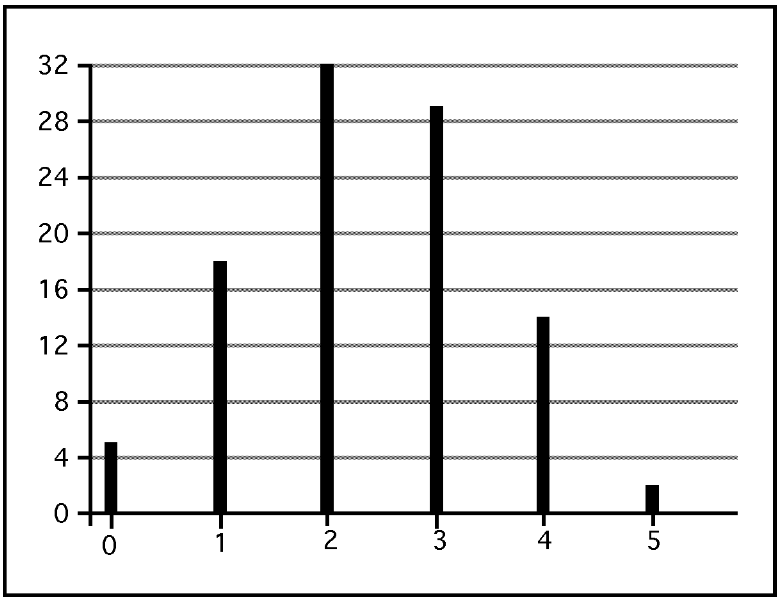
	Jam.	Parf.	Souv.	T. souv.	Tjrs
Eff.	25	55	35	50	75



Variable numérique discrète : même type de construction ; graduation régulière sur l'axe des abscisses.

Exemple : nombre de garçons dans des familles de 5 enfants

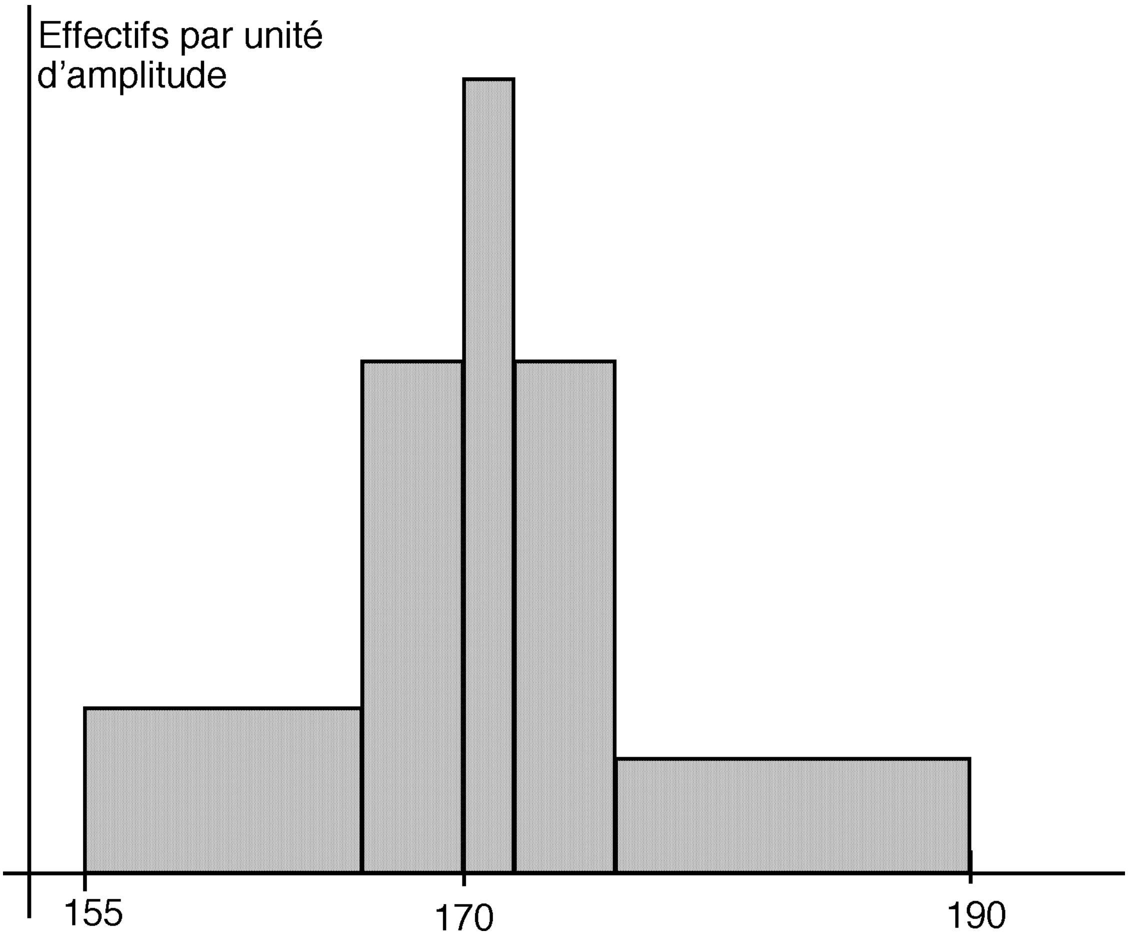
Mod.	0	1	2	3	4	5
Eff.	5	18	32	29	14	2



Variable numérique regroupée en classes : histogramme.

