ministère éducation nationale

ÉduSCOL [dgesco]

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

*Mathématiques*

*Baccalauréats professionnels*

# Ressources pour la classe
# Statistique et probabilités

## *- Commentaires et recommandations -*



*Juin 2009*

# STATISTIQUE ET PROBABILITÉS

## *COMMENTAIRES ET RECOMMANDATIONS*

## Les motivations

Un apprentissage précoce, puis régulier, des situations aléatoires est une nécessité pour répondre à un besoin social et professionnel de plus en plus prononcé dans ce domaine. De plus, cet apprentissage de l'aléatoire favorise la comparaison de notre enseignement avec celui d'autres pays de l'OCDE. L'enjeu est d'importance. Il s'agit de donner un sens rationnel aux notions de « risque », de « sondage », de « preuve statistique », de « différence significative »..., aidant à la compréhension de situations généralement empruntes d'incertitude et à la prise de décision en contexte aléatoire. Pour décrypter le monde moderne, participer au débat démocratique, exercer son esprit critique, optimiser ses activités professionnelles, « l'honnête homme » du XXIe siècle doit être éduqué aux méthodes statistiques et aux probabilités.

## Les choix généraux

Les précédents programmes de baccalauréat professionnel ne laissaient qu'une très faible place aux probabilités, pour certaines spécialités seulement, et avec une approche fondée sur le dénombrement des cas possibles. Cette approche a montré ses limites face aux enjeux décrits précédemment. Les nouveaux programmes des sections professionnelles s'inscrivent donc, dans ce domaine, dans la continuité de l'approche des probabilités initiée par le programme de la classe de troisième, mis en place à la rentrée 2008. La notion de probabilité s'approprie plus aisément par l'expérimentation et l'observation des fréquences, en répétant indépendamment l'expérience aléatoire. L'utilisation des T.I.C. (calculatrice ou tableur) favorise cet apprentissage en facilitant l'observation de la « loi des grands nombres ».
Compte tenu des enjeux qu'il présente en termes de formation de base, le domaine statistique - probabilités fait partie du tronc commun des différentes spécialités de baccalauréat professionnel.
Dans le domaine de la statistique descriptive, le module de statistique à une variable de seconde professionnelle constitue essentiellement une consolidation des contenus des programmes de collège, à affermir dans le cadre de situations, y compris lors de l'étude du module concernant les fluctuations des échantillons. En première professionnelle, le module de statistique à une variable ajoute aux indicateurs de dispersion mobilisables, l'écart type et l'écart interquartile. La statistique à deux variables est introduite en terminale professionnelle à l'aide des outils numériques, calculatrice et tableur, et en liaison avec des préoccupations non mathématiques, de la vie courante ou professionnelle et des sciences physiques.
Dans le domaine de l'aléatoire, l'objectif en seconde professionnelle est de comprendre, par l'expérimentation, que le « hasard » suit des lois. Les fluctuations des fréquences sont étudiées sur des échantillons aléatoires de même taille. Ceci permet de prendre conscience de l'esprit de la statistique, c'est-à-dire de la « variabilité naturelle » inhérente à un résultat statistique, et précise la notion de probabilité, estimée lorsqu'on augmente la taille de l'échantillon.
En première professionnelle est quantifiée plus précisément cette variabilité, en observant (sous certaines conditions) que si la fréquence d'un caractère dans une population est $p$, alors plus de 95 % des échantillons aléatoires de taille $n$ prélevés dans cette population donneront une fréquence de ce caractère comprise entre $p - \frac{1}{\sqrt{n}}$ et $p + \frac{1}{\sqrt{n}}$. Cette connaissance permet d'exercer un regard critique sur les données statistiques : il faut considérer comme « significative » (c'est-à-dire représentative d'une cause non aléatoire) une fréquence observée en dehors de cet intervalle. Plusieurs exemples d'activités de classe illustrent concrètement ce propos dans ce document. Un aperçu plus théorique de ces questions, pour le professeur, est aussi développé.
En terminale professionnelle, se mettent en place le langage et les méthodes (arbres, tableaux, diagrammes) permettant un calcul élémentaire des probabilités.
Cette progression facilite la poursuite d'études en sections de techniciens supérieurs, où, dans certaines spécialités, sont introduits les intervalles de confiance, les tests statistiques et les probabilités conditionnelles.

