

PROBABILITES ET VARIABLES

Le 10 Février 1999

{LICENCE 98-99. C1-M4. Cours de }

PRESENTATION.....	2
1. HASARD OU PROBABILITE.....	2
<u>1.1. Construction d'un échantillon</u>	<u>3</u>
2. GENERALITES ET CONCEPTS ELEMENTAIRES ET STATISTIQUES.....	3
<u>2.1. Les différents types de statistiques</u>	<u>3</u>
<u>2.2. Les différents types de recherche.....</u>	<u>4</u>
<u>2.3. Les variables.....</u>	<u>4</u>
<u>2.3. Les différents types d'échelles de mesure</u>	<u>4</u>
2.3.1. Les variables nominales ou qualitatives.....	4
2.3.2. Les variables ordinales.....	5
2.3.3. Les variables d'intervalles.....	6
2.3.4. Les variables numériques ou échelle de rapport.....	7
2.3.5. Les autres variables.....	8
<u>2.3. Les caractéristiques entre deux variables.....</u>	<u>8</u>
3. VARIABLES ALEATOIRES.....	8
<u>3.1. Notions de probabilité</u>	<u>9</u>
<u>2.4. Significativité au niveau d'une probabilité</u>	<u>11</u>
3. QUESTIONS DE COURS.....	12
BIBLIOGRAPHIE.....	12

PRESENTATION

Il convient d'établir une distinction entre "les statistiques" et "la statistique".

Les statistiques sont constituées de relevés de données caractérisant de multiples populations, par exemple la population des footballeurs français licenciés. On établit alors la liste des licenciés de la Fédération Française de Football et on porte en regard de chaque nom les caractéristiques que l'on souhaite mettre en évidence :

club d'appartenance, nombre de sélections, âge, taille, poids, performance au 100 mètres et au saut en hauteur...

La méthode statistique ne s'arrête pas à l'enregistrement des données. Elle cherche à fournir une "image", une représentation de la population étudiée, à caractériser celle-ci par différents indices chiffrés, à la rattacher si possible à un modèle mathématique théorique. Travaillant le plus souvent sur un échantillon, le chercheur pourra déterminer ce qu'il est en droit de dire sur la population-mère, en généralisant avec précaution certains des résultats observés sur cet échantillon.

Dans la première partie de ce chapitre, nous allons tout d'abord examiner les différents types de variables que l'on rencontre dans l'observation expérimentale des activités physiques et des sports, et nous allons les distinguer par le "niveau de mesure". L'utilisation statistique de ces observations suppose que l'on puisse assimiler ces variables à des variables aléatoires, variables que nous définirons un peu plus loin, après avoir rappelé quelques notions de probabilité.

1. HASARD OU PROBABILITE

Même si cela se passe encore au niveau des jeux, la probabilité laisse moins de chance au hasard. La probabilité répond à la question : combien de chances pour que cela arrive ?

- quand $p = 0$ l'événement est impossible ;
- quand $p = 1$ l'événement est certain.

L'échantillon est représentatif quand il est une image en modèle réduit de la population. Cette image n'est jamais rigoureusement fidèle.

Par exemple, la probabilité pour qu'une pièce tombe six fois de suite sur son côté pile :

$$p (1/2)^6 = 1/64 = 0,015625.$$

Plus on effectuera de lancers et plus on va se rapprocher de 50/50%.

Population	Droitiers	Gauchers	Fréquence cumulée
N = 10	10	0	10
N = 10	5	5	20
N = 10	10	0	30
N = 10	10	0	40
N = 18	12	6	58
TOTAL	47	11	58

Donc 11 gauchers pour 47 droitiers sur un population prise dans l'amphi de 58 personnes. Mais si cette évaluation avait été faite sur des escrimeurs ou des tennis man, on obtiendrait certainement des résultats différents puisque ces disciplines comportent beaucoup de gauchers proportionnellement.

Plus l'échantillon sera grande (minimum 100) meilleure sera l'idée de la proportion parente, c'est à dire de la population générale à laquelle on fait référence.

Un échantillon faible n'est valable que par rapport à lui-même.

