Cours 2 : Statistiques descriptives

Table des matières

Lecture

Obligatoire : Document sur l'utilisation du logiciel SPSS.

Suggérée : Howell, Chapitre 2, sections 2.1 et 2.2, 2.4 et 2.5, 2.7 à 2.9 jusqu'à la sous section « La moyenne et la variance en tant qu'estimateurs » exclusivement.

Objectifs

Pouvoir comprendre la notion de statistiques descriptives, connaître les plus usuels (de tendance centrale : les moyennes, la médiane, le mode; de dispersion : écart type, variance; l'asymétrie et la kurtose). Pouvoir calculer des statistiques descriptives et en faire des graphiques.

Section 1. Savoir n'est pas comprendre

Les distributions de fréquences et leurs représentations graphiques que nous avons vues au cours précédente donnent un aperçu de la répartition d'un ensemble de données. De plus, elles offrent aux chercheurs une façon empirique de vérifier la validité de leurs données. Cependant, ce n'est qu'un premier pas. Il faut ensuite obtenir des prises sur ces données brutes, des valeurs facilement communicables pour qu'un lecteur éventuel puisse se faire une idée des données sans devoir les énumérer. Ceci est le rôle de la statistique descriptive.

Pour comprendre l'importance de ces prises, imaginons un être surnaturel qui pourrait connaître pour n'importe quel moment dans le passé la position de la planète Mars. Il n'est pas clair que cette entité pourra dire où sera Mars dans un mois. En effet, pour extrapoler, il faut savoir comment généraliser nos connaissances antérieures (pour obtenir par exemple la loi du mouvement de Newton), puis évaluer des paramètres (le poids de Mars, du soleil). Ainsi, une connaissance parfaite d'un phénomène n'implique pas une compréhension des processus à l'œuvre.

De la même façon, si cet être connaît toutes les positions et les vitesses des atomes d'eau contenues dans un verre, est-ce que cela implique qu'il connaisse sa température? Encore une fois, non. La température d'un liquide reflète la vitesse de déplacement moyenne des atomes le composant. Or connaître la vitesse d'un atome particulier n'informe en rien sur la vitesse moyenne. Il faut compiler ces vitesses individuelles de façon à en extraire une information plus significative (la température).

En psychologie, si nous mesurons chez mille individus le temps nécessaire pour identifier un visage, nous allons obtenir mille mesures différentes. Que peut-on conclure? Que nous avons tous des processus différents pour reconnaître les visages? Nous sommes loin de l'idée d'une loi. Il faut plutôt chercher à identifier ce qui est commun à l'ensemble des participants. L'utilisation de statistiques descriptives permet d'atteindre ce but.

Section 2. Statistiques de la tendance centrale

Les statistiques de la tendance centrale (ou encore les mesures de la tendance centrale) ont pour objectif de donner une idée de la localisation des données brutes (i. e. la localisation de leur distribution). Les données sont-elles généralement grandes? Petites? Plusieurs mesures de la tendance centrale existent, dont la plus fréquente est la moyenne arithmétique (souvent appelée tout simplement la moyenne). Dans tous les cas, une mesure de la tendance centrale indique si la distribution est située plus à droite ou plus à gauche de l'échelle.

Dans l'exemple de la Figure 1, la taille (en cm) de deux échantillons (fictifs) a été obtenu chez 400 individus de sexe féminin et masculin respectivement. On voit en regardant les distributions que la distribution des tailles chez les femmes est légèrement décalée vers la gauche par rapport à celles des hommes. Toutes les mesures de tendance centrales devraient refléter ce point.

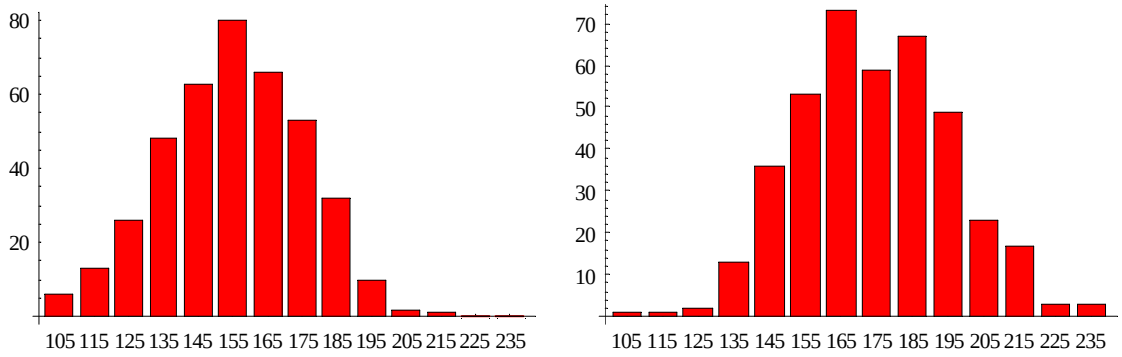


Figure 1 : Deux distributions pour la taille de femmes (à gauche) et d'hommes (à droite)

Lorsque l'on calcule le **Mode** ($\dot{X}$), la **Médiane** ($\breve{X}$), la **Moyenne arithmétique** ($\overline{X}$), la **Moyenne géométrique** ($\overset{\circ}{X}$) et la **Moyenne harmonique** ($\tilde{X}$), on obtient les résultats suivants (voir le lexique pour la définition de ces mesures) :

Statistique	**Femme**	**Homme**
$\dot{X}$	155	165
$\breve{X}$	155.8	172.9
$\overline{X}$	155.2	174.4
$\overset{\circ}{X}$	154.2	172.4
$\tilde{X}$	152.9	171.1

Comme on le voit, les cinq mesures de la tendance centrale indiquent bien que la distribution des femmes est légèrement plus à gauche que celle des hommes.

