Cours-TD Probabilités Statistiques

S3

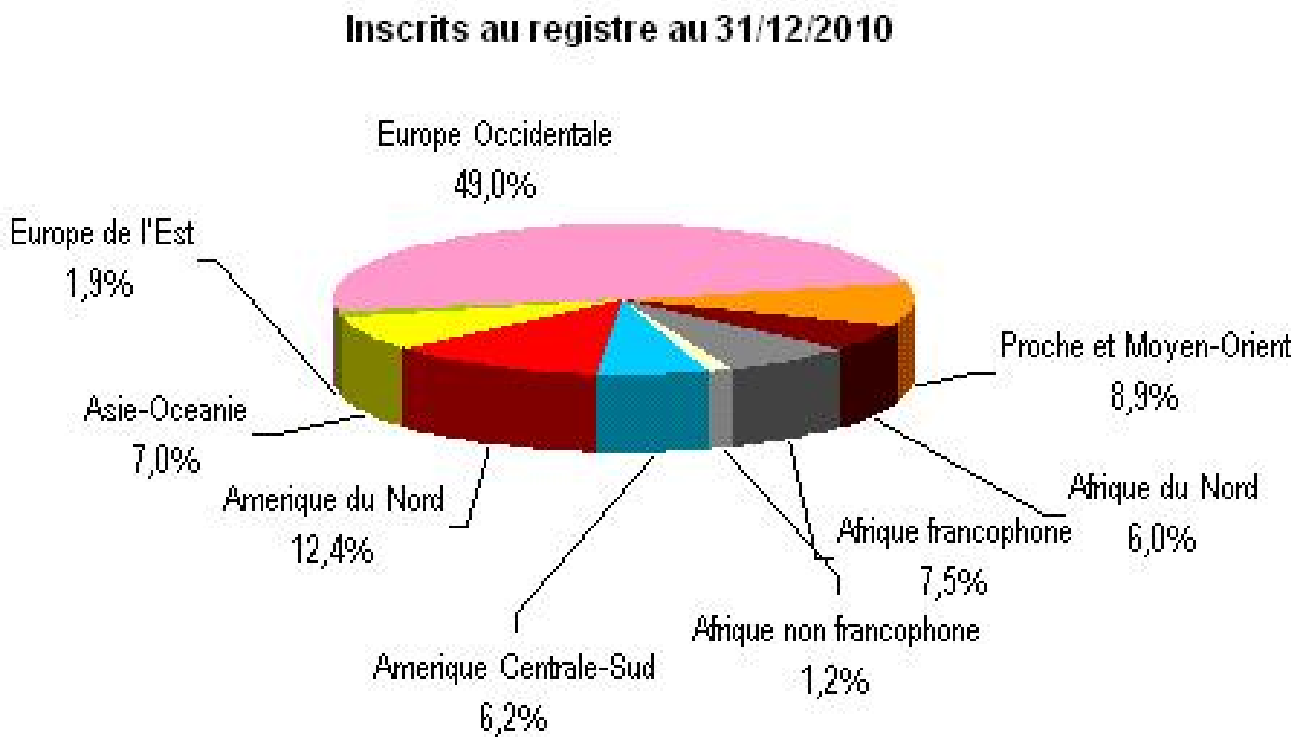
16 Décembre 2020

Introduction

La Statistique peut être définie comme la sciences des données, de leur collecte, de leur analyse et de leur présentation. S'il peut sembler facile de collecter quelques données et de calculer quelques valeurs, comme des moyennes, puis de les interpréter abusivement, il est en revanche beaucoup plus complexe de le faire dans un cadre scientifique rigoureux. Tout scientifique, qu'il soit informaticien, biologiste, chimiste, économiste, etc., et même tout citoyen, est confronté à des données et statistiques qu'il doit savoir analyser avec recul. L'étude en pratique des performances de différents composants informatiques se fait par des études statistiques : comment collecter les données (quels tests faire), comment les analyser (que permettent-ils de conclure) et comment les présenter. Il est facile de se tromper sur l'un des trois points et d'arriver à des conclusions sans fondement.