1.1. Construction d'un échantillon

Les méthodes les plus fréquemment utilisées pour construire un échantillon sont :

L'échantillonnage au hasard : dans ce cas on tire "n" sujet (ou classe de sujets) au hasard dans la population parente.

L'échantillonnage stratifié : on retient toutes les classes de sujets composant la population parente (CSP ; Sexe ; Rural-Urbain).

L'échantillonnage stratifié pondéré : les strates sont retenues proportionnellement à leur représentation à l'intérieur de la population parente.

2. GENERALITES ET CONCEPTS ELEMENTAIRES ET STATISTIQUES

2.1. Les différents types de statistiques

On distingue deux types de statistiques :

Statistiques descriptives	Statistiques comparatives
La moyenne	Le chi carré ou k^2
La médiane	Le "t" de Student
L'écart-type	L'analyse de variance
La variance	Les corrélations
Le mode	La régression
Les intervalles de confiance	L'analyse en composante principale
La distribution oercentile (ou centillonnage)	L'analyse factorielle des correspondances

Dans les statistiques descriptives on décrit sans analyse.

Dans les statistiques comparatives on décrit dans le but d'analyser.

2.2. Les différents types de recherche

On distingue deux types de recherches :

La recherche corrélacionnelle qui consiste à mettre en relation des variables sans chercher à les influencer. On se contente de mesurer et de rechercher la relation.

La recherche expérimentale qui consiste en une manipulation des variables afin de voir les retombées en comparaison avec d'autres données.

2.3. Les variables

Tous les objets statistiques que l'on mesure, contrôle, évalue.

Deux types de variables : quantitatif et qualitatif (qu'on qualifie).

Selon P. Parlebas et Cyffer (1992) : "une variable est ce qui varie. Il s'agit effectivement du caractère ou d'un ensemble de caractères dont on note les variations. Les états différents et discernables pris par les variations sont appelés : "les modalités" ou les "valeurs" de la variable. Selon la discipline ou le domaine de variation, on emploiera des termes différents : variable, trait, attribut, paramètre, dimension, propriété.

L'âge peut-être traité comme une variable, de même que la taille ou la catégorie socio-professionnelle mais aussi le niveau d'habileté motrice, la valeur de la performance motrice, la structure de communication des jeux sportifs ou le type d'intervention pédagogique.

- Les variables dépendantes sont celles qui sont simplement enregistrées ou mesurées. Ce sont par exemple les performances motrices des sujets observés.
- Les variables indépendantes sont celles qui sont manipulées par l'expérimentateur.

2.3. Les différents types d'échelles de mesure

2.3.1. Les variables nominales ou qualitatives

Une variable nominale est très peu structurée. Elle distribue la population étudiée en classes d'équivalence : ainsi, les numéros des maillots des footballeurs de division nationale, observés tel jour, distinguent-ils les joueurs qui possèdent par exemple le numéro 1 (gardien de but) ou le numéro 10 (distributeur de jeu). Mais ces classes ne sont pas hiérarchisées. Les numéros sont de simples étiquettes et ne représentent aucune valeur numérique : le numéro 1 n'est pas meilleur que le numéro 10, le numéro 10 ne vaut pas deux fois le numéro 5 ; il ne s'agit donc que d'un code conventionnel. De même, les numéros minéralogiques d'immatriculation des véhicules ne sont pas des grandeurs mathématiques, mais des étiquettes qui regroupent par exemple dans la même catégorie les automobiles du même département.

Une variable nominale produit une partition. Elle définit, sur les éléments de l'ensemble considéré, une relation réflexive, symétrique et transitive, c'est-à-dire une relation d'équivalence. C'est ce que l'on obtient lorsque l'on prend en compte les différentes nationalités des athlètes participant aux Jeux Olympiques : tous les compétiteurs peuvent être regroupés en plusieurs dizaines de classes d'équivalence répondant à leur nationalité respective, sans omission ni répétition (dans l'hypothèse où chaque athlète n'est inscrit que sous une seule nationalité).