La médiane et le mode sont des statistiques qui sont surtout utiles quand la distribution contient des valeurs extrêmes puisque ces mesures sont peu influencées par des scores marginaux. En économie par exemple, le revenu médian est beaucoup plus utilisé que le revenu moyen, considérant qu'il existe une poignée de personnes qui ont des revenus dépassant les milliards de dollars (scores extrêmes).

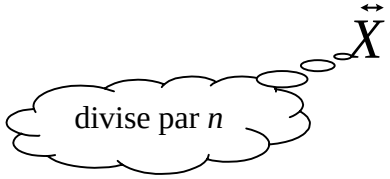
Encadré ⓿ Note sur la nomenclature.

Vu le nombre important de symbole que nous allons manipuler, il est important d'avoir une nomenclature uniforme.

Dans tout ce qui suit, nous utilisons une lettre de la fin de l'alphabet en majuscule pour dénoter des échantillons, telles **X**, **Y**, **Z**.

Lorsqu'une statistique est calculée sur un échantillon, nous ajoutons un symbole sur la lettre dénotant l'échantillon. Des exemples de statistiques calculées sur l'échantillon **X** sont $\dot{X}$, $\breve{X}$, $\overline{X}$, etc. Contrairement à **X** qui représente un ensemble de plusieurs valeurs, $\overset{?}{X}$ représente une valeur unique pour un échantillon donné.

Dans le cas de l'écart type (voir cours suivant), nous utilisons $\vec{\vec{X}}$. Or, il existe deux façons de calculer l'écart type. Pour les distinguer, nous ajoutons à la gauche du symbole une étiquette, soit *n* ou *n – 1*: ${}_n\vec{\vec{X}}$, ${}_{n-1}\vec{\vec{X}}$. L'étiquette ne représente pas une opération mathématique, seulement une indication:



Dans le passé, et sur beaucoup de calculatrices, ces symboles sont utilisés:

${}_n\vec{\vec{X}}$	S	S_n	σ_n
${}_{n-1}\vec{\vec{X}}$	s	S_{n-1}	σ_{n-1}

Leur principal défaut est de ne pas dire s'il s'agit de l'écart type pour l'échantillon **X** ou **Y**; cette ambiguïté n'existe pas avec ${}_n\vec{\vec{X}}$ vs. ${}_n\vec{\vec{Y}}$.

Une alternative à la médiane et au mode sont les moyennes. Il existe trois façons de moyenner les observations d'un échantillon, la moyenne géométrique, la moyenne harmonique, et la moyenne arithmétique. En règle générale, on observe que $\tilde{X} < \mathring{X} < \overline{X}$. Les moyennes utilisent toujours toutes les données brutes. Ainsi, chacune exerce une influence sur la moyenne obtenue (d'où l'importance de vérifier la validité des données extrêmes).

Les moyennes géométriques et harmoniques sont utilisées dans des situations particulières (et virtuellement jamais en psychologie). Par exemple, les économistes qui n'aiment pas utiliser la médiane vont utiliser la moyenne géométrique qui ressemble un peu à la moyenne (arithmétique) mais est un peu moins affectée par les données extrêmes (telles les milliardaires).

La moyenne arithmétique (appelée moyenne dans la suite) possède des propriétés mathématiques intéressantes (que nous expliquerons en détails dans l'encadré ❼ au cours 5) : il s'agit d'une statistique efficace et sans biais. Pour ces raisons, la très grande majorité des tests statistiques sur la tendance centrale sont en fait des tests sur la moyenne.

La moyenne se calcule suivant cette formule simple sur les données brutes, $\frac{1}{n}\sum_i \mathbf{X}_i$.

Cette expression peut se réorganiser comme suit : $\sum_i \frac{1}{n}\mathbf{X}_i$, où le $\frac{1}{n}$ indique tout simplement que cette données est présente 1 fois sur *n*. Si certaines données brutes $\mathbf{X}_i$ se répetent, il est possible de gagner du temps en utilisant plutôt : $\sum_i f\ \mathbf{X}_i$ où *f* est la fréquence relative de la données brute (i. e. la proportion de fois qu'elle a été observée).

Encadré ❶ Comment faire un graphe

Quand vient le temps de présenter vos résultats, une façon très efficace consiste à présenter des graphes de vos statistiques descriptives (en règle générale, la moyenne, quoique l'écart type est aussi présenté à l'occasion. Pour faire des graphes qui soient clairs, il y a certains points qu'il ne faut pas oublier :

- Tous les graphiques doivent avoir un titre (contrairement à ceux trouvés dans ces notes de cours) commencent en général par « Figure x : … ».
- Les axes doivent avoir une indication de la variable illustrée ainsi que, le cas échéant, de son unité de mesure entre parenthèses (par exemple, temps (ms)). De plus, le système métrique doit être utilisé dans tous les cas.
- Les points doivent utiliser la majorité de l'espace sur le graphe.
- Si l'abscisse est un échelle de type I, utiliser de préférence un graphe en histogramme; si l'échelle est de type II, utiliser de préférence une courbe.

Voici à la Figure 2 quelques exemples de graphes présentant des statistiques pour des études où il n'y a qu'une seule V. I.