Un histogramme est formé de rectangles adjacents :
– dont la base est proportionnelle à l'amplitude de la classe
– dont l'aire est proportionnelle à l'effectif de la classe

Classe	Eff.	Amplitude	Densité
$[155, 166[$	8	11	0.73
$[166, 170[$	9	4	2.25
$[170, 172[$	7	2	3.5
$[172, 176[$	9	4	2.25
$[176, 190]$	7	14	0,5



Fonction de répartition

Définitions

X une variable statistique numérique
Population d'effectif total N.

Effectif cumulé croissant d'une valeur numérique x *quelconque* : nombre d'individus $n_c(x)$ pour lesquels la variable X est *strictement inférieure* à x.

Fréquence cumulée croissante :

$$F(x) = \frac{n_c(x)}{N}$$

Fonction de répartition : fonction numérique F telle que, pour tout x,

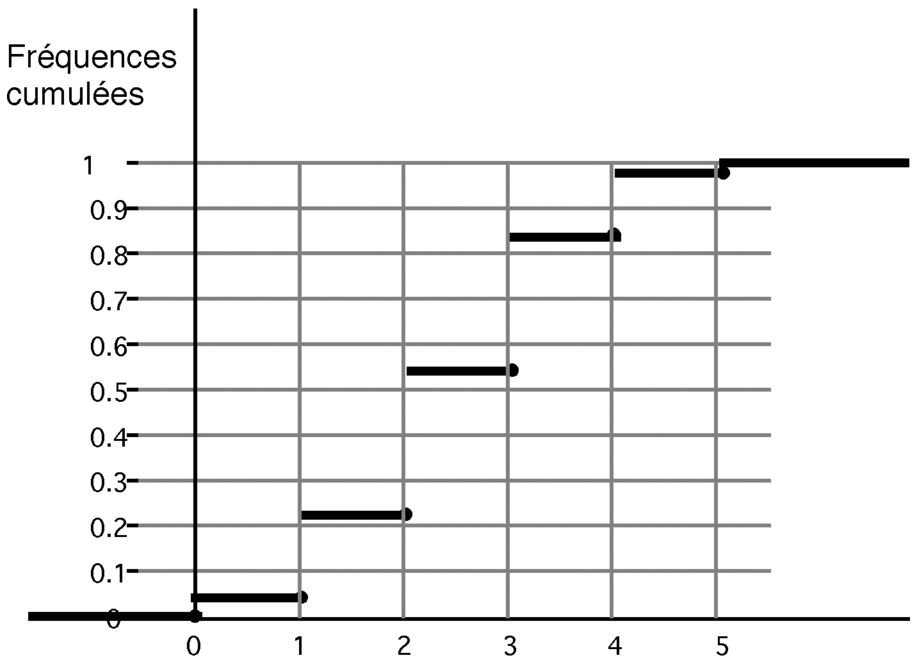
$$F(x) = \frac{n_c(x)}{N}$$

Représentation graphique, cas d'une variable discrète

Exemple : nombre de garçons dans des familles de 5 enfants

Mod.	0	1	2	3	4	5	
Eff.	5	18	32	29	14	2	
Eff. Cum.	0	5	23	55	84	98	100
Fréq. Cum.	0	5%	23%	55%	84%	98%	100%

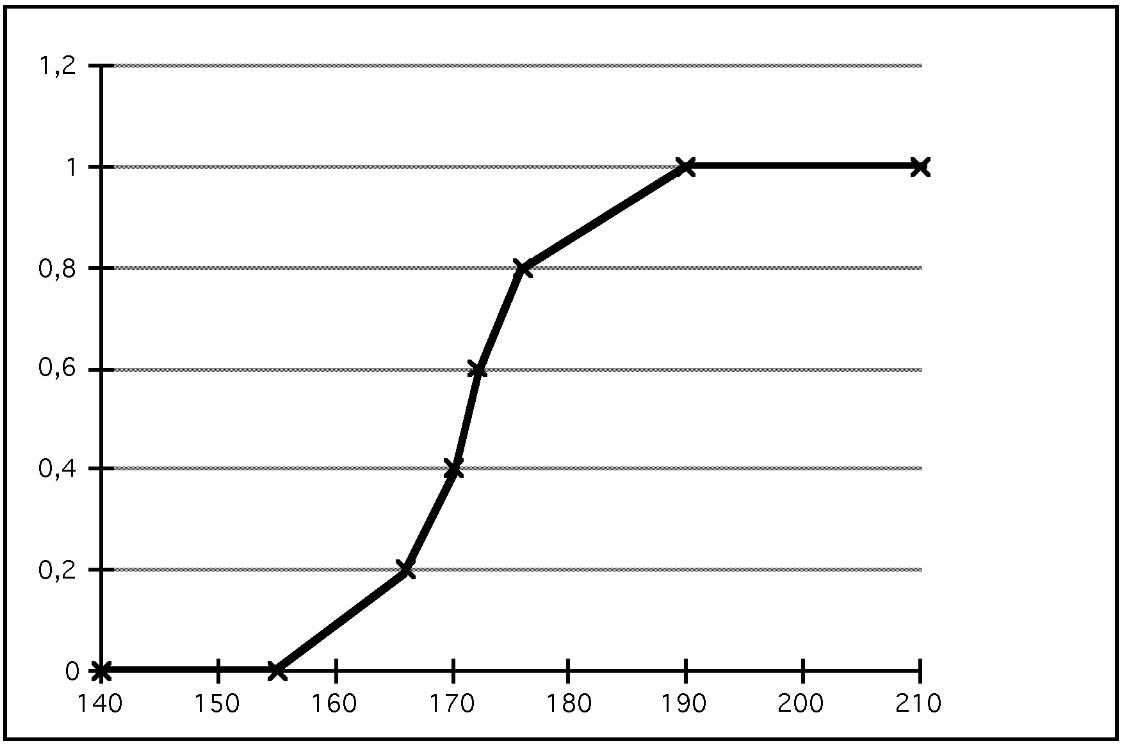
Si $x \leq 0$ $\quad F(x) = 0$
Si $0 < x \leq 1$ $\quad F(x) = 0.05$
Si $1 < x \leq 2$ $\quad F(x) = 0.23$
Si $2 < x \leq 3$ $\quad F(x) = 0.55$
Si $3 < x \leq 4$ $\quad F(x) = 0.84$
Si $4 < x \leq 5$ $\quad F(x) = 0.98$
Si $x > 5$ $\quad F(x) = 1$



