Introduction aux statistiques et probabilités

Vincent Denoël

MATH0067-1

Last update : 15 septembre 2023

Informations Générales

Où télécharger les mises à jours de ce document ?

`http://hdl.handle.net/2268/129786`

(cliquez sur l'URL ci-dessus et acceptez la connexion au site ORBi)

Où ? Quand ?

Le cours est organisé le mardi, de 8h30 à 12h30 (1er quadri), horaire : voir CELCAT

Me contacter ?

Bureau : B52/3, +1/422
Téléphone : 04/366.29.30
Mail : v.denoel@ulg.ac.be

Objectifs du cours et plan pédagogique ? Examen ?

Cours 15h+15h, valorisation 2 ECTS (plus de détails dans les engagements pédagogiques du cours)

- pas d'exercice personnel, évaluation "bonus" pendant les cours
- Examen écrit portant sur les notions théoriques et pratiques

Informations Générales

Où télécharger les mises à jours de ce document ?

`http://hdl.handle.net/2268/129786`

(cliquez sur l'URL ci-dessus et acceptez la connexion au site ORBi)

Où ? Quand ?

Le cours est organisé le mardi, de 8h30 à 12h30 (1er quadri), horaire : voir CELCAT

Me contacter ?

Bureau : B52/3, +1/422
Téléphone : 04/366.29.30
Mail : v.denoel@ulg.ac.be

Objectifs du cours et plan pédagogique ? Examen ?

Cours 15h+15h, valorisation 2 ECTS (plus de détails dans les engagements pédagogiques du cours)

- pas d'exercice personnel, évaluation "bonus" pendant les cours
- Examen écrit portant sur les notions théoriques et pratiques

Informations Générales

Où télécharger les mises à jours de ce document ?

`http://hdl.handle.net/2268/129786`
(cliquez sur l'URL ci-dessus et acceptez la connexion au site ORBi)

Où ? Quand ?

Le cours est organisé le mardi, de 8h30 à 12h30 (1er quadri), horaire : voir CELCAT

Me contacter ?

Bureau : B52/3, +1/422
Téléphone : 04/366.29.30
Mail : v.denoel@ulg.ac.be

Objectifs du cours et plan pédagogique ? Examen ?

Cours 15h+15h, valorisation 2 ECTS (plus de détails dans les engagements pédagogiques du cours)

- pas d'exercice personnel, évaluation "bonus" pendant les cours
- Examen écrit portant sur les notions théoriques et pratiques

Informations Générales

Où télécharger les mises à jours de ce document ?

`http://hdl.handle.net/2268/129786`
(cliquez sur l'URL ci-dessus et acceptez la connexion au site ORBi)

Où ? Quand ?

Le cours est organisé le mardi, de 8h30 à 12h30 (1er quadri), horaire : voir CELCAT

Me contacter ?

Bureau : B52/3, +1/422
Téléphone : 04/366.29.30
Mail : v.denoel@ulg.ac.be

Objectifs du cours et plan pédagogique ? Examen ?

Cours 15h+15h, valorisation 2 ECTS (plus de détails dans les engagements pédagogiques du cours)

- pas d'exercice personnel, évaluation "bonus" pendant les cours
- Examen écrit portant sur les notions théoriques et pratiques

Informations Générales

Où télécharger les mises à jours de ce document ?

`http://hdl.handle.net/2268/129786`
(cliquez sur l'URL ci-dessus et acceptez la connexion au site ORBi)

Où ? Quand ?

Le cours est organisé le mardi, de 8h30 à 12h30 (1er quadri), horaire : voir CELCAT

Me contacter ?

Bureau : B52/3, +1/422
Téléphone : 04/366.29.30
Mail : v.denoel@ulg.ac.be

Objectifs du cours et plan pédagogique ? Examen ?

Cours 15h+15h, valorisation 2 ECTS (plus de détails dans les engagements pédagogiques du cours)

- pas d'exercice personnel, évaluation "bonus" pendant les cours
- Examen écrit portant sur les notions théoriques et pratiques

Motivations

Pourquoi un cours de stats/probas ?

Cours
Urban planning & transportation (Master II), Travail de fin d'Etudes (Master II)

Débouchés de la formation :
Bureau d'Architecture (statistiques descriptives, interprétation de données, ...)
Administration (statistiques, enquêtes, recensements, ...)
Recherches (expériences, tests d'hypothèses,...)

Acquis et compétences :
Connaissance des outils statistiques
+ éléments listés ci-dessus

Organisation Du Cours

Le cours est calqué sur le bouquin de référence :
Statistique théorique et appliquée, Tome 1, Pierre Dagnelie, Ed. De Boeck.

objectifs : favoriser l'apprentissage autodidacte, la consultation de références externes

Table des matières :

- Collecte des données
- Statistique descriptive
- Lois de probabilité
- Statistique inférentielle

Les travaux pratiques se feront "on-the-fly", par groupes de deux étudiants ($\rightarrow$ venir avec un ordinateur portable personnel).

Organisation Du Cours

Le cours est calqué sur le bouquin de référence :
Statistique théorique et appliquée, Tome 1, Pierre Dagnelie, Ed. De Boeck.

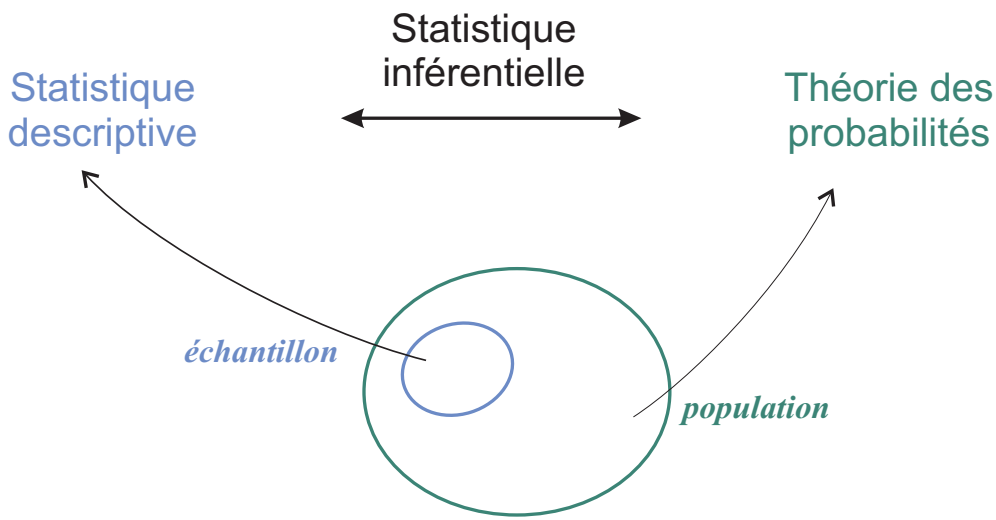
objectifs : favoriser l'apprentissage autodidacte, la consultation de références externes

Table des matières :

- Collecte des données
- Statistique descriptive
- Lois de probabilité
- Statistique inférentielle

Les travaux pratiques se feront "on-the-fly", par groupes de deux étudiants ($\rightarrow$ venir avec un ordinateur portable personnel).

Organisation Du Cours



Programme Prévisionnel

	8:30-12:30	
1	19/09/23	Collecte données - Stat descriptives 1-D (jusque stats 1-D, moyenne)
2	26/09/23	Stat descriptives 1-D (dispersion & suite)
2	3/10/23	Stat descriptives 2-D (régression linéaire + exercices)
3	10/10/23	Loi de probabilité - Axiomes - Bayes
4	17/10/23	Loi de probabilité - PDF CDF - exercices
5	24/10/23	Loi de probabilité - exercices
-	31/10/23	Toussaint - respiration
6	7/11/23	Inférence statistique
7	14/11/23	Inférence statistique
8	21/11/23	-
9	28/11/23	-
10	5/12/23	-
11	12/12/23	-
12	19/12/23	-

Pas d'exercice personnel. Tests possibles en fin de séance, avec valorisation non pénalisante.

Partie I : Collecte des Données

Acquis d'apprentissage :

- comprendre les différences entre enquête et expérimentation, et leurs spécificités
- maitriser le vocabulaire dédicacé
- mettre au point rigoureusement un recensement, une enquête, une expérience

Généralités I

Enquête : ensemble des opérations qui ont pour but de collecter de façon organisée des informations relatives à un groupe *d'individus* ou *d'éléments* observés dans leur milieu ou dans leur cadre habituel.

Individus, Eléments : personnes, animaux, plantes (ou groupes de personnes, animaux, plantes), entreprises, machines,...

Population : ensemble (complet !) des unités auxquelles on s'intéresse.

- si l'enquête touche l'ensemble de la population : enquête *complète, exhaustive* $\rightarrow$ *recensement*
- si l'enquête touche une partie de la population : enquête *partielle* ou enquête *par échantillonnage* $\rightarrow$ *sondage*

Échantillon : partie de la population qui est réellement étudiée (! choix de cette fraction de la population)

Généralités II

▷ Avant une enquête, il est important de bien définir les *individus* et la *population* → délimitation de l'enquête (buts & objectifs)

Exemple : recensement national de population humaine (étude individuelle de chacun des groupes de personnes qui vivent en commun, dans un même logement). Q : partir de notion de famille ? De ménage ? Définir ces notions. Quid des militaires ? Couvents ? Quid des étrangers de passage pour une courte durée ? Des autochtones à l'étranger ?

→ important de bien délimiter l'enquête (au cas par cas)
nb : important de bien maîtriser tous les éléments de l'enquête (reproductibilité)

Généralités III

▷ Avant une enquête, il est important de bien définir les *observations* et la *méthode de collecte des données* (moyens)

Observations qualitatives (état civil, profession) : définir tous les termes utilisés
Observations quantitatives (nombre de pièces d'une maison, superficie d'une annexe) : définir tous les termes utilisés et le mode de détermination des valeurs numériques données (protocole de mesures)

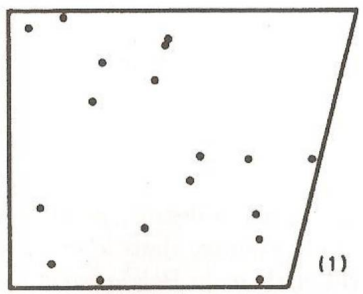
Important de fixer la date d'observation ou la période couverte par l'enquête (surtout lors d'observation de progression de maladie/pandémie)

Envoi d'un questionnaire par la poste, d'enquêteurs, ...
nb : utilité d'une pré-enquête, ou enquête-pilote

Méthodes Échantillonnage I

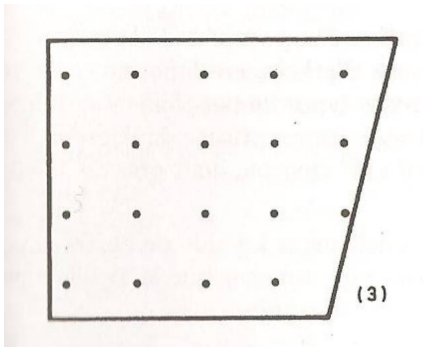
1. ***Échantillonnage (complètement) aléatoire :*** Choisir une à une et indépendamment les unes des autres chacune des unités qui seront observées en donnant à chacune des chances égales d'être choisies.

exemple : numéroter les unités puis tirer au hasard



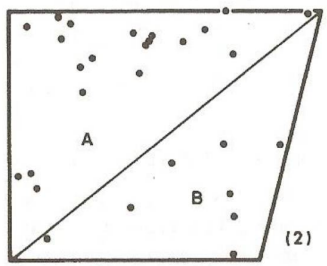
Méthodes Échantillonnage II

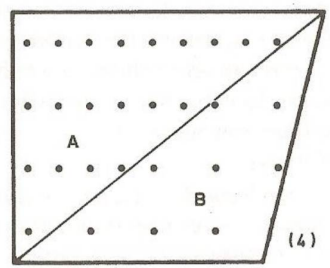
2. ***Échantillonnage systématique :*** choisir une première unité au hasard, puis les autres de façon systématique, selon une règle choisie au préalable.
exemple : dans une liste de personnes, choisir un nom au hasard, puis une personne sur 20.
→ facilement applicable en 2-D (sélection de parcelles représentatives)
→ plus simple à mettre en œuvre que l'échantillonnage aléatoire, demande également un échantillon de plus petite taille à qualité de résultat égale
→ attention lorsqu'il existe une périodicité dans l'arrangement initial des unités



Méthodes Échantillonnage III

3. ***Échantillonnage stratifié :*** subdiviser la population en *strates*, puis sélection des individus dans chaque strate indépendamment les unes des autres. La sélection dans chaque strate se fait soit par échantillonnage complètement aléatoire, soit par échantillonnage systématique.
→ idéale lorsque la population est hétérogène et que l'on désire représenter chacune des composantes
→ idéalement, définir les strates de façon à ce qu'elles soient toutes aussi homogènes que possible
exemple : échantillonnage dans une commune, avec différentes catégories socio-professionnelles (! les strates ne sont pas choisies en fonction des catégories socio-professionnelles - pas besoin donc de les identifier).





Méthodes Échantillonnage IV

4. ***Échantillonnage à deux ou plusieurs degrés - Échantillonnage en grappes :*** identifier plusieurs types d'unités statistiques correspondant aux deux ou plusieurs degrés. Échantillonner (de façon arbitraire ou systématique) à chaque niveau, en arborescence.
exemple : échantillonnage de commune, puis de rue, puis de maison.
exemple : caractérisation de la résistance d'un sol sur chantier – unité de premier degré : échantillon de sol – unité de second degré : type d'essai (écrasement, triaxial, plaque...)
→ avantage en cas d'enquête géographique : cela limite les déplacements
→ perte de précision ; pour une même taille d'échantillon, l'échantillonnage par grappes est moins précis qu'un échantillonnage complètement aléatoire

Méthodes Échantillonnage V

5. ***Échantillonnage par la méthode des quotas*** : donner une composition aussi semblable que possible à celle de la population, en fonction de quelques critères de classification considérés a priori comme particulièrement importants, mais sans définir de façon précise la manière dont les individus devront être choisis à l'intérieur de chacune des classes.
exemple : choisir 15 ouvrières âgées de 20 à 30 ans en Belgique (laisser le choix aux enquêteurs)
→ utilisé largement dans les sondages d'opinion
→ plus rapide et plus facile que l'échantillonnage complètement aléatoire
→ erreurs possibles liées au choix des enquêteurs

La réalisation d'un échantillonnage ne peut se faire valablement que si on possède au départ, pour l'ensemble de la population, un minimum d'informations qui constituent la ***base d'échantillonnage*** ou ***base de sondage*** (listes, répertoires, documents cartographiques, photos aériennes, etc)

Important que ces documents soient de qualité (à jour, sans répétition,...) : la qualité de l'échantillonnage dépend de la qualité des documents qui ont servi de base à l'échantillonnage.

nb : la base de sondage ne doit pas nécessairement s'étendre à toute la population ; par exemple, dans le cas d'un sondage à deux degrés, on imagine une liste complète de toutes les communes (unité de premier degré), et pour les communes choisies uniquement, une description des unités de second degré.

Taille d'Échantillon I

Elle est soit fixée en valeur absolue (nombre d'unités observées) ou en valeur relative (fraction du nombre total d'individus dans la population).
En valeur absolue → ***taille*** ou ***effectif de l'échantillon***
En valeur relative → ***intensité d'échantillonnage, intensité de sondage***

La précision des résultats obtenus à l'issue d'une enquête par échantillonnage dépend à la fois de l'importance de l'échantillonnage et du caractère plus ou moins homogène ou hétérogène de la population (la précision est d'autant meilleure que l'échantillon est de taille importante et que la population est plus homogène).

nb : à rediscuter après le chapitre sur les statistiques inférentielles, car la taille d'échantillon infuence fortement l'écart-type sur l'estimation d'une propriété de population - cf. test d'hypothèses.

Généralités I

Expérimentation : réalisation d'une ou plusieurs expériences ou d'un ou plusieurs essais où l'on suppose que l'apparition des faits que l'on désire étudier est volontairement provoquée, dans des conditions maitrisées, au moins partiellement.

L'expérimentation est souvent plus efficace que l'observation par l'enquête → favoriser dans la mesure du possible l'expérimentation lorsque c'est réalisable.

Toute expérience doit être l'objet d'une préparation et d'une planification minutieuse.

Le ***plan d'expérience*** ou ***protocole expérimental*** regroupe l'ensemble des questions à étudier : (i) définition du but et des conditions d'expérimentation, (ii) définition des facteurs que l'on désire étudier, (iii) définition des unités expérimentales, (iv) définitions des observations à réaliser, (v) définition du dispositif expérimental.

Généralités II

▷ **définition du but et des conditions d'expérimentation**
⇜⇝ définition de la population-parent dans le cas d'une enquête par échantillonnage
Objectif : que les résultats de l'expérience puissent s'appliquer à un ensemble plus vaste que simplement celui des échantillons testés
Important donc de bien définir la population ciblée, et que les échantillons choisis en soient un exemple représentatif

Importance de la variabilité spatiale et temporelle (discuter selon que l'on veut la mettre en évidence ou non)
Importance des échelles de temps et de longueur (ex : statistiques météo)

Généralités III

▷ **définition des *facteurs* = caractéristiques propres de l'expérience (new !)**
nb : facteurs qualitatifs .vs. facteurs quantitatifs
Modalités (variantes ou niveaux) : éléments individuels qui constituent un facteur (ex : différentes variétés de différents produits phytosanitaires, différentes doses d'engrais, etc.)

Facteurs qualitatifs : généralement les variantes sont définies a priori, en même temps que le but de l'expérience
Facteurs quantitatifs : généralement les niveaux sont choisis de façon à définir une progression arithmétique ou géométrique

Dans le cas d'une expérience à plusieurs facteurs, associer chaque facteur à chaque autre → expérience organisée de façon factorielle

Généralités IV

▷ **définition des unités expérimentales**
⇜⇝ définition de l'unité de base dans le cas d'une enquête par échantillonnage

L'unité peut être naturelle (arbre, animal), artificielle (parcelle de terrain). Dans ce cas, il est important de définir adéquatement la surface/la forme des unités à tester.
Nombre de répétitions : nombre de fois que chaque unité est testée

(▷ définitions des observations à réaliser) comme pour l'enquête par échantillonnage

▷ définition du ***dispositif expérimental*** : il détermine la façon dont les objets seront associés aux unités de base
1. répartir les objets tout à fait au hasard parmi les unités expérimentales → dispositif complètement aléatoire
2. réunir les unités expérimentales en groupes aussi homogènes que possible et répartir les objets au hasard à l'intérieur de ces groupes

Généralités V

(***blocs***) → dispositif à blocs aléatoires complets (si chaque bloc constitue une répétition aléatoire complète)
3. il existe encore des méthodes différentes : hypercube latin, dispositifs en blocs incomplets, en parcelles divisées

Exemples I

1. Exemple d'expérience à deux facteurs sur céréales

▷ Les différents types de données

a. Observations quantitatives

- ***Dénombrements*** ou ***comptages*** : nombre entier non négatif → variables discontinues et discrètes
- ***Mesures*** ou ***mensurations*** : variables continues (attention aux unités et à la précision)

b. Observations qualitatives

Elles concernent des caractères ou des attributs que possèdent les individus

nb : souvent codé sous forme numérique comme des variables quantitatives discontinues.

ex : 0 ou 1 dans le cas d'une variable ***binaire***.

Prudence dans l'interprétation de l'échelle de valeurs et dans l'utilisation d'opérations mathématiques !

Généralités II

c. Autres types de données

- ***Rang*** : numéro d'ordre des individus classés selon l'ordre croissant de la caractéristique considérée

ex : tests de dégustation, classement par expert, etc.

→ les valeurs ne sont pas indépendantes les unes des autres.

- ***Données directionnelles ou circulaires*** : angle(s), avec la particularité que $0° \equiv 360°$

ex : direction du vent

Généralités III

▷ **L'enregistrement et le traitement des données**

a. Sous forme manuscrite : sur un carnet, calepin
! soin pour éviter les erreurs de relecture et transcription
si besoin utiliser des ***feuilles de pointage***

b. Traitement des données : à l'aide de logiciel d'analyse comme R, Statistica, Excel, Matlab
vérifier l'introduction des données (double encodage)

c. Enregistrement direct des données (sur PC portable, iPad, ...)

Exemples et Exercices I

Imaginez une étude statistique par enquête.

Par écrit, décrivez le but et les conditions de l'enquête, les unités d'observation, le mode d'échantillonnage, la taille des échantillons, la nature des échantillons, le mode de collecte des données,...
15-20 minutes

Imaginez une étude par expérimentation.

Par écrit, décrivez le but et les conditions de l'expérience, le plan d'expérience, les facteurs, le protocole expérimental, les unités, le dispositif expérimental...
15-20 minutes

Partie II : Statistique Descriptive

Acquis d'apprentissage :

- connaitre les différents indicateurs de position et de distribution
- maitriser notion de corrélation et d'indépendance statistique
- synthétiser des séries de données
- utiliser la régression linéaire (au sens des moindres carrés) et non linéaire

Généralités I

La statistique descriptive a pour but de présenter les données observées sous une forme telle que l'on puisse en prendre connaissance facilement. Elle peut concerner une, deux ou plusieurs variables

A une dimension, il y a trois formes condensées distinctes :

- un tableau statistique indiquant les distributions de fréquences
- paramètres scalaires visant à la réduction de données
- un diagramme des fréquences

Distributions de fréquences I

▷ Les séries de données

Présenter les données par énumération

$$x_1,\ x_2,\ \ldots,\ x_n$$

ou par ordre croissant

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

$n =$ ***effectif***, nombre total d'échantillons

Exemple 1 :
Nombre de personnes actives par secteur statistique à Liège (Age-2001.xls) :
Saint-Lambert : 508
Feronstrée : 584
Pierreuse : 207
Saint-Jean : 441
...ou de façon compacte...

508, 584, 207, 441, 629, 324, ... (179 valeurs)

Exemple 2 :
Nombre de personnes actives de moins de 20 ans par secteur statistique à Liège :
1, 1, 0, 3, 3, 2, ... (179 valeurs)

Distributions de fréquences II

▷ **Les distributions non groupées** (pour variables discrètes)
Lorsque les distributions sont nombreuses, il est utile ou nécessaire de grouper les données sous forme d'une *distribution de fréquence*.
Fréquence absolue : nombre d'occurrences d'une même valeur observée
Distribution de fréquences : couples des différentes valeurs observées x_1, x_2, ..., x_p et des fréquences correspondantes n_1, n_2, ..., n_p

$$\sum_{i=1}^{p} n_i = n$$

Fréquence relative n_i' : nombre d'occurrences d'une même valeur observée rapporté à l'effectif

$$n_i' = \frac{n_i}{n} \qquad \rightarrow \qquad \sum_{i=1}^{p} n_i' = 1$$

Distribution de fréquences cumulées N_i' (d'une valeur observée x_i) : somme des fréquences relatives correspondant à cette valeur et à l'ensemble des valeurs inférieures. nb : $N_p' = 1$.

Distributions de fréquences III

Exemple 2 :
Nombre de personnes actives de moins de 20 ans par secteur statistique à Liège (Age-2001.xls) :
1, 1, 0, 3, 3, 2, ... (179 valeurs)

Distributions de fréquences IV

Nb de personnes x_i	fréq. abs. n_i	fréq. rel. n_i'	fréq. abs. cum	fréq. rel. cum $N'(x_i)$
0	42	0.2346	42	0.2346
1	36	0.2011	78	0.4358
2	18	0.1006	96	0.5363
3	18	0.1006	114	0.6369
4	15	0.0838	129	0.7207
5	15	0.0838	144	0.8045
6	14	0.0782	158	0.8827
7	9	0.0503	167	0.9330
8	2	0.0112	169	0.9441
⋮	⋮	⋮	⋮	⋮
17	0	0.0000	177	0.9888
18	1	0.0056	178	0.9944
19	1	0.0056	179	1.0000

Distributions de fréquences V

▷ **Les distributions groupées**

Lorsque l'effectif est encore plus important, on peut condenser les tableaux statistiques en groupant les observations en ***classes*** ou ***catégories***. Elles sont définies par leurs limites. (nb : intervalle de classe et nombre de classes à choisir adéquatement).

Fréquence de classe relative n'_i : nombre d'observations d'une même valeur observée rapporté à l'effectif

nb : lorsque les intervalles de classe sont identiques, on peut représenter les classes par leurs points centraux.

nb : toujours utiliser des distributions groupées en présence de variables continues

Exemple 1 :
Nombre de personnes actives par secteur statistique à Liège (Age-2001.xls) :

508, 584, 207, 441, 629, 324, ... (179 valeurs)

Distributions de fréquences VI

Nb de personnes x_i	fréq. abs. n_i	fréq. rel. n'_i	fréq. abs. cum	fréq. rel. cum $N'(x_i)$
0-200	78	0.4358	78	0.4358
201-400	46	0.2570	124	0.6927
401-600	29	0.1620	153	0.8547
601-800	19	0.1061	172	0.9609
801-1000	4	0.0223	176	0.9832
1001-...	3	0.0168	179	1.0000

Matlab : Utiliser les fonctions `cumsum(x)` et `find(x,k)` pour réaliser les tableaux de distributions groupées.

```
listk=200:200:1000;
for i=1:length(listk);
  k=listk(i);
  disp ([k,length(find(data<=k))]);
end
```

Réduction des données, Paramètres de position I

▷ **Moyenne (arithmétique)**

La ***moyenne arithmétique*** est la somme des valeurs observées divisée par le nombre d'observations

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{ou} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{p} (n_i x_i)$$

où les x_i représentent les valeurs observées (données brutes), ou les centres des classes (distributions groupées) et les fréquences correspondantes.

nb : pour les distributions groupées, on commet une petite erreur en remplaçant chaque valeur par le point central de la classe correspondante.

Propriétés de la moyenne arithmétique

a. La somme des écarts $x_i - \overline{x}$ entre les valeurs observées et la moyenne est nulle

b. C'est par rapport à la moyenne que la somme des carrés des écarts est la plus petite

Réduction des données, Paramètres de position II

c. Le calcul de moyenne passe à travers une opération de transformation linéaire

$$x'_i = a + bx_i \quad \Rightarrow \quad \overline{x'} = a + b\overline{x}$$

la moyenne est donc affectée par un changement d'origine et d'unités.