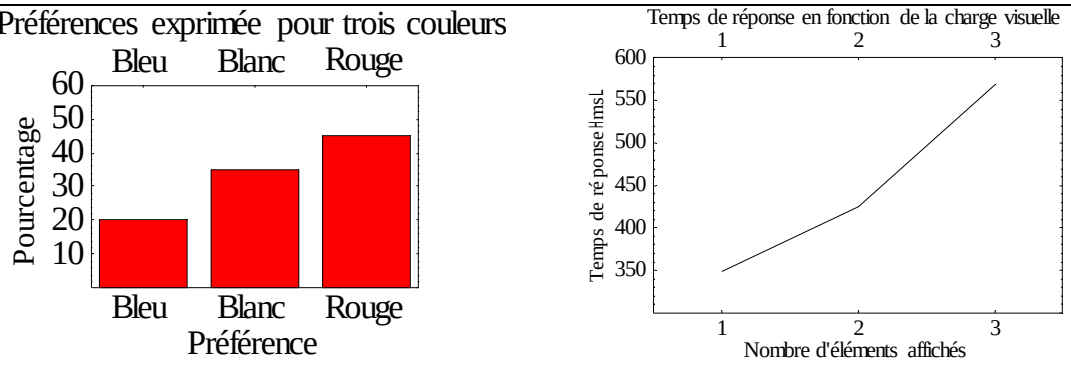


Figure 2 : Deux types de graphiques des moyennes

Quand l' étude manipule deux V. I., utilisez des histogrammes regroupés (*clustered*) ou encore plusieurs lignes, comme à la Figure 3. Dans ce cas, il ne faut pas oublier de mettre une légende (note : ces graphiques montrent l'erreur type, voir l'encadré suivant).

À l'occasion, des données avec deux V. I. peuvent aussi se prêter à un graphique en trois dimensions, comme c'est le cas dans la Figure 4.

Finalement, dans le cas où plus de 2 V. I. sont utilisées, il faut utiliser des panneaux distincts pour chaque graphique, avec une étiquette précisant le niveau d'une des V. I. sur chacun. Dans ce dernier cas, une seule légende pour l'ensemble des panneaux peut être utilisée comme c'est le cas à la Figure 5.

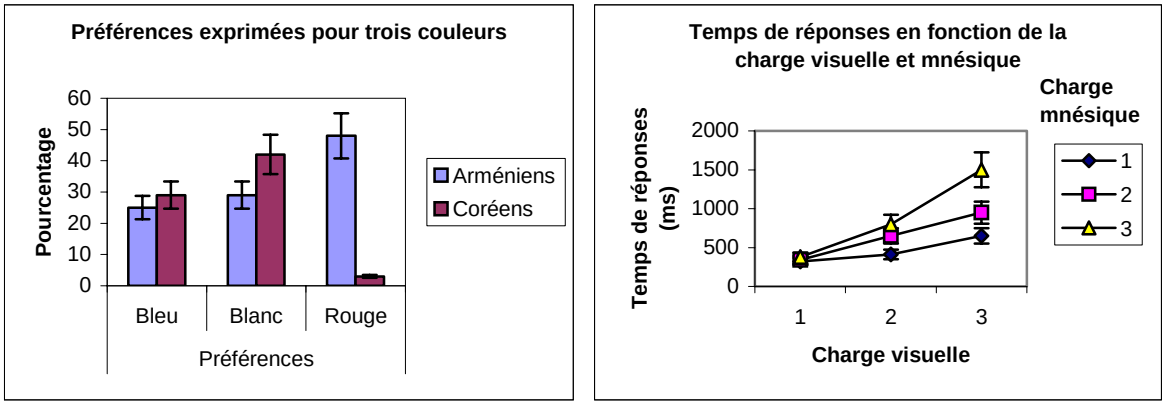


Figure 3 : Exemples de graphiques avec plus d'une V.I.

