Les statistiques aux écrits de l'agrégation

Le programme

- Statistiques descriptives univariées : indicateurs de position et indicateurs de dispersion. Représentations graphiques de données.
- Série statistique à deux variables quantitatives, nuage de points associé. Coefficient de corrélation. Droite de régression des moindres carrés.
- Estimation ponctuelle. Estimation par intervalle de confiance.

Références :

Statistique mathématique en action, Stoltz

Cours de Simon Coste `https://scoste.fr/teaching/statagreg2021.pdf`

Cours de Mathieu Mansuy `http://www.mathieu-mansuy.fr/pdf/ECG2-TP3.pdf`

Table des matières

1 Le premier point

Je me demande si mon chat a de plus grandes moustaches que la plupart des chats en France, la taille de ses moustaches étant de 6,9cm.

1.1 Statistiques descriptive univariées

Tout d'abord, je dois avoir un moyen de comparer mon chat aux autres chats.

Définition 1.1. On considère un ensemble Ω appelé *population* dont les éléments ω sont les *individus*. Un *caractère* sur la population est une application $X : \Omega \to E$ où E désigne un ensemble quelconque. Si E est un ensemble de nombres, X est un caractère *quantitatif*,

sinon X est un caractère *qualitatif*.

Exemple 1.1.

- Population : les chats en France.
- Caractère quantitatif : la taille des moustaches en centimètres.
- Caractère qualitatif : la couleur du pelage.

Définition 1.2. On appelle *n-échantillon* un n-uplet $\{\omega_1, \dots, \omega_n\} \subset \Omega$ et une *série statistique* d'un échantillon pour un caractère X donné la liste $x = (X(\omega_1), \dots, X(\omega_n))$ des valeurs prises par X sur l'échantillon.

Exemple 1.2. On reprend X le caractère donnant la taille des moustaches des chats. Pour savoir si mon chat a de grandes moustaches, j'ai demandé à tous mes amis la taille des moustaches de leur chat. Je me retrouve avec un 15-échantillon (les 15 chats de mes amis) et une série statistique donnée par les valeurs de X sur les chats de mon échantillon.

$$(7.1, 5.3, 6.0, 2.9, 7.4, 6.8, 10.2, 4.9, 7.6, 6.6, 6.0, 8.2, 3.6, 4.9, 7.1)$$

Remarque 1.1.

- L'objectif des **statistiques descriptives univariées** (ou unidimensionnelles) est de fournir des résumés synthétiques, graphiques et numériques de séries de valeurs observées sur une population ou un échantillon.
- Une série statistique brute ne permettant pas une lecture efficace des données, on souhaite la présenter de manière synthétique (tableaux, graphes,...) et en donner des caractéristiques particulières (moyenne, médiane,...).

1.2 Indicateur de position

Ensuite, je dois pouvoir positionner mon chat dans l'échantillon de chats que j'ai obtenu.

Définition 1.3. On appelle *moyenne empirique* de la série de statistiques $x = (x_1, \dots, x_n)$ le réel :

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

Remarque 1.2. On remarque la correspondance entre les notions de moyenne en statistiques et d'espérence en probabilités

$$\mathbb{E}(x) = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x) \ \sim \ \bar{x} = \sum_{i=1}^{n} m_i \cdot f_i$$

où m_i sont les valeurs prises par X i.e. sont les valeurs x_i comptés avec multiplicité et f_i sont les fréquences des m_i i.e. (nombre de fois où m_i apparaît dans x)/(cardinal de x)

Exemple 1.3. La moyenne de mon 15-échantillon est 6.3cm, mon chat se situe au dessus de la moyenne de mon échantillon.

Définition 1.4. La médiane d'une série de statistique ordonnée est un réel m partageant la série en deux séries d'effectifs égaux. Soit $(x_1 \leqslant \cdots \leqslant x_n)$ alors

- si $n = 2p - 1$, alors $m = x_p$ (valeur de milieu),
- si $n = 2p$, alors $m = \frac{x_p + x_{p+1}}{2}$.

Exemple 1.4. La médiane de mon échantillon est 6.6cm. Mon chat se situe au dessus de la médiane.

Définition 1.5. Soit $x = (x_1, \ldots, x_n)$ une série statistiques.