Les enquêtes sociologiques ont fréquemment recours aux variables nominales, tant pour l'identification des répondants : situation de famille (marié, célibataire, veuf...), catégorie socio-professionnelle (salarié agricole, agriculteur, personnel de service, ouvrier...), sexe (masculin, féminin) que pour le repérage des réponses s'appuyant sur le codage des items (0, 1, 2...). Dans les deux cas, il ne s'agit que d'étiquettes conventionnelles sans valeur numérique.

En conclusion, les variables nominales ou qualitatives font partie des classifications qualitatives (sexe, CSP...). Ne peuvent pas être classées par ordre d'importance.

2.3.2. Les variables ordinales

Dans ce cas, plus élaboré que le précédent, l'ensemble des données est doté d'une structure d'ordre, c'est-à-dire d'une relation binaire réflexive, antisymétrique et transitive.

Alors qu'une variable nominale répond à une classification, une variable ordinale correspond à un classement, ce qui signifie que les éléments sont hiérarchisés selon des rangs de préséance. On dit alors que l'on dispose d'une échelle ou encore d'une dimension.

Un classement qui attribue un rang à chaque équipe d'une compétition de sports collectifs, la hiérarchie des préférences d'une personne à l'égard d'une dizaine de spectacles sportifs, les choix sociométriques préférentiels émis par un enfant à l'égard de ses camarades, sont des illustrations de variables ordinales. Il se peut que plusieurs candidats possèdent le même rang : ces ex-aequo se regroupent alors dans des classes d'équivalence. Dans ce cas, on obtient une structure de pré-ordre total :

- d'une part, des classes d'équivalence (définissant une classification) et,
- d'autre part, un ordre sur ces classes (définissant un classement).

Dans le cas d'une variable simplement ordinale, il n'est pas possible d'apprécier les intervalles séparant deux éléments : on ne peut pas comparer l'écart séparant le deuxième du troisième à celui qui sépare le cinquième du sixième.

En conclusion, les variables ordinales : permettent de donner un ordre, de hiérarchiser les items mesurés. Classement d'une course : 1° ; 2° ; 3° etc... Peut déterminer le plus ou le moins mais ne peut pas dire de combien.

2.3.3. Les variables d'intervalles

Cette fois, les intervalles séparant deux valeurs quelconques de l'échelle sont calculables et comparables : on peut définir ici des intervalles égaux.

Cette échelle est donc nettement mieux structurée que la précédente : on y a défini une distance, un zéro-origine et une unité-étalon permettant de situer chaque élément avec précision. Cependant, la définition du zéro et de l'unité de mesure reste arbitraire. Il n'est pas possible d'attribuer un statut au rapport entre deux valeurs, alors qu'on peut le faire pour les intervalles (on peut comparer deux intervalles, les additionner ou les soustraire ; on peut dire que l'un d'entre eux vaut le tiers ou le triple d'un autre, mais on ne peut pas effectuer ces opérations sur les valeurs elles-mêmes de la variable).

Un exemple caractéristique d'échelle d'intervalles est fourni par la mesure des températures. Cette mesure peut être effectuée aussi bien sur l'échelle Celsius que sur l'échelle Fahrenheit, et les résultats bruts sont alors évidemment différents (cf. tabl. 3). Dans l'échelle Celsius, le zéro est conventionnellement associé à la température de la glace fondante et la valeur cent à la température d'ébullition de l'eau : l'unité étalon, le degré, est par définition égale à la centième partie de cet écart. Ces éléments sont tous arbitraires. Le zéro ne correspond pas en effet à l'absence de température alors que la valeur zéro dans une épreuve de tir par exemple signifiera l'absence totale d'impact sur la cible (ce qui correspond à une valeur naturelle imposée par la nature même du problème). Les valeurs correspondantes de l'échelle Fahrenheit, tout aussi conventionnelles, sont différentes des précédentes. Cependant, leurs écarts respectifs se correspondent d'échelle à échelle et conservent leurs propriétés arithmétiques internes : ainsi que le souligne le tableau suivant, l'intervalle entre C3 et C2 est égal au double de l'intervalle entre C2 et C1 (unités Celsius), tout comme l'écart entre F3 et F2 vaut le double de l'écart entre F2 et F1 (unités Fahrenheit).