# Quelques éclairages complémentaires par module

## Modules 1.1 et 1.2 du programme de seconde professionnelle

### *À propos des indicateurs statistiques moyenne et médiane*

Les indicateurs de tendance centrale que sont la moyenne et la médiane sont abordés au collège : la moyenne dès la classe de quatrième, la médiane à partir de la troisième. Il convient cependant d'en réactiver le sens ainsi que les principales propriétés et de montrer que ces indicateurs jouent un rôle complémentaire. Davantage qu'un cours théorique, indigeste et inefficace, il s'agit d'étudier l'information apportée par ces indicateurs dans des exemples bien choisis. Il est en particulier nécessaire d'envisager quelques cas où moyenne et médiane diffèrent sensiblement. Dans le cas de valeurs obtenues selon une loi normale, et pour un échantillon assez important, moyenne et médiane sont pratiquement égales. Le terme « normal » laisse entendre que cette situation est assez fréquente.
Pourtant de nombreuses distributions « non normales » existent. Un exemple fréquemment choisi, et particulièrement parlant, est celui des revenus. En 2004, d'après l'INSEE, le revenu annuel déclaré par individu était en France, en moyenne de 18 030 €, avec une médiane de 15 766 €. Un Français souhaitant comparer ses revenus à ceux de ses compatriotes considérera la médiane. Un économiste désireux de se faire une idée de la « richesse » de la France pourra multiplier la moyenne par le nombre d'individus déclarant leurs revenus. Si la question de savoir pourquoi le revenu moyen est sensiblement supérieur au revenu médian se pose, une réponse peut être apporter par l'observation, sur des exemples numériques plus simples (les notes d'une classe à faible effectif à un devoir par exemple), que la moyenne est « sensible » aux valeurs extrêmes (en l'occurrence les très gros revenus), alors que, par sa définition même, la médiane ne l'est pas.
La détermination de la médiane nécessite un tri des données, ce qui peut être parfois compliqué, mais ne l'est pas avec un outil informatique (un tableur par exemple). Pour un nombre impair de données, la médiane est la valeur centrale, après tri. Pour un nombre pair de données, plusieurs « définitions » de la médiane sont possibles. Il est en général préférable de faire une interpolation, en effectuant la demi-somme des deux valeurs centrales après tri (la médiane est considérée parfois comme la première valeur pour laquelle on atteint ou on dépasse 50 % de l'effectif).
Dans la mesure du possible, il faut éviter de calculer une moyenne ou une médiane après un regroupement des données en classes, lequel constitue une perte d'information. Les T.I.C., avec un tableur par exemple, permettent de traiter un grand nombre de données. Toutefois, pour les besoins des exercices, il arrive que l'on ne dispose que de résultats regroupés en classes, sans avoir accès à la globalité des données. Il est alors impossible de connaître la valeur exacte de la moyenne ou de la médiane. Dans ce cas il faut se contenter d'estimer la moyenne sous l'hypothèse que les données sont regroupées au centre de chaque classe (sans précision illusoire dans le résultat du calcul), et de donner la classe médiane, c'est-à-dire la classe contenant la médiane. Il est parfois possible de donner une estimation de la médiane sous l'hypothèse que les données sont uniformément réparties à l'intérieur de la classe médiane, mais il est souvent ridicule de mettre en œuvre des moyens de calculs disproportionnés eu égard au côté arbitraire de l'hypothèse effectuée. Il faut donc éviter ce type d'exercice artificiel, qui n'est pas au programme.