Graphiques

Lorsque l'on présente des résultats statistique pour comparer des proportions, il est fréquent d'utiliser des camemberts ou des histogrammes. Un exemple est donné ci-dessous :



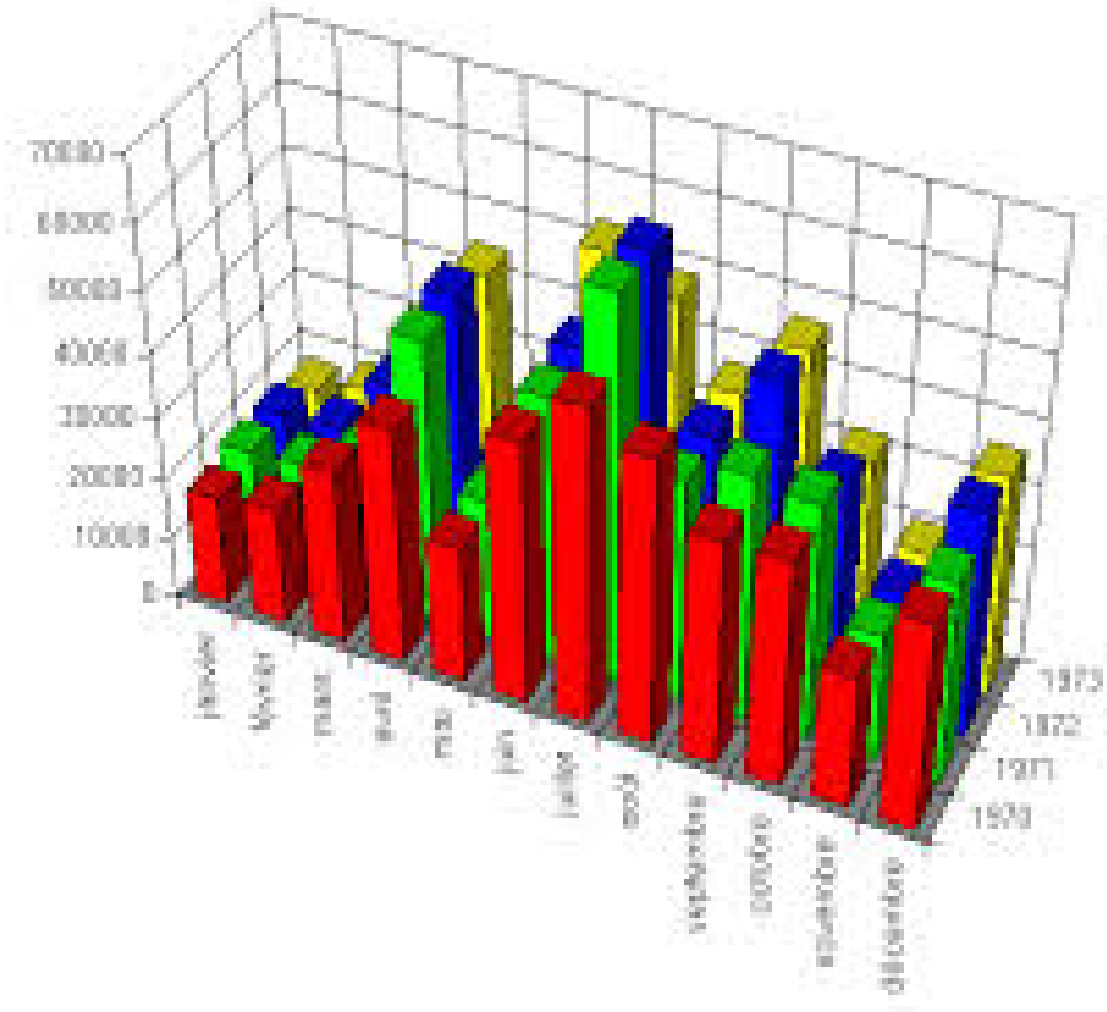
Quelles remarques peut-on faire sur cette représentation ?

Graphiques

Cette présentation, en relief, est fréquemment utilisée car *plus jolie*. Cependant, elle est tendancieuse, car ce que l'oeil compare les aires des différents secteurs, et l'utilisation de tranches (pour le relief) donne visuellement un poids plus fort aux données placées devant. Par exemple, l'Europe Occidentale occupe 50% de la surface de l'ellipse marquant le camenbert, mais bien moins de 50% de la surface totale du dessin, alors que cela devrait être le cas. De même, on peut remarque que l'Europe de l'Est paraît occuper une part moins importante que l'Afrique non francophone, ce qui n'est pas le cas. De même l'Amérique Centrale-Sud occupe plus place que le Proche et Moyen-Orient sur le dessin, alors que cela ne devrait pas être le cas. Dans un camembert en relief, les données placées sur l'avant on tendance à être surestimées. En pratique il convient donc de ne pas l'utiliser ou, lorsque cela est fait, choisir une très faible épaisseur relativement au rayon du disque et une inclinaison par trop importante, afin de minimiser le biais. De même tout découpage en tranche (en faisant ressortir une tranche) ajoute de l'épaisseur visuelle et accentue l'effet visuel. Il est aussi important de faire attention aux couleurs : des couleurs vives, comme le rouge, attirent plus l'oeil.