Matlab : `mean(x)`

Réduction des données, Paramètres de position III

▷ Moyenne géométrique

$$\overline{x}_g = \sqrt[n]{x_1 \ldots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

a. La moyenne géométrique est toujours inférieure à la moyenne arithmétique

b. Elles sont égales lorsque toutes les valeurs observées sont égales entre elles

Matlab : `geomean(x)`

Réduction des données, Paramètres de position IV

▷ Moyenne harmonique

$$\overline{x}_h = \frac{n}{\frac{1}{x_1} + \ldots \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} (1/x_i)} \quad \text{ou} \quad \overline{x}_h = \frac{n}{\sum_{i=1}^{p} (n_i/x_i)}$$

a. La moyenne harmonique est toujours inférieure à la moyenne géométrique

b. Elles sont égales lorsque toutes les valeurs observées sont égales entre elles

Matlab : `harmmean(x)`

Réduction des données, Paramètres de position V

▷ Moyenne quadratique

$$\overline{x}_q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \quad \text{ou} \quad \overline{x}_q = \sqrt{\frac{1}{n}\sum_{i=1}^{p} \left(n_i x_i^2\right)}$$

a. La moyenne quadratique est toujours supérieure à la moyenne arithmétique

b. Elles sont égales lorsque toutes les valeurs observées sont égales entre elles

Matlab : `sqrt(mean(x.^2))`

Réduction des données, Paramètres de position VI

▷ Médiane

La ***médiane*** est un paramètre de position tel qu'une moitié des observations lui sont inférieures (ou égales) et l'autre moitié lui sont supérieures (ou égales).

◦ Pour les séries statistiques ordonnées et pour les distributions non groupées,

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{si } n \text{ est impair} \\ \frac{x_{n/2}+x_{n/2+1}}{2} & \text{si } n \text{ est pair} \end{cases}$$

◦ Pour les distributions non groupées, on peut également chercher le point d'ordonnée $n/2$ (ou $1/2$) sur le polygone des fréquences cumulées.

◦ Pour les distributions groupées, ***la classe médiane*** est celle qui contient la médiane

a. Dans le cas de distributions symétriques, la moyenne est confondue avec la médiane

Réduction des données, Paramètres de position VII

b. Souvent la médiane est inférieure à la moyenne pour une distribution asymétrique gauche, et supérieure à la moyenne pour une distribution asymétrique droite

Matlab : `median(x)`

Réduction des données, Paramètres de position VIII

▷ Mode

Le ***mode*** d'une distribution non groupée est la valeur observée de fréquence maximale.
La ***classe modale*** d'une distribution groupée est la classe fréquence maximale

→ on distingue les distributions ***unimodales*** et ***plurimodales***

Matlab : `mode(x)`

Exemple 1 :
Nombre de personnes actives par secteur statistique à Liège (Age-2001.xls) :
508, 584, 207, 441, 629, 324, ... (179 valeurs)
moyenne (artihmétique) : mean(data)=306.79
moyenne géométrique : geomean(data)=166.02 (attention aux valeurs nulles)
minimum : min(data)=0
maximum : max(data)=1716

Propriété d'ordre sur les moyennes

$$min \leq \overline{x}_h \leq \overline{x}_g \leq \overline{x} \leq \overline{x}_q \leq max$$

Réduction des données, Paramètres de dispersion I

▷ **Variance, écart-type, coefficient de variation**

La ***variance*** s^2 d'une série statistique est la moyenne arithmétique des carrés des écarts par rapport à la moyenne

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{ou} \quad s^2 = \frac{1}{n}\sum_{i=1}^{p}\left[n_i\,(x_i - \overline{x})^2\right]$$

Le ***carré moyen*** a_2 d'une série est défini par

$$a_2 = \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 \quad \text{ou} \quad a_2 = \frac{1}{n}\sum_{i=1}^{p}\left[n_i x_i^2\right]$$

Propriété importante : $s^2 = \overline{x^2} - \overline{x}^2$

MATLAB : `var(x)`

[nb : on divise parfois par $n-1$ au lieu de n]

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

MATLAB : `std(x)`

Réduction des données, Paramètres de dispersion II

Le ***coefficient de variation*** cv est le rapport entre l'écart-type et la moyenne (si elle est strictement positive)

$$cv = \frac{s}{\overline{x}}$$

Réduction des données, Paramètres de dispersion III

a. l'écart-type est nul *ssi* tous les écarts par rapport à la moyenne sont nuls

b. L'écart-type est affecté par les changements d'unités, mais est indépendant d'un changement d'origine. En effet

$$x_i' = a + bx_i \quad \Rightarrow \quad s_{x'}^2 = b^2 s_x^2$$

c. Le coefficient de variation est indépendant d'un changement d'unités. En effet

$$x_i' = bx_i \quad \Rightarrow \quad cv_{x'} = cv_x$$

d. Propriété de minimum de la variance : on remarque que l'expression suivante

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

est minimale lorsque $c = \overline{x}$.

Réduction des données, Paramètres de dispersion IV

▷ Écart interquartile

Les trois ***quartiles*** q_1, q_2 et q_3 sont définis comme la médiane, par

$$N'(q_1) = \frac{1}{4} \quad ; \quad N'(q_2) = \frac{1}{2} \quad ; \quad N'(q_3) = \frac{3}{4}$$

Les trois quartiles divisent l'ensemble des observations en quatre ensembles de même effectif

MATLAB : `quantile(x,p) or prctile(x,p)`

L'***écart-interquartile*** est défini par $q_3 - q_1$

nb : il existe également les quantiles (ou fractiles) d'ordre k, qui correspondent à d'autres fractions de l'unité. D'autres cas particuliers sont les déciles et les pourcentiles.

Réduction des données, Paramètres de dispersion V

▷ **Amplitude**

L'***amplitude*** w est définie par $w = x_n - x_1$. Elle n'a de sens que pour les données brutes (sinon, perte d'info sur les valeurs extrêmes)

Souvent, $s \simeq w/4$ ou $s \simeq w/5$ (distributions en cloche)

Réduction des données, Moments d'ordres supérieur I

▷ **Les moments**

Les ***moments non centrés*** d'ordre k sont définis par

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad \text{ou} \quad a_k = \frac{1}{n}\sum_{i=1}^{p}\left(n_i x_i^k\right)$$

Matlab : `moment(x,k)`

Les ***moments centrés*** d'ordre k sont définis par

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k \quad \text{ou} \quad m_k = \frac{1}{n}\sum_{i=1}^{p}\left[n_i\,(x_i - \overline{x})^k\right]$$

a. Le moment centré d'ordre 1 est nul, $m_1 = 0$

b. Le moment non centré d'ordre 1 est la moyenne, $a_1 \equiv \overline{x}$

c. Le moment centré d'ordre 2 est égal à la variance, $m_2 = s^2$

d. Le moment non centré d'ordre 2 (le carré moyen) s'écrit aussi $a_2 = s^2 + \overline{x}^2$

e. Les moments centrés d'ordre pair représentent une dispersion

f. Les moments centrés d'ordre impair représentent une symétrie

Réduction des données, Moments d'ordres supérieur II

▷ **Coefficients de Fisher (paramètre de dissymétrie et d'aplatissement)**
Les ***coefficients de Fisher*** sont définis par

$$\gamma_3 = \frac{m_3}{s^3} \quad \text{et} \quad \gamma_4 = \frac{m_4}{s^4}$$

Matlab : `skewness(x)`, `kurtosis(x)`

Exercices

1. Donnez une représentation sous la forme d'une distribution de fréquences groupée de l'âge de la population active dans le secteur statistique de Sainte-Marie. Quel est l'âge moyen de la population active dans ce quartier ? (Age-2001.xls)

2. Quantifiez la dispersion du nombre de femmes dans les différents secteurs statistiques (Population-2001.xls)

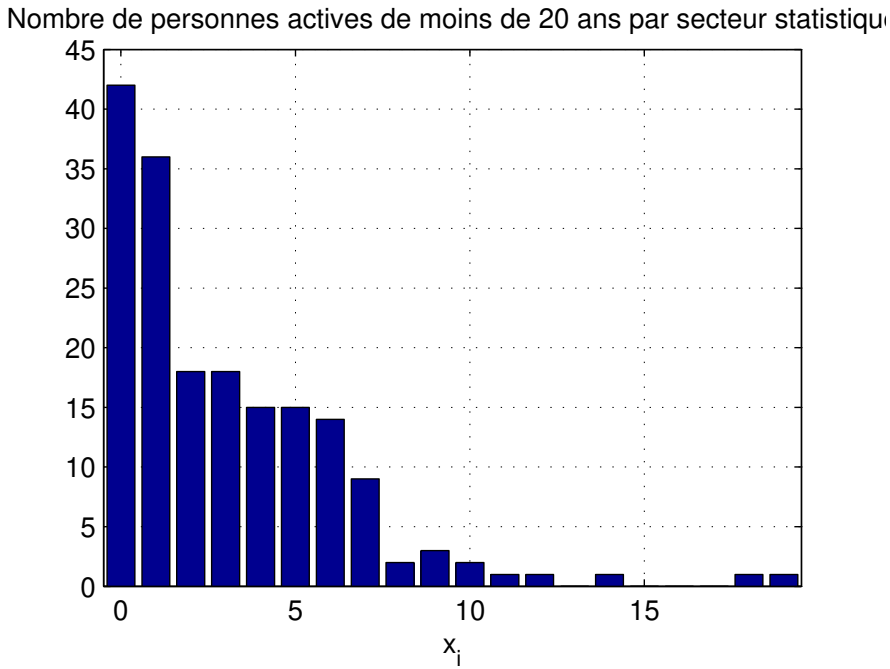
3. Dressez un tableau de distribution de fréquences du nombre de femmes dans les différents secteurs statistiques (Population-2001.xls)

Représentations graphiques I

▷ **Diagramme de fréquences non cumulées** (absolues ou relatives) - histogramme

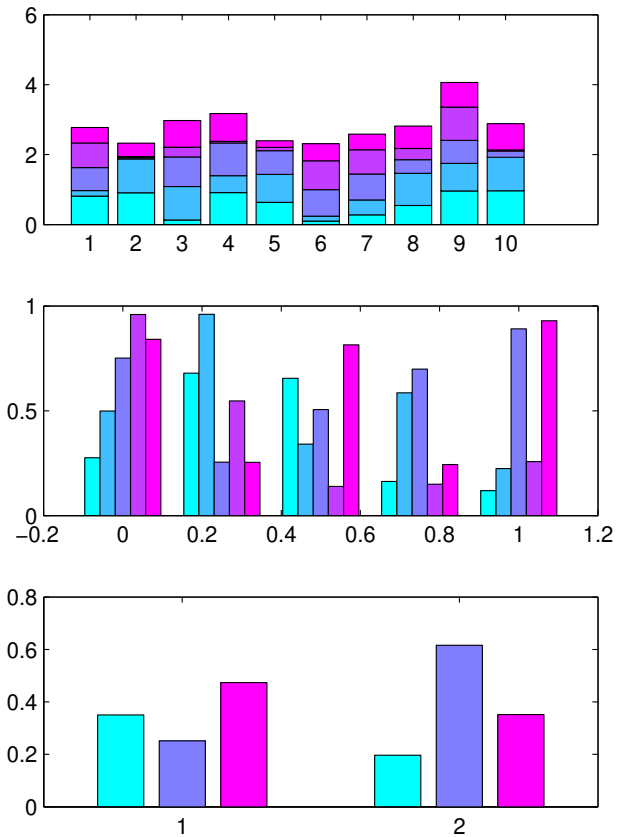
- diagrammes en bâtonnets (pour les distributions non groupées)
- histogrammes (pour les distributions groupées)
- polygone de fréquences

Représentations graphiques II



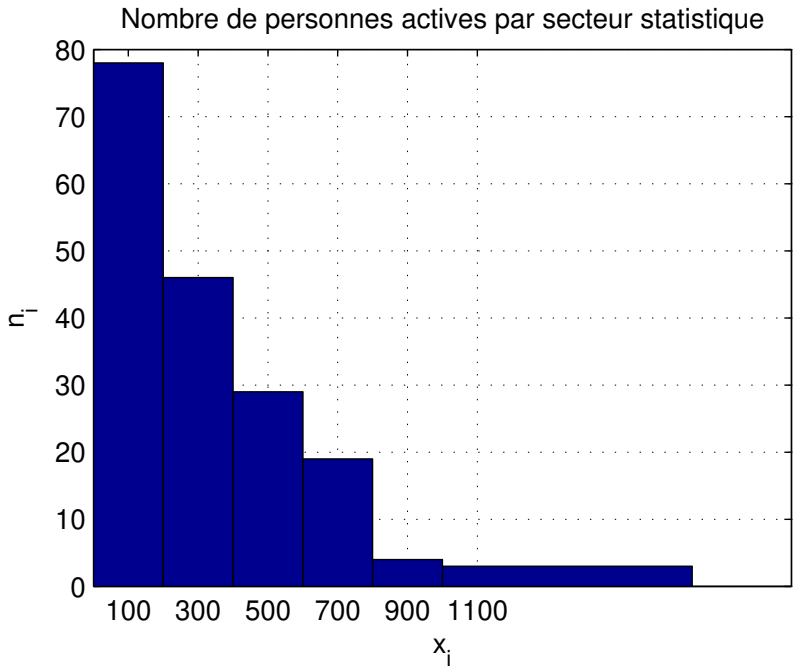
[diagramme en bâtonnets]

Représentations graphiques III



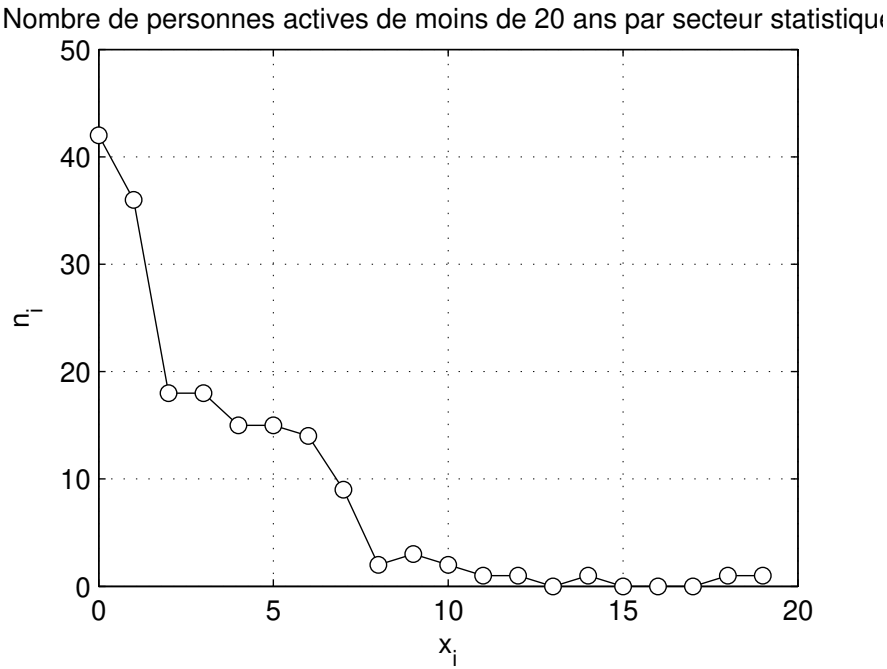
[diagrammes en bâtonnets]

Représentations graphiques IV



[histogramme]

Représentations graphiques V

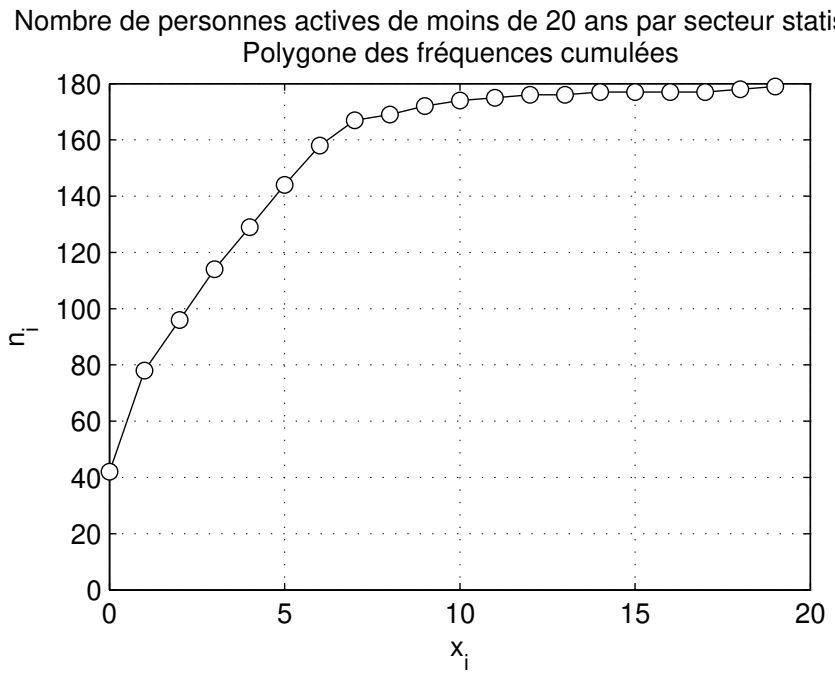


[polygone des fréquences]

Représentations graphiques VI

▷ **Diagramme de fréquences cumulées**

→ utilisation de polygones de fréquence : en escalier pour les distributions non groupées, avec une ligne brisée pour les distributions groupées

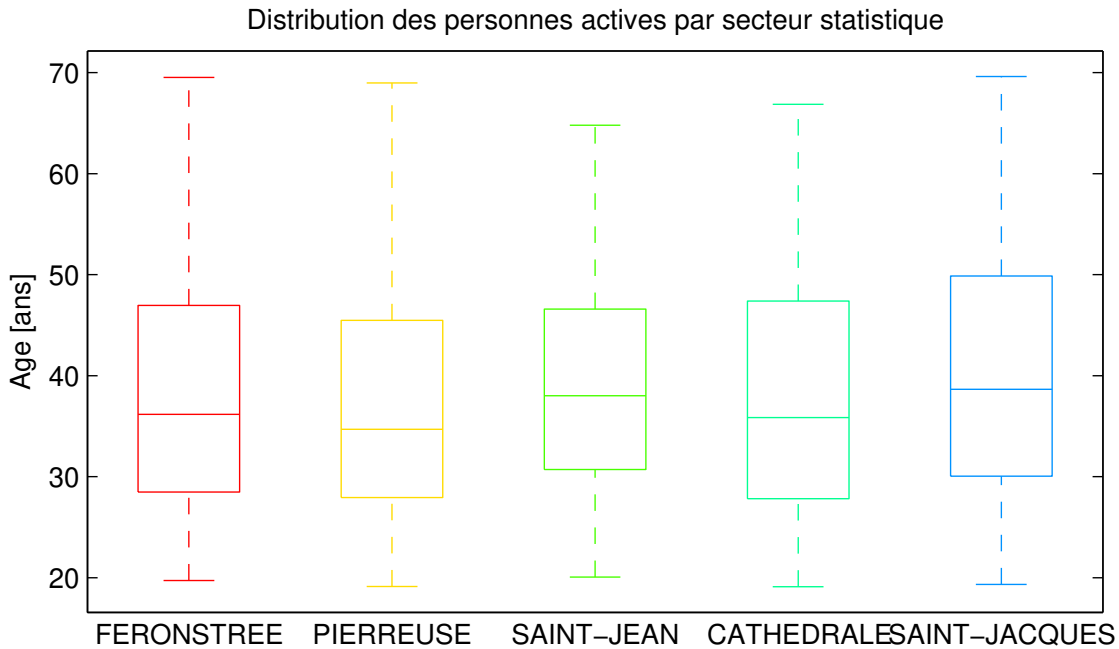


[polygone des fréquences cumulées]

Représentations graphiques VII

▷ Boxplot

Diagramme constitué de deux rectangles contigus délimités par les quartiles et de segments de droites qui s'étendent vers les valeurs extrêmes de façon à indiquer également l'amplitude.



Représentations graphiques VIII

▷ **Les principaux types de distribution de fréquence**

- distribution en cloche (zones de fréquence maximum, symétrique ou non, distribution des fréquences cumulées en S)
- distribution en I (distribution de fréquence monotonement décroissante, distribution des fréquences cumulées à concavité vers le bas)
- distribution en J (distribution de fréquence monotonement croissante, distribution des fréquences cumulées à concavité vers le haut)
- distribution en U (il existe un maximum de fréquence à chaque extrémité de la distribution)

Exercices I

1. {Matlab :} A l'aide de la fonction `randn`, générez une réalisation d'un échantillon de taille $n = 1000$ d'une variable aléatoire de moyenne 10 et d'écart-type 5. Etudiez la variabilité de la moyenne de cet échantillon, en répétant plusieurs fois ce tirage aléatoire. Discutez en fonction de la taille n de l'échantillon la variabilité sur la moyenne d'échantillon obtenue.

2. {Matlab :} Pour un échantillon de taille $n = 1000$ tiré au hasard dans une population normale, trouvez un bon compromis pour la valeur du nombre d'intervalles utilisé dans la représentation. Répétez la démarche avec $n = 10^4$ et $n = 10^5$. Evaluez votre proposition par rapport à la loi de Scott (1979), $3.5\sigma/n^{1/3}$. NB : choisissez la moyenne et l'écart-type σ de la population de façon arbitraire.

3. {Matlab :} Représentez l'histogramme du nombre de logements exclusivement destinés à l'habitation. Voir fichier data1.mat. Caractérisez la forme de la distribution (cloche, I, J, U). Représentez ces données à l'aide d'une boxplot. Source :(Logement-2001.xls)

Exercices II

4. Soit les résultats suivants, en pourcents, pour un test de connaissances :

$$67, 70, 72, 74, 76, 76, 78, 80, 82, 85$$

Calculez la médiane, la moyenne, le mode et l'écart-type. Quel pourcentage des résultats se trouvent à moins d'un écart-type de la moyenne ?

5. Dans une petite localité, on a relevé de nombre de pièces par appartement :

Nb pièces :	1	2	3	4	5	6	7
Nb appartements :	48	72	96	64	39	25	3

Le ■ nombre de pièces par appartement ■ est à considérer comme une variable aléatoire discrète à valeurs entières.

- Quel est l'effectif total ?
- Etablissez le tableau des fréquences relatives, des fréquences relatives cumulées.
- Evaluez la moyenne et l'écart-type du nombre de pièces par appartement dans cette localité.

Exercices III

6. ...

Réponses
4. moy=med=mod=76, $\sum x_i = 760$, $\sum x_i^2 = 58034$, var=27.4, std=5.23 ; 60% des données dans $\mu \pm \sigma$
5. $n = 347$, moy=3.18 pièces, c.moy=12.24pièces^2, std=1,47 pièces.

Distributions de fréquences I

▷ **Les séries de données (séries statistiques doubles)**

Une série statistique double est une suite de n couples de valeurs observées (x_i, y_i) appariées, éventuellement rangées dans l'ordre croissant d'une des deux variables :

$$x_1, \quad x_2, \ldots \quad x_n$$
$$y_1, \quad y_2, \ldots \quad y_n$$

Exemple 3 :
Nombre d'hommes et de femmes par secteur statistique à Liège (Population-2001.xls) :
Saint-Lambert : 720 hommes, 597 femmes
Feronstrée : 682 hommes, 594 femmes
Pierreuse : 274 hommes, 235 femmes
...ou de façon compacte...
(720,597), (682,594), (274,235), ... (173 valeurs)

Distributions de fréquences II

▷ **Les distributions de fréquences à deux dimensions**

Distribution de fréquences à deux dimensions : tableau statistique à double entrée dont une ligne est réservée à chaque valeur observée de x et une colonne à chaque valeur observée de y.
[groupée $\equiv$ utile lorsque l'effectif est grand (idem 1-D)]

Les fréquences n_{ij} représentent, pour chaque cellule du tableau, le nombre de couples d'observations (x_i, y_j).

nb : existence de distributions groupées ou non groupées

Distributions de fréquences III

	x_i						
y_j	0-200	201-400	401-600	601-800	801-1000	1001-...	Total
0-200	52	2	-	-	-	-	54
201-400	6	31	-	-	-	-	37
401-600	-	11	24	-	-	-	35
601-800	-	-	7	2	1	-	10
801-1000	-	-	-	13	6	1	20
1001-...	-	-	-	6	8	3	17
Total	58	44	31	21	15	4	**173**

Exemple :
Distribution de fréquences du nombre d'hommes (x_i) et de femmes (y_i) de chaque unité statistique (enquête INS).

Distributions de fréquences IV

	x_i					
y_j	0-500	501-1000	1501-2000	2001-2500	2501-...	Total
0-200	75	30	23	5	0	133
201-400	-	9	14	8	1	32
401-600	-	1	2	3	4	10
601-800	-	-	-	2	4	6
801-1000	-	-	-	1	-	1
1001-...	-	-	-	-	2	2
Total	75	40	39	19	11	**173**

Exemple :
Distribution de fréquences de la population belge (x_i) et étrangère (y_i) de chaque unité statistique (enquête INS).

Distributions de fréquences V

On définit également les ***fréquences relatives***

$$n'_{ij} = \frac{n_{ij}}{n}$$

et les ***fréquences unitaires***

$$n''_{ij} = \frac{n'_{ij}}{\Delta x_i \Delta y_j}$$

+ distribution de fréquences relatives cumulées

Distributions de fréquences VI

▷ Les distributions marginales

Les ***fréquences marginales*** $n_{i\bullet}$ et $n_{\bullet j}$ sont obtenues en calculant les totaux de chaque ligne ou chaque colonne.