En revanche, on ne peut pas établir un rapport entre les valeurs : il n'y aurait aucun sens à dire qu'une température de 10° vaut le double d'une température de 5° (il n'y fait pas deux fois plus chaud), ou le quart d'une température de 40°.

Les propriétés de cette échelle sont donc liées aux possibilités de calcul sur les intervalles et sur eux seuls : entre 5° et 10°, l'écart est égal à celui observable entre 15° et 20° et il vaut la moitié de celui qui sépare 20° et 30° (et, si l'on prend les températures correspondantes sur l'échelle Fahrenheit, on observe rigoureusement les mêmes rapports entre ces écarts).

Correspondances entre deux échelles d'intervalles arbitraires

	C1	C2	C3	C4
Echelle Celsius	10	30	70	125
	F1	F2	F3	F4
Echelle Fahrenheit	50	86	158	257

La mesure de la température varie selon l'échelle adoptée à 10° Celcius correspondent 50° Fahrenheit. Les zéros des échelles d'une part, leurs unités de mesure d'autre part, sont différents et arbitraires. En revanche, les écarts se correspondent tous et conservent leurs propriétés relatives d'échelle à échelle :

$$\frac{C3 - C2}{C2 - C1} = \frac{40}{20} = 2 \quad \text{de même que :} \quad \frac{F3 - F2}{F2 - F1} = \frac{72}{36} = 2$$

Le premier intervalle vaut le double du second, quelle que soit l'échelle adoptée.

Autrement dit, on peut passer de l'échelle Celsius (C) à l'échelle Fahrenheit (F) par une transformation de variable de type : $y = ax + b$. Ici, les paramètres ont pour valeur : $a = 1,8$ et $b = 32$.

Il vient : $F = 1,8 C + 32$

Par exemple, ainsi que l'indique le tableau, pour $C1 = 10$:

$$F1 = 1,8 C1 + 32 = 18 + 32 = 50$$

C'est ce type de conversion de variable qui est abondamment utilisé dans l'établissement des normalisations et des barèmes de notation. Lorsque la distribution d'une variable est considérée comme normale, il est habituel de la transformer en une variable normale dite réduite dont la moyenne est égale à 0 et l'écart-type à 1.

En conclusion, les variables d'intervalles permettent non seulement de donner un ordre hiérarchique ou chronologique mais aussi de quantifier et de comparer l'ampleur de différences entre les items mesurés (en général toutes les échelles de mesure).

2.3.4. Les variables numériques ou échelle de rapport

Dans ce dernier cas, l'échelle des données est complètement structurée. Les nombres associés aux observations sont des quantités avec lesquelles on peut opérer comme avec des grandeurs physiques : les opérations arithmétiques habituelles (addition, soustraction, multiplication, division) sont toutes utilisables, tant sur les valeurs et les scores que sur les intervalles.

Dans une telle échelle, le zéro n'est plus arbitraire; il dépend du problème qui l'impose naturellement, par exemple dans une épreuve de lancer de poids, de tir ou de saut en hauteur. Il est légitime d'affirmer qu'un jet de 21 mètres est égal au triple d'un jet de 7 mètres et qu'un sauteur à la perche franchissant 6 mètres saute deux fois plus haut qu'un athlète culminant à 3 mètres. Le zéro de l'échelle indique à l'évidence la valeur nulle du caractère étudié : échec dans le franchissement, aucun impact sur la cible...

L'unité de mesure reste arbitraire : ce fait est bien connu en athlétisme où certaines épreuves sont mesurées en yards (0,914 mètre), en miles (1 609 mètres) ou en mètres, ainsi que dans les épreuves de voile où le mille (1852 mètres) est souvent préféré au système métrique. Dans ce cas, il ne s'agit que d'un "changement d'échelle" où l'on passe d'une unité à l'autre par une simple transformation de type $y = ax$.

Ainsi que nous l'avons indiqué, la variable numérique autorise l'utilisation de toutes les opérations arithmétiques, aussi bien sur les scores que sur les intervalles.

