40 Variables aléatoires

40.1 Introduction

Dans le cadre des statistiques descriptives, nous avons appris à décrire une variable dans un échantillon*, par différentes distributions de fréquence et quelques valeurs caractéristiques. Ces statistiques sont caractéristiques d'un échantillon particulier, et donc se modifient d'une expérience à l'autre. Au cours de cette section, (induction) nous allons caractériser le comportement de certains types de variables dans les* populations*.*

Les variables dont la distribution peut être modélisée *a priori* dans une population sont appelées **variables aléatoires (v.a.).**

La distribution de probabilité des v.a. peut être représentée par une fonction mathématique qui associe à chaque valeur de la variable aléatoire X sa probabilité d'être observée dans la population notée $\Pi(X)$.

Cette fonction est représentée par une formule. Cette formule peut être précalculée pour un ensemble de valeurs de X.

La moyenne et la variance de la fonction de probabilité sont des valeurs constantes, caractéristiques de la population, que l'on appelle **Espérances**, ou valeurs attendues.

*La variable aléatoire mesurée (*a posteriori*) est la caractéristique numérique associée à une épreuve : la variable aléatoire est caractérisée par une* ***distribution de fréquences relatives*** *qui tend vers une* ***distribution de probabilités*** *lorsque le nombre de réalisations de l'épreuve tend vers l'infini (n→∞).*

Si le modèle décrit bien l'expérience réalisée, la distribution de probabilités (observée) correspond à la fonction de probabilités (modélisée).

40.2 Les variables aléatoires discrètes

40.2.1 Modélisation

Considérons la variable aléatoire X : le nombre de garçons par famille de trois enfants.

Moyennant certaines hypothèses nous allons pouvoir modéliser *a priori* (c'est -à -dire sans réaliser aucune observation dans un échantillon) la distribution de probabilités de X.
Il y a toujours des hypothèses à la base du modèle :

- les naissances sont indépendantes : le fait d'avoir un garçon lors d'une naissance n'influence pas la probabilité d'avoir un garçon lors d'une naissance suivante
- la probabilité d'avoir une fille = la probabilité d'avoir un garçon.

Envisageons les questions suivantes :
- Quelles sont les valeurs possibles pour X?

Dans une famille de trois enfants, le nombre de garçons peut prendre les valeurs 0,1,2,3.

Quelle est la probabilité que X = x ?

X = 0 correspond à l'événement "avoir trois filles" : $G^* \cap G^* \cap G^*$, intersection de trois événements indépendants.
$P(G^*) = 0,5$;
$P(G^* \cap G^* \cap G^*) = 0,5 * 0,5 * 0,5$;
$P(X = 0) = 0,125$

X = 1 correspond à la réalisation de la combinaison d'événements :

$(G \cap G^* \cap G^*) \cup (G^* \cap G \cap G^*) \cup (G^* \cap G^* \cap G)$

union d'événements incompatibles,
intersection d'événements indépendants.

*Soit 3 * (0,5 * 0,5 * 0,5) = 0,375*

Le même raisonnement peut être tenu pour X = 2 et X = 3.

40.2.2 Tables et graphiques

On peut donc établir la fonction de probabilité de X. La fonction de répartition plus souvent employée, est la fonction de probabilités cumulées.

La fonction de répartition est présentée sous forme de tables, les **tables de probabilités,** qui permettent de répondre rapidement à un certain nombre de questions au sujet de l'épreuve dans la population.

Exemple : la fonction de répartition indique que la probabilité d'avoir au maximum un garçon dans une famille de 3 enfants est 0,5.

x	P(X = x)	P(X ≤ x)
0	0,125	0,125
1	0,375	0,500
2	0,375	0,875
3	0,125	1
	1	

Tableau 40 -40-1 Fonction de probabilité (à gauche) et fonction de répartition (à droite),

Ces valeurs peuvent être représentées sous forme graphique, par un diagramme de barres :

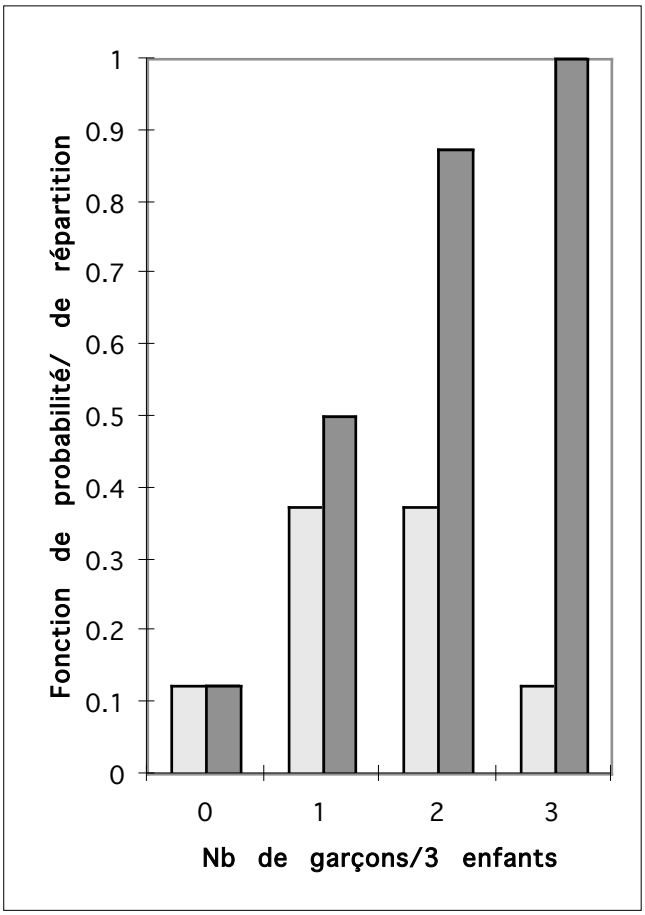

Figure 40 -1 Fonction de probabilité (en blanc) et de répartition (en gris) de la variable X = nombre de garçons par famille de trois enfants.