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \quad \text{et} \quad n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$$

On a la propriété

$$\sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = n$$

Les ***fréquences relatives marginales*** $n'_{i\bullet}$ et $n'_{\bullet j}$ sont définies par

$$n'_{i\bullet} = \frac{n_{i\bullet}}{n} \quad \text{et} \quad n'_{\bullet j} = \frac{n_{\bullet j}}{n}$$

de sorte que

$$\sum_{i=1}^{p} n'_{i\bullet} = \sum_{j=1}^{q} n'_{\bullet j} = \sum_{i=1}^{p}\sum_{j=1}^{q} n'_{ij} = 1$$

Distributions de fréquences VII

▷ Les distributions conditionnelles

Les ***fréquences conditionnelles*** correspondent aux lignes et colonnes du tableau de distribution des fréquences. On définit ainsi les fréquences conditionnelles de "y si x" et de "x si y"

$$n'_{j|i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{n'_{ij}}{n'_{i\bullet}} \quad \text{et} \quad n'_{i|j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{n'_{ij}}{n'_{\bullet j}}$$

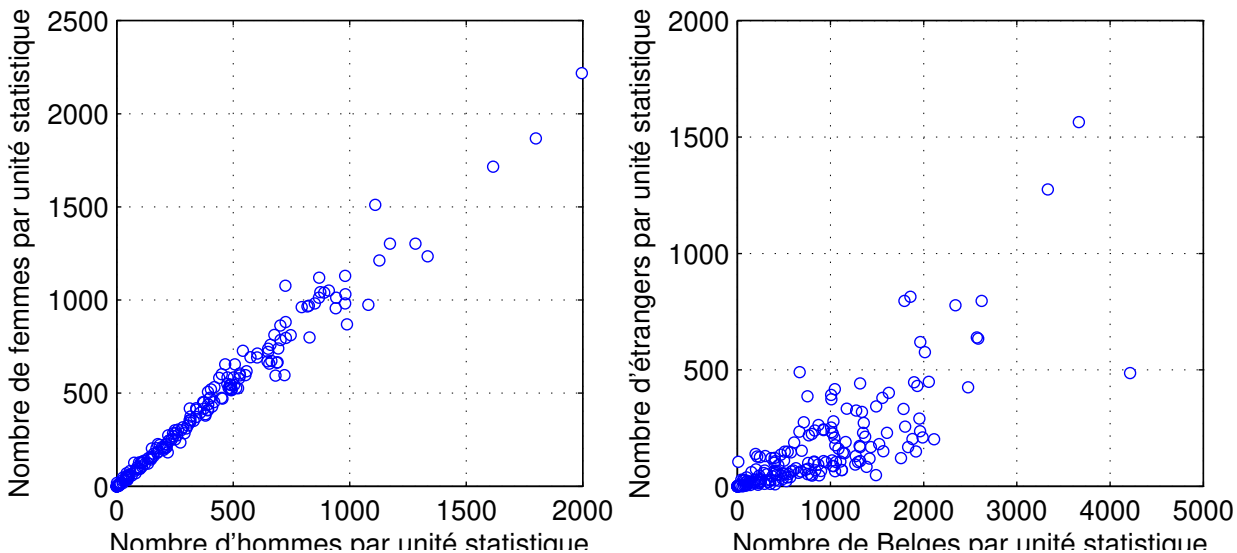
On a la propriété

$$\sum_{j=1}^{q} n'_{j|i} = \sum_{i=1}^{p} n'_{i|j} = 1$$

nb : illustrer avec l'exemple 3

Représentations graphiques I

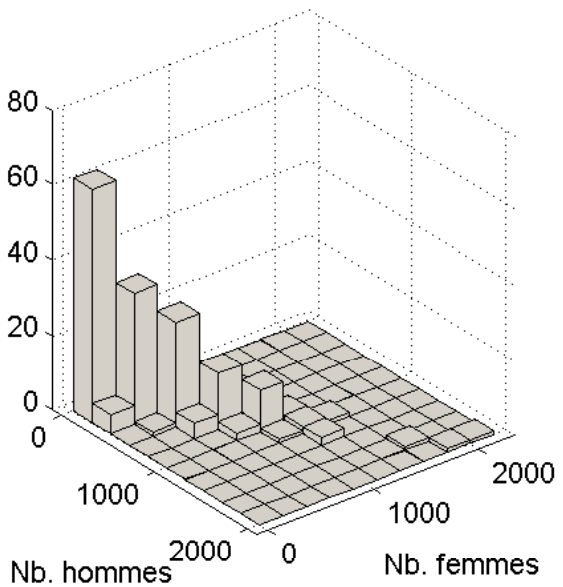
▷ Nuage de points

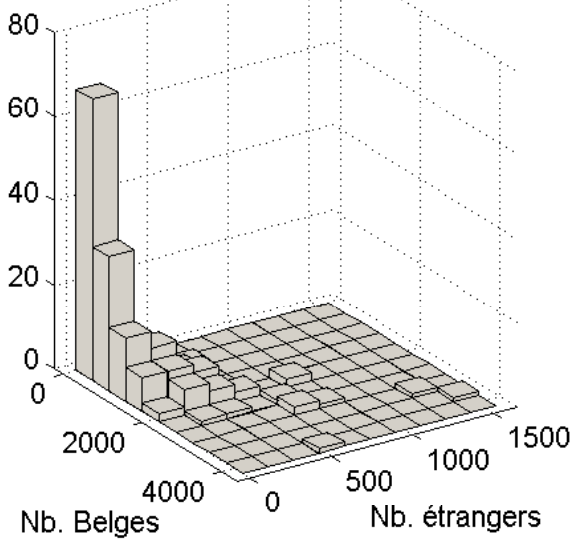


nb : introduire aussi le diagramme de régression (vs distribution)

Représentations graphiques II

▷ Histogramme à deux dimensions





Réduction des données : généralités I

Il existe deux types de paramètres. Ils visent :

- à caractériser une seule variable à la fois → distributions marginales et conditionnelles
- à caractériser les relations existant entre deux séries d'observations, considérées par valeurs appariées → caractéristiques croisées

▷ Paramètres de position

Moyennes marginales :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{et} \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Moyennes conditionnelles :

$$\overline{x}_j = \frac{1}{n_{\bullet j}}\sum_{i=1}^{p} n_{ij} x_i \qquad \text{et} \qquad \overline{y}_i = \frac{1}{n_{i\bullet}}\sum_{j=1}^{q} n_{ij} y_j$$

Réduction des données : généralités II

▷ Paramètres de dispersion

Variances marginales :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{et} \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Variances conditionnelles :

$$s_{x|j}^2 = \frac{1}{n_{\bullet j}}\sum_{i=1}^{p} n_{ij}\left(x_i - \overline{x}_j\right)^2 \qquad \text{et} \qquad s_{y|i}^2 = \frac{1}{n_{i\bullet}}\sum_{j=1}^{q} n_{ij}\left(y_j - \overline{y}_i\right)^2$$

▷ Caractéristiques croisées

le + important (!?), en tout cas spécifique aux valeurs appariées...

- les moments croisés et la covariance
- coefficient de correlation et de détermination
- régression linéaire et non linéaire

Les moments et la covariance I

Généralisation de la notion de moment → ***moments croisés***

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^k (y_i - d)^l$$

(dans le cas de données brutes)
Elles représentent les moments d'ordre k en x et d'ordre l en y, par rapport à c pour x et par rapport à d pour y

[le plus souvent, on choisit c et d égaux à 0 ou aux moyennes de x et y]
Notamment, si $c = \overline{x}$ et $d = \overline{y}$, on obtient les ***moments centrés croisés***

$$m_{kl} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k (y_i - \overline{y})^l$$

Cas particulier 1
Les variances marginales sont des cas particuliers

$$m_{20} = s_x^2 \qquad \text{et} \qquad m_{02} = s_y^2$$

Les moments et la covariance II

Cas particulier 2
La ***covariance*** est obtenue pour $k = l = 1$

$$cov(x, y) = m_{11} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Matlab : `cov(X)`

La covariance est positive ou négative selon que la relation entre les deux séries de données est croissante ou décroissante.

Démo : considérer les deux catégories de points $(x_i - \overline{x})(y_i - \overline{y}) > 0$ d'une part et $(x_i - \overline{x})(y_i - \overline{y}) < 0$ d'autre part.

nb : la covariance s'annule lorsque les deux catégories de points s'annulent.
nb : elle s'annule aussi lorsque tous les points observés sont situés sur une des deux droites $x = \overline{x}$ ou $y = \overline{y}$.

Les moments et la covariance III

a. La covariance est influencée par les changements d'unités, mais pas par les changements d'origines. En effet,

$$\left.\begin{array}{l} x_i' = a + bx_i \\ \\ y_i' = c + dy_i \end{array}\right\} \quad \Rightarrow \quad cov\left(x', y'\right) = b\, d\, cov\left(x, y\right)$$

b. La covariance est toujours inférieure (en valeur absolue) au produit des écarts-types

$$|cov\left(x, y\right)| \leq s_x s_y$$

Démo : observer que $\frac{1}{n}\sum_{i=1}^{n}\left[b\left(x_i - \overline{x}\right) - \left(y_i - \overline{y}\right)\right]^2 \geq 0$.

Le coefficient de corrélation et de détermination I

▷ Le ***coefficient de corrélation*** est défini par

$$r = \frac{cov(x, y)}{s_x s_y}$$

Matlab : `corrcoef(X)`

a. Le coefficient de corrélation possède le même signe que la covariance
b. Le coefficient de corrélation est inférieur à l'unité, en valeur absolue

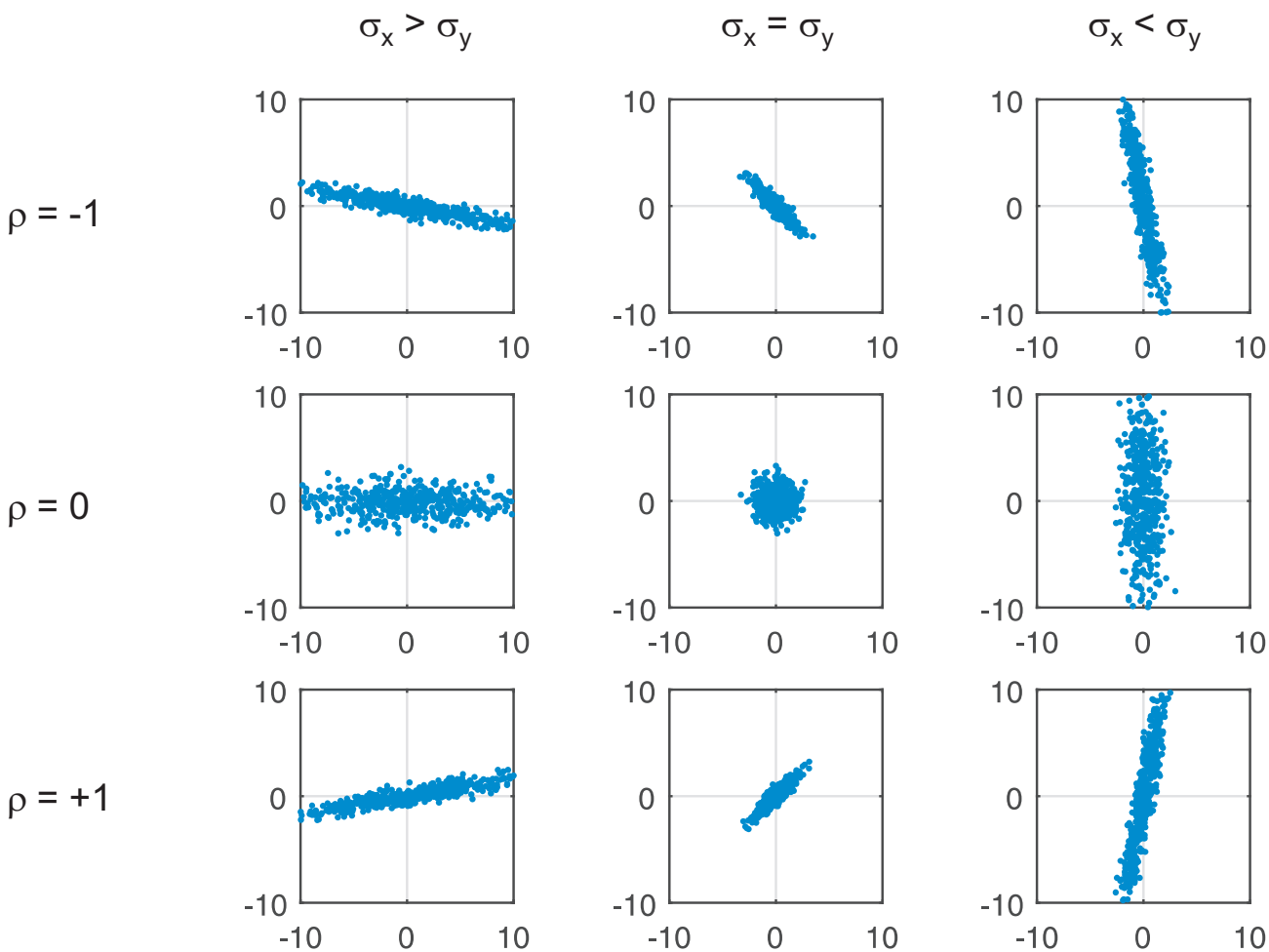
$$-1 \leq r \leq 1$$

Le coefficient de corrélation et de détermination II

Pour $r = 1$, tous les points observés se trouvent sur une droite de pente positive
Pour $r \lesssim 1$, tous les points observés se trouvent à proximité d'une droite de pente positive
Pour $0 < r < 1$, tous les points observés se trouvent sur un nuage plus ou moins dispersé autour d'une droite de pente positive
Pour $r \simeq 0$, le nuage de points est elliptiques à axes parallèles aux axes des données
Pour $-1 < r < 0$, tous les points observés se trouvent sur un nuage plus ou moins dispersé autour d'une droite de pente négative
Pour $-1 \lesssim r$, tous les points observés se trouvent à proximité d'une droite de pente négative
Pour $r = -1$, tous les points observés se trouvent sur une droite de pente négative

Le coefficient de corrélation mesure donc la *netteté* avec laquelle les points s'alignent **sur une droite**.

Le coefficient de corrélation et de détermination III



Le coefficient de corrélation et de détermination IV

a. Le coefficient de corrélation n'est pas influencé par les changements d'unité, ni par les changements d'origines. En effet,

$$\left.\begin{array}{l} x_i' = a + bx_i \\ \\ y_i' = c + dy_i \end{array}\right\} \quad \Rightarrow \quad r\left(x', y'\right) = r\left(x, y\right)$$

▷ Le ***coefficient de détermination*** est défini comme étant le carré du coefficient de corrélation

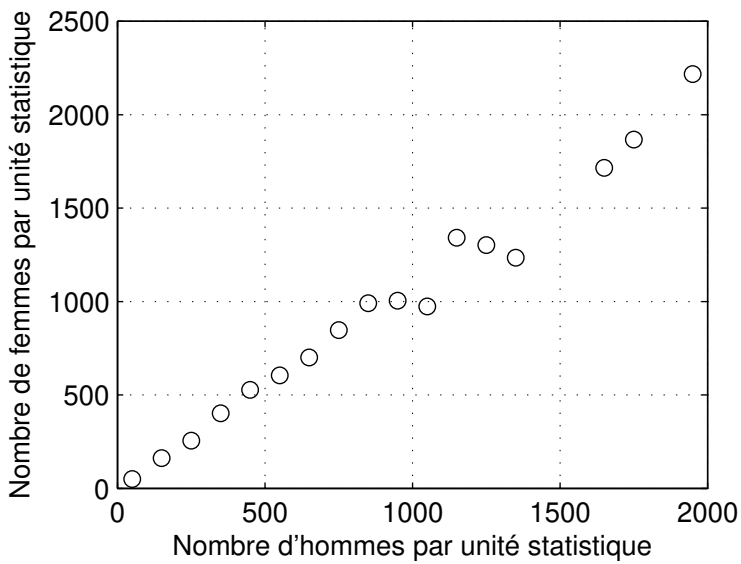
▷ Corrélation et causalité ? (si x et y sont fortement corrélés, c'est car : (i) l'un est la la cause/conséquence de l'autre, (ii) ils sont tous les deux la conséquence d'une même cause, (iii) il s'agit d'un corrélation fortuite)

Une étude a montré que « s'endormir avec des chaussures » était fortement corrélé avec « se réveiller avec un mal de tête ». *Donc s'endormir avec des chaussures cause un mal de tête.*
Discuter et commenter cette affirmation. Expliquez ce qu'est la corrélation et son lien avec la causalité.

Régression au sens des moindres carrés I

Le ***diagramme de régression de y en fonction de x*** est défini par les points moyens conditionnels $(x_i, \overline{y}_i)$.

Il donne une première idée de la façon dont la ***variable dépendante*** y varie, en moyenne, en fonction de la ***variable explicative*** x.



nb : il peut avoir une allure linéaire ou non linéaire

Régression au sens des moindres carrés II

Lorsque le diagramme de dispersion (ou le diagramme de régression) a une allure relativement linéaire, on peut tenter de préciser la relation linéaire qui lie y et $x \rightarrow$ droite de régression
La ***droite de régression*** est la droite obtenue en minimisant la somme des carrés des écarts entre les points observés et les points correspondants de la droite (méthode des moindres carrés).

Si l'équation de la droite est

$$y = a + bx,$$

la somme des carrés à minimiser est

$$\phi(a, b) = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

Le minimum est obtenu en annulant les dérivées de cette fonction par rapport à a et b

$$\frac{\partial \phi}{\partial a} = 0 \quad \text{et} \quad \frac{\partial \phi}{\partial b} = 0$$

Régression au sens des moindres carrés III

On obtient les équations normales

$$\begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0 \end{cases} \quad \text{ou} \quad \begin{cases} an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

On déduit donc que

$$\overline{y} = a + b\overline{x}$$

et donc que la droite de régression passe par le point moyen $(\overline{x}, \overline{y})$.

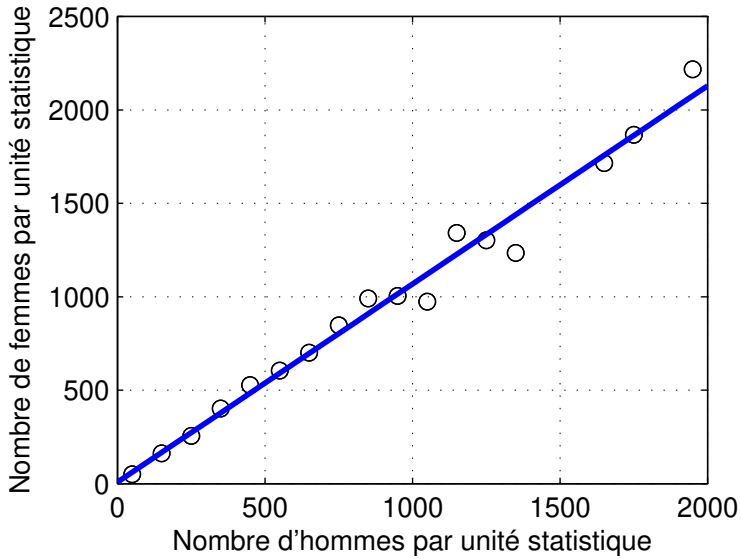
De plus,

$$b = \frac{\sum_{i=1}^{n} (x_i y_i) - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right) \left(\sum_{i=1}^{n} y_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{cov(x, y)}{s_x^2}$$

La droite de régression a donc pour équation

$$y = \overline{y} + \frac{cov(x, y)}{s_x^2} (x - \overline{x})$$

Régression au sens des moindres carrés IV



Régression au sens des moindres carrés V

▷ On appelle les ***résidus de la régression linéaire*** les écarts

$$d_i = y_i - y(x_i) = y_i - a - bx_i$$

a. les résidus sont de moyenne nulle

Les résidus donnent une indication quant à la qualité de la régression linéaire, c'est-à-dire, quant à l'existence d'une relation linéaire entre les y_i et les x_i.

Régression au sens des moindres carrés VI

▷ La ***variance résiduelle*** est précisément la valeur de la fonction que nous avons tenté de minimiser

$$s^2_{y.x} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - y\left(x_i\right)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y} - \frac{cov\left(x,y\right)}{s^2_x}\left(x_i - \overline{x}\right)\right)^2.$$

ou

$$s^2_{y.x} = s^2_y - \frac{cov^2\left(x,y\right)}{s^2_x} = s^2_y\left(1 - r^2\right) \quad \rightarrow s^2_{y.x} \leq s^2_y$$

avec comme cas limites $s^2_{y.x} = s^2_y$ (si $r = 0$) et $s^2_{y.x} = 0$ (si $r = \pm 1$).

La quantité $\frac{cov^2(x,y)}{s^2_x}$ correspond à la réduction de variance de la variable dépendante y qui est liée à la prise en considération de la variable x.
Elle est considérée comme la partie de la variance de y qui est expliquée par la régression linéaire de y en x, tandis que la variance résiduelle $s^2_{y.x}$ est la partie qui ne peut être expliquée de la sorte.
On note également que

$$r^2 = \left(\frac{cov^2\left(x,y\right)}{s^2_x}\right)/s^2_y$$

Illustrations - Exercices

1. Illustration de l'ajustement de courbes de régression dans Matlab (Basic fitting, residual plots, choix du degré du polynôme).

2. Choisissez* deux séries données appariées dans les données de l'enquête INS, et représentez les données sur un nuage de points et sur un histogramme à deux dimensions.

3. Déterminez l'équation de la droite de régression de l'une des séries en fonction de l'autre. Représentez-la sur le nuage de points.

4. Calculez la variance résiduelle.

*dont vous suspectez qu'elles sont corrélées...

Régression curvilinéaire I

Lorsque le diagramme de dispersion (ou le diagramme de régression) n'a pas une allure linéaire, on peut tenter de décrire la relation qui lie y et x par une autre expression analytique.

Le choix de cette équation peut se faire sur une base empirique ou théorique.

Exemples :

- Régression exponentielle

$$\ln y = a + bx \qquad \text{ou} \qquad y = c\, e^{bx}$$

- Régression polynomiale

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k$$

- Loi de Mitscherlich

$$y = \frac{y_{max}}{1 + e^{-a-bx}}$$

Régression curvilinéaire II

▷ dans certains cas, l'ajustement d'une courbe de régression non linéaire peut se ramener au cas d'une régression linéaire
Exemple :
La régression exponentielle

$$\ln y = a + bx$$

est linéaire en a et b.

▷ dans d'autres cas, la méthode des moidres carrés peut être appliquée sans difficulté
Exemple :
Régression polynomiale

$$y = b_0 + b_1 x + b_2 x^2 + ... + b_k x^k$$

On minimise

$$\phi(b_0, ..., b_k) = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i - ... - b_k x^k \right)^2$$

Régression curvilinéaire III

en annulant les $k+1$ dérivées partielles

$$\frac{\partial \phi}{\partial b_0} = 0, \ \frac{\partial \phi}{\partial b_1} = 0, \ ..., \ \frac{\partial \phi}{\partial b_k} = 0$$

Les équations normales s'écrivent

$$\begin{cases} b_0 n + b_1 \sum_{i=1}^n x_i + b_2 \sum_{i=1}^n x_i^2 + ... + b_k \sum_{i=1}^n x_i^k & = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 + b_2 \sum_{i=1}^n x_i^3 + ... + b_k \sum_{i=1}^n x_i^{k+1} & = \sum_{i=1}^n x_i y_i \\ \vdots & \vdots \\ b_0 \sum_{i=1}^n x_i^k + b_1 \sum_{i=1}^n x_i^{k+1} + b_2 \sum_{i=1}^n x_i^{k+2} + ... + b_k \sum_{i=1}^n x_i^{2k} & = \sum_{i=1}^n x_i^k y_i \end{cases}$$

et sont simples à résoudre.
[nb : il existe une version matricielle des équations normales.]

Matlab : `polyfit(x,y,n)`

Régression curvilinéaire IV

▷ dans certains cas, l'ajustement d'une courbe de régression non linéaire ne peut pas se ramener au cas d'une régression linéaire

Exemple :

Loi de Mitscherlich

$$y = \frac{y_{max}}{1 + c\, d^x}$$

On minimise

$$\phi\left(y_{max}, c, d\right) = \sum_{i=1}^{n} \left(y_i - \frac{y_{max}}{1 + c\, d^x} \right)^2$$

en annulant les dérivées partielles

$$\frac{\partial \phi}{\partial y_{max}} = 0, \quad \frac{\partial \phi}{\partial c} = 0 \quad \text{et} \quad \frac{\partial \phi}{\partial d} = 0$$

(nécessité d'utiliser une méthode numérique)

Exercices I

1. *[Jui13]* Le tableau ci-dessous représente la distribution de fréquences du nombre d'hommes (x_i) et de femmes (y_i) de chaque unité statistique (enquête INS).

	x_i						
y_j	0-200	201-400	401-600	601-800	801-1000	1001-...	Total
0-200	52	2	-	-	-	-	54
201-400	6	31	-	-	-	-	37
401-600	-	11	24	-	-	-	35
601-800	-	-	7	2	1	-	10
801-1000	-	-	-	13	6	1	20
1001-...	-	-	-	6	8	3	17
Total	58	44	31	21	15	4	**173**

(a) dressez le tableau des fréquences relatives

(b) donnez les tableaux des fréquences marginales (homme et femme)

Exercices II

(c) donnez le tableau des fréquences conditionnelles du nombre d'homme par unité lorsque les femmes sont en effectif compris entre 401 et 600. Discutez la signification du résultat.
(d) estimez le coefficient de corrélation de ces deux variables aléatoires

2. Deux méthodes différentes ont été utilisées pour déterminer le coefficient d'isolation d'un bâtiment. Elles ont été testées sur cinq bâtiments différents, en vue de leur comparaison. Voici les résultats obtenus (1.23 ;1.56), (2.13 ;2.43), (0.98 ;1.23) ; (0.7 ;0.65), (1.8 ; 1.65). Comment quantifier la similitude/dissimilitude entre ces deux séries de résultats ?

3. *[Sept2015]* On réalise une étude dont l'objectif de déterminer si les investissements dans l'isolation de bâtiments d'habitations existant permet de réduire les émissions calorifiques. Voici les résultats d'une étude-pilote qui présentent le coût d'investissement C en fonction de la diminution de consommation de chauffage, exprimée en litres de mazout.