Cas d'une variable répartie en classes : hypothèse d'équirépartition des effectifs à l'intérieur d'une classe.

Classes	Effect.
[155, 166[	8
[166, 170[	9
[170, 172[	7
[172, 176[	9
[176, 190]	7
Total	40

x	$n_c(x)$	$F(x)$
155	0	0
166	8	20%
170	17	40.25%
172	24	60%
176	33	80.25%
190	40	100%



Caractéristiques de position

Mode, classe modale

Mode d'une série statistique (nominale, ordinale ou numérique) : modalité correspondant à l'effectif le plus élevé.
N.B. Une série statistique peut admettre plusieurs modes.

Classe modale d'une série statistique regroupée en classes : classe qui a la plus forte densité.
N.B. : c'est la classe correspondant au rectangle de hauteur maximale dans l'histogramme.

Médiane

Variable ordinale ou numérique.

Les individus sont *classés par valeurs croissantes de la variable*. La médiane est la valeur du caractère observée sur l'individu “médian”, à savoir :

– Si N est impair, la médiane est la modalité observée sur l'individu de rang $\frac{N+1}{2}$

– Si N est pair et si le caractère est numérique, la médiane est la moyenne des modalités observées sur les individus de rangs $\frac{N}{2}$ et $\frac{N}{2} + 1$.

Dans le cas d'une variable regroupée en classes :
– La médiane est l'abscisse du point du diagramme cumulatif correspondant à la fréquence cumulée 50%.
– La médiane peut être calculée par *interpolation linéaire* :

$$Md = a_i + (a_{i+1} - a_i)\frac{0.5 - F_i}{F_{i+1} - F_i}$$

où $[a_i, a_{i+1}]$ est la classe contenant l'individu médian et F_i, F_{i+1} désignent les fréquences cumulées des valeurs a_i et a_{i+1}.

Moyenne arithmétique

Caractère numérique.

– Calcul à partir d'un tableau protocole

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

– Calcul à partir d'un tableau d'effectifs

$$\mu = \frac{1}{N}\sum_{i=1}^{k} n_i a_i = \sum_{i=1}^{k} f_i a_i$$

Exemple :

Mod.	Effect.	$n_i a_i$
0	5	0
1	18	18
2	32	64
3	29	87
4	14	56
5	2	10
Total	100	235

$$\mu = 2,35$$

– Cas d'une variable répartie en classes

On considère que la masse de chaque classe est concentrée au centre $c_i = \dfrac{a_i + a_{i+1}}{2}$ de la classe.

Exemple :

Classes	Effect.	Centres	$n_i c_i$
$[155, 166[$	8	160,5	1284
$[166, 170[$	9	168	1512
$[170, 172[$	7	171	1197
$[172, 176[$	9	174	1566
$[176, 190[$	7	183	1281
	40		6840

$$\mu = \frac{6840}{40} = 171$$

Caractéristiques de dispersion

Etendue

$x_1, x_2, \ldots, x_n$: valeurs observées d'une variable statistique numérique.

$$x_{max} = Max(x_1, x_2, \ldots, x_n)$$

$$x_{min} = Min(x_1, x_2, \ldots, x_n)$$

L'étendue de la variable est :

$$e = x_{max} - x_{min}$$

Quartiles

Soit une série statistique numérique de médiane M.

Premier quartile Q_1 : médiane de la série des observations strictement inférieures à M.

Deuxième quartile Q_2 : médiane M de la série complète.

Troisième quartile Q_3 : médiane de la série des observations strictement supérieures à M.

Cas d'une variable continue :

$$F(Q_1) = 0.25 \ ; \ F(M) = 0.5 \ ; \ F(Q_3) = 0.75$$

F est la fonction de répartition

L'écart interquartile est défini par :

$$Iq = Q_3 - Q_1$$

Représentation graphique permettant de visualiser l'étendue et les quartiles : *boîte à moustaches.*

Généralisation : déciles, centiles...

Ecart moyen

– A partir d'un tableau protocole

$$E_m = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

– A partir d'un tableau d'effectifs

$$E_m = \frac{1}{N} \sum_{i=1}^{k} n_i \, |a_i - \mu| = \sum_{i=1}^{k} f_i \, |a_i - \mu|$$

Variance et écart type

Définition : La variance est la moyenne des carrés des écarts à la moyenne.

– A partir d'un tableau protocole

$$V = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

– A partir d'un tableau d'effectifs

$$V = \frac{1}{N} \sum_{i=1}^{k} n_i (a_i - \mu)^2 = \sum_{i=1}^{k} f_i (a_i - \mu)^2$$

L'écart type est donné par : $\sigma = \sqrt{V}$.

Calcul pratique

"Moyenne des carrés moins carré de la moyenne"

$$V = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

$$V = \left(\frac{1}{N} \sum_{i=1}^{k} n_i a_i^2 \right) - \mu^2$$

Remarques
Cas d'une variable répartie en classes : utiliser les centres de classes.
Unités, effet d'un changement d'origine ou d'unités.