40.2.3 Valeurs caractéristiques.

❖ **Espérance**

Nous avons établi à partir des fréquences relatives la moyenne d'une distribution :

$$Mx = \frac{1}{n}(n_1 x_1 + n_2 x_2 + ... + n_k x_k)$$

$$Mx = (n_1/n\, x_1 + n_2/n\, x_2 + ... + n_k/n\, x_k)$$

$$Mx = \sum_{i=1}^{k} n_i/n\, x_i$$

$$\lim_{n\to\infty} Mx = \lim_{n\to\infty} \sum_{i=1}^{k} n_i/n\, x_i = \sum_{i=1}^{k} P(x_i)x_i = \mu$$

$$E(X) = \sum_{i=1}^{k} P(x_i)x_i = \mu$$

Équation 40-1

μ est l'Espérance de X : c'est une constante, caractéristique de la population, tandis que Mx est une variable, caractéristique d'un échantillon particulier.

❖ **Variance attendue**

Nous pouvons également définir la variance attendue d'une variable aléatoire, en partant de la définition de la variance dans un échantillon :

$\sigma^2 = E(X - \mu)^2 = VAR(X)$ Équation 40-2

$E(X - \mu)^2 = E(X^2) - 2E(X)\,\mu + E(\mu)^2 = E(X^2) - 2\mu\,\mu + \mu^2 = E(X^2) - \mu^2$

$\sigma^2 = E(X^2) - \mu^2 = VAR(X)$ Équation 40-3

Calcul des valeurs caractéristiques

Reprenons l'exemple du nombre de garçons par famille de trois enfants, et calculons l'Espérance et la Variance Attendue de la distribution de probabilité établie :

X	p(x)	x p(x)	$x^2 p(x)$
0	1/8	0	0
1	3/8	3/8	3/8
2	3/8	6/8	12/8
3	1/8	3/8	9/8
∑	1	12/8	24/8

$$E(X) = \mu = \sum_{i=1}^{k} P(x_i) x_i = \frac{12}{8} = 1.5$$

40.3 Les variables aléatoires continues

40.3.1 Introduction

❖ **Densité de probabilité**

Supposons qu'une firme pharmaceutique souhaite établir la probabilité qu'un individu de la population présente un taux de fibrinogène[1] supérieur à 4.5.

La première démarche possible est de prélever un échantillon assez grand que pour estimer cette probabilité à partir de la fréquence relative de l'échantillon

> *Nous savons qu'une variable continue peut être répartie en classes, et que l'on peut établir une distribution de fréquences (ou de fréquences relatives, de densité de fréquences relatives) de ces différentes classes.*

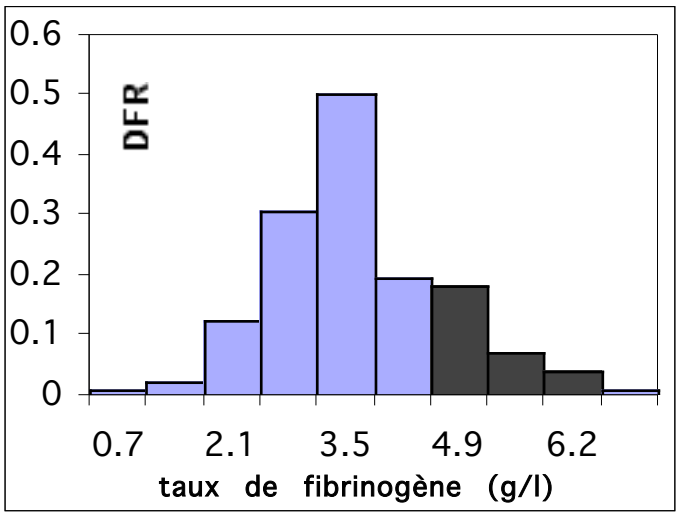

Figure 40 -2 Petit échantillon de taux de fibrinogène. La proportion de la population > 4.5 g/l est estimée par la fréquence relative.

[1] *Fibrinogène : Protéine du plasma sanguin, qui se transforme en fibrine lors de la coagulation.*

Lorsque n augmente :

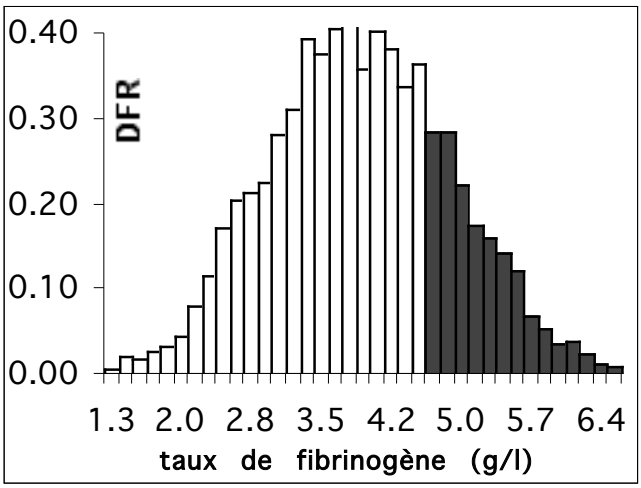

Figure 40 -3 Grand échantillon de taux de fibrinogène. La proportion de la population > 4.5 g/l est estimée par la fréquence relative.

- la surface, et donc la hauteur des différents rectangles se stabilisent, car la fréquence relative tend vers une valeur constante.

- le nombre de classes augmente, puisqu'il est proportionnel à n, et donc l'intervalle de classe diminue. La probabilité que X prenne des valeurs comprises dans un intervalle donné correspond à la somme de plusieurs rectangles.

- à la limite, lorsque n - > ∞, la distribution de densité de fréquences relatives (d.f.r) tend vers une fonction de densité de probabilité, (f.d.p.) représentée par une courbe, et la probabilité que X prenne des valeurs comprises dans un intervalle donné correspond à la surface sous la courbe :

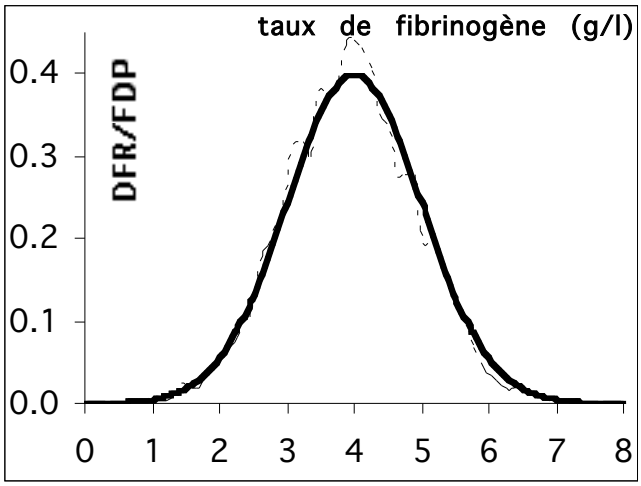

Figure 40 -4 Modèle du taux de fibrinogène dans la population (FDP). Le trait pointillé (DFR, grand échantillon) indique que le modèle « colle » bien à l'expérience.