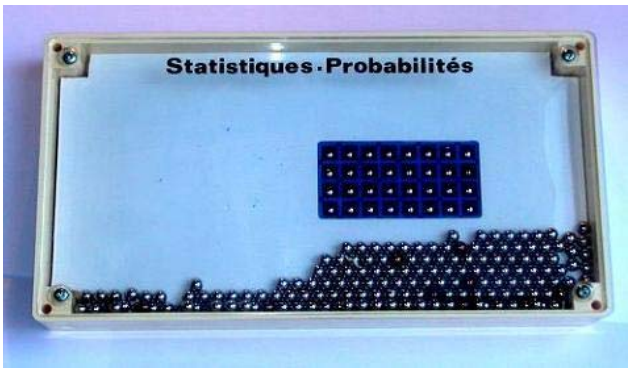
### *À propos des indicateurs statistiques étendue et quartiles*

L'étude de la dispersion est au cœur de l'activité statistique. S'il n'y avait pas de dispersion, la statistique serait de peu d'utilité. La notion d'écart type étant d'une compréhension plus difficile, elle n'est introduite qu'en première professionnelle. Il s'agit, en seconde professionnelle, de consolider l'approche de la dispersion déjà faite au collège avec comme outils l'étendue et les quartiles.
L'utilisation de l'étendue comme indicateur de dispersion peut paraître bien sommaire, cet indicateur étant obtenu simplement à partir de deux des données et une élémentaire soustraction. Mais cette simplicité est une qualité qui fait de l'étendue un indicateur de dispersion très utilisé dans les applications, en particulier en cours de fabrication industrielle ou à la réception de pièces manufacturées. Sur un échantillon aléatoire prélevé, l'étendue est calculée pour juger de la qualité. Il y a peu de risques d'erreurs de calcul. Le cas

échéant, la « robustesse » du procédé est améliorée en éliminant les éventuelles valeurs « aberrantes » avant le calcul de l'étendue.
Les quartiles, correspondant à des proportions de la population, sont conceptuellement assez simples à comprendre. Là encore il existe plusieurs définitions. La plus simple consiste à considérer que le premier quartile (respectivement le troisième quartile) correspond, après tri des données dans un ordre croissant de valeurs, à la première donnée pour laquelle on atteint ou dépasse 25 % de l'effectif (respectivement 75 %). Calculatrice et tableurs utilisent généralement d'autres définitions, fondées sur des interpolations, mais les différences éventuelles des résultats sont généralement peu significatives et ne posent pas de problème du point de vue de l'interprétation statistique qui peut être faite. Il peut ne pas y avoir de totale cohérence entre la définition choisie pour la médiane et celle qui correspondrait au deuxième quartile. Du point de vue de l'utilisation statistique, ceci est sans importance et ne doit pas être un sujet soulevé par le professeur.
La connaissance du premier et du troisième quartile (c'est-à-dire de la « boîte » du diagramme en boîte à moustaches) permet de situer la moitié (50 %) « centrale » de la population. Selon l'étendue de cette « moitié centrale », simplement « visualisée » en seconde et qui correspond à l'écart interquartile calculé en première, il est possible de comparer la dispersion de deux populations.