Organisation des calculs

– Cas d'une variable discrète

Mod.	Effect.	$n_i a_i$	$n_i a_i^2$
0	5	0	0
1	18	18	18
2	32	64	128
3	29	87	261
4	14	56	224
5	2	10	50
Total	100	235	681

$$\mu = 2.35 \ ; \ V = 6.81 - 2.35^2 = 1.29 \ ; \ \sigma = 1.13$$

– Cas d'une variable répartie en classes

Classes	Effect.	Centres	$n_i c_i$	$n_i c_i^2$
$[155, 166[$	8	160.5	1 284	206 082
$[166, 170[$	9	168	1 512	254 016
$[170, 172[$	7	171	1 197	204 687
$[172, 176[$	9	174	1 566	272 484
$[176, 190]$	7	183	1 281	234 423
	40		6 840	1 171 692

$$\mu = 171$$

$$V = \frac{1171692}{40} - 171^2 = 29292.3 - 29241 = 51.3$$

$$\sigma = \sqrt{51.3} = 7,16 \text{ cm}$$

Introduction à la notion de loi théorique

Etudier une série expérimentale, pour quoi faire ?

– Résumer les observations

– Situer un individu par rapport à son groupe
Ex. : test spatial. Mais se pose alors le problème de la représentativité de ce groupe...

– Comment utiliser les résultats d'une étude expérimentale pour énoncer des lois générales ?

Hasard, aléatoire, variabilité individuelle

– Un physicien qui recherche la valeur d'une grandeur physique, recherche un *nombre*.
Exemple : résistivité du cuivre. Si un morceau de métal n'a pas cette résistivité, ce n'est pas du cuivre...

– En biologie, en psychologie, les résultats obtenus prennent en compte une *variabilité individuelle* plus ou moins importante.
Exemple : le QI moyen de la population humaine est 100. Un être vivant a un QI de 95. Que peut-on en conclure ?

Autre aspect : recherche d'une information dans un bruit de fond.

Modéles théoriques du hasard

Formalisation de l'effet du hasard : probabilités.

Les lois théoriques permettent, dans certaines situations pratiques de justifier des conclusions telles que :

– le hasard suffit seul à expliquer la variabilité

– le hasard ne suffit pas comme explication ; il y a "autre chose".

Deux situations-types de l'effet du hasard :

– Variabilité due à des effets nombreux, de faible amplitude, additifs : loi normale

– Caractère de type dichotomique : succès/échec. Nombre de succès sur N individus tirés au hasard : loi binomiale.

Loi Binomiale

Combinaisons

$n!$ (lire *factorielle* n) désigne : $1 \times 2 \times \ldots \times n$

Le nombre de manières de choisir k éléments parmi n éléments est appelé *nombre de combinaisons de* n *éléments pris* k *à* k.

Il est noté C_n^k. On a : $C_n^k = \dfrac{n!}{k!\,(n-k)!}$.

Epreuve et loi de Bernouilli

Variable statistique à 2 modalités : 1, 0 ou succès/échec

p : fréquence de la modalité "succès"

Caractéristiques : $\mu = p$, $\sigma^2 = p(1-p)$

Loi binomiale : exemple introductif

Un QCM : 3 questions et 4 réponses dont une seule correcte par question.
Population très nombreuse.
Variable X : nombre de réponses correctes données par le sujet.
Si les sujets répondent au hasard, quelle est la fréquence de $X = 2$?

Loi binomiale

Soit n un nombre entier.

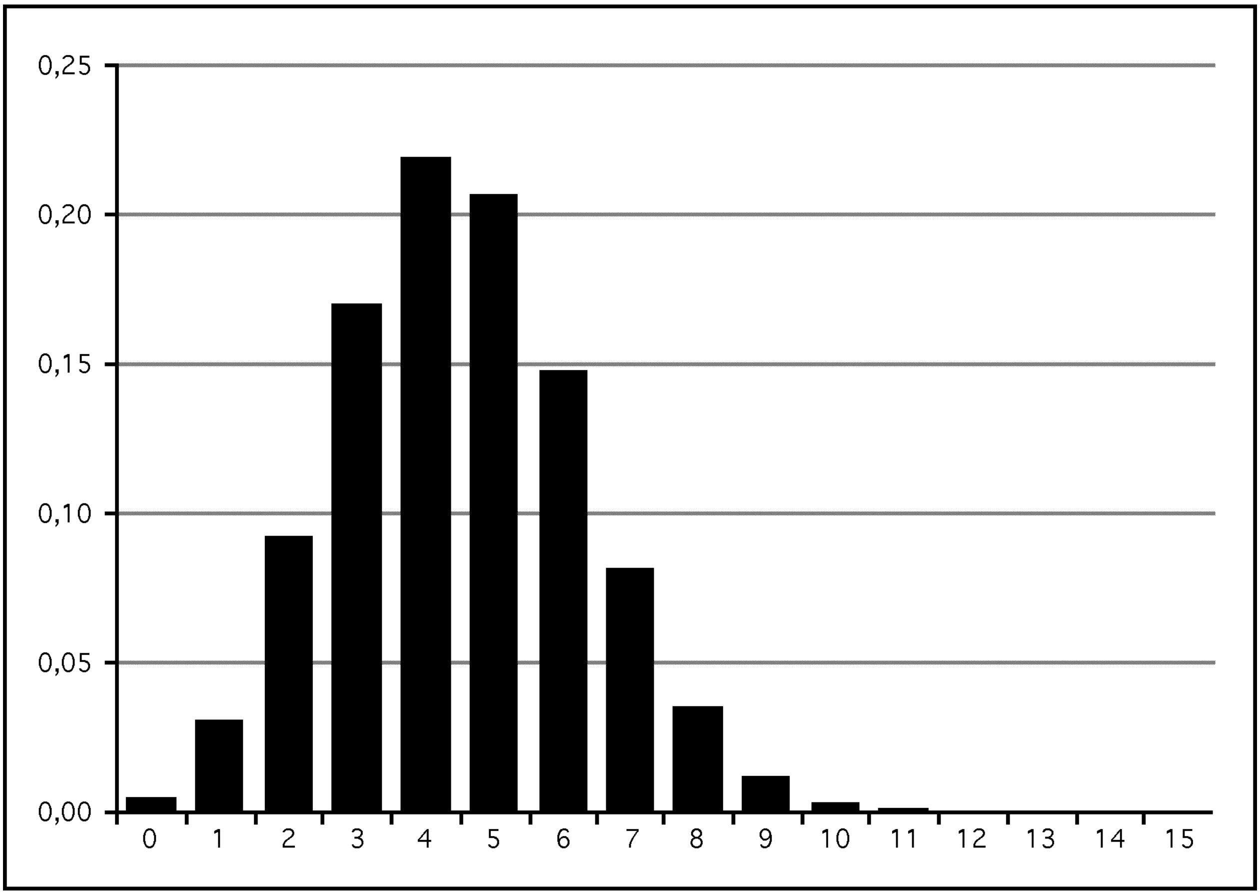
La variable X, “nombre de succès observés lorsqu'on répète n fois, de façon indépendante, une expérience de Bernouilli de paramètre p” suit une *loi binomiale de paramètres n et p*.