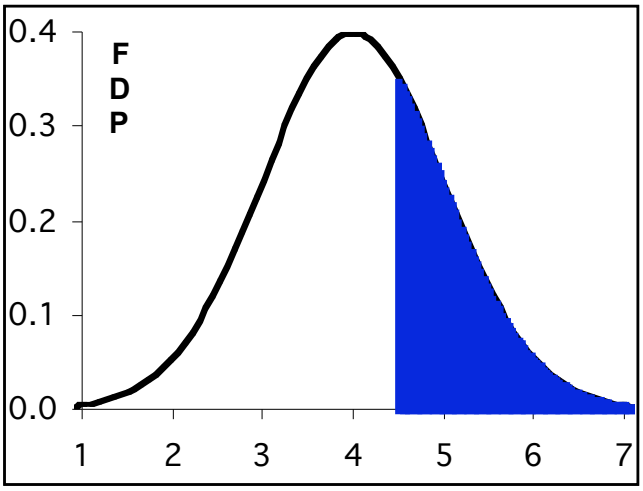

Figure 40 -5 Taux de fibrinogène dans la population (FDP). La proportion de la population > 4.5 g/l est donnée par la surface sous la courbe.

$$f.d.p. = \lim_{n \to \infty} d.f.r$$

Remarquons que puisque la variable est continue, elle peut prendre une infinité de valeurs dans un intervalle aussi petit que l'on veut. La probabilité que la variable prenne exactement une valeur déterminée est donc nulle :

P(X = x) = 0

Pour une variable continue,

P(X < x) = P (X ≤ x)

" la probabilité de sélectionner au hasard dans la population un individu dont le taux de fibrinogène est 4,5 g/l?", signifie en fait : " …compris entre 4,45 et 4,55 g/l ". Dans la majorité des cas on s'intéressera plutôt à " la probabilité de sélectionner au hasard dans la population un individu dont le taux de fibrinogène est au moins 4,5 g/l?", soit P (X ≥ 4,5), parce que cela a plus de sens de parler d'un taux élevé, que d'un taux précis.

❖ **Valeurs caractéristiques**

Pour une variable discrète, la moyenne est définie comme :

$E(X) = \Sigma x_i P(x_i)$

ce qui devient, en remplaçant $\Sigma x_i P(x_i))$ par sa valeur pour une variable continue :

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(X) dx$$

Équation 40-4

et la variance attendue $VAR(X) = E(X - \mu)^2$ devient, en remplaçant $\Sigma x_i P(x_i))$ par sa valeur pour une variable continue :

$$VAR(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(X) dx$$

Équation 40-5

Nous allons voir que pour certaines variables aléatoires caractéristiques, il est possible de caractériser la fonction de densité de probabilité par une fonction f(X) déterminée, et à partir de ces définitions, d'en retrouver leurs paramètres.

40.3.2 La variable aléatoire normale

http://www.fundp.ac.be/biostats/biostat/modules/module70/index.html- module_70

❖ **Utilité**

La v.a. normale est un modèle qui permet de décrire la distribution de probabilité de nombreuses v.a. continues, ou de leur logarithme. Par exemple, la taille ou le poids des organismes, l'erreur de mesure, la concentration de nombreuses substances sanguines (taux de fibrinogène, cholestérol sanguin…), de substances dans l'eau (nitrites, nitrates)...

Ce modèle décrit aussi la distribution de la moyenne de tous les échantillons (parfois à partir d'une taille critique).

Le modèle normal est l'un des plus utilisés pour calculer des probabilités dans le domaine biomédical.

❖ Fonction de probabilité

La fonction de densité de probabilité associée à cette variable est la fonction de Gauss -Laplace, qui représente la "courbe en cloche" bien connue.

Son équation est la suivante :

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Équation 40-6

Retenir cette équation n'est pas primordial, mais il est important de bien saisir la signification de ses paramètres.

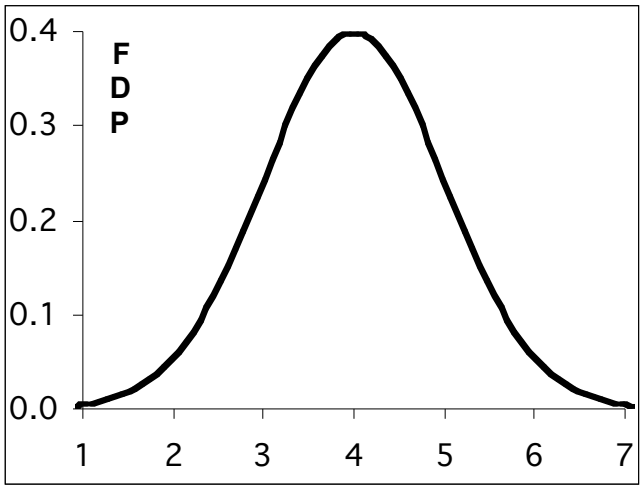

Figure 40 -6 Représentation graphique du modèle de la fonction de probabilité normale, décrite par l'équation de Gauss -Laplace.

❖ Valeurs caractéristiques

Quelles sont les caractéristiques importantes de cette fonction?
Notons d'abord qu'elle est caractérisée par deux paramètres : μ et σ, la moyenne et l'écart -type (ou σ^2 la variance).
Elle est parfaitement symétrique (médiane = moyenne)
f(X)→0 lorsque X→∞.
Le maximum (dérivée première nulle) correspond au point X = μ.
Il y a deux points d'inflexion (dérivée seconde nulle) qui correspondent aux points $\mu \pm\sigma$. La courbe sera donc d'autant plus étalée que σ est grand.
68% de la surface sous la courbe sont compris entre ces deux valeurs.
Au point x = μ correspond l'ordonnée f(X)

$$F(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

Équation 40-7

Cette valeur est donc d'autant plus petite que σ est grand.