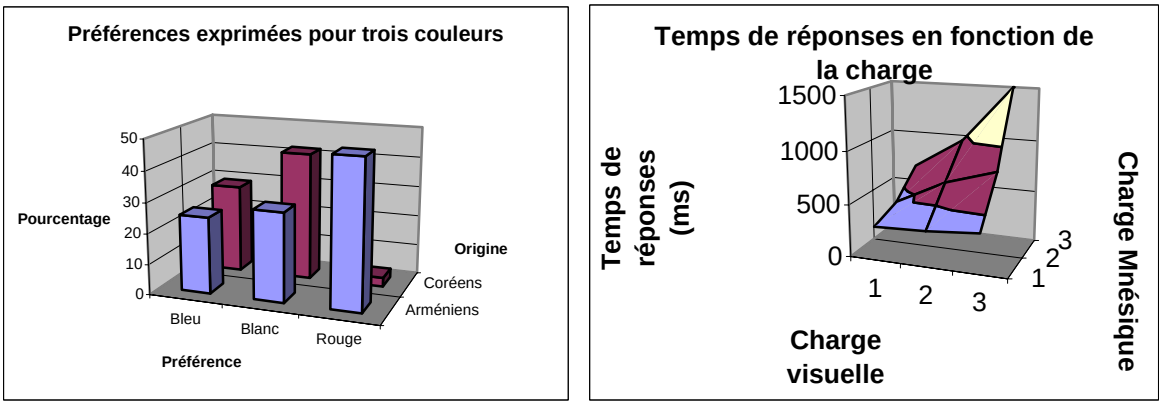


Figure 4 : Exemples de graphiques en trois dimensions

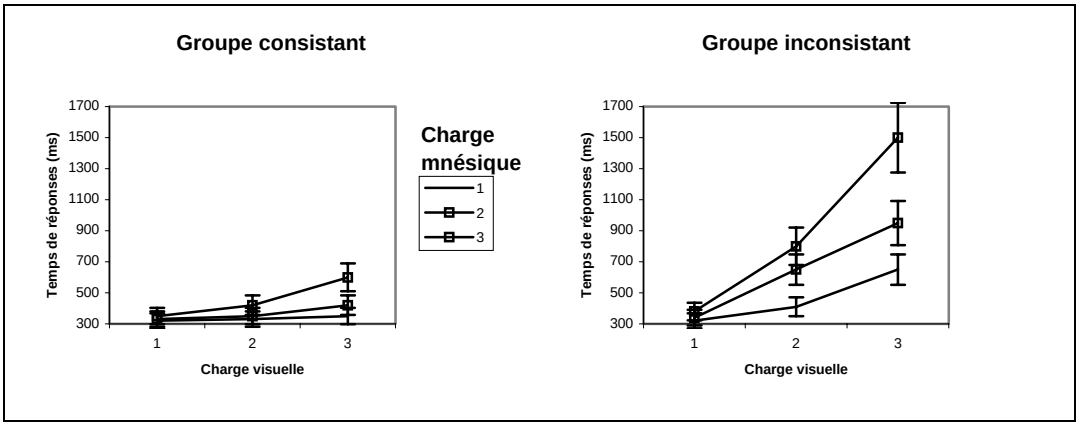


Figure 5 : Temps de réponses en fonction du groupe (consistant et inconsistant) et des charges visuelle et mnésique

Tous les graphes de cet encadré ont été faits avec Excel. SPSS possède aussi la possibilité de faire des graphes –et est souvent plus rapide- tout comme de nombreux autres logiciels.

Section 3. Statistiques de la variabilité

Les mesures de tendances centrales vues précédemment sont informatives, mais insuffisantes pour décrire une distribution. Il est aussi utile de connaître la dispersion des données.

Il existe plusieurs façons de calculer la dispersion des données brutes. Par exemple, on pourrait calculer la distance entre les deux extrêmes (l'étendue, que nous avons vu dans le cours 1). Cependant, seulement deux données sont utilisées (Min(**X**) et Max(**X**)), rendant cette mesure très sensible aux erreurs d'échantillonnage (données extrêmes). Une autre façon de mesurer la variabilité serait de calculer la moyenne des distances entre toutes les paires de scores. Cependant, nous serions confrontés à un nombre astronomique de paires de scores possibles (pour n données, il existe $n \times (n - 1) / 2$ pairs, un nombre qui devient rapidement énorme; essayez avec $n = 100$).

La méthode la plus usitée prend comme point de départ que la moyenne se situe au centre de la distribution. On peut donc l'utiliser comme point de référence. Imaginons que l'on calcule la distance entre chaque point $\mathbf{X}_i$ et la moyenne des points $\overline{X}$. Si on fait la somme de toutes ces distances et divisons par n, noté $\frac{1}{n}\sum_i (\mathbf{X}_i - \overline{\mathbf{X}})$, nous n'aurons malheureusement pas une valeur de dispersion. En effet, la somme des distances entre chaque donnée brute et sa moyenne est toujours nulle. En effet, en terme mathématique :

$$\begin{aligned}
\frac{1}{n}\sum_i (\mathbf{X}_i - \overline{\mathbf{X}}) &= \frac{1}{n}\left(\sum_i \mathbf{X}_i - \sum_i \overline{\mathbf{X}}\right) \\
&= \frac{1}{n}\left(n\overline{\mathbf{X}} - \sum_i \overline{\mathbf{X}}\right) \\
&= \frac{1}{n}\left(n\overline{\mathbf{X}} - n\overline{\mathbf{X}}\right) \\
&= \frac{1}{n}0 = 0
\end{aligned}$$

Autrement dit, à cause de la position centrale de la moyenne, les distances négatives des données plus petites que la moyenne sont exactement contrebalancées par les distances positives des données plus grandes. Pour vous en convaincre, faites le test avec ces données : **X** = {1, 2, 3, 4, 5, 6, 7}.

(Si on ignore la multiplication par $1/n$, ce résultat stipule que la sommes des écarts à la moyenne donne toujours zéro. C'est un résultat qui va revenir souvent par la suite pour simplifier des formules plus complexes.)

Pour contourner le problème, nous élevons chaque distance au carré, obtenant ainsi une série de carrés ayant tous des valeurs positives. Le résultat est appelé la variance, dont la formule est

$$\frac{1}{n}\sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})^2$$

Comme nous allons le voir dans l'encadré ❽ du cours 5, la variance d'un échantillon est biaisée. En effet, même si l'échantillon reflète dans une certaine mesure la variabilité de la population dont il est tiré, il est probable que parmi ce petit nombre de données brutes (par rapport à la taille de la population entière), les données les plus extrêmes soient sous représentées (simplement parce qu'il y en a peu dans la population). En conséquence, la variabilité de la population sera sous-estimée par la variabilité de l'échantillon.

Pour éviter ce biais, il faut augmenter la valeur de cette estimation. Cependant, cette correction doit s'atténuer lorsque la taille de l'échantillon est très grand. Cette correction est donc fonction de *n*. On démontrera à l'encadré ❽ du cours 5 que la correction adéquate est de multiplier la variance de l'échantillon par $\frac{n}{n-1}$ de façon à obtenir une variance qui reflète le fait que notre échantillon soit forcément affecté par une espèce de régression vers la moyenne. Si *n* est petit, la correction est appréciable et la variance estimée de la population est plus grande. Si *n* est très grand, la correction devient négligeable. Dans la suite, l'on va distinguer la variance corrigée pour le biais d'un échantillon, notée $_{n-1}\ddot{X}^2$ de la variance biaisée d'un échantillon, notée $_{n}\ddot{X}^2$.

Prenez le temps de vérifier que votre calculatrice de poche peut calculer la variance d'un échantillon corrigée pour le biais (parfois, le bouton est noté S_{n-1}^2 ou encore σ_{n-1}^2). C'est la seule mesure d'intérêt.