Aussi se prête-t-elle aux investigations statistiques les plus élaborées. Il faut reconnaître que le domaine des performances physiques et sportives offre un champ particulièrement propice à l'identification de variables métriques : durées, distances franchies (hauteur, longueur...), cibles atteintes, nombre de touches, poids soulevés. Et lorsque le zéro est arbitraire (épreuve à cotation subjective), il est aisé d'interpréter les performances selon une variable d'intervalles (gymnastique, patinage, plongeon...). Cette abondance de variables numériques dans le domaine sportif tend peut-être à incliner les motriciens à généraliser abusivement la procédure métrique à des situations motrices qui ne s'y prêtent pas véritablement.

En conclusion, les variables de ration ou échelle de rapport possèdent les mêmes qualités que les variables d'intervalles mais en plus possèdent un zéro absolu (temps, espace etc...), alors que les variables d'intervalles possèdent un zéro arbitraire (température).

2.3.5. Les autres variables

On reconnaît encore :

- Les variables continues : c'est une variables telle qu'entre deux valeurs quelconques, il est toujours possible de situer une valeur intermédiaire (ex : temps chronométré d'une course).
- Les variables discontinues ou discrètes : c'est une variable non continue qui varie non pas de façon progressive mais en effectuant des sauts sur un ensemble, lui même discret, en passant d'une valeur ponctuelle à une autre valeur ponctuelle arrêtée (ex : performance d'un concours de saut en hauteur).

2.3. Les caractéristiques entre deux variables

- L'intensité : c'est la "force" avec laquelle deux variables sont reliées. Plus l'intensité de la relation entre deux variables est élevée, plus on peut prévoir la valeur de l'une en fonction de la valeur de l'autre.
- La fiabilité : c'est la probabilité de trouver une relation similaire à celle mise en évidence. Si l'expérience était à nouveau menée sur d'autres échantillons issus de la même population (et plus encore à la population totale ou population parente dont elle est issue).

3. VARIABLES ALEATOIRES

Les variables que nous venons de décrire ne pourront être soumises au traitement statistique que si nous pouvons les assimiler à des variables aléatoires, notion qu'il convient de préciser. Pour cela, il nous faut, au préalable, rappeler quelques notions de probabilité.

3.1. Notions de probabilité

Nous avons tous une notion intuitive de la probabilité. Jetons une pièce de monnaie, la probabilité d'apparition de pile est de : 1/2.

Jetons un dé : la probabilité d'apparition du 5 est 1/6.

A partir de cette notion, on définit la probabilité comme le rapport du nombre de cas favorables au nombre de cas possibles.

Dans le cas du dé, il y a six cas possibles. L'apparition d'un point déterminé, par exemple le 5, constitue un cas favorable.

Nous disons :

$$\text{Pr (5)} = \frac{\text{nbre de cas favorables}}{\text{nbre de cas possibles}} = \frac{1}{6}$$

De même, la probabilité d'apparition d'un point pair est :

$$\text{Pr (point pair)} = \frac{3}{6} = \frac{1}{2}$$

Il s'agit là d'une probabilité a priori dont la définition prête cependant à critique : elle suppose les différents cas possibles équiprobables, elle implique donc comme acquise la notion de probabilités égales.

Sans entrer dans les difficultés rencontrées dans la définition exacte de la probabilité, nous devons reconnaître que nous avons également une notion intuitive de l'équiprobabilité, et que nous la rattachons à la fréquence d'apparition du cas considéré.

Si, par exemple, le 6 sort 10 fois de suite dans les lancers d'un dé, nous aurons tendance à penser, voire à affirmer, que le dé est pipé. Le dé est considéré comme correct si, en moyenne, chaque point sort une fois sur six.

Il est souvent impossible, dans la pratique, de définir les cas possibles ou également possibles. On a recours à une définition de type empirique qui est la limite vers laquelle tend la fréquence expérimentale lorsque le nombre de cas observés devient de plus en plus grand.

Exemple : probabilité qu'un homme de 50 ans atteigne sa 70^{ème} année, probabilité qu'un garçon de 16 ans réussisse à sauter 4,90 m en longueur.

