Statistique descriptive: distribution à un seul caractère

Rania RAIS

Première année licence informatique de gestion

Octobre 2020

Introduction

- Statistique : La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation dans le but d'informer et d'aider à la prise de décision.

- Les statistiques sont aujourd'hui utilisées dans tous les secteurs d'activité :
  - Industrie : fiabilité, contrôle qualité, ...
  - Economie et finance : sondages, enquête d'opinion, assurance, marketing
  - Santé, environnement, ...
  - Partout où l'on dispose de données

Définitions

- L'ensemble de personnes ou d'objets équivalents étudié s'appelle la population.
  Exemples : les étudiants de 19 à 23 ans.
- Chaque objet d'une population s'appelle un individu ou unité statistique.
  Exemples : personne humaine, automobile, entreprise, pays ,...
- Les caractéristiques que l'on observe sur chacun des individus s'appellent des variables.
  Exemples : sexe, âge, taille, nombre d'enfants,...
- Modalité d'un caractère : c'est l'ensemble des valeurs que peut prendre un caractère donné.
  Exemples : sexe $\{M,F\}$, taille $\{[1.5m,2m]\}$, nombre de filles : $\{1,2,3,4,...\}$.

Définitions

Il existe deux types de variables :

1. Variables qualitatives : caractéristiques non numériques (sexe, couleur des yeux, secteur d'activité, marque de voitures...).
2. Variables quantitatives : qui représentent des grandeurs mesurables (des relevés de poids, de température, de prix,...).
   - Une variable quantitative discrète est une variable dont les valeurs sont en nombre fini.
   - Une variable quantitative continue est une variable qui peut prendre toutes les valeurs possibles dans un intervalle.

Définitions

- **Recensement** : étude de tous les individus d'une population. Difficile en pratique lorsque les populations sont grandes pour des questions de coût et de temps.
- **Sondage** : recueil d'une partie de la population. La partie des individus étudiées s'appelle l'échantillon. Le recueil d'un échantillon à partir de la population initiale se fait par des techniques statistiques, appelées méthodes d'échantillonnage.

La série d'observations recueillies s'appelle série statistique. Elle est généralement recopier dans un tableau de données.

La statistique descriptive

On notera que les termes : statistique descriptive, statistique exploratoire et analyse des données sont quasiment synonymes.

1. Objectifs :
   - résumer, synthétiser l'information contenue dans une série statistique.
   - mettre en évidence ses propriétés. .
2. Synthèse de l'information :
   - Tableaux,
   - Graphiques (box-plots, histogrammes, diagramme en bâtons...).
3. Statistique descriptive :
   - Analyse statistique uni-variée : 1 seul caractère X.
   - Analyse statistique bi-variée : deux caractères X et Y.

Définitions

Soient x_i une modalité d'un caractère X, k est le nombre de valeurs de X.

- L'effectif total est le nombre d'individus appartenant à la population.
- L'effectif de x_i (noté n_i) est le nombre d'individus présentant cette modalité. On a donc
$$\sum_{i}^{k} n_i = N.$$
- La fréquence de x_i est la proportion d'individus de la population totale qui présente cette modalité.
$$f_i = \frac{n_i}{N} \qquad ; \qquad \sum_{i=1}^{k} f_i = 1.$$
- L'effectif cumulé de x_i (noté N_i) est égal à la somme des effectifs des modalités qui lui sont inférieures ou égales, soit $N_i = \sum_{a=1}^{i} n_a$.
- La fréquence cumulée de x_i (notée F_i) est égale à la somme des fréquences des modalités qui lui sont inférieures ou égales, soit
$$F_i = \sum_{a=1}^{i} f_a = \frac{N_i}{N}.$$

Exemple 1 : variable quantitative discrète

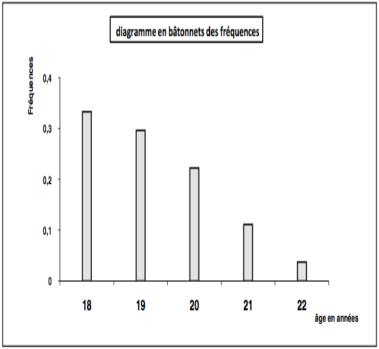
- **But** :Quel est l'âge moyen d'un étudiant de première année en Belgique.
- La population comporte un certain nombre d'individus. Il est évidemment difficile d'interroger tous les étudiants de première année.
- Pour l'exemple, on se limitera donc à un échantillon de 27 étudiants.
- Résultats :

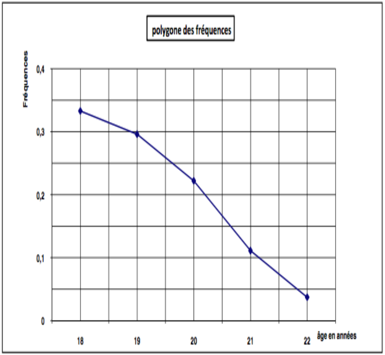
18	20	19	20	21	18	20	18	18
21	19	19	19	18	18	18	21	20
19	19	20	19	18	20	22	19	18

Soit le tableau suivant :

i	x_i	n_i :effectifs	f_i : fréquences	N_i :effectifs cumulés	F_i : fréquences cumulées
1	18	9	$9/27 \approx 33.3\%$	9	$9/27$
2	19	8	$8/27 \approx 29.63\%$	17	$17/27$
3	20	6	$6/27 \approx 22.2\%$	23	$23/27$
4	21	3	$3/27 \approx 11.1\%$	26	$26/27$
5	22	1	$1/27 \approx 3.7\%$	27	1

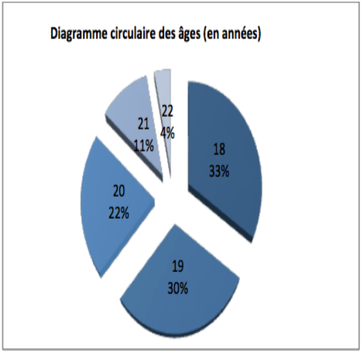
Représentations graphiques :





Représentation graphique :

fréquences f_i en %	fréquences f_i en °
33,3 %	33,3.3,6 = *120°*
29,6 %	106,7°
22,2 %	80°
11,1 %	40°
3,7 %	13,3°



Exemple 2 : variable quantitative continue

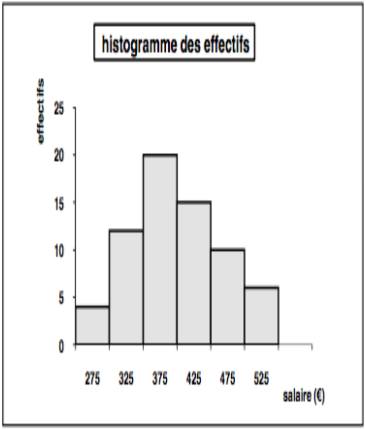
- Soit un tableau reprenant les salaires mensuels bruts en euros :270 275 300 455 642 . . . . . . (67 valeurs différentes).
- Si on classait ces données comme dans le premier exemple, on obtiendrait un très grand tableau avec une colonne d'effectif presque toujours égale à 1 (peu de valeurs sont identiques).
- On regroupe les données en classes.