- Le *premier quartile* de x est la plus petite valeur q_1 de x pour laquelle 25% des éléments lui sont inférieurs ou égales.
- Le *troisième quartile* de x est la plus petite valeur q_3 de x pour laquelle 75% des éléments lui sont inférieurs ou égales.

Remarque 1.3. On pourrait en définir d'autres : déciles, centiles,...

Exemple 1.5. Le premier quartile de mon échantillon est 5 et le troisième est 7. Mon chat est entre le premier et le troisième quartile.

1.3 Indicateur de dispersion

Maintenant, j'aimerais comprendre davantage mon échantillon, voir s'il est très concentré autour de sa moyenne, etc...

Définition 1.6. Soit $x = (x_1, \ldots, x_n)$ une série statistiques et $\bar{x}$ sa moyenne empirique.

- On appelle *variance empirique* de x le réel $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
- On appelle *écart-type empirique* de x le réel $\sigma = \sqrt{v}$.

Remarque 1.4. Comme on s'en doute, la variance empirique mesure la dispersion de la série statistique autour de sa moyenne et comme en probabilités la formule de Koenig-Huygens est valable.

Exemple 1.6. La variance de mon échantillon est de $3,15$cm et son écart-type de 1,78cm.

Définition 1.7.

- La *distance inter-quartile* est le réel $q_3 - q_1$.
- L'*étendue* d'une série statistique est la différence entre la plus grande et la plus petite valeur prise.

Remarque 1.5. La distance inter-quartile est la longueur de l'intervalle inter-quartile $[q_1, q_3]$ lequel contient la moitié des valeurs de la série, réparties autour de la moyenne. Pour représenter graphiquement l'étendue et la distance interquartile on représente souvent une "boîte à moustache" (box-plot).

Exemple 1.7. L'étendue de mon échantillon est (10,2 - 2,9) = 7.3cm. Sa distance interquartile est de 2cm.

1.4 Représentations graphiques de données

Finalement, je veux rendre jolies et visuelles toutes les données que j'ai accumulées jusqu'à présent. Pour ce faire, il y a deux choses que l'on connait depuis longtemps :

- Le diagramme en bâtons (sur l'axe des abscisses les valeurs possibles et sur l'axe des ordonnés l'effectif ou la fréquence de chaque valeur).
- L'histogramme (sur l'abscisse division de l'étendue en intervalles bien choisis non obligatoirement de même amplitude et sur l'ordonné nombre d'individu par intervalle).

2 Le deuxième point

Le problème qui se pose dans les séries statistiques à deux variables est principalement celui du lien qui existe ou non entre chacune des variables. Par exemple, je veux savoir si la taille des moustaches de chat est liée à l'âge du chat.

2.1 Série statistique à deux variables quantitatives et nuage de points associé.

Définition 2.1. Une *série statistique à deux variables quantitatives* est une série pour laquelle deux caractères quantitatifs sont relevés sur une même population. Les valeurs de cette série sont notées $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$ et elle se représente en général par un tableau ou un *nuage de points* : la représentation graphique dans un repère orthogonal d'une série statistique à deux variables quantitatives, formé par les points de coordonnées (x_i, y_i)

Exemple 2.1. Je demande à mes amis de me donner en plus l'âge de leur chat. Je me retrouve avec un 15-échantillon du type

$$\{(7.1, 9), (5.3, 5), (6.0, 6), (2.9, 2), ...\}$$

Remarque 2.1. Si les deux caractères sont bien quantitatifs, il existe un point particulier : *le point moyen* $G \coloneqq (\bar{x}, \bar{y})$ défini comme le point des moyennes des deux caractères.

Remarque 2.2. Le nuage de points associé à une série statistique à deux variables donne donc immédiatement des informations de nature qualitatives. Pour avoir des informations quantitative, on introduit la notion suivante.

Définition 2.2. Le *problème de l'ajustement* est l'établissement d'une fonctionnelle entre les deux séries.

Exemple 2.2. Par exemple, si les points du nuage de points associé à la série statistique à deux variables semblent être alignés on peut se demander quelle serait la meilleure droite pour représenter cet alignement.
Si l'âge et la taille des moustaches sont effectivement ajustables par une droite, on pourrait faire des estimations de l'âge d'un chat à partir de la donnée de la taille de ses moustaches x_0 en prenant l'ordonnée y_0 tel que (x_0, y_0) appartienne à la droite.

