Université d'Evry Val-d'Essonne Paris-Saclay **Année universitaire 2019-2020**
L2 SDV - MSV31

Statistiques descriptives

Les savoir-faire pour les évaluations

1. Savoir définir le cadre d'étude d'un problème : Population, échantillon, variable(s), modalités. Savoir identifier le mode
2. Savoir reconnaitre les types de variables : quantitative/qualitative, discrète/continue, ordinale/nominale
3. Savoir calculer les 4 types de moyennes (observées) : arithmétique, géométrique, harmonique, quadratique (et reconnaître laquelle utiliser !)
4. Savoir calculer variance et écart-type avec la définition
5. Et savoir utiliser le théorème de König-Huygens (la démonstration n'est pas à connaître)
6. Savoir qu'il existe une formule de la variance recomposée de deux groupes. Ne pas connaître la formule par cœur, seulement savoir qu'elle existe et qu'elle permet de comparer statistiquement deux moyennes
7. Connaître la formule de la transformation affine pour la moyenne et la variance
8. Savoir calculer les quantiles avec la définition donnée. En particulier, médiane, quartiles, déciles et écart interquartile. Savoir aussi calculer l'étendue
9. Savoir construire une boîte à moustache (boxplot)
10. Savoir construire un histogramme et un diagramme en bâton

---

Résumé et compléments

Les moyennes

Moyenne arithmétique

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Moyenne géométrique

$$\overline{x}_G = \sqrt[n]{x_1 \dots x_n}$$

Moyenne harmonique

$$\overline{x}_H = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$

Moyenne quadratique

$$\overline{x}_Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Inégalité entre les moyennes

Théorème : Si $x_1, ..., x_n > 0$, alors

$$\min(x_1, \dots, x_n) \leq \overline{x}_H \leq \overline{x}_G \leq \overline{x} \leq \overline{x}_Q \leq \max(x_1, \dots, x_n)$$

Variance et ses propriétés

Variance (observée) :

$$Var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Ecart-type (observé) :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Théorème de König-Huygens

$$Var = \overline{x^2} - \overline{x}^2$$

avec

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 \quad \text{et} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

Démonstration :

$$Var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}2x_i\overline{x} + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 = \overline{x^2} - 2\overline{x}\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}n\overline{x}^2$$

$$= \overline{x^2} - 2\overline{x}\,\overline{x} + \overline{x}^2 = \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 = \overline{x^2} - \overline{x}^2$$

Variance recomposée de deux groupes

Données de taille n_1, moyenne $\overline{x_1}$ et variance σ_1^2
Données de taille n_2, moyenne $\overline{x_2}$ et variance σ_2^2
Données de taille $n_1 + n_2$, moyenne $\overline{x}$ variance σ^2

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1(\overline{x_1} - \overline{x})^2 + n_2(\overline{x_2} - \overline{x})^2}{n_1 + n_2}$$

avec

$$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

Démonstration :
La première série statistique est composée des nombres $x_1, ..., x_{n_1}$
La deuxième série statistique est composée des nombres $x_{n_1+1}, ..., x_{n_1+n_2}$
La moyenne de la série globale $x_1, ..., x_{n_1+n_2}$ est notée $\overline{x}$.
On a

$$\overline{x} = \frac{1}{n_1 + n_2}\sum_{i=1}^{n_1+n_2}x_i$$

On peut vérifier que

$$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

Concernant la variance globale, on a par définition :

$$\sigma^2 = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}(x_i-\overline{x})^2 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=n_1+1}^{n_1+n_2}(x_i-\overline{x})^2\right)$$

Pour faciliter le calcul, on multiplie l'égalité par n_1+n_2, on obtient :

$$(n_1+n_2)\sigma^2 = \sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=n_1+1}^{n_1+n_2}(x_i-\overline{x})^2$$

$$= \sum_{i=1}^{n_1}((x_i-\overline{x_1})+(\overline{x_1}-\overline{x}))^2 + \sum_{i=n_1+1}^{n_1+n_2}((x_i-\overline{x_2})+(\overline{x_2}-\overline{x}))^2$$

$$= n_1\sigma_1^2 + 2\sum_{i=1}^{n_1}(x_i-\overline{x_1})(\overline{x_1}-\overline{x}) + n_1(\overline{x_1}-\overline{x})^2$$

$$+n_2\sigma_2^2 + 2\sum_{i=n_1+1}^{n_1+n_2}(x_i-\overline{x_2})(\overline{x_2}-\overline{x}) + n_2(\overline{x_2}-\overline{x})^2,$$

en utilisant la définition des moyennes $\overline{x_1}$ et $\overline{x_2}$ chacune des deux sommes restantes s'annule et le résultat est démontré. En effet, on a par exemple :

$$\sum_{i=1}^{n_1}(x_i-\overline{x_1})(\overline{x_1}-\overline{x}) = (\overline{x_1}-\overline{x})\sum_{i=1}^{n_1}(x_i-\overline{x_1}) = (\overline{x_1}-\overline{x})\left(\sum_{i=1}^{n_1}x_i - n_1\overline{x_1}\right) = (\overline{x_1}-\overline{x})\,(n_1\overline{x_1} - n_1\overline{x_1}) = 0$$

