Enseignement de la statistique par l'analyse et réciproquement

Pr. Dr. Daniel Justens
Haute Ecole Francisco Ferrer
UER "mathématique appliquée"
IREM de Bruxelles
Place Anneessens, 11
1000 Bruxelles
Belgique
daniel.justens@brunette.brucity.be

Abstract

L'enseignement de la statistique descriptive se révèle d'une grande complexité. Il est possible de combiner les cours de statistique descriptive du secondaire et certaines parties des cours de *calculus*, illustrant ce dernier et donnant de la statistique une vision plus structurée et facilement orientable vers les applications réelles. Nous introduisons les notions de variables discrètes pragmatiques, discrètes regroupées et continues. Pour la représentation graphique de ces dernières, nous abandonnons la notion mal comprise et génératrice d'erreurs, d'histogramme pour introduire systématiquement des fonctions densités de fréquence observées. Ces dernières, en escaliers, introduisent naturellement l'intégration de manière interprétable. Le passage à la fonction de répartition observée illustre la notion de primitive. La paramétrisation se fait par le biais de l'analyse : les paramètres de position réalisent le minimum d'une certaine définition de l'erreur totale et les valeurs de ces minima introduisent les paramètres de dispersion. Le passage à la modélisation continue se fait à partir de l'interprétation des paramètres discrets.

1. Introduction

Assez paradoxalement, l'enseignement de la statistique descriptive se révèle d'une grande complexité. Certes, les techniques mathématiques utilisées sont élémentaires, mais l'intégration obligatoire dans un contexte réel des problèmes traités pose souvent problème. La difficulté semble double pour nos étudiants, cumulant l'incompréhension de la "situation problème" rencontrée et la maîtrise imparfaite du modèle qui doit lui être appliqué.

Nous montrons qu'il est possible, et même dans certains cas souhaitable, de combiner les cours de statistique descriptive du secondaire (ou du premier cycle de l'enseignement supérieur) et le cours d'analyse, illustrant ce dernier au moyen d'exemples concrets et donnant de la statistique une vision plus structurée et partant plus facilement orientable vers les applications réelles.

Dans cet article, nous supposons toutes les notions de base connues et ne nous focalisons que sur les présentations innovantes ou non classiques. Nous ne rappelons aucune des définitions usuelles standard[1].

2. Regroupements, tableaux et représentations graphiques

Nous introduisons les notions de variables discrètes pragmatiques, discrètes regroupées et continues en insistant sur l'identité de traitement de ces dernières et en signalant la possibilité d'établir ici un pont avec l'analyse non standard et la notion d'infiniment petit.

Une distinction doit être faite ici, lors du regroupement, entre les points de vue mathématique et statistique, distinction qui, bien comprise, améliore la compréhension de concepts délicats comme la continuité.

Pour mieux introduire cette différence, nous tablons sur la notion d'*information* Une variable pourra être traitée de manière discrète (et sera traitée pragmatiquement en discret) si pour chaque valeur observable, une information suffisante est disponible ou encore, ce qui revient au même, un nombre jugé suffisant d'observations a été réalisé. Dans tous les autres cas, un traitement en continu et donc un regroupement doit être réalisé. Le tableau discret a la forme :

Valeurs	Effectifs	Fréquences
x_1	n_1	f_1
x_i	n_i	f_i
x_k	n_k	f_k

Insistons ici aussi sur la notion de précision de la mesure dans la distinction "continu-discret". Une variable discrète peut être mesurée exactement[2]. Une variable continue se mesure dans tous les cas avec un certain niveau de précision. Sans trop d'excès d'intuition, on peut voir une mesure réellement continue comme une *appartenance à un intervalle*.

On remarque alors une certaine similitude entre le *discret regroupé*[3] et le *continu véritable*: la valeur discrète regroupée est représentée continûment par un intervalle dont cette valeur est le centre; la valeur continue n'est pas exactement déterminée et se révèle donc une appartenance à un intervalle. Par simplification, on note alors le centre de cet intervalle comme étant la mesure effectuée avant que de procéder au regroupement des valeurs *proches* par réunion de ces intervalles en *k* classes de longueurs bien choisies, éventuellement différentes quand le contexte l'exige. Cette façon de procéder évite toute erreur de regroupement[4].