Graphiques

Que pensez vous de la représentation de la figure ?



Paradoxe de Simpson

On considère deux lycées, notés A et B. On observe les taux de réussites suivants dans les deux lycées.

	Lycée A	Lycée B
garçons	70%	71.05%
filles	75%	80%

1. Pouvez vous dire quel lycée à le meilleur taux de réussite ?
2. On considère maintenant les effectifs des deux lycées donnés ci-dessous. Calculer le taux de réussite de chaque lycée. Qu'en déduire ?

	Lycée A	Lycée B
garçons	100	190
filles	100	10

Paradoxe de Simpson

On considère deux lycées, notés A et B. On observe les taux de réussites suivants dans les deux lycées.

	Lycée A	Lycée B
garçons	70%	71.05%
filles	75%	80%

1. Non,on ne peut pas car on ne peut pas faire des moyennes de pourcentages.
2. On considère maintenant les effectifs des deux lycées donnés ci-dessous. Calculer le taux de réussite de chaque lycée. Qu'en déduire ?

	Lycée A	Lycée B
garçons	100	190
filles	100	10

Paradoxe de Simpson

On considère deux lycées, notés A et B. On observe les taux de réussites suivants dans les deux lycées.

1. Pouvez vous dire quel lycée à le meilleur taux de réussite ?
2. On considère maintenant les effectifs des deux lycées donnés ci-dessous. Calculer le taux de réussite de chaque lycée. Qu'en déduire ?

	Lycée A	Lycée B
garçons	100	190
filles	100	10

Taux de réussite du Lycée A : $\dfrac{70+75}{200} = 72,5\%$

Taux de réussite du Lycée B : $\dfrac{135+8}{200} = 71,5\%$

Contrairement à ce que l'on aurait pu penser c'est le lycée A qui a un meilleur taux de réussite.

Paradoxe de Simpson

Un exemple réel provenant d'une étude médicale sur le succès de deux traitements contre les calculs rénaux permet de voir le paradoxe sous un autre angle. La première table montre le succès global et le nombre de traitements pour chaque méthode.

Taux de succès (succès/total)

Traitement A	Traitement B
78 % (273/350)	83 % (289/350)

Cela semble révéler que le traitement B est plus efficace. Maintenant, en ajoutant des données concernant la taille des calculs rénaux, la comparaison prend une autre tournure :

Taux de succès (succès/total) (à gauche petits calculs, à droite gros calculs)

Traitement A	Traitement B
93 % (81/87)	87 % (234/270)

Traitement A	Traitement B
73 % (192/263)	69 % (55/80)

L'information au sujet de la taille des calculs a inversé les conclusions concernant l'efficacité de chaque traitement. Le traitement A est maintenant considéré comme plus efficace dans les deux cas. Le traitement le plus efficace peut être déterminé grâce à l'inégalité entre les deux rapports (succès/total). Le rebroussement de cette inégalité, qui conduit au paradoxe, se produit à cause de deux effets concurrents : La variable supplémentaire (ici la taille) a un impact significatif sur les rapports. Les tailles des groupes qui sont combinés quand la variable supplémentaire est ignorée sont très différentes.

Paradoxe de Simpson

Retournons à la sécurité routière. Considérons la différence du nombre d'accidents sur un parcours Paris-Besançon entre rouler avec une petite citadine et une grosse berline (les chiffres sont inventés).

	citadines	berlines
nombre de trajets	1300	2500
nombre d'accidents	10	11
taux d'accidents	0.76%	0.44%

Avec ces chiffres, la berlines semble beaucoup plus sûre que la citadine. Mais si :

nombre de trajets sur autoroute	300	2000
nombre d'accidents sur autoroute	1	6
taux d'accidents sur autoroute	0.33%	0.3%
nombre de trajets sur nationale	1000	500
nombre d'accidents sur nationale	9	5
taux d'accidents sur nationale	0.9%	1%

En fait, les deux types de voitures semblent avoir a peu près le même taux d'accident, les premiers résultats sont biaisés par le fait que beaucoup plus de personnes roulant avec une petite citadine prennent plutôt la nationale. Mais il faudrait aussi prendre en compte les conditions de route (pluie, nuit, etc.) pour pouvoir conclure quelque chose. Connaître les variables significatives à prendre en compte pour faire des probabilité demande un expertise certaines, à la fois en statistique mais surtout sur le sujet d'étude.