La variance étant une mesure au carré, on rapporte souvent la racine carrée de la variance, que l'on appelle l'écart type non-biaisé (ou corrigé pour le biais) d'un échantillon, et noté $_{n-1}\ddot{X}$.

Une façon simple de bien comprendre ce qu'est la variabilité mesurée par l'écart type $_{n}\ddot{X}$ est de se poser la question suivante : Supposons que je prends une mesure de mon échantillon au hasard, à quelle distance de la moyenne se trouvera-t-il approximativement?

On a déjà vu que dans l'ensemble, la déviation à la moyenne s'annule; il faut une approche qui ne tienne pas compte du signe de la déviation. On va donc considérer le carré (car élever au carré enlève le signe) puis prendre la racine carrée. En moyenne, la distance entre une donnée quelconque et sa moyenne est donnée par :

$$\overline{X_i - \overline{\mathbf{X}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})^2} = \textit{par définition}\ \sqrt{{}_{n}\ddot{X}^2} = {}_{n}\ddot{X}$$

Autrement dit, en prenant une donnée au hasard, elle a toutes les chances d'être à ± un écart type de la moyenne de l'échantillon. À partir d'une donnée unique, l'erreur que vous faites pour estimer la moyenne est de plus ou moins un écart type, en moyenne.

Encadré ❷ L'erreur type.

Une estimation basée sur un échantillon restreint de données brutes qui exclut donc la majorité des valeurs de la population contient une certaine marge d'erreur. Cette erreur, nous l'appelons l'erreur type. Il existe deux sources d'imprécision pour estimer l'erreur type. a) Imaginons que nous choisissons aléatoirement deux échantillons de même taille à l'intérieur d'une même population. Nous obtiendrons assurément deux moyennes légèrement différentes tout simplement parce que nos échantillons ne sont pas identiques. Or, la dispersion de ces moyennes dépend de la taille des échantillons sélectionnés. Des échantillons extrêmement petits ont une plus grande variabilité et sont donc imprécis pour estimer la moyenne de la population. D'un autre côté, si on choisit deux échantillons très grands, les moyennes qui en résultent varieront très peu. Pour exprimer ceci, on dira que l'erreur d'estimation est inversement proportionnelle à n (c'est à dire proportionnel à $1 / n$).

b) Le deuxième déterminant de la précision d'un estimé est la variabilité qui existe à l'intérieur même de la population. Si la population ne contient que des mesures constantes, les échantillons seront composés de cette même constante, et la variance sera zéro, ce qui signifie pas d'erreur dans l'estimé. Par contre, si la variabilité est très grande dans la population, nos deux échantillons seront aussi sans aucun doute très différents. De fait, les échantillons reflètent plus ou moins bien la dispersion de la population. Ainsi, l'erreur type sera proportionnelle à la variance de la population (inconnue mais que l'on peut estimer par la variance non biaisée ${}_{n-1}\vec{\vec{X}}^2$.

La précision du calcul d'une moyenne, que l'on nomme l'erreur type (ou en anglais *Standard error* parfois traduit pas erreur standard), notée SE, dépend donc de ces deux facteurs, que l'on peut tout simplement multiplier. Pour avoir une erreur qui soit dans la même unité que la moyenne, on extrait la racine carrée. On obtient donc :

$$SE_{\bar{X}} = \sqrt{\frac{{}_{n-1}\vec{\vec{X}}^2}{n}} = \frac{{}_{n-1}\vec{\vec{X}}}{\sqrt{n}}$$

Il est à noter que la formule d'erreur type varie selon le type de statistique dont l'on veut une marge d'erreur. Pour connaître l'erreur type de d'autres statistiques, voir Cramér. Par exemple :

$$SE_{\tilde{X}} = \sqrt{\frac{\pi}{2n}}\,{}_{n-1}\overleftrightarrow{X} \qquad SE_{{}_{n-1}\vec{\vec{X}}^2} = \sqrt{\frac{2}{n}}\,{}_{n-1}\vec{\vec{X}}^2 \qquad SE_{{}_{n-1}\vec{X}} = \frac{1}{\sqrt{2n}}\,{}_{n-1}\vec{\vec{X}}$$

Il est commode (quoique rarement fait) de rapporter dans un texte la moyenne plus ou moins l'erreur type (par exemple, la longueur est de 224 mm ± 14 mm). Comme on le verra dans le cours 4, l'erreur type est en fait très proche de la méthode du test t. De plus, il est très fortement recommandé de mettre dans tout graphique représentant des moyennes une barre d'erreur dont la hauteur est donnée par l'erreur type. Des logiciels comme SPSS peuvent calculer automatiquement cette barre d'erreur sur demande.

Section 4. Relations fondamentales sur les moments statistiques

Il existe quelques formules clefs qui vous donnent la moyenne ou la variance quand vous transformez les valeurs de votre échantillon à l'aide d'une constante multiplicative *a* et d'une constante additive *b*. Dans la suite, je dénote la moyenne $\overline{X}$ par $E(\mathbf{X})$, et la variance ${}_{n-1}\vec{X}^{2}$ par Var(**X**).

$$E(a\mathbf{X}+b) = aE(\mathbf{X})+b$$
$$Var(a\mathbf{X}+b) = a^2 Var(\mathbf{X})$$
$$Var(\mathbf{X}) = E(\mathbf{X}^2) - E^2(\mathbf{X})$$

Autrement dit, si vous additionner une constante *b* à chacune de vos données brutes, la moyenne s'en trouve affectée, mais pas la variance (car *b* n'amène aucune variabilité). N'oublions pas cette relation que nous avons déjà prouvée au cours précédent:

$$E(X - \overline{\mathbf{X}}) = 0$$

Finalement, si vous manipulez deux échantillons indépendants, il existe ces trois relations :

$$E(\mathbf{X}+\mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$$
$$E(\mathbf{X}\times\mathbf{Y}) = E(\mathbf{X}) \times E(\mathbf{Y})$$
$$Var(\mathbf{X}+\mathbf{Y}) = Var(\mathbf{X}) + Var(\mathbf{Y})$$

Ces relations seront souvent utilisées dans les preuves mathématiques.