---

[1] Le lecteur intéressé par le développement complet de notre point de vue peut consulter *La statistique par l'analyse*, ouvrage pouvant être commandé via internet sur le site www.cefal.com.
[2] Même si ce n'est pas toujours le cas.
[3] Une information insuffisante par valeur conduit à une présentation regroupée où on considère comme équivalentes les observations *proches* (un concept à définir).
[4] Les classes choisies doivent tenir compte du niveau de précision de la mesure.

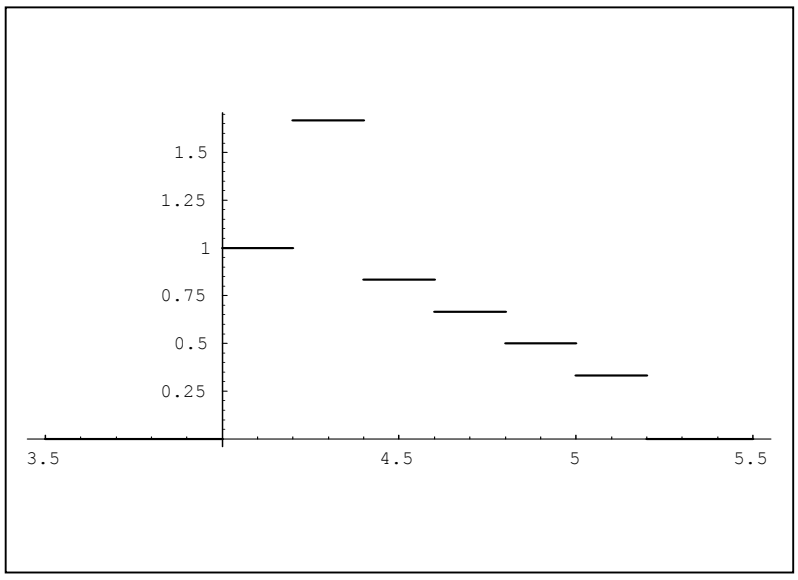
Pour la représentation graphique des variables regroupées et continues, nous abandonnons la notion trop floue et mal comprise d'histogramme pour introduire systématiquement des fonctions densités de fréquences observées. Ces dernières, en escaliers, introduisent naturellement l'intégration de manière interprétable car représentant une excellente approximation de la fréquence d'un intervalle ne coïncidant pas avec une des classes choisies. Le tableau suivant reprend une présentation générale pour les variables continues.

Valeurs	Effectifs	Fréquences	Densités
$C_1 = [a_1; a_2[$	n_1	f_1	$d_1 = \frac{f_1}{a_2 - a_1}$
$C_i = [a_i, a_{i+1}[$	n_i	f_i	$d_i = \frac{f_i}{a_{i+1} - a_i}$
$C_k = [a_k, a_{k+1}[$	n_k	f_k	$d_k = \frac{f_k}{a_{k+1} - a_k}$

Nous nous plaçons ensuite dans le cadre d'un exemple réel simple : le prix du beurre en euro. Trente marques (n=30) de qualité identique se différencient uniquement par leur prix. Nous présentons les valeurs observées déjà regroupées[5].

Intervalle de prix en euros	Effectif	Fréquence	Fréquence cumulée	Densité de fréquence
[4.00; 4.20[	6	0.200	0,200	1
[4.20; 4.40[	10	0.333	0.533	1.667
[4.40; 4.60[	5	0.167	0,700	0.833
[4.60; 4.80[	4	0.133	0.833	0.667
[4.80; 5.00[	3	0.100	0.933	0.500
[5.00; 5.20[	2	0.067	1.000	0.333

Graphiquement, la densité est une fonction en escaliers. Elle est identiquement nulle en dehors de l'intervalle de définition de la variable :



[5] Pour les méthodes de regroupement, nous renvoyons le lecteur à tout bon manuel de statistique.