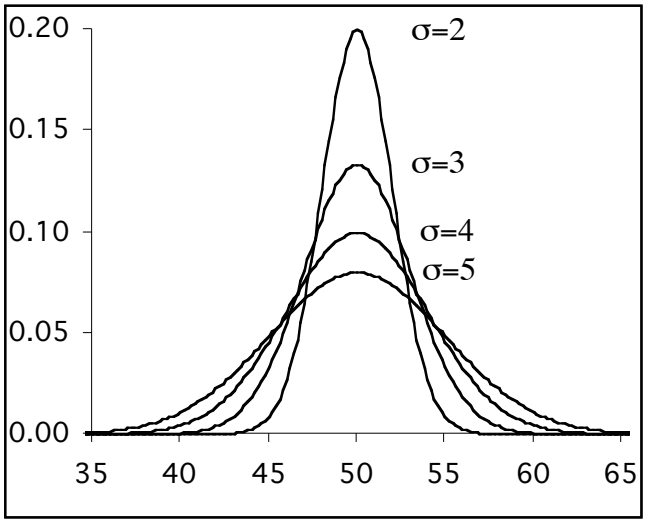

Figure 40 -7 Pour une valeur de μ = 50, on trouve donc toute une famille de courbes, correspondant à différentes valeurs de σ. Ces courbes seront plates et étalées si σ est grand, pointues et étroites si σ est petit, la surface sous la courbe restant toujours égale à 1.

La variable aléatoire normale est notée conventionnellement

X : v.a. $N(\mu;\sigma^2)$

En appliquant une transformation linéaire on trouve que :

X : v.a. $N(\mu;\sigma^2)$,
X' = a + bX : v.a. $N(a + b\,\mu;b^2\sigma^2)$

❖ Calcul de probabilité

La probabilité que X se trouve dans un intervalle donné correspond à la surface sous la courbe dans cet intervalle, et donc à l'intégrale de la fonction dans cet intervalle :

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} x \cdot f(X)dx$$

Cependant, il n'y a pas de solution analytique à cette intégrale : aucune formule ne donne $\Pi(X)$ en fonction des valeurs de x_1, x_2, μ et σ .

Si l'on a établi que le taux de fibrinogène est une v.a ; N (4 ;1), comment résoudre la question de la firme pharmaceutique : déterminer la probabilité qu'un individu présente un taux de fibrinogène ≥ 4,5 ?

Il faut recourir à une intégration numérique, c'est -à -dire estimer l'intégrale en sommant les surfaces d'un grand nombre de rectangles de base très petite.

Ce calcul étant long, les valeurs obtenues pour différents intervalles ont été tabulées. Etant donné qu'il y a une infinité de v.a. normales, chacune caractérisée par une moyenne et une variance propre, il a fallu choisir une distribution de référence, à partir de laquelle il est possible de recalculer rapidement les probabilités d'une v.a. normale particulière.

La variable de référence est la v.a. normale réduite; elle est notée Z, et son intégration numérique calculée une fois pour toutes est publiée dans une table de Z.

40.3.3 Variable normale réduite

❖ **Utilité**

La v.a. normale réduite Z N(0;1) est une variable de référence dont la distribution de probabilité est tabulée. Elle permet de calculer des probabilités relatives à une v.a. normale pour n'importe quelle combinaison de paramètres, par exemple X, v.a. N(5;2), X, v.a. N(150;25) etc…

La transformation d'une variable en normale réduite, appelée standardisation ou réduction, permet également de supprimer les unités et de donner, dans une analyse impliquant plusieurs variables, un poids indépendant du système d'unité de référence. C'est ce que nous avons réalisé pour le calcul du coefficient de corrélation.

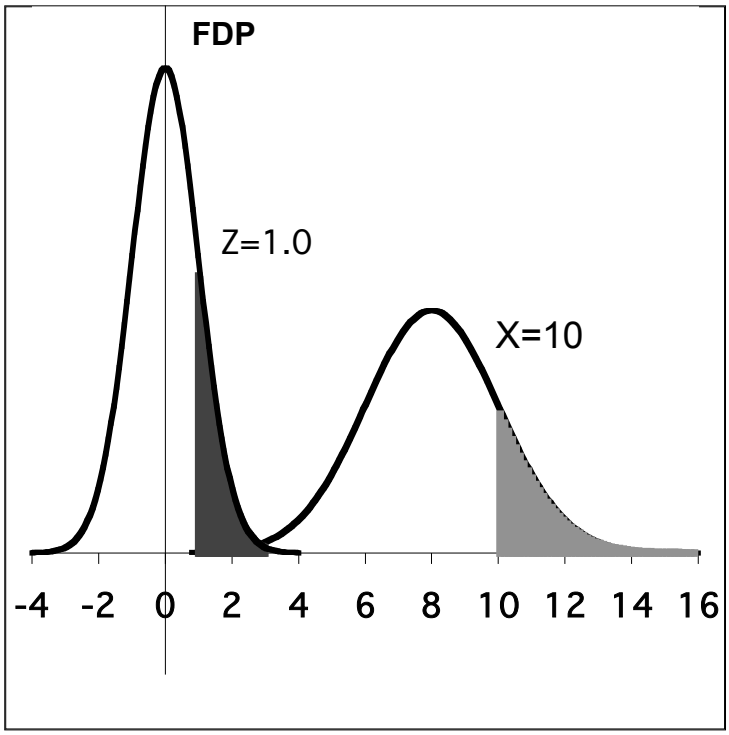

Figure 40 -8 Distribution de X v.a.N(8 ;4) et de Z v.a.N(0 ;1).P(X ≥ 10) = P(Z ≥ 1).