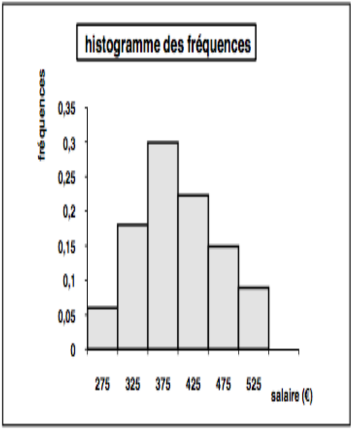
classe $N°$	classes (salaires)	centres des classes	n_i : effectifs	f_i : fréquences	F_i :fréquences cumulées
1	< 300	275	4	$4/67 \approx 6\%$	$\approx 6\%$
2	$[300; 350[$	325	12	$12/67 \approx 17.9\%$	$\approx 23.9\%$
3	$[350; 400[$	375	20	$20/67 \approx 29.9\%$	$\approx 53.8\%$
4	$[400; 450[$	425	15	$15/67 \approx 22.4\%$	$\approx 76.2\%$
5	$[450; 500[$	475	10	$10/67 \approx 14.9\%$	$\approx 91.1\%$
6	≥ 500	525	6	$6/67 \approx 9\%$	$\approx 100\%$

Les nombres 300 et 350 sont appelés respectivement borne inférieure et borne supérieure de la classe.

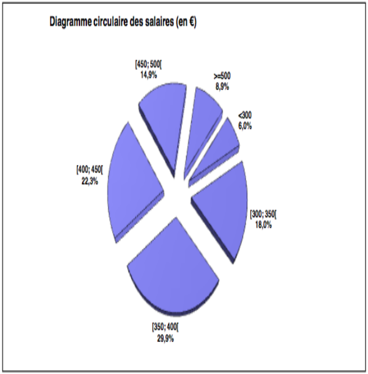
Représentations graphiques

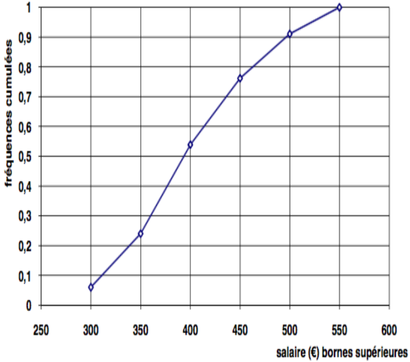
L'histogramme des effectifs ou des fréquences : on porte en abscisse les centres des classes (ou les intervalles de classe) et en ordonnée les effectifs ou les fréquences.





Représentations graphiques





Exemple 3 : variable qualitative

Modèle statistique :

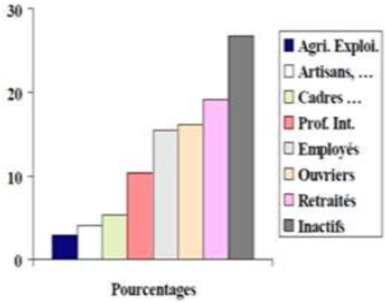
- Population :personnes agées de plus de 15 ans.
- Caractère : situation professionnelle
- Modalités : employés, ouvriers, commerçants, retraites,...

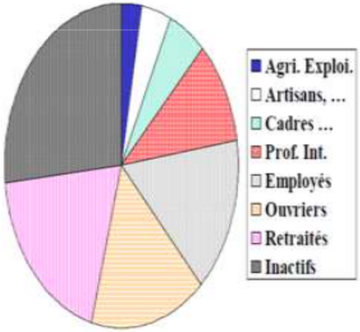
Table de distributions

CSP	Nb de personnes	Pourcentages
Agriculteurs exploitants	1268264	2.9
Artisans, commerçants et chefs d'entreprises	1757221	4.0
Cadres et professions intellectuelles supérieures	2314770	5.3
Professions intermédiaires	4593294	10.4
Employés	6771239	15.4
Ouvriers	7121812	16.2
Retraités	8429509	19.2
Inactifs divers (autres que retraités)	11741884	26.7
Ensemble	43997993	100

Représentations graphiques

Diagramme en barres **Diagramme en secteurs**





Les paramètres de position

Quand les statisticiens se trouvent en face des résultats d'une enquête, ils trouvent intéressant d'en déterminer les "tendances moyennes". Pour cela, ils disposent de plusieurs outils : la moyenne arithmétique, le mode, la médiane et les quantiles.

1) La moyenne arithmétique

- Cas d'une variable discrète :

$$\bar{X} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \sum_{i=1}^{k} f_i x_i.$$

avec

- k est le nombre de valeurs de la variable X,
- x_i sont les valeurs de la variable,
- n_i est l'effectif correspondant à la variable i,
- N est l'effectif total,
- f_i sont les fréquences.

- Cas d'une variable continue :

$$\bar{X} = \frac{\sum_{i=1}^{k} n_i c_i}{N} = \sum_{i=1}^{k} f_i c_i.$$

avec c_i représente le centre de la classe i et k le nombre de classes.

  - Exemple : En 1954, une enquête sur la répartition selon l'âge de la population agricole masculine a donné les résultats suivants :

Age en années X_i	*Centres de classe*	*Effectifs n_i*
[15; 25[	20	197
[25; 35[	30	207
[35; 45[	40	151
[45; 55[	50	189
[55; 65[	60	127
[65; 75[	70	108
75 et plus	80	21
		n = 1000

$$\bar{X} = \frac{20.197+30.207+40.151+50.189+60.127+70.108+80.21}{1000}$$

2) Le mode est la valeur de la variable (ou la classe) dont l'effectif est le plus important.

Exemples

1. Dans le cas de variable à valeurs numériques, si on reprend l'exemple des âges des étudiants, on s'aperçoit que l'âge que l'on retrouve le plus souvent est 18 (9 effectifs).
2. Dans une enquête relative au moyen de transport, on a obtenu le tableau suivant :

Moyens de transport	*Effectifs*
vélo	7
bus	10
tram	2
vélomoteur	5
à pied	6

Dans ce cas, il n'est pas possible de calculer une moyenne. On pourrait cependant se demander quel est le moyen de transport le plus utilisé.

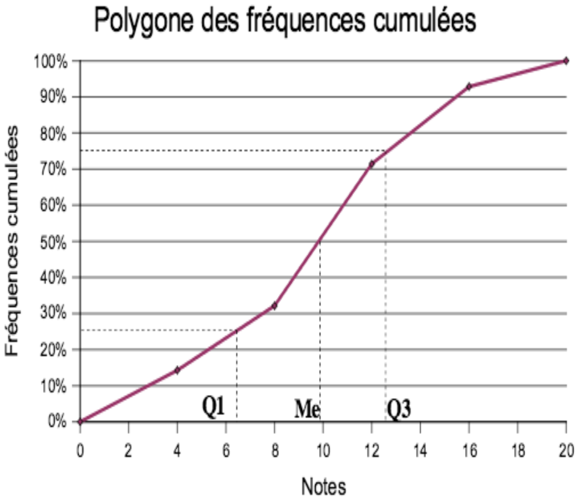
3) La médiane d'une variable statistique est la valeur de la variable (ou la classe) qui partage l'effectif en deux parties égales.