Le passage à la fonction de répartition observée doit être fait en toute compatibilité avec la partition adoptée pour le choix des classes. En optant pour des intervalles de type "fermé-ouvert", nous devons choisir une répartition représentant la fréquence des observations strictement[6] inférieures à *x*.

Le passage à la fonction de répartition observée illustre la notion de primitive. On vérifie ici que la représentation en escaliers de la densité est équivalente à une représentation polygonale de la répartition (fréquences cumulées), ce qui n'est pas toujours le cas dans la littérature.

Mathématiquement, on peut faire le raisonnement qui suit :

On peut montrer que la fonction de répartition $F(x)$ est une primitive particulière de $f(x)$, prenant l'aspect d'une fonction polygonale. Il suffit de prouver que la dérivée de $F(x)$ est égale à $f(x)$. On a par définition :

$$F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}$$

En utilisant la définition de la fonction de répartition, il vient :

$$F'(x) = \lim_{\Delta x \to 0} \frac{\text{fréquence de} [x, x+\Delta x[}{\Delta x}$$

Supposons Δx positif et x appartenant à la classe C_i. Le calcul effectué étant une limite pour Δx tendant vers 0 et la classe considérée étant ouverte à droite, le point $x+\Delta x$ peut toujours être choisi dans la même classe. Dans ce cas, la fréquence de l'intervalle considéré est donnée par $f(x) * \Delta x = d_i * \Delta x$. On en tire :

$$F'(x) = \lim_{\Delta x \to 0} \frac{d_i * \Delta x}{\Delta x} = d_i$$

On retrouve la définition de la fonction densité.

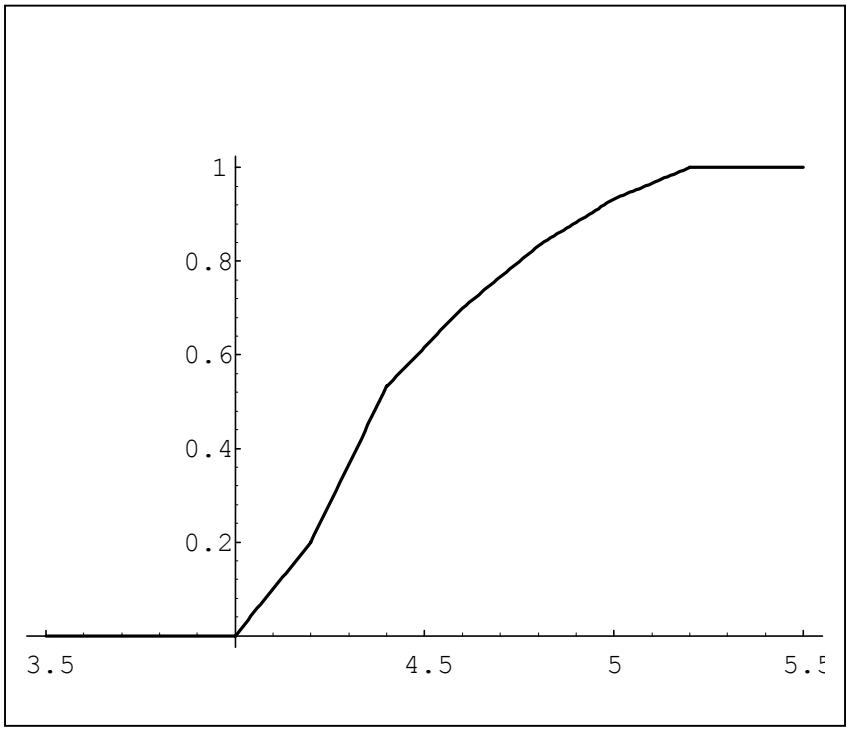
On retrouve également, illustré par le calcul de l'estimation de la fréquence observée associée à un intervalle, l'un des résultats fondamentaux du calcul différentiel et intégral, liant le calcul de primitive et le calcul d'une intégrale définie. Ce résultat est illustré dans le contexte statistique et devient aisément interprétable. En notant comme plus haut *f(x)* la densité et *F(x)* la répartition observée, on montre facilement que pour tout *a, b* réels, on a:

$$\int_a^b f(x)dx = F(b) - F(a)$$

La représentation de la fonction de répartition observée ne pose aucun problème : la densité étant une constante, sa primitive est donc du premier degré. La valeur de *F(x)* aux points de subdivision choisis (bornes de classes) est évidemment connue et coïncide avec les fréquences cumulées classiques. Entre deux points connus, l'hypothèse de densité constante conduit à une représentation graphique linéaire.

[6] Le choix d'intervalle "ouvert-fermé" conduit au choix alternatif : la fréquence des observations inférieures ou égales.

Dans le cadre de l'exemple numérique choisi, on obtient :



3. Paramètres de position et de dispersion

La paramétrisation peut se faire également par le biais de l'analyse classique : les paramètres de position réalisent le minimum d'une certaine définition (on constate entre autre que celle-ci est arbitraire ce qui n'est pas sans intérêt pour la compréhension en profondeur de la pertinence de l'utilisation de certains paramètres) de l'erreur totale (somme des carrés des erreurs pour l'introduction de la moyenne et somme des valeurs absolues des erreurs pour la médiane) et les valeurs de ces minima respectifs introduisent les paramètres de dispersion. On remarque que l'introduction justifiée de ces notions usuelles permet d'insister sur leurs limites et fait apparaître naturellement certaines de leurs propriétés. Une réflexion didactique en passant : le calcul de la médiane peut se faire soit analytiquement, ce qui n'est pas intéressant statistiquement puisque la notion intuitive de médiane n'apparaît pas dans la démonstration, soit par la méthode des emboîtements beaucoup plus riche de ce point de vue et également beaucoup plus simple.

Mathématiquement, on a ici une application élémentaire d'un calcul d'optimum pour la moyenne et l'introduction d'un artifice de calcul pour celui de la médiane.

3.1 Paramètres de position usuels

Optons pour la définition classique de Gauss qui agrège les erreurs en sommant leurs carrés. On parle alors de méthode des moindres sommes de carrés[7]. Notons a le résumé constant à construire, réalisant le minimum de notre agrégat.

[7] Cette méthode est parfois appelée plus simplement "méthode des moindres carrés", sous-entendant ainsi que l'agrégat ne pouvait être qu'une somme. Les travaux de Rousseeuw (1984 : voir [4]) nous ont montré que ce point de vue ne pouvait être défendu.

$$ET = \sum_{i=1}^{n}(x_i - a)^2$$

Vue comme fonction de a, ET est du second degré à concavité positive. Cette fonction admet donc un minimum unique qu'il convient de calculer. On a successivement :

$$\begin{aligned} \frac{dET}{da} &= \sum_{i=1}^{n} \frac{dET}{da}(x_i - a)^2 \\ &= \sum_{i=1}^{n} 2(x_i - a)(-1) \end{aligned}$$

Un calcul élémentaire nous conduit à constater que cette condition équivaut à la suivante : la somme des erreurs est nulle. Ce point est à retenir pour la construction d'une variance non biaisée. On retrouve bien entendu la définition classique de la moyenne :

$$\begin{aligned} \sum_{i=1}^{n}(x_i - a) &= 0 \\ \sum_{i=1}^{n} x_i - [na] &= 0 \\ a &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Classiquement, cette valeur se note $\bar{x}$ et représente la *moyenne arithmétique* des observations. On note donc :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

On constate que ce paramètre de position est en fait la somme pondérée uniformément de toutes les observations. Chaque valeur mesurée apporte la même part d'information dans la construction du résumé.

Un raisonnement quelque peu différent permet la construction de la médiane[8]. En utilisant la méthode des emboîtements, la définition classique de la médiane apparaît mais en plaçant le paramètre dans un contexte global en le liant à la métrique des valeurs absolues.

