Cours : séance 4 (lundi 3 février)

Cinquième partie 5. Statistiques et probabilités.

Correction des derniers exercices de statistique.

Exercice 5.8 – Moyenne par classes.

Dans cette situation, la distinction entre ouvrier et cadre n'est pas pertinente. En effet la question est simplement celle des salaires. L'information intéressante est l'effectif de chaque *classe* de salaire. On regroupe les salariés en trois classes : la classe C_1 où le salaire est dans $[1000; 1500[$, la classe C_2 où le salaire est dans $[1500; 3000[$ et la classe C_3 où le salaire est dans $[3000; 8000[$.

1. On fait comme si tous les individus d'une classe donnée avaient pour salaire le centre de l'intervalle de salaire correspondant à la classe. Pour C_1 : $s_1 = 1,25$; pour C_2 : $s_2 = 2,25$; pour C_3 : $s_1 = 5,5$. Si on note N_1, N_2, N_3 les effectifs des classes C_1, C_2, C_3 la formule pour la moyenne "par classe" donne

$$\text{mpc} = \frac{\text{effectif de } C_1 \times \text{salaire "central"} s_1 + \text{effectif de } C_2 \times \text{salaire "central"} s_2 + \text{effectif de } C_3 \times \text{salaire "central"} s_3}{\text{effectif total}}$$

$$= \frac{N_1 \times s_1 + N_2 \times s_2 + N_3 \times s_3}{N_1 + N_2 + N_3}$$

(a) Dans l'entreprise $E1$, les effectifs des classes C_1, C_2, C_3 sont $250, 165, 30$. Dans l'entreprise $E2$, les effectifs des classes C_1, C_2, C_3 sont $200, 150, 50$.

On obtient donc $m_1 = \dfrac{250 \times 1,25 + 165 \times 2,25 + 30 \times 5,5}{445} = \dfrac{848,75}{445} \simeq 1,907$.

Et $m_2 = \dfrac{200 \times 1,25 + 150 \times 2,25 + 50 \times 5,5}{400} = \dfrac{862,5}{400} \simeq 2,156$.

(b) On reprend le calcul avec les effectifs indiqués pour les seuls ouvriers :

On obtient alors $o_1 = \dfrac{250 \times 1,25 + 150 \times 2,25 + 0 \times 5,5}{400} = \dfrac{650}{400} = 1,625$.

Et $o_2 = \dfrac{200 \times 1,25 + 100 \times 2,25 + 0 \times 5,5}{300} = \dfrac{475}{300} \simeq 1,583$.

(c) Enfin on reprend le calcul avec les effectifs indiqués pour les seuls cadres :

On obtient alors $c_1 = \dfrac{0 \times 1,25 + 15 \times 2,25 + 30 \times 5,5}{45} = \dfrac{198,75}{45} \simeq 4,417$.

Et $c_2 = \dfrac{0 \times 1,25 + 50 \times 2,25 + 50 \times 5,5}{100} = \dfrac{387,5}{100} = 3,875$.

2. Le salaire moyen de l'ensemble des salariés dans l'entreprise $E2$ est effectivement supérieur à celui dans l'entreprise $E1$. Et il est également vrai qu'en moyenne et les ouvriers, et les cadres de l'entreprise $E1$ sont mieux payés que dans l'entreprise $E2$. Mais le raisonnement du PDG1 est inexact : il affirme que la comparaison des moyennes sur les classes des ouvriers et des cadres permet de conclure quant à la moyenne générale, ce qui est une erreur, c'est bien le PDG2 qui dit vrai. L'explication est que, bien que le salaire moyen des cadres dans l'entreprise $E2$ soit moindre que l'entreprise $E1$, l'effectif des cadres dans l'entreprise $E2$ est nettement supérieur (en valeur absolue ET en proportion) à celui dans l'entreprise $E1$, ce qui permet de tirer la moyenne générale des salaires au dessus de celle de l'entreprise $E1$.

Exercice 5.9 – **Lecture d'un diagramme et médiane.**

1. (a) 12 archers ont gagné 6 points.