– ses modalités sont 0, 1, ..., n
– la fréquence de la modalité k est donnée par :

$$f_k = b(n, p, k) = C_n^k \, p^k \, (1-p)^{n-k}$$

Caractéristiques

$\mu = np$; $\sigma^2 = np(1-p)$



Loi Normale

Qu'est-ce qu'une variable théorique continue ?
Pourquoi des lois théoriques continues ?

– Variables "naturellement" continues (ex. taille)

– Variables regroupées en classes

– Approximation de lois discrètes.

Comment est définie une loi théorique continue ?

Loi d'une variable observée : histogramme, fonction de répartition.

Pour une loi théorique continue :
– densité f ; courbe $y = f(x)$
– Fonction de répartition F ;
courbe cumulative $y = F(x)$

Fréquence de la classe $[a, b[$: $F(b) - F(a)$
Notée aussi : $P(a \leq X < b)$.

$F(a)$: fréquence de la classe $]-\infty, a[$ ou $X < a$
Notée aussi : $P(X < a)$

De même, pour la classe notée $[b, +\infty[$ ou $X \geq b$, $P(X \geq b) = 1 - F(b)$.

Une loi théorique intéressante : la loi normale

Problème : Une loi théorique approchant bien les distributions expérimentales dans lesquelles la dispersion de la variable résulte d'effets *nombreux, additifs, indépendants, du même ordre de grandeur.*

Situation type : planche de Galton

Loi normale centrée réduite

Vocabulaire :
centrée : $\mu = 0$ réduite : $\sigma = 1$

Une variable Z suit la loi normale centrée réduite si ses caractéristiques sont les suivantes :

Ensemble des modalités : $]-\infty, +\infty[$
Densité :

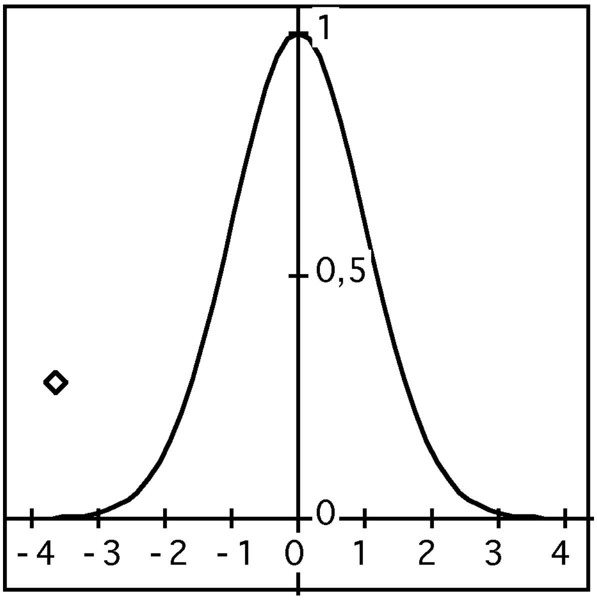
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Fonction de répartition

Utilisation de cette loi : tables numériques

Exemples :
$P(-1 \leq Z < 1) = 68\%$
$P(-2 \leq Z < 2) = 95\%$
$P(-3 \leq Z < 3) = 98\%$

Densité de la loi normale centrée réduite



Loi normale – cas général

Une variable statistique X de moyenne μ et d'écart-type σ suit la loi (est distribuée selon la loi) normale de paramètres μ et σ si la variable Z définie par :

$$Z = \frac{X - \mu}{\sigma}$$

suit une loi normale centrée réduite.

Convergence de la loi binomiale vers la loi normale

Passage du discret au continu: *correction de continuité*

A la modalité k de la loi binomiale, on fait correspondre la classe $[k - 0.5, k + 0.5[$.