Exercices III

C [k€]	6.3	7.9	11.0	13.4	9.8	10.2	6.3	7.4	5.5	10.1	12.1
Δq [ℓ]	268	266	326	560	728	360	242	306	248	332	432

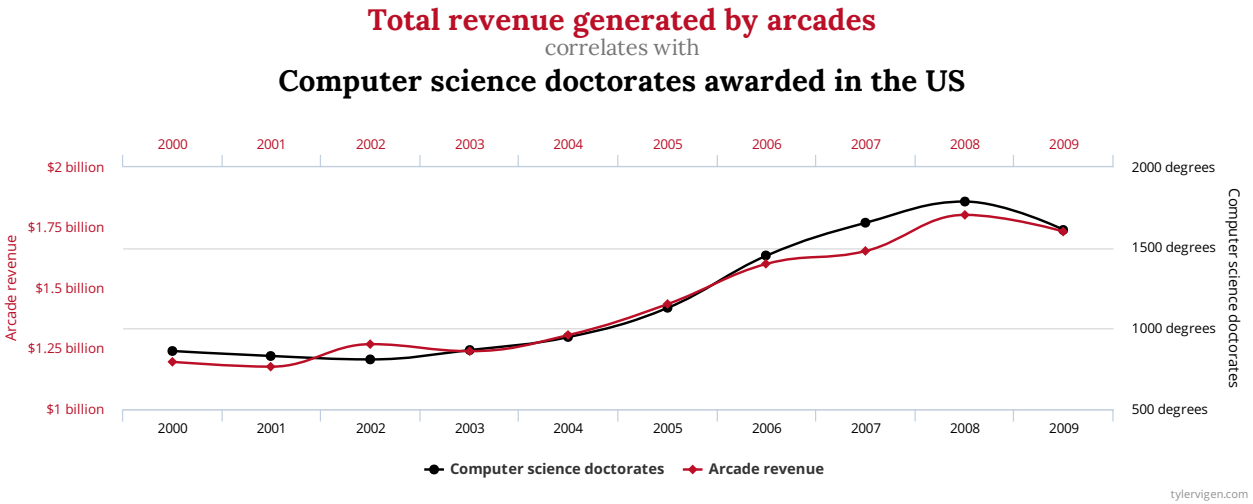
Quantifiez la corrélation existant entre les deux grandeurs et concluez quant aux résultats de cette étude.

Exercices IV

1.

```
xy = [ 52 2 0 0 0 0; 6 31 0 0 0 0; 0 11 24 0 0 0; 0 0 7 2 1 0; 0 0 0 13 6
1; 0 0 0 6 8 3];
n = sum(xy(:));
disp ('f relatives')
disp(xy / n)
disp ('Marginale x')
x = sum(xy); disp([1:6; x/n]')
disp ('Marginale y')
y = sum(xy'); disp([1:6; y/n]')
disp ('Conditionnelle')
disp([(1:6)' xy(:,3)/sum(xy(:,3))])
v = 100:200:1100;
xmoy = v*x'/n; ymoy = v*y'/n;
disp ('Moyennes'), disp([xmoy ymoy])
x2moy = (v.^2)*x'/n; y2moy = (v.^2)*y'/n;
disp ('Ecarts-types'), stdx = sqrt(x2moy-xmoy^2); stdy = sqrt(y2moy-ymoy^2);
disp([stdx stdy])
covxy = sum(sum( (v'*v).*xy/n )) - xmoy * ymoy; r = covxy / (stdx*stdy);
disp ('Coefficient de correlation') disp(r)
```

Exercices V



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
x	861	830	809	867	948	1129	1453	1656	1787	1611
y	1.196	1.176	1.269	1.24	1.307	1.435	1.601	1.654	1.803	1.734

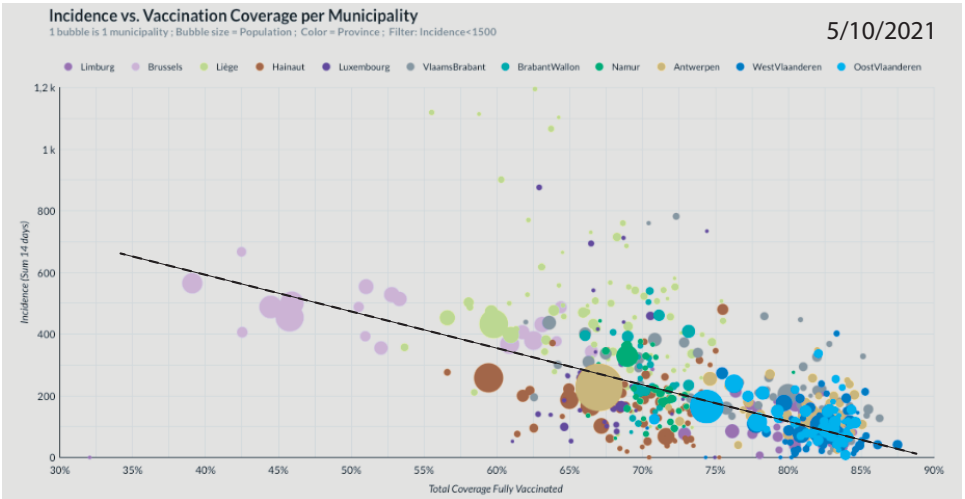
x = Nb. doct, y=Revenu $\cdot 10^9\$$ — Corrélation : $r = 0.9851$

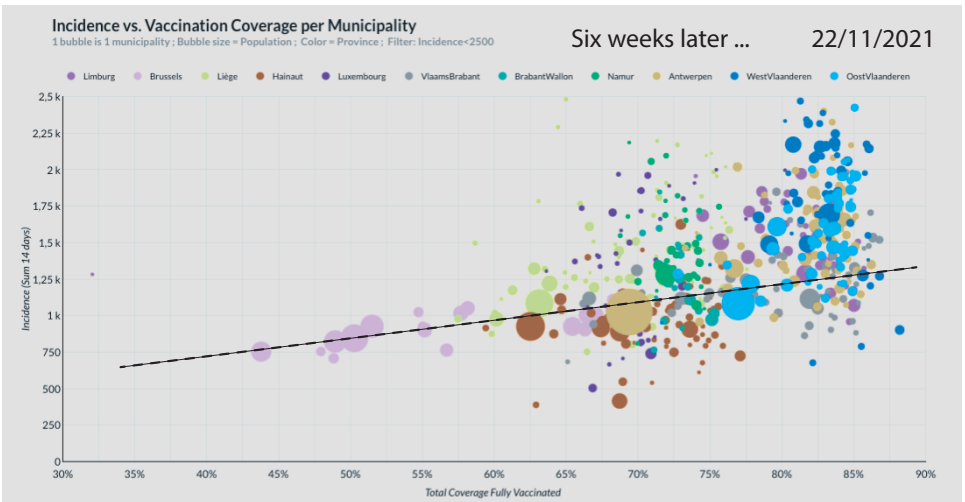
Les doctorants qui font un doctorat en sciences informatiques (aux Etats-Unis) sont de vrais gamers ! Discutez.

Exercices VI

3è Bac AR, 2023-2024

V. Denoël





Exercices VII

(i) Peut-on dire que l'incidence de la pandémie de covid-19 est influencée par la vaccination ? (ii) Quid des petites communes de la province de Liège à forte incidence ? (iii) Discutez la question de causalité à partir de ce graphique.

Partie III : Lois de Probabilité

Acquis d'apprentissage :

- classification des différentes lois de probabilité
- importance de la distribution normale et théorème limite central
- modélisation probabiliste à l'aide de variable aléatoire
- opérations algébriques et transformation de variables aléatoires

Notion de probabilité I

Différence entre statistiques et probabilités ?

- relation entre statistiques et probabilités
- probabilité conditionnelle
- indépendance stochastique
- espérance mathématique
- distributions de probabilité essentielles

Notion de probabilité II

▷ ***Définition classique de la probabilité***

La définition de la probabilité est liée aux notions d'expérience et d'évènement aléatoire. Une ***expérience*** est dite ***aléatoire*** quand on ne peut pas en prévoir exactement le résultat, en raison du fait que tous les facteurs qui déterminent ces résultats ne sont pas parfaitement maitrisables. Un ***évènement aléatoire*** est un évènement qui peut ou non se réaliser au cours d'une expérience aléatoire.

Exemple : tirage d'une carte d'un paquet de carte
expérience = tirage de la carte
événement = obtenir une coeur

Si m résultats peuvent se produire avec des chances égales au cours de l'expérience aléatoire, et si k de ces résultats conduisent à l'évènement A, on définit la probabilité d'occurrence de l'évènement A par le rapport du nombre de cas favorables au nombre de cas équiprobables possibles

$$P(A) = \frac{k}{m}$$

P(coeur) = 13/52 = 1/4

Notion de probabilité III

Problème 1 : on définit la notion de probabilité à partir de la notion d'égales probabilités... (!?)

Problème 2 : cette "définition" ne fonctionne que si on est capable de définir des cas "équiprobables" et dénombrables.

Notion de probabilité IV

▷ ***Définition fréquentiste de la probabilité***
On répète une expérience (un grand nombre) n fois et on *dénombre* le nombre de fois que l'évènement A est observé, sa fréquence absolue n_A. On définit la probabilité associée à l'évènement A par

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} = \lim_{n\to\infty} n'_A$$

nb : l'existence de cette limite est garantie par le phénomène de stabilité des fréquences relatives (régularité stochastique)

$\to$ probabilité = version idéalisée de la fréquence relative

Problème 1 : Elle ne peut être déterminée que d'une façon approchée... nb : est-ce vraiment un problème ? quelle est la masse d'un objet ?

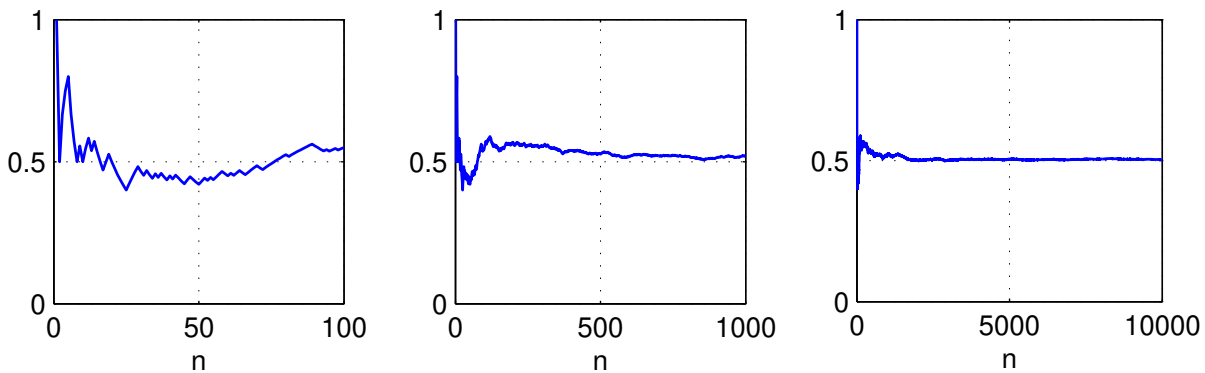
$\rightsquigarrow$ lien avec les statistiques descriptives : propriétés d'une population "théoriquement infinie". La théorie des probabilités vise à caractériser cette population théorique(ment).

Notion de probabilité V

Exemple : Jet d'une pièce de monnaie

Définition classique de la probabilité : on admet 1/2 comme valeur pour la probabilité d'obtenir "face" (2 résultats équiprobables)

Définition *fréquentiste* de la probabilité : *pour une pièce donnée*, on peut déterminer la fréquence relative associé à l'évènement "obtenir face".



Notion de probabilité VI

MATLAB : On peut utiliser la fonction `randi(imax,M,N)` pour générer un tableau de `MxN` nombres entiers tirés aléatoirement entre 1 et `imax`.

Generate integer values from the uniform distribution on the set 1 :10.

```
r = randi(10,100,1);
```

Generate integer values drawn uniformly from -10 :10.

```
r = randi([-10 10],100,1);
```

Exemple du jet de la pièce :

```
N=10000; results = randi(2,N,1)-1;
plot(cumsum(results)./(1:N)')
xlabel ('n'); grid on
```

Les axiomes de base

a. La probabilité est toujours comprise entre 0 et 1

$$0 \leq P(A) \leq 1$$

(puisque la fréquence relative est toujours comprise entre 0 et 1)

b. La probabilité d'un événement qui doit nécessairement se réaliser est unitaire → ***événement certain***
(nombre de solution équiprobables = 1)

c. Soit deux événements A et B associés à la même expérience aléatoire, mais ne pouvant pas se produire simulténament (***événements exclusifs***)
exemple : A = obtenir une coeur / B = obtenir une pique
contre-exemple : A = obtenir une coeur / B = obtenir une dame

Notion de probabilité VIII

Si on observe n_A réalisations de l'événement A et n_B réalisations de l'événement B (mutuellement exclusifs), avec $n_A + n_B \leq n$, parmi un nombre n de réalisations de l'expérience stochastique, on a

$$n'_A = \frac{n_A}{n} \quad , \quad n'_B = \frac{n_B}{n} \quad , \quad n'_{A \text{ ou } B} = \frac{n_A + n_B}{n}$$

et donc

$$n'_{A \text{ ou } B} = n'_A + n'_B$$

Par extension (à la limite $n \to +\infty$),

$$P(A \text{ ou } B) = P(A) + P(B)$$

(axiome d'additivité)

e. Si on peut identifier les m seuls événements exclusifs A_1, A_2, ...A_m possibles d'une expérience (***événements complémentaires***), on a

$$P(A_1 \text{ ou } A_2 \text{ ou } ...A_m) = P(A_1) + P(A_2) + ... + P(A_m) = 1$$

Notion de probabilité IX

f. Dans le cas de deux évènements A et B, qui *ne sont pas nécessairement mutuellement exclusifs*,

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ et } B)$$

car

$$P(A \text{ ou } B) = P(A \text{ sans } B) + P(B \text{ sans } A) + P(A \text{ et } B)$$

où $P(A \text{ sans } B) = P(A) - P(A \text{ et } B)$ et $P(B \text{ sans } A) = P(B) - P(A \text{ et } B)$.

Probabilité conditionnelle et indépendance stochastique I

Soit une expérience aléatoire qui peut mener à la réalisation (ou la non-réalisation) de deux évènements A et B, non nécessairement exclusifs. A l'issue de $n \gg$ répétitions de cette expérience, on observe

n_{11} réalisations de A et B	n_{12} réalisations de A sans B
n_{21} réalisations de B sans A	n_{22} non-réalisations de A et B

La fréquence conditionnelle relative de A sous la condition B est

$$n'_{A|B} = \frac{n_{11}}{n_{11} + n_{21}} = \frac{n_{11}}{n_{\bullet 1}} = \frac{n\,(A \text{ et } B)}{n\,(B)}.$$

De même

$$n'_{B|A} = \frac{n_{11}}{n_{11} + n_{12}} = \frac{n_{11}}{n_{1\bullet}} = \frac{n\,(A \text{ et } B)}{n\,(A)}$$

Probabilité conditionnelle et indépendance stochastique II

Par extension, on définit (lorsque $P(B) \neq 0$), ***la probabilité conditionnelle de A sous la condition B par***

$$P(A|B) = \frac{P(A \text{ et } B)}{P(B)} \quad \text{et} \quad P(B|A) = \frac{P(A \text{ et } B)}{P(A)}.$$

e. **Propriété de multiplicativité (ou, des probabilités composées) :**

$$P(A \text{ et } B) = P(A)\,P(B|A) = P(B)\,P(A|B)$$

Pour m évènements :

$$\begin{aligned} P(A_1 \text{ et } A_2 \text{ et } \ldots \text{ et } A_m) \quad = \quad & P(A_1)\,P(A_2|A_1)\,P[A_3|(A_1 \text{ et } A_2)] \\ & \cdots P[A_m|(A_1 \text{ et } \ldots \text{ et } A_{m-1})] \end{aligned}$$

Probabilité conditionnelle et indépendance stochastique III

	$x_i \equiv X$						
$y_j \equiv Y$	0-200	201-400	401-600	601-800	801-1000 (A)	1001-...	Total
0-200	52	2	-	-	-	-	54
201-400	6	31	-	-	-	-	37
401-600	-	11	24	-	-	-	35
601-800	-	-	7	2	1	-	10
801-1000 (B)	-	-	-	13	6	1	20
1001-...	-	-	-	6	8	3	17
Total	58	44	31	21	15	4	**173**

Exemple : Distribution de fréquences du nombre d'hommes (x_i) et de femmes (y_i) de chaque unité statistique (enquête INS).

Questions : déterminer

$$P(A), \quad P(B), \quad P(A \text{ et } B), \quad P(A|B), \quad P(B|A), \quad P(A|\text{non-}B)$$

Probabilité conditionnelle et indépendance stochastique IV

L'indépendance stochastique

L'évènement A est dit ***stochastiquement indépendant*** de l'évènement B si la probabilité de voir A se réaliser ne dépend pas de la réalisation ou non-réalisation de B :

$$P(A|B) = P(A|non - B)$$

On démontre que, si cette condition est satisfaite,

$$\begin{aligned} P(A|B) &= P(A|non - B) = \frac{P(A \text{ et } B)}{P(B)} = \frac{P(A \text{ et } non - B)}{P(non - B)} \\ &= \frac{P(A \text{ et } B) + P(A \text{ et } non - B)}{P(B) + P(non - B)} = P(A) \end{aligned}$$

De plus,

$$P(A \text{ et } B) = P(A)\,P(B)$$

Du coup

$$P(B|A) = \frac{P(A \text{ et } B)}{P(A)} = P(B) = \frac{P(B) - P(A \text{ et } B)}{1 - P(A)} = \frac{P(non - A \text{ et } B)}{P(non - A)}$$

Probabilité conditionnelle et indépendance stochastique V

Donc : si A est stochastiquement indépendant de B, alors B est aussi stochastiquement indépendant de A (**réciprocité de l'indépendance stochastique**) .

Pour m évènements stochastiquement indépendants :

$$P(A_1 \text{ et } A_2 \text{ et } \ldots \text{ et } A_m) \quad = \quad P(A_1)\,P(A_2)\cdots P(A_m)$$

Exemple : Distribution de fréquences du nombre d'hommes (x_i) et de femmes (y_i) de chaque unité statistique (enquête INS).
Questions : comparer

$$P(A).P(B), \quad \text{et} \quad P(A \text{ et } B)$$

Probabilité conditionnelle et indépendance stochastique VI

Exemple : deux jets d'un dé à six faces

On lance deux fois un dé à 6 faces. On considère les évènements suivants :

- A : le résultat du premier jet est impair,
- B : le résultat du second jet est impair,
- C : la somme des résultats est impaire.

Vérifier que ces trois évènements sont stochastiquement indépendants deux à deux, mais qu'ils ne sont pas indépendants lorsque les trois évènements sont considérés simultanément.

Probabilité conditionnelle et indépendance stochastique VII

Théorème de Bayes

$$P(A|B) = \frac{P(A \text{ et } B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P(B|non-A)\,P(non-A)}$$

[utile dans les méthodes d'inférence statistique]

On appelle $P(A)$ ***la probabilité a priori*** et e $P(A|B)$ la ***probabilité a posteriori.***

[Exemple (célèbre exemple de Hays et Winckler, 1971)].
Une personne se réveille pendant la nuit à cause d'un mal de tête. Elle avale ce qu'elle pense être un calmant, en avalant dans l'obscurité un comprimé prélevé dans un flacon trouvé à portée de main. Peu de temps après, elle se sent plus mal et de nouveaux symptômes apparaissent. Elle se rend compte que trois flacons se trouvaient à portée de main, dont deux contiennent le calmant qu'elle voulait prendre et le troisième contient une substance toxique. Cette personne se pose la question de savoir si, par mégarde, elle n'aurait pas avalé un comprimé de la substance toxique.
[nb : Son médecin l'a informée que le calmant qu'elle a voulu prendre peut provoquer dans 5 cas sur 100 des symptômes semblables à ceux qu'elle a ressentis et que la substance toxique provoque de tels symptômes dans 80% des cas.]

Probabilité conditionnelle et indépendance stochastique VIII

Soit A (et non-A) l'absorption du calmant (de la substance toxique) et soit B l'apparition de symptomes.

On sait que

$$P(B|A) = \frac{1}{20} \quad \text{et} \quad P(B|non-A) = \frac{4}{5}$$

Si on suppose que le choix aléatoire parmi les trois flacons se fait de façon uniforme (***probabilités a priori***),

$$P(A) = \frac{2}{3} \quad \text{et} \quad P(non-A) = \frac{1}{3}.$$

Le théorème de Bayes donne (***probabilités a posteriori***)

$$P(A|B) = \frac{\frac{1}{20}\cdot\frac{2}{3}}{\frac{1}{20}\cdot\frac{2}{3}+\frac{4}{5}\cdot\frac{1}{3}} = \frac{1}{9} \quad \text{et} \quad P(non-A|B) = \frac{\frac{4}{5}\cdot\frac{1}{3}}{\frac{4}{5}\cdot\frac{1}{3}+\frac{1}{20}\cdot\frac{2}{3}} = \frac{8}{9}$$

càd les probabilités d'avoir ingéré (ou non) la substance toxique, sachant que la personne a manifesté les symptomes supplémentaires.

Variable aléatoire 1-D et distribution théorique I

Une ***variable aléatoire*** X est une variable associée à une expérience aléatoire, servant à caractériser le résultat de cette expérience. Elle est *discrète* ou *continue* en fonction de la nature de la variable.

▷ **variables aléatoires discrètes**

[nb : on suppose que la variable ne peut prendre que des valeurs entières - par extension, on peut traiter d'autres variables discrètes, voire non numérales, qualitatives]

exemple : âge d'un individu

A chaque valeur x_i que peut prendre la variable aléatoire X, il correspond une probabilité

$$P(x_i) = \text{prob}(X = x_i)$$

La fonction $P(x)$ est appelée ***distribution de probabilité.***
La ***fonction de répartition*** $F(x)$ est la distribution cumulée des probabilités, càd

$$F(x_i) = \text{prob}(X \leq x_i)$$

Variable aléatoire 1-D et distribution théorique II

- puisque les résultats sont mutuellement exclusifs,

$$\sum_{i=0}^{+\infty} P(x_i) = 1$$

- la fonction de répartition est une fonction en escalier avec

$$0 \leq F(x_i) \leq 1$$

avec $F(-\infty) = 0$ et $F(+\infty) = 1$.

Exemple : jet d'un dé à six faces
Exemple : distribution uniforme discontinue

si on répète l'expérience suffisamment,
les distributions observées tendent
vers les distributions de probabilité.

Variable aléatoire 1-D et distribution théorique III

▷ **variables aléatoires continues**
[nb : la variable peut prendre des valeurs réelles sur un intervalle donné]
exemple : poids d'un individu

On peut dans certains cas déterminer la probabilité d'observer une valeur comprise dans un intervalle donné $[x; x + \Delta x]$

$$\text{prob}\,(x < X \leq x + \Delta x)$$

En général, cette probabilité tend vers 0 pour $\Delta x \to 0$ (la probabilité d'obtenir exactement un résultat donné est généralement nulle). La notion de distribution de probabilité n'a plus de sens.
Par contre, la notion de *fonction de répartition* garde tout son sens

$$F(x) = \text{prob}\,(X \leq x)$$

La probabilité d'observer X dans un intervalle est

$$\text{prob}\,(x < X \leq x + \Delta x) = F(x + \Delta x) - F(x)$$

Si $F(x)$ est dérivable,

Variable aléatoire 1-D et distribution théorique IV

$$\lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = p_X(x)$$

La fonction

$$p_X(x) = \frac{dF}{dx}$$

est appelée la ***densité de probabilité*** de la variable aléatoire X. La quantité $p_X(x)\,dx$ représente la probabilité d'observer une réalisation de X dans un intervalle de longueur dx centré sur x.

a. En raison de la définition, la fonction de répartition est un primitive de la densité de probabilité :

$$F(x) = \int_{-\infty}^{x} p_X(t)\,dt$$

b. Normalisation de la densité de probabilité

$$\int_{-\infty}^{+\infty} p_X(x)\,dx = F(+\infty) = 1$$

Variable aléatoire 1-D et distribution théorique V

Relation entre statistiques et probabilités :
A la limite, pour un effectif infiniment grand ($n \to +\infty$), l'histogramme normé d'une distribution tend à se rapprocher de la densité de probabilité.

Exemple : distribution uniforme continue sur l'intervalle [a ;b]

$$p_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon} \end{cases}$$

Exemple : distribution normale

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Variable aléatoire 1-D et distribution théorique VI

MATLAB : On peut utiliser la fonction `Y=PDF(NAME,X,A,B)` pour estimer la densité de probabilité de variables continues et discontinues de distributions courantes.

MATLAB : On peut utiliser la fonction `Y=CDF(NAME,X,A,B)` pour estimer la fonction de répartition de variables continues et discontinues de distributions courantes.

'beta' or 'Beta'	'bino' or 'Binomial'
'chi2' or 'Chisquare'	'exp' or 'Exponential'
'ev' or 'Extreme Value'	'f' or 'F'
'gam' or 'Gamma'	'gev' or 'Generalized Extreme Value'
'gp' or 'Generalized Pareto	'geo' or 'Geometric
'logn' or 'Lognormal'	'norm' or 'Normal'
'poiss' or 'Poisson'	'rayl' or 'Rayleigh'
't' or 'T'	'unif' or 'Uniform'
'unid' or 'Discrete Uniform'	'wbl' or 'Weibull'

Exercices I

1. Soit X une variable aléatoire normale (de moyenne nulle et d'écart-type unitaire). Déterminez la probabilité que la variable X soit comprise dans des intervalles $[-n\sigma, n\sigma]$, avec $n = 1, 2, 3$. Utilisez, éventuellement, la fonction `cdf` de Matlab.

2. L'âge d'une population est supposé être distribué selon une distribution normale de moyenne $\mu = 55$ ans et d'écart-type $\sigma = 10$ ans. Quelle est la probabilité de dépasser l'âge de 70 ans ?
(nb : quelle est la probabilité d'avoir un âge inférieur à 0 ! ?)