(b) Tous sauf cinq archers - c'est à dire 75 archers - ont gagné au moins 3 points.

(c) Il y a 34 archers qui ont gagné ≤ 6 points, 14 archers qui ont gagné 7 points, et 32 archers qui ont gagné ≥ 8 points. Le score médian est donc de 7 points.

2. (a) Il nous faut calculer le score moyen pour le club A :

$m_A = \dfrac{5\times 2+9\times 3+8\times 5+12\times 6+14\times 7+6\times 8+8\times 9+18\times 10}{80} = \dfrac{547}{80} \simeq 6,83.$

Le club B a donc un meilleur score moyen que le A.

(b) Les 10 meilleurs archers du club A ont tous gagné 10 points, la moyenne de leur score est donc 10, supérieure à la moyenne correspondante pour le club B.

Exercice 5.10 – **Organisation des données.**

1. Il faut commencer par réorganiser les données. Par ordre croissant on obtient :

$$12-16-16-16-16-17-17-17-21-21-21-21-22-22-22-22-22-$$

$$23-23-23-25-26-28-31-32-56-57-68-70-84\ .$$

que l'on pourrait aussi présenter par un diagramme en bâtons, ou le long d'une ligne, en portant un point sur la ligne pour chaque salaire, avec au dessus le salaire, et au dessous l'effectif des salariés ayant ce salaire. Ou encore :

12 (1); 16 (4); 17 (3); 21 (4); 22 (5); 23 (3); 25 (1); 26 (1); 28 (1); 31 (1); 32 (1); 56 (1); 57 (1); 68 (1); 70 (1); 84 (1)

Les trois classes de salaires sont C_1 ($\leq 17k$euros), C_2 ($18k$euros$\leq\leq 35k$euros) et C_3 ($\geq 36k$euros). Les effectifs correspondants sont $8, 17, 5$. Les angles correspondants sont donc $\alpha_1 = \dfrac{8}{30}360 = 96$ degrés, $\alpha_2 = \dfrac{17}{30}360 = 204$ degrés, $\alpha_3 = \dfrac{5}{30}360 = 60$ degrés. D'où le camembert.

2. Avec un regroupement en classes définies par un intervalle d'étendue $5k$euros, les effectifs sont

$10 \leq$ Salaire < 15 : 1
$15 \leq$ Salaire < 20 : 7
$20 \leq$ Salaire < 25 : 12
$25 \leq$ Salaire < 30 : 3
$30 \leq$ Salaire < 35 : 2
$55 \leq$ Salaire < 60 : 2
$65 \leq$ Salaire < 70 : 1
$70 \leq$ Salaire < 75 : 1
$80 \leq$ Salaire < 85 : 1

Il y a beaucoup de classes qui son vides ! ($35 \leq$ Salaire < 40 ; $40 \leq$ Salaire < 45 ; $45 \leq$ Salaire < 50 ; $50 \leq$ Salaire < 55 ; $60 \leq$ Salaire < 65 ; $75 \leq$ Salaire < 80)

3. Pour avoir moins de classes creuses ou même vide, on élargit l'étendue des salaires de chaque classe. Par exemple si on impose une étendue de $15k$euros, les effectifs deviennent

$10 \leq$ Salaire < 25 : 20
$25 \leq$ Salaire < 40 : 5
$40 \leq$ Salaire < 55 : 0
$55 \leq$ Salaire < 70 : 3
$70 \leq$ Salaire < 85 : 2

L'histogramme est moins creux...mais il montre encore l'éclatement de l'échelle des salaires.

4. Il y a 12 salariés dont le salaire (annuel) est $\leq 21k$ euros, 5 salariés dont le salaire est $= 22k$ euros et 13 salariés dont le salaire est $\geq 23k$ euros. Le salaire médian est donc de $22k$ euros : la moitié des salariés a un salaire $\leq 22k$euros, et l'autre moitié a un salaire $\geq 22k$euros.

Passons aux quartiles. Notre série a été réordonnée par valeurs croissantes.