Le calcul d'une moyenne arithmétique dans le cas d'une présentation en tableau est également une bonne application du calcul intégral. Le passage au calcul de la moyenne dans le cas d'une variable réellement discrète permet l'extension de la notion au contexte continu.

En effet, pour une variable discrète pouvant prendre *k* valeurs différentes, la relation précédente devient :

[8] Nous renvoyons le lecteur intéressé au développement complet présenté dans *La statistique par l'analyse* (op. cit.). Il convient ici de classer les observations par ordre non décroissant et de travailler par inclusion successives d'intervalles.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^* = \frac{1}{n}\sum_{i=1}^{k} n_i.x_i = \sum_{i=1}^{k} f_i.x_i$$

On peut donc interpréter la moyenne comme la somme pondérée par leurs fréquences des valeurs différentes observées. Dans un contexte continu, et dans le cas d'une variable définie sur un intervalle [*a, b*], cette définition conduit à la relation[9] :

$$\bar{x} = \int_a^b x f(x) dx$$

Des calculs élémentaires conduisent à la formule classique de la moyenne dans le cas continu tout en illustrant de manière utile les intégrations de polynômes élémentaires. En notant a_i les bornes des k classes choisies et l_i leurs longueurs, on a :

$$\begin{aligned}
\bar{x} &= \sum_{i=1}^{k} \int_{a_i}^{a_{i+1}} x d_i dx \\
&= \sum_{i=1}^{k} \left[\frac{x^2}{2}\right]_{a_i}^{a_{i+1}} d_i \\
&= \sum_{i=1}^{k} \frac{1}{2}(a_{i+1}^2 - a_i^2) d_i
\end{aligned}$$

La factorisation de la différence de deux carrés fait apparaître le centre de classe et sa longueur. En utilisant la définition de la densité, une simplification donne :

$$\begin{aligned}
\bar{x} &= \sum_{i=1}^{k} (a_{i+1} - a_i) \frac{(a_{i+1} + a_i)}{2} \frac{f_i}{l_i} \\
&= \sum_{i=1}^{k} f_i.c_i
\end{aligned}$$

3.2 Paramètres de dispersion

Notre présentation permet en outre l'introduction de plusieurs variances. Considérons tout d'abord une variable discrète. Avec les notations introduites plus haut, on peut définir la variance comme la répartition sur l'ensemble des observations (n) de l'erreur totale minimale. On a :

[9] On peut considérer que le produit *f(x) dx* mesure la fréquence de l'observation x qui consiste en fait en l'appartenance à [*x* – *dx*/2 ; *x* + *dx*/2[.

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{k} n_i.(x_i - \bar{x})^2 \\ &= \sum_{i=1}^{k} f_i.(x_i - \bar{x})^2 \end{aligned}$$

Cette définition est biaisée. On sait en effet que la construction de la moyenne arithmétique passe par l'annulation de la somme des erreurs. Cette condition nous fait perdre un degré de liberté : comme la somme est nulle, la dernière peut être construite à partir des (*n*-1) premières. On corrige alors le paramètre[10] :

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i.(x_i - \bar{x})^2$$

Le passage au continu se fait d'une manière identique à ce que nous avons présenté pour la moyenne : calcul d'une somme continue (intégrale) pondérée par les fréquences observées. On vérifie tout d'abord que :

$$\begin{aligned} s^{*2} &= \int_a^b (x - \bar{x})^2.f(x)dx \\ &= \int_a^b x^2.f(x)dx - 2\bar{x}\int_a^b x.f(x)dx + \bar{x}^2\int_a^b f(x)dx \\ &= \int_a^b x^2.f(x)dx - 2\bar{x}^2 + \bar{x}^2 \\ &= \int_a^b x^2.f(x)dx - \bar{x}^2 \end{aligned}$$

On calcule ensuite la variance de manière directe.