σ^2 = variance **intrapopulation** + variance **interpopulation**

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2} + \frac{n_1(\overline{x_1}-\overline{x})^2 + n_2(\overline{x_2}-\overline{x})^2}{n_1+n_2}.$$

$(n_1+n_2)\sigma^2$ = SCE = somme des carrés des écarts

$$SCE_{\text{total}} = SCE_{\text{residu}} + SCE_{\text{facteur}}$$

Étudier $\dfrac{SCE_{\text{facteur}}}{SCE_{\text{residu}}}$ pour savoir si $\overline{x_1}$ et $\overline{x_2}$ sont "les mêmes" ! Possibilité de test statistique...

Transformations affines

Soient a et b des nombres réels. Pour les données $y_i = ax_i + b$ on obtient

$$\overline{y} = a\overline{x} + b \quad \text{et} \quad Var(y) = a^2 Var(x)$$

Les quantiles

On note $x_{(1)}, x_{(2)}, ..., x_{(n)}$, avec des indices entres parenthèses, les données $x_1, ..., x_n$ triées par ordre croissant.

Quantiles en statistique descriptive

Le **quantile d'ordre** $\alpha \in]0,1[$, noté q_α, est défini comme étant "la plus petite valeur" de la série permettant d'avoir au moins αn nombres inférieurs (ou égaux) à lui-même.

$$q_\alpha = \begin{cases} x_{(\lceil \alpha n \rceil)} & \text{si} \quad \alpha n \notin \mathbb{N} \\ \dfrac{x_{(\alpha n)} + x_{(\alpha n+1)}}{2} & \text{si} \quad \alpha n \in \mathbb{N} \end{cases}$$

avec $\lceil \cdot \rceil$ désignant l'opération partie entière supérieure. (Exemples $\lceil 2.6 \rceil = 3$, $\lceil 23.11 \rceil = 24$.)

Si $\alpha n \in \mathbb{N}$ le quantile n'est pas une valeur de la série statistique, d'où l'utilisation des guillements pour "plus petite valeur".

Principaux quantiles

— 4 parts égales, on a 3 **quartiles** $(Q_1, Q_2, Q_3) = (q_{\frac{1}{4}}, q_{\frac{2}{4}}, q_{\frac{3}{4}})$

— 10 parts égales, on a 9 **déciles** $(D_1, ..., D_9) = (q_{\frac{1}{10}}, ..., q_{\frac{9}{10}})$

— 100 parts égales, on a 99 **centiles** $(C_1, ..., C_{99}) = (q_{\frac{1}{100}}, ..., q_{\frac{99}{100}})$

La médiane Me = indicateur de position centrale :

$$Me = Q_2 = D_5 = C_{50} = q_{\frac{1}{2}}$$

Ecart interquartile :

$$IQ = Q_3 - Q_1$$

Boîte à moustache (boxplot)

1. Le long d'un axe gradué, on représente un **rectangle** délimité par les quartiles Q_1 et Q_3 et coupé en deux par un trait à hauteur de Me
2. Cette boîte est complétée par des **moustaches** (des traits reliant la boîte) s'arrêtant à la valeur $x_{(a)}$ réalisant le minimum des $x_{(i)}$ dans $[Q_1 - 1.5\,IQ, Q_1]$ et $x_{(b)}$ réalisant le maximum des $x_{(i)}$ dans $[Q_3, Q_3 + 1.5\,IQ]$.
3. Les valeurs en dehors des moustaches sont repérées par **des croix ou des ronds**. Ce sont des valeurs dites extrêmes (outliers).

Représentations graphiques

Histogrammes

= Rectangles adjacents, uniquement pour les **variables continues**

1. L'axe des abscisses est découpé en classes successives : $[a_0, a_1[$, $[a_1, a_2[$,...,$[a_{K-1}, a_K[$
2. h_i est la hauteur de la classe i ($[a_{i-1}, a_i[$) d'effectif n_i donné par

$$n_i = \#\{x_k \text{ tels que } x_k \in [a_{i-1}, a_i[\}$$

et h_i est défini par

$$h_i = \frac{n_i}{a_i - a_{i-1}}$$

Diagrammes en bâtons

= Bâtons adjacents, uniquement pour les **variables discrètes**
La hauteur de chaque bâton est proportionnelle au nombre d'occurrences de la modalité placée sous le bâton