Paradoxe de Simpson

Ce qu'il est important de retenir, c'est qu'il est facile (et malheureusement usuel) d'utiliser des représentations graphiques trompeuses. Par ailleurs, il est parfois difficile d'obtenir les valeurs pertinentes permettant d'en tirer des conclusions utiles. Deux exemples : il y a en France plus de commotions cérébrales dues à des accidents de la route sur des piétons que sur des vélos. Certaines associations d'utilisateur de vélos l'utilise pour que le port du casque ne soit pas rendu obligatoire. Mais ces chiffres doivent être rapportés aux volume du trafic vélo et du trafic piétons. Se pose alors la question de choisir la mesure du trafic (en temps passé ou en kilomètres parcourus ?) et de savoir le mesurer (actuellement on ne sait le faire qu'avec un facteur 100 près). Toujours sur la sécurité routière, il est par exemple aussi difficile de comparer la dangerosité des routes entre deux pays comme la France et le Royaume-Uni. On peut bien entendu compter le nombre d'accidents mortels selon des critères identiques, mais doit on rapporter ce total au nombre de véhicule ? au nombre d'habitants ? au nombre de kilomètres parcourus ? au nombre de kilomètres de route ? Par ailleurs, le trafic n'est pas du tout le même en France (pays de transits entre l'Europe du Nord et du Sud) et le Royaume-Uni qui est une île. De même, on sait que les conditions météorologiques influencent sensiblement le nombre d'accidents. Comment le prendre en compte ? En conclusion, il faut en statistique avoir les bonnes données (cela demande une expertise métier en général), en nombre suffisant, et savoir restituer ces résultats convenablement.

Phénomène de Rogers

Le phénomène de Rogers peut traduire qu'il est possible, avec de même données, de donner des résultats aux conclusion différentes. Cela montre qu'il est possible de truquer des résultat sans pour autant truquer des données, par un moyen purement mathématique. Imaginons un IUT dans lequel il y a deux groupes de niveaux A et B. Les étudiants du groupe A sont, en général, meilleur que ceux du groupe B. Une année, les groupes A a eu 14.2 de moyenne et le groupe B a eu 8 de moyenne.
La seconde année, les notes du groupe A sont : 16, 15, 15, 13, 11. les notes du groupe B sont 13, 8, 6 et 3. La moyenne du groupe A est de 14 et celle du groupe B de 7.5. Les résultats semblent donc moins bon que l'année précédente. L'enseignant décide alors de faire passer le plus mauvais étudiant du groupe A dans le groupe B. Les notes du groupe A deviennent alors 16, 15, 15 et 13, soit une moyenne de 14.75. Celle du groupe B sont 13, 11, 8, 6 et 3, soit une moyenne de 8.2. De cette façon, les moyennes des groupes A et B ont augmenté !

Paramètres statistiques

Moyenne

On considère la série statistique suivante :

Valeurs	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

L'effectif total est : $n_1 + n_2 + ... + n_p = N$.

La moyenne de cette série est le nombre noté $\bar{x}$ défini par :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + ... + n_p x_p}{N}.$$

Cas particuliers :

- ⋆ Si la série contient une seule fois chacune des valeurs x_1, x_2,...,x_p alors : $\bar{x} = \frac{x_1 + x_2 + ... + x_p}{p}$.
- ⋆ Si f_1, f_2,...,f_p sont les fréquences associées, alors $\bar{x} = x_1 f_1 + x_2 f_2 + ... + x_p f_p$.
- ⋆ Lorsque les valeurs de la série sont regroupées en classes, on utilise les centres de classes pour calculer une approximation de la valeur moyenne.

Remarque : La moyenne est un indicateur de position : elle donne une tendance centrale de la série. Cependant, elle n'indique rien quant à la dispersion des valeurs de la série autour de la moyenne. Lorsque les valeurs sont très dispersées, la moyenne ne représente pas la série de manière significative.

Linéarité de la moyenne

Lorsque toutes les valeurs de la série sont transformées par une fonction affine $x \mapsto mx + p$, la moyenne de la nouvelle série est $m\bar{x} + p$.

Paramètres statistiques

Variance et écart-type

On appelle **variance** d'une série statistique la moyenne des carrés des écarts des valeurs à la moyenne. On note V ce nombre.
On appelle **écart-type** de la série le nombre positif σ défini par $\sigma = \sqrt{V}$.