Remarque : Pour déterminer la médiane, il existe diverses formules dans la littérature. On se limitera à déterminer la médiane sur le diagramme des fréquences cumulées (la médiane est l'abscisse correspondant à une fréquence de 50 %).

4) Les quantiles que l'on rencontre le plus souvent sont (définis à partir de la courbe des fréquences cumulées) :

- La médiane Me correspond à une fréquence cumulée de 50%.
- Les quartiles Q_1 , $Q_2 = Me$ et Q_3 correspondent aux fréquences cumulées 25%, 50%, et 75%. Ils partagent l'ensemble des observations en 3 parties de même effectif.

Exemple On considère la représentation graphique des notes du DS en polygone des fréquences cumulées.



Les paramètres de dispersion

Dans le cas d'une variable quantitative, on appelle :

- L'étendue d'une série statistique est la différence entre la valeur maximum et la valeur minimum de la série.
- L'intervalle interquartile $[Q_1, Q_3]$ ou l'écart interquartile $EIQ = Q_3 - Q_1$ mesure la dispersion des valeurs observées autour de la médiane.
- Ecart : la valeur absolue de la différence entre la moyenne et une valeur de la variable.
- Ecart moyen : la moyenne de la série des écarts de tous les individus de la population.

Les paramètres de dispersion

- La variance est la moyenne de la série des carrés des écarts entre la moyenne et les valeurs de la variable de tous les individus de la population.

$$V(X) = \frac{1}{N}\sum_i n_i(x_i - \bar{X})^2 = \frac{1}{N}\sum_i n_i x_i^2 - \bar{X}^2 = \sum_i f_i(x_i - \bar{X})^2$$

- L'écart-type est la racine carrée de la variance noté par σ.
- On désigne par $\left[\bar{X} - \sigma ; \bar{X} + \sigma\right]$ l'intervalle moyen. On dit qu'en moyenne, les valeurs observées se trouvent dans l'intervalle moyen.

Exemple 4 :

Pour étudier le nombre d'enfants dans les familles amiénoises, on a interrogé 1000 familles et on a obtenu les résultats suivants :

Nombre d'enfants	0	1	2	3	4	5	6	7	8
Nombre de familles	162	240	297	208	62	16	8	5	2

Modèle statistique :

- Population : les familles.
- Variable X : le nombre d'enfants, quantitative discrète.
- Echantillon de $N = 1000$ familles.

Soit le tableau suivant :

Valeur x_i	Effectif n_i	Eff. Cum. N_i	Fréquence f_i	Fréq. Cum. F_i	$n_i x_i$	$n_i x_i^2$
0	162	162	0,162	0,162	0	0
1	240	402	0,240	0,402	240	240
2	297	699	0,297	0,699	594	1188
3	208	907	0,208	0,907	624	1872
4	62	969	0,062	0,969	248	992
5	16	985	0,016	0,985	80	400
6	8	993	0,008	0,993	48	288
7	5	998	0,005	0,998	35	245
8	2	1000	0,002	1	16	128
Total	1000		1		1885	5353

Soit le tableau suivant :

Valeur x_i	Effectif n_i	Eff. Cum. N_i	Fréquence f_i	Fréq. Cum. F_i	$n_i x_i$	$n_i x_i^2$
0	162	162	0,162	0,162	0	0
1	240	402	0,240	0,402	240	240
2	297	699	0,297	0,699	594	1188
3	208	907	0,208	0,907	624	1872
4	62	969	0,062	0,969	248	992
5	16	985	0,016	0,985	80	400
6	8	993	0,008	0,993	48	288
7	5	998	0,005	0,998	35	245
8	2	1000	0,002	1	16	128
Total	1000		1		1885	5353

- Moyenne : $\bar{X} = \dfrac{1}{1000} \times 1885 = 1.885.$
- Variance : $V(X) = \dfrac{1}{1000} \times 5353 - (1.885)^2 \approx 1.79.$
- Ecart-type : $\sigma(X) \approx 1.34$
- Intervalle moyen :$[\bar{X} - \sigma; \bar{X} + \sigma] = [0.54; 3.23].$
  En moyenne, les famille ont un nombre d'enfants compris entre 1 et 3 enfants.

Merci pour votre attention

Statistique descriptive bi-variée

Rania RAIS

Première année licence informatique de gestion

Octobre 2020

Analyse statistique bi-variée

- On considère une population sur laquelle on étudie deux variables (ou caractères) X et Y.
- On étudiera donc des séries statistiques à deux variables ; autrement dit un couple de variables (X, Y).
- X et Y pouvant être de nature différente : qualitative, quantitative discrète ou continue. On note $(x_i)_{i=1,\dots,k}$ les k modalités de X et $(y_i)_{i=1,\dots,l}$ les l valeurs de Y.
- Les deux variables X et Y sont mesurées simultanément sur chacun des N individus de la population. On notera $n_{i,j}$ l'effectif correspondant au couple (x_i, y_j).

- Pour chaque indice i, l'effectif $n_{i,.}$ est le nombre total d'observations de la modalité x_i de X quelle que soit la modalité de Y. C'est-à-dire

$$n_{i,.} = \sum_{j=1}^{l} n_{i,j} = \text{total de la ligne}\, i.$$

- Les k couples $(x_i, n_{i,.})$ définissent la distribution marginale de la variable X.
- Pour chaque indice j, l'effectif $n_{.,j}$ est le nombre total d'observations de la modalité y_j de Y quelle que soit la modalité de X. C'est-à-dire

$$n_{.,j} = \sum_{i=1}^{k} n_{i,j} = \text{total de la colonne}\, j.$$

- Les l couples $(y_j, n_{.,j})$ définissent la distribution marginale de la variable Y.

Définitions

- La fréquence du couple (x_i, y_j) est
$$f_{i,j} = \frac{n_{i,j}}{N}; \qquad \sum_i \sum_j f_{i,j} = 1.$$
- La fréquence marginale de x_i est
$$f_{i,.} = \frac{n_{i,.}}{N}; \qquad \sum_i f_{i,.} = 1.$$
- La fréquence marginale de y_j est
$$f_{.,j} = \frac{n_{.,j}}{N}; \qquad \sum_j f_{.,j} = 1.$$
- La fréquence conditionnelle de x_i sachant que $Y = y_j$ est
$$f_{x_i|y_j} = \frac{n_{i,j}}{n_{.,j}}.$$
- La fréquence conditionnelle de y_j sachant que $X = x_i$ est
$$f_{y_j|x_i} = \frac{n_{i,j}}{n_{i,.}}.$$

Indépendance

- X est dite indépendante de Y si les variations de Y n'entrainent pas des variations de X.
- Si X est indépendante de Y alors Y est indépendante de X. On dit que X et Y sont indépendantes.
- Les deux variables sont indépendantes si et seulement si

$$f_{ij} = f_{i,.} \times f_{.,j}$$

Exemple : étude de deux variables quantitatives

Une entreprise employant 100 femmes relève pour chaque femme son âge, noté X, et le nombre de journées d'absence durant le mois de janvier, noté Y.