Variable aléatoire 2-D et distribution théorique I

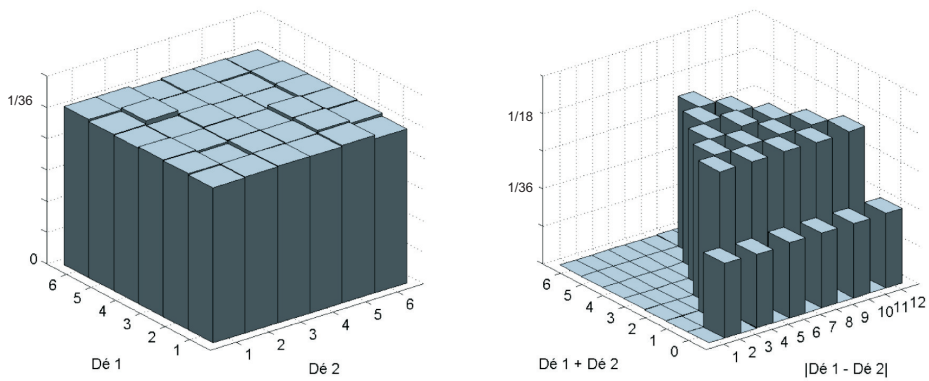
Dans de nombreux cas, les résultats d'une expérience sont caractérisés par deux (ou plusieurs) variables aléatoires. Elles peuvent être *discrètes* ou *continues* (ou mixtes) en fonction de la nature de la variable.

▷ **variables aléatoires discrètes**

A chaque couple de valeurs $(X = x, Y = y)$ correspond une probabilité

$$P(x, y) = \text{prob}(X = x \text{ et } Y = y)$$

La fonction $P(x, y)$ est appelée ***distribution de probabilité.***



Variable aléatoire 2-D et distribution théorique II

La ***fonction de répartition*** $F(x,y)$ correspondante est la distribution cumulée des probabilités, càd

$$F(x,y) = \text{prob}(X \leq x \text{ et } Y \leq y).$$

[forme idéalisée des distributions de fréquences cumulées]

- puisque les résultats sont mutuellement exclusifs,

$$\sum_{x=0}^{+\infty}\sum_{y=0}^{+\infty} P(x,y) = 1$$

- la fonction de répartition est une fonction "en escalier" avec

$$0 \leq F(x,y) \leq 1$$

avec $F(0,0) = 0$ et $F(+\infty,+\infty) = 1$.
De plus, en sommant par rapport à une des variables, on retrouve les distributions marginales

$$P_X(x) = \sum_{y=0}^{+\infty} P(x,y) \quad ; \quad P_Y(y) = \sum_{x=0}^{+\infty} P(x,y)$$

Variable aléatoire 2-D et distribution théorique III

▷ **variables aléatoires continues**

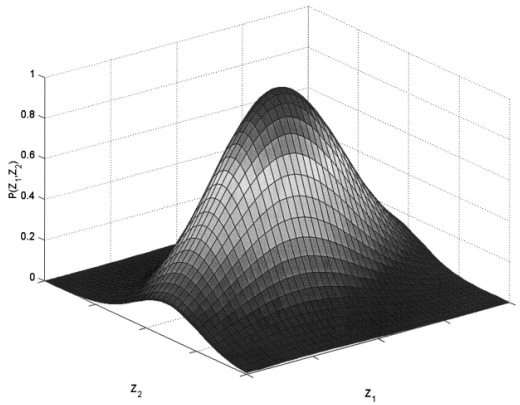
Si les variables aléatoires X et Y sont continues, la probabilité d'observer une valeur de X comprise dans l'intervalle $[x; x + \Delta x]$ et une valeur de Y comprise dans l'intervalle $[y; y + \Delta y]$

$$\text{prob}\left(x < X \leq x + \Delta x \text{ et } y < Y \leq y + \Delta y\right)$$

tend en général vers 0 pour $\Delta x \to 0$ et pour $\Delta y \to 0$. La notion de distribution de probabilité n'a plus de sens.
Par contre, la notion de *fonction de répartition* garde tout son sens

$$F\left(x, y\right) = \text{prob}\left(X \leq x \text{ et } Y \leq y\right)$$

Variable aléatoire 2-D et distribution théorique IV



Si $F(x,y)$ est dérivable, on définit

$$p_{XY}(x,y) = \frac{\partial^2 F}{\partial x \partial y}$$

comme étant la ***densité de probabilité conjointe*** des variables aléatoires X et Y. La quantité $p_{XY}(x,y)\,dx\,dy$ représente la probabilité d'observer une réalisation de X dans un intervalle de longueur dx centré sur x, ainsi qu'une réalisation de Y dans un intervalle de longueur dy centré sur y.

Variable aléatoire 2-D et distribution théorique V

a. Normalisation de la densité de probabilité

$$\int_{-\infty}^{+\infty}\int p_{XY}(x,y)\,dx\,dy = 1$$

b. En intégrant par rapport à l'une des variables aléatoires, on retrouve les densités de probabilité marginales :

$$p_X(x) = \int_{-\infty}^{+\infty} p_{XY}(x,y)\,dy \quad ; \quad p_Y(y) = \int_{-\infty}^{+\infty} p_{XY}(x,y)\,dx$$

Exemple : distribution uniforme continue à deux dimensions

Independance stochastique I

Par extension de la définition de l'indépendance stochastique, on dit que deux variables sont stochastiquement indépendantes si...

▷ **variables aléatoires discrètes**

$$P(X = x_i \text{ et } Y = y_i) = P(X = x_i)\, P(Y = y_i) \quad \rightarrow \quad P_{XY}(x_i, y_i) = P_X(x_i)\, P_Y(y_i)$$

▷ **variables aléatoires continues**

$$p_{XY}(x, y) = p_X(x)\, p_Y(y)$$

Independance stochastique II

Exemple : jet de deux dés : distribution de la somme et du résultat le plus élevé

→ les résultats des deux dés sont indépendants l'un de l'autre

	x_i						
y_j	1	2	3	4	5	6	Total
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Total	1/6	1/6	1/6	1/6	1/6	1/6	**1**

Independance stochastique III

soit $z = x + y$ et $v = \max(x, y)$

$\rightarrow$ les deux variables v et z ne sont pas indépendantes

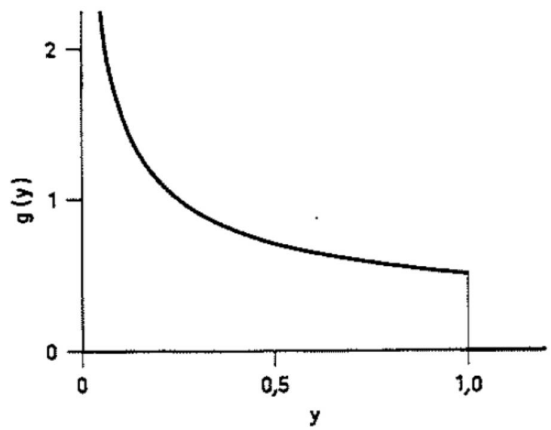
	v						
z	1	2	3	4	5	6	Total
2	1/36						1/36
3		2/36					2/36
4		1/36	2/36				3/36
5			2/36	2/36			4/36
6			1/36	2/36	2/36		5/36
7				2/36	2/36	2/36	6/36
8				1/36	2/36	2/36	5/36
9					2/36	2/36	4/36
10					1/36	2/36	3/36
11						2/36	2/36
12						1/36	1/36
Total	1/36	3/36	5/36	7/36	9/36	11/36	**1**

Transformations de variables aléatoires I

Une transformation d'une variable aléatoire est elle-même une variable aléatoire. Connaissant $p_X(x)$ et une transformation $y(x)$, quelle est la distribution $p_Y(y)$?

Soit $F_X(x)$ et $F_Y(y)$ les fonctions de répartition de X et Y.

$$F_X(x) = \text{prob}(X \leq x) \quad ; \quad F_Y(y) = \text{prob}(Y \leq y)$$



Transformations de variables aléatoires II

Si la transformation $Y(X)$ est monotone croissante,

$$F_Y(y) = \text{prob}(y(X) \leq y) = \text{prob}(X \leq x(y)) = F_X[x(y)]$$

Donc

$$p_Y(y) = \frac{d}{dy} F_X[x(y)] = X'(y) \frac{dF_X[x(y)]}{dx} = x'(y)\, p_X[x(y)]$$

- Si la transformation $Y(X)$ est monotone décroissante, $p_Y(y) = |x'(y)|\, p_X[x(y)]$
- Si la transformation est linéaire, la forme de la distribution est conservée (voir exemples)

Transformations de variables aléatoires III

Transformation linéaire d'une variable aléatoire uniforme

Soit $Y = a + bX$, et X une variable aléatoire à distribution uniforme sur $[0; 1]$

$$p_X(x) = \begin{cases} 1 & \text{si } 0 \leq x \leq 1 \\ 0 & \text{sinon} \end{cases}$$

On a

$$\begin{aligned} y = a + bx \quad &\rightarrow \quad y'(x) = b \\ x = \frac{y-a}{b} \quad &\rightarrow \quad x'(y) = \frac{1}{b} \end{aligned}$$

Donc

$$p_Y(y) = \begin{cases} \frac{1}{b} & \text{si } 0 \leq \frac{y-a}{b} \leq 1 \\ 0 & \text{sinon} \end{cases} \quad \rightarrow \quad p_Y(y) = \begin{cases} \frac{1}{b} & \text{si } a \leq y \leq a+b \\ 0 & \text{sinon} \end{cases}$$

Une transformation linéaire d'une variable uniforme reste donc uniforme

Transformations de variables aléatoires IV

Transformation linéaire d'une variable aléatoire normale
Soit $Y = a + bX$, et X une variable aléatoire à distribution normale de moyenne μ et d'écart-type σ

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Donc

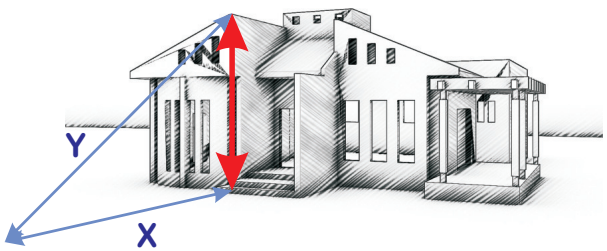
$$p_Y(y) = \frac{1}{b}\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\left(\frac{y-a}{b}-\mu\right)^2}{2\sigma^2}} = \frac{1}{(b\sigma)\sqrt{2\pi}} e^{\frac{-(y-(a+\mu b))^2}{2(b\sigma)^2}}$$

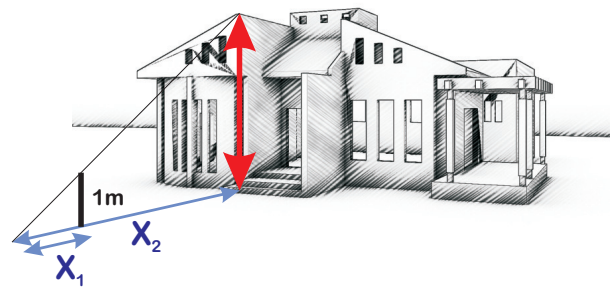
Une transformation linéaire d'une variable normale reste donc normale

La variable transformée est de moyenne $a + \mu b$ et écart-type $b\sigma$
Donc, si on considère $b = \sigma^{-1}$ et $a = -\mu/\sigma$, on réduit la variable X à une variable Y de moyenne nulle et d'écart-type unitaire.

Exercices I

1. On veut comparer deux méthodes différentes pour faire le relevé de la hauteur d'un bâtiment, à savoir (i) par une méthode géométrique basée sur la formule de Pythagore, (ii) par une méthode géométrique basée sur la formule de Thalès (voir figure ci-dessous).





Sachant que les cotes relevées sont $X = 8500\text{mm}$, $Y = 9450\text{mm}$, $X_1 = 1000\text{mm}$ et $X_2 = 4130\text{mm}$ et que les grandeurs mesurées peuvent s'apparenter à des variables normales d'écart-type $\sigma = 1mm$ (lié à la précision de la mesure), déterminez quelle méthode donne la mesure la plus fiable.

Exercices II

2. Déterminez la distribution du carré d'une variable uniforme continue sur l'intervalle $[0, 1]$ dont la densité est

$$p_X(x) = \begin{cases} 1 & \text{si } 0 \leq x \leq 1 \\ 0 & \text{sinon} \end{cases}$$

Validez vos observations à l'aide de Matlab.

3. Déterminez la distribution du carré d'une variable normale de moyenne unitaire ($\mu = 1$) et d'écart-type variable ($\sigma = 0.1,\ 1,\ 10$) dont la densité est

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

Validez vos observations à l'aide de Matlab.

4. Distribution de probabilité du rapport entre la contrainte et la résistance d'une barre de treilis

$$Safety = \frac{N}{A f_y}$$

Exercices III

où N, A et f_y sont des variables gaussiennes représentant l'effort axial, la section droite et la limite élastique.

Théorème Central Limite I

Somme de deux variables aléatoires continues indépendantes
Soit $Z = X + Y$. La densité de probabilité de z est donnée par

$$p_Z(z) = \int_{-\infty}^{+\infty} p_X(x)\, p_Y(z-x)\, dx = \int_{-\infty}^{+\infty} p_X(z-y)\, p_Y(y)\, dy$$

Démo :

$$\begin{aligned} F_Z(z) &= \text{prob}(Z \leq z) = \text{prob}(X + Y \leq z) \\ &= \int_{-\infty}^{+\infty} \text{prob}(x < X \leq x + dx)\, \text{prob}(X + Y \leq z \,|\, x < X \leq x + dx)\; dx \\ &= \int_{-\infty}^{+\infty} p_X(x)\, \text{prob}(Y \leq z - x)\; dx = \int_{-\infty}^{+\infty} p_X(x)\, F_Y(z-x)\; dx \end{aligned}$$

Théorème Central Limite II

Exemple 1 : somme de deux variables uniformes

$$p_X(x) = \begin{cases} 1 & \text{si } 0 \leq x \leq 1 \\ 0 & \text{sinon} \end{cases} \quad ; \quad p_Y(y) = \begin{cases} 1 & \text{si } 0 \leq y \leq 1 \\ 0 & \text{sinon} \end{cases}$$

donc

$$p_Z(z) = \int_{-\infty}^{+\infty} p_X(x)\, p_Y(z-x)\, dx = \int_{\Omega} dx$$

où Ω est le domaine englobant les valeurs de x telles que $0 \leq x \leq 1$ et $0 \leq z - x \leq 1$, càd $0 \leq x \leq 1$ et $z - 1 \leq x \leq z$.
Donc, $z \leq 1 \rightarrow 0 \leq x \leq z$ et $1 \leq z \leq 2 \rightarrow z - 1 \leq x \leq 1$. Et donc,

$$p_Z(z) = \begin{cases} z & \text{si } 0 \leq z \leq 1 \\ 2 - z & \text{si } 1 \leq z \leq 2 \\ 0 & \text{sinon} \end{cases}$$

Théorème Central Limite III

Exemple 2 : somme de deux variables normales

$$p_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{\frac{-(x-\mu_X)^2}{2\sigma_X^2}} \quad ; \quad p_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}} e^{\frac{-(y-\mu_Y)^2}{2\sigma_Y^2}}$$

donc

$$p_Z(z) = \frac{1}{\sqrt{2\pi}\sigma_Z} e^{-\frac{(z-\mu_Z)^2}{2\sigma_Z^2}}$$

avec $\mu_Z = \mu_X + \mu_Y$ et $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$

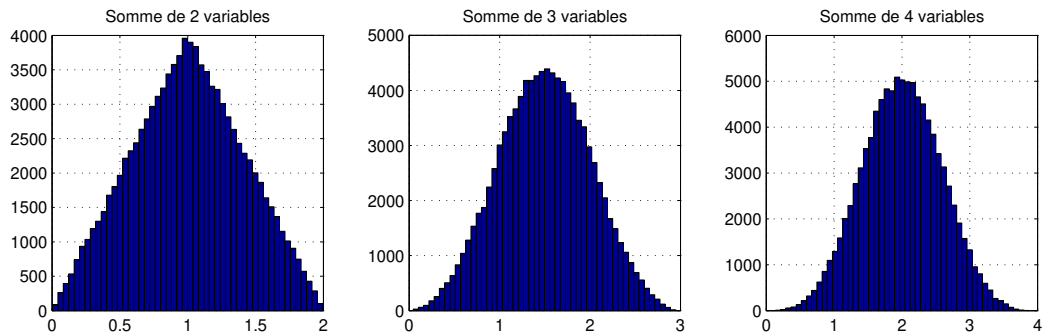
Une somme de variables aléatoires normales reste normale.

En fait, toute combinaison linéaire de variables normales reste normale

Théorème Central Limite IV

Théorème Central Limite

Toute somme de variables aléatoires indépendantes tend vers une variable aléatoire gaussienne lorsque le nombre de termes dans la somme tend vers l'infini.



Exemple : distribution d'une somme de n variables aléatoires (n=1,2,3,...)

Espérance mathématique I

On appelle ***espérance mathématique*** d'une variable aléatoire X la grandeur

$$E[X] = \int_{-\infty}^{+\infty} x\, p_X(x)\, dx$$

nb : analogie avec la définition de la moyenne $\overline{x} = \frac{1}{n} \sum x_i n_i$.

Soit $Y(X)$ une fonction de X. L'espérance mathématique de Y est donc obtenue par

$$E[Y(X)] = \int_{-\infty}^{+\infty} y\, p_Y(y)\, dy = \int_{-\infty}^{+\infty} y\, X'(y)\, p_X(X(y))\, dy = \int_{-\infty}^{+\infty} Y(x)\, p_X(x)\, dx$$

Espérance mathématique II

a. Transformation linéaire de variables aléatoires

$$E[a + bX] = a + b\,E[X]$$

b. Espérance mathématique d'une somme ou différence

$$E[X \pm Y] = E[X] \pm E[Y]$$

c. Espérance mathématique d'un produit (si les variables sont stochastiquement indépendantes)

$$E[XY] = E[X]\;E[Y]$$

Paramètres des distributions à une dimension I

La moyenne d'une distribution de probabilité est l'espérance mathématique de la variable aléatoire X correspondante

$$m = E[X]$$

La médiane est telle que

$$F(\tilde{m}) = \frac{1}{2}$$

Le mode correspond au maximum relatif de la densité de probabilité

La variance d'une distribution de probabilité est l'espérance mathématique du carré de la variable aléatoire X déjaugée de sa moyenne

$$\mu_2 = \sigma^2 = E\left[(X-m)^2\right] = \int_{-\infty}^{+\infty} (x-m)^2 \, p_X(x)\, dx$$

L'écart-type est la racine carrée de la variance. Le coefficient de variation est le rapport entre l'écart-type et la moyenne.

Paramètres des distributions à une dimension II

Les moments centrés sont définis par

$$\mu_k = E\left[(X-m)^k\right] = \int_{-\infty}^{+\infty} (x-m)^k \; p_X(x)\, dx$$

Paramètres des distributions à une dimension III

Propriétés de la moyenne et de la variance
Soit $Y = a + bX$

$$\begin{aligned} m_Y &= a + b m_X \\ \sigma_Y^2 &= b^2 \sigma_X^2 \end{aligned}$$

Soit $S = X_1 \pm X_2$ (avec X_1 et X_2 stochastiquement indépendants)

$$\begin{aligned} m_S &= m_{X_1} \pm m_{X_2} \\ \sigma_S^2 &= \sigma_{X_1}^2 + \sigma_{X_2}^2 \end{aligned}$$

Paramètres des distributions à une dimension IV

L'inégalité de Bienaymé-Tchebychev

Quelle que soit la distribution de probabilité de la variable X, de moyenne m et d'écart-type σ, et quelle que soit la quantité positive k, on peut démontrer que

$$P\left(|X - m| \geq k\sigma\right) \leq \frac{1}{k^2}$$

$\rightarrow$ la proportion d'individus s'écartant de la moyenne de plus de k fois l'écart-type est toujours inférieure $1/k^2$.

Démo : soit c une constant positive et Ω le domaine dans lequel $(X - m)^2 \geq c$.

Paramètres des distributions à une dimension V

$$\begin{aligned}\sigma^2 &= E\left[(X-m)^2\right] = \int_{-\infty}^{+\infty} (X-m)^k \, p_X(x)\, dx \geq \int_{\Omega} c\, p_X(x)\, dx \\ &= c\int_{\Omega} p_X(x)\, dx = c\,\text{prob}\left[(X-m)^2 \geq c\right]\end{aligned}$$

En posant $c = k^2\sigma^2$, on obtient

$$\sigma^2 \geq k^2\sigma^2\,\text{prob}\left[(X-m)^2 \geq k^2\sigma^2\right]$$

(CQFD)

Exercices I

1. Calculer la moyenne et l'écart-type d'une distribution continue sur $[a; b]$.

2. Un industriel fabrique des axes et des coussinets de 20 mm de diamètre moyen. Supposons que le diamètre extérieur A des axes soit une variable aléatoire normale d'écart-type $\sigma_A = 0.05$mm et que le diamètre intérieur C des coussinets soit une variable normale d'écart-type $\sigma_C = 0.07$mm. Dans quelle proportion des cas les axes ne pourront-ils pas pénétrer dans les coussinets ? Que deviendrait cette proportion si on prévoyait, en moyenne, un jeu de 0.1mm, en donnant aux axes un diamètre moyen de 19.9mm ? Quel jeu faudrait-il prévoir pour réduire cette probabilité de défaillance à 1% ?

3. Le revenu brut annuel des personnes de la population active de la ville de Harelbeke peut être assimilé à une variable aléatoire normale de moyenne 45100€ et d'écart-type 12600€. Quelle fraction de la population active possède un revenu supérieur à 65000€ ?

4. *[Jui2013]* Un bâtiment non-mitoyen, de forme cubique de côté $c = 6.5$ mètres, est isolé par des matériaux isolants sur les façades ainsi que sur sa toiture plate. Pour des raisons techniques, ces isolants sont différents. Pour les façades, la transmission thermique est $U_1 = 1.15\ W/m^2K$, avec un coefficient

Exercices II

de variation de 10%, alors qu'elle est de $U_2 = 1.8\ W/m^2K$ pour la toiture, avec un coefficient de variation de 10% également (ces variables ont des distributions normales). Pour une paroi de transmission U, la déperdition calorifique par unité de surface s'exprime par

$$W = U \cdot \Delta T$$

où $\Delta T = T_i - T_o = 30°C$ est la différence de température intérieur-extérieur.
∘ quelle est la densité de probabilité de la demande énergétique de ce bâtiment ?
∘ quelle est la probabilité qu'une chaudière de $P = 10kW$ soit insuffisante pour garder le bâtiment dans un régime de chauffage stationnaire ?

Réponses :
1. $\mu = \frac{a+b}{2}$; $\sigma = \frac{|b-a|}{2\sqrt{3}}$.
2. (i) $p = 0.5$; (ii) $p = 12.3\%$; (iii) 0.2mm.
3. 5.7%
4. $\mu_e = 8112$W, $\sigma_e = 626$W ; probabilité d'insuffisance = 0.13%

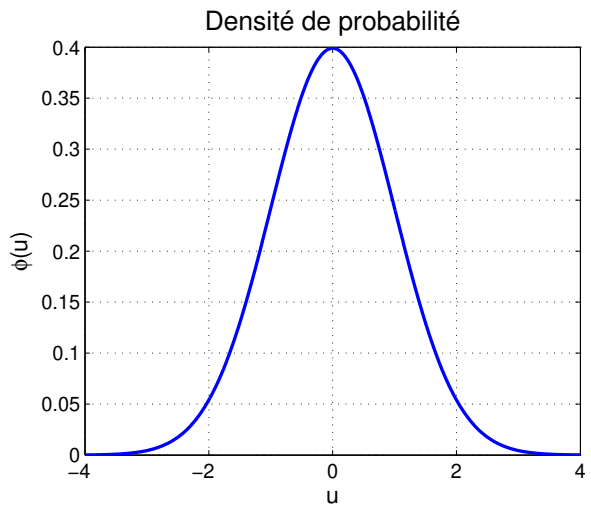
Distribution Normale I

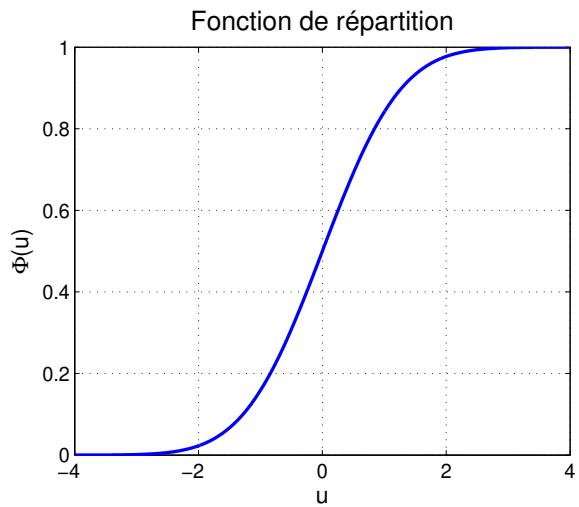
Une variable X possède une ***distribution normale réduite*** lorsque sa densité de probabilité s'écrit

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

La fonction de répartition correspondante est

$$\Phi(u) = \int_{-\infty}^{u} \phi(x)\, dx$$





Distribution Normale II

Matlab : densité de probabilité

On peut utiliser la fonction `Y=PDF('norm',u,0,1)` pour estimer la densité de probabilité d'une variable normale réduite.

Matlab : fonction de répartition

ERF Error function.