D'abord on cherche une valeur du salaire Q_1 telle que au moins $\frac{1}{4}(= 25\%)$ de l'effectif a un salaire $\leq Q_1$, et de plus Q_1 est la plus petite valeur possible avec cette propriété. Quel est le plus petit effectif qui englobe un quart de l'effectif total ?

On a $30/4 = 7,5$. Un effectif (entier !) $\geq 7,5$ est en fait ≥ 8.

Ainsi le premier quartile est la huitième valeur de la série.

Le 8ème salaire dans l'ordre croissant vaut $17k$euros. Donc $Q_1 = 17k$euros.

Le troisième quartile est le plus petit salaire Q_3 telle que au moins $\frac{3}{4}(= 75\%)$ de l'effectif a un salaire $\leq Q_3$. Quel est le plus petit effectif qui englobe trois quarts de l'effectif total ?

On a $\frac{3}{4}30 = 22,5$. Un effectif (entier !) $\geq 22,5$ est en fait ≥ 23.

Ainsi le troisième quartile est la vingt-troisième valeur de la série.

Le 23ème salaire dans l'ordre croissant vaut $28k$euros. Donc $Q_3 = 28k$euros.

5. La boîte à moustache, grâce aux deux rectangles, fait bien ressortir *l'intervalle interquartile*, c'est à dire $Q_3 - Q_1$. Par définition même des quartiles il y a la moitié des employés dont le salaire est dans l'intervalle $[Q_1; Q_3]$. Dans ce cas on voit que l'intervalle $[Q_1; Q_3]$ représente une assez petite fraction de l'étendue des salaires de cette entreprise ().

6. D'abord réorganisons la série des salaires dans l'ordre croissant :

$$11 - 15 - 17 - 18 - 18 - 22 - 22 - 23 - 23 - 24 - 24 - 25 - 25 - 26 - 26 - 27 - 27 - 31 - 42 - 45$$

Puisque c'est la répartition des salaires que l'on veut comparer il vaut mieux utiliser un deuxième diagramme à moustache. En choisissant l'échelle pour que l'intervalle entre les salaires extrêmes soit le même.

L'effectif total est 20. Donc $\frac{1}{4}20 = 5$. Pour englober un quart de la série, il faut aller jusqu'à la 5ème valeur, donc le 1er quartile est 18.

La moitié de l'effectif est 10. Donc la médiane est $\frac{\text{10ème salaire } + \text{11ème salaire}}{2} = \frac{24 + 24}{2} = 24.$

Enfin $\frac{3}{4}20 = 15$. Pour englober trois quarts de la série, il faut aller jusqu'à la 15ème valeur, donc le 3ème quartile est 26.

Quand on trace les boîtes à moustache en remettant à l'échelle pour que les étendues de salaires aient la même longueur, on voit que l'intervalle interquartile de la deuxième entreprise occupe une proportion plus grande de l'intervalle des salaires. Donc les salaires sont moins concentrés autour de la médiane, ils sont mieux répartis sur toute l'étendue.

Exercice 5.11 – Organisation des données.

1. Le total des salaires des hommes est 7450 (de sorte que le salaire moyen des hommes est $1862,5$). Le total des salaires des trois femmes est $3 \times 1700 = 5100$. Pour obtenir en rajoutant un salaire la même moyenne il suffit de rajouter un salaire qui compense la différence entre 7450 et 5100, soit 2350.

2. Il y a trois salaires, mettons x, y, z dans l'ordre croissant. Le salaire médian est donc y, ce qui nous donne $y = 1875$. La somme des trois salaires est $x + y + z = 5100$ (voir question précédente). Donc $x + 1875 + z = 5100$, soit $x + z = 3225$. Enfin l'étendue des salaires est de 1000 euros, donc $z - x = 1000$. En ajoutant les deux relations on obtient $2z = 4225$, d'où $z = 2112,5$ puis $x = 3225 - z = 1112,5$.

Voici le contexte dans lequel apparaissent les probabilités (au concours et au collège).