X \ Y	0	1	2	3
$[20, 30[$	0	0	5	15
$[30, 40[$	0	15	20	0
$[40, 50[$	15	10	5	0
$[50, 60[$	0	5	5	5

Exemple

Calculs des effectifs $n_{i,.}$ et $n_{.,j}$; $i, j = 1 \ldots 4$:

$X \backslash Y$	0	1	2	3	$n_{i,.}$
$[20, 30[$	0	0	5	15	
$[30, 40[$	0	15	20	0	
$[40, 50[$	15	10	5	0	
$[50, 60[$	0	5	5	5	
$n_{.,j}$					

Exemple

Calculs des effectifs $n_{i,.}$ et $n_{.,j}$; $i, j = 1 \dots 4$:

$X \backslash Y$	0	1	2	3	$n_{i,.}$
$[20, 30[$	0	0	5	15	20
$[30, 40[$	0	15	20	0	35
$[40, 50[$	15	10	5	0	30
$[50, 60[$	0	5	5	5	15
$n_{.,j}$	15	30	35	20	

Exemple

Calculs des fréquences marginales $f_{i,.}$:

$X \backslash Y$	0	1	2	3	$f_{i,.}$
[20,30[	0	0	5	15	
[30,40[	0	15	20	0	
[40,50[	15	10	5	0	
[50,60[	0	5	5	5	

Exemple

Calculs des fréquences marginales $f_{i,.}$:

X \ Y	0	1	2	3	$f_{i,.}$
[20,30[	0	0	5	15	
	0	0	0.05	0.15	0.2
[30,40[	0	15	20	0	
	0	0.15	0.2	0	0.35
[40,50[	15	10	5	0	
	0.15	0.1	0.05	0	0.3
[50,60[	0	5	5	5	
	0	0.05	0.05	0.05	0.15

Exemple

Calculs des fréquences conditionnelles $f_{y_j|x_i}$:

X \ Y	0	1	2	3	$n_{i,.}$
[20,30[	0	0	5	15	20
	0	0	1/4	3/4	
[30,40[	0	15	20	0	35
	0	3/7	4/7	0	
[40,50[	15	10	5	0	30
	1/2	1/3	1/6	0	
[50,60[	0	5	5	5	15
	0	1/3	1/3	1/3	

Moyennes des distributions marginales

- Moyenne de X :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{k} n_{i,.}\, x_i$$

- Moyenne de Y :

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^{l} n_{.,j}\, y_j$$

- Cas de l'exemple précédent : les x_i sont les centres des classes.

$$\bar{X} = \frac{1}{100}(20 \times 25 + 35 \times 35 + 30 \times 45 + 15 \times 55) = 39.$$
$$\bar{Y} = \frac{1}{100}(15 \times 0 + 30 \times 1 + 35 \times 2 + 20 \times 3) = 1.6.$$

Variances des distributions marginales

- Variance et écart-type de X :

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{k} n_{i,.}\, x_i^2\right) - \bar{X}^2 \quad \text{et} \quad \sigma(X) = \sqrt{V(X)}.$$

- Variance et écart-type de Y :

$$V(Y) = \left(\frac{1}{N}\sum_{j=1}^{l} n_{.,j}\, y_j^2\right) - \bar{Y}^2 \quad \text{et} \quad \sigma(Y) = \sqrt{V(Y)}.$$

- Cas de l'exemple précédent :

$$V(X) = 1615 - 39^2 = 94 \quad \text{donc} \quad \sigma(X) \approx 9.69.$$

$$V(Y) = 3.5 - 1.6^2 = 0.94 \quad \text{donc} \quad \sigma(Y) \approx 0.97.$$

Moyenne et variance des distributions conditionnelles

- Moyenne de X sachant $Y = y_j$:

$$\bar{X}_{|Y=y_j} = \frac{1}{n_{.,j}} \sum_{i=1}^{k} n_{i,j} x_i.$$

- Variance de X sachant $Y = y_j$:

$$V(X_{|Y=y_j}) = \frac{1}{n_{.,j}} \sum_{i=1}^{k} n_{i,j} x_i^2 - \bar{X}^2_{|Y=y_j}.$$

- Cas de l'exemple précédent : déterminer

$$\bar{X}_{|Y=1} =$$

$$V(X_{|Y=1}) =$$

Moyenne et variance des distributions conditionnelles

- Moyenne de X sachant $Y = y_j$:

$$\bar{X}_{|Y=y_j} = \frac{1}{n_{.,j}} \sum_{i=1}^{k} n_{i,j} x_i.$$

- Variance de X sachant $Y = y_j$:

$$V(X_{|Y=y_j}) = \frac{1}{n_{.,j}} \sum_{i=1}^{k} n_{i,j} x_i^2 - \bar{X}^2_{|Y=y_j}.$$

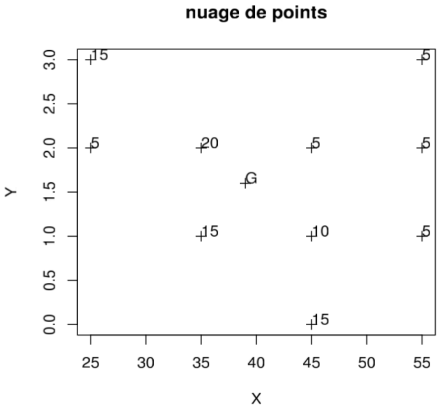
- Cas de l'exemple précédent :

$$\bar{X}_{|Y=1} = \frac{25 \times 0 + 35 \times 15 + 45 \times 10 + 55 \times 5}{30} \approx 41.67.$$

$$V(X_{|Y=1}) = \frac{35^2 \times 15 + 45^2 \times 10 + 55^2 \times 5}{30} - 41.67^2 \approx 55.28.$$

Représentation graphique

- On représente graphiquement cette série bi-variée par un nuage de points de coordonnées (x_i, y_j).
- Le centre de gravité du nuage est alors le point de coordonnées $(\bar{X}; \bar{Y})$.