`Y = ERF(X)` is the error function for each element of X. X must be real. The error function is defined as : erf(x) = 2/sqrt(pi) * integral from 0 to x of exp(-t^2) dt.

$$\begin{aligned}\Phi(u) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{u/\sqrt{2}} e^{-t^2}dt \\ &= \frac{1}{2}+\frac{1}{\sqrt{\pi}}\int_{0}^{u/\sqrt{2}} e^{-t^2}dt = \frac{1+\operatorname{erf}\left(u/\sqrt{2}\right)}{2}\end{aligned}$$

Distribution Normale III

Propriétés

$$\begin{aligned} m_U &= E[U] = 0 \\ \sigma_U^2 &= E\left[(U - m_U)^2\right] = 1 \\ \gamma_3 &= 0; \quad \gamma_4 = 3; \end{aligned}$$

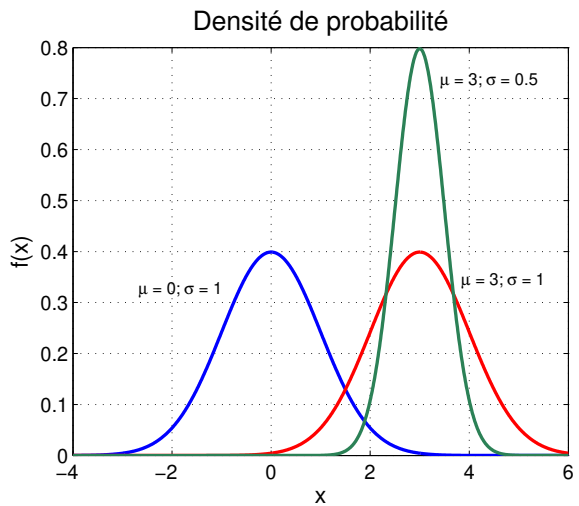
u	$P(U \leq u) = \Phi(u)$	$P(\lvert U\rvert \leq u) = 2\Phi(u) - 1$	$P(\lvert U\rvert > u) = 2[1 - \Phi(u)]$
1	0.8413	0.6826	0.3174
1.645	0.95	0.90	0.10
1.960	0.975	0.95	0.05
2.326	0.99	0.98	0.02
2.576	0.995	0.99	0.01
3.090	0.999	0.998	0.002
3.291	0.9995	0.999	0.001

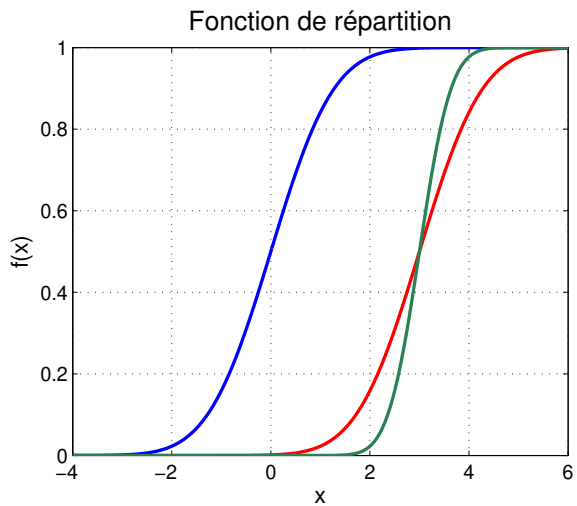
Distribution Normale IV

Une variable X possède une ***distribution normale*** lorsque sa densité de probabilité s'écrit

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

On peut vérifier que $E[X] = \mu$ et $E\left[(X-\mu)^2\right] = \sigma^2$





Matlab : densité de probabilité
On peut utiliser la fonction `Y=PDF('norm',x,moy,sigma)` pour estimer la densité de probabilité d'une variable normale de moyenne `moy` et d'écart-type `sigma`.

Propriétés

a. Si X est une variable aléatoire normale de moyenne m_x et d'écart-type σ_x, toute transformation linéaire

$$Y = a + bX$$

possède également une distribution normale de moyenne et d'écart-type

$$m_y = a + b\, m_x \qquad \text{et} \qquad \sigma_y = |b|\, \sigma_x$$

b. En particulier, si X est une variable aléatoire normale de moyenne m_x et d'écart-type σ_x, la variable réduite

$$U = \frac{X - m_x}{\sigma_x}$$

est une variable normale réduite (moyenne nulle et variance unitaire).

c. La somme de deux ou plusieurs variables normales est également une distribution normale dont la moyenne est la somme des moyenne et dont la variance est la somme des variances.

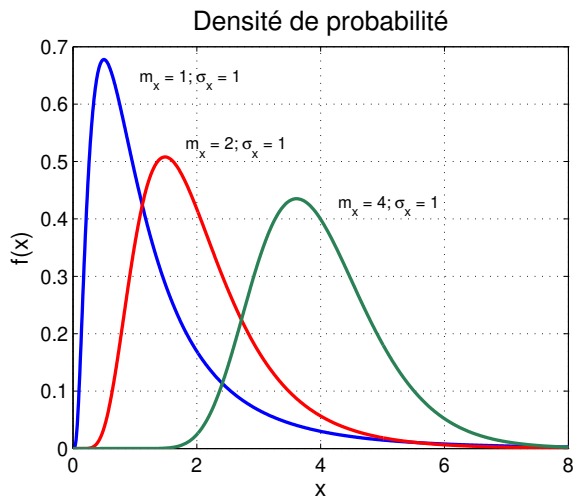
d. Une proprité importante des variables aléatoires est celle du théorème central limite.

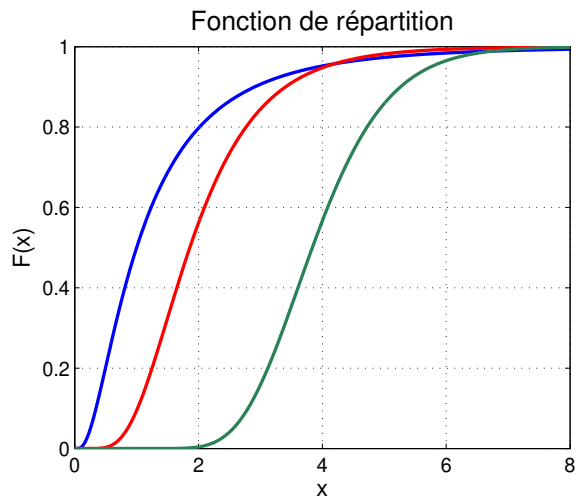
Distribution Lognormale I

Une variable X possède une ***distribution log-normale*** lorsque son logarithme $Y = \ln X$ a une distribution normale. L'intervalle de variation d'une telle variable s'étend de 0 à l'infini. On démontre que sa densité de probabilité s'exprime par

$$p_X(X) = \frac{1}{\sqrt{2\pi} X \sigma_y} e^{-\frac{1}{2}\left[\frac{(\ln X - m_y)}{\sigma_y}\right]^2}$$

où m_y et σ_y sont les deux paramètres de la distribution normale de la variable Y.





Distribution Lognormale II

MATLAB : densité de probabilité

On peut utiliser la fonction `Y=PDF('logn',X,my,sy)` pour estimer la densité de probabilité d'une variable log-normale dont les moyenne et écart-type de la distribution normale sous-jacente sont `my` et `sy`.

Propriétés

$$m_x = E[X] = \int_0^{+\infty} x\, p_X(x)\, dx = e^{m_y + \sigma_y^2/2}$$

$$\sigma_x^2 = E\left[(X - m_x)^2\right] = e^{2m_y + \sigma_y^2}\left(e^{\sigma_y^2} - 1\right)$$

a. Si X est une variable lognormale, toute transformation de puissance du type

$$Z = a\, X^b$$

est également une variable lognormale.

Distribution Lognormale III

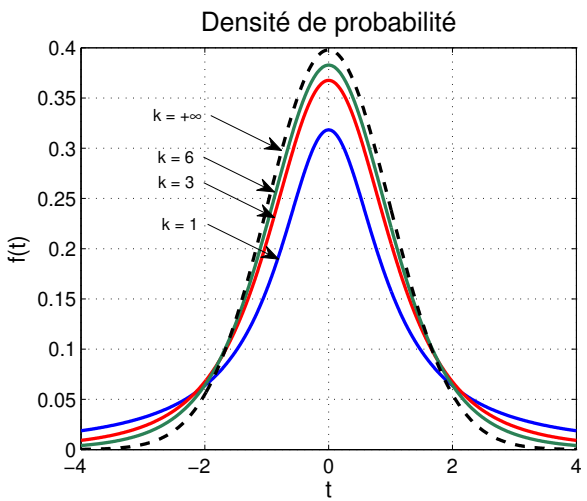
Utilité : on utilise souvent la distribution log-normale pour représenter des grandeurs qui sont nécessairement positives (longueurs, masses, dimensions, etc.)

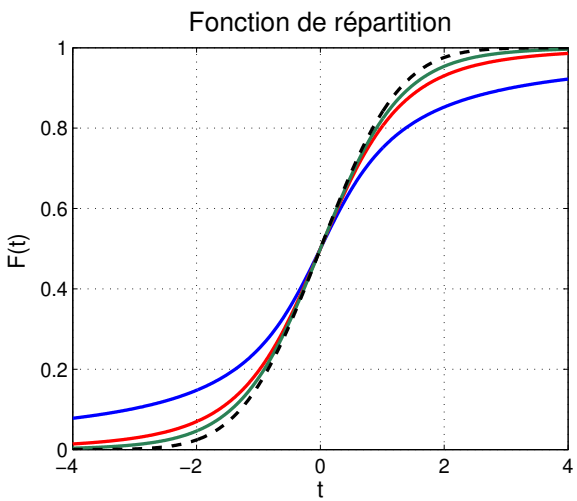
Distribution t de Student I

Les ***distributions t de Student*** sont caractérisées par une densité de probabilité qui s'exprime par

$$p(t) = c\left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

dans laquelle t varie continûment de $-\infty$ à $+\infty$, k est une constante entière positive appelée ***nombre de degrés de liberté*** et c une constant entière positive (dépendant de k) assurant la propriété d'unicité.





Distribution t de Student II

MATLAB : densité de probabilité
On peut utiliser la fonction `Y=PDF('t',t,k)` pour estimer la densité de probabilité d'une variable à distribution t dont le nombre de degrés de liberté est `k`.

Propriétés

$$\begin{aligned} m &= 0 \\ \sigma^2 &= \frac{k}{k-2} \xrightarrow[k\gg]{} 1 \\ \gamma_4 &= 3\frac{k-2}{k-4} \xrightarrow[k\gg]{} 3 \end{aligned}$$

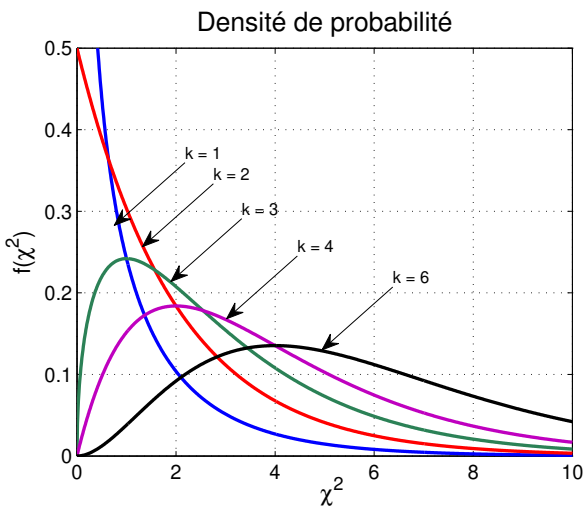
a. La distribution t est asymptotiquement normale (réduite) lorsque $k \to +\infty$.

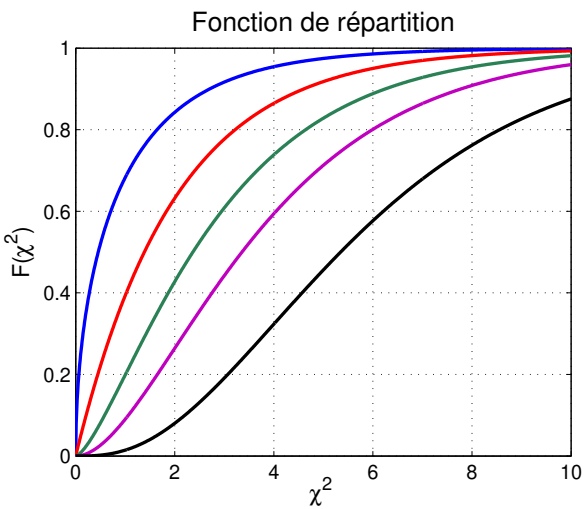
Distribution χ^2 de Pearson I

Les ***distributions χ^2 de Pearson*** sont caractérisées par une densité de probabilité qui s'exprime par

$$p_X(x) = c\, x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad \text{ou} \quad p_{\chi^2}(\chi^2) = c\chi^{k-2} e^{-\frac{\chi^2}{2}}$$

dans laquelle la variable X varie continûment de 0 à $+\infty$, k est une constante entière positive appelée ***nombre de degrés de liberté*** et c une constant entière positive (dépendant de k) assurant la propriété d'unicité.





Distribution χ^2 de Pearson II

MATLAB : densité de probabilité
On peut utiliser la fonction `Y=PDF('chi2',X2,k)` pour estimer la densité de probabilité d'une variable χ^2 dont le nombre de degrés de liberté est `k`.

Propriétés

$$\begin{aligned} m &= k \\ \sigma^2 &= 2k \\ \gamma_3 &= 8/k \\ \gamma_4 &= 3 + 12/k \end{aligned}$$

a. Distribution en "i" lorsque $k = 1$ ou $k = 2$. Distribution "en cloche" lorsque $k \geq 3$.

Distribution χ^2 de Pearson III

b. Les distributions χ^2 sont additives : une somme de deux distributions χ^2 de degrés de libertés k_1 et k_2 est une autre distribution χ^2 de degré de liberté $k_1 + k_2$.

c. Les distributions χ^2 sont donc asymptotiquement normales.

d. La densité de probabilité du carré X d'une variable normale réduite U, i.e. $X = U^2$, est une variable χ^2 de paramètre $k = 1$. Une variable χ^2 de paramètre $k = 1$ est donc le carré d'une variable normale réduite.

e. Si deux variables U et X sont stochastiquement indépendantes et si elles ont respectivement une distribution normale réduite et une distribution χ^2 de paramètre k, alors, la variable aléatoire

$$t = \frac{U}{\sqrt{\frac{X}{k}}}$$

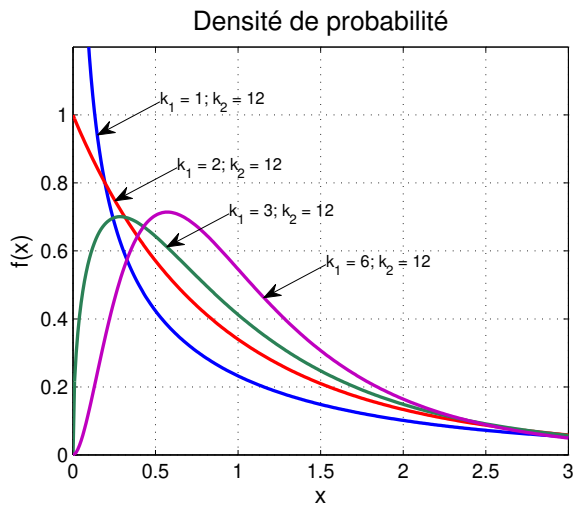
possède une distribution t de Student de paramètre k.

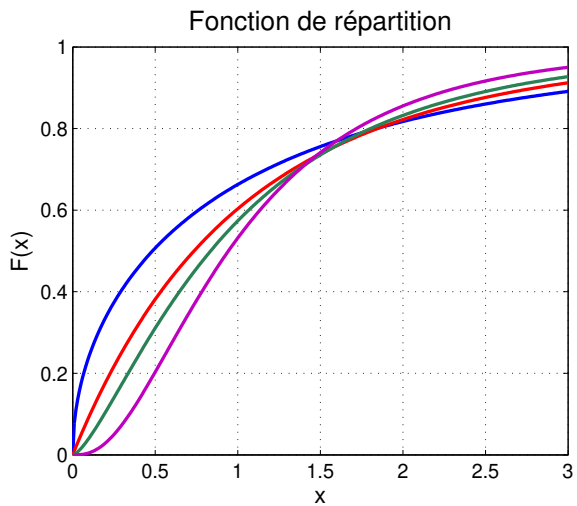
Distribution F de Fisher-Snedecor I

Les ***distributions F de Fisher-Snedecor*** sont telles que

$$p(x) = c\, x^{\frac{k}{2}-1} (k_1 x + k_2)^{-\frac{k_1+k_2}{2}}$$

dans laquelle la variable X varie continûment de 0 à $+\infty$, k_1 et k_2 sont deux constantes entières positives appelée ***nombre de degrés de liberté*** et c une constant entière positive (dépendant de k) assurant la propriété d'unicité.





Distribution F de Fisher-Snedecor II

Matlab : densité de probabilité
On peut utiliser la fonction `Y=PDF('F',X,k1,k2)` pour estimer la densité de probabilité d'une variable χ^2 dont les degrés de liberté sont `k1` et `k2`.

Propriétés

$$\begin{aligned} m &= \frac{k_2}{k_2-2} \\ \sigma^2 &= 2k_2^2 \frac{k_1+k_2-2}{k_1(k_2-2)^2(k_2-4)} \end{aligned}$$

a. $m \to 1$ et $\sigma \to 0$ lorsque $(k_1, k_2) \to +\infty$.

b. $\gamma_3 \to 0$ et $\gamma_4 \to 3$ lorsque $(k_1, k_2) \to +\infty$.

Distribution F de Fisher-Snedecor III

c. Si deux variables Y_1 et Y_2 sont stochastiquement indépendantes et si elles ont respectivement des distributions χ^2 de paramètres k_1 et k_2, alors, la variable aléatoire

$$X = \frac{\frac{Y_1}{k_1}}{\frac{Y_2}{k_2}}$$

est une variable de Fisher-Snedecor de paramètres k_1 et k_2.

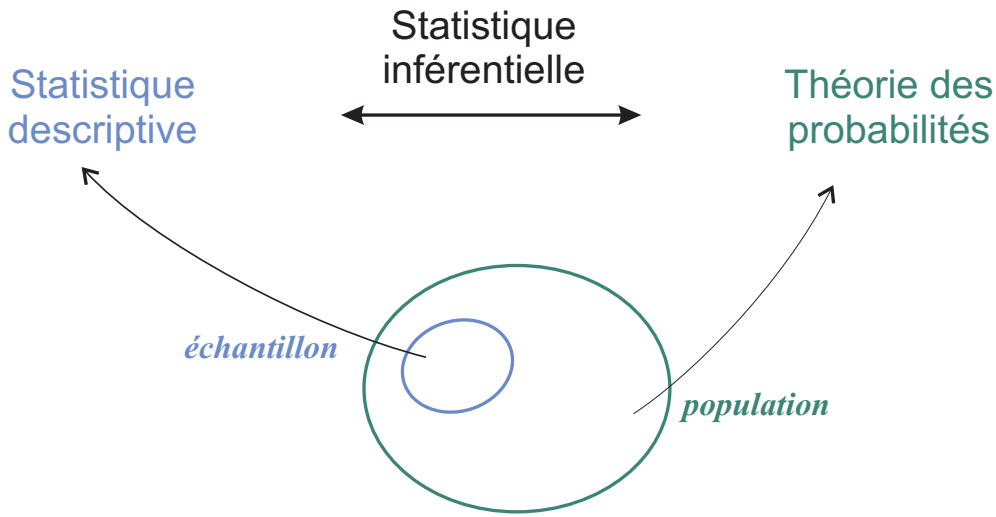
d. Si X est une variable F de paramètres (k_1, k_2), alors $1/X$ est une variable F de paramètres (k_2, k_1).

Partie IV : Statistique Inférentielle

Acquis d'apprentissage :

- faire le lien entre statistiques descriptives et lois de probabilité, échantillon et population
- prendre conscience que tout ce que l'on observe au quotidien n'est jamais qu'une réalisation d'une expérience aléatoire
- apprendre à jauger le faculté de pouvoir extrapoler vers la population complète, les résultats d'un expérience sur un fraction de celle-ci
- connaitre les principes généraux relatifs aux tests d'hypothèse

Contexte & Objectifs



◦ distribution d'échantillonage :
population supposée connue statistiquement → distribution de grandeurs statistiques (contexte : échantillonnage *aléatoire simple*)

◦ tests d'hypothèse
échantillon sondé / mesuré → population (avec intervalle de confiance)

exemple : âge dans la classe

Quelques distributions d'échantillonnage I

▷ Distribution d'échantillonnage de la moyenne

Supposons que, dans une population infinie, on ait prélevé par échantillonage aléatoire simple un *premier* échantillon de n observations, et que l'on en calcule la moyenne correspondante

$$x_1, x_2, ..., x_n \quad \rightarrow \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Si on prélève un *second* échantillon, on obtient a priori un résultat différent

$$x'_1, x'_2, ..., x'_n \quad \rightarrow \quad \overline{x}' = \frac{1}{n}\sum_{i=1}^{n} x'_i.$$

A chaque échantillon choisi correspondra une autre moyenne. Toutes ces moyennes observées

$$\overline{x}, \overline{x}', \overline{x}'', \overline{x}''', ...$$

sont les réalisations d'une variable aléatoire $\overline{X}$ qui est fonction des n observations X_1, X_2, ... X_n, telle que

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Quelques distributions d'échantillonnage II

Tout comme les variables aléatoires X_1, X_2, ... X_n, la nouvelle variable $\overline{X}$ possède une distribution de probabilité. C'est la ***distribution d'échantillonnage de la moyenne***.

Cadre : on suppose que les variables X_1, X_2, ... X_n sont identiquement distribuées et stochastiquement indépendantes (échantillonnage aléatoire simple) . Leurs moyennes sont notées m et leurs variances sont notées σ^2.

Peut on déterminer la distribution de probabilité de $\overline{X}$?

Quelques distributions d'échantillonnage III

Moyenne de la distribution de $\overline{X}$

$$E\left[\overline{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = m$$

Variance de la distribution de $\overline{X}$

$$\text{var}\left[\overline{X}\right] = \text{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}\left[X_i\right] = \frac{\sigma^2}{n}$$

La moyenne et l'écart-type (ou l'erreur-standard) **de la moyenne** d'un échantillon aléatoire simple sont donc donnés par

$$E\left[\overline{X}\right] = m \quad \text{et} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Quelques distributions d'échantillonnage IV

Exemple : les X_i sont distribués selon une distribution uniforme sur $[a, b] = [-1, 1]$; on a pu collecter $n = 2$ échantillons $\rightarrow (X_1, X_2)$

X_1	X_2	$\frac{X_1+X_2}{2}$
-0.802	0.625	-0.0885
0.792	0.847	0.8195
0.124	-0.912	-0.3940
-0.187	-0.632	-0.4095
-0.654	0.812	0.0790
$\vdots$	$\vdots$	$\vdots$

$N = 100 \quad \rightarrow \quad E\left[\overline{X}\right] = 0.0064, \sigma_{\overline{X}} = 0.4061$
$N = 1000 \quad \rightarrow \quad E\left[\overline{X}\right] = -0.0011, \sigma_{\overline{X}} = 0.4025$
$N = 10000 \quad \rightarrow \quad E\left[\overline{X}\right] = -0.0042, \sigma_{\overline{X}} = 0.4055$

Théoriquement : $E\left[\overline{X}\right] = 0, \sigma_{\overline{X}} = \frac{1}{\sqrt{n}} \frac{b-a}{2\sqrt{3}} = 0.4082$

Quelques distributions d'échantillonnage V

Exemple : les X_i sont distribués selon une distribution uniforme sur $[a, b] = [-1, 1]$; on a pu collecter $n = 4$ échantillons $\rightarrow (X_1, X_2, X_3, X_4)$

X_1	X_2	X_3	X_4	$\frac{X_1+X_2+X_3+X_4}{4}$
-0.802	0.625	-0.147	0.762	0.1095
0.792	0.847	0.986	-0.345	0.5700
0.124	-0.912	0.467	-0.951	-0.318
-0.187	-0.632	-0.356	0.843	-0.0830
-0.654	0.812	-0.845	0.476	-0.0527
$\vdots$	$\vdots$			$\vdots$

$N = 100 \quad \rightarrow \quad E\left[\overline{X}\right] = 0.0089, \sigma_{\overline{X}} = 0.3017$

$N = 1000 \quad \rightarrow \quad E\left[\overline{X}\right] = -0.0122, \sigma_{\overline{X}} = 0.2940$

$N = 10000 \quad \rightarrow \quad E\left[\overline{X}\right] = -0.0024, \sigma_{\overline{X}} = 0.2874$

Théoriquement : $E\left[\overline{X}\right] = 0, \sigma_{\overline{X}} = \frac{1}{\sqrt{n}} \frac{b-a}{2\sqrt{3}} = 0.2887$

Quelques distributions d'échantillonnage VI

▷ **Sur la forme de la distribution de la moyenne $\overline{X}$**

a. Si la population-parent est distribuée normalement, la moyenne observée est également distribuée normalement (conservation de la normalité par transformation linéaire)

b. Si la population-parent n'est pas distribuée normalement, la moyenne observée tend à devenir normale lorsque la taille d'échantillon n est grande (théorème central limite).

Quelques distributions d'échantillonnage VII

▷ Distribution d'échantillonnage de la variance

Supposons que, dans une population infinie, on ait prélevé par échantillonage aléatoire simple un *premier* échantillon de n observations, et que l'on en calcule la variance correspondante

$$x_1, x_2, ..., x_n \quad \rightarrow \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 .$$

Si on prélève un *second* échantillon, on obtient a priori un résultat différent

$$x'_1, x'_2, ..., x'_n \quad \rightarrow \quad s'^2 = \frac{1}{n}\sum_{i=1}^{n}(x'_i - \overline{x}')^2 .$$

A chaque échantillon choisi correspondra une autre variance. Toutes ces variances observées

$$s^2, s'^2, s''^2, s'''^2, ...$$

sont les réalisations d'une variable aléatoire S^2 qui est fonction des n observations X_1, X_2, ... X_n, telle que

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

Quelques distributions d'échantillonnage VIII

Tout comme les variables aléatoires X_1, X_2, ... X_n, la nouvelle variable S^2 possède une distribution de probabilité. C'est la ***distribution d'échantillonnage de la variance***.

Cadre : on suppose que les variables X_1, X_2, ... X_n sont identiquement distribuées et stochastiquement indépendantes (échantillonnage aléatoire simple) . Leurs moyennes sont notées m et leurs variances sont notées σ^2.