2.2 Coefficient de corrélation et droite de régression des moindres carrés

On considère une série statistique à deux variables $(x, y) = \{(x_i, y_i)\}_{1 \leqslant i \leqslant n}$ représentée par un nuage justifiant un ajustement affine.

Définition 2.3. Dans le plan muni d'un repère orthonormal, la droite d'équation $y = ax+b$ est la droite de régression de y en x de la série statistique si la norme suivante est minimale :

$$\|y - (ax + b)\|_2^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

Définition 2.4. On appelle covariance de la série le réel

$$cov(x, y) \coloneqq \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Remarque 2.3. On a $cov(x, x) = Var(x)$

Propriété 2.5. La droite de régression de y en x a pour équation $y = ax + b$ où

$$a = \frac{cov(x, y)}{cov(x, x)} \qquad\qquad b = \bar{y} - a\bar{x}$$

Remarque 2.4. Le point moyen G appartient toujours à la droite de régression

On veut maintenant savoir si on a eu raison de penser à un ajustement affine, i.e. on veut mesurer la qualité de notre ajustement.

Définition 2.6. Le coefficient de corrélation linéaire de la série statistique à deux variables est

$$r \coloneqq \frac{cov(x, y)}{\sqrt{cov(x, x)cov(y, y)}}$$

Remarque 2.5. Ce coefficient est à valeur entre 1 et -1. Plus il se rapproche de 1 en valeur absolue, meilleur est l'ajustement affine. Idéalement, $r = \pm 1$ et dans ce cas la droite passe par tous les points du nuage. Si r est positif, les variables augmentent ensemble, sinon, lorsque l'une augmente l'autre diminue.

Remarque 2.6. En pratique, il existe une autre valeur indiquant la qualité de notre estimation : la p-value.

La p-value est une mesure de la probabilité utilisée pour tester une hypothèse. Le but du test d'hypothèse est de déterminer s'il existe suffisamment de preuves pour soutenir une certaine hypothèse concernant nos données. En fait, nous formulons deux hypothèses : l'hypothèse nulle H_0 et l'hypothèse alternative H_1. Dans le cas de l'analyse de la corrélation, l'hypothèse nulle est généralement que la relation observée entre les variables est le pur fruit du hasard (le coefficient de corrélation r est vraiment zéro, il n'existe pas de relation affine). L'hypothèse alternative est que la corrélation mesurée est légitimement présente dans nos données (le coefficient de corrélation est différent de zéro).

La p-value désigne la probabilité d'observer un coefficient de corrélation différent de zéro dans les données de notre échantillon lorsqu'en fait l'hypothèse nulle est vraie. Une faible p-value impliquerait de rejeter l'hypothèse nulle. En général, le seuil de rejet d'une hypothèse nulle est une p-value de 0,05. Ainsi, si on a une p-value inférieure à 0,05, on rejete l'hypothèse nulle en faveur de l'hypothèse alternative selon laquelle le coefficient de corrélation est différent de zéro. ATTENTION, une grande p-value ne permet pas de confirmer l'hypothèse H_0 mais permet simplement de dire qu'on a pas assez de preuve pour la rejeter !

3 Le troisième point

3.1 Estimation ponctuelle

L'estimation ponctuelle d'un paramètre inconnu θ consiste à calculer une valeur "probable" de θ compte tenu des observations. Cela pré-suppose qu'on a déjà fait une hypothèse probabiliste sur les données du type "les données observées sont les réalisations d'une variable aléatoire dont la loi dépend du paramètre θ". Soit un phénomène aléatoire que l'on peut rapprocher d'une loi connue mais de paramètre θ inconnu. Par exemple une loi normale $\mathcal{N}(\theta, 1)$.

Définition 3.1. Ici, on appelera *n-échantillon* un n-uplet $D_n = ((X_1, Y_i) \dots, (X_n, Y_n))$ de couple de variables aléatoires réelles indépendantes et identiquement distribuées où $X = (X_i)_{1 \leqslant i \leqslant n} \in \mathcal{X}$ est une série statistique et $Y_i \in \mathcal{Y}$ la donnée observée sur l'individu X_i.