Définitions

- La covariance est définie par :

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j - \bar{X}\bar{Y}.$$

- Le coefficient de corrélation entre X et Y est la covariance divisée par les deux écart-types $\sigma(X)$ et $\sigma(Y)$:

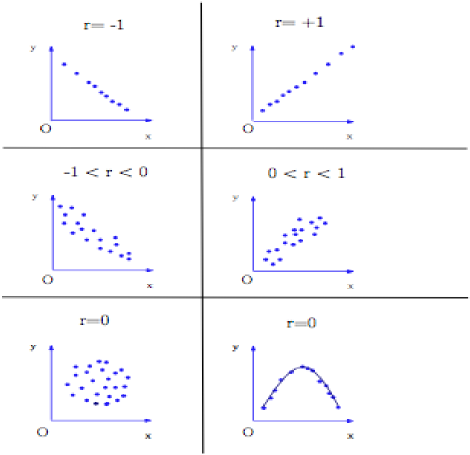
$$r(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}.$$

- Propriétes
  - $r(X,Y) \in [-1,1]$
  - $r(X,Y) = r(Y,X)$
  - $r(X,X) = 1.$

Coefficient de corrélation

- Si $r = 1$, les points sont alignés sur une droite de pente positive.
- Si $r = -1$, les points sont alignés sur une droite de pente négative.
- Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrèmes -1 et 1, plus la corrélation entre les variables est forte.
  - Si $|r|$ proche de 1 : corrélation linéaire. La liaison est considérée comme forte si $|r| > 0.9$.
  - Si $|r|$ proche de 0 : pas de corrélation linéaire.
- Remarque : Quand la corrélation linéaire des données est trés forte, on peut faire de la prévision.

Illustrations



La forme du nuage obtenu peut indiquer le type de dépendance possible entre X et Y.

Régression linéaire

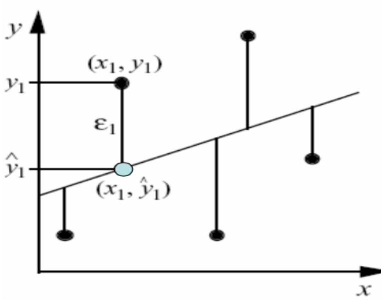
- La régression est une des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives.
- Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle.
- On s'intéresse à la relation entre deux variables quantitatives. On parlera dans ce cas d'une régression simple en exprimant une variable en fonction de l'autre.
- Le modèle de régression linéaire simple : Soit un échantillon de N individus. Pour un individu $i(i = 1, \cdots N)$, on a observé :
  - x_i la valeur de la variable quantitative X.
  - y_i la valeur de la variable quantitative Y.

  On veut étudier la relation entre ces deux variables.

Régression linéaire

- Principe des moindres carrés :

  Soit $\hat{y} = ax + b$ la droite de régression de y en x.
  - Le but est de faire passer cette droite, à travers le nuage de points, de façon à ce que la différence $y - \hat{y}$ soient les plus faibles possible pour l'ensemble des points.
  - La différence $\varepsilon_i = y_i - \hat{y}_i$ porte le nom de résidu pour l'observation i.



Régression linéaire

- Le principe des moindres carrés consiste à choisir les valeurs a et b qui minimisent la somme des carrées des écarts

$$\sum_i \varepsilon_i^2.$$

- La droite des moindres carrés (ou de régression) de y en x a pour équation

$$D : \hat{y} = \widehat{a}x + \widehat{b}, \quad \text{avec } \widehat{a} = \frac{cov(X,Y)}{V(X)} \text{ et } \widehat{b} = \bar{Y} - \widehat{a}\bar{X}.$$

- La valeur de Y prédite par la régression au point x_i est notée $\hat{y}_i$ et vaut $\hat{y}_i = \hat{a}x_i + \hat{b}$.
- La droite des moindres carrés (ou de régression) de x en y a pour équation

$$D' : \hat{x} = \widehat{a'}y + \widehat{b'}, \quad \text{avec } \widehat{a'} = \frac{cov(X,Y)}{V(Y)} \text{ et } \widehat{b'} = \bar{X} - \widehat{a'}\bar{Y}.$$

- Les droites D et D' se coupent donc au point $G(\bar{X}, \bar{Y})$.

Exemple

La ville de Gatineau veut modéliser les fluctuations de la concentration d'ozone dans l'air à midi pour le mois de Juillet. On est donc en présence de deux variables, l'une qui est la température à midi le mois de juillet (X) qui est parfaitement observable et connue et l'autre qui la concentration de l'ozone (Y) qui est moins connue. On va donc essayer de modéliser cette dernière en fonction de la première. La ville a commencé par choisir au hazard 10 journées des mois de Juillet et noter ces deux variables, elle obtient les résultats suivants :

X ($°C$)	23.9	32.8	23.7	7.4	19.7	30.7	28.9	25.4	28.9	27.5
Y (ppm)	118.9	143.8	116.2	87.6	100.9	134.8	98.7	78.6	109.2	102.6

Exemple

Pour déterminer l'équation de régression on a besoin de calculer $\bar{X}, V(X)$, $\bar{Y}$ et $cov(X, Y)$.

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} x_i \approx 24.89; \qquad V(X) = \frac{1}{10}\sum_{i=1}^{10} x_i^2 - \bar{X}^2 \approx 47.07;$$

$$\bar{Y} = \frac{1}{10}\sum_{i=1}^{10} y_i \approx 109.13; \qquad cov(X, Y) = \frac{1}{10}\sum_{i=1}^{10} x_i y_i - \bar{X}\bar{Y} \approx 75.05.$$

En appliquant les formules précédentes, on obtient

$$\widehat{a} = 1.59 \quad \text{et} \quad \widehat{b} = 69.55.$$

La droite de régression est alors :$\widehat{y} = 1.59x + 69.55$. Par exemple, si on veut prédire la concentration d'azote un jour de Juillet où il fait $28\,°C$ à midi, on peut le formaliser comme suit :

$$\widehat{y} = 1.59 \times 28 + 69.55 = 114.08\, ppm.$$

Merci pour votre attention

Introduction aux probabilités

Rania RAIS

Première année licence informatique de gestion

Novembre 2020

1)Généralités

Définition

Une expérience est aléatoire est une expérience dont :

- *elle conduit à plusieurs résultats possibles,*
- *on peut décrire tous ces résultats,*
- *on ne peut pas prévoir le résultat de façon certaine.*

Exemple

- La valeur lue sur la face du dé que l'on jette.
- Lancer d'une pièce.
- Jeu de fléchettes sur un cercles.

Définition

- *Un évènement élémentaire est un résultat possible d'une expérience. Il est en général noté par ω.*
- *L'espace des possibles ou univers décrit tous les résultats possibles de l'expérience. Il est en général noté par Ω.*
- *Un évènement (non nécessairement élémentaire) est un sous-ensemble de Ω ou une réunion d'événements élémentaires.*

Exemple

On lance un dé : $\Omega = \{1, 2, 3, 4, 5, 6\}$. On peut s'intéresser à l'événement A"On obtient un chiffre pair".
$A = \{2, 4, 6\}$ c'est un exemple d'évènement non nécessairement élémentaire.

Remarque : Dans ce cours Ω sera soit

- Un ensemble discret : $\Omega = \{\omega_1, \ldots, \omega_n\}$ contenant un nombre fini d'éléments ou bien $\Omega = \{\omega_1, \ldots, \omega_n, \ldots\}$ infini dénombrable.
- Un ensemble continu : Ω est un ensemble infinie non-dénombrable.

