RESSOURCE : STATISTIQUES DESCRIPTIVES

Les **statistiques descriptives** servent à décrire les caractéristiques d'un ensemble d'observations que l'on appelle **échantillon**. C'est l'étape préliminaire à toute modélisation statistique. Il s'agit de *prendre contact* avec les observations, résumer l'information disponible, mettre en évidence d'éventuelles tendances ou valeurs extrêmes. Chaque caractéristique est appelée **variable**.

Exemple conducteur : Un jardinier souhaite connaître les caractéristiques des gousses de quatre espèces. Il ramasse 252 gousses et relève sur chacune d'elles sa masse, sa taille, son espèce, le nombre de graines de la gousse et si la gousse est suffisamment grande pour être vendue selon les normes en vigueur (1 si oui et 0 si non). Voici un extrait des données :

	masse	taille	graines	espece	norme
1	28.6	19.1	4	glycine blanche	1
2	20.6	14.8	3	glycine blanche	1
3	29.2	19.7	5	glycine blanche	1
4	32.0	21.1	7	glycine blanche	1
5	24.5	19.4	4	glycine blanche	1
6	29.0	19.5	4	glycine blanche	0

Ici un individu est une gousse, l'échantillon est donc composé de 252 individus et de 4 variables : masse, taille, espèce, norme et nombre de graines.

On étudie d'abord chacune des variables séparément, on parle alors de statistiques descriptives **univariées**. On étudie ensuite le lien entre des couples de variables, on met alors en œuvre des statistiques descriptives **bivariées**.

On distingue deux types de variables :

- Une variable est dite **quantitative** quand toutes ses valeurs sont numériques. Une variable quantitative sera dite **discrète** si elle prend un nombre fini (ou dénombrable) de valeurs (par exemple, toutes les valeurs entières positives). Elle sera dite **continue** si elle prend toutes les valeurs d'un intervalle.
- les **variables qualitatives** : une variable est dite qualitative lorsque ses valeurs possibles sont des catégories, appelées **modalités**.

Exemple : Dans le jeu de données ci-dessus, les variables masse et taille sont des variables quantitatives continues, le nombre de graines est une variable quantitative discrète et l'espèce est une variable qualitative prenant 4 modalités différentes dont la première est "glycine blanche", norme est une variable qualitative dont les deux modalités sont 0 et 1.

Remarque : on voit sur cet exemple qu'une variable peut être de type qualitative mais posséder des modalités numériques.

STATISTIQUES UNIVARIÉES DES VARIABLES QUALITATIVES

2.1 Table de comptage

La description d'une variable qualitative consiste en l'énumération de ces modalités et de l'occurrence de chacune de ces modalités. Cette information est souvent résumée sous la forme d'une **table de comptage** ou d'une **table de fréquence**.

Exemple : Dans notre cas, la variable présente 4 modalités (Glycine blanche, Glycine violette, Bignone et Laurier Rose) avec respectivement 54, 56, 70 et 72 gousses pour chaque modalité. On obtient donc les tables suivantes :

bignone	glycine blanche	glycine violette	laurier rose
70	54	56	72

bignone	glycine blanche	glycine violette	laurier rose
0.2777	0.2142	0.2222	0.2857

La première table est une table de comptage (*ex* : on a compté 70 bignones dans ce jeu de données), la seconde est une table de fréquence (*ex* : les glycines violettes représentent environ 22% de ce jeu de données). On peut en déduire que la répartition des données est **équilibrée** selon les espèces.

2.2 Lecture de graphiques

La table d'une variable qualitative est souvent remplacée lorsqu'il n'y a pas trop de modalités par des graphiques, en général des **diagrammes en bâtons** ou des **diagramme en camembert**.

Exemple : dans notre cas, la variable espece peut être représentée par l'un ou l'autre des graphes ci-dessous :

STATISTIQUES UNIVARIÉES DES VARIABLES QUANTITATIVES

3.1 Des indicateurs de position

Les indicateurs de positions servent à qualifier la position d'un individu au sein du jeu de données : est-il proche de l'individu moyen ou au contraire atypique ?

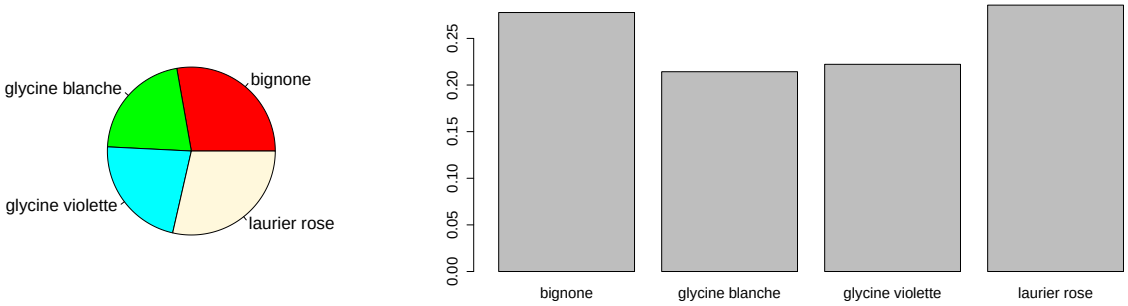


FIGURE 1 – Diagramme en camembert (à gauche) et diagramme en bâtons (à droite) de la variable espece

Deux premiers indicateurs de positions sont le **minimum** et le **maximum** d'une variable.