Remarques : L'écart-type est un indicateur de dispersion : c'est un nombre positif qui mesure les variations des valeurs de la série autour de la moyenne $\bar{x}$.
Dans la pratique, on obtient la valeur approchée de σ grâce à la calculatrice.

Paramètres statistiques

Médiane et quartiles (rappels)

- La **médiane** (notée M_e) est la valeur du caractère pour lequel au moins 50% des caractères lui sont inférieurs ;
- Le **premier quartile** (notée Q_1) est la valeur du caractère pour lequel au moins 25% des caractères lui sont inférieurs ;
- Le **troisième quartile** (noté Q_3) est la valeur du caractère pour lequel au moins 75% des caractères lui sont inférieurs.

Paramètres statistiques

Méthode de calcul : On range les N valeurs par ordre croissant. Puis pour déterminer la médiane :

- si N est impair, M_e est la valeur centrale,
- si N est pair, M_e est la moyenne des deux valeurs centrales.

Pour déterminer Q_1 : on calcule $\frac{N}{4}$, le rang de Q_1 est l'entier supérieur ou égal au résultat obtenu.
Pour déterminer Q_3 : on calcule $\frac{3}{4}N$, le rang de Q_3 est l'entier supérieur ou égal au résultat obtenu.

Exemples :

- La médiane de la série de valeurs : 16 - 11 - 22 - 19 - 21 - 12 - 18 est ...
- La médiane de la série de valeurs :7 - 7 - 8 - 9 - 10 - 10 - 11 - 11 - 11 - 12 - 14 - 15 est ...
- Les quartiles Q_1 et Q_3 de la série de valeurs : 18 - 16 - 22 - 22 - 24 - 26 - 18 - 20 - 18 - 20 sont ...
- La médiane des données de ce tableau est :

Nombre de spams	0	1	2	3	4	5	6
Effectif	3	1	5	3	1	2	1
ECC							

Le premier quartile est : ...

Le troisième quartile est : ...

Paramètres statistiques

Exemples :

- Pour trouver la médiane de la série de valeurs : 16 - 11 - 22 - 19 - 21 - 12 - 18, on les range dans l'ordre croissant puis comme il y a 7 valeurs, la médiane est la 4ème valeur, donc $\boxed{18}$
- La médiane de la série de valeurs :7 - 7 - 8 - 9 - 10 - 10 - 11 - 11 - 11 - 12 - 14 - 15 est $\boxed{10{,}5}$ (moyenne entre la 6ème et 7ème valeur)
- Pour trouver les quartiles Q_1 et Q_3 de la série de valeurs : 18 - 16 - 22 - 22 - 24 - 26 - 18 - 20 - 18 - 20, on range les valeurs dans l'ordre croissant. Il y a 10 valeurs, donc Q_1 est la 3ème valeur, soit $\boxed{18}$ et Q_3 est la 8ème valeur, soit $\boxed{22}$.
- La médiane des données de ce tableau est la moyenne entre la 8ème et la 9ème valeur, soit $\boxed{2}$

Nombre de spams	0	1	2	3	4	5	6
Effectif	3	1	5	3	1	2	1
ECC	3	4	9	12	13	15	16

Le premier quartile est la 4ème valeur soit $\boxed{1}$.

Le troisième quartile est la 12ème valeur soit $\boxed{3}$.

Paramètres statistiques

Indices de dispersion

- On appelle **intervalle interquartile** l'intervalle $[Q_1 ; Q_3]$.
- On appelle **écart interquartile** la différence $Q_1 - Q_3$.

Remarques :

⋆ L'intervalle interquartile contient environ 50% des valeurs de la série.

⋆ La médiane et les quartiles sont des indicateurs de position.
L'écart interquartile est un indicateur de dispersion de la série.

⋆ La médiane et les quartiles, contrairement à la moyenne et l'écart-type, ne sont pas influencés par les valeurs extrêmes de la série. Ils dépendent seulement du nombre de valeurs de la série.

Selon les cas, on utilise le couple (moyenne/écart-type) ou bien le couple (médiane/écart interquartile) pour comparer des séries statistiques.
Plus l'écart-type d'une série est faible, plus les valeurs de la série sont « resserrées » autour de la moyenne.
Plus l'écart interquartile est faible, plus les valeurs de la série sont « resserrées » autour de sa médiane.

Pour finir

Pour la séance du mercredi de la rentrée : Préparer les exercices 1, 2, 3 et 4 du sujet de partiel 2014 que vous trouverez sur coursinfo.

Merci de votre attention,
bonne fin de journée et bonnes vacances ! !