Il existe un vocabulaire propre aux évènements :

Notation	Vocabulaire ensembliste	Vocabulaire probabiliste
Ω	ensemble plein	évènement certain
$\varnothing$	ensemble vide	évènement impossible
ω	élément de Ω	évènement élémentaire
A	sous-ensemble de Ω	évènement
$\omega \in A$	ω appartient à A	ω réalise A
$A \subset B$	A inclus dans B	A implique B
$A \cup B$	réunion de A et B	Soit A est réalisé, soit B est réalisé, soit ils sont simultanément réalisés
$A \cap B$	intersection de A et B	A et B sont simultanément réalisés
A^c ou $\bar{A}$	complémentaire de A	A n'est pas réalisé
$A \cap B = \varnothing$	A et B disjoints	A et B sont incompatibles

Définition

Soit Ω un ensemble quelconque et soit $\mathcal{P}(\Omega)$ l'ensemble de ses parties. Une probabilité $\mathbb{P}$ sur $\mathcal{P}(\Omega)$ est une application à valeurs dans $[0, 1]$ vérifiant :

1. $\mathbb{P}(\Omega) = 1.$
2. *Si pour tout $n \in \mathbb{N}$, $\quad A_n \in \mathcal{P}(\Omega)$ et sont deux à deux incompatibles*

$$\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

Le triplet $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ est appelé espace probabilisé.

Proposition

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. $\mathbb{P}(\varnothing) = 0$.
3. $\mathbb{P}$ *est une fonction croissante i.e pour tout couple d'évènements* (A, B) *tel que* A *implique* B*, on a* $\mathbb{P}(A) \leq \mathbb{P}(B)$.
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

2)Probabilités conditionnelles et indépendance On notera $\mathbb{P}(A/B)$ la probabilité de A sachant B s'est réalisé.

Définition

Soit Ω *muni d'une probabilité* $\mathbb{P}$ *et* $A, B \subset \Omega$ *tel que* $\mathbb{P}(B) \neq 0$. *Alors la probabilité de* A *sachant* B *est définie par*

$$\mathbb{P}(A/B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Exemple

Parmi 10 pièces mécaniques, 4 sont défectueuses. On prend successivement deux pièces au hasard dans le lot (sans remise). Quelle est la probabilité pour que les deux pièces soient correctes.

Soient A "La première pièce est bonne" et B "La seconde pièce est bonne".

Comme il y a 6 pièces bonnes sur 10, $\mathbb{P}(A) = \frac{3}{5}$.

Lorsque l'on a retiré une pièce bonne, il reste 5 pièces bonne sur 9.

On a donc $\mathbb{P}(B/A) = \frac{5}{9}$.

On conclut que la probabilité cherchée est

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B/A) = \frac{3}{5} \times \frac{5}{9} = \frac{1}{3}.$$

Proposition

"Formule des probabilités composées"

1. *Si* A *et* B *sont deux événements tels que* $\mathbb{P}(B) \neq 0$ *alors*

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A/B).$$

2. *Si* $A_1, A_2, \ldots, A_n$ *sont* n *événements tels que* $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \neq 0$ *alors*

$$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2/A_1)\mathbb{P}(A_3/A_1 \cap A_2)\ldots\mathbb{P}(A_n/A_1 \cap A_2 \ldots \cap A_{n-1}).$$

Définition

Une famille $(A_i)_{i \in I}$ $(I \subset \mathbb{N})$ *d'évènements deux à deux incompatibles et tels que* $\bigcup_{i \in I} A_i = \Omega$ *est appelée système complet d'évènements.*

Proposition

"Formule des probabilités complètes (ou totales)"

Soit Ω muni d'une probabilité $\mathbb{P}$ et $(A_i)_{i\in I}$ $(I \subset \mathbb{N})$ un système complet d'évènements de probabilités non nulle. Alors, pour tout évènement B

$$\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(B \cap A_i) = \sum_{i\in I} \mathbb{P}(A_i)\mathbb{P}(B/A_i).$$

Un cas particulier :

Soit un évènement A tel que $\mathbb{P}(A) \neq 0$ et $\mathbb{P}(A^c) \neq 0$. Considérons le système complet d'évènements $\{A, A^c\}$ on obtient :

$$\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B/A) + \mathbb{P}(A^c)\mathbb{P}(B/A^c).$$

Application

Une population animale comporte 1/3 de mâles et 2/3 de femelles. L'albinisme frappe 6% des mâles et 0.36% des femelles. Déterminer la probabilté pour qu'un individu pris au hasard (dont on ignore le sexe) soit albinos.

Soient A" mâle" et $\bar{A}$ "femelle" constituent un système complet d'évènements.

Soit B "albinos" et $\bar{B}$ "non albinos".

D'après la fomule des probabilités totale,

$$P(B) = P(B/A)P(A) + P(B/\bar{A})P(\bar{A}).$$

Donc, $P(B) = 0.06 \times 1/3 + 0.0036 \times 2/3 = 0.0224.$

Proposition

"Formule de Bayes"
Soit Ω muni d'une probabilité $\mathbb{P}$ et $(A_i)_{i \in I} (I \subset \mathbb{N})$ un sytème complet d'évènements de probabilité non nulle. Alors, pour tout évènement B de probabilité non nulle on a $\forall j \in I$

$$\mathbb{P}(A_j/B) = \frac{\mathbb{P}(A_j)\mathbb{P}(B/A_j)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_j)\mathbb{P}(B/A_j)}{\sum_{i \in I} \mathbb{P}(A_i)\mathbb{P}(B/A_i)}.$$

Application

Dans une population pour laquelle un habitant sur cent est atteint d'une maladie génétique A, on a mis au point un test de dépistage. Le résultat du test est soit positif (T) soit négatif $(\bar{T})$. On sait que $\mathbb{P}(T/A) = 0.8$ et $\mathbb{P}(\bar{T}/\bar{A}) = 0.9$.
On soumet un patient au test. Celui-ci est positif. Quelle est la probabilité que ce patient soit attent de la maladie A soit $\mathbb{P}(A/T)$?

Application

Dans une population pour laquelle un habitant sur cent est atteint d'une maladie génétique A, on a mis au point un test de dépistage. Le résultat du test est soit positif (T) soit négatif ($\bar{T}$). On sait que $\mathbb{P}(T/A) = 0.8$ et $\mathbb{P}(\bar{T}/\bar{A}) = 0.9$.
On soumet un patient au test. Celui-ci est positif. Quelle est la probabilité que ce patient soit attent de la maladie A soit $\mathbb{P}(A/T)$?

D'après la formule de Bayes :

$$\mathbb{P}(A/T) = \frac{\mathbb{P}(A \cap T)}{\mathbb{P}(T)} = \frac{\mathbb{P}(T/A)\mathbb{P}(A)}{\mathbb{P}(T/A)\mathbb{P}(A) + \mathbb{P}(T/\bar{A})\mathbb{P}(\bar{A})}.$$

d'où

$$\mathbb{P}(A/T) = \frac{0.01 \times 0.8}{0.8 \times 0.01 + 0.1 \times 0.99} = 0.075.$$

Dire que l'évènement A est indépendant de B revient à dire que la réalisation de B ne modéfie pas $\mathbb{P}(A)$ c'est à dire qu'elle n'apporte aucune informations sur l'éventuelle réalisation de A.