Section 5. Autres moments statistiques et leur représentation visuelle

Pour bien décrire une distribution de données brutes, il est indispensable de rapporter la moyenne et l'écart type. En règle générale, ces deux statistiques sont les plus signifiantes, et les plus à même d'être retrouvées dans un article scientifique. Si on illustre une distribution quelconque (ses fréquences obtenues ou encore sa fonction de masse PDF, si connue) en utilisant un graphique avec une courbe continue, on peut localiser visuellement la moyenne en trouvant le point où la distribution serait en équilibre. Dans les deux exemples de la Figure 6, la moyenne se trouve à 170.

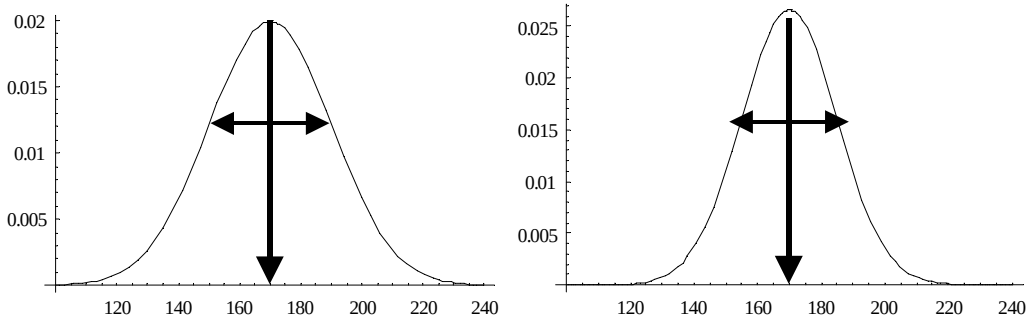


Figure 6 : le centre de gravité d'une distribution de données

On peut obtenir une appréciation visuelle de l'écart type en regardant la largeur de la distribution. Dans les exemples précédents, il est très clair que les données du graphe de droite présentent une plus grande variabilité que celles de gauche. Il est cependant plus difficile de voir que l'écart type est 20 dans le premier cas et de 15 dans le second. Dans ces exemples, les données étaient réparties symétriquement, ce qui fait que la moyenne coïncide avec la médiane et le mode. De plus, le centre de gravité se trouve au centre. Par contre, les distributions ne sont pas toujours symétriques, comme dans les deux exemples de la Figure 7.

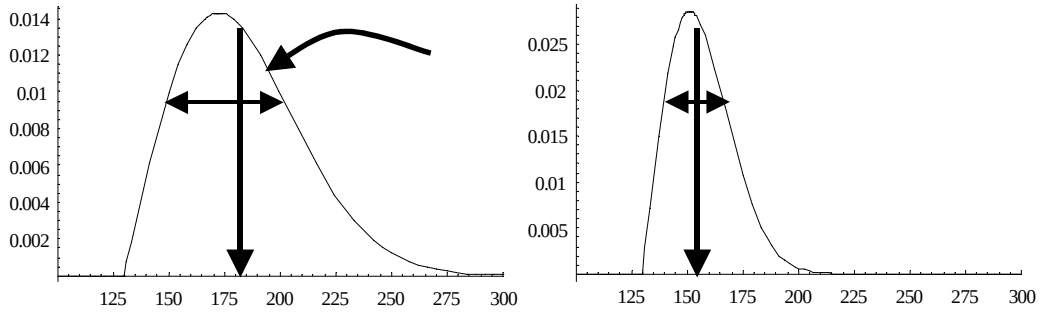


Figure 7 : Deux distributions asymétriques

Encore une fois, la moyenne est le centre de gravité, légèrement à droite du mode puisque la partie droite s'étend beaucoup plus que l'autre. L'écart type se mesure aussi assez bien visuellement, étant de 13,9 et 6.9 respectivement. Cependant, il devient clair que l'asymétrie de ces distributions est un aspect important des données et qu'il faudrait rapporter une statistique mesurant cet état de fait. L'exemple de la Figure 8 donne un exemple ou l'asymétrie est encore plus extrême. Lorsqu'une distribution n'est pas symétrique, le mode, la moyenne et la médiane diffèrent.

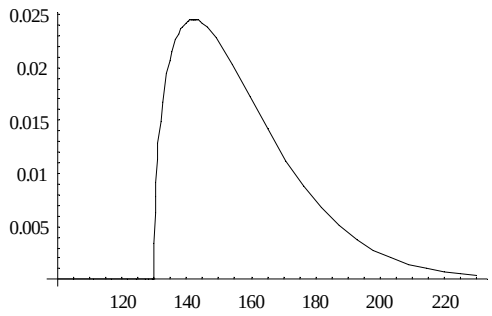


Figure 8 : Une distribution très asymétrique

Pour quantifier l'asymétrie (*Skewness* en anglais), on utilise la formule suivante :

$$\mathbf{X} = \frac{\frac{1}{n}\sum_i (\mathbf{X}_i - \overline{X})^3}{{}_n\vec{X}^3}$$

Le résultat nous indique la direction de l'asymétrie. Il existe trois cas possibles : Si $\mathbf{X} >$ 0, l'asymétrie est positive et la distribution s'étale plus vers les valeurs élevées de la variable.