Peut on déterminer la distribution de probabilité de S^2 ?

Quelques distributions d'échantillonnage IX

Moyenne de la distribution de S^2

$$\begin{aligned} E\left[S^2\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2-\left(\overline{X}-m\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}E\left[(X_i-m)^2\right]-E\left[\left(\overline{X}-m\right)^2\right] \\ &= \sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2 \end{aligned}$$

Variance de la distribution de S^2

Dans le cas d'une distribution de population normale ($\gamma_3 = 0\,;\ \gamma_4 = 3$), on démontre que

$$\operatorname{var}\left[S^2\right] = E\left[\left(S^2-\frac{n-1}{n}\sigma^2\right)^2\right] = 2\frac{n-1}{n^2}\sigma^4$$

De façon générale, pour une population arbitraire,

$$\operatorname{var}\left[S^2\right] = \frac{n-1}{n^2}\frac{\gamma_4(n-1)-n+3}{n}\sigma^4$$

Quelques distributions d'échantillonnage X

La moyenne et l'écart-type (ou l'erreur-standard) **de la variance** d'un échantillon aléatoire simple sont donc donnés par

$$E\left[S^2\right] = \frac{n-1}{n}\sigma^2 \quad \text{et} \quad \sigma_{S^2} = \frac{\sqrt{2\,(n-1)}}{n}\sigma^2$$

Quelques distributions d'échantillonnage XI

Exemple : les X_i sont distribués selon une distribution uniforme sur $[a, b] = [-1, 1]$; on a pu collecter $n = 2$ échantillons $\rightarrow (X_1, X_2)$

X_1	X_2	$\frac{X_1+X_2}{2}$	var (X_1, X_2)
-0.802	0.625	-0.0885	0.5091
0.792	0.847	0.8195	0.0008
0.124	-0.912	-0.3940	0.2683
-0.187	-0.632	-0.4095	0.0495
-0.654	0.812	0.0790	0.5373
$\vdots$	$\vdots$	$\vdots$	

$N = 100 \quad \rightarrow \quad E\left[S^2\right] = 0.1567, \sigma_{S^2} = 0.1899$
$N = 1000 \quad \rightarrow \quad E\left[S^2\right] = 0.1718, \sigma_{S^2} = 0.1988$
$N = 10000 \quad \rightarrow \quad E\left[S^2\right] = 0.1634, \sigma_{S^2} = 0.1949$

Théoriquement : $E\left[S^2\right] = \frac{2-1}{2}\frac{(b-a)^2}{12} = 0.1667$, $\sigma_{S^2} =$??? (dépend de γ_4)

Quelques distributions d'échantillonnage XII

Exemple : les X_i sont distribués selon une distribution uniforme sur $[a, b] = [-1, 1]$; on a pu collecter $n = 4$ échantillons $\rightarrow (X_1, X_2, X_3, X_4)$

X_1	X_2	X_3	X_4	$\frac{X_1+X_2+X_3+X_4}{4}$	var (X_1, X_2, X_3, X_4)
-0.802	0.625	-0.147	0.762	0.1095	0.3970
0.792	0.847	0.986	-0.345	0.5700	0.2841
0.124	-0.912	0.467	-0.951	-0.318	0.3913
-0.187	-0.632	-0.356	0.843	-0.0830	0.3111
-0.654	0.812	-0.845	0.476	-0.0527	0.5041
$\vdots$	$\vdots$			$\vdots$	$\vdots$

$N = 100 \quad \rightarrow \quad E\left[S^2\right] = 0.2781, \sigma_{S^2} = 0.1668$
$N = 1000 \quad \rightarrow \quad E\left[S^2\right] = 0.2568, \sigma_{S^2} = 0.1469$
$N = 10000 \quad \rightarrow \quad E\left[S^2\right] = 0.2530, \sigma_{S^2} = 0.1526$

Théoriquement : $E\left[S^2\right] = \frac{4-1}{4}\frac{(b-a)^2}{12} = 0.25$, $\sigma_{S^2} =$??? (dépend de γ_4,)

Quelques distributions d'échantillonnage XIII

▷ **Sur la forme de la distribution de la variance S^2**

a. Si la population-parent est distribuée normalement, la variance observée S^2 est telle que

$$\frac{nS^2}{\sigma^2}$$

possède une distribution χ^2 de paramètre $k = n - 1$.
[admis sans démonstration]

b. Si la population-parent n'est pas distribuée normalement, la variance observée S^2 tend à devenir normale lorsque la taille d'échantillon n est grande.

c. Si la population-parent est distribuée normalement, la variable aléatoire

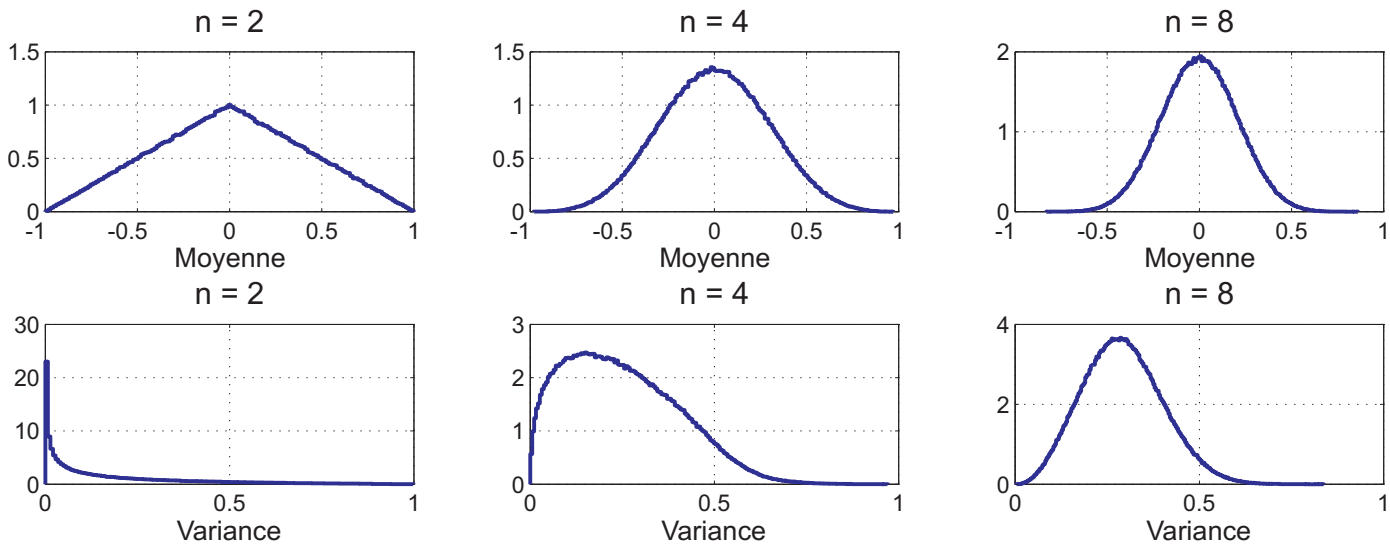
$$(\overline{X} - m) / \sqrt{S^2/(n-1)}$$

possède une distribution t de Student avec un paramètre $k = n - 1$.
[quotient d'une variable normale réduite et de la racine d'une variable χ^2]

Quelques distributions d'échantillonnage XIV

Exemple : les X_i sont distribués selon une distribution uniforme sur $[a, b] = [-1, 1]$

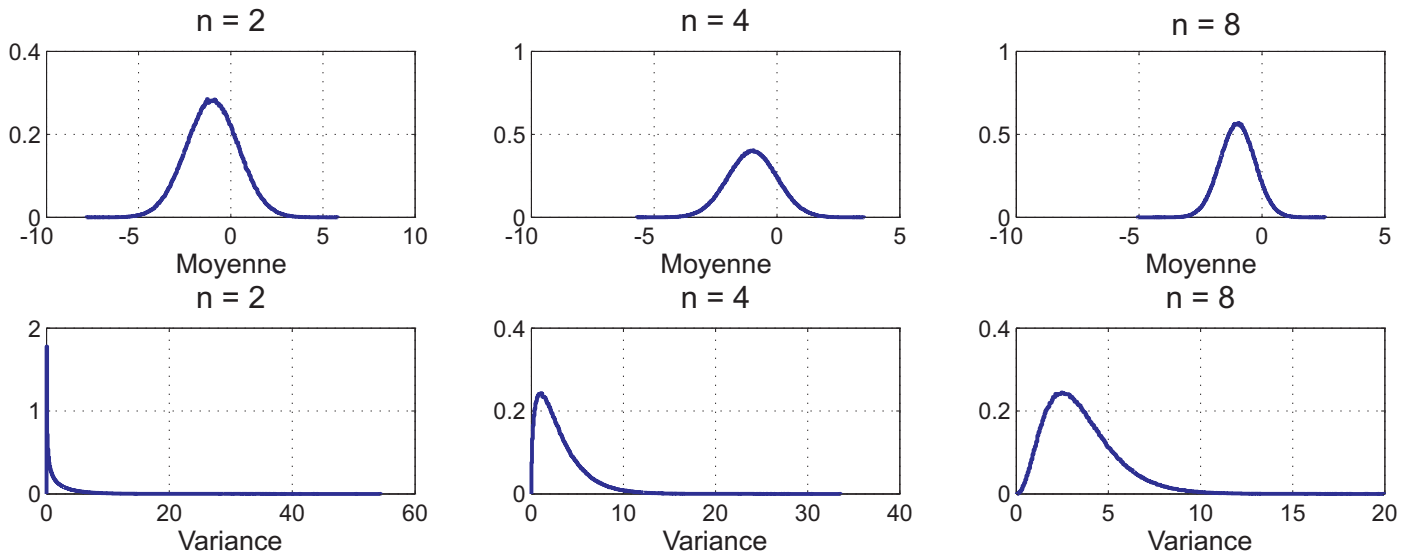
Distributions d'échantillonnage de la moyenne et de la variance des X_i pour $n = 2, 4, 8$.



Quelques distributions d'échantillonnage XV

Exemple : les X_i sont distribués selon une distribution normale réduite

Distributions d'échantillonnage de la moyenne et de la variance des X_i pour $n = 2, 4, 8$.



Quelques distributions d'échantillonnage XVI

Autres distributions

rem : on parle également d'autres types de distributions, celles :

- de la médiane

$$E\left[\tilde{X}\right] = \tilde{m} \quad ; \quad \text{var}\left[\tilde{X}\right] \simeq \frac{1}{4nf^2\left(\tilde{m}\right)}$$

donc, pour une population normale, $\text{var}\left[\tilde{X}\right] \simeq \pi\sigma^2/\left(2n\right)$.

- de l'amplitude
- du mode
- des valeurs extrémales
- ...

Principes généraux I

▷ La connaissance théorique des distributions d'échantillonnage

A tout paramètre γ d'une population quelconque, on peut associer une série (infinie) de valeurs observées g, g', g'', ... qui pourraient être calculées à partir d'échantillons successifs de même effectif, prélevés indépendamment les uns des autres, dans des conditions identiques. Ces valeurs sont des valeurs observées d'une variable aléatoire G et cette variable est une fonction des différentes variables aléatoires qui peuvent être associées à chacun des individus des échantillons

$$G = G(X_1, X_2, ...X_n)$$

En supposant que l'échantillon est aléatoire et simple, on peut calculer la moyenne et la variance de G

$$m_G = E[G] \quad ; \quad \sigma_G^2 = E\left[(G - m_G)^2\right]$$

et, si possible, la distribution complète (***distribution d'échantillonnage***).

Principes généraux II

En fonction de considérations théoriques, les distributions d'échantillonnage peuvent être connues :

(i) soit d'une manière exacte,

exemple : si la population-parent est normale, la distribution de la moyenne d'échantillon est normale, celle de la variance est χ^2, celle du rapport moyenne/variance est t.

(ii) soit de façon approchée (p.ex. normalité asymptotique).

exemple : si la taille de l'échantillon est "suffisamment grande", on peut approcher les distributions de la moyenne, de la variance ou de la médiane par des distributions normales. Il suffit donc dans ce cas d'estimer la moyenne et la variance de la distribution.

Principes généraux III

Exemple : les X_i sont distribués selon une distribution uniforme sur $[a, b] = [-1, 1]$.

(i) Quelle est la probabilité que la moyenne de $n = 4$ nombres tirés au hasard soit supérieure à 0.5 ?

(ii) Quelle est la probabilité que la variance de $n = 8$ nombres tirés au hasard soit supérieure à 0.5 ?
(nb : le coefficient d'aplatissement d'une distribution uniforme sur $[a; b]$ vaut 1.8, $\forall a, \forall b > a$)

Principes généraux IV

Solution

(i) La distribution d'échantillonnage de la moyenne donne

$$E\left[\overline{X}\right] = m = 0 \quad ; \quad var\left[\overline{X}\right] = \frac{\sigma^2}{n} = \frac{1}{4}\frac{2^2}{12} = 0.0833$$

La distribution de la moyenne peut être approchée par une distribution normale de moyenne nulle et d'écart-type égal à 0.2887. La probabilité de dépassement de 0.5 vaut donc

$$prob\left(\overline{X} \geq 0.5\right) = 1 - \Phi\left(\frac{0.5}{0.2887}\right) = 4.16\%$$

(ii) La distribution d'échantillonnage de la variance donne

$$E\left[S^2\right] = \frac{n-1}{n}\sigma^2 = \frac{7}{8}\frac{2^2}{12} = 0.2917$$

$$var\left[S^2\right] = \frac{n-1}{n^2}\frac{\gamma_4\left(n-1\right) - n + 3}{n}\sigma^4 = 0.01155$$

Principes généraux V

La distribution de la variance peut être approchée par une distribution normale de moyenne 0.2917 et d'écart-type égal à 0.1074. La probabilité de dépassement de 0.5 vaut donc

$$prob\left(S^2 \geq 0.5\right) = 1 - \Phi\left(\frac{0.5 - 0.2917}{0.1074}\right) = 2.62\%$$

Principes généraux VI

▷ **La connaissance empirique des distributions d'échantillonnage**
(i) Lorsque la détermination de la distribution d'échantillonnage est difficile sur des bases théoriques, on peut en obtenir une estimation sur base des méthodes de simulation (simulation de *Monte Carlo*). → génération de séries de nombres aléatoires et utilisation de la statistique descriptive.

(ii) Lorsque la distribution de la population-parent n'est pas connue, il est possible d'utiliser des ***méthodes de rééchantillonnage*** (p.ex. Jackknife, Bootstrap)

Dans la méthode du Jackknife, on élimine chacun des individus constituant l'échantillon initial de sorte à former n sous-échantillons d'effectif $n-1$. On calcule ensuite la valeur du paramètre γ pour ces n sous-échantillons et on étudie ainsi la distribution des valeurs obtenues.

Dans la méthode du Bootstrap, on forme à partir d'un échantillon d'effectif n une série de sous-échantillons d'effectif n (également !), en

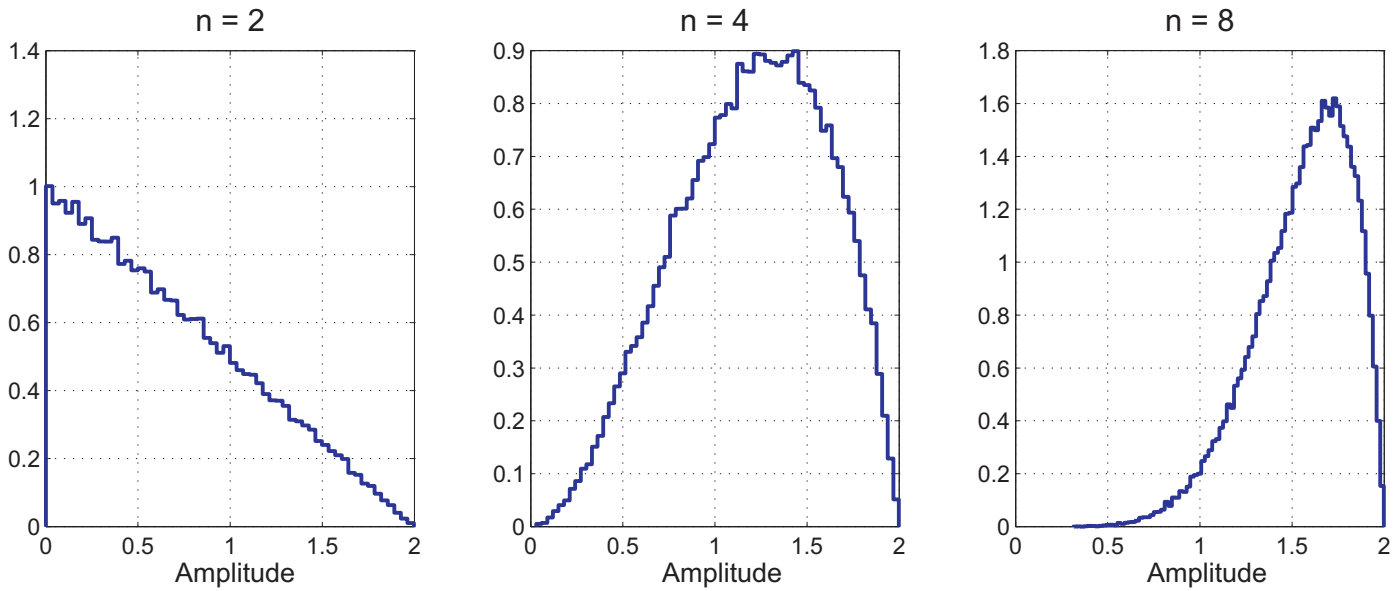
Principes généraux VII

procédant à un tirage aléatoire, **avec** remise. On calcule ensuite la valeur du paramètre γ pour les sous-échantillons formés et on étudie ainsi la distribution des valeurs obtenues.

! grande dépendance vàv de la qualité de l'échantillon initial !

Principes généraux VIII

Exemple : utilisation d'une méthode de simulation pour représenter la distribution d'échantillonnage de l'amplitude (max-min) d'un échantillonnage de n valeurs uniformément choisies entre -1 et 1.



calculé par simulation de Monte-Carlo ($N = 10^5$).

Principes généraux IX

▷ **L'importance de la notion de distribution d'échantillonnage**

Toute personne qui travaille par enquête non exhaustive, ou expérimentation, doit être consciente que les résultats qu'elle obtient ne sont pas une image intangible de l'ensemble des individus auxquels elle s'intéresse.

◦ Une répétition de la même enquête ou de la même expérience dans des conditions identiques devrait toujours conduire, tout naturellement, à des résultats différents de ceux de la première enquête ou expérience.
◦ il est bon d'avoir une certaine connaissance de l'ordre de grandeur des différences qui peuvent exister entre les résultats de plusieurs répétitions éventuelles d'une même enquête, ou d'une même expérience, réalisée dans les mêmes conditions ($\rightarrow$ utiliser les distributions d'échantillonnage).

Principes généraux X

En termes de coefficients de variations,

$$CV_{\overline{X}} = \frac{\sigma_{\overline{X}}}{E\left[\overline{X}\right]} = \frac{\sigma/\sqrt{n}}{m} = \frac{CV_X}{\sqrt{n}}$$

où $CV_{\overline{X}}$ est la variabilité de la moyenne d'un échantillon et CV_X est la variabilité de la variable étudiée.
Si la population-parent possède une distribution normale,

$$CV_{S^2} = \frac{\sigma_{S^2}}{E\left[S^2\right]} = \frac{\frac{\sqrt{2(n-1)}}{n}\sigma^2}{\frac{n-1}{n}\sigma^2} = \sqrt{\frac{2}{n-1}}$$

$$CV_S = \frac{\sigma_S}{E\left[S\right]} \simeq \frac{1}{\sqrt{2n-3}} \simeq \frac{1}{\sqrt{2n}}$$

Exemple
Soit un échantillon aléatoire et simple d'effectif $n = 10$ à 100, issu d'une population supposée normale. Le coefficient de variation de la population est $CV_X = 10\%$ (usuel).
Le coefficient de variation de l'échantillon vaudra 3.16% (resp. 1%) pour les échantillons d'effectifs $n = 10$ (resp. $n = 100$)

$$CV_{\overline{X}} = \frac{0.10}{\sqrt{10}} = 3.16\% \quad CV_{\overline{X}} = \frac{0.10}{\sqrt{100}} = 1.00\%$$

Lois des Grands Nombres I

▷ La loi (faible) des grands nombres

On s'attend à ce que les écarts $\overline{X} - m$ soient d'autant plus petits que les effectifs des échantillons sont grands. On peut démontrer (Bienaymé-Tchebytchev) que

$$\lim_{n \to +\infty} P\left(\left|\overline{X} - m\right| \leq \varepsilon\right) = 0$$

pour toute valeur de ε, aussi petite soit-elle.

Exercices I

1. Que valent la moyenne et l'écart-type des résultats qu'on peut obtenir, si on choisit au hasard et indépendamment 10 nombres entiers pouvant aller de 1 à 9 et qu'on calcule la moyenne des 10 nombres choisis ?
[un même nombre peut être choisi plusieurs fois et tous les nombres ont la même probabilité d'être choisis]
Quelle est approximativement la probabilité que la moyenne des 10 nombres choisis soit supérieure à 6 ?

2. L'âge d'une grande population est représenté par une distribution normale de moyenne 65 ans et d'écart-type 18 ans. On interroge 20 personnes, choisies au hasard par un échantillonnage complètement aléatoire.
2.1. Que valent la moyenne et l'écart-type de l'âge moyen dans l'échantillon ?
2.2. Que valent la moyenne et l'écart-type de la variance dans l'échantillon ?
2.3. Quelle est la probabilité que l'âge moyen d'un groupe de 30 personnes choisies au hasard soit supérieur à 70 ans ?
2.4. Combien de personnes faut-il sélectionner pour que la moyenne de leur âge soit supérieure à 40 ans, avec 1% de défaillance ?

Calculez les solutions des problèmes puis validez et illustrez vos résultats à l'aide de Matlab (par simulation).

Exercices II

3. *[Jui2013]* Des statistiques sur un très grand nombre de fournisseurs indiquent que le prix de vente du mètre cube de béton est une variable aléatoire lognormale de moyenne 100€ et d'écart-type 22€.
Afin de réduire ses chances de passer commande à la centrale la plus chère (on imagine qu'il n'est pas possible d'obtenir de devis), un architecte décide, pour chaque nouveau chantier, de passer commande en part égale à deux centrales choisies arbitrairement parmi ce grand nombre de fournisseurs.
Après un grand nombre de chantiers réalisés,
- quel est le prix moyen qu'il a payé pour un mètre cube de béton ?
- quel est la variabilité du prix payé autour de cette moyenne ?
- aurait-il été plus intéressant de faire la même chose, mais en répartissant sa commande entre 4 fournisseurs ? Comment changent dans ce cas les estimations ci-dessus ?

4. Chaque année, le jogging des 10km de Liège attire de nombreux participants. Cette année, la vitesse des coureurs avait une distribution normale, avec une moyenne de 11,45 km/h et un écart-type de 1,7 km/h.
(i) quelle est la probabilité que 15 coureurs pris au hasard sur les 500 participants courent en moyenne à plus de 14km/h ?
(ii) quelle est la probabilité que la vitesse moyenne de 3 coureurs pris au hasard soit supérieure à 12 km/h ?

Exercices III

(iii) quelle est la probabilité que la variance des vitesses de 3 coureurs pris au hasard soit supérieure à 2 $(\text{km/h})^2$?

Réponses

2. (1) moy=65 ans, std=4.02 ans ; (2) moy=308 ans^2, std=99.9 ans^2 ; (3) prob=6.43% ; (4) $n = 3$ personnes

2.2

2.3

2.4

3

4

Les problèmes d'estimation I

Objectif : estimer, à partir d'un échantillon, les valeurs numériques des paramètres de la population (moyenne, variance, etc.). Comment chiffrer, en probabilité, l'estimation que l'on peut donner.

▷ **Estimation de la moyenne**

La meilleure estimation de la moyenne m d'une population qui puisse être déduite d'un échantillon aléatoire et simple, est (à première vue), la moyenne $\overline{x}$ de l'échantillon :

$$\hat{m} = \overline{x}$$

En effet, pour l'ensemble des échantillons qui pourraient être rencontrés, on doit s'attendre à retrouver, en moyenne, la "vraie" valeur de la population

$$E\left[\overline{X}\right] = m$$

nb : la dispersion des différentes estimations de la moyenne, autour de cette moyenne m, est mesurée par l'erreur-standard de la moyenne $\sigma_{\overline{X}} = \sigma/\sqrt{n}$.

Les problèmes d'estimation II

Autres estimateurs de la moyenne de population

(i) On pourrait choisir d'utiliser la médiane de l'échantillon $\tilde{x}$ (au lieu de la moyenne), comme estimateur de la moyenne de la population

$$\hat{m} = \tilde{x}$$

On retrouve également, en moyenne, la vraie valeur de la moyenne de la population (car $E\left[\tilde{X}\right] = m$). Cependant la dispersion est plus importante car $\sigma_{\tilde{X}} \simeq \sigma\sqrt{\pi/(2n)} > \sigma/\sqrt{n}$ (si $n > 2$).
$\rightarrow$ l'estimation $\hat{m} = \overline{x}$ est plus efficace que $\hat{m} = \tilde{x}$

En d'autres mots, pour avoir la même erreur-standard, il faudrait avoir un échantillon d'effectif n' plus grand lorsqu'on utilise la médiane comme estimateur

$$\frac{\sigma^2}{n} = \frac{\pi\,\sigma^2}{2n'} \qquad \rightarrow \quad n' = \frac{\pi}{2}n \simeq 1.57\,n$$

Les problèmes d'estimation III

(ii) On pourrait utiliser, comme estimation de la moyenne de population, la moyenne des valeurs extrémales

$$\hat{m} = \frac{x_{min} + x_{max}}{2} = \frac{x_1 + x_n}{2}$$

(***méthode du mid-range***). Cette méthode est cependant fortement affectée par la présence d'*outliers* dans les mesures.