Définition

Soit Ω muni d'une probabilité de $\mathbb{P}$ deux évènements A et B sont dits indépendants si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Remarque

L'indépendance de A et B se caractérise aussi par les relations

$$\mathbb{P}(A/B) = \mathbb{P}(A) \quad \text{ou} \quad \mathbb{P}(B/A) = \mathbb{P}(B).$$

Cette notion d'indépendance s'étend à plus de deux évènements.

Définition

Les évènements $A_i, i \in \{1, \ldots, m\}$ sont dits indépendants si

$$\forall I \subset \{1, \ldots, m\}, \qquad \mathbb{P}(\bigcap_{i \in I} A_i) = \prod_{i \in I} \mathbb{P}(A_i).$$

Attention : Il ne suffit pas que $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_m) = \prod_{i=1}^{m} \mathbb{P}(A_i)$ pour que les évènements soient indépendants.

Exemple :

Pour que trois évènements A_1, A_2 et A_3 soient indépendants, il ne suffit pas qu'ils soient deux à deux indépendants. L'indépendance se traduit par :

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2); \quad \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_3);$$

$$\mathbb{P}(A_2 \cap A_3) = \mathbb{P}(A_2)\mathbb{P}(A_3) \quad \text{et} \quad \mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3).$$

Fin

Merci pour votre attention

Variables aléatoires discrètes

Rania RAIS

Première année licence informatique de gestion

Décembre 2020

Définition

On appelle variable aléatoire discrète une application

$$X : \Omega \longrightarrow F \subset \mathbb{R}$$

où F est un ensembe fini ou infini dénombrable tel que pour tout $k \in F$, $\{\omega \in \Omega : X(\omega) = k\}$ est un évènement. On note l'évènement de façon concise $\{X = k\}$.

Example

On lance trois fois une pièce et on s'intéresse au nombre X de fois où la face Pile apparaît. A chaque événement élémentaire ω, on associe $X(\omega)$.

ω	PPP	PPF	PFP	FPP	FFP	FPF	PFF	FFF
$X(\omega)$	3	2	2	2	1	1	1	0

On observe que plusieurs événements donnent la méme valeur. On peut les regrouper et obtenir des événements qui correspondent à des valeurs distinctes de X.

k	3	2	1	0
$\{X=k\}$	$\{PPP\}$	$\{PPF, PFP, FPP\}$	$\{PFF, FPF, FFP\}$	$\{FFF\}$

Définition

Soit X une variable aléatoire discrète. On appelle loi de probabilité de X l'application $\mathbb{P}_X$ définie par :

$$\begin{array}{rccl} \mathbb{P}_X : & X(\Omega) & \longrightarrow & [0,1] \\ & k & \longmapsto & \mathbb{P}(X=k). \end{array}$$

Proposition

Si X est une variable aléatoire discrète dont la loi de probabilité est $\mathbb{P}_X$ alors $\mathbb{P}_X$ vérifie les deux propriétes suivantes :

1. $\forall k \in X(\Omega), \quad \mathbb{P}_X(k) \geq 0,$
2. $\displaystyle\sum_{k \in X(\Omega)} \mathbb{P}_X(k) = 1.$

Définition

On appelle fonction de répartition d'une variable aléatoire X la fonction

$$\begin{array}{rccc} F_X : & \mathbb{R} & \longrightarrow & [0,1] \\ & t & \longmapsto & \mathbb{P}(X \leq t). \end{array}$$

On dit que deux variables aléatoires X et Y ont la même loi si elles ont la même fonction de répartition $F_X = F_Y$.

Proposition

Soit F_X la fonction de répartition d'une variables aléatoire

1. F_X *est croissante sur* $\mathbb{R}$ *i.e* $\forall (a,b) \in \mathbb{R}^2, a \leq b \Rightarrow F_X(a) \leq F_X(b)$.
2. $\lim\limits_{t \to -\infty} F_X(t) = 0$ *et* $\lim\limits_{t \to +\infty} F_X(t) = 1$.
3. *Si* $a < b$ *alors* $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

Remarque : Si X est une variable aléatoire discrète dont les valeurs sont x_i avec $i = 1, 2, \cdots$ supposées rangées par ordre croissant alors la fonction de répartition F_X prend les valeurs :

$$F_X(t) = \begin{cases} 0 & \text{pour } t < x_1, \\ \mathbb{P}(X = x_1) & \text{pour } t \in [x_1, x_2[, \\ \vdots & \\ \mathbb{P}(X = x_1) + \mathbb{P}(X = x_2) + \cdots + \mathbb{P}(X = x_k) & \text{pour } t \in [x_k, x_{k+1}[, \\ \vdots & \end{cases}$$

La fonction de répartition F_X est constante par morceaux, ayant pour points de discontinuité $\{x_i, i \in \mathbb{N}\}$.

Exemple

On considére l'événement ω :"lancer 3 fois d'une piéce de monaie".
$X(\omega)$:="nombre de Piles de l'événement ω".

nombre de piles	$\mathbb{P}(X = k)$	F_X
0	$1/8$	$1/8$
1	$3/8$	$4/8$
2	$3/8$	$7/8$
3	$1/8$	1

Définition

Soit $X : \Omega \longrightarrow F \subset \mathbb{R}$ une variable aléatoire discrète. L'espérance de X est, si elle existe, notée $E(X)$ et définie par :

$$E(X) = \sum_{k \in F} k\,\mathbb{P}(X = k).$$

L'espérance représente la valeur moyenne prise par le variables X.

Proposition

1. *Pour toute constante c, $E(c) = c$.*
2. *L'espérance est une application linéaire, i.e pour toute variables X et Y telles que $E(X)$ et $E(Y)$ sont bien définies, pour toute $\alpha, \beta \in \mathbb{R}$. $E(\alpha X + \beta Y)$ est bien définie et on a*

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$

Proposition

1. *Soit* $X : \Omega \longrightarrow F \subset \mathbb{R}$ *une variable aléatoire discréte réelle et* $f : \mathbb{R} \longrightarrow \mathbb{R}$ *la variable* $f(X)$ *est une v.a.d. Alors* $E(f(X))$ *est bien définie ssi*

$$\sum_{k \in F} \mid f(k) \mid \mathbb{P}(X = k) < +\infty$$

et dans ce cas on a

$$E(f(X)) = \sum_{k \in F} f(k)\mathbb{P}(X = k)$$

Définition

On dit que la v.a.d X admet une variance si la v.a.d $(X-E(X))^2$ admet une espérance. On note alors

$$V(X)=E(X-E(X))^2=\sum_{k\in F}(k-E(X))^2\mathbb{P}(X=k)=E(X^2)-E(X)^2.$$

L'écart type de X noté par σ_X est la racine carrée de sa variance.