Règle

En pratique, on peut remplacer la loi binomiale de paramètres n et p par la loi normale de paramètres $\mu = np$ et $\sigma = \sqrt{np(1-p)}$ lorsque :

$$n \geq 30, \ np \geq 15, \ n(1-p) \geq 15$$

Exemple : loi binomiale telle que $n = 30$ et $p = 0.5$

$P(X = 12) = 0.0806.$
$P(11.5 \leq \tilde{X} < 12.5) = 0.0800$

$F(X \leq 12) = 0.1808.$
$F(\tilde{X} < 12.5) = 0.1807$

Analyses bivariées - Introduction

Jusqu'à présent : études portant sur une seule variable.

Etude simultanée de deux variables nominales :

Analyse croisée de deux variables (par ex. questionnaire d'enquête)

– Loisir préféré et sexe

– Opinion sur l'immigration et sensibilité politique

Question posée : les deux variables sont-elles *indépendantes* ou *dépendantes* ?

Outil : analyse d'un tableau de contingence à l'aide d'un test du χ^2.

Etude de la liaison entre deux variables numériques

– Population d'étudiants. Variables : note de février et note de juin. Lien éventuel ?

– Population de sujets : lien entre taille et poids

Question posée : Y a-t-il un lien, une *corrélation* entre les deux variables ?

Outil : étude de la corrélation linéaire entre les deux variables

Analyse d'un tableau de contingence
Test du χ^2

Exemple : préférences des publics masculin et féminin.
Effectifs observés

	H	F	Total
Comédie	90	75	165
Drame	50	45	95
Variétés	160	80	240
Total	300	200	500

Goûts dépendants du sexe ?

Effectifs attendus (ou théoriques) si indépendance :

Dans chaque case : $\text{effectif} = \dfrac{\text{total ligne} \times \text{total colonne}}{\text{total général}}$

	H	F
Comédie	99	66
Drame	57	38
Variétés	144	96

Calcul de la "distance" du χ^2

Mod.	n_{ij}	t_{ij}	$\dfrac{(n_{ij}-t_{ij})^2}{t_{ij}}$
H.C.	90	99	0.82
H.D	...		0.86
H.V.			1.78
F.C.			1.23
F.D.			1.29
F.V.			2.67
			8.64

Test du χ^2

– 500 personnes : échantillon

– 2 sources de variation : effet du sexe, hasard

– *Si seul le hasard est en cause,* la distance suit une *loi du χ^2 à 2 ddl.*

– On se fixe un seuil de 5% (par exemple)

– *Si seul le hasard est en cause,* on a seulement 5% de chances d'observer un χ^2 supérieur à la *valeur critique* $\chi^2_c = 5.991$.

– Or on a observé : $\chi^2_{obs} = 8.64$.

– Conclusion : différence de goûts selon le sexe.

Résumé

Tableau de contingence : effectifs observés n_{ij}
Totaux par ligne : $n_{i.}$ par colonne : $n_{.j}$
Total général : N ou $n_{..}$
l lignes et c colonnes

Effectifs théoriques : tableau (t_{ij}) avec :

$$t_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}} = \frac{total\ \ ligne \times total\ \ colonne}{total\ \ g\acute{e}n\acute{e}ral}$$

Distance du χ^2 :

$$\chi^2_{obs} = \sum_{i,j} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Test proprement dit :

– Hypothèse : le hasard est seul en cause. Les variables sont indépendantes.

– On fixe un seuil α (5%, 1%, ...)

– Lecture de la table : valeur critique χ_c^2 pour le seuil α et $(l-1)(c-1)$ ddl

– Intervalles d'acceptation et de rejet

– Comparaison de χ_{obs}^2 et de χ_c^2

– Conclusion :

  – Si $\chi_{obs}^2 < \chi_c^2$, indépendance acceptée
  – Si $\chi_{obs}^2 > \chi_c^2$, indépendance rejetée ;
  les variables dépendent l'une de l'autre.