❖ **Standardisation (ou réduction)**

Disposant de la valeur $P(z_1 \leq Z \leq z_2)$, on doit donc retrouver x_1 et x_2 telles que

$P(x_1 \leq X \leq x_2) = P(z_1 \leq Z \leq z_2)$

Pour réduire une distribution quelconque $N(\mu,\sigma^2)$ en v.a. normale réduite N(0;1), il faut effectuer les transformations suivantes :

1. centrer les valeurs , c'est -à -dire retirer la moyenne à chaque valeur
x' = x -μ donc
X' : v.a. $N(0,\sigma^2)$

2. Standardiser les valeurs centrées, c'est -à -dire les exprimer en unités écart -type :

$z = (x - \mu)/\sigma$ ou $x = z\sigma + \mu$ avec $P(X \leq x) = P(Z \leq (x - \mu)/\sigma)$

$$z_i = \frac{x_i - \mu}{\sigma}$$ **Équation 40-8**

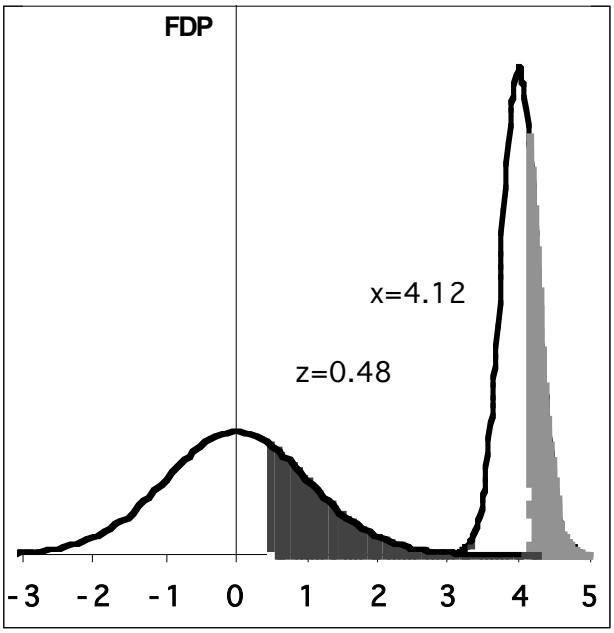

Figure 40 -9 Distribution de X v.a.N(4 ;0.0625) et de Z v.a.N(0 ;1). σ = 0.25 ; P(X ≥ 4,12) = P(Z ≥ 0,48).

Cette transformation permet donc de passer d'une valeur particulière d'une variable normale quelconque à la valeur correspondante de la variable normale réduite, et donc de se référer à sa table de probabilité.

❖ Valeurs caractéristiques

Suivant notre transformation, nous obtenons bien une v.a. N(0;1). En effet,

$$E(Z) = E\left(\frac{x-\mu}{\sigma}\right) = E\left(\frac{x}{\sigma} - \frac{\mu}{\sigma}\right) = E\left(\frac{x}{\sigma}\right) - E\left(\frac{\mu}{\sigma}\right) =$$

$$\frac{1}{\sigma}E(x) - \left(\frac{\mu}{\sigma}\right) = \frac{1}{\sigma}\mu - \left(\frac{\mu}{\sigma}\right) = \left(\frac{\mu}{\sigma}\right) - \left(\frac{\mu}{\sigma}\right) = 0$$

$$VAR(Z) = VAR\left(\frac{x-\mu}{\sigma}\right) = VAR\left(\frac{x}{\sigma} - \frac{\mu}{\sigma}\right) = VAR\left(\frac{x}{\sigma}\right) - VAR\left(\frac{\mu}{\sigma}\right) =$$

$$VAR\left(\frac{x}{\sigma}\right) - 0 = \left(\frac{1}{\sigma^2}\right)VAR(X) = \left(\frac{1}{\sigma^2}\right)\sigma^2 = 1$$

❖ Propriétés

Etant donné la symétrie des valeurs autour de 0, on trouve :

$P(Z \leq -z) = P(Z \geq z)$

La probabilité totale étant égale à 1, on trouve :

$P(Z \geq z) = 1 - P(Z \leq z)$

La probabilité que Z = z étant nulle, on trouve :

$P(z_1 \leq Z \leq z_2) = P(Z \leq z_2) - P(Z \leq z_1)$

❖ Tables et graphiques

Ces propriétés permettent de calculer la probabilité de n'importe quel intervalle sur Z en se référant à la fonction de répartition $p(Z \leq z)$ tabulée pour $Z \geq 0$

0,900	0,950	0,975	0,990	0,995
1,282	1,645	1,960	2,326	2,576

Tableau 40 -40-2 Extrait de la table Z. La première ligne donne une sélection des valeurs de probabilités $P(Z \leq z)$. La seconde ligne donne la valeur de Z correspondant à cette probabilité.

Pour des raisons de mise en page, cette table constitue la dernière ligne de la table t de Student (degrès de liberté ∞).
Une table plus complète est disponible dans les annexes de ce cours. Elle est un peu plus complexe à lire et est rarement nécessaire.
Pour des calculs spécifiques, nous vous renvoyons aux fonctions du tableur Excel.

Reprenons les paramètres du taux de fibrinogène, supposé être une v.a. N(4 ;1).

Quelle est la probabilité de trouver dans cette population un individu dont le taux de fibrinogène est supérieur à 4,5 ?

Calculer la valeur de z

A la valeur 4,5 dans cette distribution, correspond la valeur $(4,5 - 4)_{/1}$ = 0,5 dans la distribution normale réduite.

Définir la probabilité recherchée

$P(X \geq 4,5) = P(Z \geq 0,5)$

Calcul de la probabilité

$P(Z \geq 0,5) = 1 - P(Z \leq 0,5) = 1 - 0,691 = 0,309$

On peut aussi répondre à la question : entre quelles limites 95% des individus de la population sont -ils situés?

Rechercher les valeurs de Z correspondant à cette probabilité

$P(z_1 \leq Z \leq z_2) = 0,95$
en utilisant la symétrie de la courbe :
$P(Z \leq z_2) = 0,975$
$z_2 = 1,96$ et donc $z_1 = -1,96$

Calculer les valeurs x1 et x2 correspondantes

$x_1 = -1,96 * 1 + 4 = 2,04$
$x_2 = 1,96 * 1 + 4 = 5,96$

Ce calcul justifie notre approximation en statistique descriptive : 95% des valeurs environ sont comprises entre Mx±2S, ce qui donne ici $\mu \pm 2\sigma = 4 \pm 2$.