On dit qu'elle a une longue queue à droite. Si $\mathbf{X} = 0$, la distribution est parfaitement symétrique. Si $\mathbf{X} < 0$, l'asymétrie est négative, et la queue est plus longue à gauche. [Notez que SPSS multiplie par $\frac{n^2}{(n-1)(n-2)}$ le résultat ci-haut; la valeur retournée par le logiciel est donc légèrement différente, mais le signe positif ou négatif reste le même.]

Finalement, il peut arriver que l'aplatissement de la distribution soit inusité. Une statistique qui donne un indice de cet aplatissement est la kurtose (*Kurtosis*). Elle se calcule avec la formule :

$$\mathbf{X} = \frac{\frac{1}{n}\sum_i (\mathbf{X}_i - \bar{X})^4}{{}_n\ddot{X}^4}$$

Une kurtose de 3 indique une rondeur typique de la distribution en pointillé (qu'on appelle mésocurtique). Si $\mathbf{X}$ est inférieur à 3, ceci indique une distribution plus pointue et plus haute (qu'on appelle platycurtique). Dans le cas contraire (qu'on appelle leptokurtique), les queues de la distribution s'étendent plus loin que pour la distribution normale, à écart type équivalent. Comme la valeur 3 est une indication de kurtose neutre, certains auteurs (et SPSS) recommandent de soustraire 3 à la formule ci-haut. Dans la Figure 9, les trois distributions ont une moyenne de 0, un écart type de 1, et sont symétriques (*skewness* = 0) mais varient selon la kurtose.

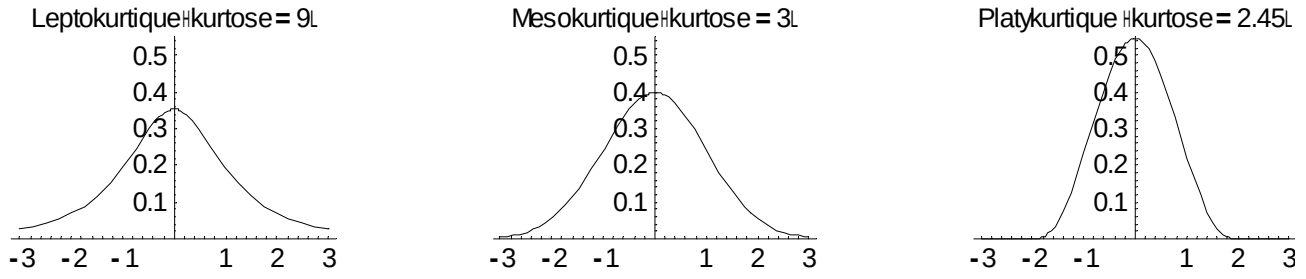


Figure 9 : Trois exemples de kurtose

Les statistiques présentées jusqu'à présent sont aussi appelées des mesures de la position (moyenne), de l'échelle (écart type), de l'asymétrie (*Skewness*) et de l'aplatissement. De plus, l'échelle, l'asymétrie et la kurtose sont aussi appelées des moments μ de degré r (2, 3, et 4) respectivement pour la raison que dans leurs calculs, on utilise la somme des écarts à la moyenne élevée à la puissance r : $\mu_r = \frac{1}{n}\sum_i (X_i - \bar{X})^r$. Comme on l'a dit précédemment, la somme des écarts à la moyenne (sans exposant) donne zéro. On a donc que $\mu_1 = 0$. De plus, on a ces relations

$${}_n\ddot{X}^2 = \mu_2$$

$$\mathbf{X} = \frac{\mu_3}{{}_n\ddot{X}^3}$$

$$\mathbf{X}_{n} = \frac{\mu_4}{\vec{X}^4}$$

Il existe d'autres moments de niveau supérieur (une infinité en fait), mais ceux-ci deviennent de plus en plus abstraits, et imperceptibles quand on inspecte visuellement une distribution.

Section 6. Quantiles

Il peut arriver lorsque la distribution des données est particulière ou encore quand les hypothèses des chercheurs portent sur celle-ci que rapporter seulement les trois premières statistiques ne soit pas suffisant. Dans ce cas, une image vaut mille mots : rapportez le graphe de la distribution des fréquences observées.

Une alternative quelque fois utilisée, qui remonte au début du siècle quand les méthodes pour réaliser des graphes étaient moins facilement accessibles, est de rapporter les quantiles (parfois appelé les *N*-tiles). L'idée est de rapporter quelques points le long de la distribution cumulative des fréquences. Avec ces quelques points, le lecteur peut extrapoler pour obtenir la distribution complète. Le nombre de points à rapporter est variable, mais souvent, on utilise les quartiles (N = 4 points), les déciles (N = 10 points) et les centiles (N = 100 points). Bien entendu, le nombre de points rapportés N doit être nettement inférieur au nombre d'observations n, pour que les valeurs des quantiles soient stables.