### *À propos des notions de probabilité et de fluctuation d'une fréquence selon les échantillons*

Pour comprendre les capacités attendues des élèves dans ce module du programme de seconde, et la façon dont il est possible de pratiquer en classe, le plus éclairant est sans doute de consulter les exemples d'activités présentées dans ce document, et qui ont été pratiquées dans des classes. Des éléments plus théoriques, présentés aussi dans ce document, permettent d'avoir ensuite un peu de recul.
Il paraît judicieux de commencer par étudier des exemples simples, se ramenant à des lancers successifs d'une pièce de monnaie, ou d'un dé, ou à des tirages dans une urne. Tout d'abord à l'aide de vraies expériences, réalisées par exemple avec un sac contenant des billes ou une bouteille contenant des perles de couleur.
Sur l'image suivante, la boîte contient 5 % de billes dorées, considérées comme défectueuses. Les 32 alvéoles, en secouant la boîte, constituent des échantillons aléatoires sur lesquels sont comptées les billes défectueuses ce qui permet d'étudier les fluctuations d'échantillonnage. La bouteille, dont les parois sont peintes, contient des perles rouges et jaunes. Seul le fond fait apparaître 5 perles. En secouant la bouteille, d'autres échantillons de 5 perles sont observables. En multipliant le nombre d'échantillons, la proportion des rouges et des jaunes dans la bouteille peut être estimée et utilisée pour illustrer ainsi l'approche fréquentiste d'une probabilité.



La partie expérimentale est suivie de l'utilisation de simulations avec la calculatrice ou le tableur (voir paragraphe II-3 pour des précisions techniques sur la simulation). L'utilisation d'un outil de simulation permet tout d'abord de multiplier les expériences à taille d'échantillon fixée, puis ensuite d'observer que l'ampleur des fluctuations a tendance à diminuer lorsque la taille de l'échantillon augmente.

Dans un second temps, il est essentiel de considérer des exemples issus de la vie courante, de l'économie, de la technologie, de l'environnement ou des thématiques proposées au BOEN (voir les exemples d'activités qui suivent). Ces exemples illustrent les enjeux de ce qui n'est pas un jeu.

## Module 1.1 du programme de première professionnelle

### *À propos de l'écart type et des boîtes à moustaches*

L'écart type est l'indicateur de dispersion à associer à la moyenne, alors que l'écart interquartile est celui associé à la médiane.
Le couple (moyenne, écart type) trouve sa pertinence en liaison avec les probabilités et en particulier avec la loi normale, qui est étudiée dans de nombreuses spécialités de techniciens supérieurs. Notamment, pour une distribution normale (courbe de Gauss), environ 95 % des valeurs sont situées autour de la moyenne à plus ou moins deux écarts types. Les distributions normales étant fréquemment rencontrées, en particulier dans les domaines professionnels industriels (pour le contrôle de qualité ou les erreurs de mesure), les élèves de première professionnelle peuvent étudier la proportion des termes d'une série appartenant à l'intervalle $[\,\bar{x}-2\sigma\,,\ \bar{x}+2\sigma\,]$.
L'écart type est pénible à calculer sans les T.I.C. et assez difficile à comprendre. Son calcul ne présente aucun intérêt à une époque où les T.I.C. permettent une obtention quasi immédiate (du moins dans le cas d'une calculatrice ; mais dans le cas d'un tableur, ce n'est vrai que si les valeurs ne sont pas regroupées, une petite procédure étant nécessaire dans le cas de valeurs regroupées). Il est donc hors de question de calculer un écart type plus ou moins « à la main » à l'aide d'un tableau. La chose importante est de travailler sur sa signification, par exemple, dans le cas de distributions relativement normales, en calculant la proportion des valeurs appartenant à l'intervalle $[\,\bar{x}-2\sigma\,,\ \bar{x}+2\sigma\,]$, ou en comparant la dispersion de deux séries d'écarts types différents. Il est intéressant de faire remarquer que deux séries de même écart type peuvent, si elles sont éloignées d'une distribution normale, avoir une distribution très différente, comme le montre l'exemple ci-dessous. C'est l'occasion de rappeler l'intérêt d'un graphique, qui peut montrer davantage qu'un simple résumé numérique.