La **moyenne empirique** d'un échantillon $x_1, \ldots, x_n$ est définie par :

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

Elle est très sensible aux valeurs extrêmes.

La **médiane** d'un échantillon $x_1, \ldots, x_n$ est le réel m tel que le nombre d'observations qui la précèdent est égal au nombre d'observations qui la suivent. Dans le cas d'un nombre pair d'observations, la médiane n'est pas définie de manière unique mais une médiane de l'échantillon peut être définie par la moyenne entre la $n/2$-ième valeur et la $(n/2+1)$-ième valeur. Comme la moyenne, la médiane décrit un individu moyen du jeu de données.

Exemple : La médiane des cinq premières valeurs de la variable masse est de 28.6 (3-ème valeur par ordre croissant). Une médiane des six premières valeurs du jeu de données est la moitié entre 28.6 (3-ième valeur) et 29 (4-ième valeur) soit 28.8.

La médiane est un cas particulier de ce que l'on appelle **quantile**. Le **quantile d'ordre p** est la valeur x_p telle qu'il y ait une proportion p des observations qui soient inférieures ou égales à x_p.

En particulier si $p = 1/4$ ou $p = 3/4$ les quantiles associés sont appelés **premier quartile** et **troisième quartile**, si $p \in \{1/10, \ldots, 9/10\}$ ils sont appelés les **déciles** et pour $p \in \{1/100, \ldots, 99/100\}$ les **percentiles**.

Exemple : Supposons que le premier quartile de la variable taille vaut 11 et que son huitième décile (ou 80-ième percentile) vaut 20. Cela signifie que dans le jeu de données relevé, un quart des gousses ont une taille

inférieure à 11 mm tandis que 20% des gousses ont une taille supérieure à 20 mm.

3.2 Des indicateurs de dispersion

Les indicateurs de dispersions servent à caractériser l'ensemble du jeu de données : est-il peu ou très hétérogène / dispersé ?

Le premier indicateur de dispersion est la **variance empirique** d'un échantillon $x_1, \ldots, x_n$: elle permet de mesurer la dispersion autour de leur moyenne. Elle est définie par :

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

On trouve également la définition suivante :

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

On parle alors de **variance sans biais** ou **variance corrigée** (voir cours Estimation Statistique).

L'**écart-type empirique** s_x est la racine carrée de la variance :

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

L'écart-type a l'avantage de s'exprimer dans les mêmes unités que les observations. Il est donc plus facile à interpréter que la variance, on peut notamment regarder son ordre de grandeur par rapport à la valeur moyenne. La variance est plutôt utilisée pour les développements mathématiques.

Exemple : Supposons que la variance des différentes variables quantitative soit donnée ci-dessous :

	taille	masse	graines
variance	13.4264	81.028	2.924

Alors, en prenant la racine carrée de ces valeurs, leurs écarts-type valent (on donne également les valeurs moyennes) :

	taille	masse	graines
écart-type	3.664	9	1.71
moyenne	13.37	11.13	1.274

Comparativement aux valeurs moyennes on peut dire que :

- La taille a une dispersion moyenne
- Les masses des gousses présentent une forte hétérogénéité

Deux derniers indicateurs de dispersions sont :

- **L'étendue** : différence entre le maximum de la variable et son minimum
- **L'étendue interquartile** : différence entre le troisième quartile de la variable et son premier quartile

3.3 Lecture de graphiques

Les indicateurs numériques de dispersion et de position peuvent aussi être résumés graphiquement. Voici les graphiques les plus usuellement rencontrés.

Un **histogramme** est une figure constituée de rectangles juxtaposés dont la base correspond à l'amplitude de chaque classe et dont la surface est proportionnelle à l'effectif de cette classe. La **boite à moustache** (**boxplot** en anglais) est un graphique où sont représentés la médiane (tiret central), le premier et le troisième quartile (valeurs extrêmes du "rectangle central"), le premier et le neuvième décile (les "bouts de la moustache") ainsi que quelques valeurs extrêmes parfois, représentées par des points. Ces deux graphiques donnent une idée de l'allure globale de la distribution de la variable.

Exemple : Voici les graphiques associés à l'exemple des graines. Le premier histogramme est un histogramme des comptages (frequency en anglais) tandis que le deuxième est un histogramme des fréquences.