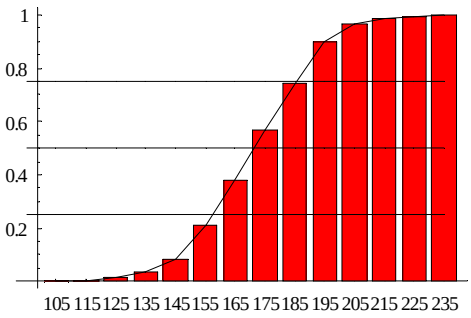


Figure 10 : les quartiles d'une distribution de données obtenus avec le graphe de la distribution cumulative

Si vous rapportez les quartiles (N = 4), vous devez rapporter les points tels que 25%, 50%, et 75% des données observées y soient inférieures. Ces points sont les frontières entre le premier quart des données et le second, entre le second et le troisième, et entre le troisième et le dernier quart des données. De façon, générale, pour un N choisi, vous rapportez N - 1 valeurs situant les fréquences relatives $1/N$, $2/N$, … $(N\text{-}1)/N$. Les quantiles s'obtiennent aisément du graphique des fréquences cumulatives (que vous intrapolez en reliant ensemble les centres de classes). Dans l'exemple sur la gauche, où l'on voit les fréquences cumulatives relatives (i.e. entre 0 et 1) de la taille des hommes vues au cours précédent, localisez les points sur l'abscisse tels que l'ordonnée soit 25%, 50% et 75%. Notez en passant que le point où la fréquence relative est de 50% est par définition la médiane. Dans notre exemple, on trouve les valeurs {157, 172, 185}. Plusieurs logiciels peuvent faire ce travail d'intrapolation efficacement.

Section 7. Conclusion

Exercices

1. Si la moyenne $\overline{X} = \frac{1}{n}\sum \mathbf{X}_i$, il s'ensuit que :
   a) $n\overline{X} = \sum \mathbf{X}_i$
   b) $n = \sum \mathbf{X}_i / \overline{X}$
   c) $1 = \frac{1}{n\overline{X}}\sum \mathbf{X}_i$
   d) Toutes ces réponses
   e) Aucune de ces réponses
2. Si $\overline{X}$ =55, et $\sum \mathbf{X}_i$ =825, combien de sujets ont participé à l'étude :
   a) n = 35
   b) n = 18
   c) n = 15
   d) n = 25
   e) Impossible de le déterminer.
3. Si n = 25, et $\sum \mathbf{X}_i$ =525, que vaux $\overline{X}$:
   a) 25
   b) 21
   c) 18
   d) 15
   e) Impossible à déterminer.
4. Lorsque nous organisons un ensemble de données en ordre croissant et que nous indiquons à côté de ces données la fréquence qui y est associée, nous construisons :
   a) Une distribution de fréquence
   b) Un histogramme
   c) Un diagramme en bâton
   d) Un graphe des fréquences
   e) Aucune de ces réponses.
5. Soit une variable dont une des données brute est 6. Quelle est la valeur réelle de ce score :
   a) Entre 5.5 et 6.5
   b) Dépend de la précision de l'instrument de mesure
   c) Dépend de l'erreur type.
   d) 6
   e) b et c sont corrects
6. Généralement, en combien de classes les données doivent-elles être regroupées :
   a) 5 à 10
   b) 10 à 20
   c) 5 à 15
   d) 15 à 30
   e) 10 à 15
7. Soit une série de données dont la plus basse est 19 et la plus élevée 82. Supposons que vous décidiez de regrouper ces données en 8 classes. Quelle sera l'étendu de chaque intervalle de classe :
   a) 8
   b) 9
   c) 10
   d) 19
   e) Nous n'avons pas assez d'informations.
8. Qu'est-ce que la médiane :
   a) Une valeur qui divise une série de données en deux groupes d'effectifs égaux
   b) Une valeur égale au centre d'équilibre d'une distribution

c) Le rang milieu si l'on ordonne les données et leur donne un rang successif

d) a et b

e) a et c

f) a, b, et c

9. À propos de l'écart type et de la variance :

a) C'est la même chose

b) La variance est l'écart type élevé au carré

c) L'écart type est la variance au carré

d) Il n'y a pas de relation entre eux.

e) a, b, et c sont corrects.

10. La variance échantillonale est un indice de l'homogénéité des observations :

a) Vrai

b) Faux

11. La moyenne est une mesure de la tendance centrale. Quels sont les avantages associés à cette mesure :

a) elle permet de synthétiser un ensemble de mesures

b) Elle permet de comparer l'individu à l'ensemble de son groupe

c) Elle permet de comparer des groupes entre eux.

d) Toutes ces réponses

e) Aucune de ces réponses.

12. La moyenne est un paramètre lorsqu'elle est calculée sur toute la population

a) Vrai

b) Faux

13. Les quartiles divisent une distribution en combien de classes :

a) 100

b) 25

c) 3

d) 4

e) 10

14. Les effectifs inclus dans chaque quartile sont égaux

a) Vrai

b) Faux

15. Soit l'échantillon $\mathbf{Z}$ = {9,8,7,7,7,5,5,5,5,4,4,3,3,2,1,1},

a) Quelle est la moyenne :

b) Quel est le mode :

c) Quel est l'écart type

d) Quelle est la moyenne harmonique :

e) Quelle est le premier quartile :

16. Soit 1 m = 3.3 pieds, et ces données $\mathbf{X}$ = {1, 2, 3, 4, 7, 9} en mètres.

a) Calculer $\overline{\mathbf{X}}$ en mètre.

b) Soit $\mathbf{Y}_i = 3.3\ \mathbf{X}_i$, la distance en pieds. Calculer $\overline{\mathbf{Y}}$

c) Calculer l'écart type ${}_{n-1}\ddot{\mathbf{X}}$ en mètres.

d) Calculer l'écart type ${}_{n-1}\ddot{\mathbf{Y}}$ en pieds.

17. Si la moyenne $\overline{\mathbf{Z}}$ est de 10, et la moyenne des scores élevés au carré $\overline{\mathbf{Z}^2}$ est de 230, calculez la variance biaisée ${}_{n}\ddot{\mathbf{Z}}^2$.

18. Supposons que la taille moyenne des homes est de 1.75 m avec un écart type de 20 cm, et celle des femmes, de 1.65 avec un écart type de 20 cm. Si l'on suppose un nombre égale d'hommes et de femmes,

a) Calculer la taille moyenne de l'humanité

b) Calculer l'écart type de l'humanité.