**Série 1**

| valeurs | effectifs |
|---|---|
| 2 | 60 |
| 6 | 60 |
| 10 | 60 |
| 14 | 60 |
| 18 | 60 |

$N = 300$
$\bar{x} = 10$
$Me = 10$
$e = 16$
$\sigma \approx 5{,}66$

**Série 2**

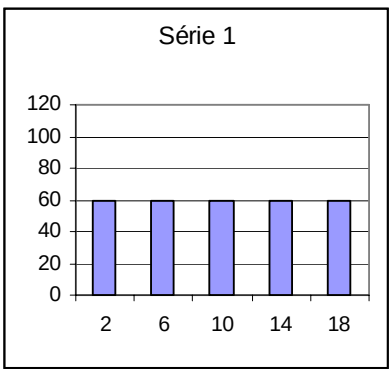
| valeurs | effectifs |
|---|---|
| 2 | 70 |
| 6 | 20 |
| 10 | 120 |
| 14 | 20 |
| 18 | 70 |

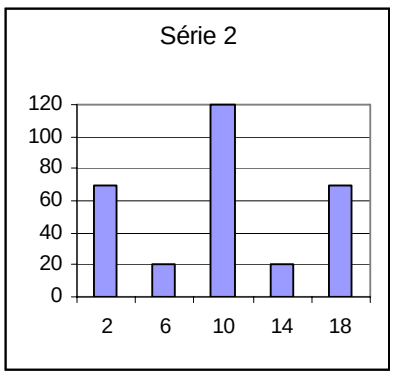
$N = 300$
$\bar{x} = 10$
$Me = 10$
$e = 16$
$\sigma \approx 5{,}66$

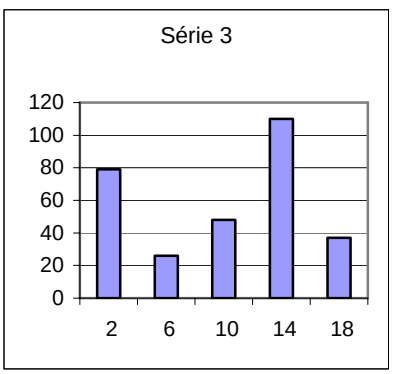
**Série 3**

| valeurs | effectifs |
|---|---|
| 2 | 79 |
| 6 | 26 |
| 10 | 48 |
| 14 | 110 |
| 18 | 37 |

$N = 300$
$\bar{x} = 10$
$Me = 10$
$e = 16$
$\sigma \approx 5{,}66$



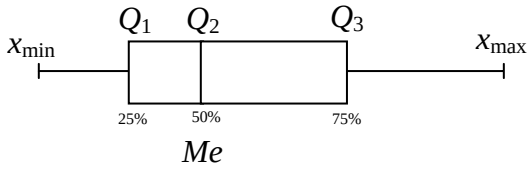




Le couple (médiane, écart interquartile), sans apporter les mêmes renseignements que le précédent, convient à toutes les distributions et est en particulier peu sensible aux valeurs extrêmes. Dans des

domaines où de nombreuses distributions non normales sont rencontrées, par exemple dans les secteurs professionnels tertiaires, il est privilégié et souvent associé à une représentation graphique en boîte à moustaches.

Il existe plusieurs types de diagrammes en boîte. La « boîte » est un rectangle limité par le premier et le troisième quartile et où figure la médiane. Les « moustaches » en revanche peuvent s'achever aux valeurs extrêmes, le minimum et le maximum de la série, ou aux premier et neuvième déciles, 10 % et 90 %, mais la notion de décile est hors programme.




La réalisation des diagrammes en boîte n'est pas exigible des élèves, mais ils doivent pouvoir lire et interpréter de tels graphiques. La convention retenue quant aux valeurs extrêmes des moustaches est indiquée.

## Module 1.1 du programme de terminale professionnelle

### *À propos de la statistique à deux variables*

Dans ce module, l'accent est mis sur l'utilisation des T.I.C. pour obtenir rapidement une représentation graphique du nuage de points et une droite (ou autre courbe) d'ajustement. Ceci permet de travailler davantage sur l'interprétation et l'utilisation de cet ajustement.
La méthode de Mayer, peu performante et absente des moyens habituels de calcul, n'est pas à envisager.