La probabilité est un nombre toujours compris entre 0 et 1.

□ $p = 0$ caractérise un événement impossible,

□ $p = 1$ caractérise un événement certain.

Sans nous attarder sur le calcul des probabilités, il nous paraît indispensable de proposer quelques notions fondamentales.

Notation : on est souvent conduit à préciser les circonstances dans lesquelles l'événement aléatoire considéré peut se produire :

si A désigne l'événement aléatoire,

si h désigne les circonstances on écrit $\Pr(A/h)$ que l'on énonce : Probabilité de A si h ou, pour être rigoureux : probabilité d'apparition de l'événement A lorsque les conditions h sont réalisées.

On se contente d'écrire $\Pr(A)$ lorsqu'il n'y a aucune confusion possible (cas du lancer d'un dé par exemple).

a) Probabilités complémentaires ou contraires

On représente par \bar{A} l'événement contraire de A.

On a :

$$\Pr(\bar{A}) = 1 - \Pr(A)$$

Exemple : Probabilité qu'un dé ne donne pas la valeur 5 :

$$\Pr(\bar{5}) = 1 - \Pr(5) = 1 - \frac{1}{6} = \frac{5}{6}$$

2.4. Significativité au niveau d'une probabilité

C'est une mesure estimée du degré auquel le résultat est vrai.

Ex : p (level en anglais) représente le niveau d'erreur.

Si $p = 0,05$, cela veut dire qu'on peut accepter le résultat avec 5% d'erreurs. Plus p est bas et plus le niveau de significativité est élevé. Il existe donc 5% pour que la relation soit due au hasard ---> 5% d'erreur et 95% de chance de relation juste.

- à $p = 0,05$ ($p = .05$) : les résultats sont statistiquement significatifs avec une probabilité d'erreur de 5%, ce qui n'est pas négligeable.
- à $p = 0,01$ ($p = .01$) : les résultats sont statistiquement significatifs.
- à $p = 0,001$ ($p = .001$) : les résultats sont statistiquement très significatifs.

NB : il faut se rappeler que ces références ne sont que des conventions arbitraires, basées sur une expérience générale de la recherche.

La probabilité d'un événement se définit comme le rapport entre le nombre des cas favorables à l'arrivée (l'apparition) de cet événement et le nombre total des cas possibles (ex : une pièce monnaie, la probabilité de faire pile est : $p = 1/2 = 0,5$; pour un dé à six faces nous avons une chance sur six de faire "4" : $p = 1/6 = 0,167$; pour tirer une "dame de cœur" au hasard d'un jeu de 32 cartes : $p = 1/32 = 0,031$).

Il apparaît donc que la probabilité est un nombre compris entre 0 et 1. La probabilité d'un événement impossible est égale à 0. La probabilité d'un événement certain est égale à 1.

La significativité d'un événement est fortement liée à la taille de l'échantillon, donc plus l'échantillon est faible et moins celui-ci sera significatif.

Si la relation est forte, un faible échantillon peut suffire.

Si la relation est faible, il faut un large échantillon.

La fonction qui fait apparaître la significativité est donnée par la loi normale.

Le chercheur énoncera une hypothèse expérimentale. Elle peut être suivant le cas une hypothèse de différence ou de non différence (H_1).

Le statisticien énonce toujours une hypothèse de non référence : nous n'avons pas apporté la preuve d'une différence ; les différences peuvent s'expliquer par le hasard de l'échantillonnage. C'est l'hypothèse nulle (H_0).

3. QUESTIONS DE COURS

Quelle est la probabilité d'apparition d'un événement certain, d'un événement impossible ?

Quelles sont les différentes manières de procéder à un échantillonnage ?

Qu'est-ce qu'une variable ordinale ?

Quels sont les différents niveaux de significativité d'une relation entre deux variables ?

Qu'est-ce qu'une hypothèse nulle ?

BIBLIOGRAPHIE

Parlebas, P & Cyffers, B (1992). Statistique appliquée en milieu éducatif - Paris INSEP.

Langouet, G & Porlier, J-C (1981). Mesure et statistique en milieu éducatif - Paris ESF.