(iii) A la place, on peut utiliser ***une moyenne élaguée***, définie par

$$\hat{m} = \frac{1}{n-2} \sum_{i=2}^{n-1} x_i$$

▷ **Estimation de la variance**

On pourrait croire que la meilleure estimation de la variance σ^2 d'une population est simplement donnée par la variance s^2 d'un échantillon... mais non ! On obtiendrait sinon, en moyenne, une valeur différente de la variance de la population, puisque

$$E\left[S^2\right] = \frac{n-1}{n}\sigma^2.$$

On corrige cette erreur ***systématique*** (un biais) en définissant

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

de sorte que l'espérance mathématique soit bien égale à σ^2. Dans le cas d'une population normale, l'erreur-standard de cette estimation est

$$\sqrt{\operatorname{var}\left[\frac{nS^2}{n-1}\right]} = \frac{n}{n-1}\sqrt{\operatorname{var}\left[S^2\right]} = \frac{n}{n-1}\frac{\sqrt{2\,(n-1)}}{n}\sigma^2 = \sqrt{\frac{2}{n-1}}\sigma^2.$$

Les problèmes d'estimation V

Rem : $\hat{\sigma}$ n'est pas nécessairement une bonne estimation de l'écart-type. On peut montrer que les résultats obtenus produisent systématiquement un écart-type trop faible (de l'ordre de 6% pour 5 individus, de l'ordre de 3% pour 10 individus, de l'ordre de 1% pour 20 à 30 individus)

Les problèmes d'estimation VI

▷ Principes généraux de l'estimation

Considérons une population quelconque dont la distribution de probabilité dépend d'un paramètre γ, ainsi qu'un échantillon d'effectif n de cette population. On appelle ***estimateur*** du paramètre γ, toute fonction G des valeurs observées (ou de certaines de ces valeurs), susceptible de servir à estimer γ par

$$G = G(X_1, X_2, ...X_n)$$

∘ la première qualité d'un "bon" estimateur est l'absence d'erreur systématique, ou de biais. Donc, en moyenne, on doit idéalement retrouver

$$E[G] = \gamma$$

→ estimateur ***non biaisé*** (ou ***impartial***).
Exemple : la moyenne d'échantillon est un estimateur impartial de la moyenne de la population ; la variance d'échantillon est un estimateur biaisé de la variance de la population.

Les problèmes d'estimation VII

◦ la seconde qualité d'un "bon" estimateur est de posséder une précision suffisante. Cette précision est mesurée par le moment d'ordre 2 par rapport à la valeur théorique

$$E\left[(G-\gamma)^2\right]$$

Pour des estimateurs non biaisés, ce moment se confond avec la variance de la distribution d'échantillonnage

$$E\left[(G-\gamma)^2\right] = E\left[(G-m_G)^2\right] = \sigma_G^2$$

Exemple : la moyenne d'échantillon est un estimateur plus précis que la médiane.

◦ la précision des estimateurs est cependant limitée. A tout paramètre γ correspond une valeur minimum de $E\left[(G-\gamma)^2\right]$ en-dessous de laquelle il est impossible de descendre (quelle que soit la fonction G choisie). L'estimateur G qui permet de minimiser cette variance est dit ***de variance minimum***.

Les problèmes d'estimation VIII

Dans le cas d'estimateur non biaisés, on peut démontrer que cette variance minimum est

$$\frac{1}{n\int_{-\infty}^{+\infty}\left(\frac{\partial \ln f(x;\gamma)}{\partial \gamma}\right)^2 f(x;\gamma)\,dx}$$

(cette valeur ne dépend donc que de la distribution de la population et de l'effectif d'échantillon).

Un but poursuivi consiste donc à trouver un estimateur G qui permet d'atteindre cette valeur minimale, mais cette solution n'existe pas toujours.
A défaut, certains estimateurs sont asymptotiquement de variance minimum, càd que cette variance minimum peut être atteinte pour une taille d'échantillon tendant vers l'infini.

∘ une autre qualité d'un "bon" estimateur est d'être relativement insensible aux valeurs anormales (outliers).
Exemple : la moyenne des deux valeurs extrémales (max et min) d'échantillon est plus sensible aux outliers que la médiane d'échantillon.

Les problèmes d'estimation IX

Exemple : Estimation de la variance d'une distribution normale
Puisque la distribution de probabilité est donnée par

$$f\left(x,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}},$$

la variance minimale vaut

$$\frac{1}{n\int_{-\infty}^{+\infty}\left(\frac{\partial \ln f(x;\gamma)}{\partial\gamma}\right)^2 f\left(x;\gamma\right)dx}=\frac{2\sigma^4}{n}.$$

L'estimateur

$$\hat{\sigma}^2=\frac{n}{n-1}S^2$$

n'est qu'asymptotiquement efficace puisque

$$\text{var}\left[\hat{\sigma}^2\right]=\left(\frac{n}{n-1}\right)^2\text{var}\left[S^2\right]=\left(\frac{n}{n-1}\right)^2\frac{2\left(n-1\right)\sigma^4}{n^2}=\frac{2\sigma^4}{n-1}.$$

Méthode du Maximum de Vraisemblance I

La méthode du maximum de vraisemblance est une méthode qui permet de déterminer les estimateurs qui possèdent les qualités discutées avant.
Elle consiste à choisir, comme estimation de tout paramètre γ, la valeur la plus vraisemblable, çàd celle qui a la plus forte probabilité de provoquer l'apparition des valeurs réellement observées.

La fonction de vraisemblance est la probabilité relative aux valeurs observées $x_1, ..., x_n$, exprimée en fonction du paramètre γ de la distribution de la population. Pour un échantillon aléatoire et simple, la fonction de vraisemblance s'écrit

$$L(\gamma; x_1, ...x_n) = f(x_1; \gamma) f(x_2; \gamma) ...f(x_n; \gamma)$$

Les estimateurs du maximum de vraisemblance correspondent au maximum de cette fonction (définition). On les obtient en résolvant

$$\frac{\partial L}{\partial \gamma} = 0$$

Méthode du Maximum de Vraisemblance II

ou en annulant la dérivée de son logarithme

$$\frac{\partial \ln L}{\partial \gamma} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\gamma)}{\partial \gamma} = 0$$

a. On peut démontrer que cette méthode fournit toujours des estimateurs de variance minimum.

b. La distribution des estimateurs obtenus est toujours asymptotiquement normale.

c. Les estimateurs obtenus ne sont pas toujours non biaisés.

Méthode du Maximum de Vraisemblance III

Exemple : Estimation de la moyenne et de la variance d'une distribution normale
Puisque la distribution de probabilité est donnée par

$$f\left(x,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}},$$

$$\begin{aligned}\ln L=\sum_{i=1}^{n}\ln f\left(x_i;\gamma\right) &= -\sum_{i=1}^{n}\left[\frac{1}{2}\ln\left(2\pi\sigma^2\right)+\frac{\left(x_i-m\right)^2}{2\sigma^2}\right]\\ &= -\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2}\sum_{i=1}^{n}\frac{\left(x_i-m\right)^2}{\sigma^2}\end{aligned}$$

et donc l'annulation de la dérivée par rapport à la moyenne donne

$$\frac{\partial \ln L}{\partial m}=\sum_{i=1}^{n}\frac{\left(x_i-m\right)}{\sigma^2}=0$$

soit,

$$\hat{m}=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

Méthode du Maximum de Vraisemblance IV

L'annulation de la dérivée par rapport à la variance donne

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m)^2}{\sigma^4} = 0$$

soit,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 .$$

Méthode des Moments I

La méthode des moments est une autre approche permettant de déterminer les estimateurs pour les k paramètres d'une distribution d'une population. Elle consiste à imposer l'égalité entre les k premiers moments estimés de la population (exprimés en fonction de k paramètres) aux k premiers moments de l'échantillon.

En termes de moments non centrés, on obtient ainsi un système d'équations

$$\begin{cases} \hat{\alpha}_1(\hat{\gamma}_1, ..., \hat{\gamma}_k) & = a_1 \\ \vdots & \\ \hat{\alpha}_k(\hat{\gamma}_1, ..., \hat{\gamma}_k) & = a_k \end{cases}.$$

a. La distribution des estimateurs obtenus est toujours asymptotiquement normale.

[Exemples]

Intervalles de confiance I

Principe de l'estimation :

Objectif: estimer, à partir d'un échantillon, les valeurs numériques des paramètres de la population (moyenne, variance, etc.). Comment chiffrer, en probabilité, l'estimation que l'on peut donner.

Donner la valeur estimée ne suffit pas. Compléter par :
- l'erreur-standard ;
- la détermination d'un intervalle (autour de la moyenne) dont on a de bonnes raisons de croire qu'il contient la "vraie" valeur du paramètre recherché → ***intervalle de confiance***

Soit le paramètre γ et l'estimateur G. On cherche $G_1 \leq G \leq G_2$ tels que l'intervalle de confiance $[G_1; G_2]$ a une forte probabilité de contenir γ. On se donne un ***degré de confiance*** (ou niveau de confiance) proche de l'unité, noté $1 - \alpha$ (avec $\alpha \ll 1$) et on détermine G_1 et G_2 de façon à ce que

$$P(G_1 \leq \gamma \leq G_2) = 1 - \alpha.$$

Souvent, $\alpha = 0.05$, $\alpha = 0.01$ ou $\alpha = 0.001$.

Intervalles de confiance II

Donc, en affirmant que l'intervalle $[G_1; G_2]$ contient la vraie valeur de γ, on commet une erreur dont la probabilité est α.

Attention : le risque total α peut être réparti d'une infinité de façons différentes (une équation, deux inconnues).

$$P(\gamma < G_1) + P(G_2 < \gamma) = \alpha.$$

Souvent, on divise le risque en deux parties égales

$$P(\gamma < G_1) = P(G_2 < \gamma) = \frac{\alpha}{2}.$$

Intervalle de confiance de la moyenne (d'une population normale)

Nous avons déjà un estimateur $\overline{X}$ de la moyenne. L'intervalle de confiance de la moyenne $\left[\overline{X}_1; \overline{X}_2\right]$ est défini par

$$P\left(m < \overline{X}_1\right) = P\left(\overline{X}_2 < m\right) = \frac{\alpha}{2}.$$

Soit

$$\overline{X}_1 = \overline{X} - d_1 \qquad ; \qquad \overline{X}_2 = \overline{X} + d_2$$

On cherche donc d_1 et d_2 tels que

$$P\left(\overline{X} - m > d_1\right) = P\left(m - \overline{X} > d_2\right) = \frac{\alpha}{2}.$$

[nb : pour une population normale, on sait que $\overline{X}$ est une variable aléatoire normale de moyenne m et d'écart-type $\sigma/\sqrt{n}$]

Dans le cas d'une population normale, la variable réduite

$$U = \frac{\overline{X} - m}{\sigma/\sqrt{n}}$$

Intervalles de confiance IV

est une variable normale réduite. On cherche donc d_1 et d_2 tels que

$$P\left(U > \frac{d_1}{\sigma/\sqrt{n}}\right) = P\left(U < \frac{-d_2}{\sigma/\sqrt{n}}\right) = \frac{\alpha}{2}.$$

Si on note $u_{1-\alpha/2}$ le nombre tel que $\Phi\left(u_{1-\alpha/2}\right) = 1 - \alpha/2$.

$$\frac{d_1}{\sigma/\sqrt{n}} = \frac{d_2}{\sigma/\sqrt{n}} = u_{1-\alpha/2} \quad \rightarrow \quad d_1 = d_2 = u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

L'intervalle de confiance est donc

$$I = \left[\hat{m} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \hat{m} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

L'écart-type σ de la population-parent n'est pas toujours connu. On peut le remplacer par son estimation $\hat{\sigma}$ (si $n \gtrsim 30$) de sorte que

$$I = \left[\hat{m} - u_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}; \hat{m} + u_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

Quelques remarques générales

Les développements peuvent être étendus facilement aux intervalles de confiance de n'importe quel paramètre γ de distribution normale (ou asymptotiquement normale) et dont l'erreur standard σ_G est connue :

$$I = \left[\hat{g} - u_{1-\alpha/2}\sigma_G ; \hat{g} + u_{1-\alpha/2}\sigma_G\right]$$

◦ Lorsque l'erreur standard σ_G n'est pas connue, on peut éventuellement obtenir une estimation en la remplacant par $\hat{\sigma}_G$, ce qui est valable lorsque l'effectif n est grand (! pas toujours 30, cela dépend de la distribution).

◦ Lorsque la distribution du paramètre γ n'est pas normale, on peut reproduire le même raisonnement (ex : intervalle de confiance de la variance d'une population normale).

Intervalles de confiance VI

Attention à l'interprétation erronnée. L'intervalle de confiance :
- N'EST PAS l'intervalle tel qu'il y ait une probabilité $1-\alpha$ d'y trouver la moyenne de la population,
- est tel que, en moyenne, si on répétait l'opération d'échantillonnage un grand nombre de fois, la vraie valeur de la moyenne aurait une probabilité $1-\alpha$ de se trouver dans l'intervalle.

Estimation du Nombre d'observations I

Question : quel est le nombre minimum d'observations à réaliser pour atteindre, lors de l'estimation des paramètres, une précision donnée, avec un *degré de confiance* $1-\alpha$?

Dans le cas de la moyenne, on choisit n de façon à ce que l'écart $\left|\overline{X}-m\right|$ ne dépasse une valeur fixée d (choisie a priori) qu'avec une probabilité α, càd

$$d = u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \rightarrow \qquad n = u_{1-\alpha/2}^2\frac{\sigma^2}{d^2}$$

Exemple, si $\alpha = 0.05$, il faut que $n \geq 3.84\sigma^2/d^2$.

nb : on peut remplacer σ par $\hat{\sigma}$.

Exercices I

1. Les résultats au test d'anglais de base en première année d'ingénieur sont distribués suivant une loi normale de moyenne $m = 62$ et d'écart-type $\sigma = 25$. On fait passer le test à 30 nouveaux étudiants. Quelle est la probabilité pour que la moyenne de l'échantillon soit comprise entre 70 et 80 [%] ?

2. La taille d'un ensemble de gymnastes est représenté par une distribution normale de moyenne 155cm et d'écart-type 20cm. On sélectionne 30 gymnastes via un échantillonnage aléatoire. Déterminez
(i) la moyenne et l'écart type de la taille moyenne de l'échantillon
(ii) la moyenne et l'écart type de la variance de l'échantillon
(iii) la probabilité que la taille moyenne d'un groupe de 10 personnes soit inférieure à 130cm
(iv) admettons que la moyenne de la population (155cm) soit inconnue, mais que l'on désire l'estimer à partir d'un échantillon. Quelle est la taille de l'échantillon à considérer pour que l'on obtienne un intervalle de confiance de 10cm (avec un degré de confiance de 0.95) ?

Exercices II

Réponses

1. $p = 4,002\%$.
2. (i) $E\left[\overline{X}\right] = 155$ cm ; $\sigma_{\overline{X}} = 3.651$ cm ; (ii) $E\left[S^2\right] = 386.7$ cm^2 ; $\sigma_{S^2} = 101.5$ cm^2 ; (iii) $p = 4 \cdot 10^{-5}$; (iv) $n \geq 62$.

Principes généraux I

Test d'hypothèse ou ***test de signification :*** vérification, à partir des données d'un ou de plusieurs échantillons, la validité de certaines hypothèses relatives à une (ou plusieurs) populations.

Il existe différents tests :

▷ ***test d'ajustement*** : vérification si un échantillon observé peut être considéré comme extrait d'une population donnée.
ex : une variable observée/sondée est-elle extraite d'une population qui possède une distribution normale ? - voir si les écarts entre les fréquences relatives observées et la densité de probabilité normale peuvent être dus au hasard de l'échantillonnage, ou s'ils doivent être attribués à d'autres facteurs

▷ ***test d'indépendance*** : contrôler, à partir d'un échantillon, l'indépendance stochastique de deux ou plusieurs critères de classification (généralement qualitatifs).
ex : le type de profession au sein d'une population est-il indépendant de la situation d'état civil ?

▷ ***test de conformité*** : visent également à comparer un échantillon observé et une population théorique, mais dans un but plus restreint : il

Principes généraux II

vérifie si un échantillon donné peut être considéré comme extrait d'une population possédant (non pas une distribution entièrement spécifiée mais) seulement une moyenne donnée ou une variance donnée, ou tout autre paramètre.
(souvent utilisé pour vérifier si un échantillon prélevé satisfait à des prescriptions normatives)
→ test de conformité de moyenne, test de conformité de variance, test de conformité de coefficient de corrélation, etc.
On vérifie si la différence entre la valeur observée et la valeur théorique peut être attribuée au hasard de l'échantillonnage ou non.

▷ ***test d'égalité ou d'homogénéité*** : comparer entre elles différentes populations, à l'aide d'un même nombre d'échantillons.
→ test d'égalité de moyenne, test d'égalité de variance, test d'égalité de coefficient de corrélation, etc.

Principes généraux III

On vérifie si la différence entre les valeurs observées (moyennes, p.ex.) de deux séries de mesures peut être attribuée au hasard de l'échantillonnage ou non.

Principes généraux IV

Principes

1. Emettre un hypothèse à tester (***hypothèse nulle*** H_0)

2. On mesure *l'écart observé* entre (i) certaines caractéristiques de l'échantillon et de la population (ajustement et conformité) ou (ii) entre certaines caractéristiques dans différents échantillons (égalité).

3. On calcule la probabilité d'observer un écart aussi important, en supposant que l'hypothèse nulle est vraie.

Si cette probabilité est élevée, on considère l'hypothèse comme plausible et on l'accepte (au moins provisoirement).
Si cette probabilité est faible, i.e. inférieure à un ***niveau de signification*** préalablement fixé, l'écart observé apparaît comme peu compatible avec l'hypothèse nulle, et on la rejette.

Principes généraux V

Test d'hypothèse ≡ sorte de démonstration par l'absurde en probabilité. On rejette éventuellement l'hypothèse émise au départ, parce qu'on arrive à une situation peu vraisemblable.

Le hasard de l'échantillonnage peut évidemment fausser les conclusions et quatre possiblités doivent être envisagées :
- accepter l'hypothèse nulle alors qu'elle est vraie (ok),
- rejeter l'hypothèse nulle alors qu'elle est vraie (α-error),
- accepter l'hypothèse nulle alors qu'elle est fausse (β-error),
- rejeter l'hypothèse nulle alors qu'elle est fausse (ok).

α-error : ***erreur de première espèce*** $RH_0|H_0$

$$\alpha = P(RH_0|H_0)$$

β-error : ***erreur de seconde espèce*** $AH_0|H$

$$\beta = P(AH_0|H)$$

Principes généraux VI

En pratique, on se donne également une limite supérieure du risque de première espèce α (le plus souvent 5%, 1% ou 0.1%). [nb : cette limite est également le niveau de signification du test, qui permet de définir la ***condition de rejet*** de l'hypothèse nulle.]

TEST D'ÉGALITÉ DE MOYENNES I

Supposons que nous ayons deux populations normales, de moyennes inconnues mais d'écart-types identiques que l'on peut supposer être égaux à σ. On prélève un échantillon dans chaque population et on utilise le résultat de cet échantillonnage pour déterminer si les moyennes des deux populations ont des raisons d'être identiques.
Soit m_1 et m_2 les moyennes des deux populations.

L'hypothèse nulle s'écrit

$$H_0 : m_1 = m_2 \qquad \text{ou} \qquad H_0 : m_1 - m_2 = 0$$

On note $\overline{x}_1$ et $\overline{x}_2$ les moyennes des deux échantillons (d'effectif n chacun).
[l'hypothèse nulle doit être acceptée si les deux moyennes observées sont proches l'une de l'autre ; elle doit être rejetée si les deux moyennes observées sont fort différentes]
Le problème revient à fixer la limite de rejet, en fonction du niveau de signification choisi.

TEST D'ÉGALITÉ DE MOYENNES II

Les moyennes calculées sont des valeurs observées de deux variables aléatoire $\overline{X}_1$ et $\overline{X}_2$. Ces variables sont normales (moyennes m_1 et m_2, variance σ^2/n) et indépendantes.

La différence $\overline{X}_1 - \overline{X}_2$ est donc une variable aléatoire normale de moyenne $m_1 - m_2$ et de variance $2\sigma^2/n$. Si l'hypothèse nulle est vraie, la variable

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{2/n}}$$

possède donc une distribution normale réduite. La probabilité d'observer une différence de moyennes au moins égale à $\overline{x}_1 - \overline{x}_2$ (en valeur absolue), s'écrit

$$\begin{aligned} P\left(\left|\overline{X}_1 - \overline{X}_2\right| \geq |\overline{x}_1 - \overline{x}_2|\right) &= P\left(|U| \geq \frac{|\overline{x}_1 - \overline{x}_2|}{\sigma\sqrt{2/n}}\right) \\ &= 2\left[1 - \Phi\left(\frac{|\overline{x}_1 - \overline{x}_2|}{\sigma\sqrt{2/n}}\right)\right] \end{aligned}$$

TEST D'ÉGALITÉ DE MOYENNES III

On rejette l'hypothèse nulle si cette probabilité est inférieure ou égale au niveau de signification α choisi.

nb : puisque σ est inconnu, on peut le remplacer par $\hat{\sigma} = \sqrt{\left(\hat{\sigma}_1^2 + \hat{\sigma}_2^2\right)/2}$.

Exemple I

On désire comparer les revenus des personnes physiques en Flandre et en Wallonie. Pour ce faire, on a choisi un échantillon de 50 personnes dans chaque communauté. Si on obtient pour ces deux échantillons des revenus moyens de 38900€ (en Flandre) et 37900€ (en Wallonie), peut-on admettre qu'il existe réellement une différence de revenu entre les deux communautés ?
nb : les écart-types calculés dans les deux cas sont inconnus mais peuvent être estimés à partir d'un coefficient de variation de 10%.

Hypothèse $H_0 : m_1 = m_2$

$$\hat{\sigma}_1 = 0.1 \times 38900 = 3890€ \qquad \hat{\sigma}_2 = 0.1 \times 37900 = 3790€$$

Donc $\hat{\sigma} = 3840€$. On a donc

$$P\left(\left|\overline{X}_1 - \overline{X}_2\right| \geq 1000\right) = P\left(|U| \geq \frac{1000}{3840\sqrt{2/50}}\right) = P\left(|U| \geq 1.3020\right) = 0.1868$$

Exemple II

Il y a donc une probabilité de 18.7% d'obtenir, par le simple fait du hasard de l'échantillonnage, une différence de moyenne supérieure à 1000€.
Nous sommes donc amenés à accepter l'hypothèse nulle. (Les deux revenus ne diffèrent pas significativement)

Même question, mais supposons que nous ayons réalisé les mesures sur des échantillons d'effectif $n = 200$ et non plus $n = 50$.

$$P\left(|\overline{X}_1 - \overline{X}_2| \geq 1000\right) = P\left(|U| \geq \frac{1000}{3840\sqrt{2/200}}\right) = P\left(|U| \geq 2.6039\right) = 0.0092$$

L'hypohèse nulle doit donc être acceptée au niveau de signification $\alpha = 0.001$, mais rejettée au niveau de signification $\alpha = 0.01$.

Exercices I

1. En vue de comparer deux méthodes d'enseignement, on a réparti, de façon complètement aléatoire, 50 étudiants en deux classes de 25 étudiants, et on a appliqué les deux méthodes d'enseignement séparément à chacune des deux classes. On a ensuite soumis l'ensemble des élèves à un même examen et on a obtenu les résultats suivants : $\overline{x}_1 = 12$, $\sigma_1^2 = 9.56$, $\overline{x}_2 = 13.40$, $\sigma_2^2 = 8.40$.
Faut-il considérer qu'en moyenne, la deuxième méthode d'enseignement donne de meilleurs résultats que la première ?
[utilisez un niveau de signification α =1%]

Réponse : non

Commandes Matlab I

Statistiques descriptives

`find(x,k)`: recherche des données dans un vecteur
`cumsum(x)`: somme cumulée
`mean(x)`: moyenne arithmétique du vecteur x
`geomean(x)`: moyenne géométrique des données du vecteur x
`harmmean(x)`: moyenne harmonique des données du vecteur x
`median(x)`: médiane des données du vecteur x
`mode(x)`: mode des données du vecteur x

`var(x)`: variance des données du vecteur x
`std(x)`: écart-type des données du vecteur x
`quantile(x,p)`: quartiles des données du vecteur x correspondant aux fractions cumulées `p`
`prctile(x,p)`: quartiles des données du vecteur x correspondant aux fractions cumulées `p` (exprimées entre 0 et 100)
`moment(x,k)`: moment centré d'ordre `k` des données du vecteur x
`min(x)`: minimum des données du vecteur x
`max(x)`: maximum des données du vecteur x
`sort(x)`: données du vecteur x triées depuis la plus petite vers la plus grande

`hist(x)`: histogramme du vecteur x (données brutes)
`bar(x)`: diagramme en bâtonnets (x = fréquences de classes)
`boxplot(x)`: diagramme en boxplot des données brutes x

Commandes Matlab II

`boxplot(x,'whisker',w)`: diagramme en boxplot des données brutes x (gestion des données aberrantes)
`hist3(x,y)`: histogramme des données appariées x et y
`plot(x,y)`: graphe de la fonction explicite y(x)
`stem(x,y)`: graphe de la fonction explicite y(x) de type bâtonnets avec une bulle
`surf(x,y,z)`: graphe de la fonction explicite z(x,y)

`corrcoef(X)`: coefficient de correlation
`cov(X)`: covariance
`polyfit(X)`: ajustement d'un polynome

Probabilités

`rand(N,1)`: génération d'un vecteur de `N` valeurs distribuées uniformément
`randn(N,1)`: génération d'un vecteur de `N` valeurs distribuées normalement

`pdf(name,X,A)`: densité de probabilité de différentes distributions
`cdf(name,X,A)`: fonction de répartition de différentes distributions
`erfinv(X)`: fonction inverse de Φ (fct répartition normale réduite)

Lexique des définitions

Les définitions et termes à connaître sont indiqués en ***magenta*** dans les transparents de la présentation.

Lectures complémentaires I

P. Dagnelie
Statistique théorique et appliquée
de Boeck, 2nd édition, 1998.