Exemple 3.1. Phénomène étudié : le marché immobilier.

- $\mathcal{X}$: l'ensemble des caractéristiques des logements sur Rennes,
- $X_i \in \mathcal{X}$: un p-uplet de données sur un logement à Rennes $(x_1, x_2, x_3, \dots) = (m^2$, nb de chambres, nb de salles de bains, ...),
- $\mathcal{Y}$: l'ensemble des prix possibles,
- $Y_i \in \mathcal{Y}$: le prix donné au logement X_i.

Définition 3.2. On appelle *prédicteur* une application mesurable $f : \mathcal{X} \to \mathcal{Y}$. On appelle *estimateur* de θ toute application mesurable Z qui à un n-échantillon associe un prédicteur.

Exemple 3.2. Dans l'exemple précédent, f est une fonction qui, à partir d'un p-uplet de données sur un logement à Rennes, donne une "bonne" estimation de son prix.

Remarque 3.1. On peut prendre un peu tout et n'importe quoi comme estimateur : pour θ la moyenne d'une gaussienne, on peut prendre $\bar{X}, \min(X_i), \max(X_i), X_1 - X_n, \exp(X_1^2 - 1), ...$ Le but est donc d'en trouver un meilleur que les autres. Mais que veut dire "meilleur" ?

Définition 3.3. Soit $\hat{\theta}_n$ un estimateur de θ admettant une espérance. Le *biais* de l'estimateur est $\mathbb{E}(\hat{\theta}_n) - \theta$. Il est dit *sans biais* si son biais est nul et est dit *asymptotiquement sans biais* si $\lim_{n \to \infty} \mathbb{E}(\hat{\theta}_n) - \theta = 0$.

Définition 3.4. Un estimateur $\hat{\theta}_n$ est *faiblement convergent* (ou *faiblement consistant*) si $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$. Il est dit *fortement consistant* si $\hat{\theta}_n \xrightarrow{p.s.} \theta$.

Définition 3.5. Si l'estimateur $\hat{\theta}_n$ admet un moment d'ordre 2, le *risque quadratique* de $\hat{\theta}_n$ est

$$RQ(\hat{\theta}_n, \theta) \coloneqq \mathbb{E}[(\hat{\theta}_n - \theta)^2].$$

Remarque 3.2. Si l'estimateur est sans biais, le risque quadratique de celui-ci est sa variance. En général, le risque quadratique décrit la dispersion de l'estimateur autour de θ. On retient alors ce risque comme un indicateur de la qualité d'un estimateur.

Remarque 3.3. Puisque on a l'égalité

$$RQ(\hat{\theta}_n, \theta) = \Big(\mathbb{E}(\hat{\theta}_n) - \theta\Big)^2 - Var(\hat{\theta}_n),$$

pour que le risque soit faible (et donc pour que l'estimateur soit de bonne qualité), il faut que son biais et sa variance soient faibles.

Propriété 3.6. Si $\hat{\theta}_n$ admet un risque quadratique tendant vers 0 en l'infini, alors $\hat{\theta}_n$ est consistant.

Remarque 3.4. Lorsque l'on cherche à comparer les performances de deux estimateurs, il suffit souvent de calculer leurs risques quadratiques : on choisira systématique l'estimateur dont le risque quadratique tend vers 0 le plus rapidement.

Remarque 3.5. On pourrait remplacer partout θ par $g(\theta)$ avec g une application quelconque. En effet, parfois il est plus aisé d'estimer une fonction de θ, par exemple $1/\theta$, que θ lui-même.

3.2 Estimation par intervalle de confiance

Lors de l'estimation ponctuelle, on cherche une valeur "probable" à attribuer à θ. L'estimation par intervalle de confiance consiste à trouver un intervalle dans lequel il est "probable" que θ soit.

Définition 3.7. Soit $\alpha \in [0, 1]$. On appelle *intervalle de confiance de niveau* $1 - \alpha$ pour θ tout intervalle I_n tel que $\mathbb{P}(\theta \in I_n) = 1 - \alpha$ dont les bornes sont des variables aléatoires indépendantes de θ.