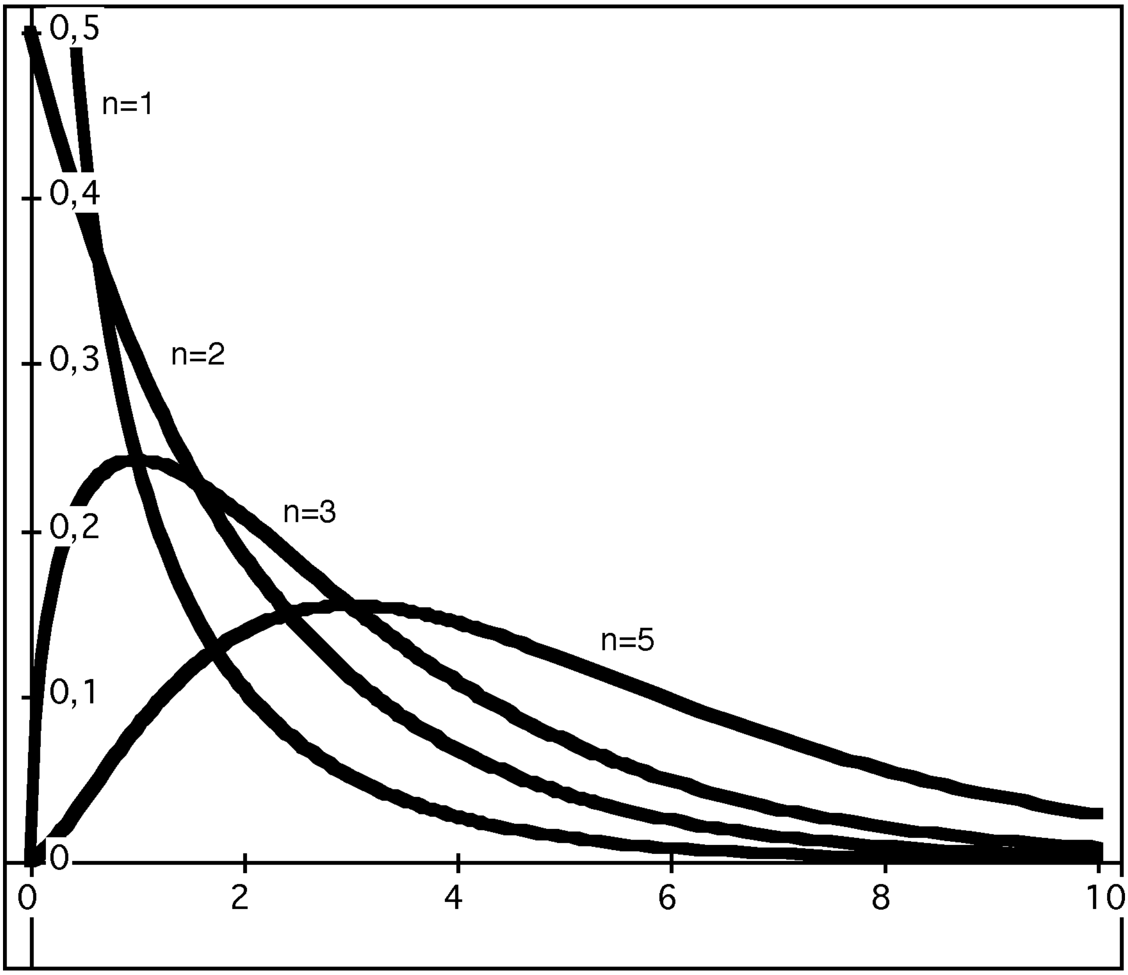
Remarques

– Condition sur les effectifs théoriques minimaux

– Correction de Yates

– D'autres utilisations du test du χ^2

Distributions du χ^2



Corrélation linéaire

Exemple :
– Sujets : étudiants s_i
– Variables : note de février x_i, note de juin y_i
– Y a-t-il un lien entre ces deux notes ?

Données :

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
...	...	...

Nuage de points

Covariance des variables X et Y

$$Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

ou

$$Cov(X, Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\ \overline{y}$$

Coefficient de corrélation de Bravais Pearson

$$\rho = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}$$

Remarques

– Il existe des relations non linéaires

– Corrélation n'est pas causalité

Echantillonnage

Exemple : échantillon - fluctuations d'échantillonnage

Distribution d'échantillonnage

Population des individus
Caractère X, moyenne μ, écart type σ

Population des échantillons de taille n.
Variable $\overline{X}$: moyenne observée sur un échantillon.
Distribution de $\overline{X}$: *distribution d'échantillonnage*.

Théorème de la limite centrée

La variable $\overline{X}$, moyenne observée sur un échantillon de taille n, a pour paramètres :

$$\text{Moy}(\overline{X})=\mu \ \ ; \ \ \text{Var}(\overline{X})=\frac{\sigma^2}{n}$$

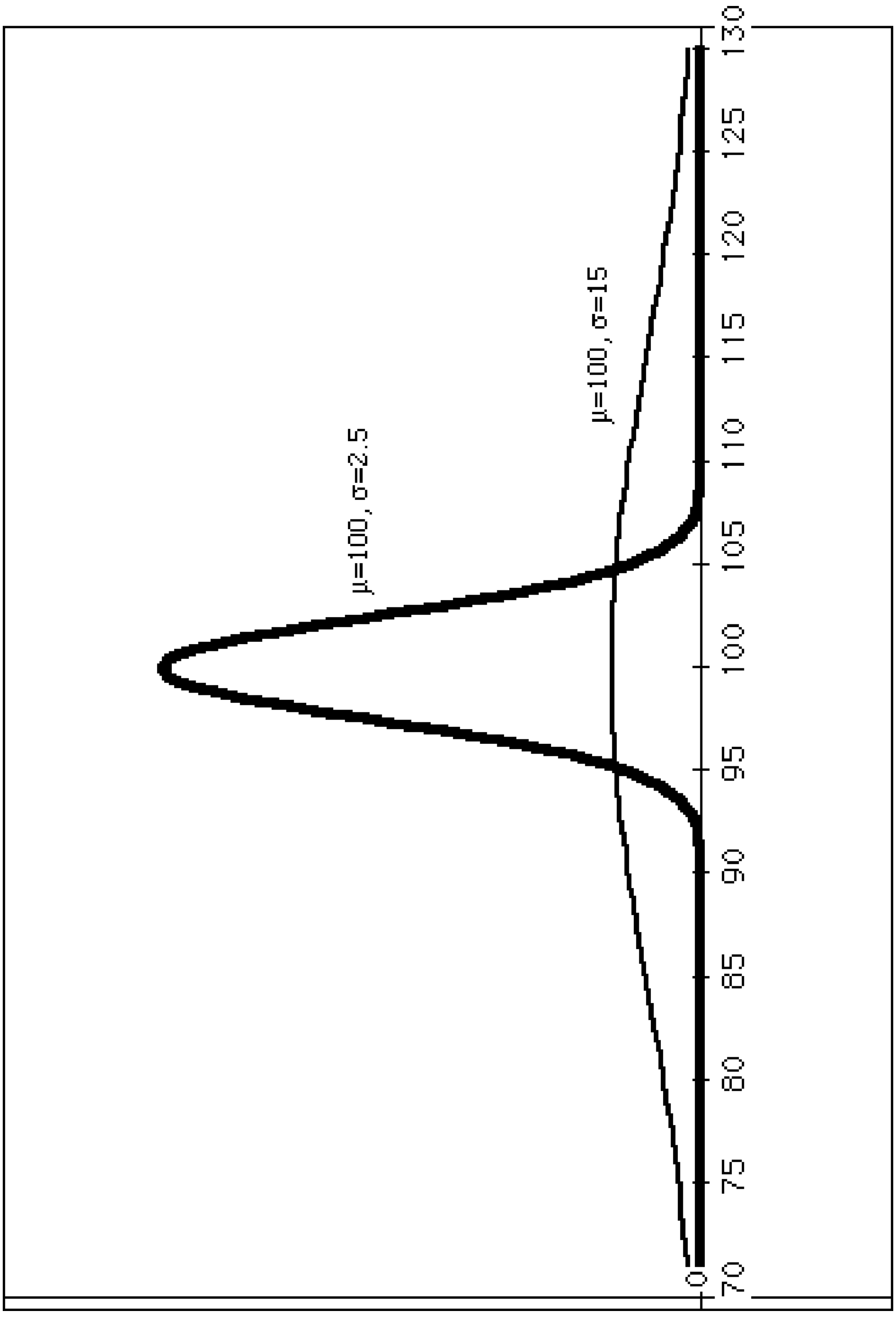
Si n est assez grand ($n \geq 30$), $\overline{X}$ est approximativement distribuée selon une loi normale.