[10] Remarquons ici que la nouvelle définition est plus proche de l'intuition pour des échantillons de faible effectif. Considérons par exemple le cas d'une mesure isolée. La formule initiale de la variance donne zéro. La formule modifiée conduit à l'indétermination 0/0. Cette dernière est justifiable : en effectuant une seule observation, on n'a aucune idée de la variabilité de la mesure effectuée. Pour obtenir *une* mesure de variabilité, il faut *deux* observations.

$$
\begin{aligned}
s^{*2} &= \int_a^b x^2.f(x)dx - \bar{x}^2 \\
&= \sum_{i=1}^{k} \int_{a_i}^{a_{i+1}} x^2.f(x)dx - \bar{x}^2 \\
&= \sum_{i=1}^{k} \int_{a_i}^{a_{i+1}} x^2.d_i dx - \bar{x}^2 \\
&= \sum_{i=1}^{k} d_i \left[\frac{x^3}{3}\right]_{a_i}^{a_{i+1}} - \bar{x}^2 \\
&= \sum_{i=1}^{k} d_i \frac{1}{3}\left[a_{i+1}^3 - a_i^3\right] - \bar{x}^2
\end{aligned}
$$

Cette expression est relativement peu usuelle. On peut développer la différence de cube et introduire une ultime simplification en détaillant la densité mais le résultat obtenu n'est guère sympathique. En introduisant une variance discrétisée, calculée sur base de la concentration des observations de chaque classe sur son centre, on arrive au résultat suivant :

$$s^2 = \sum_{i=1}^{k} f_i.(c_i - \bar{x})^2$$

En développant l'expression comme plus haut, on obtient :

$$
\begin{aligned}
s^2 &= \sum_{i=1}^{k} f_i.(c_i - \bar{x})^2 = \sum_{i=1}^{k} f_i.c_i^2 - \bar{x}^2 \\
&= \sum_{i=1}^{k} f_i \left[\frac{(a_{i+1} + a_i)}{2}\right]^2 - \bar{x}^2 \\
&= \sum_{i=1}^{k} f_i \frac{1}{4}\left[a_{i+1}^2 + 2a_i * a_{i+1} + a_i^2\right] - \bar{x}^2
\end{aligned}
$$

Pour arriver au résultat attendu, on calcule enfin :

$$
\begin{aligned}
s^{*2} - s^2 &= \sum_{i=1}^{k} f_i \frac{1}{12} \left[a_{i+1}^2 - 2a_i * a_{i+1} + a_i^2 \right] \\
&= \frac{1}{12} \sum_{i=1}^{k} f_i \left[a_{i+1} - a_i \right]^2 \\
&= \frac{1}{12} \sum_{i=1}^{k} f_i l_i^2
\end{aligned}
$$

Assez curieusement, ce dernier généralise la correction de Sheppard ... au signe près puisque dans cette dernière, la variance calculée en continu est inférieure à la variance calculée par discrétisation.

4. Conclusion provisoire

Les quelques exemples que nous avons introduits montrent les avantages d'une présentation simultanée. Il est clair que toute la statistique n'est pas contenue dans l'analyse et inversement que toute l'analyse ne sera pas accessible au moyen d'une présentation de ce type. Nous pensons néanmoins que ce petit pas va dans la bonne direction. L'expérience de cet enseignement simultané est en cours en première candidature en sciences commerciales au département économique de la Haute Ecole Ferrer. Nous ne manquerons pas de donner notre analyse des résultats obtenus quand bien même cette dernière devait s'avérer décevante.

BIBLIOGRAPHIE SOMMAIRE

[1] DROESBEKE, F. (2001) : *Eléments de statistique.* 4e édition. Bruxelles, Editions de l'Université de Bruxelleset Paris, Editions Ellipse.

[2] HAESBROECK, G – HENRY, V (2004) : *Pratique de la statistique descriptive : résolution et interprétation de problèmes.* Presses Ferrer/Céfal. ISBN 2-87130175-1.

[3] JUSTENS, D. (2004) : *La statistique par l'analyse.* Presses Ferrer/Céfal. ISBN 287130176-X.

[4] ROUSSEUW, PJ (1984) : Least Median os Sum of Squares Regression, *Journal of the American Statistical Association,* 79, 871-880.