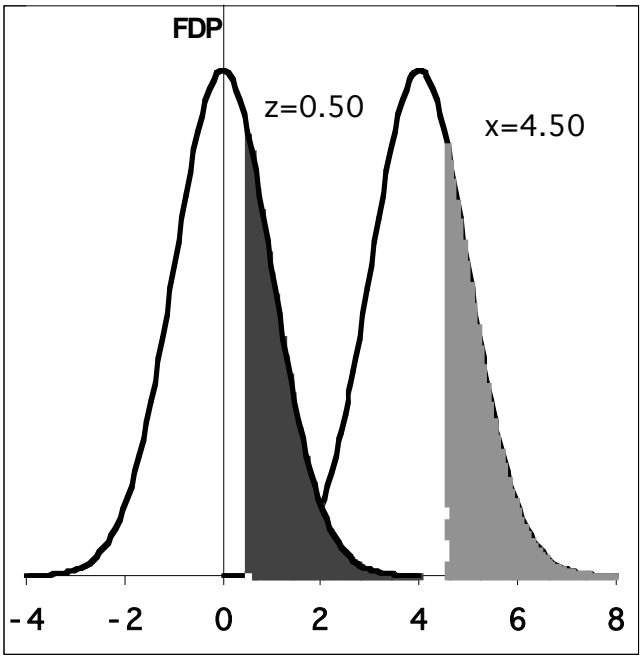

Figure 40 -10 Taux de fibrinogène N(4 ;1). P(X ≥ 4.5) et P(Z ≥ 0,5),

Supposons que la distribution du taux de cholestérol dans le sang (mg/100ml) pour la population humaine soit bien connue et caractérisée par une v.a.N(210;1600).

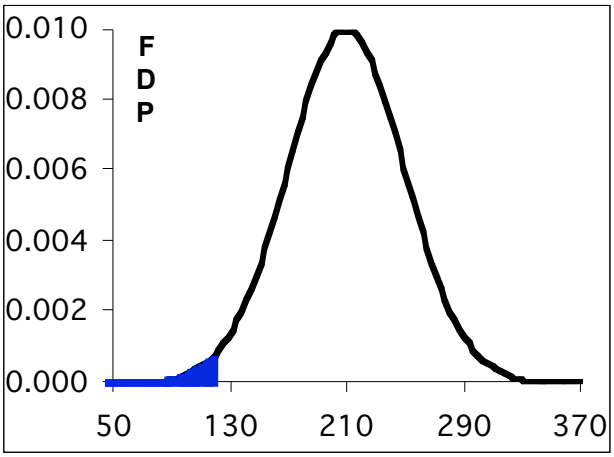

Figure 40 -11 Distribution du taux de cholestérol dans le sang v.a.N(210;1600) mg/100ml.
I

Quelle est la probabilité de trouver dans cette population un individu dont le taux est inférieur à 120 mg/100ml?

Calculer la valeur de z : à la valeur 120 mg/100ml, correspond la valeur $(120 - 210)/40 = -2,25$ dans la distribution normale réduite.

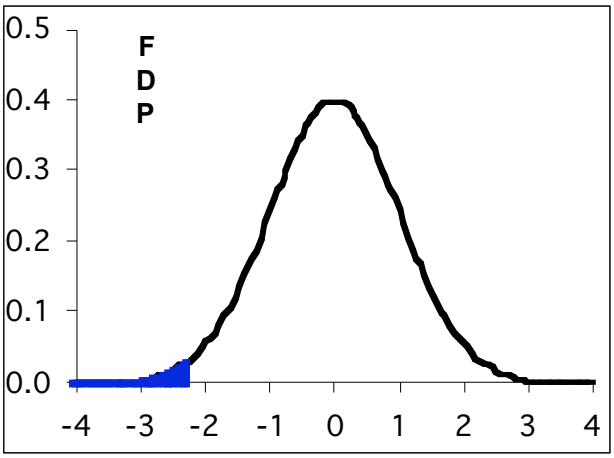

Figure 40 -12 Distribution de Z. La partie hachurée représente P(Z ≤ -2.25).

Définir la probabilité recherchée

$P(X \leq 120) = P(Z \leq -2,25)$

Calcul de la probabilité

En utilisant la symétrie de la v.a.normale, on trouve :

$P(Z \leq -2,25) = P(Z \geq 2,25) = 1 - P(Z \leq 2,25) = 1 - 0,988 = 0,012$

Entre quelles limites inférieure et supérieure 95% des individus de la population sont -ils situés?

Rechercher les valeurs de Z correspondant à cette probabilité

$P(z1 \leq Z \leq z2) = 0,95$
en utilisant la symétrie de la courbe :

$P(Z \leq z2) = 0,975$
$z2 = 1,96$ et donc $z2 = -1,96$

Calculer les valeurs x1 et x2 correspondantes

$x1 = -1,96 * 40 + 210 = 131,6$
$x2 = 1,96 * 40 + 210 = 288,4$

Dans le tableur Excel, la fonction LOI.NORMALE(x;μ;σ;VRAI) renvoie P(X ≤ x) pour X v.a. N(μ;σ)
Dans notre exemple,
LOI.NORMALE(120;210;40;VRAI) renvoie 0,012.

LOI.NORMALE.INVERSE(π;μ;σ) renvoie X v.a. N(μ;σ) pour π = P(X ≤ x)
LOI.NORMALE.INVERSE(0,975;0;1)
renvoie 1,96

Prenez garde, les conventions d'Excel varient d'une fonction à l'autre.

40.3.4 Distribution de Khi -Carré

http://www.fundp.ac.be/biostats/biostat/modules/module90/index.html - module_90

❖ Utilité

La variable Khi -Carré est un modèle qui exprime la distribution de sommes de carrés d'écarts standardisés.
Elle permet de calculer la probabilité d'observer des écarts dus au hasard entre des fréquences observées et celles prévues par une loi de probabilité.
Elle permet également de décrire la distribution de la variance des échantillons pris dans une population.
C'est avec la distribution normale un des outils les plus utilisés en statistiques biomédicales.

❖ Principe

Imaginons un modèle qui répartisse les observations en deux catégories, par exemple mâles et femelles, dans une population de sex -ratio[2] 0,5.

P(mâle) = 1/3, P(femelle) = 2/3; ratio = 0,5

Comptons la fréquence des mâles et des femelles dans un échantillon (n = 87) et la fréquence théorique attendue suivant la répartition 1/3, 2/3.