Il s'agit TOUJOURS d'étudier les résultats d'une expérience aléatoire. Dans une telle expérience il y a plusieurs "résultats possibles", bien décrits. Les résultats possibles sont appelés issues, notées ω dans ce qui suit. ATTENTION, les issues sont les résultats de l'expérience aléatoire, mais il y a une autre sorte de "résultats" dans les énoncés. Car dans un deuxième temps on s'intéresse en général à un caractère ou une "mesure" effectuée sur une issue. Cette mesure peut être un nombre, mais pas forcément. Il y a donc un deuxième type de résultat : ce que donne la mesure de l'issue. Pour éviter les confusions à partir de maintenant j'appelle issue le premier type de résultat : les résultats de l'expérience aléatoire.

Pour traiter correctement un exercice de probabilité il est indispensable de commencer par clarifier (1) quelle est l'expérience aléatoire, (2) quelles sont les issues possibles, (3) quelle est la mesure effectuée sur les issues (ou même quelles sont les mesures effectuées...).

Quand on fait l'expérience on ne sait pas avec certitude quelle issue elle aura. Le but des probabilités est de calculer les "chances" que l'on a d'obtenir telle ou telle issue.

L'ensemble de toutes les issues possibles de l'expérience aléatoire est appelé l'univers (des possibles), souvent noté Ω. Dans nos exemples il y aura toujours un nombre fini d'issues possibles : l'univers est un ensemble fini, que l'on peut représenter par une "patate".

Un ensemble d'issues possibles est appelé un événement. Les événements sont analogues aux classes en statistique. Et on peut représenter un événement comme une sous-patate de la patate universelle.

L'événement certain contient toutes les issues possibles : c'est l'univers entier.

L'événement impossible ne contient aucune issue : c'est l'ensemble vide.

Un événement consistant en une unique issue est appelé événement élémentaire. On n'entoure qu'une croix.

Dans les exercices de probabilités, voici comment sont décrits les événements en général. A partir de chaque issue ω on fait une observation, une mesure $f(\omega)$. Cette mesure est décrite au tout début de l'énoncé ; ATTENTION de ne pas confondre l'issue ω et le résultat de la mesure, $f(\omega)$. Cela changerait l'univers et fausserait le calcul des probabilités. Alors un événement A sera décrit par une propriété de $f(\omega)$: A est l'ensemble des issues ω telles que $f(\omega)$ a une certaine propriété. On identifie très vite l'événement et la propriété qui le caractérise. On peut alors manipuler les propriétés pour produire de nouveaux événements. Par exemple si A correspond à la propriété $\mathscr{P}$ et B à $\mathscr{Q}$, on obtient un nouvel événement décrit par la propriété "ω vérifie les deux propriétés $\mathscr{P}$ et $\mathscr{Q}$" (c'est en fait l'intersection $A \cap B$), ou alors la propriété "ω vérifie la propriété $\mathscr{P}$ ou ω vérifie la propriété $\mathscr{Q}$" (c'est en fait la réunion $A \cup B$).

Si A est un événement caractérisé par une propriété, l'événement contraire de A est celui qui est caractérisé par la propriété contraire (la négation de la propriété). On note $\overline{A}$ l'événement contraire de A. Les sous-patates correspondantes partitionnent l'univers en deux sous-ensembles disjoints.

Evidemment le contraire du contraire de A, c'est A ! (moins par moins ça fait plus).

Deux événements A et B sont incompatibles s'il n'existe aucune issue ω commune à A et B. Aucun résultat de l'expérience aléatoire n'a à la fois les caractéristiques de A et les caractéristiques de B : la propriété caractérisant A est incompatible avec la propriété caractérisant B. Cela revient à dire que la propriété qui définit B est un cas particulier de la négation de la propriété qui caractérise

A. Autrement dit B est contenu dans $\overline{A}$. Ou encore que l'événement constitué des issues qui vérifient la propriété de A ET la propriété de B est impossible.

Passons aux probabilités. L'univers Ω, fini, est supposé constitué de n issues : $\Omega = \{\omega_1, \cdots, \omega_n\}$.