L'écart-type (ou la variance) mesure la dispersion de la v.a X autour de sa valeur moyenne $E(X)$.

Proposition

1. *La variance d'une variable aléatoire constante est nulle.*
2. *Soient $\alpha,\beta\in\mathbb{R}$, on a $V(\alpha X+\beta)=\alpha^2V(X)$.*
3. *Soient X et Y deux v.a.* $V(X+Y)=V(X)+V(Y)+2\mathrm{cov(X,Y)}$, *avec* $\mathrm{cov(X,Y)=E(XY)-E(X)E(Y)}$.

Remarque

Deux variables qui ont méme loi ont méme espérance et méme variance.

Exemple : On reprend l'exemple précédent.

$$E(X) = \sum_{k=0}^{3} k\,\mathbb{P}(X = k) =$$

$$V(X) = \sum_{k=0}^{3} k^2\mathbb{P}(X = k) - E(X)^2 =$$

$$\sigma_X =$$

Remarque

Deux variables qui ont méme loi ont méme espérance et méme variance.

Exemple : On reprend l'exemple précédent.

$$E(X) = \sum_{k=0}^{3} k\,\mathbb{P}(X = k) = 3/2.$$

$$V(X) = \sum_{k=0}^{3} k^2\mathbb{P}(X = k) - E(X)^2 = 3/4.$$

$$\sigma_X = \frac{\sqrt{3}}{2}.$$

Définition

On appelle moment non centré d'ordre $r \in \mathbb{N}^*$ la quantité, lorsqu'elle existe :

$$m_r(X) = E(X^r).$$

Le moment centré d'ordre $r \in \mathbb{N}^*$ est :

$$\mu_r(X) = E[X - E(X)]^r.$$

Exemple : On reprend l'exemple précédent.

$m_2(X) =$

$\mu_2(X) =$

Définition

On appelle moment non centré d'ordre $r \in \mathbb{N}^*$ la quantité, lorsqu'elle existe :

$$m_r(X) = E(X^r).$$

Le moment centré d'ordre $r \in \mathbb{N}^*$ est :

$$\mu_r(X) = E[X - E(X)]^r.$$

Exemple : On reprend l'exemple précédent.

$$m_2(X) = E(X^2) = \sum_{k=0}^{3} k^2 \, \mathbb{P}(X = k) = 3.$$

$$\mu_2(X) = E[X - E(X)]^2 = V(X) = 3/4.$$

II)Lois discrétes usuelles

1)Loi de Bernoulli : $\mathcal{B}(1, p); \quad p \in]0, 1[$

- **Définition**
  On dit X suit la loi de Bernoulli de parmètre p (on note $X \hookrightarrow \mathcal{B}(1, p)$) si elle est à valeurs dans {0,1} avec $\mathbb{P}(X = 1) = p$ et $\mathbb{P}(X = 0) = 1 - p$. Cette loi modélise l'issue d'une expérience en ne s'intéressant qu'au "succés" ou à l' "echec" de l'expérience.

- **Exemple** :
  On considère l'expérience aléatoire suivante : jet d'une pièce de monaie. On associe à cette expérience la variable aléatoire X qui prend 1 si le résultat est "Pile" et 0 sinon.

- **Proposition** :
  Soit X une variable aléatoire telle que $X \hookrightarrow \mathcal{B}(1, p)$, on a :

$$E(X) = p \quad \text{et} \quad V(X) = p(1 - p).$$

2)Loi binomiale : $\mathcal{B}(n,p); \quad n \in \mathbb{N}^*, \quad p \in]0,1[.$

- **Définition**

  Soit X la variable aléatoire qui représente le nombre de succés obtenus lors des n épreuves d'un schéma de Bernoulli. Alors on dit que X suit la loi binomiale de paramètre (n,p) (on note $X \hookrightarrow \mathcal{B}(n,p)$) si elle est à valeurs dans $\{0,1,\cdots,n\}$ avec

  $$\forall i = 0,\cdots,n \qquad \mathbb{P}(X=i) = C_n^i\, p^i(1-p)^{n-i}.$$

  "La probabilité d'obtenir i succés au cours des n répétitions"

- **Proposition** : Si X est une variable aléatoire de loi $\mathcal{B}(n,p)$ alors

  $$E(X) = np \quad \text{et} \quad V(X) = np(1-p).$$

3)Loi uniforme discrète

- **Définition**
  On dit que X suit une loi uniforme discrète sur un ensemble fini de cardinal N : $\{1, \cdots, N\}$ (on note $X \hookrightarrow \mathcal{U}_N$) si elle est à valeurs dans $\{1, \cdots, N\}$ avec

$$\forall i \in \{1, \cdots, N\}, \qquad \mathbb{P}(X = i) = \frac{1}{N}.$$

  Cette loi modélise l'issue d'une expérience où les résultats sont équiprobables.

- **Exemple** : La distribution des chiffres obtenus au lancer d'un dé (non pipé) suit une loi uniforme.

- **Proposition** : Si X est une variable aléatoire de loi $\mathcal{U}_N$ alors

$$E(X) = \frac{N+1}{2} \quad \text{et} \quad V(X) = \frac{N^2-1}{12}.$$

4)Loi géométrique : $\mathcal{G}(p); \quad p \in]0,1[$

- **Définition**
  On dit que la variable aléatoire X suit une loi géométrique de paramètre p (on note $X \hookrightarrow \mathcal{G}(p)$) si elle est à valeurs dans $\mathbb{N}^*$ avec

$$\forall i \in \mathbb{N}^*, \quad \mathbb{P}(X = i) = p(1-p)^{i-1}.$$

  La loi géométrique est la loi du premier succés.

- **Exemple** :
  On lance une pièce truquée jusqu'à ce qu'on obtienne "Pile". On associe à cette expérience la variable aléatoire X correspond au premier jet dans lequel le résultat de l'expérience donne "Pile".

- **Proposition** : Soit $X \hookrightarrow \mathcal{G}(p)$, on a

$$E(X) = \frac{1}{p} \quad \text{et} \quad V(X) = \frac{1-p}{p^2}.$$

5)Loi de Poisson : $\mathcal{P}(\lambda); \quad \lambda > 0$

- **Définition**
  On dit que la variable aléatoire X suit une loi de Poisson de paramètre λ (on note $X \hookrightarrow \mathcal{P}(\lambda)$) si elle est à valeurs dans $\mathbb{N}$ avec

$$\forall i \in \mathbb{N}, \quad \mathbb{P}(X = i) = \mathrm{e}^{-\lambda} \frac{\lambda^i}{i!}.$$

- **Exemple** : La loi de Poisson est utilisée pour modéliser le comptage d'un évènement rare c'est à dire des évènements ayant une faible probabilité de réalisation : maladie rare, accident, pannes,...

- **Proposition** : Soit X une variable aléatoire qui suit la loi de Poisson de paramètre $\lambda > 0$ on a :

$$E(X) = V(X) = \lambda.$$

Merci pour votre attention