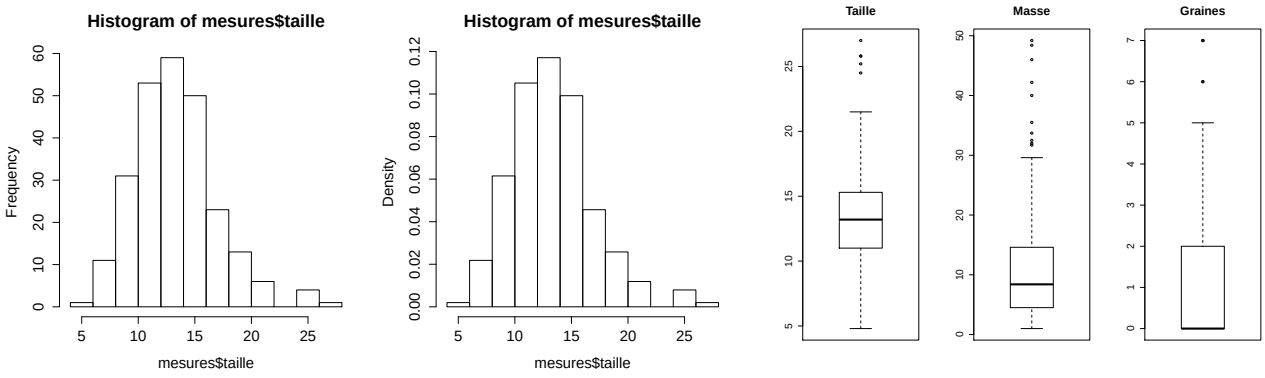


FIGURE 2 – Deux histogrammes de la taille à gauche et trois boites à moustache à droite

Sur l'histogramme on lit que la variable taille semble symétrique, présente une forme en cloche (voir cours Variable Gaussienne), sa dispersion est moyenne, quelques valeurs extrêmes au delà de 25 grammes sont présentes. On retrouve ces informations sur la première boite à moustache. On peut également noter que la variable masse est plus asymétrique, avec une forte proportion de graines de faible masse, de même pour la variable graines : la moitié des gousses relevées ne présentant pas de graines. Enfin on peut lire par exemple les valeurs suivantes : une médiane de l'ordre de 13 pour la taille (50% des gousses ont dont une taille inférieure à 13 mm), un premier quartile de l'ordre de 5 pour la masse (25% des gousses ont dont une taille inférieure à 5g) ou encore un neuvième décile de l'ordre de 5 pour le nombre de graines (seul un dixième des gousses présentent plus de 5 graines).

POUR ALLER PLUS LOIN : STATISTIQUES BIVARIÉES

4.1 Croisement de deux variables qualitatives

Le croisement de deux variables qualitatives se résume généralement dans une **table de contingence** : chaque case renferme le nombre d'individus correspondant au croisement entre une certaine modalité de la première variable et une certaine modalité de la seconde variable.

Exemple : Voici une table de contingence pour les variables espèce et norme :

	bignone	glycine blanche	glycine violette	laurier rose
aux normes	34	53	51	69
pas aux normes	1	2	3	3

L'étude des tables de contingence relève d'un domaine appelé Plans d'expérience, voir le cours dédié.

4.2 Croisement d'une variable quantitative avec les modalités d'une variable qualitative

L'étude du croisement entre une variable quantitative et une variable qualitative consiste simplement en la comparaison des indicateurs de position, de dispersion, ou des graphiques associées de la variable quantitative en fonction de chaque modalité.

Exemple : Voici un tableau de statistiques de la variable masse par rapport aux modalités de la variable espece :

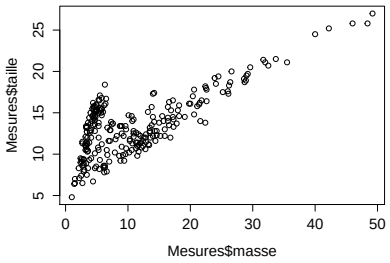
	bignone	glycine blanche	glycine violette	laurier rose
moyenne	7.5	8.9	17	21
écart-type	9	9.2	10.5	14
maximum	19	20	48	50

On remarque que les 4 espèces présentent des masses différentes, la glycine blanche semble posséder les graines les plus légères tandis que le laurier rose semble posséder les plus lourdes, avec une masse moyenne, mais également une masse maximum supérieure. Leur dispersion est quant à elle comparable.

4.3 Croisement de deux variables quantitatives

Lorsqu'on étudie le comportement de deux variables quantitatives sur un même ensemble d'individus la représentation graphique utilisée est le **nuage de points**, par exemple :

Le **coefficient de corrélation** est utilisé pour caractériser le **lien linéaire** entre deux variables quantitatives x et y. Cette notion est vue en détail dans le cours dédié.


Mesures$taille
5
10
15
20
25
0
10
20
30
40
50
Mesures$masse