On fait l'hypothèse que chaque événement A possède une probabilité $p(A)$: c'est un nombre de $[0;1]$, qui doit prédire la "chance" que l'événement a de se produire. Probabilité proche de 0, faible chance de se produire. Probabilité proche de 1, forte chance de se produire.

En particulier on demande que si $A = \emptyset$ est l'événement impossible, alors $p(A) = 0$. Et si $A = \Omega$ est l'événement certain, alors $p(A) = 1$.

On veut que $p(\overline{A}) + p(A) = 1$, ou encore $p(\overline{A}) = 1 - p(A)$.

On veut aussi que si A et B sont deux événements incompatibles alors $p(A \text{ ou } B) = p(A) + p(B)$ (cela permet de calculer la probabilité d'un événement compliqué C en le décomposant sous la forme $C =$ "ou bien A, ou bien B").

Attention si A et B ne sont pas incompatibles, alors en général $p(A \text{ ou } B) < p(A) + p(B)$.

Pour simplifier, lorsque $A = \{\omega\}$ est un événement élémentaire on note $p(\omega)$ sa probabilité au lieu de $p(\{\omega\})$...

Alors si $A = \{\omega_1, \omega_2, \ldots, \omega_k\}$ on a

$$p(A) = p(\omega_1) + \cdots + p(\omega_k)$$

En particulier la somme des probabilités de toutes les issues possibles vaut 1 :

$$p(\omega_1) + \cdots + p(\omega_n) = 1$$

Dans (presque) tous les exercices on suppose que tous les événements élémentaires ont la même probabilité (hypothèse d'équiprobabilité). Donc en fait $\boxed{p(\omega) = \dfrac{1}{n}}$.

Et si un événement A est constitué de k issues alors $p(A) = \dfrac{1}{n} + \cdots + \dfrac{1}{n}$, donc $\boxed{p(A) = \dfrac{k}{n}}$, i.e.

$$\boxed{p(A) = \frac{\text{Nombre d'issues réalisant } A}{\text{Nombre total d'issues}} = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas total}}}$$

Ainsi dans la pratique les calculs de probabilité sont des manipulations de fractions $\dfrac{k}{n}$ où n est le nombre d'issues possibles (donc fixé quand le cadre est fixé), et k est le nombre d'issues d'un événement (qui, lui, change au cours de l'exercice).

Exercice 5.2 – Vocabulaire des issues et des événements.

Il faut toujours clarifier l'univers des possibles au début d'un problème de probabilités ! Ici la partie aléatoire du processus, c'est, après avoir fait tourner la roue, d'obtenir un des sept secteurs tracés dessus (en position haute par exemple). Donc les issues possibles sont les sept secteurs angulaires de la roue, mettons $s_1, s_2, s_3, s_4, s_5, s_6, s_7$, avec s_1 le secteur où est marqué L, s_2 le secteur où est marqué O, etc...

1.

- Il y a donc bien 7 issues possibles. Donc c'est VRAI.

Ensuite on effectue une "mesure" à partir du secteur issu de l'expérience aléatoire : simplement on lit la lettre qui y est écrite. Il y a six résultats de lecture possible, puisque le E est lu au secteur s_4 et au secteur s_7.

- "Obtenir une consonne" est une propriété de la lecture effectuée sur une issue, ce n'est pas une issue (FAUX).

- "Obtenir une consonne" est l'événement A réalisé par les issues s_1, s_3, s_5, $A = \{s_1, s_3, s_5\}$ (VRAI).

- L'événement "Obtenir une lettre du mot VICTOIRE" est réalisé par les issues $s_2, s_3, s_4, s_5, s_6, s_7$. Trois quelconques de ces issues réalisent donc l'événement. Donc c'est VRAI. Mais si on considère que l'affirmation signifie : il y a exactement trois issues qui réalisent l'événement, alors c'est FAUX (puisqu'il y en a six).

2. "Obtenir une consonne" et "Obtenir une voyelle" sont des événements contraires.

"Obtenir une lettre du mot MAMAN" est un événement impossible.

"Obtenir une lettre du mot ETOILER" est un événement certain.