Calculons un écart quadratique entre les fréquences observées et théoriques, standardisé par la fréquence théorique :

$$\frac{(f_{obs} - f_{th})^2}{f_{th}} : \frac{(23-29)^2}{29} = 1.24$$

et rassemblons les valeurs dans un tableau :

	mâles	femelles	total
f_i observée	23	64	87
f_i théorique	29	58	87
écart quadratique standardisé	1.24	0.62	1.86

Tableau 40 -40-3 Fréquences observées et théoriques, écart quadratique, standardisé par la fréquence théorique, totaux.

[2] *Sex -ratio : rapport numérique des sexes à la naissance (mâles/femelles). (Dans l'espèce humaine il est d'environ 105 garçons pour 100 filles soit 1.05).*

Les fréquences observées fi correspondent approximativement à la valeur de X = Po(μ), expression dans laquelle $\mu = np_i$ avec n la taille de l'échantillon et p_i la probabilité d'appartenir à la catégorie i.

La variance de cette fréquence observée est donc Var(X) = $\mu = np_i$.

La quantité $\frac{(f_{obs} - f_{th})}{\sqrt{f_{th}}}$ est donc approximativement une variable Z(0 ;1).

L'écart global entre les observations et le modèle est calculé par la statistique[3] :

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(f_{i\ obs} - f_{i\ th})^2}{f_{i\ th}}$$ **Équation 40-9**

qui suit approximativement une distribution théorique

$$\chi^2_{k-1} = \sum_{i=1}^{k} Z_i^2$$

expression dans laquelle k représente le nombre de catégories et k -1 le nombre de degrés de liberté, dont dépend la forme de la courbe.

Si l'on répète l'observation un grand nombre de fois, on obtiendra différentes fréquences, et différentes valeurs de χ^2_{obs} .

Echantillon N°	mâles	femelles	χ^2_{obs}
1	23	64	1,860
2	29	62	0,088
3	25	60	0,588
4	25	73	2,699
5	32	63	0,005
6	37	66	0,311
7	32	74	0,472

Tableau 40 -40-4 Répétition de l'expérience, fréquences observées et valeurs de χ^2_{obs}.

Comparons les valeurs obtenues pour χ^2_{obs} et χ^2_{k-1}

[3] χ *: vingt -deuxième lettre de l'alphabet grec, se prononçant khi. Dans les références statistiques, s 'écrit aussi chi -carré,khi -deux….*

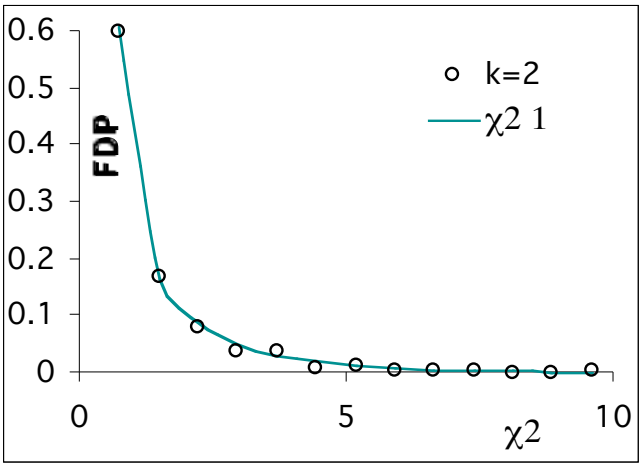

Figure 40 -13 Comparaison des écarts quadratiques standardisés à la distribution théorique de khi -carré avec un degré de liberté.

❖ **Exemple**

Imaginons un modèle qui répartisse les observations en trois catégories, par exemple les produits AA, Aa (ou aA) et aa de plusieurs croisements hétérozygotes, de probabilité 25%, 50% et 25% respectivement. Effectuons 5 fois l'expérience qui consiste à relever la fréquence de chaque génotype[4].

N°	AA	Aa	aa	χ2obs
1	27	49	20	1,06
2	17	53	31	4,13
3	27	46	22	0,62
4	22	46	27	0,62
5	28	53	17	3,12

Tableau 40 -40-5 Répétition de l'expérience, fréquences observées et valeurs de χ^2_{obs}.

[4] *Génotype : patrimoine génétique d'un individu dépendant des gènes hérités de ses parents. Phénotype : (du grec. phainein, montrer, et tupos, marque) : ensemble des caractères somatiques apparents d'un individu, qui expriment l'interaction du génotype et du milieu.*

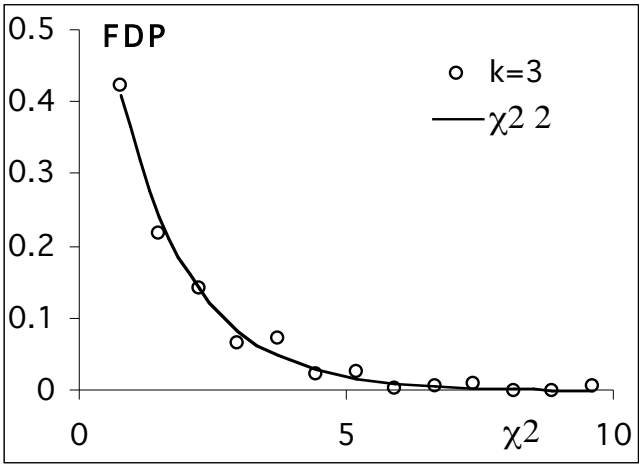

Figure 40 -14 Comparaison des écarts quadratiques standardisés à la distribution théorique de khi -carré avec deux degrés de liberté.

Considérons la probabilité a priori de 10 acides aminés[5] de se trouver dans une hélice alpha[6] et dénombrons leur fréquence dans 4 protéines.

Acide aminé	**Proba - bilite**	**Protéines**			
		1	**2**	**3**	**4**
Ile	0,03	4	3	3	3
Asn	0,05	8	5	8	4
Val	0,07	12	7	8	4
Thr	0,07	5	8	7	9
Tyr	0,07	9	7	10	5
Leu	0,13	8	15	15	12
Pro	0,13	14	13	12	13
Glu	0,15	9	18	16	17
Gly	0,15	17	12	16	10
Met	0,15	12	16	18	20
total	**1**	**98**	**104**	**113**	**97**
χ^2_{obs}		12,32	1,49	2,35	6,40

Tableau 40 -40-6 Probabilités, fréquences observées et χ2 observé pour 10 acides aminés répertoriés dans les hélices alpha de 4 protéines.