Cas d'une proportion

X : variable dichotomique, modalités 0 et 1
Fréquence de la modalité 1 : p

F fréquence observée sur un échantillon de taille n

$$\text{Moy}(F) = p \ \ ; \ \ \text{Var}(F)=\frac{p(1-p)}{n}$$


μ=100, σ=2.5
μ=100, σ=15
0
70
75
80
85
90
95
100
105
110
115
120
125
130

Estimation de paramètres

Variable X sur une population nombreuse
Moyenne μ, écart type σ inconnus

On a tiré un échantillon:
– Taille n
– Moyenne observée sur l'échantillon : $\overline{X}_{obs}$
– Ecart type de l'échantillon : s.

Estimation ponctuelle de la moyenne

$$\hat{\mu} = \overline{X}_{obs}$$

Estimation ponctuelle de la variance et de l'écart type

Variance corrigée, définie par :

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

Estimation de la moyenne par un intervalle de confiance

Problème : obtenir une affirmation du type : μ est comprise entre ... et ...

En fait, plutôt : avec 95% de degré de confiance, j'estime que μ est comprise entre ... et ...

Exemple

Test sur une population. Score X.
μ et σ inconnus

Echantillon :
$n = 49$
$\overline{X}_{obs} = 25$
$s = 5.94$; $s^2 = 35.26$

Estimation ponctuelle de σ :

$$s_c^2 = \frac{49}{48} \times 35.26 = 36 \ ; \ s_c = 6$$

Raisonner sur une observation

Raisonner sur la distribution d'échantillonnage : intervalle de confiance

- $\overline{X}$: var. "moyenne sur un échantillon de taille 49"
- variable normale, moyenne μ
- Ecart type estimé S tel que : $S^2 = \frac{s_c^2}{n} = \frac{36}{49}$
- Une valeur observée : $\overline{X}_{obs} = 25$

Variable normale centrée réduite associée à $\overline{X}$: Z

$$Z = \frac{\overline{X} - \mu}{S} \ ; \ \overline{X} = SZ + \mu$$

Avec 95% de degré de confiance, on estime que la valeur Z_{obs} associée à $\overline{X}_{obs}$ vérifie :

$$-1.96 \leq Z \leq 1.96$$

Calcul des valeurs extrêmes pour μ :

Pour $\overline{X}_{obs} = 25$ et $Z = -1.96$, $\mu = 26.68$
Pour $\overline{X}_{obs} = 25$ et $Z = 1.96$, $\mu = 23.32$

Conclusion

Avec un degré de confiance de 95%, on estime que :

$$23.32 \leq \mu \leq 26.68$$

Remarques et compléments

– Si on veut donner une formule générale :

$$\overline{X}_{obs} - \frac{s_c}{\sqrt{n}} z_\alpha \leq \mu \leq \overline{X}_{obs} + \frac{s_c}{\sqrt{n}} z_\alpha$$

où z_α est déterminé par :

$$P(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha$$

– Normalité de la distribution parente
– Petits échantillons : on procède autrement si $n < 30$.
– Il existe une méthode analogue dans le cas des proportions

Tests d'hypothèse

On reprend l'exemple précédent :

Test sur une population. Score X
μ et σ inconnus
Echantillon :

$n = 49$
$\overline{X}_{obs} = 25$
$s = 5.94$; $s^2 = 35.26$

Problème : Peut-on affirmer presque sûrement que, sur la population tout entière, la moyenne μ est supérieure à 20 ?

- Hypothèse nulle H_0 : $\mu = 20$
  Hypothèse alternative H_1 : $\mu > 20$
- Seuil choisi : 1% par exemple
- Statistique de test :

$$Z = \frac{\overline{X} - 20}{E} \quad \text{avec} \quad E^2 = \frac{s_c^2}{n}$$

$$Z = \frac{7}{6}(\overline{X} - 20)$$

- Distribution de la statistique de test : loi normale centrée réduite
- Règle de décision et valeur critique : $z_c = 2.33$
- Mise en œuvre du test et conclusion :

Ici : $Z_{obs} = 5.83$. On rejette donc H_0.