[5] *Acide aminé : substance organique ayant une fonction acide et une fonction amine. Vingt acides aminés sont les constituants fondamentaux des protéines.*

[6] *Hélice alpha : structure secondaire d'une protéine, plus ou moins longue, dans laquelle les acides aminés forment un angle caractéristique d'un pas d'hélice.*

Si l'on répète l'expérience sur un plus grand nombre de protéines, on observe une distribution de χ^2_{obs} qui peut se comparer à une distribution théorique de χ^2 avec 9 degrés de liberté.

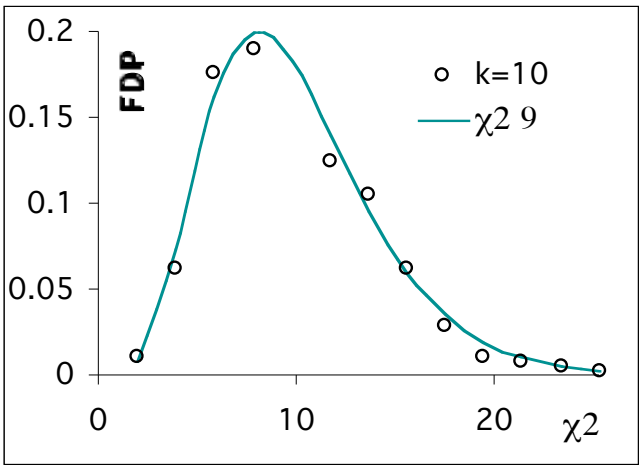

Figure 40 -15 Comparaison des écarts quadratiques standardisés à la distribution théorique de khi -carré avec 9 degrés de liberté.

❖ **Tables et graphiques**

La variable χ^2 est utilisée pour décrire la distribution de sommes des carrés des écarts.

Les degrés de liberté déterminent la forme de la courbe et dépendent du nombre de catégories dans lesquelles les fréquences sont dénombrées.

Plus le nombre de degrés de liberté augmente, plus χ^2 tend vers une v.a. Normale.

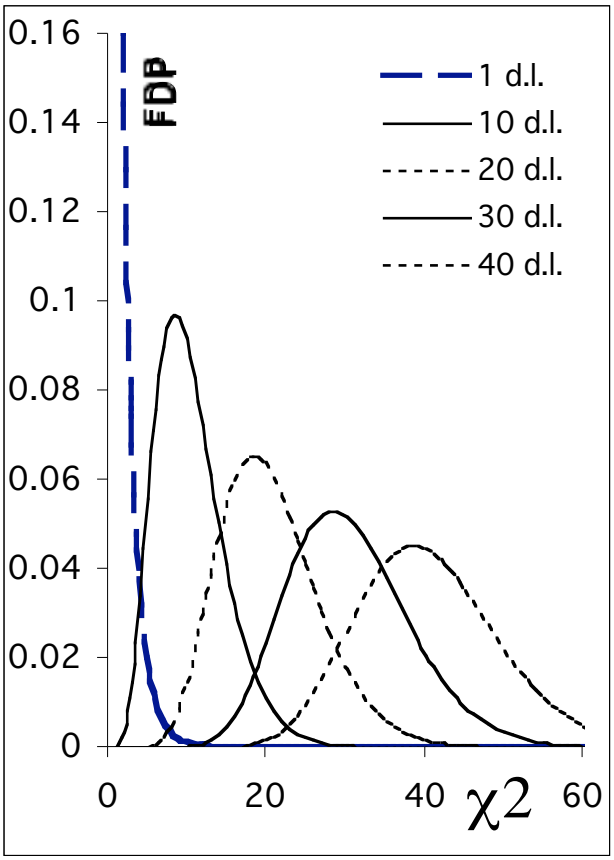

Figure 40 -16 Comparaison de fonctions χ^2 avec différents nombres de degrés de liberté. La distribution de χ^2 avec un petit nombre de degrés de liberté est fortement asymétrique.

La table de χ^2 est généralement présentée de la façon suivante :

	0,9	0,95	0,975	0,99
1	2,71	3,84	5,02	6,63
2	4,61	5,99	7,38	9,21
3	6,25	7,81	9,35	11,34
4	7,78	9,50	11,14	13,28

Tableau 40 -40-7 Extrait de la table de χ^2
En tête de colonne, les probabilités π , en tête de ligne, les degrés de liberté (k). Chaque case comprend $P(\chi^2_\kappa \leq \chi^2_{\kappa;\pi})$.

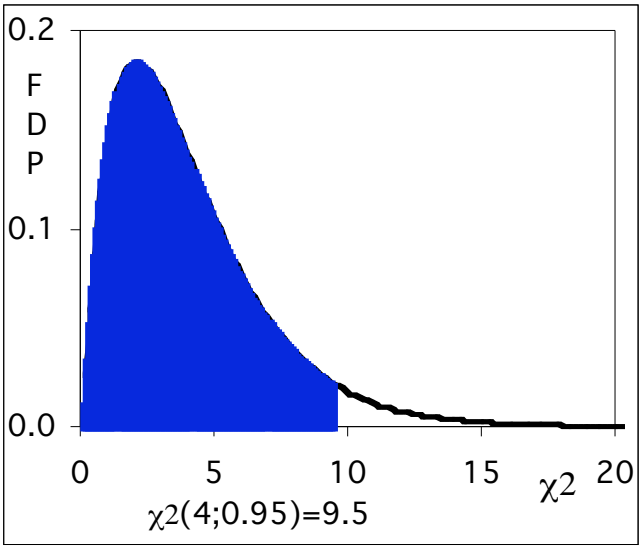

Figure 40 -17 Illustration de la probabilité reprise dans la table, 4 d.l. π= 0.95.

Dans le tableur Excel, la fonction LOI.KHIDEUX(x;dl) renvoie

$P(\chi^2 dl \geq x)$.

Dans notre exemple,
1 -LOI.KHIDEUX(9.5 ;4) renvoie 0,95.